CONTRIBUTORS

Opening Lecture

P. B. HIRSCH

Electron Optics and Instrumentation

A. SEPTIER
R. CASTAING
A. V. CREWE
U. VALDRÈ
A. KÜBLER

Diffraction Contrast and Applications

A. HOWIE
R. GEVERS
L. M. BROWN
M. J. MAKIN
M. J. GORINGE
C. R. HALL

Transfer of Image Information and Phase Contrast

F. LENZ
F. THON
A. C. VAN DORSTEN
C. R. HALL
R. H. WADE
D. WOHLLEBEN

ELECTRON MICROSCOPY IN MATERIAL SCIENCE

"Ettore Majorana" International Centre for Scientific Culture
1970 International School of Electron Microscopy
a NATO Advanced Study Institute
Sponsored by CNR-MPI
Erice, April 4-18

EDITOR
U. VALDRÈ

1971

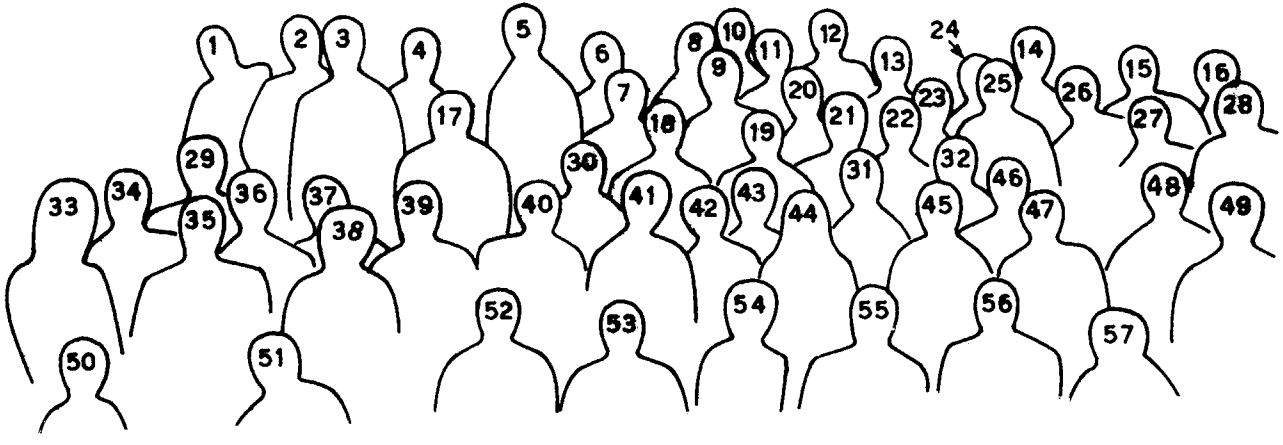
ACADEMIC PRESS NEW YORK AND LONDON

ACADEMIC PRESS INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS INC. (LONDON) LTD.
Berkeley Square House, London W.1

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 70-157136

PRINTED IN ITALY



1. D. Wohlleben
2. A. V. Crewe
3. E. Angeli
4. Z. S. Basinski
5. H. P. Karnthaler
6. D. Krahl
7. F. Thon
8. P. Tischer
9. D. Shechtman
10. R. B. Scarlin
11. H. Schultz
12. A. J. Craven
13. C. P. Cutler
14. T. H. Webster
15. M. Sarracino
16. Mrs. M. Sarracino
17. S. Niedzwiedz
18. A. C. van Dorsten
19. L. Stagni
20. J. J. Burton
21. H. A. Ferwerda
22. K. G. Wold
23. C. English
24. Mrs. G. Servi
25. A. Stern
26. G. Pozzi
27. J. Torres
28. F. Guglielmi
29. G. Franceschetti
30. Mrs. F. Pedrielli
31. C. Severin
32. I. Emanuilov
33. Mrs. A. Howie
34. A. Howie
35. G. Servi
36. A. Giarda
37. E. Bonetti
38. Mrs. Lopez de Souza Santos
39. M. Nobile
40. J. Kumar
41. P. H. Gargya
42. B. E. Wynne
43. P. G. Merli
44. Miss I. Salerno
45. S. Mahajan
46. J. H. Evans
47. J. P. Guigay
48. C. Colliex
49. U. Valdrè
50. D. Howie
51. F. Mezzetti
52. Mrs. I. Dumler
53. F. Pedrielli
54. Mrs. J. L. Brown
55. L. M. Brown
56. J. L. Brown
57. Mrs. L. M. Brown

Foreword

The activities of the "Ettore Majorana" International Centre for Scientific Culture started in 1963 with the Courses of the International School of Subnuclear Physics and have since extended to include the Courses of other National and International Schools. This year we have a new international School which has been established in one of the most important and promising fields of modern research, i.e.: electron microscopy.

This new School is, for the Centre, a further step in developing the promotion of international collaboration among scientists actively working in those disciplines which are likely to bring fundamental contributions to our understanding of the world around us. The research work needed to reach this understanding requires a high degree of specialization. But specialization often implies lack of communication even among various branches of the same discipline. It is our duty not to forget that the simplest or most complicated phenomenon is studied in order to acquire a new contribution to the universality of our knowledge.

I am sure that the electron microscopists will contribute a great deal to help the Centre in pursuing its aims.

A. ZICHICHI
Director of the Centre

Introduction

Application of electron microscopy in material science has registered a rapid growth in the past ten years and many have felt the need for a school with the following purposes:

i) to bring scientists up to date with a refresher course on the latest developments in conventional and non conventional electron microscopy;

ii) to favour contacts between the three main groups of scientists interested in electron microscopy (*i.e.* instrument designers, experts in electron diffraction and contrast theory, and users of electron microscopes) in order to allow each group to be acquainted with the problems and perspectives in the other fields;

iii) to stimulate discussions on specialized topics which are likely to bring fundamental improvements in electron microscopy.

To meet this need an International School of Electron Microscopy was held in Erice in 1970 and was attended by 65 scientists from 17 different countries.

The programme was mainly devoted to: new developments in electron optics and electron microscopy instrumentation; basic ideas on diffraction contrast and applications (partly as fully solved problems) to material science; recent progress on the transfer of image information and on phase contrast, with particular emphasis on high resolution microscopy and Lorentz microscopy.

The School has proved to be very successful and the many requests received have suggested the publication of the lectures, which are contained in this book. I hope the reader will benefit from this book as the participants did from the Course.

The School was sponsored by the Italian National Research Council (CNR), the Italian Ministry of Public Education (MPI), the North Atlantic Treaty Organization (NATO), the Regional Sicilian Government (ERS) and the Organizing Committee of the Rome Conference for Electron Microscopy. The financial support of these Institutions is gratefully acknowledged.

I wish to thank the lecturers and the many colleagues and friends who helped with encouragement and advice in the organization of the School; I am particularly indebted to Drs. C. Colliex, M. J. Goringe, A. Howie, G. Pozzi and Miss I. Salerno.

THE EDITOR

Bologna. March 1971

The Impact of Transmission Electron Microscopy in the Science of Materials

P. B. HIRSCH

Metallurgy Department, University of Oxford - Oxford, England

1. Historical introduction.

In the 1940's a number of commercial (50 ÷ 100) kV electron microscopes became available with resolution for routine operations of about 100 Å, and down to (25 ÷ 30) Å under optimum operating conditions. These instruments were powerful tools for the biologists, who could (although not without some difficulty) prepare specimens suitable for observation by transmission. The metallurgists and physicists were rather slow in utilizing these new and powerful instruments, mainly because of the difficulties of specimen preparation, and because the studies were in any case limited to surface topography by replica methods; in addition the motivation for electron microscope observation was perhaps rather stronger for the biologists, who had a clear appreciation of the need for studies of ultra-fine structure in biological specimens.

In 1949 Heidenreich showed that it was possible to produce thin sections of aluminium by etching, thin enough for direct observation by transmission; these sections revealed the substructure in cold-worked aluminium. In this classic work Heidenreich showed *a*) that sufficiently thin metal films could be produced, *b*) how thickness and orientation variations give rise to contrast effects, in terms of the dynamical theory of electron diffraction, *c*) that metallurgically significant results could be obtained in this way ([1]).

Although this work showed great promise, there was an incubation period of 7-8 years before the transmission technique was developed further. The reason for this long interval was at least in part due to actual and to some

extent imagined difficulties associated with the preparation of thin specimens; it was commonly thought at the time, it would be impossible to « see through » thicknesses of metal greater than a few 100 Å, and this acted as a deterrent.

Then in the mid-1950's dislocations were observed by transmission electron microscopy (TEM), in stainless steel ([2]) and in aluminium ([3]), and G. P. zones in Al 4 % Cu ([4]). A number of factors contributed to the very rapid development of TEM from this time onwards: 1) Important advances in specimen preparation techniques, in particular of the electropolishing technique due to Bollmann ([2]). 2) Improvements in resolution of the electron microscopes, down to 25 Å in routine operation and a few Å under optimum operating conditions. 3) Availability of double condenser systems. 4) Development of the theory of contrast of images of crystal defects, following the work of Whelan on stacking faults ([5]). 5) Development of specimen manipulation techniques, *e.g.* goniometer stages etc., to which Valdrè has made many important contributions. (See Valdrè and Goringe in this volume.)

2. Applications of TEM in materials science.

Apart from Heidenreich's early pioneering studies, TEM has now been used extensively in studies of materials over a period of about 15 years. We shall now consider the contribution the technique has made to the advancement of knowledge in this field.

Materials science is concerned with the structure, properties, production and application of materials of all kinds. One most important aspect is the relation between properties and microstructure on all levels, from macroscopic down to atomic dimensions.

Microstructure includes not only the size, shape and nature of the constituents, *e.g.* grains, precipitates etc., but also deviations from regularity of the crystal lattice, *i.e.* lattice defects. Optical microscopy can be used to study microstructures down to $\sim 1\ \mu m$, or somewhat less using special techniques. At the lower end of the scale the crystal structure can be determined using X-ray diffraction.

In the intermediate range X-ray diffraction can give some of the answers, but often only in the form of statistical information; an example is the use of X-ray line broadening measurements in studies of the cold-worked state. The low-angle X-ray diffraction technique certainly yields important and essential data about the structure of small precipitates etc., but it does not

give direct information about the distribution of particles within a grain, or about the presence of defects.

The electron microscope has essentially filled the gap in this important size range. The information which can be obtained includes:

1) The size, shape and distribution of microstructural entities, *e.g.* precipitates, transformation products in general, etc. These can of course be studied by replica techniques, and much valuable information has and is being obtained in this way. For the study of slip lines replica techniques are invaluable. However for work on precipitation and transformation products TEM can be used without having to rely on differential etching methods, structures can be observed on a smaller scale, no modifications due to etching take place, and associated lattice strains and defects can be observed directly. Coupled with the selected-area diffraction method the TEM technique is very powerful.

2) Lattice defects and strains can be observed directly at high resolution.

3) Dynamic observations can be made of structural changes, although the results must of course be interpreted very carefully as they are not necessarily typical of the behaviour in bulk material.

4) Special contrast effects, *e.g.* due to Lorentz deflection from magnetic fields, can give additional information, in this case about magnetic domain structure.

We shall now survey briefly the fields in which the TEM technique has made particular impact.

2·1. Dislocation theory.

In the late 1940's and the 1950's there was intense development of this theory by solid state physicists. Experimental observations were rather difficult to make, although some beautiful techniques were developed, including etching, and decoration methods. There is little doubt that in this area TEM has made a most important contribution in putting the theory firmly on the map; many theoretical models were confirmed, but it soon become evident that the variety of defects and their interaction in three dimensions are so numerous that the experiments could guide the further development of the theory. Furthermore, dislocation theory is essential for the understanding of mechanical properties, a very important field in metallurgy. By enabling

dislocations to be studied in the most complex materials TEM played an important part in leading to the general acceptance of the theory in metallurgy.

It is not necessary to quote examples, there are a number of review articles and books to which reference may be made (*e.g.* ([6-8])). It might be mentioned however that the technique yields not only information on dislocation reactions, configurations, distributions etc. but important parameters such as the stacking fault energy can be obtained directly from the micrographs.

2·2. **Mechanical properties.**

In this area correlation experiments of dislocation structure as a function of strain, temperature of the deformation etc. have yielded important data on the distribution of dislocations under a variety of conditions. Some workers have claimed that the electron microscope observations, particularly the early studies, may have confused the subject of work-hardening rather than clarifying it. It is fair to say that whereas the interpretation of the structures observed in relation to their effect on hardening is still controversial, at least TEM has given a reasonably detailed picture of the dislocation structure; which, alas, is rather more complex than had been envisaged in most work-hardening theories.

There is a vast literature on the dislocation arrangements in various materials treated in different ways. In annealing studies, TEM has furnished important data on processes such as polygonization, or on mechanisms of recrystallization, particularly in polycrystalline materials. The structures in fatigue hardened metals or in specimens deformed in creep have been studied in considerable detail. In the study of deformation of more complex materials, *e.g.* two-phase alloys, the technique enables unique information to be obtained about dislocation particle interactions, leading to considerable clarification in this field. In short the technique is indispensable in the development of our understanding of mechanical properties of all kinds of materials. (Applications of TEM in metallurgy are discussed by Brown in this volume.)

2·3. **Point defects and dislocations: quench-hardening.**

When a metal is quenched from a temperature close to the melting point, a high concentration of vacancies is retained in supersaturation at the low temperature. On annealing, the vacancies coagulate into clusters, which can

assume a variety of forms depending on the nature and crystal structure of the metal, purity, quenching rates, etc. With TEM the nature of the clusters, *e.g.* dislocation loops, tetrahedra of stacking faults, cavities etc. has been revealed in considerable detail. The technique essentially opened up a new field here, concerned with structure of point defect clusters, which could not have been studied by any other method at present available.

2·4. **Radiation damage.**

TEM was being developed just at the time when the need to understand radiation damage in materials used in reactors became acute. In this field it has been possible to determine the small point defect clusters formed by neutron or ion irradiation, giving important evidence relevant to radiation damage theory.

Furthermore important problems in reactor materials technology could be investigated and solved. For example swelling of uranium fuel elements was found to be associated with pores formed at grain boundaries and small gas bubbles in the grain interior.

The shape change is probably due to interstitial and vacancy loops forming on different planes causing anisotropic expansion and contraction. Applications in this area are discussed by Makin in this volume. There is no doubt that the TEM techniques is an indispensable tool in this important field.

2·5. **Phase transformations.**

This is another area in which the TEM technique has been particularly fruitful. The sizes and distribution of zones and precipitates have been determined in age-hardening alloys, and the coherency strains associated with precipitates studied (see Brown, this volume). Precipitation on grain boundaries, stacking faults and dislocations has been revealed in many systems. Platelets of precipitate in type-I diamond were observed directly, thereby confirming the interpretation of X-ray studies which had been carried out over a number of years. Another application has been to the study of antiphase domain boundaries in ordered alloys, which have been revealed in great detail. In the field of martensite transformation, much information has been obtained on the structure of martensite plates; in a number of cases twinning has been observed, and this is important in connection with the theory of martensite transformations.

2·6. **Kinetic studies.**

These can be carried out either *in situ* in the electron microscope, or by monitoring the structure after various times of annealing etc. of the specimen outside the microscope. The shrinkage and growth of prismatic loops, faulted or unfaulted, have been studied for a variety of metals and conditions. The shrinkage experiments demonstrated the dislocation « climb » process for the first time, and later experiments on faulted and unfaulted loops were used to determine the stacking faults energy. The growth of loops under certain conditions has been shown to be related to surface oxidation, and illustrates the effect of lattice defects on this process ([9,10]). Precipitation processes have been followed *in situ* in the microscope, and observation on the break-up of dipoles into loops and other related experiments gave direct evidence for the important mechanism of pipe diffusion. Provided care is taken in the interpretation of the experiments on thin foils, very valuable data can be obtained in this way.

2·7. **Surface layer studies.**

This is an important area in solid-state physics, and particularly in the microelectronics field. TEM has given important results on the structure and growth of nuclei of a layer on a substrate, on the nature of lattice defects in these thin films and how they are formed, on how a continuous film is formed etc. The nature of interface dislocations has been studied in a number of cases. There is little doubt that the technique has helped considerably to establish a realistic model of growth of thin films, and it is an important tool for determining the actual structure of surface layers to be correlated with their properties (for review see ([11])).

2·8. **Magnetic properties.**

Antiferromagnetic domain structures have been studied in considerable detail, the contrast arising from the twin boundaries between twin related regions of crystal produced by the lattice structure transformation below the Néel temperature. Ferromagnetic domains in thin foils have been studied in great detail using the technique of Lorentz microscopy pioneered by Fuller and Hale ([12]) and Boersch and Raith ([13]). One of the discoveries was that

of magnetization ripple, and the nature of domain structures has been revealed at very high resolution. An important feature of this technique is that the magnetization process can be followed *in situ* in the electron microscope, giving unique information in this way. More recently, observations on Heusler alloys have shown strong pinning of the magnetic domain walls by the anti-phase domain boundaries in the ordered alloy ([14]); information of this kind is difficult to obtain by any other way. (For reviews see ([15]) and Wade and Wohlleben, this volume.)

2·9. Miscellaneous applications.

TEM has been applied in other fields and to many materials and problems not included in the above brief and quite incomplete survey. Applications to polymer and mineral studies might be mentioned as well as investigations of solid state or surface reactions of interest in chemistry. Special techniques for particular problems have been developed; *e.g.* the electric field distribution across a *p-n* junction can be determined using a method analogous to Lorentz microscopy for magnetic domain structures ([16]).

There is one important point which should be mentioned. In the course of the development of the TEM technique, the image contrast theory had to be worked out. To achieve this the dynamical theory of electron diffraction from perfect and imperfect crystals was extensively studied and developed. The effect of inelastic scattering is also now understood much better as a result of the intense activity in this field. These advances also helped in the development of contrast theories for other related techniques, *e.g.* X-ray topography and more recently the scanning electron microscope channelling technique ([17]); the high-energy electron diffraction theoreticians have also turned their attention to the field of low-energy electron diffraction (LEED), and important contributions are being made by them as well as by the band theory groups in this area.

3. Conclusions.

There is no doubt that the TEM technique has had a profound influence in the development of physical metallurgy during the last (10÷15) years. It is hard to see how our knowledge of, for example, radiation damage in

materials, could have progressed at anything like the rate it has done during this period without the availability of this technique. It is clear that TEM is now established in metallurgy as one of the basic techniques alongside optical microscopy, X-ray diffraction, and electron probe X-ray microanalysis.

In solid-state physics the impact has perhaps been somewhat less in relation to the field as a whole. Nevertheless it has had an important impact in dislocation theory, theory of point defect clusters, radiation damage, and structure of thin films. In magnetism there have also been important contributions to domain structures in ferromagnetic thin films, and it is clear that much of the basic work on domain structures had already been done using other methods.

In mineralogy relatively little work has been carried out so far, partly because of the difficulty of specimen preparation. But high-voltage electron microscopy should help to open up this field.

4. Future prospects.

In physical metallurgy TEM will remain an essential technique for many years. The advent of high-voltage electron microscopy will extend its application to new materials, to investigations in controlled environment, and to studies of electron irradiation damage. There are many problems waiting to be solved, both on the fundamental aspects and on technologically important materials.

There is little doubt that new techniques will be developed, which will increase the power of TEM further. Much progress is already being made with high resolution techniques, *e.g.* by the Japanese using direct lattice resolution methods. The new « weak beam » technique ([18]), is already yielding important results on the separation of partials in dislocations in f.c.c. metals, and on the structure of dislocations in complex ordered alloys; it is a very promising high resolution technique. (See Howie and see Goringe and Hall in this volume.)

The methods which are being developed for biologically important structures, whereby images can be improved by correcting for intrumental aberrations using data from a through-focus series of pictures (see Lenz and Thon in this volume), should have important applications in the study of materials. There is an urgent need to develop techniques for studying small nuclei, clusters of point defects and clusters of atoms (see Hall in this volume).

Recent experiments using the high-voltage electron microscopy have shown that it is possible to determine scattering factors with an accuracy apparently greater than that possible using X-ray diffraction (see Howie, this volume); this technique should certainly be explored.

Finally, there is an urgent need to develop further the technique of combined energy analysis and electron microscopy, or using « filtered » electrons (see Castaing, this volume). This method has already revealed segregation at grain boundaries in some Al-Mg alloys ([19]), but further exploration is necessary; a high-intensity electron gun plus a monochromator may well be needed to make this technique more widely applicable. There is little doubt that an urgent need in materials studies is for a technique capable of yielding composition analysis of very small volumes of material, and of interfaces, on a scale from a few to $\sim$ 1000 Å. It may well turn out that a scanning technique utilising a field emission type gun (see Crewe, this volume) will provide the means for this.

REFERENCES

1) R. D. Heidenreich: *Journ. Appl. Phys.*, **20**. 993 (1949).
2) W. Bollmann: *Phys. Rev.*, **103**, 1588 (1956).
3) P. B. Hirsch, R. W. Horne and M. J. Whelan: *Phil. Mag.*, **1**, 677 (1956).
4) R. B. Nicholson, G. Thomas and J. Nutting: *Brit. Journ. Appl. Phys.*, **9**, 25 (1958).
5) M. J. Whelan and P. B. Hirsch: *Phil. Mag.*, **2**, 1121, 1303 (1957).
6) A. Howie: *Metallurgical Reviews*, **6**, 467 (1961).
7) S. Amelinckx: *Solid State Phys.*, Supplement 6, Academic Press (1964).
8) P. B. Hirsch, A. Howie, R. B. Nicholson, D. W. Pashley and M. J. Whelan: *Electron Microscopy of Thin Crystals*, Butterworths (1965).
9) R. Hales, P. S. Dobson and R. E. Smallman: *Metal Science Journ.*, **2**, 224 (1968).
10) J. A. Johnston, P. S. Dobson and R. E. Smallman: *Crystal Lattice Defects*, **1**, 47 (1969).
11) D. W. Pashley: *Adv. in Phys.*, **14**, 327 (1965).
12) H. W. Fuller and M. E. Hale: *Journ. Appl. Phys.*, **31**, 238, 1699 (1960).
13) H. Boersch and H. Raith: *Naturwiss.*, **46**, 574 (1959).
14) J. Jakubovics: unpublished.
15) P. J. Grundy and R. S. Tebble: *Adv. in Phys.*, **17**, 153 (1968).
16) J. M. Titchmarsh, A. J. Lapworth and G. R. Booker: *Phys. Stat. Sol.*, **34**, K 83 (1969).
17) G. R. Booker, A. M. B. Shaw, M. J. Whelan and P. B. Hirsch: *Phil. Mag.*, **16**, 1185 (1967).
18) D. J. H. Cockayne, I. L. F. Ray and M. J. Whelan: *Phil. Mag.*, **20**, 1265 (1969).
19) S. L. Cundy, A. J. F. Metherell, R. B. Nicholson, P. N. T. Unwin and M. J. Whelan: *Proc. Roy. Soc.*, A **307**, 267 (1968).

Geometrical Electron Optics

A. SEPTIER

Institut d'Electronique Fondamentale, Laboratoire associé au CNRS,
Faculté des Sciences - Orsay, France

1. Electrostatic lenses.

1·1. Introduction.

During the last few years, we have seen spectacular developments in high voltage electron microscopes. The electrons have energies from hundreds of keV to several MeV as they pass through the lenses and relativistic effects, which are regarded as a minor perturbation in most of the books on electron microscopy, then become dominant. I feel that it will be useful, therefore, to recapitulate some formulae that describe the variation in the mass and velocity of the particles as the energy is increased.

The quantity

$$W = m_0 c^2 \tag{1}$$

is known as the rest energy (m_0, rest mass; c, velocity of light).

The kinetic energy of an electron is then given by

$$T = mc^2 - m_0 c^2 = e\varphi_0 , \tag{2}$$

in which φ_0 denotes the potential difference between the point of observation and the cathode from which the electrons are emitted. The mass m, however, varies with the velocity v. Writing $\beta = v/c$, we have

$$m = m_0(1-\beta^2)^{-\frac{1}{2}} , \tag{3}$$

so that

$$T = m_0 c^2 \{(1-\beta^2)^{-\frac{1}{2}} - 1\} . \tag{4}$$

Expanding eq. (4) as a power series, we find that, for small values of β, we recover the « nonrelativistic » formula for T:

$$T = \tfrac{1}{2} m_0 v^2 . \tag{5}$$

We can express m, v and mv (the momentum) as functions of φ_0. Writing

$$\frac{T}{m_0 c^2} = \frac{e\varphi_0}{m_0 c^2} = 2\varepsilon\varphi_0 , \tag{6}$$

we obtain

$$\left.\begin{aligned} m &= m_0(1 + 2\varepsilon\varphi_0) , \\ \beta &= \frac{v}{c} = \frac{\sqrt{4\varepsilon\varphi_0(1+\varepsilon\varphi_0)}}{1+2\varepsilon\varphi_0} , \\ \frac{p}{c} &= \frac{mv}{c} = 2m_0\sqrt{\varepsilon\varphi_0(1+\varepsilon\varphi_0)} . \end{aligned}\right\} \tag{7}$$

The quantity φ_0^*,

$$\varphi_0^* = \varphi_0(1 + \varepsilon\varphi_0) , \tag{8}$$

is usually referred to as the « relativistic potential ».

The electron velocity v, which at low energies is given by

$$v = \sqrt{2e\varphi_0/m_0}$$

becomes at high energies

$$v = \sqrt{\frac{2e\varphi_0}{m_0}}\frac{\sqrt{1+\varepsilon\varphi_0}}{1+2\varepsilon\varphi_0} \quad \text{and} \quad mv = \sqrt{2m_0 e\varphi_0^*} . \tag{9}$$

As φ_0 increases, v tends rapidly towards c and we can conveniently divide the energy range into three regions:

1) nonrelativistic, $v \propto \sqrt{\varphi_0}$ ($\varepsilon\varphi_0 < 0.1$, say, and hence $\varphi_0 < 100$ kV);

2) relativistic;

3) ultra-relativistic, $v = c$; this zone corresponds to values of $\varphi_0 > 2$ MV.

The variation of m with energy introduces a term $\mathrm{d}m/\mathrm{d}t$ into the equations of motion of the particle, but the basic equation of dynamics,

$$F = \frac{\mathrm{d}}{\mathrm{d}t}(mv) \tag{10}$$

remains valid, with $F = eE$ in an electrostatic system and $F = ev \times B$ in a magnetic system. (E and B denote the electric field and magnetic induction respectively; $B = \mu_0 H$).

1·2. Equation of motion in an electrostatic lens.

The electrons, which have been accelerated through a potential φ_0, pass through a region in which the potential varies $\varphi = \varphi(r, \theta, z)$. We consider here only systems possessing rotational symmetry about the optic axis, $0z$; this implies that $\partial\varphi/\partial\theta = 0$. In the absence of free charges, φ satisfies Laplace's equation $\nabla^2\varphi = 0$. For electron beams of low current density, we may assume that the potential distribution is not affected by the passage of the electrons, but in the case of intense beams, Poisson's equation, $\nabla^2\varphi = -\varrho/\varepsilon_0$, must be solved. ($\varrho$ is the density of charge in the beam.)

If the expression for the potential along the axis, $\varphi(0, z) = \varphi_0(z)$, is known (the axis is the line $r = 0$), we can use the fact that

$$\varphi(r, z) = \varphi_0(z) - \frac{1}{4}r^2\varphi_0''(z) + \frac{1}{64}r^4\varphi_0''''(z) - \ldots \tag{11}$$

The origin of potential is taken at the cathode.

Close to the axis, we neglect terms of fourth and higher order in eq. (11) and find

$$E_r = -\partial\varphi/\partial r = \tfrac{1}{2}r\varphi_0''(z) = -\tfrac{1}{2}r\,\partial E_z/\partial z\,. \tag{12}$$

The radial force, $F_r = eE_r$, that acts on the electrons is therefore proportional to their distance from the axis (and to $\varphi_0''(z)$). By analogy with the optics of glass lenses, we recognize that this property corresponds to the action of a lens without aberrations. Equation (12) is, however, valid only in the immediate vicinity of the axis. If the r^4-term in (11) is not negligible, the force will contain a term in r^3 and rays far from the axis will be deflected more than those close to the axis. This is one of the causes of the « third order » aberrations, spherical aberration in particular.

With the aid of eq. (10), we can establish the equation of transverse motion of the electron:

$$\frac{\mathrm{d}}{\mathrm{d}t}(mv_r) = eE_r = -e\partial\varphi(r,z)/\partial r\,. \tag{13}$$

Knowing that

$$v = v_z\left\{1+\left(\frac{\mathrm{d}r}{\mathrm{d}z}\right)^2\right\}^{\frac{1}{2}} = \frac{\mathrm{d}z}{\mathrm{d}t}(1+r'^2)^{\frac{1}{2}},$$

we can replace m and $\mathrm{d}t$ by functions of $\varphi(r,z)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[m_0(1+2\varepsilon\varphi)\frac{v}{(1+r'^2)^{\frac{1}{2}}}\frac{\mathrm{d}r}{\mathrm{d}z}\right] = eE_r\,, \tag{14}$$

$$\frac{\sqrt{2e\varphi^*/m_0}}{(1+2\varepsilon\varphi)(1+r'^2)^{\frac{1}{2}}}\frac{\mathrm{d}}{\mathrm{d}z}\left[\sqrt{2e\varphi^*/m_0}\cdot(1+r'^2)^{-\frac{1}{2}}\cdot r'\right]+\frac{e}{m_0}\frac{\partial\varphi}{\partial r} = 0\,, \tag{15}$$

with $r' = \mathrm{d}r/\mathrm{d}z$. This general equation must be used if trajectories far from the axis or steeply inclined ($\mathrm{d}r/\mathrm{d}z > 0.1$, say) are to be studied. If, however, we restrict ourselves to the conditions of Gaussian optics (r small, $\mathrm{d}r/\mathrm{d}z \ll 1$ and hence $v_z \simeq v$), we obtain a simpler equation, from which we can deduce the « first order » trajectories:

$$\frac{1+\varepsilon\varphi_0(z)}{1+2\varepsilon\varphi_0(z)}r'' + \frac{\varphi_0'(z)}{2\varphi_0(z)}r' + \frac{\varphi_0''(z)}{4\varphi_0(z)}r = 0\,. \tag{16}$$

For low-energy electrons,

$$r'' + \frac{\varphi_0'(z)}{2\varphi_0(z)}r' + \frac{\varphi_0''(z)}{4\varphi_0(z)}r = 0\,. \tag{16 bis}$$

Writing $R = r(\varphi_0(z))^{\frac{1}{4}}$, eq. (16 bis) becomes

$$R'' + \frac{3}{16}\left(\frac{\varphi_0'(z)}{\varphi_0(z)}\right)^2 R = 0\,. \tag{17}$$

The term in $\varphi_0''(z)$ has vanished, which often makes the equation simpler to solve.

Equation (16) must be used in calculating the optical properties of the

accelerating systems of magnetic microscopes operating at very high voltage. In electrostatic microscopes, however, breakdown imposes a limit on the voltage that can be employed, and the accelerating voltage does not exceed (50 ÷ 70) kV. Equations (16 bis) or (17) can then be used.

Remarks:

i) At the cathode, we have $\varphi_0(z) = 0$; the factors φ'/φ and φ''/φ are infinite, so that computer calculation is impossible. Moreover, the electrons that leave the cathode with small but not zero energy $W_0 (W_0 \simeq 10^{-1}$ eV) are emitted into a solid angle of 2π sr so that the condition $\mathrm{d}r/\mathrm{d}z \ll 1$ is no longer satisfied. The same problem arises in any region of space in which $\varphi_0(z)$ is zero or negative; the electrons are reflected back towards the source, since $v_z = 0$ if $\varphi_0(z) = 0$ and $\mathrm{d}r/\mathrm{d}z$ is infinite at the point of reflection.

ii) At low energies, the factor e/m_0 does not appear (eq. (16 bis)) so that for the same value of the potential $\varphi_0(z)$, the trajectories of charged particles of the same kinetic energy T will be identical. Ions having different masses and charges, from the ion source of a mass-spectrograph for example, can thus be brought to a focus at the same point.

iii) Examination of eqs (16 bis) or (17) reveals an extremely interesting feature of electrostatic lenses. If the potentials applied to the electrodes from a single voltage source vary with time, the trajectories will remain unaltered (provided that all the variations are *in phase*).

iv) In Gaussian optics, the trajectory equations are linear. It can easily be shown that any rotationally symmetric electrostatic system is stigmatic for pairs of conjugate points situated on or close to the axis, just as in light optics; the elementary lens formulae of Descartes or Newton can be applied once the positions of the foci and principal planes have been established. The Lagrange-Helmholtz relation is also valid, in the form

$$\sqrt{\varphi_0} r_0 r_0' = \sqrt{\varphi_i} r_i r_i' = \text{const}, \tag{18}$$

in which r and r_i denote the radius of the beam in the object and image planes, and r_0', r_i' denote the slope of the ray that intersects the axis in these planes (Fig. 1).

If $\varphi_0 = \varphi_i$, the quantity rr' remains constant and so we cannot reduce r and r' *simultaneously* by means of a « unipotential » lens. If, on the contrary, $\varphi_i \gg \varphi_0$, we have $r_i r_i' \ll r_0 r_0'$ (as in an accelerating tube).

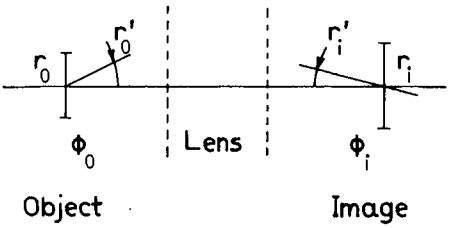


Fig. 1.

The quantity $\sqrt{\varphi}$ behaves just like the refractive index n in glass optics. We may, for example, write

$$n(z) = \sqrt{\frac{2e\varphi_0(z)}{m_0 c^2}}, \qquad \text{or} \qquad n(z) = \frac{mv}{m_0 c}.$$

v) The object and image focal lengths are related by

$$\sqrt{\varphi_0} f_0 = \sqrt{\varphi_i} f_i . \tag{19}$$

1·3. The properties of some electrostatic lenses.

In practice, only three types of lenses are used:

i) Unipotential (einzel) lenses.

ii) Immersion lenses, accelerating or retarding.

iii) Cathode lenses, or immersion objectives.

1·3.1. *The unipotential or einzel lens.* – The potential is the same on either side of the lens and is equal to the accelerating voltage of the electrons, φ_0.

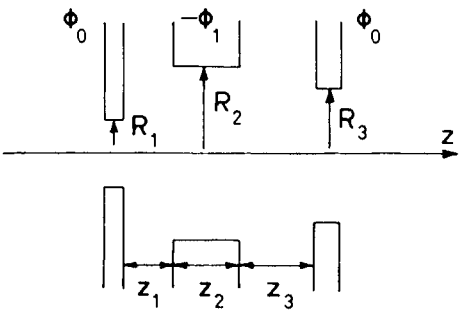


Fig. 2.

The lens consists of three diaphragms (Fig. 2); the two outer electrodes are earthed and the central electrode is connected to a variable voltage supply (the polarity of which is such that the particles are retarded).

The lens is very often symmetrical about its mid-point O, and the curve representing $\varphi_0(z)$ is then symmetric (Fig. 3). Nevertheless, we may have $Z_1 \neq Z_3$ and $R_1 \neq R_3$, with electrodes of complex shape. The function is then unsymmetric.

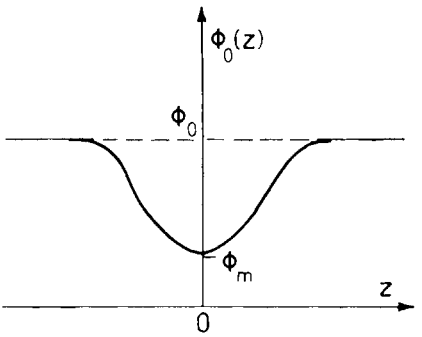


Fig. 3.

This type of lens has provoked a considerable quantity of theoretical and experimental work. Before the advent of the computer, $\varphi_0(z)$ was determined in an electrolytic tank or by means of analogue networks (devices yielding a solution of Laplace's equation).

Equation (16 bis) was then integrated, either step-by-step or by representing $\varphi_0(z)$ by an approximate analytic function. The second method is less accurate, but has the advantage that analytic expressions for $r(z)$ are obtained, in which the role of the various lens parameters can be seen.

Nowadays, the potential $\varphi(r, z)$ within a set of diaphragms can be computed by the method of relaxation; the potential distribution need be known only over a surface bounding the region of interest.

In a lens of given geometry, the only variable parameter is the potential $(-\varphi_1)$ applied to the central electrode. If we take the origin of potential at the cathode, the outer electrodes have potential $+\varphi_0$ and the excitation of the lens can be characterized by $R = \varphi_m/\varphi_0$. φ_m denotes the minimum potential on the axis,

$$\varphi_m = \varphi_0 - k\varphi_1$$

(k is a constant, less than unity, determined by the geometry alone).

Before we examine the variation of the lens convergence with φ_1, we must first consider briefly the definitions of the cardinal elements of a lens. We assume that the lens occupies a limited region of space of length L. We denote the values of r and r' at the exit from the lens by r_s and r_s' respectively for an incident ray parallel to the axis (r_0, $r_0' = 0$). We have (Fig. 4)

$$f_i = -r_0/r_s' ; \quad SF_i = -r_s/r_s' ; \quad SH_i = SF_i - f_i . \tag{20}$$

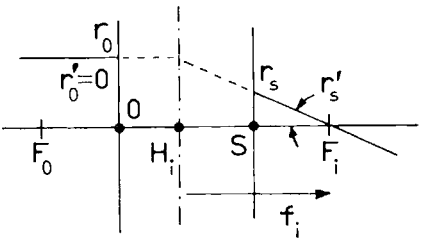


Fig. 4.

Whatever the form of $\varphi_0(z)$, the object and image focal lengths, defined in this way, are equal, but $|OF_0| = |SF_i|$ only when the lens is symmetric.

These cardinal elements are the *asymptotic* elements, which involve only quantities outside the lens. In reality, if the trajectory intersects the axis at a point within the lens, the asymptotic focus is different from the real focus F_i' (Fig. 5).

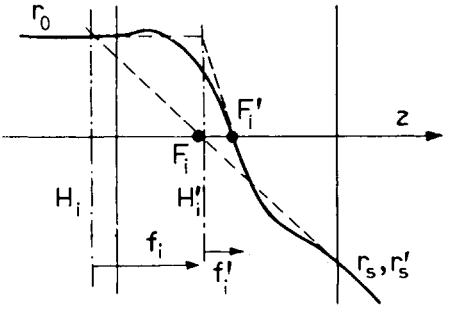


Fig. 5.

If we know the slope of the trajectory at F_i', an « immersion » focal length, f_i', can be defined. Unlike the case of magnetic lenses, these immersion elements cannot be used to design an objective, because the potential distri-

bution will be considerably altered, especially in symmetry, by the insertion of an object into the lens.

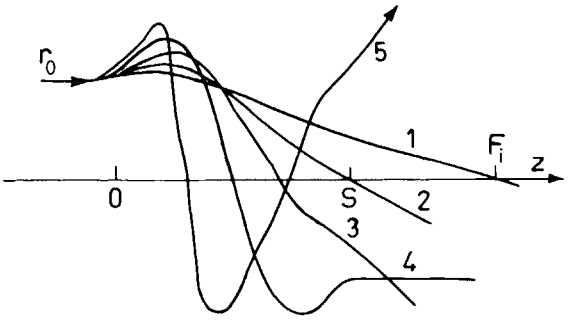


Fig. 6.

When $|\varphi_1|$ is increased, the trajectories are modified as shown in Fig. 6. The focal length f_i first decreases, then passes through a minimum and finally tends to infinity, which corresponds to a telescopic system (the ray emerges parallel to the axis). In Fig. 7, which shows f as a function of R, we see that there is a series of ever narrower regions in which the focal length takes alternate signs.

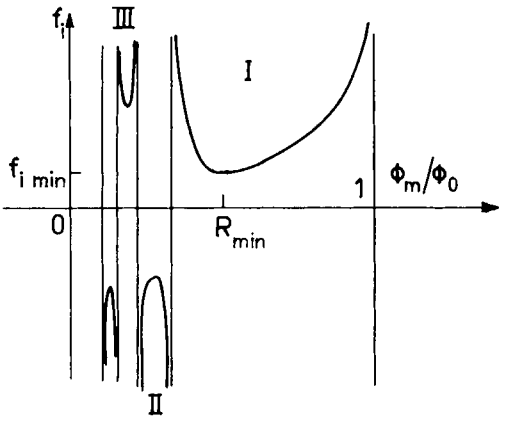


Fig. 7.

There is thus an infinite number of regions, corresponding to an increasing number of real foci. When $R<0$, the electrons are reflected and the lens becomes an electrostatic mirror, initially convergent when the incident elec-

trons reach the central zone, and subsequently divergent, when R is very negative and the trajectories are reflected in the outer region where the force F_r is positive.

If an image is required, we use region I of the diagram, and operate near to the minimum value of f_i ($R = R_{\min}$). For an objective, an electrode geometry must then be selected that will give a real focus outside the lens. For a projector on the other hand, the asymptotic foci may be immersed since the object is virtual. The minimum focal length is approximately equal to the distance between the diaphragms.

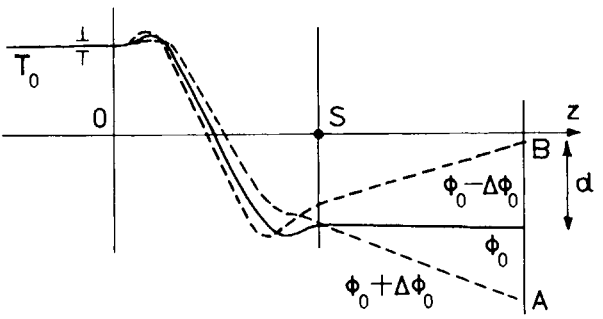


Fig. 8.

In these lenses, the principal planes are always crossed. In unsymmetrical lenses, the optical centre roughly coincides with the minimum of $\varphi_0(z)$.

The transition between I and II is used in analyzers of the Möllenstedt type, but there slit lenses are used, which are convergent in one direction only. We see (Fig. 8) that for a very small variation of R, the slope of the emergent ray corresponding to a ray T_0 incident parallel to the axis varies very rapidly. A variation in R can be obtained either by altering φ_1 or by changing φ_0. If the electrons that arrive along T_0 have an energy dispersion $\pm \Delta\varphi_0$ about φ_0, they will be distributed between A and B on the final screen, and the separation d is effectively proportional to $\Delta\varphi_0$. In this way the energy spectrum of the electrons that have passed through the object can be displayed, and the characteristic energy losses in particular can be studied.

1'3.2. *Immersion lenses.* – By definition, we have $\varphi_0 \neq \varphi_i$. These lenses may be accelerating or retarding.

Three-electrode lenses. If the three electrodes of the lens are held at potentials φ_0, φ_1 and φ_i, pairs of values of the ratios φ_1/φ_0 and φ_i/φ_0 can be obtained which enable us to vary the convergence for a given pair of conjugate points, chosen from the outset (fixed object and image). We thus have the equivalent of a zoom objective. The whole image space must obviously be held at the potential φ_i of the third electrode.

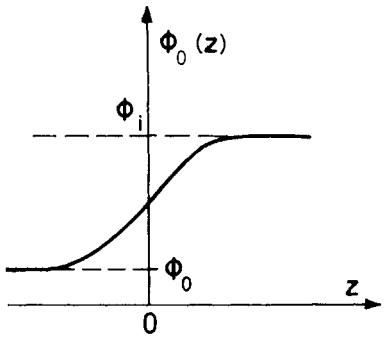


Fig. 9.

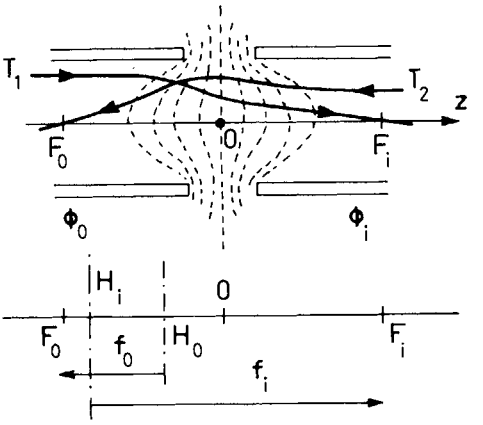


Fig. 10.

Two-tube or two-diaphragm lenses. These lenses are extensively used in oscilloscopes. The distribution $\varphi_0(z)$ corresponding to two tubes of the same

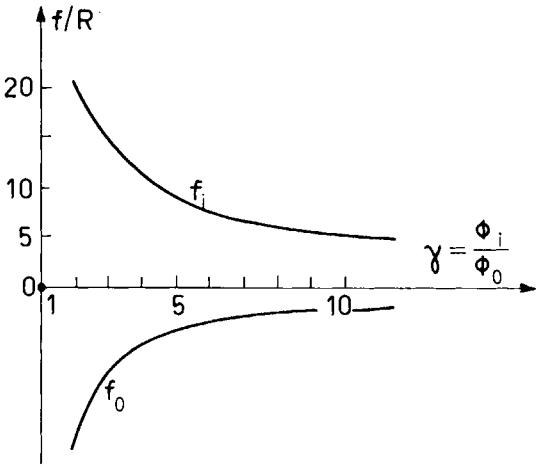


Fig. 11.

radius is shown in Fig. 9 and the equipotential surfaces are shown in Fig. 10, together with two trajectories T_1 and T_2 parallel to the axis. T_1 and T_2 have different appearances: T_1 is first convergent, then divergent, while T_2 is first divergent, then convergent.

In these lenses, we always have $f_0/f_i = \sqrt{\varphi_0/\varphi_i}$. The principal planes always lie in the low energy region. The convergence increases with the ratio $\gamma = \varphi_i/\varphi_0$ for a given value of φ_0 (Fig. 11).

Here too, the optical properties have been extensively studied, especially in the last few years by designers of electrostatic accelerating tubes. In such a tube, containing N lenses say, each lens supports the same potential difference $\Delta\varphi$ ((100÷250) kV). The ratio γ therefore decreases gradually from input to exit, tending towards unity: the last lenses are only very weakly convergent. The optical properties of the accelerating tube, regarded as a thick optical system, are obtained with the aid of matrix algebra. Each lens, like any linear system, may be characterized by a matrix containing four elements, relating r_s and r_s' to r_0 and r_0'. If such a matrix is denoted by T_j, we have

$$|T| = |T_N| \dots |T_j| \dots |T_2||T_1| . \tag{21}$$

A lens of this kind can equally well provide a family of electrostatic mirrors; for a given φ_0, we have merely to make the second tube negative with respect to the electron source.

13.3. *The immersion objective.* – This is the name given to the optical system with which an image of a plane surface emitting electrons can be obtained. Immersion objectives are employed in emission microscopes (using thermal emission or secondary emission caused by ion or photo-electric bombardment). The first electrode K is a plane cathode (at zero potential). Two types of objectives are used.

α) *The three-electrode objective.* The electrons are emitted from K with a very small initial energy φ_0 (the most probable value lies between a few tenths of an eV and a few eV) and are then accelerated by an anode A to a final potential φ_A. An intermediate electrode, the Wehnelt W, is held at the potential V close to that of K (for electrons, it is usually negative) and this enables us to vary the field E_0 at K and the convergence of the system between wide limits. By varying V, the image can be focused on the screen. The function $\varphi_0(z)$ now has the form shown in Fig. 12. Sets of curves are available, giving the characteristics of objectives of this types for a range of geometries.

We see (Fig. 13) that the crossover C of the beam lies close to the opening in the anode, and that the useful aperture of the beam of electrons leaving M with $\alpha_0 = r_0' \leqslant 90°$ can be reduced by placing an aperture in this plane; the

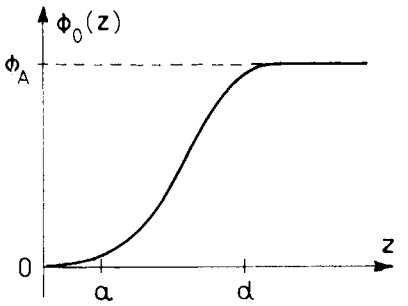


Fig. 12.

width of the energy spectrum of the electrons will likewise be reduced. When the distance between the object and the Wehnelt is varied, and V is adjusted to focus onto E, we see that E_0 remains effectively constant (with $E_0 \simeq 0.35\varphi_A/d$).

This very simple objective operates with a relatively low field E_0, therefore, since φ_A cannot exceed (30÷50) kV, with $d_{\min} = (2.5 \div 3)$ mm.

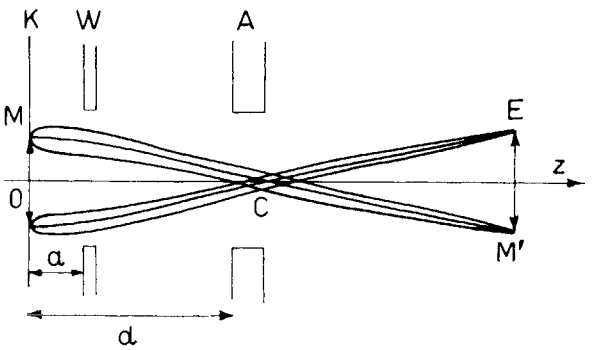


Fig. 13.

We can however show that the resolving power of an immersion objective is given by an expression of the form

$$\delta = K\varphi_0/E_0\,, \qquad K \simeq 1\,, \tag{22}$$

for $0 < \alpha_0 < 90°$. It is very difficult in practice to use an aperture plate, and it is wiser to try to increase E_0.

β) The high-field objective. In order to reduce δ, E_0 must be increased. This is achieved by using a more complicated system consisting of a cathode lens with two electrodes followed by a convergent (electrostatic or magnetic)

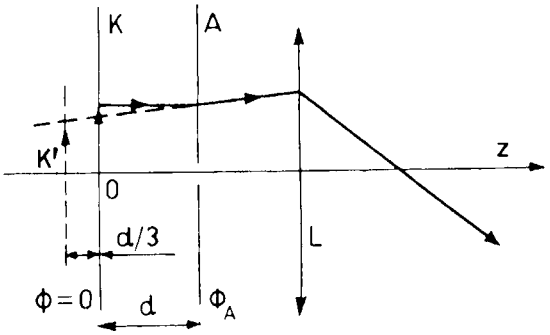


Fig. 14.

lens. The lens KA is divergent. If the anode opening is very small, the image of K falls at K', with $OK' = -d/3$, and the magnification is $\frac{2}{3}$ (Fig. 14). A high magnification at the final image is achieved simply by arranging that the object focus of L is close to K'.

In this objective, the field E_0 is given by

$$E_0 \simeq \varphi_A/d\,. \tag{23}$$

For $d = 2$ mm and $\varphi_A = 50$ kV, we have $E_0 = 250$ kV/cm, giving a minimum improvement of 3 over the preceding objective. All the high resolution images ((100÷200) Å) that have been obtained have used this system.

γ) Electrostatic mirrors. The three electrode objective is regularly employed as an electrostatic mirror in the mirror microscope. The cathode has simply to be held at a potential φ_k slightly more negative than that of the source producing the beam. The trajectories of electrons reflected in the immediate vicinity of the cathode surface are very sensitive to small local deformations of the equipotentials. These may be caused by surface relief, or by potential differences between neighbouring points on the surface (crystals,

adsorbed layers) or by small magnetic perturbations. By adjusting φ_k and V, we can then obtain « pseudoimages » of the perturbed region, in which the perturbations correspond to contrast (Fig. 15).

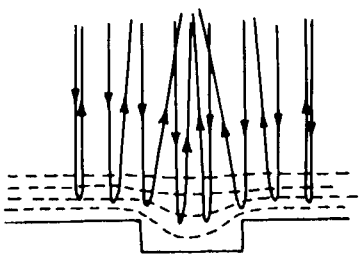


Fig. 15.

1'4. Aberrations of electrostatic lenses.

The main aberration of lenses is spherical aberration, which limits the resolving power of objective lenses. Moreover, Scherzer has shown that, for rotationally symmetric lenses, the spherical aberration cannot be corrected, as it can in the case of glass lenses. It is for this reason that efforts have been made to find lenses with as little spherical aberration as possible.

For a high magnification objective, in which the object lies very close to the object focus, the spherical aberration for a point on the axis can be defined by a relation of the form (Fig. 16)

$$\varrho = MC_s\alpha^3\,, \tag{24}$$

where M is the magnification and C_s the spherical aberration constant.

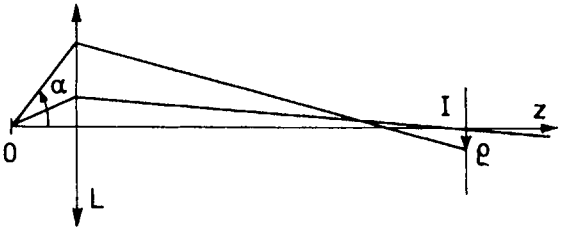


Fig. 16.

Referred back to the object plane, this aberration corresponds to a disc of radius

$$\delta = C_s \alpha^3 . \tag{25}$$

The constant C_s can be calculated by various methods, which lead to equivalent results:

i) a perturbation method based on Fermat's principle, in which a generalized refractive index is decomposed into terms of zero, second and fourth order;

ii) the trajectory method, in which fourth order terms in the potential are retained in the equations of motion, and the effect of terms in r'^2 is considered.

Writing $u = \varphi_0'(z)/\varphi_0(z)$, we obtain an expression of the form

$$C_s = \frac{1}{64\varphi_0^{\frac{1}{2}}} \int_{z_0}^{z_i} \varphi_0^{\frac{1}{2}}(z)(4u'^2 + 3u^4 - 5u^2u' - uu'')r_\alpha^4 \mathrm{d}z , \tag{26}$$

in which φ_0 is the potential in object space, z_0 and z_i are the abscissae of the object and the image respectively and $r_\alpha(z)$ is the first order trajectory that satisfies the conditions $r_\alpha(z_0) = 0$ and $r_\alpha'(z_0) = 1$. Equation (26) can be converted by partial integration into various forms. In particular, the quantity in brackets can be written as a sum of squared terms, which shows immediately that C_s cannot be zero.

The coefficient C_s can be reduced by making the lens smaller, which also reduces the focal length of the lens. This process is limited by breakdown between the electrodes, however, for a given value of φ_0, and by the requirement that the foci of the lenses must lie outside the electrodes. Alternatively, the shape of the electrodes may be varied until the minimum value of C_s is attained. Unsymmetrical three-electrode lenses are much better than symmetrical ones, provided that the zone in which $\varphi_0(z)$ varies more rapidly is near to the object, when we have objectives in mind.

The best values of C_s are four or five times greater than those of magnetic lenses with their foci outside the lens, and 20÷50 times greater than those of the best magnetic lenses with their foci within the lens. It is for this reason that electrostatic lenses are used in transmission electron microscopy only in simplified instruments of modest resolution ((50÷100) Å), operating at maximum accelerating voltages of 30 to 50 kV.

2. Magnetic lenses.

2·1. Principle.

A typical round magnetic lens is an electromagnet consisting of a winding carrying an adjustable current I_c, a yoke and pole pieces made of a soft magnetic material (Fig. 17). The electrons travel through an axial hole bored through the centre.

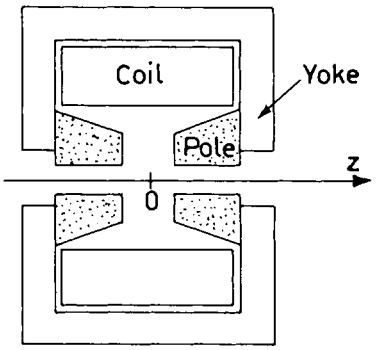


Fig. 17.

The magnetic field produced in the gap is the sum of two terms: the field produced by the poles B_p, and the field of the coil itself B_c. Under normal circumstances, B_c is much smaller than B_p. As I_c is increased, saturation in the magnetic circuit restricts B_p to (20÷25) kG, the exact value depending on the nature of the magnetic material employed. If higher fields are required, B_c must be increased. Ruska has recently succeeded in operating a standard objective with

$$B_m = B_{\text{total}} = 27 \text{ kG}.$$

(We shall see later that if superconducting windings are employed, fields $B_m \simeq (50 \div 80)$ kG can be attained.)

The curve representing the axial distribution of induction $B_0(z)$ is bell-shaped, and the half-width is slightly smaller than the gap S. The half-width increases slowly as the poles saturate.

In such a lens, the principal component of the induction B is parallel to the optic axis (Fig. 18), and this has no effect at all on trajectories that are

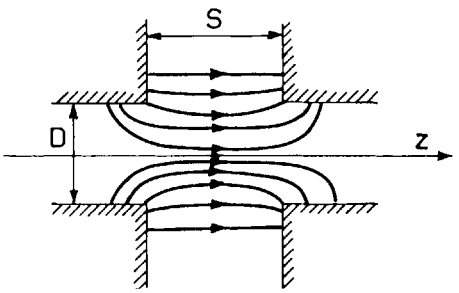


Fig. 18.

exactly parallel to the axis. The force F that B exerts on a particle of velocity v is given by

$$F = ev \times B,$$

so that $F = 0$ if $v \| B$. At the beginning of the lens, however, the radial component B_r causes the particle to rotate about the axis, giving it an azimuthal velocity $v_\theta = r\,\mathrm{d}\theta/\mathrm{d}t$. The field B_z then exerts a radial force $F_r = ev_\theta B_z$ which is always towards the axis.

2·2. The motion of particles in a magnetic lens.

In a magnetic field a particle gains no energy, so that $m = \text{const} = m_0(1 + 2e\varphi_0)$. Setting out from the general equations, namely

$$F = ev \times B = \frac{\mathrm{d}}{\mathrm{d}t}(mv), \quad \text{and} \quad B = \operatorname{curl} A \tag{27}$$

(where A is the vector potential, which here has a single component A_θ), it is easy to derive the equations of motion

$$r'' + \frac{1 + r'^2}{2(m_0/e)\varphi_0^*} - A_\theta^2 \left(A_\theta \frac{\partial A_\theta}{\partial r} - r' A_\theta \frac{\partial A_\theta}{\partial z} \right) = 0, \tag{28}$$

$$\mathrm{d}\theta/\mathrm{d}t = -eA_\theta/mr \qquad (m \neq m_0).$$

Close to the axis, we have

$$A_\theta = \frac{r}{2} B_0(z) - \frac{r^3}{16} B_0''(z) + \dots ,$$

where $B_0(z)$ is the axial distribution of B_z. For small values of r, and $r'^2 \ll 1$ (the conditions characterizing Gaussian optics), we obtain simpler equations:

$$r'' + [eB_0^2(z)/8m_0\varphi_0^*]r = 0 , \tag{29}$$

$$\theta' + (e/8m_0\varphi_0^*)^{\frac{1}{2}} B_0(z) = 0 . \tag{30}$$

φ_0^* denotes the relativistic accelerating voltage. Equation (30) gives the total rotation $\Delta\theta$ of the particle,

$$\Delta\theta = -\int_{z_0}^{z_t} \left(\frac{e}{8m_0\varphi_0^*}\right)^{\frac{1}{2}} B_0(z)\,\mathrm{d}z . \tag{31}$$

For an unsaturated lens, in which the coil is carrying a total current of nI A, this reduces to

$$\Delta\theta = -(e/8m_0\varphi_0^*)^{\frac{1}{2}} \mu_0 nI \tag{32}$$

when z_0 and z_t lie outside the lens. n is the number of turns in the coil.

If now we consider the meridian plane rOz containing the trajectory of the incident particle, and regard it as rotating within the lens at a rate given by eq. (30), then eq. (29) allows us to determine the radial motion of the particle in this rotating plane.

Lenses with no overall rotation can be designed by using two identical gaps in succession, in which the fields B_z are in opposite directions. The combination is always convergent, since the convergence is related to the square of the field: $F_r = -\frac{1}{2}(e^2/m) r B_0^2(z)$.

2·3. Optical properties.

In order to solve eq. (29), we need therefore to know only the axial distribution $B_0(z)$ and this can be measured accurately. Acomputer can then be used, given the measured values of $B_0(z)$, and the maximum field $B_{\max}$

or φ_0 can be varied. If an accuracy of a few percent is adequate, $B_0(z)$ may be represented by an approximate function that allows eq. (29) to be solved analytically. This procedure has the advantage that simple expressions for the cardinal elements (focal lengths, positions of the foci) and aberration coefficients (spherical and chromatic) are obtained. For symmetric lenses, a function of the form

$$B_0(z) = B_m/[1 + (z/a)^2] \tag{33}$$

is selected, which has the same maximum value B_m as the experimental curve, and the same area:

$$\left(\int_{-\infty}^{+\infty} B_0(z)\,\mathrm{d}z \right)_{\text{exp}} = \int_{-\infty}^{+\infty} \frac{B_m}{1 + (z/a)^2}\,\mathrm{d}z = \pi a B_m .$$

This condition determines the length a (the half-width of the theoretical bell-shaped curve). The solution of eq. (29) is to be found in Glaser's articles. Writing

$$\Omega^2 = 1 + k^2 \qquad k^2 = eB_m^2 a^2/8m_0\,\varphi_0^*$$

we finally obtain

$$f_0 = a/\sin(\pi/\Omega) \qquad z_{F_0} = a\,\mathrm{ctg}(\pi/\Omega) . \tag{34}$$

The principal planes are crossed. The two focal lengths (object and image) are always equal in magnetic lenses, since the electrostatic potential is constant. An object can be inserted into the magnetic field without disturbing $B_0(z)$, and strongly excited lenses can therefore be employed, in which F_0 is « immersed » in the field (Fig. 19).

A study of the variation of f_0 with Ω shows that f_0 decreases indefinitely as Ω increases (Fig. 20). In particular, $f_0 = a$ if $\Omega = 2$ ($k^2 = 3$) and the foci coincide at $z = 0$, the centre of the lens. If the voltage φ_0^* is too high, it is not possible to reach $\Omega = 2$, since saturation restricts B_m to about 25 kG.

The function (33) extends from $-\infty$ to $+\infty$ (whereas $B_0(z)$ is bounded in space), and the focus F_0 given by eq. (34) is thus always immersed in the field. When $B_0(z)_{\text{theoret}} \leqslant 10^{-2} B_m$, however, we may regard the effect of $B_0(z)$ as negligible, and say that we are « outside » the lens specified by eq. (33).

For a projector lens, asymptotic cardinal elements must be defined, in terms of the slope of the ray $r(z)$ at infinity (Fig. 19). We find

$$f_1 = a\Omega/\sin(\pi\Omega), \qquad z_{F_1} = a\Omega \operatorname{ctg}(\pi\Omega). \tag{35}$$

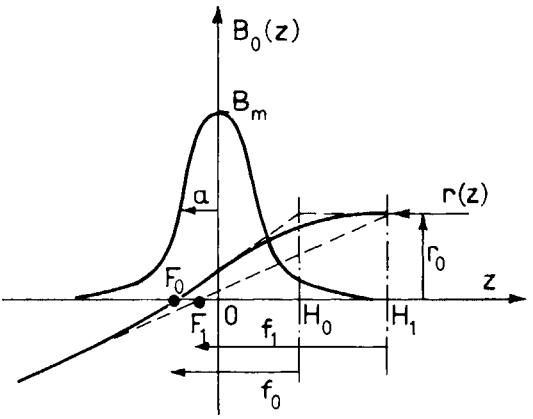


Fig. 19.

This focal length passes through a minimum as Ω is increased (Fig. 20) and becomes infinite at $k^2 = 3$. For this particular excitation, a beam incident parallel to the axis intersects the latter at $z = 0$ and emerges parallel to the axis (Fig. 21). If we place the object at $z = 0$, the part of the field for which $z < 0$ behaves as a condenser, and we have the very singular type of lens known as the « condenser-objective ». This lens has been studied by Riecke

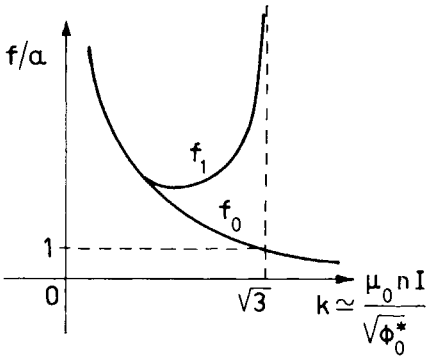


Fig. 20.

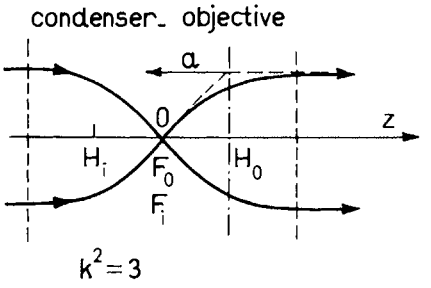


Fig.21.

and Ruska and is now used in various high resolution microscopes and in particular, in the first high voltage electron microscope (HVEM) at Toulouse. Its spherical aberration coefficient is moreover very small:

$$C_s = 0.3a$$

(in Ruska's instrument, $C_s = 0.45$ mm).

2·4. Aberrations of magnetic lenses.

If we retain terms of third order in the expansion of A_θ, and terms in r'^2, the spherical aberration coefficient C_s can be calculated by a perturbation method. If $r_\alpha(z)$ is the Gaussian trajectory satisfying the boundary conditions $r_\alpha(z_0) = 0$, $r'_\alpha(z_0) = 1$, we find

$$C_s = \frac{1}{96}\frac{e}{m_0\varphi_0^*}\int_{z_0}^{z_i}\left\{\frac{2e}{m_0\varphi_0^*}B_0^4(z) + 5B_0'^2(z) - B_0(z)B_0''(z)\right\}r_\alpha^4(z)\,\mathrm{d}z\,. \tag{36}$$

Here again, quite an accurate value of C_s can be obtained by representing $B_0(z)$ by the bell-shaped curve (33). Equation (36) can then be integrated, and for a high magnification objective we find

$$\frac{C_s}{a} = \left(\frac{\pi k^2}{4\Omega^3} - \frac{1}{8}\frac{4k^2-3}{4k^2+3}\sin\frac{2\pi}{\Omega}\right)\operatorname{cosec}^4\left(\frac{\pi}{\Omega}\right). \tag{37}$$

As the excitation is increased, C_s first falls rapidly, and then remains nearly constant between $k^2 = 2$ and $k^2 = 7$, after which it slowly increases. The minimum ($C_s = 0.25a$) occurs at excitations too high to be used in practice, for which the object must be situated beyond the centre of the lens. For $k^2 = 3$ (which is the practical limit, corresponding to the condenser-objective), we have $C_s \simeq 0.3a$.

The chromatic aberration coefficient C_c can also be calculated; C_c is defined by

$$\delta = C_c\alpha\Delta\varphi_0^*/\varphi_0^* \tag{38}$$

and we find

$$C_c = \int_{z_0}^{z_i} r_\alpha'^2\,\mathrm{d}z\,. \tag{39}$$

For the bell-shaped model

$$C_c = \frac{\pi k^2}{2\Omega^3} \operatorname{cosec}^2 \left(\frac{\pi}{\Omega}\right). \tag{40}$$

The coefficient C_c passes through a minimum value of $C_c/a \simeq 0.6$ at $k^2 = 4$. For $k^2 = 3$, we have $C_c/a = 1.8$.

If greater accuracy is required, eqs (36) and (39) are integrated numerically. A considerable effort has been devoted to the task of optimizing the dimensions of magnetic lenses, so that the smallest possible values of C_s and C_c are obtained. For a given convergence, it is always best to use short strongly excited lenses. Thus for 100 keV electrons, a value of $C_s \simeq 0.5$ mm can be attained with $f = 1.5$ mm. This displays clearly the great superiority of magnetic lenses over electrostatic lenses.

2·5. Magnetic lenses for high voltage microscopes.

2·5.1. *Normal lenses.* – In a given lens, with a fixed value of B_m, the focal length f_0 increases rapidly as φ_0^* is increased (Fig. 20). It is then advantageous to increase a in order to keep k constant. The focal length f_0 increases, since $f_0/a = \text{const}$, but less rapidly than before, and the spherical aberration remains much smaller.

If a is increased, the dimensions of the poles also increase and if B_m is to remain constant (B_m is made as high as possible, 20 kG for example), the current nI must be raised and hence the size of the coil must be increased. The lens eventually becomes very big and extremely heavy. As an example of this, we give the characteristics of the objective of the Toulouse microscope, $\varphi_0 = 3$ MV ($\varphi_0^* \simeq 12$ MV).

Gap: 12 mm; bore: 12 mm; cobalt iron pole pieces; outer diameter: 930 mm; height: 490 mm. Number of turns in the coil, $n = 34000$. Weight: 2240 kg. Power dissipated: 4 kW.

Table I shows how f_0 and C_s vary as a function of B_m; the values have been *computed* from the *measured* values of $B_0(z)$.

The chromatic aberration coefficient C_c is about 7.3 mm at maximum excitation. For $B_m > 20$ kG, the poles saturate and the curve $B_0(z)$ becomes gradually broader. Condenser-objective operation corresponds to $k^2 = 3.6$, if k is defined in terms of the measured value of the half-width, a. This lens can operate in the condenser-objective mode up to $\varphi_0 = 1680$ kV.

TABLE I.

I (A)	B_m (G)	f (mm)	C_s (mm)
0.75	19 145	15.2	28.8
1.0	22 160	12.1	13.0
1.25	24 300	10.9	8.25
1.5	26 140	10.5	6.2

2·5.2. *Superconducting lenses.* – Superconducting materials having very high critical fields, H_c, have now been known for about a decade. (The critical field H_c is the value of the magnetic field beyond which the material reverts to normal behaviour.) H_c depends on the current I that is flowing in the wire, since the field of I itself is added to the ambient field. For niobium-titanium wires, for example, which are commonly used, $H_c > 100$ kOe and for Nb_3Sn, $H_c > 200$ kOe. It is therefore possible to design coils producing axial inductions of several tens of kG. The current density in the winding can reach 10^5 A/cm^2, whereas the limit is $100 \div 1000$ with water-cooled copper.

The coils must operate in liquid helium ($T = 4.2$ °K), since the material is superconducting only beneath its critical temperature T_c ($T_c \simeq (12 \div 18)$ °K).

Various types of lenses are possible:

i) iron-free lenses;
ii) coils with an outer casing;
iii) lenses with ferromagnetic poles.

Iron-free lenses have the advantage that their properties can be rapidly and easily calculated, since the curve $B(z)$ (which is given by analytic expression) can be replaced by a Glaser bell-shaped curve to a very good approximation. The field B_m is increased and the width of the fringing fields reduced by enclosing the coil in a thick cobalt steel casing; this saturates locally close to the axis, but the curves $B_0(z)$ are nevertheless narrower (Fig. 22). The axial extent of $B_0(z)$ can also be reduced by using a superconducting screen (NbTi tubes).

The coils are wound in wire ($(0.025 \div 0.25)$ mm diameter) or ribbon in the case of Nb_3Sn (thin copper ribbon covered with $(30 \div 40)$ μm of superconductor). Another construction technique has been explored in Siegel's laboratory: the Nb_3Sn is evaporated under vacuum onto thin platinum discs

in the form of concentric circular rings (Fig. 23); these discs are piled up to form the coil, and small lenses can then be assembled.

In all these lenses, the field is essentially produced by the current flowing

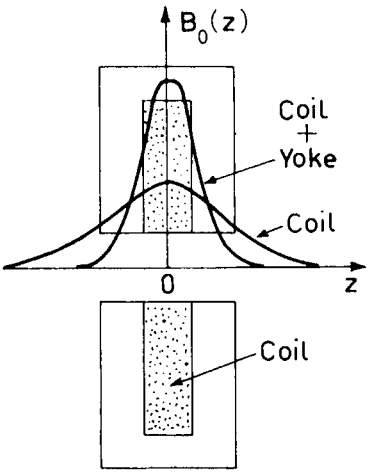


Fig. 22.

in the superconductors; the rotational symmetry can hence be disturbed by irregularities in the winding and by deformations arising during cooling. For this reason, there is a tendency to return to fully screened lenses having

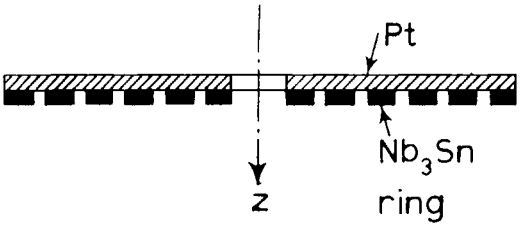


Fig. 23.

ferromagnetic pole pieces. For $B_m > 20$ kG, the poles saturate, but the field in the gap is much higher than that of the coil alone:

$$B_m \simeq B_{\text{coil}} + B_{\text{sat}}$$

($B_{sat} \simeq (24 \div 25)$ kG for permendur); the curve $B_0(z)$ also remains much narrower than that produced by the coil alone. It is possible to increase B_m by using materials that have much higher saturation magnetizations, B_{sat}, than cobalt steel; some of the rare earth metals (dysprosium and holmium) have this property, with $B_{sat} \simeq 34$ kG. It is then easy to obtain $B_m > 60$ kG in the gap, but the curve $B_0(z)$ are then very different from the Glaser bell-shaped distribution (Fig. 24) and the properties must be calculated on a computer.

Superconducting lenses can be used in two different situations:

i) In ordinary microscopes ($\varphi_0 \leqslant 200$ kV), with $B_m \leqslant 20$ kG. The coil is then very small, even allowing for the cryostat. Very short focal lengths can be achieved ($f < 2$ mm). Here, the spatial quality of the field is completely determined by that of the poles. If the coil is short-circuited when the current necessary to produce the desired field is flowing in it, the current subsequently remains perfectly stable since the flux through the coil is constant. Any contribution to the chromatic aberration due to fluctuations in I is thus eliminated, and long exposures are possible.

ii) In high voltage microscopes. With $B \simeq 50$ kG and $a = 6$ mm, we should have a lens with the following characteristics (for Glaser's bell-shaped distribution):

φ_0 (MV)	φ_0^* (MV)	k^2	f_0 (mm)	C_s (mm)
2	6	5.8	6.3	1.5
3	12	2.9	6.5	1.8
5	30	1.15	7.5	3.6

Figure 24 shows the characteristics of a lens with dysprosium pole pieces and a gap of 6 mm. Lenses with soft iron pole pieces have been studied in various laboratories. Very high voltage microscopes could be built with these lenses ($\varphi_0 \geqslant 5$ MV) but the real aberrations of the lenses, in which the coil field is dominant, have not yet been measured. At the present time, only the new Toulouse HVEM has an electron source suitable for such measurements; for this reason, we plan to measure the aberrations with lithium ions at a few keV, since these are equivalent to electrons at several MeV. The lenses can then be excited at their nominal value.

Another advantage of superconducting lenses is their lightness: a lens weighing 4 or 5 kg (including the cryostat) could replace the 2200 kg objective of a 3 MV microscope.

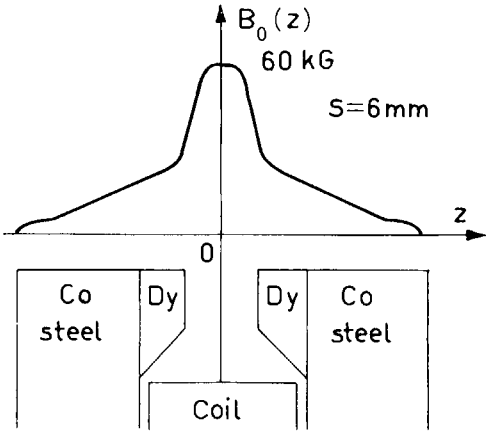


Fig. 24.

A serie of papers surveying current research in superconducting lenses is to be found in the Proceedings of the European Conference on Electron Microscopy (Rome, 1968) and of the International Conference on Electron Microscopy (Grenoble, 1970).

3. Quadrupole lenses.

3·1. Introduction.

Quadrupole lenses, which are still known as « strong-focusing lenses », constitute highly astigmatic systems, with which it is *a priori* impossible to obtain a highly magnified image of an object. They are extensively used to guide and focus very high energy beams, emerging from large particle accelerators for example.

Nevertheless, they have two possible applications in ordinary electron optics.

i) They may be used in association with octopoles, to provide correction of the aperture aberration of round lenses.

ii) They may be combined to form objective or projector lenses, equivalent to round lenses, for use in very high voltage microscopy. These optical systems, which can be corrected for spherical aberration by the use of octopoles, offer an alternative to superconducting lenses.

3·2. Optical properties of quadrupole lenses.

3·2.1. *Field distribution.* – We have seen that in round lenses, the field produced by the electrodes or poles is essentially longitudinal; the focusing action, caused by E_r or by the effect of B_z on v_θ (the azimuthal velocity created by B_r), is thus a differential effect. In quadrupole lenses, the field is transverse, except in the fringing fields at the ends, so that the focusing force is much stronger than in round lenses.

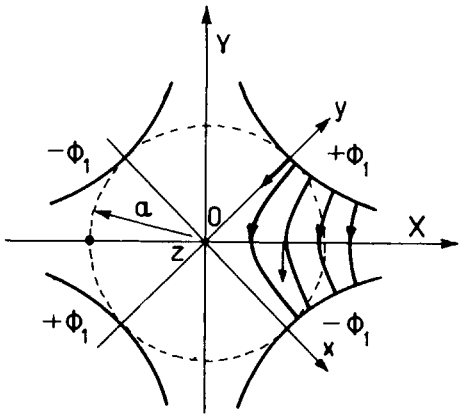


Fig. 25.

A quadrupole lens consists of four electrodes parallel to the optic axis, and has four radial symmetry planes (Fig. 25), *xOz*, *yOz*, *XOz* and *YOz*. In such a system, which we shall first of all assume to be of infinite extent in the *z*-direction, the scalar potential has the form

$$\varphi(r, \theta) = \sum_{n=1}^{\infty} A_n r^{2n} \sin 2n\theta \,, \tag{41}$$

where $n = 1, 3, 5 \ldots (2k + 1)$. The θ-origin is taken along *OX*. In a real lens, in which the radial force is proportional to r, $\varphi(r, \theta)$ must contain only

the second order term in r. Such a potential would be created by electrodes of hyperbolic shape, infinitely extended in the X and Y directions. In practice, hyperbolic segments of finite extent would be used, and $\varphi(r, \theta)$ would then contain higher order terms. Nevertheless, it is easy to suppress the term in r^6 by a judicious choice of the electrode dimensions, so that over a considerable axial region we may represent $\varphi(r, \theta)$ by

$$\varphi(r, \theta) = A_2 r^2 \cdot \sin 2\theta .$$

If $\pm \varphi_1$ denotes the potentials applied to the electrodes and a the radius of the inscribed circle (see Fig. 25), the following expressions are obtained for the potential:

$$\left.\begin{aligned} \varphi(r, \theta) &= -\frac{\varphi_1}{a^2} r^2 \sin 2\theta , \\ \varphi(x, y) &= -\frac{\varphi_1}{a^2}(x^2 - y^2), \\ \varphi(X, Y) &= \frac{2\varphi_1}{a^2} XY . \end{aligned}\right\} \tag{42}$$

For a magnetic lens, we should have

$$\varphi_1 = \mu_0 nI , \tag{43}$$

in which nI is the total number of ampere-turns flowing in the coil wound on the pole; the poles will be alternatively north and south.

In a real lens, of finite length, variations in the potential at the ends can be taken into account by writing

$$\varphi(x, y, z) = -\frac{\varphi_1}{a^2}(x^2 - y^2) \cdot k(z) \quad \text{or} \quad \varphi(X, Y, z) = \frac{2\varphi_1}{a^2} XY \cdot k(z) , \tag{44}$$

in which $k(z)$ is a function equal to unity at the centre ($z = 0$) and zero outside the lens (Fig. 26).

In studying the first-order optical properties, we replace this function by a rectangle of length L such that

$$L = \int_{-\infty}^{+\infty} k(z)\, \mathrm{d}z . \tag{45}$$

If the lens is very short ($L \leqslant 2a$, for example), $k(z)$ is a bell-shaped curve, which can be represented by a function of the form

$$k(z) = 1 \Big/ \left[1 + \left(\frac{z}{b}\right)^2\right]^2. \tag{46}$$

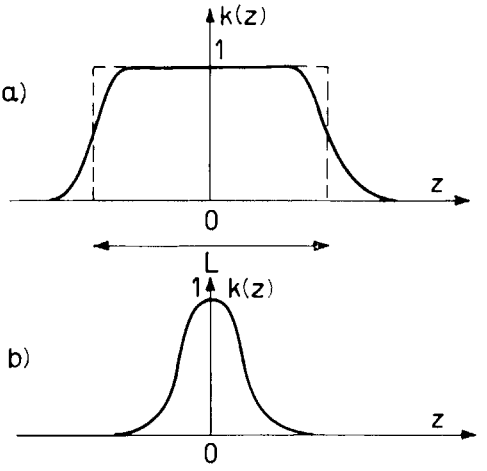


Fig. 26.

Examination of the field lines (or of the expression for the field components) in such a lens shows that there are privileged radial planes in which the force exerted on a particle is towards the axis. These are the planes xOz and yOz for an electrostatic lens and the planes XOz and YOz for a magnetic lens. With the conventions of Fig. 25, this force will have a convergent effect for an electron in xOz (or XOz) and a divergent effect in yOz (or YOz). In every other radial plane, the force is transverse but has two components, F_r and F_θ, so that the particles will be rotated. A quadrupole lens therefore constitutes a doubly cylindrical system, equivalent to a convergent lens in one direction and to a divergent lens in the direction at right angle.

If now we place eight identical electrodes parallel to the axis, so that they form a system with eight geometrical symmetry planes and four electrical symmetry planes, we obtain an octopole lens (Fig. 27). The potential is given, in the vicinity of the axis by

$$\varphi(r, \theta) = A_4 r^4 \sin 4\theta + \ldots$$

The force on a particle varies as r^3. This explains why these lenses can be used to correct the spherical aberration of round lenses. Here again, however, the transverse force is purely radial only in certain privileged planes, and is alternately convergent and divergent every 45° around the axis.

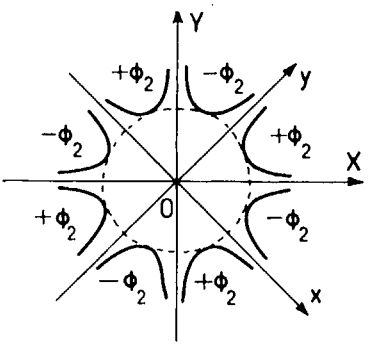


Fig. 27.

3'2.2. *Equations of motion.* – Applying the law

$$\frac{\mathrm{d}}{\mathrm{d}t}(mv) = eE \quad \text{or} \quad \frac{\mathrm{d}}{\mathrm{d}t}(mv) = ev \times B$$

in the planes xOz, yOz for the electrostatic case and XOz, YOz for the magnetic case, and assuming that the scalar potential is given by eqs (42) and that $r'^2 \ll 1$, we obtain uncoupled sets of equations

$$\left.\begin{aligned} x'' + \beta_E^2 k(z) x = 0\,, \\ y'' - \beta_E^2 k(z)\, y = 0\,, \end{aligned}\right\} \quad \text{or} \quad \left.\begin{aligned} X'' + \beta_M^2 k(z) X = 0\,, \\ Y'' - \beta_M^2 k(z)\, Y = 0\,, \end{aligned}\right\} \tag{47}$$

with

$$\left.\begin{aligned} \beta_E^2 &= \frac{eK_E(0)}{mv^2} = \frac{K_E(0)(1 + 2\varepsilon\varphi_0)}{2\varphi_0^*}\,, \\ \beta_M^2 &= \frac{eK_M(0)}{mv} = K_M(0)\sqrt{\frac{e}{2m_0\varphi_0^*}}\,. \end{aligned}\right\} \tag{48}$$

$K_E(0)$ and $K_M(0)$ denote the radial gradient absolute values of the electric and magnetic *fields*, respectively, in the centre of the lens:

$$K_E(0) = \frac{2\varphi_1}{a^2}\,, \qquad K_M(0) = \frac{2\mu_0 nI}{a^2}\,. \tag{49}$$

If we use the rectangular model approximation, eqs (47) are easily integrated since $k(z) = 1$. We obtain for example

$$\left.\begin{aligned} X &= X_0 \cos \beta_M z + \frac{X_0'}{\beta_M} \cdot \sin \beta_M z , \\ Y &= Y_0 \cosh \beta_M z + \frac{Y_0'}{\beta_M} \sinh \beta_M z . \end{aligned}\right\} \tag{50}$$

For a lens of length L, the quantities X_s, X_s', Y_s and Y_s' can be expressed in terms of the initial conditions, X_0 X_0', Y_0 and Y_0' by means of the relations

$$\left.\begin{aligned} \begin{pmatrix} X_s \\ X_s' \end{pmatrix} &= \begin{vmatrix} \cos \beta L & (1/\beta) \sin \beta L \\ -\beta \sin \beta L & \cos \beta L \end{vmatrix} \begin{pmatrix} X_0 \\ X_0' \end{pmatrix} = |T_X| \begin{pmatrix} X_0 \\ X_0' \end{pmatrix}, \\ \begin{pmatrix} Y_s \\ Y_s' \end{pmatrix} &= \begin{vmatrix} \cosh \beta L & (1/\beta) \sinh \beta L \\ +\beta \sinh \beta L & \cosh \beta L \end{vmatrix} \begin{pmatrix} Y_0 \\ Y_0' \end{pmatrix} = |T_Y| \begin{pmatrix} Y_0 \\ Y_0' \end{pmatrix}. \end{aligned}\right\} \tag{51}$$

This matrix formalism, in which $|T_X|$ and $|T_Y|$ denote the transfer matrices of the lens, allows us to calculate rapidly the optical properties of complex systems formed by joining several lenses. The overall transfer matrix is obtained by multiplying the individual matrices. The matrix corresponding to field-free space (a drift space) of length D is given by

$$|T| = \begin{vmatrix} 1 & D \\ 0 & 1 \end{vmatrix} .$$

We recall that in a transfer matrix of the form

$$|T| = \begin{vmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{vmatrix} ,$$

we always have

$$T_{11} T_{22} - T_{12} T_{21} = 1 , \tag{52}$$

provided that the electrostatic potential is the same on either side of the lens.

The convergence of the system, C, is given by

$$C = \frac{1}{f} = -T_{21} \tag{53}$$

and the position of the focus with respect to the exit plane z_s by

$$\overline{SF} = z_F - z_S = -T_{11}/T_{21}. \tag{54}$$

3'2.3. *The cardinal elements of a single lens.* – By considering a trajectory incident parallel to the axis, and setting $\beta L = k$, we obtain (Fig. 28)

$$\left.\begin{aligned} f'_X/L &= (k \sin k)^{-1}, \quad & f'_Y/L &= -(k \sinh k)^{-1}, \\ \overline{SF}'_X/L &= \frac{\cos k}{k \sin k}, \quad & \overline{SF}'_Y/L &= -\frac{\cosh k}{k \sinh k}. \end{aligned}\right\} \tag{55}$$

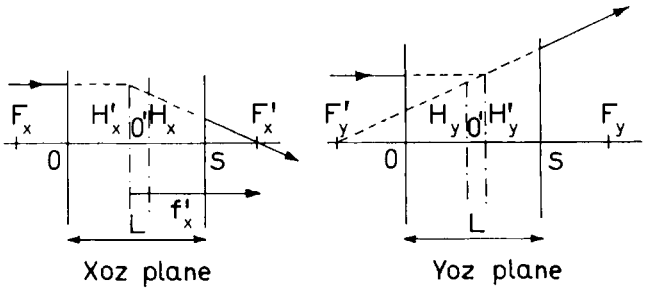


Fig. 28.

The object elements are such that $\overline{SF}'_X = -\overline{OF}_X$, $\overline{SF}'_Y = -\overline{OF}_Y$, $f'_X = -f_X$, and $f'_Y = -f_Y$. The principal planes are crossed.

In order to obtain an optical system that is convergent in all directions, we combine two crossed quadrupoles, Q_1 and Q_2 (the convergent plane C of Q_1 coincides with the divergent plane D of Q_2): this yields a quadrupole doublet.

3'2.4. *The cardinal elements of a doublet.* – When the excitations of the two lenses, which we can characterize by k, are chosen arbitrarily, the object foci of the doublet F_X and F_Y, are not coincident (nor are the image foci, F'_X and F'_Y), and the focal lengths are unequal: (Fig. 29)

$$f'_X \neq f'_Y.$$

Nevertheless, $f_Y = -f'_Y$ and $f_X = -f'_X$. When the lenses are identical ($L_1 = L_2$, $a_1 = a_2$) and equally excited ($k_1 = k_2$), we do have

$$f_X = f_Y,$$

but the X and Y foci are separated. We shall now show that a doublet consisting of two identical lenses, with $k_1 = k_2$, can be made equivalent to a round lens. For this, we merely write

$$\overline{SF}_X = \overline{SF}_Y ,$$

that is

$$(T_{11})_X = (T_{11})_Y . \tag{56}$$

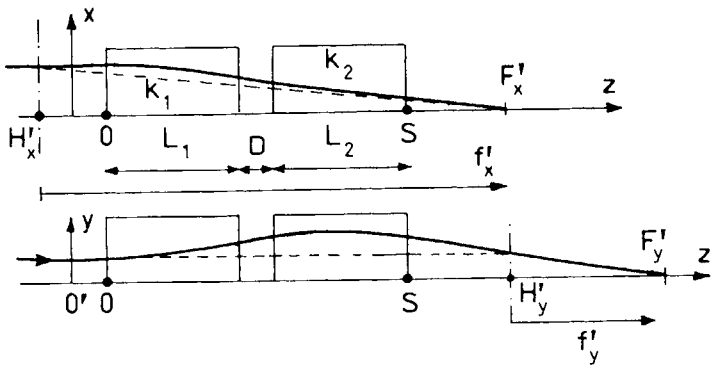


Fig. 29.

Writing $D/L = \lambda$, we find

$$1/\lambda = -\tfrac{1}{2}k(\operatorname{ctg} k + \operatorname{ctgh} k) . \tag{57}$$

For $D = 0$, this equation gives: $k = \pi$

$$\frac{f_X}{L} = \frac{f_Y}{L} = \frac{f}{L} = \frac{1}{\pi \sinh \pi} , \qquad \frac{\overline{SF}'_X}{L} = \frac{\overline{SF}'_Y}{L} = \frac{-1}{\pi \operatorname{tgh} \pi} \simeq \frac{-1}{\pi} . \tag{58}$$

The focal length is very short, and unfortunately the asymptotic foci fall well inside the lens. Examination of the trajectories (Fig. 30) shows that the real immersed foci O' and O'' do not coincide in the X and Y planes. This doublet cannot therefore be used as a high magnification immersion objective or as a demagnifying lens for forming a real probe. It could on

the other hand, be used as a projector. We have indeed succeeded in obtaining good quality images in an electrostatic microscope using a quadrupole projector.

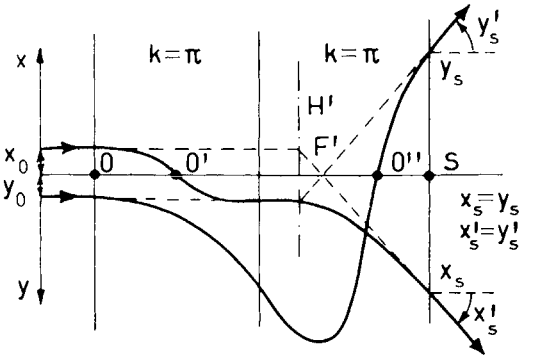


Fig. 30.

3·3. Quadrupole systems suitable as objectives.

An optical system equivalent to a round lens for all points on the axis must contain at least four lenses, combined in such a way that the quadruplet has a plane of geometrical symmetry (P). The two central lenses have the same length L_2 (Fig. 31). If (P) is also a plane of electrical symmetry, the foci

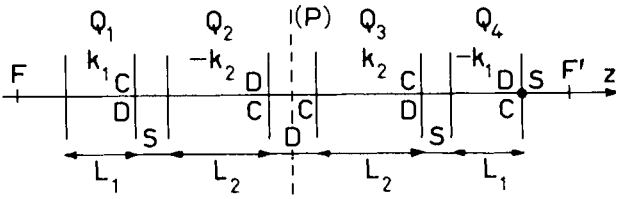


Fig. 31.

will always be immersed. If, however, (P) is a plane of antisymmetry, it is possible to find pairs of excitations k_1 (for Q_1 and Q_4) and k_2 (for Q_2 and Q_3, which are crossed), such that F and F' lie outside the lenses. The plane XOz is then convergent-divergent-convergent-divergent (CDCD) and YOz is (DCDC). The terms in the transfer matrix can be simplified to some extent

by setting $S = 0$. This type of quadruplet has been extensively studied, and the cardinal elements have been obtained as functions of the geometrical parameters L_2/L_1, D/L_1 and S/L_1. In general, there are several regions of the $k_1 - k_2$ plane for which $SF' > 0$. The focal length of such a quadrupole may be very short: $f/L_1 \simeq 10^{-2}$ for example, but the excitations then become high. With $L_2 = 2L_1$, $S = 0$ and $D/L_1 = 0.155$, we find $f/L_1 = 4 \cdot 10^{-2}$, $z_F = 0.26 L_1$ with $k_1 = 1.95$ and $k_2 = 4.5$. Figure 32 shows the form of trajectories in-

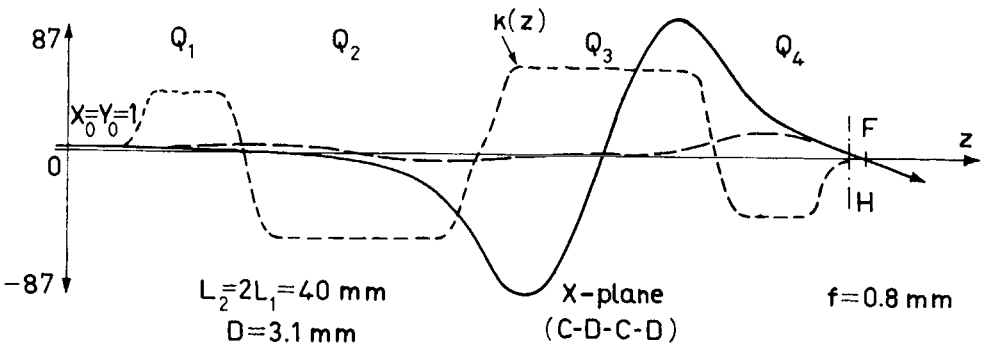


Fig. 32.

cident parallel to the axis. For 3 MV electrons ($\varphi_0^* = 12$ MV), the field gradients in the lenses would have to be $1.1 \cdot 10^4$ G/cm and $1.5 \cdot 10^4$ G/cm respectively, for $a = 4$ mm, $L_1 = 2$ cm and $L_2 = 4$ cm; we then have $f = 0.8$ mm and $\overline{SF} = 6.4$ mm. This shows vividly how superior these lenses are to round lenses, so far as convergence is concerned. The aberration coefficients are much higher, however, and octopole correctors would be necessary.

3·4. The aperture aberrations of quadrupole systems and their correction.

When rotational symmetry is abandoned, four coefficients become necessary in general to describe the aperture aberration. For a beam incident parallel to the axis, the transverse aberration in the Gaussian image plane of a pseudostigmatic system (two coincident foci) will be given by

$$\Delta X = C_X \alpha^3 + C_{XY} \alpha \beta^2,$$

$$\Delta Y = C_Y \beta^3 + C_{YX} \beta \alpha^2,$$

and in this special case, $C_{XY} = C_{YX}$. α and β denote the semi angular apertures in image space, in the XOz and YOz planes respectively, In this case, three coefficients are adequate; expressions for them, and also for the other aberration coefficients (coma, distortion, ...), are to be found in articles published by P. W. Hawkes.

Examination of Fig. 32 shows that the outermost trajectories of the beam are much farther from the axis in the CDCD plane than in the DCDC plane. They are hence more sensitive to the perturbations caused by the third order terms (and even by fifth order terms) which we have so far neglected in the potential and by the steep slope of the rays. Even for a « pure » quadrupole field (eqs (42)), the effect of the fringing fields has to be taken into account; if the field variation at the ends of the lens is characterized by $k(z)$, the field B_z varies as $k'(z)$ and term in $k''(z)$ must be included so that the potential function is indeed a solution of Laplace's equation. In the x-y co-ordinates, for example, we have

$$\varphi(x, y, z) = \frac{\varphi_1}{a^2} k(z)(x^2 - y^2) - \frac{k''(z)}{12} \frac{\varphi_1}{a^4} (x^4 - y^4) + \text{terms of sixth order} . \quad (59)$$

Using (59) and the relation

$$\frac{dz}{dt} = v[1 + (x'^2 + y'^2)]^{-\frac{1}{2}}$$

we obtain the « third-order equations of motion », the solutions of which are written $x^{(3)}(z)$ and $y^{(3)}(z)$. If we regard the aberrations as a perturbation of the Gaussian trajectories, $x^{(1)}(z)$ and $y^{(1)}(z)$, we can set

$$\varepsilon(z) = x^{(3)}(z) - x^{(1)}(z) ,$$
$$\eta(z) = y^{(3)}(z) - y^{(1)}(z) ,$$

and we obtain equations of the following form:

$$\varepsilon'' + k(z)\varepsilon = S_x(x, y, x', y', z) .$$

We replace the unknown functions x, y, x' and y' on the right-hand side by the Gaussian solutions, this enables us to solve the equations more easily, though only approximately. If greater precision is required, the third-order

equations are integrated directly on a computer. The expressions for the aberration coefficients given by Hawkes, which are easy to compute, can thus be used once trajectories have been obtained by integrating the Gaussian equations.

All the calculations show that the aberration coefficients of strongly excited quadrupoles, which could be used in electron microscopy, are higher than those of round lenses. Measurements made by Dhuicq and Septier in particular confirm this result. For the objective of Fig. 31, for example, we find that with $f = 0.8$ mm

$$C_Y/f = 39 \quad C_{XY}/f = 72 \quad C_X/f = 1.4 \cdot 10^5 \quad \text{(magnetic)},$$

$$C_y/f = 12 \quad C_{xy}/f = 550 \quad C_x/f = 0.5 \cdot 10^5 \quad \text{(electrostatic)}.$$

The three coefficients can be cancelled simultaneously by placing octopoles at suitable points in the quadrupole system. We describe briefly the principle of this correction technique, which can also be used with round lenses. In a region in which the beam is rotationally symmetric, an octopole is placed, excited in such a way that it corrects the terms in $\alpha\beta^2$ or $\beta\alpha^2$ in the planes $\theta = \pm 45°$ (which bisect the co-ordinate axes xyz); it worsens the aberrations in the plane xOz and yOz. In regions where the beam is highly elliptical, two

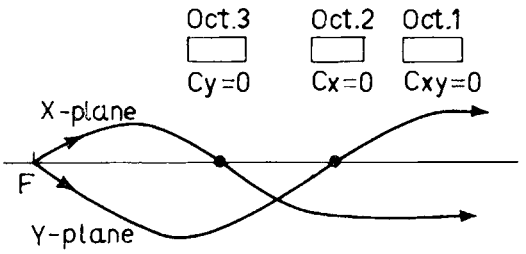


Fig. 33.

other octopoles are placed; these correct the α^3 terms (when $y = 0$, $x \neq 0$ for example: *i.e.* close to a line focus parallel to Ox) and the β^3 terms (when $y \neq 0$, $x = 0$: line focus parallel to Oy): see Fig. 33. The octopole potentials required for correction can be produced either by separate lenses or by introducing extra poles or electrodes into the quadrupoles. For the quadruplet of Fig. 31, Dhuicq has adopted the solution described by Deltrap, in which

each pole of the quadrupole is cut in two and fitted with two types of coils, so that both quadrupole excitations ($\pm \varphi_2$) and octopole excitations ($\pm \varphi_4$) can be produced (Fig. 34).

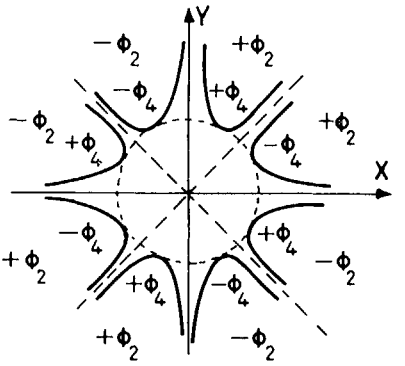


Fig. 34.

Many theoretical and experimental studies have been devoted to this problem, with a view to obtaining objectives capable of focusing (5÷10) MeV electrons, or correctors for round lenses, or microprobes of high current density. (We recall that the last demagnifying lens of a microprobe must have its focus outside the lens, so that the lowest values of C_s that can be obtained with a strongly convergent objective cannot be attained.)

Considerable efforts to optimize the solution are still necessary, which will reveal which systems have the smallest aberration coefficients and are least sensitive to small misalignments of the electrodes; for, if we are to take full advantage of the correction of spherical aberration, the new aberrations introduced by imperfect symmetry or misalignment must be zero or negligible.

3·5. Correction of chromatic aberration.

It is possible to obtain achromatic quadrupole systems, by combining magnetic and electrostatic lenses. As a simple example we consider a mixed lens, consisting of four magnetic poles and four electrodes held at electrostatic potentials $\pm \varphi_1$, arranged as shown in Fig. 35. The lenses produce both electrostatic and magnetic field gradients, K_E and K_M, in the

same region of length L. The equation of motion in the xOz plane is now

$$x'' + \left(\frac{eK_E}{mv^2} + \frac{eK_M}{mv}\right) x = 0\,, \quad \text{or} \quad x'' + A(v)x = 0\,,$$

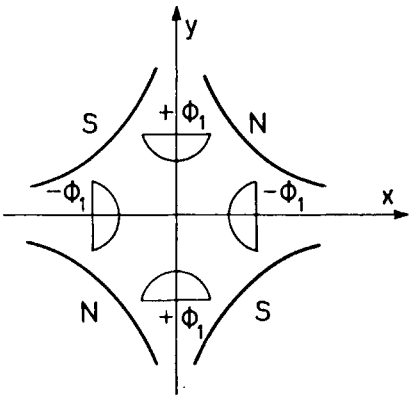


Fig. 35.

for nonrelativistic electrons. The system will be achromatic if $\partial A(v)/\partial v = 0$, or

$$\frac{2eK_E}{mv^2} + \frac{eK_M}{mv} = 0\,,$$

that is

$$K_E = -\frac{v}{2} \cdot K_M\,.$$

We thus require (see (48)):

$$A(v) = \frac{\beta_M^2}{2}\,.$$

If we consider the y-equation we obtain the same condition. This implies that if the magnetic lens is convergent in the plane xOz, the electrostatic lens must be divergent. The overall convergence of the lens is thus less than that of the magnetic lens alone.

This interesting property has been confirmed experimentally. We can show that the correction can also be obtained when the lenses are separated along

the axis. At the present time, arrangements are being sought with which *both* the chromatic aberration *and* the spherical aberration can be cancelled. Unfortunately, the need for electrostatic lenses prevents us from exploiting this property at very high voltages.

4. Prisms optics.

4·1. Introduction.

In particle optics, there is a whole range of situations in which we have to analyse the energy spectrum or the different components of a heterogeneous beam. We then use deflector systems with dispersive properties and, by analogy with glass optics, we call such systems « prisms ». We shall find that these prisms also possess interesting focusing properties.

In its simplest form, a prism consists of a limited region of space containing a homogeneous electric or magnetic field: a plane electrostatic condenser, or a pair of parallel magnetic poles very close together.

Such a system is used to obtain fairly weak angular deflections: the electron beam in a cathode-ray tube is deflected in this way.

In order to obtain larger deflections and a high dispersion it is better to use long systems with a large radius of curvature, in which the mean trajectory is a segment of circle centred at O, and of radius R.

Electrostatic prisms will thus consist of portions of cylindrical or spherical condensers and magnetic prisms of circular sector magnets. The angle at the centre of the prism, Φ, is then equal to the deflection imposed on the mean trajectory of the beam in the symmetry plane of the system.

4·2. Simple prisms.

4·2.1. *Electrostatic deflection.* – In a parallel plate condenser, supporting a potential difference $V = Ed$ (see Fig. 36), the field E deflects a beam injected along the Oz axis through an angle α given by:

$$\alpha = \frac{1}{2}\frac{Vl}{\varphi_0 d} = \frac{elE}{mv_0^2}.$$

l is the length of the condenser and φ_0 the accelerating voltage. The trajectory is a parabola and the emerging beam seems to come from the prism centre O'. Inside the condenser, the origin being taken in O, we obtain:

$$y = \frac{eEz^2}{2mv_0^2}.$$

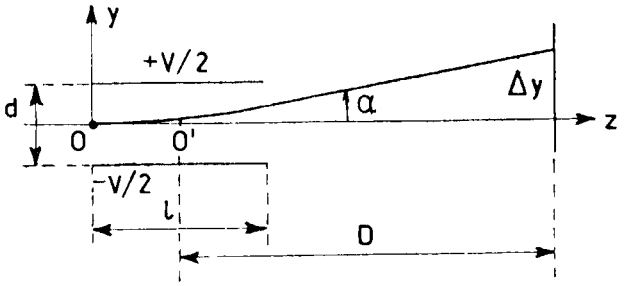


Fig. 36.

At a distance D from O', the deviation is

$$\Delta y = \frac{V}{\varphi_0}\frac{Dl}{2d}.$$

4·2.2. *Magnetic deflection.* – We only consider the case of a homogeneous magnetic field B, extending over a limited distance l along the axis Oz, (Fig. 37).

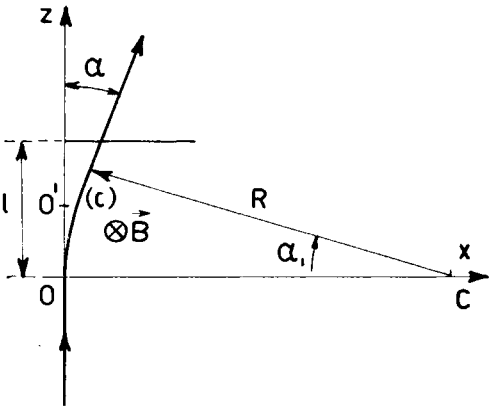


Fig. 37.

The trajectory (c) is a circle of radius R, with

$$R = \frac{mv_0}{eB} = \frac{\sqrt{(2m/e)\varphi_0}}{B} \quad \text{and} \quad x \simeq \frac{eBz^2}{2mv_0}.$$

The deviation angle being small, we have

$$\alpha \simeq \frac{l}{R} = \frac{lBe}{mv_0} = \frac{lB}{\sqrt{(2m/e)\varphi_0}}.$$

4·2.3. *Parallel electrostatic and magnetic fields.* – B and E are paralle to the Oy axis, and act on the incident beam $O'z$. The two deflections are superimposed independently. Taking the ratio x/y, we obtain

$$\frac{x}{y} = \frac{B}{E} v_0 .$$

A monochromatic beam will mark one point on a screen set up at right angle to the Oz axis at $z = z_0$. If now the beam contains electrons spread. over a wide range of velocities it will be dispersed along a curve on the screenr The equation of this curve is derived by elimination of v_0 from the above equations

$$\frac{x^2}{y} = \frac{eB^2}{2mE} z_0^2 .$$

The inhomogeneous beam trace a parabola $y = kx^2$ on the screen (o photographic plate).

If the beam is made of positive ions, of different kinds corresponding to different values of the ratio (e/m), each kind of ion gives a parabola on a photographic plate. (Parabola mass spectrometer.)

4·2.4. *Crossed electric and magnetic fields* (Wien filter). – The beam always follows the Oz axis, but we have now $E \parallel Ox$ and $B \parallel Oy$: both deflections are in the xOz plane, and the direction of B is chosen in order to counterbalance the electric deflection for a given value of v_0

$$x_E = \frac{eEz^2}{2mv_0^2} = -x_B = -\frac{eBz^2}{2mv_0} .$$

In this case the ratio E/B have to be fixed to a special value $E/B = v_0$.

Particles of velocity v_0 are not deflected; by putting (Fig. 38) a narrow slit at the exit of the system, we obtain a velocity filter or a mass filter. By varying the ratio E/B, a velocity (or mass) spectrum may be recorded beyond the slit. This type of filter is extensively used for studying the energy spectrum of electron beams, which have passed through a thin film (or a gas). Its resolution, which depends on the width of the output slit, can attain 10^{-1} eV.

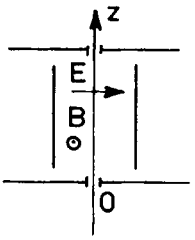


Fig. 38.

4·3. Magnetic prisms.

4·3.1. *Description-fields.* – We consider a magnet consisting of two identical poles in the form of circular sectors of angle Φ, centre C, and mean radius r_0, terminated by the plane faces AB and $A'B'$ (Fig. 39). The magnetic field is constant along an arbitrary circle (C) of radius r lying in the symmetry plane (the « median » plane). On (C_0), we have $B = B_0$. We assume

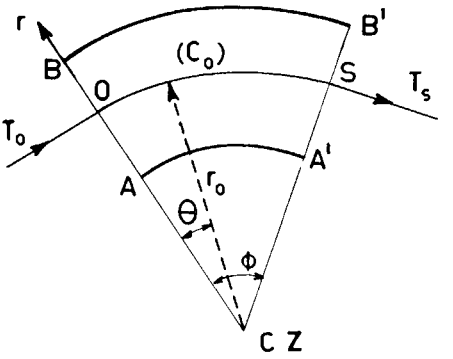


Fig. 39.

that $B = 0$ outside the end faces, and that $B =$ constant along (C) for $0 \leqslant \theta \leqslant \Phi$ (rectangular model approximation).

The field B is parallel to the axis CZ, in the median plane. Instead of the standard cylindrical co-ordinates r, θ, with axis CZ, it is more convenient to use the reduced dimensionless co-ordinates x, z, θ, defined by

$$x = (r - r_0)/r_0 , \qquad z = \frac{Z}{r_0} , \qquad \theta = s/r_0 .$$

(C_0) plays the role of a curved axis, which extends outside the sector as two straight lines T_0 and T_8; $\mathrm{d}s$ is an element of length along (C_0).

Solving the Laplace equation in the gap between the poles, and retaining only the first terms in the expression giving the two components of the field, we obtain

$$B_z = (1 - nx)B_0 = B_0[1 - n(r - r_0)/r_0] , \qquad B_x = -B_0 nz = -B_0 nZ/r_0 ,$$

n is called the « index » of the field. $n = 0$ corresponds to a homogeneous field ($B_x = 0$ in this case, and $B_z = B_0$).

4·3.2. *First-order trajectories* (Gaussian optics). – The curve (C_0) is the path of a particle of charge e, mass m and momentum $p_0 = mv_0$, such that $B_0 r_0 = p_0/e$.

Particles of momentum $p_1 \neq p_0$ can then also follow circular paths (C) of radius $r_1 \neq r_0$, satisfying the equation

$$B(x_1)r_1 = p_1/e .$$

We now study the trajectories of particles of momentum p_0, injected near the curved axis $(C_0) = x$, $z \ll 1$, with small slopes $(x'^2, z'^2 \ll 1)$. Furthermore we may write $v_\theta = v_0$, and so

$$v_\theta = r\,\mathrm{d}\theta/\mathrm{d}t = v_0 ,$$

or

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} \simeq \frac{1}{1+x}\left(\frac{\mathrm{d}\theta}{\mathrm{d}t}\right)_0 = \frac{\omega_0}{1+x} .$$

ω_0 is known as the « cyclotron frequency » of the particle in the field B ($\omega_0 = eB_0/m$). In the axis $(x, 0, z)$, the equation of the motion takes the fol-

lowing form

$$\left.\begin{aligned} &d^2x/d\theta^2 + (1-n)x = 0\,, \\ &d^2z/d\theta^2 + nz = 0\,. \end{aligned}\right\} \tag{60}$$

For $0 < n < 1$, the solution of (60) can be expressed in terms of circular functions: the prism is convergent in both the x and z directions (that is in both the horizontal plane H, and the « vertical » plane V). If $n = 0$ (homogeneous field), there is no focusing in V plane, but only focusing in H plane. We observe that the two equations are identical when $n = \frac{1}{2}$, so that convergence is the same in H and V: the prism *is equivalent to a round thick lens.*

4·3.3. *Optical properties of the prism*, with entry and exit faces normal to the mean trajectory (C_0). – The solutions of (60) are of the form

$$x = A_H \cos\left(\sqrt{1-n}\,\theta\right) + B_H \sin\left(\sqrt{1-n}\,\theta\right),$$

$$z = A_V \cos\left(\sqrt{n}\,\theta\right) + B_V \sin\left(\sqrt{n}\,\theta\right).$$

The constants A and B can easily be expressed as a function of the initial conditions at the entry face, and in the planes H and V respectively, we can define the transfer matrices of the prism. We only give them in two special cases: $n = 0$ and $n = \frac{1}{2}$.

Homogeneous field ($n = 0$). The matrices $|H|$ and $|V|$ take the simple form (in reduced co-ordinates x, z, θ)

$$|H| = \begin{vmatrix} \cos\Phi & \sin\Phi \\ -\sin\Phi & \cos\Phi \end{vmatrix}, \qquad |V| = \begin{vmatrix} 1 & \Phi \\ 0 & 1 \end{vmatrix}.$$

In the plane V the prism is equivalent to a drift space of length $L = r_0\Phi$ and has no effect on the trajectories; in the plane H, it behaves like a thick lens. Returning to the unscaled variables, we can easily show that (see Fig. 40)

$$f'_H = -f_H = r_0/\sin\Phi\,,$$

$$g'_H = -g_H = r_0 \operatorname{ctg}\Phi\,,$$

$$s'_H = -s_H = -r_0 \operatorname{tg}(\Phi/2)\,.$$

The principal planes lie at P, the point at which T_0 and T_s intersect. *For* $\varphi = 90°$, we have $f'_H = r_0$, $g'_H = 0$.

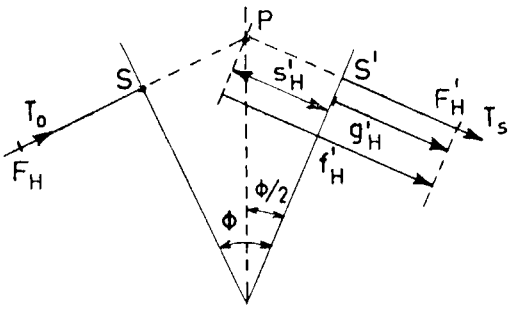


Fig. 40.

Consider now an object P (Fig. 41), at a distance p from S, and its image Q, distant $S'Q = q$ from S'. In reduced co-ordinates, it is easy to calculate the complete transfer matrix $|T|$ between P and Q:

$$|T| = \begin{vmatrix} 1 & q/r_0 \\ 0 & 1 \end{vmatrix} |H| \begin{vmatrix} 1 & p/r_0 \\ 0 & 1 \end{vmatrix}.$$

For a ray emerging from P ($x_P = 0$) at slope $x'_P \neq 0$, we see that at Q, $x_Q = x'_P T_{12} = 0$ so that $T_{12} = 0$ and

$$\operatorname{tg} \Phi = -\frac{(p+q)/r_0}{1-(pq/r_0^2)}.$$

From this relation a very simple rule giving the position of Q (P being known), can be deduced: P, C and Q are collinear (Fig. 41).

Knowing that $T_{12} = 0$ (conjugate points), we have $x_Q = T_{11} x_P$.

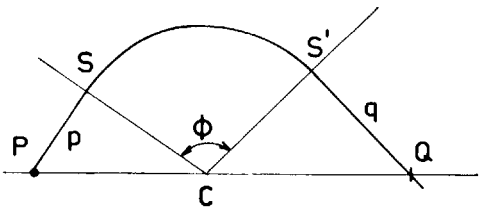


Fig. 41.

The linear magnification G_H is given by

$$G_H = T_{11} = \cos \Phi - \frac{q}{r_0} \sin \Phi ,$$

and for the symmetrical case ($p = q$), we have $G_H = -1$

$$p = r_0 \operatorname{ctg}(\Phi/2) .$$

Second case ($n = \frac{1}{2}$). – In both H and V planes, we have

$$f' = -f = r_0 \sqrt{2}/\sin(\Phi/\sqrt{2}) ,$$
$$g' = -g = r_0 \sqrt{2} \operatorname{ctg}(\Phi/\sqrt{2}) .$$

If $\Phi = \pi/\sqrt{2} \simeq 127°$, the foci are situated at S and S', and $f' = r_0\sqrt{2}$. Symmetrical operation ($p = q$) is obtained for

$$p = q = r_0 \sqrt{2} \operatorname{ctg}(\Phi/2\sqrt{2}) \quad \text{and} \quad G_H = G_V = -1 .$$

For a given angle Φ, and the same radius of curvature r_0, a prism of index $n = \frac{1}{2}$ is *less convergent* in the H plane than the corresponding prism with $n = 0$, but has the advantage of convergence in the V plane.

The field with index n will be obtained with poles having hyperbolic cross-section; in order to simplify their mechanical construction, the hyperbola is usually replaced by its tangent at the point ($x = 0$, $z = h$). The pole pieces are then portions of cones of revolution, with meridian

$$z = \frac{h}{r_0}(1 + nx) .$$

For $n = \frac{1}{2}$, the slope of the tangent is (Fig. 42)

$$\operatorname{tg} \alpha = h/2r_0 .$$

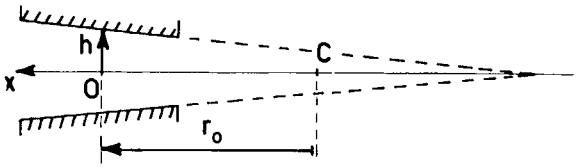


Fig. 42.

4'3.4. *Prism with inclined entry and exit faces.* – Let us now suppose that the entry and exit faces are turned (about S and S') through angles α and β respectively (Fig. 43). By convention, the angle of rotation is taken to be

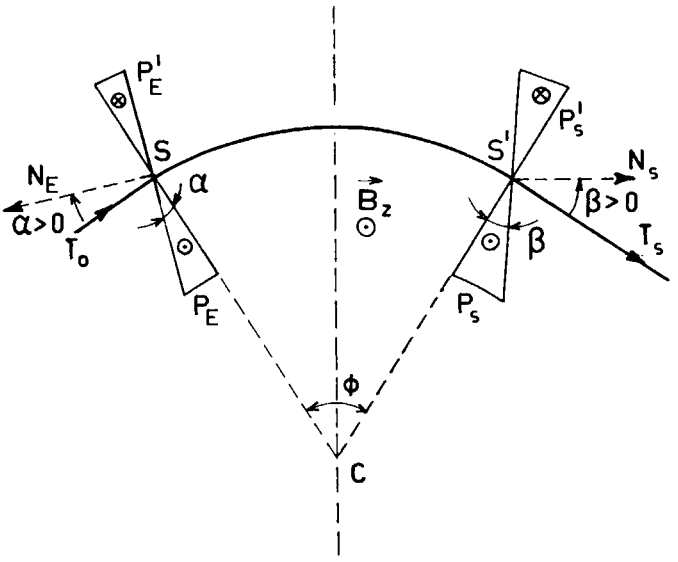


Fig. 43.

positive when the normal to the tilted face is outside the mean trajectory $T_0(C_0)T_s$. For simplicity we only consider the prism with $n = 0$. The rotation of the faces modify the convergence in the H plane. The whole prism is equivalent to a prism (P) of angle Φ, having its faces normal to T_0, T_s, with two small thin prisms P_E, P_S added in S and S' and two prisms P'_E, P'_S removed. In P_E, P_S, the field B_z has the same amplitude and sense as the main field in (P); we may also consider that we *add* P'_E, P'_S, but in these prisms B_z has the same amplitude as in (P), but is in the opposite direction. So, each double prism $(P_E P'_E, P_S P'_S)$ will be equivalent to a thin lens of focal length

$$f_{E_H} = -r_0 \operatorname{ctg}\alpha\,, \quad f_{S_H} = -r_0 \operatorname{ctg}\beta$$

and the total transfer matrix of the real prism (P') with inclined end faces is therefore given by

$$|H'| = \begin{vmatrix} 1 & 0 \\ \operatorname{tg}\beta & 1 \end{vmatrix} |H| \begin{vmatrix} 1 & 0 \\ \operatorname{tg}\alpha & 1 \end{vmatrix}.$$

In a symmetrical prism ($\alpha = \beta$), one obtains

$$|H'| = \begin{vmatrix} \dfrac{\cos(\Phi - \alpha)}{\cos\alpha} & \sin\Phi \\ -\dfrac{\sin(\Phi - 2\alpha)}{\cos^2\alpha} & \dfrac{\cos(\Phi - \alpha)}{\cos\alpha} \end{vmatrix}.$$

When $\Phi = \alpha + \beta$, the entry and exit faces are parallel: the prism (P') then behaves as a « parallel plate »; in the H plane, its convergence is zero since $H_{21} = 0$. If $\alpha + \beta < \Phi$ the prism is convergent in the H plane. It is more tedious to obtain the transfer matrix V; following the same method, one can show that the double prisms behave like thin lens, having focal distances

$$f_{Ev} = + r_0 \operatorname{ctg}\alpha, \qquad f_{Sv} = + r_0 \operatorname{ctg}\beta$$

and

$$|V'| = \begin{vmatrix} 1 & 0 \\ -\operatorname{tg}\beta & 1 \end{vmatrix} |V| \begin{vmatrix} 1 & 0 \\ -\operatorname{tg}\alpha & 1 \end{vmatrix}.$$

A tilted face has the same effect as a thin strong focusing lens: divergent in the H plane, and convergent in the V plane (or conversely, depending on the sign of α or β). The prism (P') with $n = 0$ becomes convergent in the V plane, and remains convergent in the H plane. With (P') we can achieve stigmatic operation, but, unlike the case $n = \frac{1}{2}$, the system is stigmatic for only one particular pair of conjugate planes, since the cardinal elements (f and g) are different in H and V. In the case of a symmetric system ($\alpha = \beta$, $p = q$), the values of $\operatorname{tg}\alpha$ and p/r_0 for which this type of operation is possible are given by $T_{H12} = 0$, $T_{V12} = 0$. We obtain

$$\operatorname{tg}\alpha = r_0/p = \tfrac{1}{2}\operatorname{tg}(\Phi/2).$$

If $\Phi = \pi/2$, $\operatorname{tg}\alpha = \frac{1}{2}$ (or $\alpha = 26° 34'$), $p = 2r_0$.
A device incorporating a stigmatic prism of this type ($\Phi = \pi/2$) has been successfully used by Castaing and Slodzian to transport the image in ion microscopy, and to act as a mass selector simultaneously.

4'3.5. *Trajectories for particles of momentum* ($p_0 + \Delta p$). – We now consider a particle incident along T_0, but having a momentum

$$p = p_0 + \Delta p = p_0(1 + \Delta p/p_0) = p_0(1 + \delta).$$

It is easy to find the trajectory in (P), by the following method: the circular equilibrium trajectory corresponding to p will be a circle (C) of radius $r = r_0 + \mathrm{d}r$ such that

$$(r_0 + \mathrm{d}r)B_z = r_0(1 + x)B_z = r_0(1 + x)(1 - nx)B_0 = \frac{p_0(1 + \delta)}{e},$$

x being small, one finds

$$x \simeq \delta/(1 - n).$$

With respect to the new curved axis (C), the initial values where the particle trajectory T_0 enters the prism are as follow: $x_0 = -\delta/(1-n)$, $x_0' = 0$; using the transfer matrix $|H|$ of the prism, and then coming back to the curved axis (C_0), corresponding to p_0, we find at the exit of the prism

$$x_s = \left\{1 - \left[\cos\left(\sqrt{1-n}\,\Phi\right)\right]\frac{\delta}{1-n}\right\}, \qquad x_s' = \frac{\delta}{\sqrt{1-n}}\sin\left(\sqrt{1-n}\,\Phi\right).$$

Case $n = 0$	*Case* $n = \frac{1}{2}$
$x_s = \delta(1 - \cos\Phi)$ $x_s' = \delta \sin\Phi$	$x_s = 2\delta[1 - \cos(\Phi/\sqrt{2})]$ $x_s' = 2\delta \sin(\Phi/\sqrt{2})$

Momentum dispersion. The *dispersive* power of a prism is its ability to separate, in the image plane, particles of slightly different momenta, originating in the same object point on the axis ($x_0 = 0$) by an amount x_i, and is characterized by the quantity

$$D_p = x_i p_0/\Delta p = x_i/\delta.$$

Using the expressions of x_s and x_s', x_i may be known, at the image plane ($S_Q' = q$), and we find

$$D_p = \frac{1}{1-n}(1 - G_H),$$

where G_H is the linear magnification in the H plane. In a symmetric case ($p = q$, $G_H = -1$), we obtain

$$D_p = 2, \qquad \text{for } n = 0, \quad \alpha = \beta = 0,$$

$$D_p = 4, \qquad \text{for } n = \tfrac{1}{2}, \quad \alpha = \beta = 0,$$

and

$$D_p = 4\,, \qquad \text{for } n = 0\,, \quad \alpha = \beta \neq 0\,.$$

The dispersion is doubled by the use of inclined faces, or of a field with $n = \frac{1}{2}$.

Mass dispersion. When the incident beam consists of ions of different masses, all carrying the same charge e, and accelerated through the same potential φ_0, the different masses can be separated. A mass dispersion can be defined by the relation

$$D_m = x_i m_0/\Delta m\,,$$

where Δm is the difference in mass between two ions in the vicinity of the mass m_0. If $\Delta m \ll m_0$, the relation $mv = p = \sqrt{2em\varphi_0}$ implies

$$\frac{\Delta p}{p} = \frac{1}{2}\frac{\Delta m}{m} \quad \text{and hence } D_m = \frac{1}{2} D_p\,.$$

4·3.6. *Aberrations.* – We have obtained the first order optical properties with the aid of a rectangular model; we have neglected:

— terms of higher order in the expansions of B_z and B_r,

— transverse velocities v_z and v_x,

— the effect of the component B_θ that is present in the stray fields, and which affects v_z and v_x,

— the curvature of the entry and exit faces of the prism.

The dominant aberrations terms that appear when we take all these factors into account are of second order, and they have been extensively studied. Just as for lenses, the most important aberration, the defect that limits the resolving power of the prism is the aperture aberration in the H plane.

Taking into account the momentum spread δ, Δx_i is given by

$$\Delta x_i = M_{11}x_0^2 + M_{12}x_0x_0' + M_{13}x_0'^2 + M_{14}z_0^2 + M_{15}z_0z_0' + M_{16}z_0'^2 + \\ + N_{11}x_0\,\delta + N_{12}x_0'\,\delta + N_{13}\,\delta^2\,.$$

The M_{1k} corresponds to geometrical aberrations, and N_{1j} to chromatic aberrations. In devices, such as mass or velocity spectrometers, in which a high

resolving power is required in the H plane, sources (objects) that are very narrow in the x direction are employed; the terms in x_0^2 and $x_0 x_0'$ can then be neglected. Furthermore, the optics can be designed so that z_0 is also small and that the incident beam is parallel to the H plane ($z_0' = 0$). The principal aperture aberration term is then given by

$$\Delta x_i = M_{13} x_0'^2 \qquad (M_{13} < 0) .$$

It can be shown that

$$|M_{13}| = \tfrac{1}{2}(|G_H| + 1/G_H^2) .$$

Figure 44 shows the trajectories corresponding to $x_0 = 0$, $x_0' \neq 0$, in the vicinity of the image plane. M_{13} depends on the radi of curvature R_1 and R_2 of the faces of the prism; these curvatures can be selected in such a way that the coefficient vanishes.

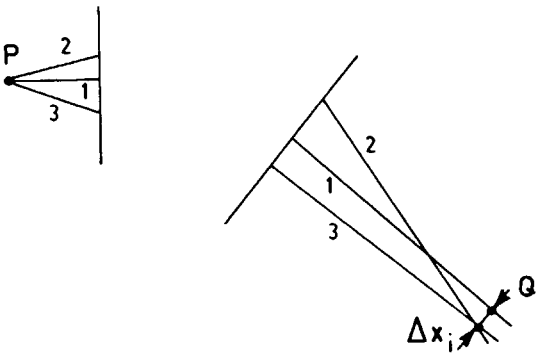


Fig. 44.

Resolving power of the prism. In a prism free of aberrations, all particles of the same momentum p_0, coming from a slit source of width a_0 in the H plane, situated at P, will be concentrated in an « image slit » of width a_i situated at Q (the image of P). We have $a_i = |G_H| a_0$.

In a first approximation, the particles of momentum $p_i = p_0 + \Delta p_i$ will be focused over an image of the same width a_i at a distance $x_i = D_p(\Delta p_i / p_0)$ from the axis.

If we have a single collector behind an analysing slit of width s placed on the axis, we may pass particles of different momenta across this slit by

varying B_0. Particles of momentum p_0 can be completely separated from particles of momentum $p_0 + \Delta p_i$ if

$$r_0 x_i = a_i + s = |G_H| a_0 + s .$$

If now we consider the effect of the aperture aberration, this condition becomes

$$r_0 x_i = a_i + s + M_{13} x_0'^2 r_0 = D_p r_0 \frac{\Delta p_i}{p_0} .$$

The resolving power, that is defined by $R_p = p_0/\Delta p_i$ is then given by the expression (in unscaled co-ordinates)

$$R_p = \frac{D_p r_0}{|G_H| a_0 + s + x_0'^2 (r_0/2)\left(|G_H| + 1/G_H^2\right)} .$$

Case $n = 0$, symmetrical case ($|G_H| = 1$)

$$\alpha = \beta = 0: \quad R_p = \frac{2r_0}{a_0 + s + r_0 x_0'^2} = \frac{2r_0}{\mathscr{D}} , \qquad \alpha = \beta \neq 0: \quad R_p = \frac{4r_0}{\mathscr{D}} .$$

Case $n = \frac{1}{2}$, symmetrical case: $R_p = 4r_0/\mathscr{D}$.

When the prism is used for mass separation, we can define a corresponding resolving power $R_m = m_0/\Delta m$, and we have $R_m = \frac{1}{2} R_p$.

4·4. Electrostatic prisms.

We shall examine now very briefly this family of prisms, the methods used for their study being the same as those used for magnetic prisms.

We restrict our study to nonrelativistic particles.

4·4.1. *Description.* – An electrostatic prism consists of a portion of a condenser, with two metal electrodes having an axis of revolution Z and a plane of mechanical and electrical symmetry (called the H plane, as before). The electrodes coincide with part of the surface of a torus (see Fig. 45), and they therefore have a double curvature: in the H plane and in the vertical plane (r, Z). The angle of the prism is Φ and the mean radius of curvature in the H plane is r_0. The electrodes are held at potentials $\pm V_1$,

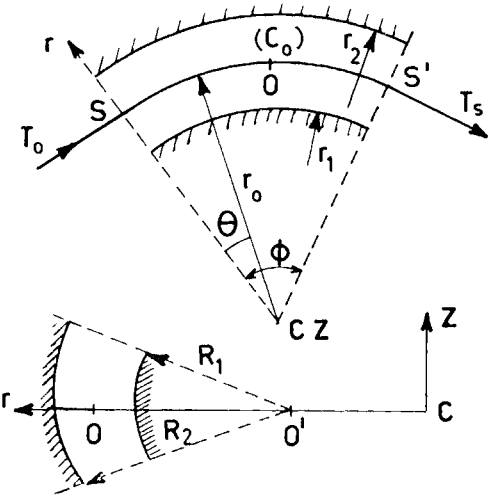


Fig. 45.

symmetric with respect to earth ($V = 0$); the mean equipotential surface ($V = 0$) intersects H along a circle of radius r_0. We assume that the separation of the electrodes $\Delta r = r_2 - r_1 = R_2 - R_1$ (see Fig. 45) is small in comparison with r_1 and R_1. So, we have $r_0 = (r_2 + r_1)/2$ and $R_0 = (R_2 + R_1)/2$.

We adopt a rectangular model for the field. ($E = 0$ outside the prism, and $E = \text{const}$ within it). In the H plane, the field is wholly radial.

In Gaussian optics, and using reduced co-ordinates, E_r may be written in the form

$$E_r = E_0(1 - nx),$$

$$E_z = -E_0(1 - n)z,$$

where $n = 1 + r_0/R_0$.

For a cylindrical condenser ($R_0 = \infty$), we have $n = 1$, ($E_z = 0$), and for a spherical one ($R_0 = r_0$), $n = 2$.

4·4.2. *Trajectories. Optical properties.* – A particle of mass m, charge e and kinetic energy $W_0 = \frac{1}{2}mv_0^2 = e\varphi_0$ travelling in the H plane along T_0, enters the prisms at S and follows the circle (C_0), if the value of E_0 is such that

$$\frac{mv_0^2}{r_0} = eE_0, \quad \text{or} \quad E_0 r_0 = 2\varphi_0.$$

The velocity of the particle will not remain equal to v_0, except on (C_0). Solving the equations of motion, we finally obtain the equation of paraxial trajectories in reduced co-ordinates

$$\left.\begin{aligned} &\frac{d^2x}{d\theta^2} + (3-n)x = 2\beta\,, \\ &\frac{d^2z}{d\theta^2} + (n-1)z = 0\,, \end{aligned}\right\} \quad (\beta = \Delta v/v_0)\,.$$

The trajectories oscillate about the curved axis (C_0) provided that $(3-n)>0$ and $(n-1)>0$. The transfer matrices can be easily obtained, for a monoenergetic beam. Unlike the case of magnetic prisms, the mass of the particles does not occur in the motion equations. Particles of different masses that have been accelerated through the same voltage φ_0 follow the same path through the prism: electrostatic prisms cannot be used to separate particles of different masses emitted by an ion source. They can however be used as *velocity analyzers* when all the particles are *of the same type*. In this case, we have

$$\frac{\Delta v}{v_0} = \frac{1}{2}\frac{\Delta\varphi}{\varphi_0}\,.$$

The optical elements and the energy dispersion D_e can be calculated. Returning to real co-ordinates, we obtain (see Fig. 46)

H-plane:

$$\left.\begin{aligned} f'_H &= -f_H = r_0/\sin(\omega_H\theta)\,, \\ g'_H &= -g_H = r_0/\mathrm{tg}\ (\omega_H\theta)\,, \end{aligned}\right\} \qquad \omega_H = \sqrt{3-n}\,,$$

V-plane:

$$\left.\begin{aligned} f'_V &= -f_V = r_0/\sin(\omega_V\theta)\,, \\ g'_V &= -g_V = r_0/\mathrm{tg}\ (\omega_V\theta)\,, \end{aligned}\right\} \qquad \omega_V = \sqrt{n-1}\,.$$

The electrostatic prism, like the magnetic prism is equivalent to two centred systems of length $L = r_0\Phi$. If ω_H and ω_V are real, both systems are convergent.

The properties of cylindrical prisms are thus similar to those of the magnetic prism with homogeneous field, since $f_V = g_V = \infty$ $(n = 1)$. The foci

are in the entry and exit planes ($g_H = 0$), for $\Phi_1 = \pi/2\sqrt{2}$. By joining two prisms of angle Φ_1, we can focus particles from a point source at S at the point S' (see Fig. 47).

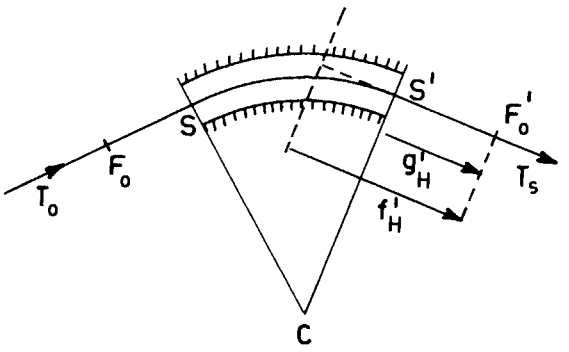


Fig. 46.

The image of S will be a segment of a straight line (in first order), since there is no convergence in the V plane, and the total deflection will be

$$\Phi = 2\Phi_1 = \pi/\sqrt{2} \simeq 127^\circ .$$

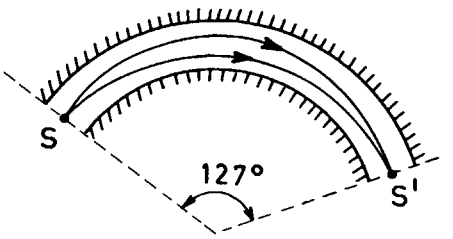


Fig. 47.

The spherical prism is equivalent to a rotationally symmetric optical system, since $\omega_H = \omega_V = 1$ (like the magnetic prism with index $n = \frac{1}{2}$) and we have

$$f'_H = r_0/\sin\Phi ,$$
$$g'_H = r_0/\mathrm{tg}\,\Phi .$$

Particles from S can be focused at S' if $\Phi = 180^\circ$, but we now have a point image.

The dispersion D_e is defined, in the Gaussian image plane by

$$D_e = \frac{\Delta x_i}{\Delta\varphi/\varphi_0}.$$

It can be shown that

$$D_e = (1 - G_H)/\omega_H^2.$$

With symmetric operations, $D_e = 1$ for a cylindrical prism, $D_e = 2$ for a spherical one, and the real separation d_c between particles of energies φ_0 and $\varphi_0 + \Delta\varphi$ in the Gaussian image is given by

$$d_c = r_0 \Delta x_i = r_0 D_e \Delta\varphi/\varphi_0 = r_0 \Delta\varphi/\varphi_0 \quad \text{(cylindrical)},$$

or

$$d_c = 2r_0 \Delta\varphi/\varphi_0 \quad \text{(spherical)}.$$

Aberrations of electrostatic prisms are not yet known in the general case, as they are for the magnetic prisms. We can only give an approximate value of the resolution R_e in a prism in which the image of S is formed at S', neglecting all the other aberration terms in comparison with the aperture aberration term $M_{13}x_0'^2$.

Spherical condenser ($\Phi = 180°$, $|G_H| = 1$)

$$R_e = \frac{r_0 D_e}{a_0 + s + M_{13} r_0 x_0'^2} = \frac{2r_0}{a_0 + s + 2r_0 x_0'^2},$$

for example, with $r_0 = 5$ cm, $a_0 = 0.25$ mm, $s = 0.25$ mm (circular holes as object and image), we have

$$R_e \simeq 200(1 - 200\, x_0'^2).$$

Cylindrical condenser ($\Phi = 127°$, $|G_H| = 1$)

$$R_e = r_0/(a_0 + s + \tfrac{4}{3} r_0 x_0'^2).$$

With the same values of r_0, a_0 and s (slits):

$$R_e \simeq 100(1 - 133\, x_0'^2).$$

4·4.3. *Vertical focusing in a cylindrical prism.* – Vertical focusing may be present only if the mean equipotential surface $V = 0$ (corresponding to the energy φ_0 of the particles has a nonzero curvature (R_0 finite). It is possible to obtain such a curvature with the help of two supplementary electrodes, at the top and the bottom of the cylindrical prism (Fig. 48), held at a common

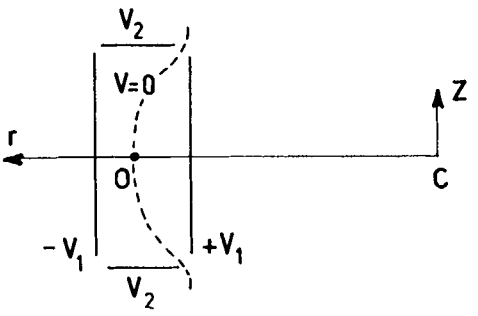


Fig. 48.

potential V_2, slightly different of $V = 0$. Measurements in an electrolytic tank give a value of R_0, for different values of V_2 (and $\pm V_1$).

REFERENCES

W. GLASER: *Grundlagen der Elektronenoptik*, Springer (1952).

P. GRIVET: *Electron Optics* (1965). (English edition, revised by A. Septier), Pergamon. A second edition, completed by A. Septier, will be published in 1971.

P. W. HAWKES: *Quadrupole Optics* (1966); *Springer Tracts in Modern Physics*, vol. **42**.

A. SEPTIER ed.: *Focusing of charged particles* (2 vol.), Academic Press (1967). This book contains 21 chapters, covering the whole field of particle optics, including systems with curved axis, and space charge effect. (More than 1200 references are given in these books.)

Proceedings of the 4th European Regional Conference on Electron Microscopy, Rome 1968 (Rome, 1968).

Proceedings of the 7th International Conference on Electron Microscopy, Grenoble 1970 (Paris, 1970).

Problems on Geometrical Electron Optics

A. SEPTIER

Institut d'Electronique Fondamentale, Laboratoire associé au CNRS
Faculté des Sciences - Orsay, France

1. Electrostatic lenses.

1'1. Problems.

Problem 1. Obtain by means of Laplace's equation

$$\frac{\partial^2 \varphi}{\partial r^2} + \frac{1}{r}\frac{\partial \varphi}{\partial r} + \frac{\partial^2 \varphi}{\partial z^2} = 0$$

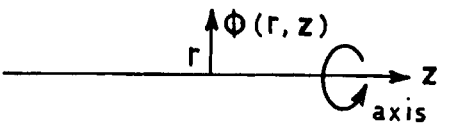


Fig. 1.

a series solution for the scalar potential $\varphi(r, z)$ in a cylindrical region near the symmetry axis Oz. (Eq. (11) of p. 17.)

Problem 2. Using for $\varphi_0(z)$ in a two-cylinder lens the simplified function given on the Fig. 2, solve the trajectory equation (16 bis) (see p. 18)

$$r'' + r'\frac{\varphi'}{2\varphi} + r\frac{\varphi''}{4\varphi} = 0\,,$$

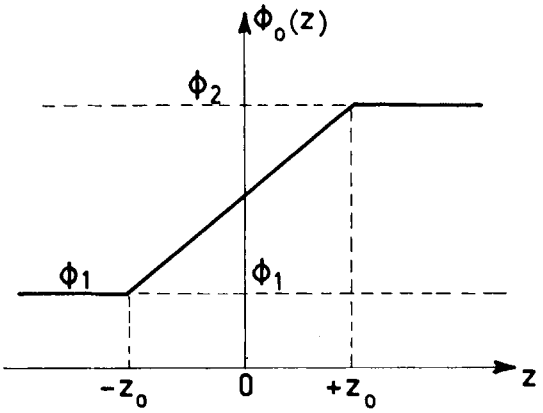


Fig. 2.

a) in a short interval enclosing the field discontinuity at $(-z_0)$ or $(+z_0)$;

b) between $(-z_0)$ and $(+z_0)$.

Give the transfer matrix of the lens, the values of the focal distances and positions of the foci.

1·2. Solutions.

Problem 1. Due to axial symmetry, the potential function $\varphi(r, z)$ may be expressed as a power series of r, in which all odd-power terms in r must be zero:

$$\varphi(r, z) = \sum_{n=0}^{\infty} A_n(z) r^{2n}, \qquad n = 1, 2, 3 \ldots$$

If this expression is substituted in the Laplace's equation

$$\frac{\partial^2 \varphi}{\partial r^2} + \frac{1}{r}\frac{\partial \varphi}{\partial r} + \frac{\partial^2 \varphi}{\partial z^2} = 0,$$

we find

$$\sum_{n=0}^{\infty} [A_n''(z) r^{2n} + 2n(2n-1) A_n(z) r^{2n-2} + 2n A_n(z) r^{2n-2}] = 0.$$

By grouping like powers of r, it can be seen that:

$$\sum_{n=0}^{\infty} [A''_{n-1}(z) + 4n^2 A_n(z)] r^{2n-2} = 0 .$$

The coefficients A_n are then related by the following recurrence relation:

$$A_n(z) = - A''_{n-1}(z)/4n^2 = - A''_{n-1}(z)/(2n)^2 .$$

By applying this relation, the coefficients are found to be

$$A_0(z) = \varphi_0(z) ,$$

$$A_1(z) = -\frac{1}{4.1^2} \cdot \varphi''_0(z) ,$$

$$A_2(z) = -\frac{1}{4.2^2} A''_1(z) = + \frac{1}{4^2(1.2)^2} \cdot \varphi''''_0(z) ,$$

....

$$A_n(z) = (-1)^n \frac{1}{2^{2n}(n!)^2} \cdot \varphi_0^{(2n)}(z) .$$

The potential distribution, around the axis, takes the following form:

$$\varphi(r, z) = \varphi_0(z) - \frac{r^2}{4} \varphi''_0(z) + \frac{r^4}{64} \varphi''''_0(z) - \dots .$$

Problem 2. The true potential function $\varphi_0(z)$ is replaced by a sequence of regions of uniform field (Fig. 3). The integration of the motion equation may

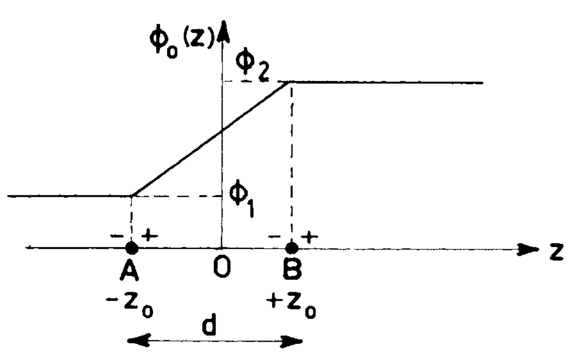


Fig. 3.

be divided into two parts: 1) the integration over an infinitesimal distance enclosing the « break points » A (and B), from A^- to A^+; 2) the integration from one end of the linear segment AB to the other.

1) At the junction between the field-free region ($z<-z_0$) and the lens, $\varphi_0''(z)$ becomes infinite, whereas $\varphi_0(z)$, $\varphi_0'(z)$, r and r' all remain finite. φ and r are continuous and they may be considered as constants: the junction is equivalent to a thin lens (only r' varies).

The integration across the junction

$$\int_{z_{A^-}}^{z_{A^+}} r'' \mathrm{d}z = -\int_{z_{A^-}}^{z_{A^+}} \left(\frac{\varphi'}{2\varphi} r' + \frac{\varphi''}{4\varphi} r\right) \mathrm{d}z$$

reduces to

$$r'_{A^+} - r'_{A^-} = -\int_{z_{A^-}}^{z_A} \left(\frac{\varphi'}{2\varphi} r' + \frac{\varphi''}{4\varphi} r\right) \mathrm{d}z - \int_{z_A}^{z_{A^+}} \left(\frac{\varphi'}{2\varphi} r' + \frac{\varphi''}{4\varphi} r\right) \mathrm{d}z .$$

From z_{A^-} to z_A:

$$\varphi' = \varphi'_{A^-}, \quad r = r_{A^-} = r_A, \quad \varphi_{A^-} = \varphi_A = \varphi_1,$$

and from z_A to z_{A^+}:

$$\varphi' = \varphi'_{A^+}, \quad r = r_{A^+} = r_A, \quad \varphi_{A^+} = \varphi_A = \varphi_1 .$$

We obtain

$$r'_{A^+} = r'_{A^-} + \frac{\varphi'_{A^-} - \varphi'_{A^+}}{4\varphi_A} .$$

We know that

$$\varphi'_{A^+} = \frac{\varphi_B - \varphi_A}{2z_0} = \frac{\varphi_2 - \varphi_1}{2z_0} \quad \text{and} \quad \varphi'_{A^-} = 0.$$

Finally

$$\left.\begin{aligned} r_{A^+} &= r_{A^-}, \\ r'_{A^+} &= -r_{A^-} \cdot \frac{\varphi_2 - \varphi_1}{8\varphi_1 z_0} + r'_{A^-} . \end{aligned}\right\}$$

At the junction B, one would have, (with $\varphi'_{B^+}=0$, $\varphi'_{B^-}=(\varphi_2-\varphi_1)/2z_0$)

$$\left.\begin{aligned} r_{B^+} &= r_{B^-}\,, \\ r'_{B^+} &= +\,r_{B^-}\cdot\frac{\varphi_2-\varphi_1}{8\varphi_2 z_0}+r'_B\,. \end{aligned}\right\}$$

The transfer matrices of these thin lenses are given by

$$|T_A|=\begin{vmatrix} 1 & 0 \\ -\dfrac{\gamma-1}{8z_0} & 1 \end{vmatrix}, \qquad |T_B|=\begin{vmatrix} 1 & 0 \\ +\dfrac{\gamma-1}{8\gamma z_0} & 1 \end{vmatrix},$$

where $\gamma=\varphi_2/\varphi_1>1$.
$|T_A|$ corresponds to a *converging* lens, of convergence

$$C_A=\frac{1}{f_A}=\frac{\gamma-1}{8z_0}$$

and $|T_B|$ to a *diverging* lens such as

$$C_B=\frac{1}{f_B}=-\frac{\gamma-1}{8\gamma z_0}\,.$$

2) Within the segment AB, $\varphi''=0$, and the general equation becomes

$$r''+\frac{1}{2}\frac{\varphi'}{\varphi}r'=0\,, \quad \text{or} \quad \varphi^{\frac{1}{2}}r''+\frac{\varphi'}{2\varphi^{\frac{1}{2}}}r'=0\,,$$

which is integrated by $r'\varphi^{\frac{1}{2}}=C$.

C is a constant determined by the values of r' and φ at the initial (or final) point of the segment AB

$$C=r'_{A^+}\cdot\varphi_A^{\frac{1}{2}}=r'_{B^-}\cdot\varphi_B^{\frac{1}{2}}\,.$$

Between A and B

$$\varphi_0(z)=\varphi_A+(z-z_A)\varphi'_{A^+}=\varphi_1+(z+z_0)\varphi'_{A^+}$$

and

$$\varphi'_{A^+}=(\varphi_2-\varphi_1)/2z_0\,.$$

A further integration yields the following value for r_B (to the left of B):

$$r_B = r_{A^+} + \int_{z_A}^{z_B} \frac{C}{[\varphi_A + (z + z_0)\varphi'_{A^+}]^{\frac{1}{2}}}\,\mathrm{d}z\,,$$

$$r_B = r_{A^+} + \frac{2C}{\varphi'_{A^+}}(\varphi_B^{\frac{1}{2}} - \varphi_A^{\frac{1}{2}}) = r_{A^+} + r'_{A^+}\cdot\frac{4\varphi_1^{\frac{1}{2}} z_0(\varphi_2^{\frac{1}{2}} - \varphi_1^{\frac{1}{2}})}{(\varphi_2 - \varphi_1)}\,,$$

or

$$\left.\begin{aligned} r_B &= r_{A^+} + r'_{A^+}\cdot\frac{4\varphi_1^{\frac{1}{2}} z_0}{\varphi_2^{\frac{1}{2}} + \varphi_1^{\frac{1}{2}}}\,, \\ r'_B &= r'_{A^+}\cdot\left(\frac{\varphi_1}{\varphi_2}\right)^{\frac{1}{2}}. \end{aligned}\right\}$$

The transfer matrix of the segment AB may be written

$$|T_{AB}| = \begin{vmatrix} 1 & \dfrac{4z_0}{1+\gamma^{\frac{1}{2}}} \\ 0 & \dfrac{1}{\gamma^{\frac{1}{2}}} \end{vmatrix}.$$

The total transfer matrix of the whole lens is obtained by multiplying the three individual matrices:

$$|T| = |T_B|\,|T_{AB}|\,|T_A|\,.$$

We recall that:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}\begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} = \begin{vmatrix} (a_{11}b_{11} + a_{12}b_{21}) & (a_{11}b_{12} + a_{12}b_{22}) \\ (a_{21}b_{11} + a_{22}b_{21}) & (a_{21}b_{12} + a_{22}b_{22}) \end{vmatrix}.$$

We have

$$|T| = \begin{vmatrix} 1 & 0 \\ \dfrac{\gamma - 1}{8\gamma z_0} & 1 \end{vmatrix}\begin{vmatrix} 1 & \dfrac{4z_0}{\gamma^{\frac{1}{2}} + 1} \\ 0 & \dfrac{1}{\gamma^{\frac{1}{2}}} \end{vmatrix}\begin{vmatrix} 1 & 0 \\ -\dfrac{\gamma - 1}{8z_0} & 1 \end{vmatrix},$$

or

$$|T| = \begin{vmatrix} \dfrac{3 - \gamma^{\frac{1}{2}}}{2} & \dfrac{4z_0}{\gamma^{\frac{1}{2}} + 1} \\ \dfrac{-3(\gamma - 1)(\gamma^{\frac{1}{2}} - 1)}{16\gamma z_0} & \dfrac{3\gamma^{\frac{1}{2}} - 1}{2\gamma} \end{vmatrix} = \begin{vmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{vmatrix}.$$

Consider (Fig. 4) an incident ray (T_1) parallel to the axis ($r_0' = 0$). Thus

$$r_S = r_0 \frac{3 - \gamma^{\frac{1}{2}}}{2},$$

$$r_S' = -\frac{3r_0}{16z_0} \frac{(\gamma - 1)(\gamma^{\frac{1}{2}} - 1)}{\gamma}.$$

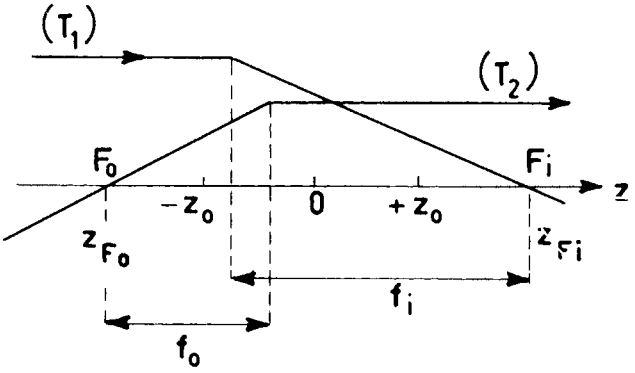


Fig. 4.

The image focal distance is given by

$$f_i = -\frac{r_0}{r_S'} = \frac{16z_0}{3} \frac{\gamma}{(\gamma - 1)(\gamma^{\frac{1}{2}} - 1)}$$

and the location of the image focal point by:

$$z_{F_i} - z_0 = -\frac{r_S}{r_S'} = \frac{8z_0}{3} \frac{\gamma(3 - \gamma^{\frac{1}{2}})}{(\gamma - 1)(\gamma^{\frac{1}{2}} - 1)}.$$

The object focal distance f_0 is known thanks to the relations

$$\frac{f_i}{f_0} = \sqrt{\frac{\varphi_2}{\varphi_1}} = \gamma^{\frac{1}{2}}.$$

Then:

$$f_0 = \frac{16z_0}{3} \frac{\gamma^{\frac{1}{2}}}{(\gamma - 1)(\gamma^{\frac{1}{2}} - 1)}.$$

A trajectory T_2 crossing the axis at F_0, leaves the lens (at B of Fig. 3) parallel to the axis ($r'_S = 0$). We may write:

$$r_S = r_0 T_{11} + r'_0 T_{12},$$

$$r'_S = r_0 T_{21} + r'_0 T_{22} = 0,$$

$$-\left(\frac{r_0}{r'_0}\right) = \frac{T_{22}}{T_{21}} = z_{F_0} - z_A = z_{F_0} + z_0 .$$

The object focal point F_0 is thus located at z_{F_0}

$$z_{F_0} = -\left(z_0 + \frac{8(3\gamma^{\frac{1}{2}} - 1)z_0}{3\gamma(\gamma - 1)(\gamma^{\frac{1}{2}} - 1)}\right).$$

2. Round magnetic lenses.

2·1. Problems.

Problem 1. In a long solenoid the function $B_0(z)$ may be approximated by a rectangular model of length L (Fig. 5), with

$$B_m L = \int_{-\infty}^{+\infty} B_0(z)\,\mathrm{d}r .$$

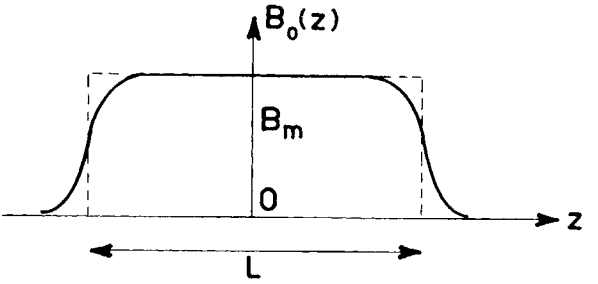


Fig. 5.

Using the eq. (29) of p. 33 determine:

— the asymptotic optical elements,
— the immersed elements.

Example:

$$B_m = 2 \text{ Tesla}, \quad L = 1 \text{ cm}, \quad \varphi_0 = 1 \text{ MV}.$$

Problem 2. In a short lens, $B_0(z)$ being given by the bell-shaped model

$$B_0(z) = B_m \Big/ \left[1 + \left(\frac{z}{a}\right)^2\right],$$

solve the equation of motion:

$$r'' + \frac{eB_0^2(z)}{8m_0\varphi_0^*} r = 0.$$

2·2. Solutions.

Problem 1. We have to integrate the equation:

$$r'' + \frac{eB_0^2(z)}{8m_0\varphi_0^*} r = 0.$$

From $z = 0$ to $z = L$, the field $B_0(z)$ is considered as constant: $B_0(z) = B_m$.

By putting

$$\frac{eB_m^2}{8m_0\varphi_0^*} = \omega^2,$$

the trajectory has the form:

$$r(z) = A \cos \omega z + B \sin \omega z$$

and its slope is given by:

$$r'(z) = -A\omega \sin \omega z + B\omega \cos \omega z.$$

The integration constants A and B are easily calculated, if we know the values

of r and r' at $z = 0$ (r_0, r'_0)

$$r(z) = r_0 \cos \omega z + \frac{r'_0}{\omega} \sin \omega z ,$$

$$r'(z) = -r_0 \omega \sin \omega z + r'_0 \cos \omega z .$$

The transfer matrix of a lens of length L may be written:

$$|T| = \begin{vmatrix} \cos \omega L & (1/\omega) \sin \omega L \\ -\omega \sin \omega L & \cos \omega L \end{vmatrix} = \begin{vmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{vmatrix} .$$

« Asymptotic » elements are given by classical relations:

$$f_1 = -1/T_{21} , \qquad \overline{SF_1} = z_{F_1} = -T_{11}/T_{21} .$$

Dimension-free elements are, with $k = \omega L$:

$$\frac{f_1}{L} = \frac{1}{k \sin k} , \qquad \frac{\overline{SF_1}}{L} = \frac{\cos k}{k \sin k} .$$

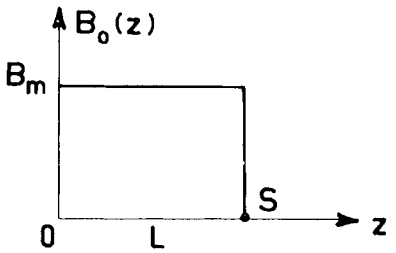


Fig. 6.

If the field B_m increases, φ_0^* being fixed, $\overline{SF_1}$ decreases, vanishes for $k = \pi/2$ and then becomes negative. For $k > \pi/2$, the asymptotic image focal point F_1 is different from the true « immersed » focal point F_0. The location of F_0 corresponds to

$$\omega z_{F_0} = \pi/2$$

(the origin being taken at O, see Fig. 6).

At $z = z_{F_0}$, the slope of a ray entering the lens parallel to the axis, is:

$$r'(z_{F_0}) = -\omega r_0 ,$$

and we can define the « immersed » focal length:

$$f_0 = -r_0/r'(z_{F_0}) = +\frac{1}{\omega},$$

or

$$\frac{f_0}{L} = \frac{1}{k} \quad \text{and} \quad \frac{z_{F_0}}{L} = \frac{\pi}{2k}.$$

Example: $B_m = 20$ kG $= 2$ Tesla, $L = 10^{-2}$ m, $e/m_0 \simeq 1.8 \cdot 10^{11}$ Coul/kg,

$$\omega = \left(\frac{1.8 \cdot 10^{11}}{8}\right)^{\frac{1}{2}} \cdot \frac{B_m}{\sqrt{\varphi_0^*}} = \frac{3 \cdot 10^5}{\sqrt{\varphi_0^*}} \quad \text{and} \quad k = \frac{3 \cdot 10^3}{\sqrt{\varphi_0^*}}.$$

For electrons accelerated under a voltage $\varphi_0 = 1$ MV $= 10^6$ V ($\varphi_0^* \simeq 2.10^6$ V) we obtain

$$k \simeq 2.1 \text{ rad} \simeq 120°,$$

$$\cos k \simeq -0.5,$$

$$\sin k \simeq +\sqrt{3}/2 \simeq +0.866,$$

$$f_1/L \simeq 0.55, \qquad f_1 \simeq 5.5 \text{ mm},$$

$$\overline{SF_1}/L \simeq -0.27, \qquad SF_1 \simeq -2.7 \text{ mm},$$

or

$$f_0/L \simeq 0.475, \qquad f_0 \simeq 4.75 \text{ mm},$$

$$\overline{SF_0} \simeq 0.77, \qquad SF_0 \simeq -2.3 \text{ mm}.$$

Problem 2. The paraxial ray equation may be written in a dimension-free form by putting:

$$x = \frac{z}{a}, \quad y = \frac{r}{a}, \quad k^2 = \frac{eB_m^2 a^2}{8 m_0 \varphi_0^*}$$

($2a$, is the half-value width of the field distribution $B_0(z) = B_m/(1 + (z/a)^2)$),

$$y'' + \frac{k^2 y}{(1 + x^2)^2} = 0. \tag{1}$$

The introduction of the new independent variable

$$x = \operatorname{ctg}\varphi \tag{2}$$

(such as $dx = -d\varphi/\sin^2\varphi$, $(1 + x^2) = 1/\sin^2\varphi$) brings this equation into the form:

$$\frac{d^2y}{d\varphi^2} + 2\operatorname{ctg}\varphi\frac{dy}{d\varphi} + k^2y = 0\,. \tag{3}$$

Using the new variable $R = r\sin\varphi$, (3) becomes, after a tedious calculation

$$\frac{d^2R}{d\varphi^2} + (1 + k^2)R = 0\,. \tag{3 bis}$$

Writing $\Omega^2 = 1 + k^2$, we obtain

$$R = A\sin\Omega\varphi + B\cos\Omega\varphi\,.$$

In terms of the original co-ordinate r:

$$r = \frac{1}{\sin\varphi}[A\sin\Omega\varphi + B\cos\Omega\varphi]\,, \qquad \text{with } \varphi = (\operatorname{arc\,ctg})\frac{z}{a}\,. \tag{4}$$

When z varies from $+\infty$ to $-\infty$, φ increases from 0 to π.

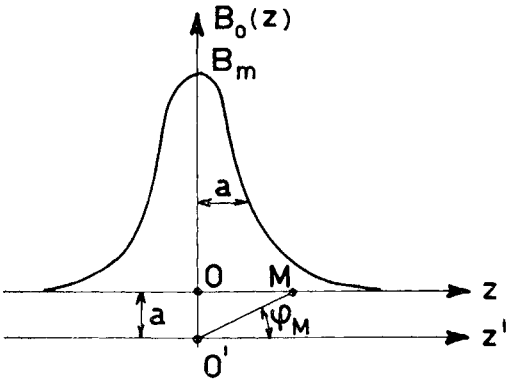


Fig. 7.

Consider a ray incident *from the right* ($z>0$), parallel to the axis. For it, $r(z)$ must remain finite for $x=+\infty$ ($\varphi=0$), so that, in (4)

$$B=0$$

and

$$r(\varphi)=\frac{A\sin\Omega\varphi}{\sin\varphi}, \tag{5}$$

r_0 being the value of r at $\varphi=0$, we obtain $A=r_0\Omega^{-1}$, and

$$r(\varphi)=\frac{r_0}{\Omega}\frac{\sin\Omega\varphi}{\sin\varphi}. \tag{6}$$

For large excitations (strong fields), the ray intersects the axis several times. The points of intersection are given by

$$r(\varphi)=0,$$

or

$$\Omega\varphi=n\pi=(1+k^2)^{\frac{1}{2}}\varphi, \qquad \text{with } n=1,2,3,\ldots,$$

and

$$\varphi_{\max}=\pi.$$

Thus the lens has one focal point for $k^2<3$, two focal points for $k^2<8$, etc. We shall consider only the case $k^2\leqslant 3$. Then $\Omega\varphi_{F_0}=\pi$, or

$$\varphi_{F_0}=\frac{\pi}{\Omega}=\frac{\pi}{\sqrt{1+k^2}}, \tag{7}$$

(φ_{F_0} is the value of φ corresponding to the immersed focal point F_0; see Fig. 8). The slope of the ray $r'=\mathrm{d}r/\mathrm{d}z$, is given by

$$r'=\frac{\mathrm{d}r}{\mathrm{d}\varphi}\cdot\frac{\mathrm{d}\varphi}{\mathrm{d}x}\cdot\frac{\mathrm{d}x}{\mathrm{d}z},$$

$$r'=\left(\frac{r_0}{\Omega}\frac{\Omega\sin\varphi\cos\Omega\varphi-\sin\Omega\varphi\cos\varphi}{\sin^2\varphi}\right)(-\sin^2\varphi)\left(\frac{1}{a}\right),$$

$$r'=-\frac{r_0}{a}\left(\sin\varphi\cos\Omega\varphi-\frac{1}{\Omega}\sin\Omega\varphi\cos\varphi\right). \tag{8}$$

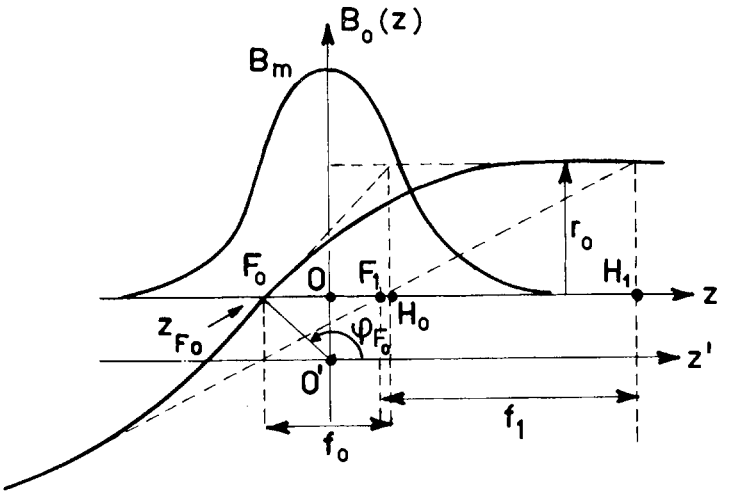


Fig. 8.

At F_0:

$$r'_{F_0} = \frac{r_0}{a} \sin \varphi_{F_0} = \frac{r_0}{a} \sin\left(\frac{\pi}{\sqrt{1+k^2}}\right). \tag{9}$$

From (9)

$$|f_0| = + \frac{r_0}{r'_{F_0}} = \frac{a}{\sin(\pi/\sqrt{1+k^2})}, \qquad z_{F_0} = a \operatorname{ctg} \frac{\pi}{\sqrt{1+k^2}}. \tag{10}$$

The immersed focus F_0 is located at the centre of the lens ($z_{F_0} = 0$) if $\varphi_{F_0} = \pi/2$; this occurs for $\Omega = 2$, or $k^2 = 3$ (condenser-objective).

The asymptotic focal distance also may be calculated. We know that:

$$|f_1| = \frac{r_0}{r'(z = -\infty)} = \frac{r_0}{r'(\varphi = \pi)}.$$

Using (8) we obtain

$$r'(\varphi = \pi) = -\frac{r_0}{a\Omega} \sin \pi\Omega$$

and

$$f_1 = \frac{a\Omega}{\sin \pi\Omega} = \frac{a\sqrt{1+k^2}}{\sin \pi\sqrt{1+k^2}}.$$

For $k^2 = 3$, we have $f_1 \to \infty$.

The asymptote to the trajectory takes the form

$$r_1(z) = Az + B = Aa\frac{\cos\varphi}{\sin\varphi} + B$$

with $A = [r'(z)]_{\varphi=\pi}$

$$r_1(z) = -\frac{r_0}{\Omega}\left[\frac{(\Omega \sin\varphi \cos\Omega\varphi - \sin\Omega\varphi\cos\varphi)\cos\varphi}{\sin\varphi}\right]_{\varphi=\pi} + B.$$

For $z = -\infty$ $(\varphi = \pi)$, we have

$$r_1(z) \equiv r(z) = \frac{r_0}{\Omega}\cdot\frac{\sin\Omega\varphi}{\sin\varphi}$$

and

$$B = \frac{r_0}{\Omega}\left[\frac{\sin\Omega\varphi + \Omega\sin\varphi\cdot\cos\varphi\cos\Omega\varphi - \sin\Omega\varphi\cos^2\varphi}{\sin\varphi}\right].$$

The position z_{F_1} of F_1 is given by $r_1(z_{F_1}) = 0$, or $z_{F_1} = -B/A$, with $A = r'_{(\varphi=\pi)}$. Finally

$$z_{F_1} = a\Omega\cdot\mathrm{ctg}\pi\Omega.$$

3. Quadrupole lenses.

3'1. Problems.

Problem 1. Obtain the expression for the transfer matrices T_X and T_Y of a symmetrical doublet, both lenses having same length L, and same excitation parameter $\beta L = k$.

The distance between lenses is $D/L = \lambda$.

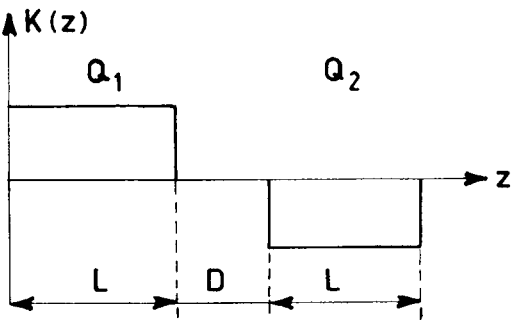


Fig. 9.

Problem 2. Show that it is possible to find a value of k for which the system is equivalent to a round lens. Calculate k in the case $D = 0$. Give the values of f and $\overline{SF}$.

Problem 3. Calculate the values of the corresponding field gradient K_0 in a magnetic lens having $L = 20$ mm, $a = 2$ mm (a, is the distance between poles and axis), for $\varphi_0 = 2$ MV.

How many ampere-turns have to be injected in the coil surrounding each pole-piece?

Problem 4. In these conditions, what would be the energy of protons having the same trajectories in the doublet?

Problem 5. – Calculate the value of k for which a ray entering the lens Q_1 parallel to the axis, leaves the lens Q_2 at S, also parallel to the axis (in both planes XOz or YOz). Give a schematic representation of the trajectories in the doublet. The maximum amplitude X_{max} or Y_{max} of the trajectories being limited by the condition $X_{max} = Y_{max} = a = 2$ mm, give the values of X_{0max} and Y_{0max}.

(We suppose $D = 0$.)

Problem 6. We suppose now $D \neq 0$. Give the expressions of D and k for which the focal points F_X and F_Y coincide at S ($\overline{SF}_X = \overline{SF}_Y = 0$). Has the problem a solution?

3·2. **Solutions.**

Problem 1. The transfer matrices of an individual lens are, in XOz and YOz planes respectively:

$$|T_X|_1 = \begin{vmatrix} \cos k & (1/\beta)\sin k \\ -\beta \sin k & \cos k \end{vmatrix} \quad \text{and} \quad |T_Y|_1 = \begin{vmatrix} \cosh k & (1/\beta)\sinh k \\ \beta \sinh k & \cosh k \end{vmatrix}.$$

For the doublet, the transfer matrices may be obtained by multiplying three matrices

$$|T_X| = |T_Y|_1 \begin{vmatrix} 1 & D \\ 0 & 1 \end{vmatrix} |T_X|_1 \quad \text{and} \quad |T_Y| = |T_X|_1 \begin{vmatrix} 1 & D \\ 0 & 1 \end{vmatrix} |T_Y|_1,$$

and one obtains:

$$|T_X| = \begin{vmatrix} (\cos k \cosh k - \beta D \sin k \cosh k - \sin k \sinh k) & \frac{1}{\beta}(\cosh k \sin k + \beta D \cosh k \cos k + \sinh k \cos k) \\ \beta(\sinh k \cos k - \beta D \sin k \sinh k - \cosh k \sin k) & (\sin k \sinh k + \beta D \sinh k \cos k + \cosh k \cos k) \end{vmatrix},$$

$$|T_Y| = \begin{vmatrix} (\cosh k \cos k + \beta D \sinh k \cos k + \sinh k \sin k) & \frac{1}{\beta}(\sinh k \cos k + \beta D \cosh k \cos k + \cosh k \sin k) \\ \beta(\sinh k \cos k - \beta D \sinh k \sin k - \cosh k \sin k) & (\cosh k \cos k - \beta D \sin k \cosh k - \sinh k \sin k) \end{vmatrix}.$$

Comparing $|T_X|$ and $|T_Y|$, we observe that the following relations hold:

$$T_{11_X} = T_{22_Y}, \quad T_{11_Y} = T_{22_X}, \quad T_{12_X} = T_{12_Y}, \quad T_{21_X} = T_{21_Y}.$$

(One matrix is therefore sufficient to describe the symmetrical doublet.)

Applying the general relation

$$\frac{1}{f} = -T_{21} \quad \left(\text{or} \quad \frac{f}{L} = -\frac{1}{LT_{21}}\right),$$

we have

$$f_X = f_Y = f$$

and

$$\frac{f}{L} = \frac{1}{k} \frac{1}{(\sin k \cosh k - \sinh k \cos k + k\lambda \sin k \sinh k)}.$$

The two focal points F_X and F_Y are not generally coincident, since

$$\frac{\overline{SF}}{L} = -\frac{T_{11}}{T_{21}}$$

(with $T_{11_X} \neq T_{11_Y}$, at least for weak values of k).

Problem 2. The symmetrical doublet is equivalent to a round thick lens, if the two following conditions are simultaneously fulfilled:

$$f_X = f_Y, \tag{11}$$

$$\overline{SF}_X = \overline{SF}_Y. \tag{12}$$

(11) is always true in a symmetrical doublet ($T_{12_X} = T_{12_Y}$),

(12) gives $T_{11_X} = T_{11_Y}$,

so we obtain:

$$\cosh k \cos k - \beta D \sin k \cosh k - \sin k \sinh k =$$

$$= \cosh k \cos k + \beta D \sinh k \cos k + \sinh k \sin k ,$$

or

$$\beta D = - \frac{2 \sinh k \sin k}{\sinh k \cos k + \sin k \cosh k} ,$$

$$\beta D = - \frac{2}{\operatorname{ctg} k + \operatorname{ctgh} k} . \tag{13}$$

In the special case $D = 0$, we have

$$\operatorname{ctg} k + \operatorname{ctgh} k \to \infty ,$$

which implies that $k = \pi$ (or $k = n\pi$).

In these conditions ($D = 0$, $k = \pi$), the focal distance and the position of the image focal point are given by:

$$\left.\begin{aligned} &\frac{f}{L} = \frac{1}{k \sinh k} = \frac{1}{\pi \sinh \pi} , \qquad \frac{\overline{SF}}{L} = \frac{\cosh k}{k \sinh k} = - \frac{1}{\pi \operatorname{tgh} \pi} \simeq - \frac{1}{\pi} , \\ &\frac{f}{L} \simeq 3 \cdot 10^{-2} , \qquad \frac{\overline{SF}}{L} \simeq - 0.32 . \end{aligned}\right\} \tag{14}$$

Problem 3. The excitation parameter $k = \beta L$ is a function of the field gradient K_0

$$k^2 = \beta^2 L^2 = K_0 L^2 \sqrt{\frac{e}{2 m_0 \varphi_0^*}} . \tag{15}$$

In the example considered, we have $L = 2 \cdot 10^{-2}$ m, $\varphi_0^* \simeq 6 \cdot 10^6$ V and, for electrons, $e/m_0 \simeq 1.8 \cdot 10^{11}$ Coul/kg.

From (15) we have

$$K_0 = \pi^2 \Big/ \left(L^2 \sqrt{\frac{e}{2 m_0 \varphi_0^*}} \right) \simeq 200 \text{ Tesla/m} = 2 \cdot 10^4 \text{ G/cm} .$$

We know that (with $\mu_0 = 4\pi \, 10^{-7}$ in MKSA unit system)

$$K_0 = \frac{2 \mu_0 (nI)}{a^2} ,$$

nI being the total intensity circulating in the coil around each pole piece. We have

$$(nI) = \frac{a^2 K_0}{2\mu_0} \simeq 320 \text{ Ampere-turns} .$$

Problem 4. We consider protons entering the doublet studied above, equivalent to a round lens, in which $K_0 = 200$ Tesla/m, and $k = \pi$.

We have

$$\left(\frac{e}{2m_0 \varphi_0^*}\right)_{\text{el}} = \left(\frac{e}{2M_0 \varphi_0^*}\right)_{\text{prot}}, \quad \text{or} \quad \frac{\varphi_{0\,\text{el}}^*}{\varphi_{0\,\text{prot}}^*} = \frac{M_0}{m_0},$$

$$(\varphi_0^*)_{\text{prot}} \simeq \frac{1}{1\,840} (\varphi_0^*)_{\text{el}}, \quad (\varphi_0^*)_{\text{H}^+} \simeq 3 \cdot 10^3 \text{ V} \simeq (\varphi_0)_{\text{H}^+} .$$

It is thus possible to simulate relativistic electrons by utilizing slow protons, for the study of the properties of strongly excited lenses, specially designed for very high tension electron microscopes.

Problem 5. For $X_0' = Y_0' = 0$, we have

$$X_S' = X_0 T_{21_X},$$

$$Y_S' = Y_0 T_{21_Y}.$$

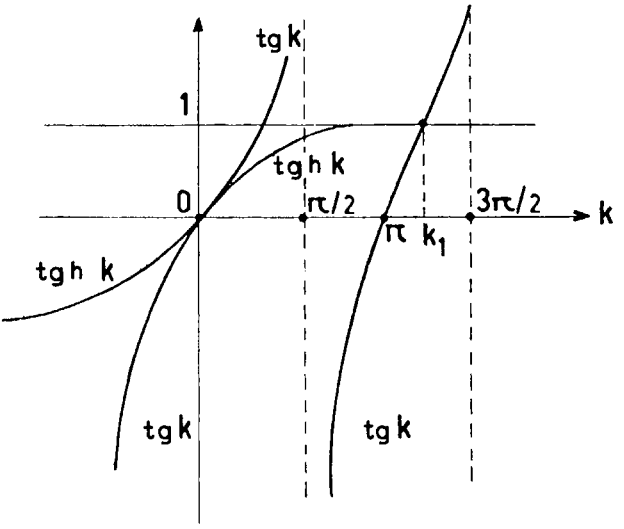


Fig. 10.

The emergent beam will be parallel to the axis, if: $T_{21_X} = T_{21_Y} = 0$, or: $\sinh k \cos k - \cosh k \sin k = 0$,

$$\operatorname{tgh} k = \operatorname{tg} k . \tag{16}$$

A value of k, solution of (16) may be obtained easily (Fig. 10)

$$k_1 \simeq \frac{5\pi}{4} .$$

The Figure 11 shows two trajectories in XOz and YOz planes respectively.

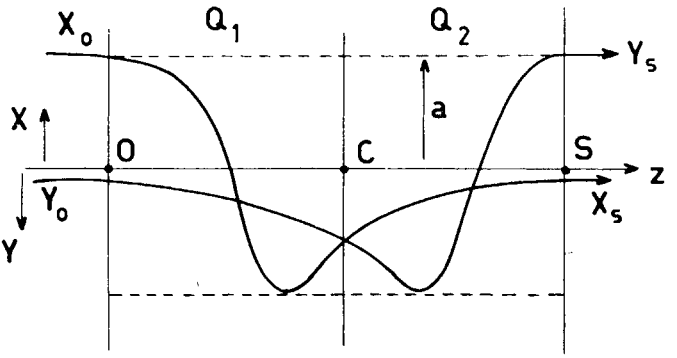


Fig. 11.

They are symmetrical about the centre C of the lens, and we have:

$$X_0 = Y_S \quad \text{and} \quad Y_0 = X_S .$$

In the first lens, $|X_{\max}| = X_0 = a$, and in the second lens $Y_S = |Y_{\max}| = a$.
We have:

$$X_S = T_{11_X} \cdot X_0 = T_{11_X} a = - Y_{0\max} .$$

Knowing that $k \simeq 5\pi/4$, we find:

$$\left(- \frac{\sqrt{2}}{2} \cosh\left(\frac{5\pi}{4}\right) + \frac{\sqrt{2}}{2} \sinh \frac{5\pi}{4} \right) a = - Y_{0\max} ,$$

or

$$Y_{0\max} = a \frac{\sqrt{2}}{2} \left(\cosh \frac{5\pi}{4} - \sinh \frac{5\pi}{4} \right) = a \frac{\sqrt{2}}{2} \exp\left[- \frac{5\pi}{4} \right] \simeq$$

$$\simeq 1.4 \cdot 10^{-2} a = 2.8 \cdot 10^{-2} \text{ mm} ,$$

$$X_{0\max} = 2 \text{ mm} .$$

This system may be used to rotate of $\pi/2$ around the axis a ribbon beam of width X_0 and thickness Y_0, such as $Y_0/X_0 \leqslant 1.4 \cdot 10^{-2}$

Problem 6. We have now

$$\frac{\overline{SF}_X}{L} = -\frac{T_{11_X}}{T_{21_X}}, \qquad \frac{\overline{SF}_Y}{L} = -\frac{T_{11_Y}}{T_{21_Y}}.$$

The condition $\overline{SF}_X = 0$, $\overline{SF}_Y = 0$ may be written as

$$\cos k \cosh k - \sin k \sinh k - \beta D \sin k \cosh k = 0, \tag{17}$$

$$\cos k \cosh k + \sin k \sinh k + \beta D \cos k \sinh k = 0. \tag{18}$$

From (17) we obtain:

$$\beta D = k\lambda = \frac{\cos k \cosh k - \sin k \sinh k}{\sin k \cosh k} = \operatorname{ctg} k - \operatorname{tgh} k. \tag{19}$$

Using (19) and (18) the relation giving k may be obtained:

$$\cos k \sin k + \cosh k \sinh k = 0,$$

or

$$\sin 2k = -\sinh 2k.$$

The problem *has no solution.*

Obtaining coincident foci at S (or beyond S) is only possible by making $k_1 \neq k_2$, but the doublet is then an astigmatic system (magnifications are different in XOz and YOz planes).

4. Prisms.

4·1. Problems.

Problem 1. In a homogeneous field prism, with $\Phi = 90°$, and $r_0 = 100$ mm, we inject electrons accelerated under a voltage φ_0. The « object » P, situated at $p = r_0/2$, has a circular cross-section of diameter $a_0 = 0.2$ mm; at this point, the divergence of the beam is given by $x_0' = 5 \cdot 10^{-2}$ rad.

a) Give the position q of the image Q of P, the magnification M, and the width (in horizontal plane) of the image, taking into account the 2nd-order aperture aberration.

b) What is the width of the image in the vertical plane?

c) A slit of width $s = 0.2$ mm is placed at Q. Calculate the resolving power R_e of the prism, used as energy analyser.

d) Give the position of the « achromatic » virtual focus.

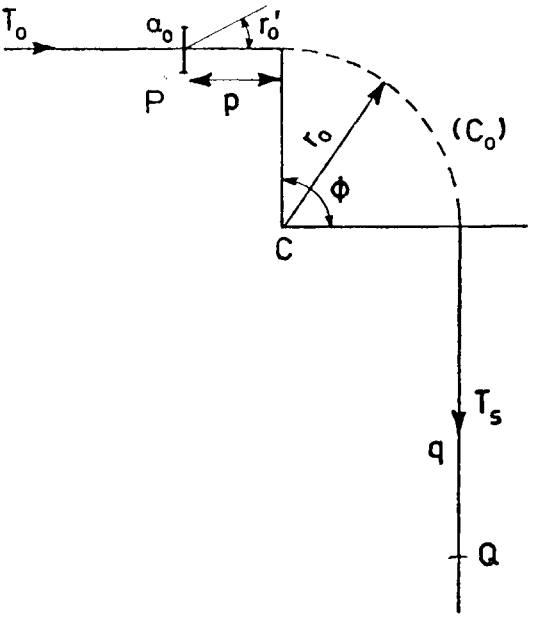


Fig. 12.

Problem 2. In order to increase the resolving power, and focus in the vertical direction, we intend to use a prism with inhomogeneous field (index $n = \frac{1}{2}$). In the case of symmetrical system ($p = q$), give the corresponding values of p and of the resolving power.

Problem 3. In the conditions given in the *Problem 1* compare the respective performances of the magnetic homogeneous field prism and the electrostatic spherical prism having the same Φ. ($\Phi = \pi/2$.)

4·2. Solutions.

Problem 1.

a) The cardinal elements in the horizontal plane H of an homogeneous field prism are given by

$$f_i = \frac{r_0}{\sin \Phi}, \qquad \overline{SF_i} = \frac{r_0}{\operatorname{tg} \Phi}.$$

The prism has no convergence in the vertical plane V. In our case ($\Phi = \pi/2$) we have (see Fig. 13)

$$f = r_0, \qquad \overline{SF_i} = 0. \tag{20}$$

The principal planes coincide at M.

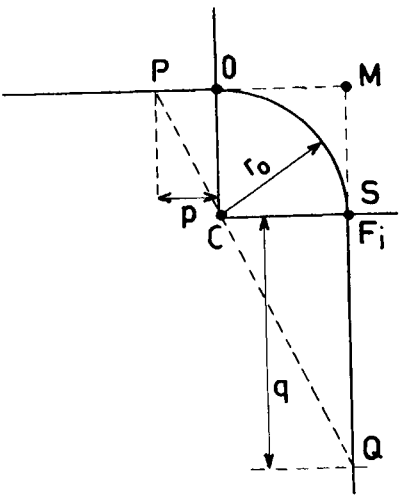


Fig. 13.

The transfer matrices between O and S are given by:

$$|H| = \begin{vmatrix} 0 & 1 \\ -1 & 0 \end{vmatrix} \quad \text{and} \quad |V| = \begin{vmatrix} 1 & \pi/2 \\ 0 & 1 \end{vmatrix}.$$

We know that P, C and Q are on a straight line, so we have:

$$pq = r_0^2, \quad \text{or} \quad q = r_0^2/p = 2r_0 \qquad (p = r_0/2).$$

These results may also be found by using the transfer matrix $|T|$ between P and a point Q located at a distance q from S:

$$|T| = \begin{vmatrix} 1 & q/r_0 \\ 0 & 1 \end{vmatrix} |H| \begin{vmatrix} 1 & p/r_0 \\ 0 & 1 \end{vmatrix},$$

$$|T| = \begin{vmatrix} -q/r_0 & 1-(pq/r_0^2) \\ -1 & -p/r_0 \end{vmatrix}. \tag{21}$$

A ray leaving $P(x_0 = 0,\ x_0' \neq 0)$ is given by

$$x_Q = \frac{x_Q}{r_0} = x_0' \left(1 - \frac{pq}{r_0^2}\right)$$

and we have: $x_Q = 0$ for all values of x_0' only if

$$pq = r_0^2 \tag{22}$$

(conjugation relation).

The magnification G may be calculated by simple geometrical considerations or by applying Newton's law:

$$G = -\frac{q}{f} = -\frac{f}{p},$$

or by using $|T|$, taking into account the fact that P and Q are conjugate points, and that (22) holds.

We have

$$x_Q = -x_0 \frac{q}{r_0} \quad \text{and} \quad G = \frac{x_Q}{x_0} = -\frac{q}{r_0}.$$

All methods give the same result:

$$G = -2.$$

The aperture aberration constant M_{13} is given by

$$M_{13} = \frac{1}{2}\left(|G| + \frac{1}{G^2}\right) = \frac{1}{2}\left(2 + \frac{1}{4}\right) = \frac{9}{8} \quad \text{and} \quad \Delta x_i = M_{13}\, x_0'^2 = \frac{9}{8} x_0'^2.$$

In unscaled co-ordinates, the total width of the image becomes:

$$(\Delta X)_Q = |G| a_0 + r_0 \Delta x_i = 2a_0 + \frac{9}{8} r_0\, x_{0\max}'^2, \qquad (\Delta X)_Q \simeq 0.68 \text{ mm}.$$

b) In the vertical plane, no focussing occurs. The true distance from P to Q is

$$L = p + r_0 \Phi + q = r_0\left(\frac{1}{2} + \frac{\pi}{2} + 2\right) \simeq 4.07 r_0 = 40.7 \text{ cm}$$

and the length h of the image line in the V plane at Q is

$$h = a_0 + 2Lx_0' \simeq 4.1 \text{ cm} .$$

c) Using the expressions for the disperion D_p and the resolving power R_p (relative to a variation $\Delta p_0/p_0$ of the momentum $p_0 = mv_0$) we find

$$D_p = (1 - G) = 3 , \qquad R_p = \frac{D_p r_0}{|G| a_0 + s + \Delta X} \simeq \frac{300}{0.88} \simeq 340 .$$

Considering now the variations in energy, related by $\Delta\varphi_0/\varphi_0$, we observe that

$$\frac{\Delta p_0}{p_0} = \frac{1}{2}\frac{\Delta\varphi_0}{\varphi_0} \quad \text{and} \quad R_e = \frac{1}{2} R_p .$$

In our system: $R_e \simeq 170$.

d) Consider particles coming from P, entering the prism at O with $x_0' = 0$, and having a momentum

$$p = p_0 + \Delta p_0 = p_0(1 + \delta) .$$

At the exit of the prism, we have (see Fig. 14):

$$x_s = \delta(1 - \cos\Phi) = \delta ,$$

$$x_s' = \delta \sin\Phi = \delta .$$

For each value of δ, the emergent particles seem to be issued from a virtual object point situated on the straight axis (T), at M.

$$\overline{SM} = -\frac{r_0 x_s}{x_s'} = -r_0 .$$

$\overline{SM}$ is independent of δ. M is a virtual chromatic image of P.

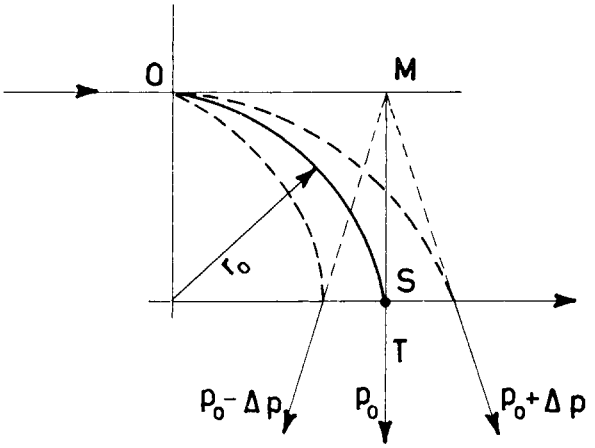


Fig. 14.

Problem 2. We have now

$$n = \frac{1}{2} \quad \text{and} \quad f = \frac{r_0\sqrt{2}}{\sin(\Phi/\sqrt{2})} = \frac{r_0\sqrt{2}}{\sin[\pi/(2\sqrt{2})]} \simeq 1.59 r_0 ,$$

$$\overline{SF_i} = g = \frac{r_0\sqrt{2}}{\operatorname{tg}(\Phi/\sqrt{2})} = \frac{r_0\sqrt{2}}{\operatorname{tg}[\pi/(2\sqrt{2})]} \simeq 0.69 r_0 .$$

In a symmetrical system ($p = q$), we have (see Fig. 15):

$$\overline{F_i Q} = -\overline{F_0 P} = f \quad \text{and} \quad \overline{SQ} = -\overline{OP} = f + g = \sqrt{2} r_0 \operatorname{ctg}\left(\frac{\pi}{4\sqrt{2}}\right),$$

or

$$p = q \simeq 2.28 r_0 .$$

The magnification is unity, in both planes H and V

$$G = G_H = G_V = -1 .$$

In the plane H, the image width is given by

$$(\Delta X)_Q = |G| a_0 + M_{13} x'^2_{0\max} = a_0 + r_0 x'^2_{0\max} .$$

We have:

$$R_p = \frac{4r_0}{a_0 + s + r_0 x'^2_{0\max}} = \frac{400}{0.65} \simeq 615 \quad \text{and} \quad R_\varphi \simeq 307 .$$

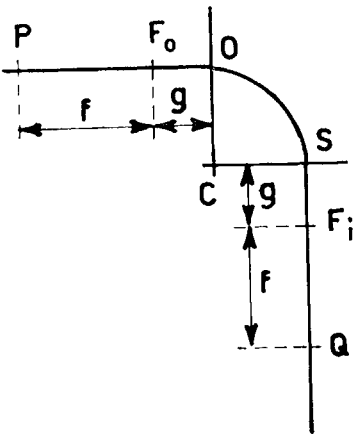


Fig. 15.

Problem 3. The cardinal elements of the spherical electrostatic prism are given by

$$f = r_0/\sin \Phi = r_0 = 100 \text{ mm} , \quad g = r_0/\operatorname{tg} \Phi = 0 .$$

In the same conditions as in *Problem 1* of Subsect. **4**·1, we have $|G| = 2$ and

$$D_e = \Delta x_i \Big/ \left(\frac{\Delta \varphi_0}{\varphi_0}\right) = (1 - G_H) = 3 .$$

For R_e, we obtain

$$R_e = \frac{r_0 D_e}{a_0 + s + M_{13} r_0 x'^2_{0\max}} \quad \text{with } M_{13} = \left(|G| + \frac{1}{G^2}\right) = 2.25 .$$

Finally:

$$R_e = \frac{300}{0.2 + 0.2 + 0.56} = \frac{300}{0.96} \sim 310 .$$

This value of R_e has to be compared to the value of R_e given in *Problem 1c)* of Subsect. **4**·2.

$$(R_e)_{\text{mag}} < (R_e)_{\text{sph. cond}} .$$

Secondary Ion Microanalysis and Energy-Selecting Electron Microscopy

R. CASTAING

Faculté des Sciences - Orsay, France

1. Dispersive microscopy using magnetic prisms.

1·1. Introduction.

The range of problems in which electron and particle optics are of use in the field of solid state physics has been extended appreciably by the introduction in the optical column of dispersive elements which allow extra information to be drawn from the images.

The purpose of this lecture is to describe the optical properties of a magnetic dispersive unit, which was first studied in our laboratory by Henry and Miss Paras ([1]), then improved and applied to energy-selecting electron microscopy ([2-4]) and to secondary ion microanalysis ([5-7]).

After a brief summary of the non-Gaussian optics of a simple magnetic prism, the properties of that dispersive unit, which uses two magnetic deflections separated by a reflection at an electrostatic mirror, will be reviewed; the special arrangements which are convenient for energy selection of electrons and for mass selection of ions will be considered separately.

The reader is expected to know the basic elements of geometrical optics, at an elementary level.

1·2. First-order focussing properties of a simple magnetic prism.

To start with, let us consider the elementary case of a simple deflection by a magnetic prism; to introduce the ideas, we will consider electrons, but the same conclusions would apply to any kind of identical charged particles.

Let us suppose that a magnet produces a uniform magnetic induction $\boldsymbol{B}$ in a region of free space limited by a dihedron whose aperture is 45° (Fig. 1); the induction vector $\boldsymbol{B}$ is perpendicular to the plane of the figure. An electron, whose initial trajectory $Z_0'Z_0$, in the field-free region outside the magnet, makes an angle of 45° with the entrance pole face, is deviated inside the gap along a circular path; if the induction has the correct value, the total deflection is 90° and the electron leaves the magnetic gap in the direction $Z_1'Z_1$, along the normal to the exit pole face.

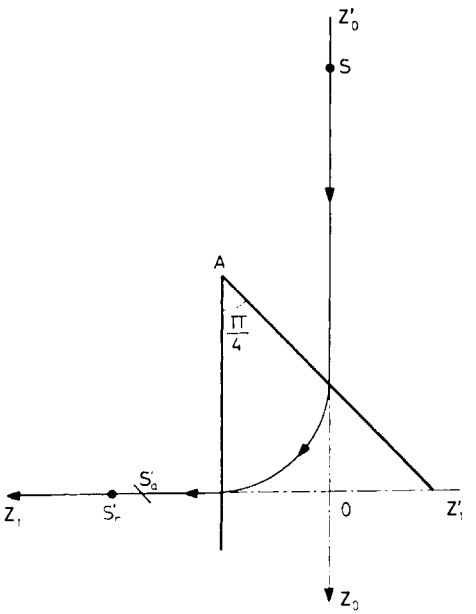


Fig. 1. – Deflection by a magnetic prism. General case. (Courtesy of *Zeitschrift für angewandte Physik*.)

If we consider now a narrow pencil of monoenergetic electrons, starting from a point source S located on the axis $Z_0'Z_0$, it is easy to show that the electron trajectories located in the plane of the figure (radial plane) will converge after the deflection by the magnetic prism onto a « radial image » S_r', whereas the trajectories lying in the plane normal to that of the figure and passing through $Z_0'Z_0$ (axial plane) will converge (because of the focussing effect of the fringing fields of the magnet) onto an « axial image » S_a'. As a result, the narrow beam coming from the point source S is transformed

through the magnetic deflection into an astigmatic beam passing through two focal lines: a radial focal line normal to the radial plane at S_r' and an axial focal line, lying in the radial plane and passing through S_a'.

The separation between the focal lines depends on the location of the point source S; in the case that we have considered, it reduces to zero for two different positions of the source S; in other words, the magnetic prism exhibits two stigmatic object points S_1 and S_2 (Fig. 2). A narrow electron

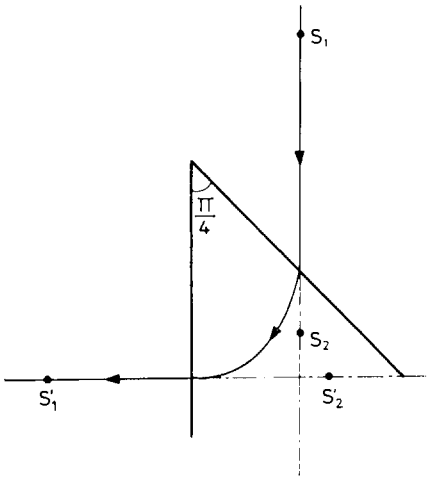


Fig. 2. – Stigmatic points. (Courtesy of *Zeitschrift für angewandte Physik*.)

beam produced at S_1 or S_2 and whose axis is $Z_0'Z_0$ will converge after going through the magnetic prism onto stigmatic image points S_1' and S_2' (Fig. 3). (S_1, S_1') is a pair of real conjugated stigmatic points, (S_2, S_2') is a pair of virtual conjugated stigmatic points.

Suppose now that $Z_0'Z_0$ is the optical axis of an electron microscope and that the image-carrying beam transmitted through the objective lens is deflected at 90° by the magnetic prism we have just considered (Fig. 4); we arrange things in such a way that the exit cross-over of the objective lens (from which the electron trajectories are originating) is located at S_1; furthermore we adjust the excitation of the objective lens in such a way that the image which is carried by the beam would be formed, if the prism was absent, at the level of S_2. If the energy of the electrons has the correct value for ensuring a 90° deflection of the mean trajectory, the emerging beam will converge onto an « image cross-over » located at S_1' and it will correspond

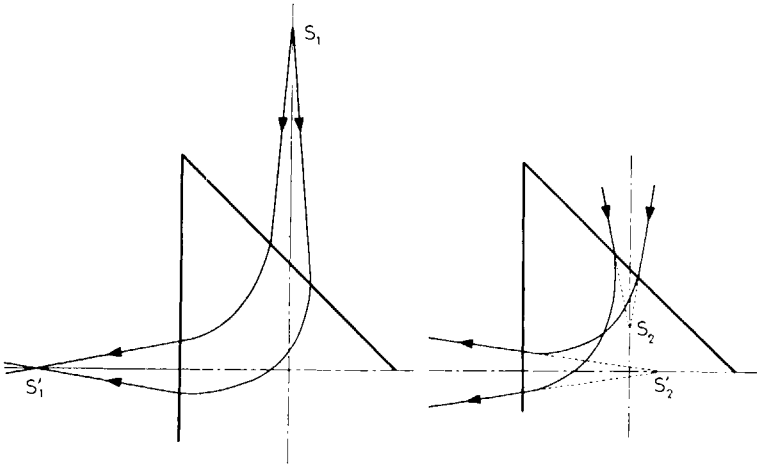


Fig. 3. – The two pairs of conjugate stigmatic points: left, real stigmatic points; right, virtual stigmatic points. (Courtesy of *Zeitschrift für angewandte Physik*.)

to a stigmatic image located at the level of S_2'. On the other hand, if the incident beam contains electrons of various energies (because of various velocity losses inside the sample, for example), the electrons of a given energy

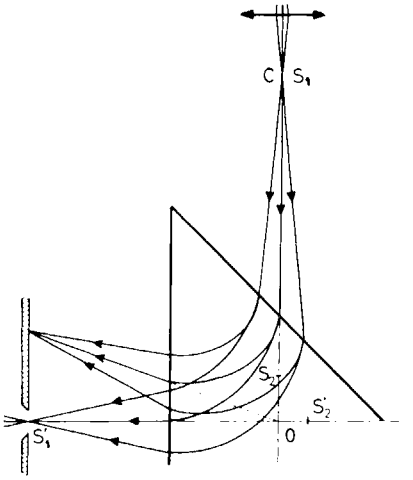


Fig. 4. – Energy filtering by simple deflection. (Courtesy of *Zeitschrift für angewandte Physik*.)

only will converge at S_1'; thus it will be possible, by means of a slit located at S_1', to isolate (by a suitable adjustment of the magnetic induction) those of the electrons which have suffered, when going through the specimen, a given amount of energy loss (Fig. 4). By means of a subsequent set of projection lenses it will be possible to project onto the fluorescent screen the plane of the image S_2'; as a result, we will observe on the screen the image of the specimen which is formed by those of the electrons which have been decelerated when going through the sample, by a given amount. Such is the principle of the magnetic filtering of velocities in electron microscopy; a similar arrangement was used by Slodzian for the purpose of mass-filtering in the original secondary ion microanalyzer ([5,7]).

Nevertheless, there is a serious flaw in such a simple filtering device: the selection slit located at S_1' cannot be infinitely narrow (which would not be convenient anyway for obtaining images bright enough!), and the admitted energy band has a width δE. Now, it is easy to see that a slight modification of the electron velocity results in a rotation of the mean emergent trajectory, in the radial plane, around the point O where the axes S_1S_2 and $S_2'S_1'$ intersect. As a result, the image which forms at the neighbourhood of S_2' is blurred by strong chromatic effects: each image point is spread over a linear energy spectrum in the radial plane if the admitted energy bandwidth is not zero. Under such conditions, it is better to adjust the excitation of the objective lens in such a way that the exit image (after the magnetic deflection) forms at the level of O, so that the chromatic aberration is reduced to the second order (it is clear that if the exit image forms around O, a slight variation of the energy of the electrons will result in a rotation of the emerging trajectories around the image points themselves, so that the image is achromatic to the first order); but in that case the image will be strongly astigmatic because O is not a stigmatic image point. In this simple arrangement, the image cannot be stigmatic and achromatic at the same time, so that the dispersive unit must be equipped with an auxiliary stigmator at the level of the exit cross-over S_1'. This was done in the first model of the secondary ion microanalyzer ([8]).

Another drawback arises from the fact that the optic axis is bent through 90° by the magnetic deflection, which is inconvenient in any application to electron microscopy; we were thus led to modifying the dispersive system to overcome those difficulties. The arrangement that Henry applied to energy filtering in electron microscopy uses a double magnetic deflection in conjunction with a reflection at an electrostatic mirror; this proved to be a convenient arrangement and a similar system was used for ion emission microscopy.

1·3. The dispersive system of the energy-selecting electron microscope.

In the energy-selecting electron microscope developed by Henry, the electron beam which diverges from the exit cross-over of the objective lens (more precisely, from the exit cross-over of a first intermediate lens) enters the first half of a triangular magnetic prism where the induction is uniform. After the first 90° deflection (the induction is adjusted to the right value) the beam leaves normally to the vertical face of the prism (Fig. 5). A reflection at an electrostatic mirror brings the beam back into the prism where a second deflection, symmetric to the first one, brings its axis back into the line of the original beam.

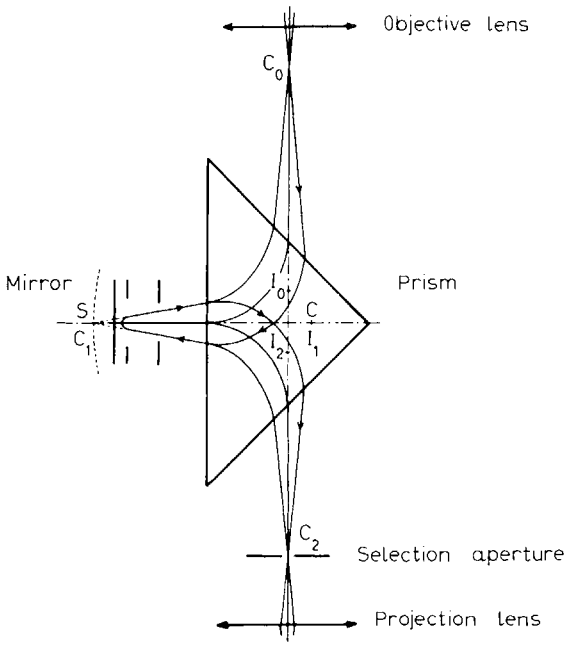


Fig. 5. – Dispersive unit of the energy-selecting microscope. (Courtesy of *Optique des Rayons X et Microanalyse.*)

1·3.1. *First-order stigmatic conditions.* – By adjusting the excitation of the previous lenses, it is easy to arrange things in such a way that the cross-over C_0 and the initial image of the sample I_0 form at the levels of the stigmatic object points (the real one and the virtual one, respectively) of the first magnetic prism. After its first deflection by the magnetic prism, the beam con-

verges towards a cross-over C_1 (we suppose for the moment that the electrons are monoenergetic), appearing to come from a stigmatic image located at the level of I_1. Then the beam is reflected by the field of an electrostatic mirror.

The mirror comprises, starting from the left (Fig. 5) a reflecting electrode whose negative potential is higher than that of the electron source, a Wehnelt electrode and an anode at ground potential. The dotted line (Fig. 5) represents the image (produced by the field of the mirror) of the « zero » equipotential surface whose potential is exactly that of the electron source. That image surface is the « reflecting surface » of the spherical mirror; electrons emitted with zero velocity from the zero equipotential surface (not represented) would converge at a cross-over which plays the role of the centre of the spherical mirror; note that in the arrangement of Fig. 5 the spherical mirror is a concave one. By positioning and exciting the mirror appropriately, its apex S is located at C_1 and its centre C at I_1. As a consequence, the mirror transforms the cross-over C_1 into itself, and the image I_1 into an inverted image located at the same point. It is clear that the second magnetic deflection will result in an exit beam corresponding to a stigmatic image located at I_2 and converging onto C_2. The exit image I_2 and the exit cross-over C_2 are symmetrical with the entrance image I_0 and the entrance cross-over C_0 about to the axis C_1I_1.

1˙3.2 *Dispersive properties.* – It may easily be shown that a slight change in the energy of the electrons results in a rotation of the mean exit trajectory around I_2; as a consequence, the image which forms at I_2 is stigmatic and achromatic to the first order.

It is worth-while to note at this point that the whole dispersive device is found to be stigmatic for all points along the axis C_0I_0 if C_1 and I_1 (the stigmatic image points of the first magnetic deflection) coincide with the stigmatic points of the mirror; as a consequence, it is not essential to locate the entrance cross-over at the stigmatic point C_0 of the first magnetic deflection; on the other hand, the entrance image has to be located at I_0 to obtain achromatism; otherwise (if the electrons were strictly monoenergetic) it could be located anywhere without causing astigmatism.

If the electrons have suffered various energy losses when going through the object, an energy loss spectrum is formed at the level of the exit cross-over C_2, whereas all the images produced by electrons which have lost various amounts of energy are superimposed (to the first order) at the level of I_2. Hence, the subsequent handling of the image carrying beam may proceed in two different ways:

a) we may adjust the subsequent projection lenses so as to project the plane of I_2 onto a selection aperture and the plane of C_2 onto the final screen; we will observe on the screen the energy loss spectrum of that part of the object which is isolated by the selection aperture: the instrument behaves as a velocity analyzer;

b) alternatively, we may isolate a narrow portion of the energy loss spectrum by means of a slit located at the level of C_2 and adjust the subsequent projection lenses so as to project the plane of I_2 onto the final screen; we will then observe on the screen the image of the object formed by those electrons which have suffered a definite energy loss ΔE (obviously, ΔE can be made equal to zero for observing the « no loss image »): the instrument behaves as an energy-selecting microscope.

It is easy to see that the resolving power of the microscope is not damaged by the introduction of the dispersive device. In fact both magnetic deflections and the electrostatic reflection introduce second-order aberrations in the image, but the effect of these aberrations is negligible if the initial magnification, at the level of I_0, is large enough. Typically, the magnification is equal to 200 at the level of I_0, so that the aberrations of the dispersive device are diminished by the same factor when reduced to the object scale; furthermore the aberrations are small because of the extremely small divergence (about 10^{-5} rad) of the beams which form each image point at the level of I_0. In practice the resolving power of the microscope is improved by the dispersive device when the elastic (no loss) image is observed, because effect of chromatic aberration in the objective lens is eliminated. The quality of the inelastic image is frequently less good (but still very good); this is probably due to the fact that the inelastically scattered electrons fill the objective aperture uniformly, leading eventually to edge effects, whereas the « elastic » ones are passing mainly through the central part of the objective aperture.

Let us consider now the quality of the energy filtering obtained when the instrument operates as an energy-selecting microscope. This quality depends on the energy bandwidth that we select by the slit located at C_2. Now, this bandwidth cannot be made equal to zero, even by making the slit infinitely narrow because of the second-order aberrations of the dispersive device at the level of C_2; those aberrations may be fairly large if a wide beam is made to converge at C_2; other things being equal, the divergence of this beam is proportional to the diameter of the imaged part of the sample.

To obtain an idea of the performance we consider a typical case where the radius of curvature of the electron trajectories inside the magnet is 4 cm

for a 100 kV accelerating voltage, and where the magnification at the level of I_0 is about 200. In such conditions, the dispersion at the level of C_2 is about one micron per electron-volt. Let us suppose that the electrons emitted by the gun are perfectly monoenergetic and that the slit located at C_2 is infinitely narrow; in such conditions the energy bandwidth that we select for the image is a little better than 1 eV when the diameter of the imaged area is equal to one micron; in principle, the bandwidth would reduce to 0.01 eV for an imaged area whose diameter would be equal to 0.1 μm. It should be emphasized at this point that what we speak about here is the total energy bandwidth which is admitted over the whole of the image. If the illuminating electrons are perfectly monoenergetic and if the selecting slit at C_2 is infinitely narrow, each point of the image receives electrons whose energy is perfectly defined. However the value of the energy which is being selected is not the same at different points of the image as a result of the second-order chromatic aberrations of the dispersive device at the level of C_2.

The situation is depicted in Fig. 6. The overall deflection of the electron trajectories is a little larger (for a given energy) when the trajectory is inclined in the radial plane with respect to the optical axis. As a consequence, the electrons which form the centre of the image have a lower energy than those which form both sides; the selected energy loss varies, on a diameter perpendicular to the induction vector $\boldsymbol{B}$, as the square of the distance between the image point and the optical axis. If we observe a large region of the object, say 5 μm, and if the value of the induction is such that the centre of the image is formed by electrons whose energy loss is 25 eV, the two sides of the image are formed by elastically scattered electrons. The image field is crossed by isoenergetic lines; in other words, an energy spectrum is superimposed onto the image; an example will be given in a next lecture ([4]). The spacing between the isoenergetic lines becomes smaller the further away the imaged region is from the optical axis (this results from the second-order chromatic aberration); for obtaining a good homogeneity of the energies across an extended region, it is preferable to align the instrument perfectly so as to bring the optical centre in the middle of the final screen; on the other hand, if we are interested in looking at a wide part of the energy spectrum on a small region of the object, we will off-centre the instrument so as to push the optical centre off the screen ([4]).

In the present experimental conditions, the electron gun (hot tungsten wire) emits electrons whose energy spectrum spreads over 0.5 eV, so that the energy bandwidth is at least 0.5 eV at each image point and 1.5 eV over the whole image of a region of the object whose diameter is 1 μm (that corresponds

to an error of ± 0.75 eV for the value of the energy loss, for a 100 kV accelerating voltage). A better accuracy could be obtained by reducing the

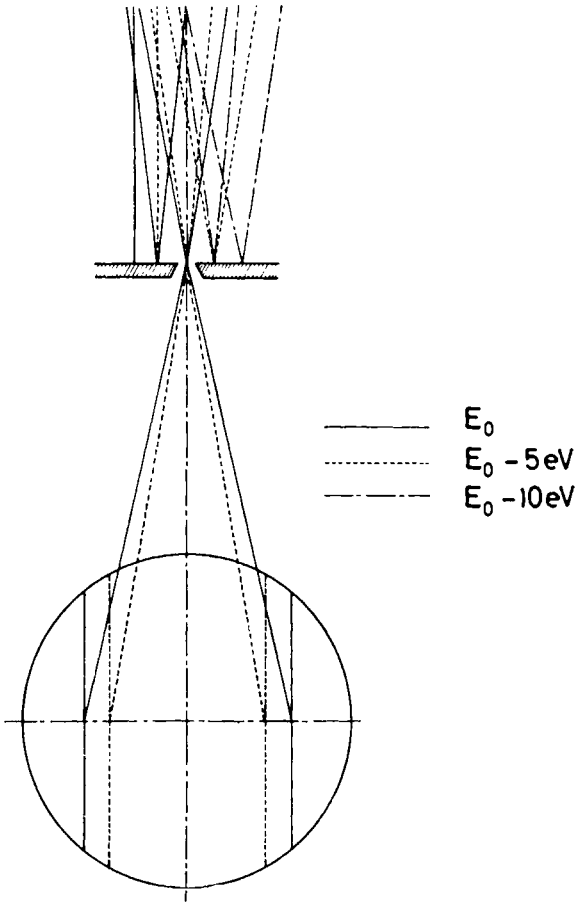


Fig. 6. – Energy loss selected by the slit at various points of the image. E_0: initial energy of the electrons. The electron trajectories are drawn in the radial plane and the induction vector in the prism points towards the observer (contrary to Fig. 5). The straight lines in the image field represent isoenergetic lines (in fact those lines are curved). (Courtesy of *Zeitschrift für angewandte Physik*.)

diameter of the imaged region, filtering the velocities of the electrons emitted by the gun and compensating the loss of brightness by an image intensifying device. Now, an energy resolution of 1 eV is good enough for eliminating inelastic electrons, as will be seen in the next lecture ([4]), because the various

inelastic processes (apart from the quasi-elastic scattering by phonons) lead to energy losses larger than 1 eV. In such conditions, the exposure time which is required for registering elastic images (or images produced by a strong plasmon loss) is about the same as in a conventional electron microscope.

1·4. The dispersive system of the secondary ion microanalyzer.

A similar arrangement may be used for filtering masses in ion emission microanalysis. The principle of that technique will be outlined in a later lecture ([7]). The polished surface of a massive sample is bombarded by a primary beam of ions; the secondary ions produced by the bombardment are focussed by an emission lens to form an image of the target surface. This total image is formed by all the secondary ions; it consists of the superimposition of various « characteristic images », produced by the various types of secondary ions; isolating one of the characteristic images by means of the dispersive device makes it possible to obtain a map of the distribution of the corresponding element (or isotope) across the bombarded area.

It would be possible to use the same optical arrangement that we used for energy selection of electrons (apart from the fact that the value of the induction is much higher in the case of ions) but some trouble would arise, in the case of heavy ions, from the fact that the magnetic deflection provides momentum filtering instead of true mass filtering. The ions are emitted from the object with initial energies which are not negligible (from zero to several tenths of electron-volts), so that the same circular path may be followed inside the prism by an ion of mass M emitted with zero velocity and a ion of mass $M-1$ emitted with an appreciable initial energy. This difficulty cannot be overcome by using a small aperture at the image focal plane of the emission lens, for such an aperture eliminates from the beam the ions which have been emitted with a large *tangential* velocity only, so that fast secondary ions emitted normally to the object surface pass through the aperture. As a consequence, superimposition of neighbouring masses would occur, especially in the case of heavy elements. The problem may be overcome by modifying slightly the optical arrangement of the dispersive device so as to eliminate from the exit beam the ions which have been emitted with too high an initial energy.

For that purpose, the position and the excitation of the electrostatic mirror are modified in such a way (Fig. 7) that its apex S is now located at I_1 and its centre C at C_1. Note that in this case the spherical mirror is a convex one

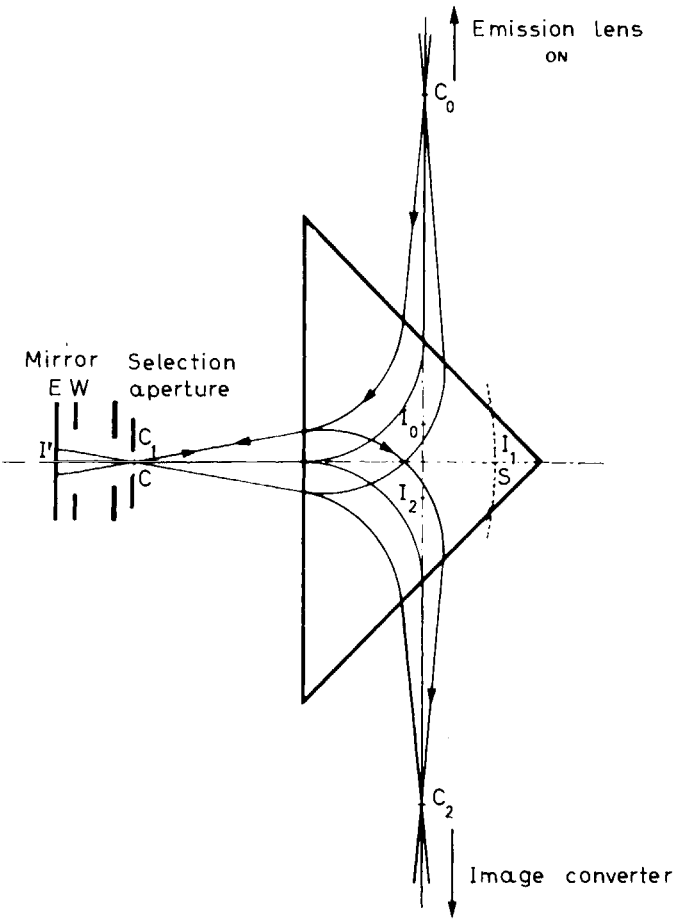


Fig. 7. – Dispersive unit of the secondary ion microanalyzer. (Courtesy of *Optique des Rayons X et Microanalyse.*)

(with a virtual reflecting surface at S). Furthermore, the potential of the reflecting electrode E is adjusted just above (one or two volts for example) that of the analysed sample. For secondary ions emitted with a very low initial velocity, the situation is practically the same as that we described for the electrons. The entrance image I_0 is transformed into a stigmatic exit image I_2 (which is now symmetric of I_0 with respect to the axis parallel to the induction vector and passing at the intersect of C_0C_2 and C_1I_1). Momentum filtering is provided by an aperture located at C_1.

But the situation is quite different for ions which have left the sample with a high initial velocity: they touch the reflecting electrode E and they are neutralized and eliminated from the exit beam; in other words, the mirror

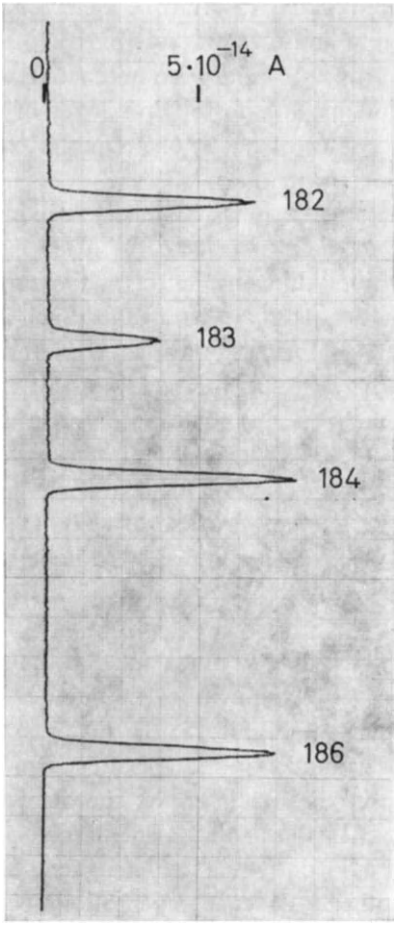


Fig. 8. – Tungsten spectrum registered in the secondary ion microanalyzer (Rouberol).

provides a low-pass energy filtering which combines with the momentum filtering of the magnet for ensuring practically that no superimposition of neighbouring masses will occur. This is illustrated by the tungsten spectrum of Fig. 8, which was obtained in the secondary ion microanalyzer by Rouberol; the neighbouring masses are perfectly isolated from one another, even for

heavy elements. The arrangement was found to operate quite satisfactorily allowing energy selection with an accuracy better than one electron-volt for secondary ions whose energy is 4 keV. Such an energy selection could be obtained by means of a spherical condenser, but the same resolution would require a very large radius of curvature (one meter or more); that is due to the fact that the emitting source (the cross-over C_0) is much larger (some tenths of a millimeter) in the secondary ion microanalyzer than in the energy-selecting electron microscope, where the cross-over of the first intermediate lens has a Gaussian diameter of about one micron, so that the selecting slit is necessarily much wider.

It should be noted that the ions are reflected in the immediate vicinity of the reflecting electrode E (at a distance of a few microns). It might be expected in such conditions that any defect at the surface of that polished electrode would damage the quality of the image. Such is not the case because an intermediate image I' is formed at the level of the electrode E, so that any local defect (a scratch of a dust for example) will affect a very small part of the image only; the defect will in fact appear superimposed on the filtered image. The situation would be quite different in the arrangement of the energy selecting microscope, if the reflection occurred in the immediate neighbourhood of the reflecting electrode; any defect would disturb the trajectories at the level of an intermediate cross-over and injure the quality of the whole image. This is one of the reasons why we used this modified arrangement for the case of ions.

To conclude, it is worth-while to point out that the necessity of focussing the image at the exact level of the achromatic point I_2 is much less stringent here than in the arrangement of Fig. 5; the first reason is that the energy selection provided by the mirror is much better than that which could be obtained from a selection slit; furthermore, the dispersion is much smaller in this arrangement than in the arrangement of Fig. 5, because of the fact that the dispersions provided by the two magnetic deflections are nearly compensating one another. Exact compensation could be obtained by locating the apex S of the mirror at the intersection of axes C_0C_2 and C_1I_1: the device as a whole would then be achromatic (the mass selection would occur anyway at the level of C_1); the exit image would exhibit a small amount of astigmatism which could be corrected easily by an auxiliary stigmator.

Examples of application of those dispersive units will be given in the following lectures ([4-7]).

REFERENCES (Section 1)

1) N. PARAS: Diplôme d'Etudes Supéricures, Paris (1961).
2) R. CASTAING and L. HENRY: *Compt. Rend. Acad. Sci. Paris*, **255**, 76 (1962).
3) L. HENRY: *Thesis*, University of Paris (1964); *Bull. Soc. Franç. Min. Crist.*, **88**, 172 (1964).
4) R. CASTAING: Section **2**.
5) G. SLODZIAN: *Thesis*, University of Paris (1963); *Ann. de Phys.*, **9**, 591 (1964).
6) R. CASTAING: *Optique des Rayons X et Microanalyse*, in *Proc. of the 4th Int. Conf. on X-Ray Optics and Microanalysis, Orsay 1965* (Hermann, Paris, 1965), p. 48.
7) R. CASTAING: Section **3**.
8) R. CASTAING and G. SLODZIAN: *Journ. de Microscopie*, **1**, 395 (1962).

2. Some applications of the magnetic filtering of energies in electron microscopy.

2·1. Introduction.

The important role of inelastic scattering in the formation of image contrast was emphasized many years ago by electron microscopists; as early as 1949, Boersch ([1,2]) succeeded in eliminating the inelastic background from electron micrographs and diffraction patterns by the use of a high-pass filter lens or a retarding grid. A more elaborate device was developed later on for the same purpose by Beaufils ([3]) in our laboratory at Toulouse; by means of a retarding grid, it was possible to isolate, for producing the image or the diffraction pattern, those electrons which had been elastically or quasi-elastically scattered by the sample. However it was clear that such high-pass filtering was not adequate for studying the inelastic scattering itself, and efforts were made in various laboratories to develop a band-pass filter which would make it possible to select in the beam, for producing the image, those electrons which had suffered a definite amount of energy loss while travelling through the sample.

The experimental technique considered here makes use of the dispersive unit, comprising two deflections at 90° and a reflection on an electrostatic mirror, whose optical properties are described elsewhere ([4]); it was investigated

initially in our laboratory by Hennequin ([5]) and Miss Paras ([6]) then developed and applied to some problems in solid state physics by Henry ([7,8]), El Hili ([9,10]) and Henoc ([11]); it is much simpler and leads to a better resolving power than the technique (based on the dispersive properties of a Möllenstedt lens) developed independently by Watanabe and Uyeda ([12]).

After a brief survey of the various scattering processes which occur when the electron beam goes through a solid sample, the main course of the lecture deals with the experimental study of the coherency of those various interaction processes which was made by Henry, El Hili and Henoc; special attention is given to the partial coherency of the electron-phonon interaction whose quantitative interpretation was given by Natta ([13]).

The reader is expected to understand the basic principles of the dynamical theory of electron diffraction. (See Howie's lectures.)

2·2. The various scattering processes that an electron may undergo in a solid sample.

Those scattering processes may be separated into three groups according to the magnitude of the energy change that is suffered by the electron as a result of the process.

i) In the first group we find the true elastic processes (no energy loss), leading to Bragg scattering in crystalline samples, to diffuse elastic scattering in amorphous samples; some amount of elastic diffuse scattering around the Bragg spots may arise in crystalline samples from the presence of imperfections such as dislocations, stacking faults or point defects. If the observed part of the sample is devoid of visible defects such as dislocations or stacking faults, point defects such as vacancies, interstitials or impurity atoms are the only source of elastic diffuse scattering around the Bragg spots.

ii) Quasi-elastic scattering arises from electron-phonon interaction; the scattering angle may be very large, but the energy change is very low: typically a few hundredths of an electron-volt. At very low temperatures (near the absolute zero), the only process is the excitation of phonons, leading to an energy loss; at higher temperatures, thermal scattering arises from the excitation or quenching of phonons and results in negative or positive energy changes. This part of the scattered beam cannot be isolated from the elastic part in our instrument by the filtering device, whose energy resolution is of the order of one electron-volt.

iii) Inelastic scattering arises mainly from:

a) Excitation of plasma oscillations, leading to more or less discrete energy losses whose value is typically 10 eV.

b) Excitation of single electrons, leading to energy losses whose spectrum extends from a few eV (outer electrons or free electrons in metals) to very large values (core electrons).

c) Bremsstrahlung leading to the continuous X-ray spectrum.

The use of an energy selecting microscope whose energy resolution is 1 eV makes it possible to isolate the inelastic scattering because of the happy circumstance that the cross-section for a true inelastic process leading to an energy loss in the range $(0 \div 1)$ eV is extremely small, as we shall see below; so that the « zero loss image » can be considered as arising from the true elastic scattering and the quasi-elastic phonon scattering only. The various inelastic processes can be separated from one another through the choice of the energy loss; furthermore, the quasi-elastic scattering can be isolated from the elastic one, in the case of crystalline samples, by selecting the electrons which have been scattered out of the Bragg spots and choosing parts of the sample which are devoid of visible defects; furthermore, the quasi-elastic thermal scattering can be disentangled from the residual elastic scattering arising from point defects through its temperature dependence.

2·3. Energy selection and « colour » electron microscopy.

We have seen in Section **1** that the energy selecting microscope makes it possible, either to project onto the final screen the energy spectrum of the electrons which have passed through a given part of the object, or alternatively to observe the image which is formed by those of the electrons which have suffered a given energy loss; the bandwidth may be better than 1 electron-volt if the diameter of the imaged area is one micron or less. Now, the main part of the inelastic scattering is due to plasmon losses whose energy spectrum is characteristic of the local composition and structure of the sample. For example, Fig. 9 shows the characteristic energy loss spectrum of aluminium; the losses are integral multiples of an elementary loss which is about 14.6 eV. It is thus possible by selecting the electrons which have suffered that energy loss which is characteristic of a given compound to form the image, to get the distribution of that compound over the imaged area. As an

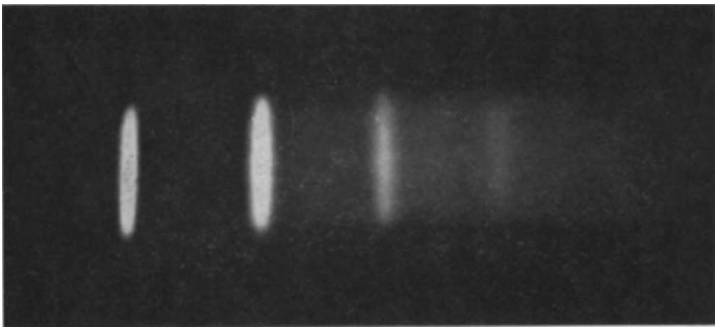


Fig. 9. – Energy loss spectrum of aluminium, as recorded on the energy-selecting microscope (Henry). (Courtesy of *Compt. Rend. Acad. Sci. Paris.*)

example, Fig. 10 shows three images obtained by El Hili ([10]) from a sample consisting of an aluminium foil which had been partially oxidized by heating in air. On the left we see the image obtained from « elastic » electrons (no loss); in the middle the image is obtained from the characteristic loss of aluminium (14.6 eV): it shows a bright part which appears in dark on the elastic image

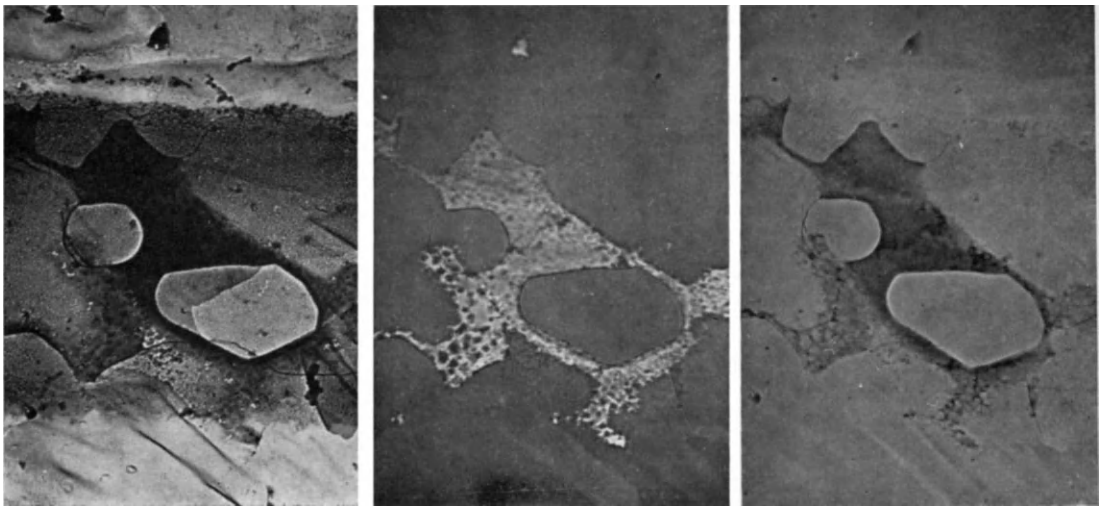


Fig. 10. – Aluminium foil, partially oxidized. Left: no loss image; centre: 14.6 eV loss image; right: 21.4 eV loss image (El Hili). (Courtesy of *Compt. Rend. Acad. Sci. Paris.*)

and on the image on the right-hand side, which is produced by electrons which have suffered a characteristic loss (22 eV) of alumina; this bright part thus corresponds to unoxidized, metallic aluminium.

Obviously, energy losses are not « characteristic » enough, nor are they known with sufficient accuracy for establishing a true method of microanalysis, but this analytical procedure is the only one which can be applied at the moment to particles whose diameter is less than 100 Å units. In fact, for the present time, such a method is more a technique of « colour » microscopy than a true microanalytical technique; nevertheless it has been applied with success by El Hili ([10]) to the identification of very small precipitates, and recently Henoc ([11]) was able to identify by this method small cavities inside a metallic sample from the change of the plasmon energy which arises from size effects in such a case.

In a bulk sample the plasmon energy is mainly governed by the electron concentration; in simple cases such as that of binary solid solutions the energy loss may be used for measuring a local concentration. Cundy, Metherell and Whelan ([14]), using a Möllenstedt filter, were thus able to demonstrate the slight change in magnesium concentration which occurs in an aluminium-magnesium alloy in the immediate vicinity of a grain boundary. Quite recently Colliex and Jouffrey ([15]) in our laboratory have used large energy losses produced by the excitation of X-ray levels for obtaining characteristic images where local brightness is controlled by the concentration of the various component elements; a similar technique, using a scanning probe, had been proposed by Hillier twenty five years ago, before the explanation by Bohm and Pines, in terms of collective excitation of plasma oscillations, of the characteristic losses first observed by Ruthemann. The technique has so far been applied only to metallurgical samples, but there is no doubt that there is a fascinating range of applications in the field of biology; for the moment, these possibilities are completely unexplored.

2·4. Experimental investigation of the coherency of the interaction of fast electrons with a solid sample.

Let us turn now to a series of experiments which were carried out in our laboratory by Henry, El Hili and Henoc in order to investigate, by means of the energy-selecting microscope, the degree of coherency of the various scattering processes which occur when a fast electron goes through a solid sample.

2˙4.1. *Fresnel fringes.* – It was generally thought ten years ago that inelastic scattering is essentially incoherent, so that any diffraction phenomenon produced by the sample itself and observable on the elastically scattered wave would be absent in the beam which has been scattered inelastically. That point of view seemed to be confirmed by the first experiments of Watanabe and Uyeda who claimed that « the inelastically scattered electrons do not

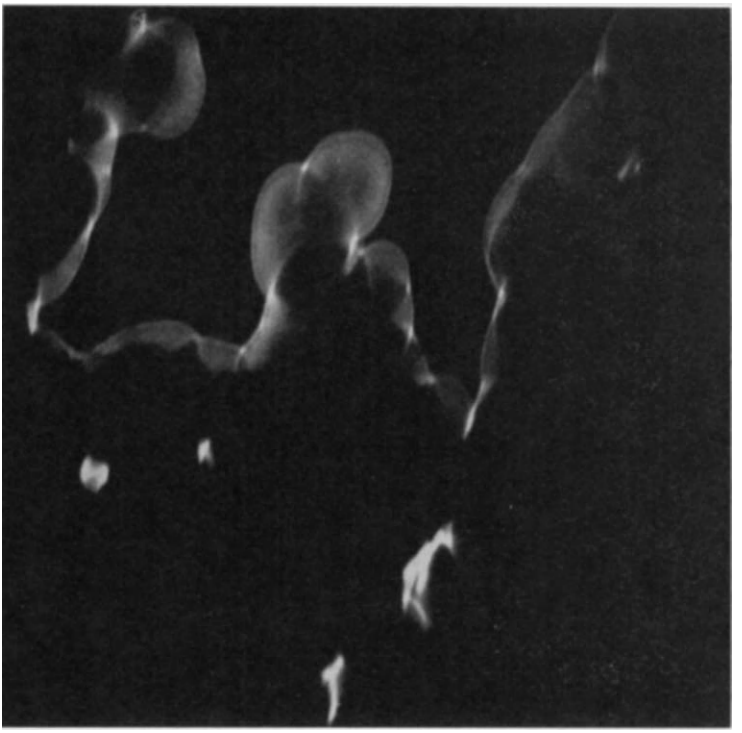


Fig. 11. – Fresnel fringe (underfocused) around carbon black particles. Inelastic image (5.9 eV loss of carbon). Magnification 150000× (Henry). (Courtesy of *Compt. Rend. Acad. Sci. Paris.*)

produce a Fresnel fringe » at the edge of the sample; as a matter of fact, Henry was able to show that the Fresnel fringe is present in the image produced by the electrons which have been scattered inelastically by plasmons; Fig. 11 shows the Fresnel fringe observed at the edge of carbon black particles; it is very faint and its contrast is quite different from what is observed on the elastic image; this is probably due to the fact that the amplitude of the scat-

tered wavelets is increasing rapidly with the distance of the scattering point (inside the sample) from the free edge, because of the increasing thickness of the specimen.

2·4.2. *Diffraction contrast.* – The energy selecting microscope was used extensively by El Hili and Henoc to study the influence of inelastic scattering on the diffraction contrast which appears on electron images of crystalline specimens.

It is well known that diffraction contrast effects such as thickness or contour fringes, dislocation and stacking fault images, are explained quite satisfactorily by the dynamical theory of electron diffraction. They are produced by the interference of the various wave fields (Bloch waves) which are excited in the crystal by the plane wave associated with the incident electron beam.

Those wave fields show phase relationships imposed by the boundary conditions at the entrance surface of the specimen. They give rise at the exit surface to a set of plane waves, travelling in vacuo, consisting of the transmitted beam and various diffracted beams; the intensity of each emerging beam depends essentially on the phase changes that the various Bloch waves undergo when passing through the sample. The phase changes themselves depend on the thickness of the specimen and on the orientation of the incident beam with respect to the crystal lattice; as a result, thickness fringes, bent contours and diffraction contrast produced by crystalline defects are observed in the image. Now, the usual dynamical theory does not take account of the inelastically scattered electrons, in spite of the fact that their contribution to the image may predominate in the case of thick specimens.

It is clear that, if the transitions that the various Bloch waves undergo through a given inelastic scattering process are independent (in other words if the final state of the crystal after the elementary interaction is not the same for the various Bloch waves), the phase coherency that the Bloch waves had with respect to one another initially is destroyed by the interaction process, so that the scattered waves are no longer able to interfere and give rise to diffraction contrast. For example, in the case of the inelastic scattering associated with plasma excitations (plasmon-electron interaction), the inelastically scattered waves will give rise to diffraction contrast on the image if, and only if, all the Bloch waves excited in the crystal by the incident beam interact with the same mode of oscillation of the plasma; in other words, all the Bloch waves must be scattered simultaneously by the excitation of the same plasmon. In such a case, the interaction is said to be « coherent » since it maintains the coherency between the various Bloch waves.

Thickness contours were observed by Watanabe and Uyeda ([12]) in an image produced by plasmon scattering, but the quality of the image was not good enough to confirm that the contrast of the thickness fringes was the same as in the elastic image; nevertheless, it could be concluded that the inelastic process involved in plasmon-electron interaction was at least partially coherent. As a tentative explanation, Heidenreich ([16]) had suggested that only a part of the interaction was coherent, namely that part which corresponds to electrons which have lost some amount of energy « outside the sample »; that amounted to saying that the Bloch waves are scattered independently inside the crystal, but that there is an appreciable probability for the occurence of an inelastic interaction before the electron penetrates the crystal (in such a case the interaction concerns the plane incident wave, before the Bloch waves are excited) or after the electron has left the crystal.

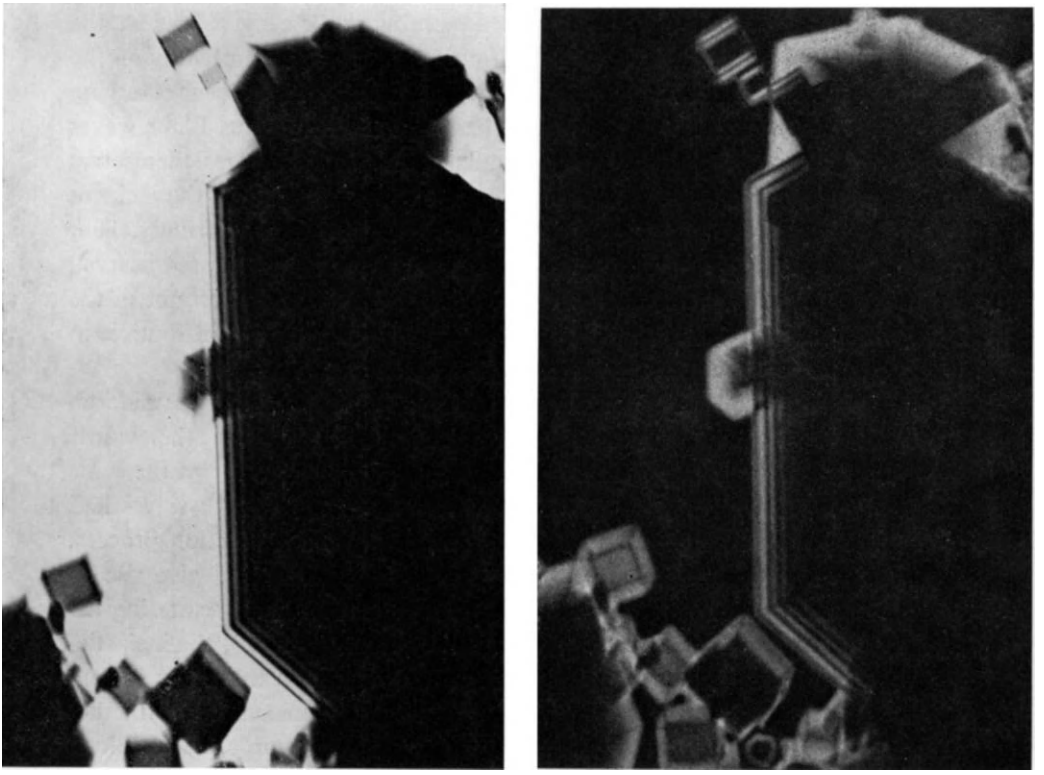


Fig. 12. – Magnesium oxide smoke particles (bright field). Left: elastic image; right: 90 eV loss image. Magnification 120000× (El Hili). (Courtesy of *Compt. Rend. Acad. Sci. Paris.*)

Now, Fig. 12, which is of thickness fringes produced by a crystal of magnesium oxide, shows that such an interpretation cannot be valid. The fringe contrast is the same in the image formed by electrons which have excited four plasmons (90 eV energy loss, Fig. 12 right) as in the elastic image (Fig. 12 left). It is clear that the relative importance of a process where the energy loss would occur in a region of a finite width located outside the crystal would become proportionally less significant for thick specimens and multiple losses.

As a matter of fact, it was observed that whatever the crystal thickness and the number of elementary losses suffered by the electrons, the dark fringes remain perfectly dark on the inelastic images, so that electron-plasmon interaction may be considered as fully coherent. The similarity of the diffraction contrast observed in elastic images and inelastic images arising from

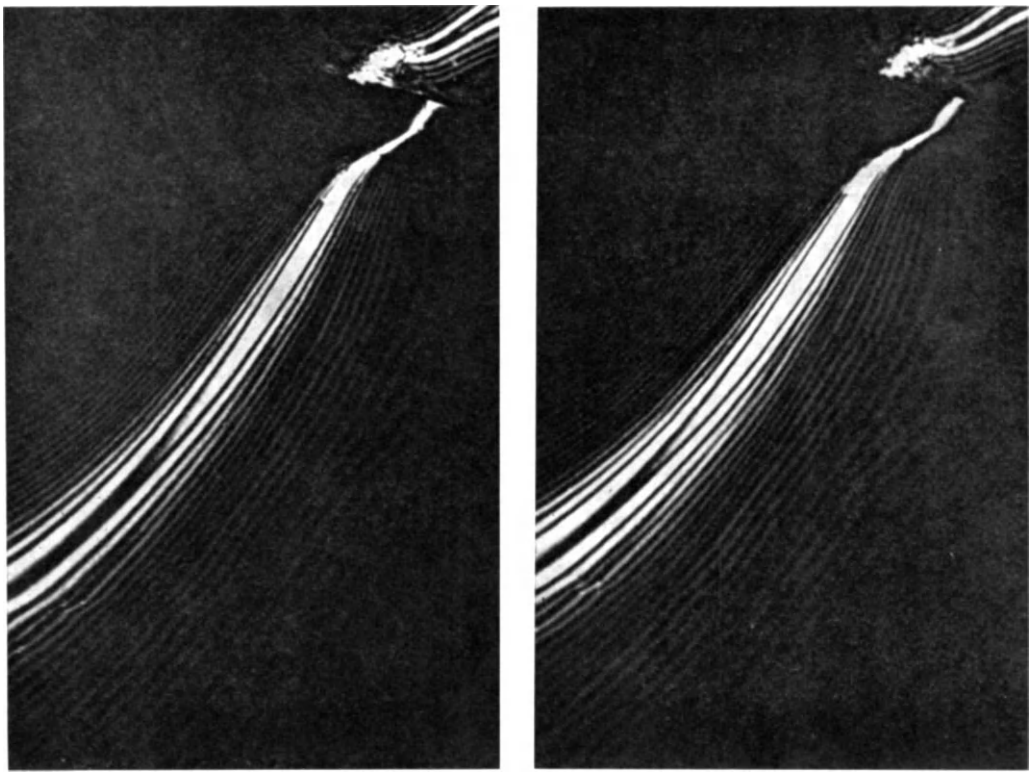


Fig. 13. – Aluminium foil (dark field). Left: elastic image; right: 14.6 eV loss image. Magnification 50000× (El Hili). (Courtesy of *Journ. de Microscopie.*)

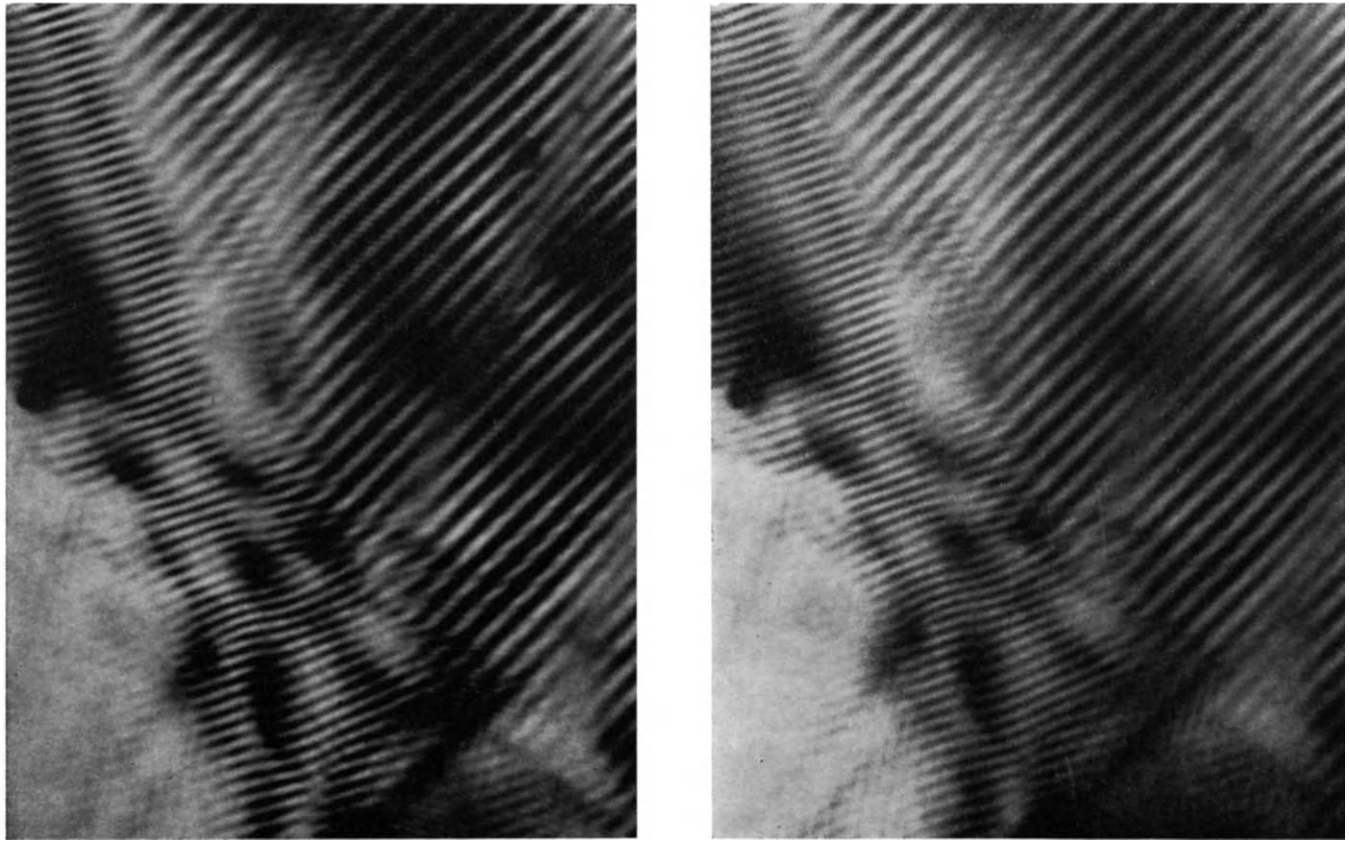


Fig. 14. – Moiré patterns (graphite). Left: elastic image; right: 28 eV loss image. Magnification 130000× (Henry). (Courtesy of *Compt. Rend Acad. Sci. Paris.*)

plasmon excitation is not limited to thickness contours. It is observed on bend contours as is shown in Fig. 13 (aluminium specimen, dark field images), on moiré patterns as is shown in Fig. 14 (graphite flakes). The contrast produced by a stacking fault, which is characterized by a symmetrical array of fringes in the bright field image and an asymmetrical one in the dark field image, is exactly the same on the elastic image and on the inelastic image arising from plasma excitation (Fig. 15). The same is true for the contrast arising from dislocations or for complicated patterns produced by the interaction of many wave fields.

All those results are in perfect agreement with the theoretical predictions

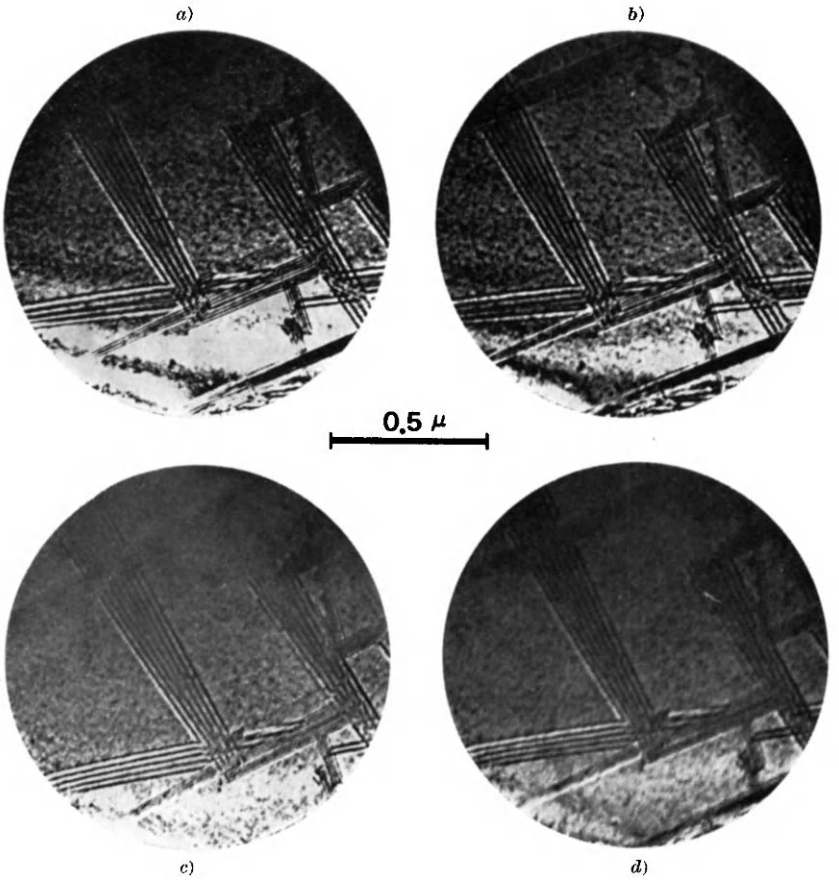


Fig. 15. – Stacking faults in a cobalt-chromium alloy. *a*) Bright field elastic image; *b*) Bright field inelastic (21 eV loss) image; *c*) Dark field elastic image; *d*) Dark field inelastic (21 eV loss) image (El Hili). (Courtesy of *Journ. de Microscopie.*)

of Howie ([17]); in the case of plasmon scattering, the interaction potential is a long range one; as a consequence, a very large proportion of the transitions of the fast electrons are intraband transitions which maintain the phase relationships between the Bloch waves, so that the fringe contrast is not modified.

Let us consider now the case of the energy losses which belong to the continuous part of the spectrum, between the characteristic losses. Those losses are very faint and we found it more convenient to use a different method for studying them. We have seen in the previous lecture ([4]) that, because of the aberrations of the magnetic filtering device, the value of the energy loss which is being selected by the slit depends on the inclination of the electron trajectory with respect to the optical axis. As a consequence, if a wide beam is used, which corresponds to an observation at low magnification of a large region of the sample, or alternatively if the instrument is off-centred so that the optical centre is outside the image, the selected energy loss may be well defined at each image point, but it decreases as the image point moves

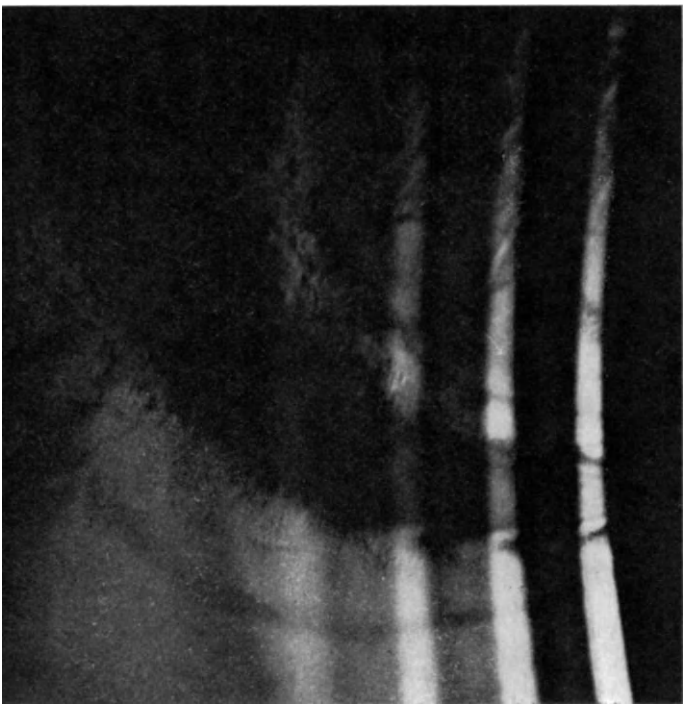


Fig. 16. – Extinction contours (aluminium) with superimposition of the energy loss spectrum. Magnification 30 000× (El Hili). (Courtesy of the *Journ. de Microscopie.*)

away from the optical centre, so that an energy spectrum is superimposed onto the micrograph. Figure 16 was obtained from an aluminium specimen by using this technique and opening the energy selecting slit slightly to lower the exposure time: the extinction fringes are clearly visible in the parts of the image which are produced by the continuous part of the energy loss spectrum. This is probably due to the fact that these losses are arising essentially from the excitation of single electrons, giving rise to electron-hole transitions. The scattering potential which is involved in such an interaction is a screened Coulomb potential of long range nature and it gives rise mainly to intraband transitions.

2·4.3. *The special case of thermal scattering.* – Quite interesting results are obtained ([10,11]) when studying the coherency of the waves which have been scattered diffusely in the vicinity of the Bragg spots. An aperture is located at the image focal plane as is shown in Fig. 17. By positioning the aperture

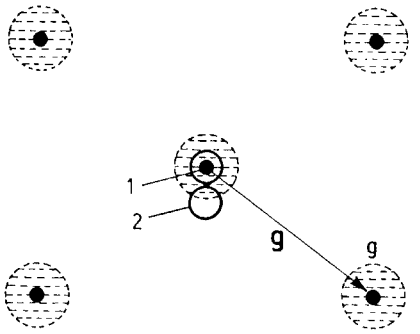


Fig. 17. – Location of the aperture at the image focal plane of the objective lens.

on the central spot or on the Bragg spot, one obtains a bright field or a dark field image, respectively; when positioning it in the vicinity of the central spot or of a Bragg spot, the image is produced by diffusely scattered electrons. This result was used by Kamiya and Uyeda for roughly separating the elastic electrons from the inelastic ones in a conventional electron microscope; we have combined that technique with the use of the filtering device, so that it is possible in each case to separate all the component images of various energies; for example, the image produced by electrons diffusely scattered near a Bragg spot may be separated into a quasi-elastically scattered image (energy loss less than 1 eV) and an inelastically scattered image produced by

the plasmon loss. It is observed that the inelastically scattered image shows the same fringe contrast as the elastic or the inelastic image which is observed when the aperture is positioned at the neighbouring Bragg spot; that results from the fact that electron-plasmon interaction maintains the coherency between the waves. But a most interesting point is that we found some amount of fringe contrast in the quasi-elastic image produced by electrons which had been diffusely scattered around the Bragg spot.

Now, if we consider the quasi-elastic processes only, the main part of the scattering out of the Bragg spots is due to the thermal diffuse scattering, in other words to electron-phonon interactions, at least in the case of pure samples reasonably devoid of surface imperfections or impurities. Plasmon scattering and core electron excitation do not contribute to the quasi-elastic image because they give rise to large energy losses; for the same reason, the major part of the single electron excitation processes in the conduction band do not need to be taken into account; we shall return to this point a little later. Thus we were led to the conclusion that some degree of coherence remains in the phonon scattered electron waves.

More precisely, we concluded that some amount of phase relationship was retained after the quasi-elastic scattering process for the various Bloch waves scattered in a given direction near the Bragg spots.

Now it is often considered that in the case of phonon scattering, the various processes which give rise to intraband and interband transitions are associated with different modes of vibration of the lattice (in other words that they correspond to the excitation or quenching of different thermal vibrations), so that they are incoherent and cannot give rise to diffraction contrast; and some experimenters proposed an alternative explanation for the observed contrast: the diffraction contrast, in the image produced by electrons scattered quasi-elastically near the Bragg spots, would arise from spurious scattering processes in amorphous oxide or contamination surface films, whereas phonon scattering would contribute to the background and result in a general lowering of the observed contrast. Obviously, scattering in the top contamination layer for instance would have the same effect as tilting the gun: a small part of the direct beam would be in such a case directed towards the aperture, giving rise to the classical extinction contrast observed on the Bragg spots.

A further series of experiments ([11,19]) has made clear that the objection was not valid.

As a matter of fact, similar effects appeared in MgO samples which obviously are devoid of oxide layers, but some spurious scattering by a

contamination film was still possible. For eliminating this possibility, the energy-selecting microscope was provided with an anticontamination device; this device operated quite satisfactorily since the image of a metallic sample remained unchanged after 20 min of electron bombardment. We observed in such conditions electropolished gold layers prepared from gold foils of high purity (better than 99.99 %); it was found again ([11]) that noticeable diffraction contrast appears in the images produced by quasi-elastic scat-

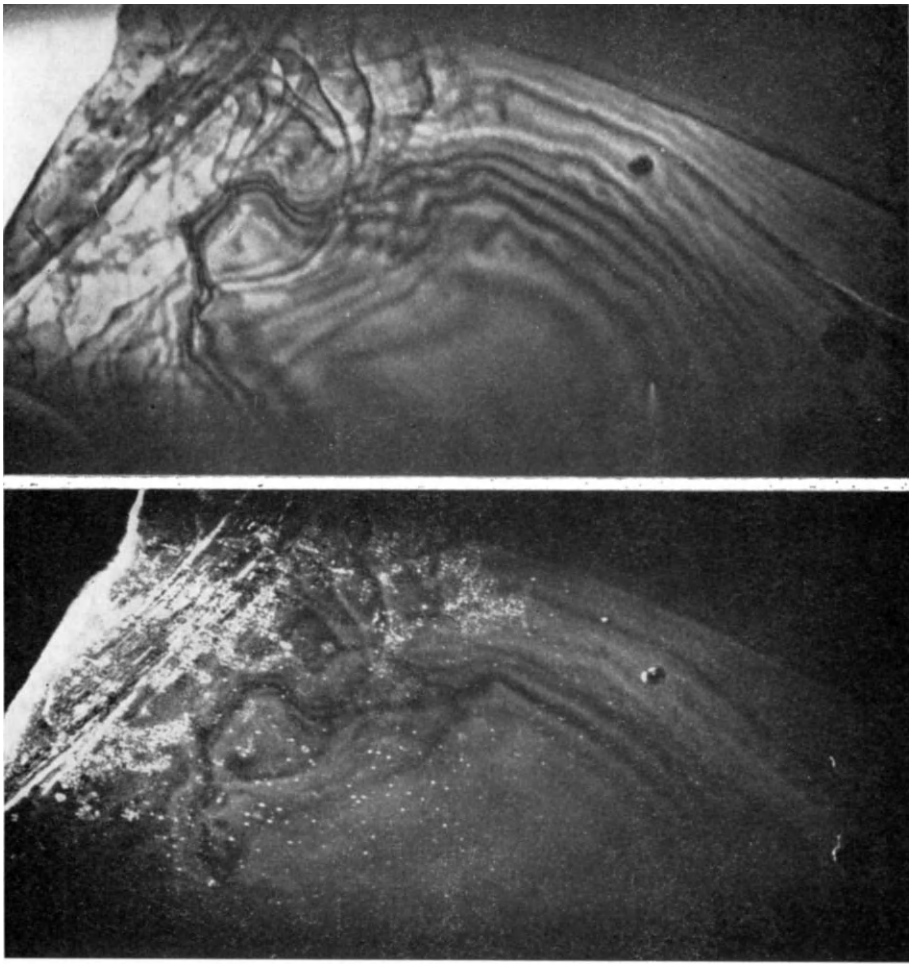


Fig. 18. – Electropolished gold foil. Top: Elastic bright field image; bottom: quasi-elastic image (diffuse scattering near the central spot). Magnification 60 000 × (Henoc). (Courtesy of *Compt. Rend. Acad. Sci. Paris.*)

tering near the Bragg spots, at least in the case where low-index reflections are excited. Figure 18 (bottom) shows bend contours so obtained on a gold sample by Henoc. The image obtained from electrons diffusely scattered in the direction $\frac{1}{10}(111)$ shows the same type of diffraction contrast as does the bright field elastic image (Fig. 18 (top)) but the contrast is much weaker. In such a case the observed contrast cannot be considered as arising from spurious scattering in amorphous surface films.

On the other hand, Henoc was able to verify that the temperature dependence of the diffuse scattering agreed with what would be expected from pure thermal scattering ([11]). Furthermore a specific criterion, based on the splitting of the fringes which occurs, if the scattering is due to a surface layer, when tilting the gun or moving the objective aperture, has confirmed directly that the observed diffuse scattering was due to the lattice itself and did not arise from spurious layers ([19]).

As a result, there is strong experimental evidence that a noticeable amount of coherency is retained in the quasi-elastic scattering processes near the Bragg spots. Let us discuss now the origin of this partial coherency.

We suppose that the incident beam is at the exact Bragg conditions for a low index reflection (say (111) for instance); the scattered beam is observed very near to the Bragg spot (say at $\frac{1}{10}(111)$), so that the two-wave approximation is valid for the primary wave and approximately valid for the scattered wave. The situation is depicted in Fig. 19. The primary wave is represented by a combination of two Bloch waves whose wave vectors originate from A^1 and A^2; the scattered wave which is selected for producing the image is represented by two Bloch waves originating from A'^1 and A'^2. In the case of the excitation of a phonon, A'^1 and A'^2 lie on the dispersion surface corresponding to an energy $E_0 - \delta E$; δE is typically equal to 0.01 eV for phonon scattering. In any case, δE is limited to about 1 eV by the filtering device, so that the dispersion surfaces are very near to one another.

The transitions we have to consider are: two intraband transitions, $A^1A'^1$ and $A^2A'^2$, and two interband transitions, $A^1A'^2$ and $A^2A'^1$. Now, as Fujimoto and Kainuma ([20]) pointed out in 1963, some lack of conservation of momentum is possible, in the case of thin samples, in the z direction perpendicular to the sample surface; the extra momentum is given up to the sample as a whole. In terms of wave vectors, the permitted error is of the order of $2\pi/d$, d being the specimen thickness. On the other hand, exact momentum conservation must prevail in the x and y directions lying in the plane of the surface. If d is of the order of the extinction distance ξ_g for instance, the permitted error δk_z is of the order of $\frac{1}{2}(A^1A^2)$ so that it is

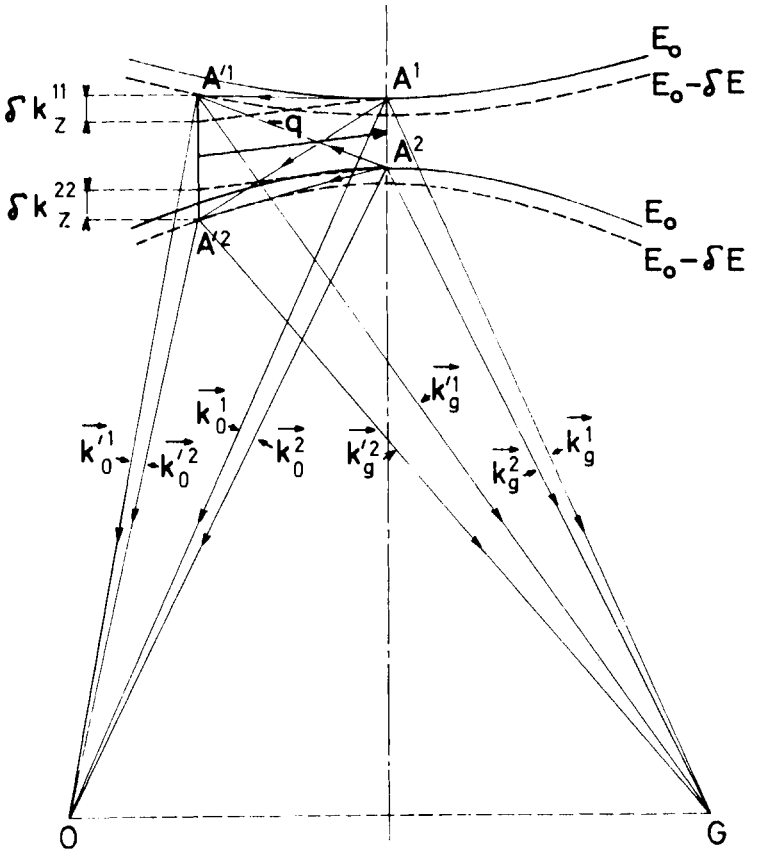


Fig. 19. – Dispersion surfaces and wave vectors.

seen easily that, if $A^1A'^1$ is about $g/10$, both intraband transitions may be associated with the excitation of the same phonon $-q$, whereas the interband transition $A^1A'^2$ for example cannot be produced with a noticeable probability amplitude by the excitation of the phonon $-q_{21} = -A^2A'^1$. As a result, in the case of thin samples (for gold and 80 kV electrons, the required thickness is $d \ll 5\,\mu$m for the (111) reflection and $d \ll 0.5\,\mu$m for the (220) reflection) the intraband transitions must be taken as coherent whereas the interband transitions are generally incoherent.

This now allows the fringe contrast produced by phonon scattering to be estimated; the calculation was made by Natta ([13]) in our laboratory, using the Born approximation, a rigid ion model for the scattering potential

and the Debye approximation for the phonons, considering one-phonon processes only.

The theoretical values of the relative intensity of the thermal scattering is found to agree with the observed value within the limit of experimental error; but the calculated contrast of the fringes is generally lower than what is observed. It should be pointed out that multiphonon processes, which make the treatment considerably more complicated, were neglected. More detailed calculations and careful experiments are necessary in order to understand the cause of the remaining discrepancies.

The observed contrast could be reinforced by some amount of scattering due to single electron excitations (we have seen that such an interation involves a long-range potential); but it may be shown ([13]) that the excitation of conduction electrons does not play any important role in such quasi-elastic processes. Rough calculations indicate that these interactions, which result in complete coherency of the scattered electron waves, give rise to a negligible intensity for the scattered waves when the energy loss lies in the range $(0 \div 1)$ eV; it is to be noted that the efficiency of the process is much higher if the selected energy loss lies in a range of the same width situated at a higher value (say between 3 and 4 eV); this explains the strong contrast we have observed in the images produced by energy losses located in the continuous spectrum, between the plasmon losses.

To end with, I would like to point out that we have considered the interband transitions as incoherent, which is correct in the present case when the q_{12} and q_{21} vibration modes are independent of one another. Now, if it were possible to realize conditions such that the mean free path of the phonons is much larger than the specimen thickness, and if the specimen surfaces were atomically flat and parallel, the vibration modes q_{12} and q_{21} would be coupled together; in other words the quantization of the phonons would involve standing waves in the z direction, leading to a high degree of coherency in the interband processes. For a clean specimen the standing waves would exhibit antinodes at the free surfaces. The calculation shows that in this case the phonon scattered image would exhibit strong diffraction contrast; but the contrast would be reversed, so that scattering near the central spot for example would give rise to the contrast of a dark field image. Furthermore the contrast would be very sensitive to the presence of surface layers modifying the vibrational modes. Such experiments would require perfect crystals with very smooth surfaces, and which would be non metallic for increasing the mean free path of the phonons; they would be observed at low temperature for reducing the phonon-phonon interactions, and the

exposure time, which is typically one or two minutes, would be increased to about one hour in the absence of an image-intensifying device; nevertheless, we believe that it would be worthwhile to try to overcome the experimental difficulties.

REFERENCES (Section 2)

1) H. BOERSCH: *Optik*, **5**, 426 (1949).
2) H. BOERSCH: *Zeits. Phys.*, **134**, 156 (1953).
3) R. BEAUFILS: *Compt. Rend. Acad. Sci. Paris*, **248**, 145 (1959).
4) R. CASTAING: Section **1**.
5) J.-F. HENNEQUIN: Diplôme d'Etudes Supérieures, Paris (1960).
6) N. PARAS: Diplôme d'Etudes Supérieures, Paris (1961).
7) R. CASTAING and L. HENRY: *Compt. Rend. Acad. Sci. Paris*, **255**, 76 (1962).
8) L. HENRY: *Thesis*, University of Paris (1964); *Bull. Soc. Franç. Minéral. Crist.*, **88**, 172 (1964).
9) R. CASTAING, A. EL HILI and L. HENRY: *Compt. Rend. Acad. Sci. Paris*, **261**, 3399 (1965).
10) A. EL HILI: *Thesis*, University of Paris (1967); *Journ. de Microscopie*, **5**, 669 (1966); **6**, 693 (1967); **6**, 725 (1967).
11) R. CASTAING, P. HENOC, L. HENRY and M. NATTA: *Compt. Rend. Acad. Sci. Paris*, **265**, 1293 (1967).
12) H. WATANABE and R. UYEDA: *Journ. Phys. Soc. Japan*, **17**, 569 (1962).
13) M. NATTA: *Journ. Phys.*, **29**, 257 (1968).
14) S. L. CUNDY, A. J. F. METHERELL and M. J. WHELAN: *Phil. Mag.*, **15**, 623 (1967).
15) C. COLLIEX and B. JOUFFREY: *Compt. Rend. Acad. Sci. Paris*, **270**, 144, 673 (1970).
16) R. D. HEIDENREICH: *Journ. Appl. Phys.*, **34**, 964 (1963).
17) A. HOWIE: *Proc. Roy. Soc.*, **271**, 268 (1963).
18) Y. KAMIYA and R. UYEDA: *Journ. Phys. Soc. Japan*, **16**, 1361 (1961).
19) P. HENOC, M. NATTA, L. HENRY and R. CASTAING: *Compt. Rend. Acad. Sci. Paris*, **267**, 756 (1968).
20) F. FUJIMOTO and Y. KAINUMA: *Journ. Phys. Soc. Japan*, **18**, 496 (1963).

3. Ion emission microanalysis.

3·1. Introduction.

The secondary ion microanalyzer which was developed ten years ago in our laboratory gives local chemical and isotopic analysis by using the secondary ion emission that accompanies the phenomenon of cathodic sputtering of a target to form the image of the surface of a sample ([1-3]).

The principle of such an analytical procedure is essentially different from that of the well-known technique of electron probe microanalysis since the particles which are used for detecting the various chemical species are ions instead of X-ray photons. But another difference lies in the operation of the instrument itself. In the original electron probe microanalyzer ([4]), point by point analysis was obtained by directing a finely focussed electron beam onto any desired point of the sample; then scanning facilities were introduced ([5]) which made it possible to get, on the fluorescent screen of an oscilloscope, an image with intensity proportional to the local concentration of the element being detected. The same mode of operation may be used in secondary ion microanalysis; the possibilities of such an « ion microprobe » will be discussed briefly in the course of this lecture. However, the mode of operation of the initial ion microanalyzer is quite different, since the distribution image of any element or isotope is obtained directly, without having to resort to a scanning procedure. The instrument is essentially a « mass-selecting » secondary ion microscope, which makes possible to « see » separately the various component elements (or isotopes) of a solid sample.

After a brief description of the instrumental arrangement, which makes use of the dispersive optical system we have described elsewhere ([6]), the main part of the lecture deals with the experimental and theoretical work carried out by Slodzian, Hennequin, Joyes, Blaise and Brochard in our laboratory for disentangling the physical processes involved in the production of the secondary ions. To end with, some examples of applications are given.

The reader is assumed to possess a basic knowledge of atomic and solid state physics, at an elementary level.

3·2. General description of the secondary ion microanalyzer.

The instrument is essentially a mass-selecting secondary ion microscope, which operates as follows.

The surface of the solid sample, which is flat and optically polished, is bombarded over an extended area (whose diameter is a little less than half a millimeter) by a primary beam of ions (generally A^+ ions whose energy is about 10 keV); the primary ion density is of the order of 100 μA per square millimeter. The sample is progressively etched down and a significant proportion of the particles which are extracted from it by the bombardment leave the surface as ions. Those « characteristic » secondary ions are formed from the atoms which were present in the first atomic layers; they leave

the bombarded surface with low energies, between zero and some tenths of electron-volts; it is therefore possible, by focusing them with an emission lens, to obtain an image of the sample surface which is produced by using all the secondary ions. This « material image » can then be separated by mass spectrography into its component « characteristic images »; each of the characteristic images is carried by a given type of ion and it brings with it a map of the distribution of the corresponding element, or isotope, across the specimen surface.

The diagram of the first experimental apparatus, which was achieved in 1962, is shown in Fig. 20. The beam which carries the total image produced

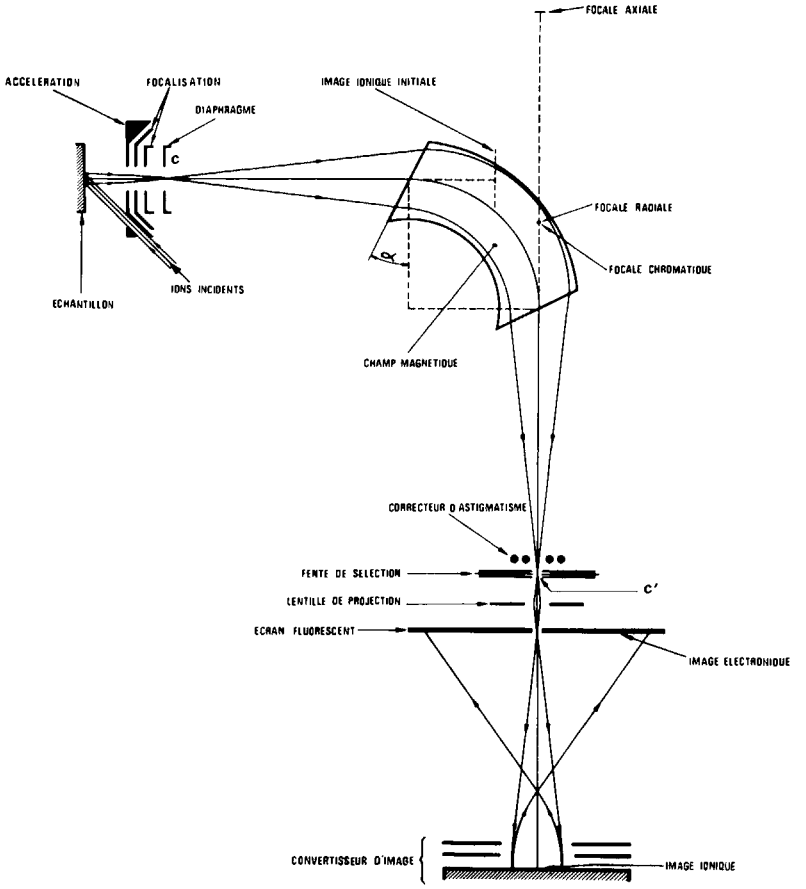


Fig. 20. – Diagram of the original microanalyzer (Slodzian). (Courtesy of *Compt. Rend. Acad. Sci. Paris.*)

by the emission lens goes through a magnetic sector where the value of the induction is chosen in such a way that the desired characteristic ions are deviated by 90°. For a convenient value of the angle of incidence of the beam relative to the pole faces, the beam originating at the cross-over of the emission lens converges towards an exit cross-over where the ions are selected by a slit. The characteristic image carried by these ions is transformed by the magnetic prism into a virtual image whose (strong) astigmatism is corrected by a stigmator. A set of lenses projects this image as a real one onto the cathode of an ion-electron converter, so that the final electron image observed on the fluorescent screen reproduces the surface distribution of the element which has been selected for analysis.

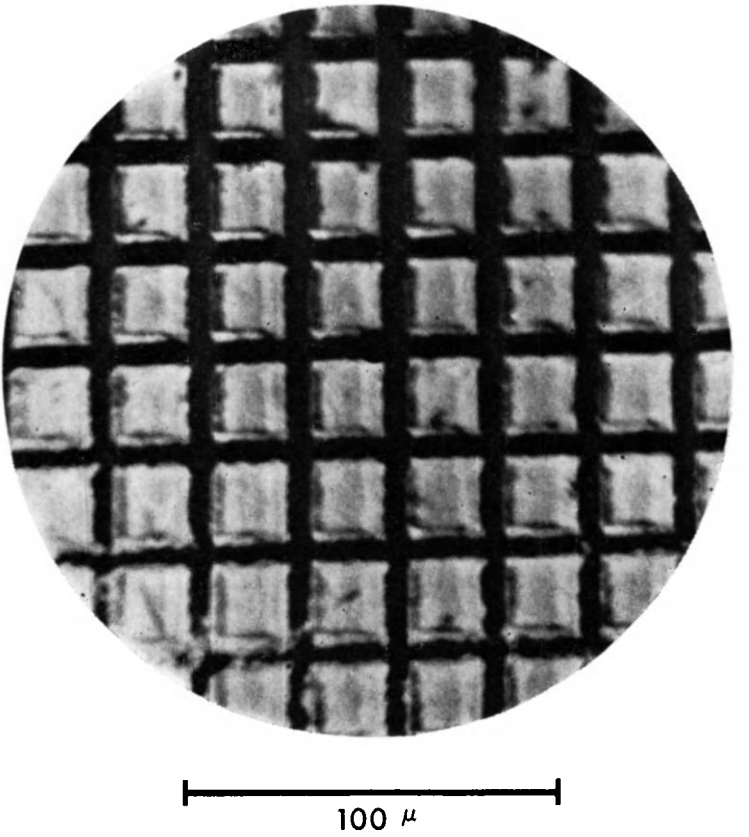


Fig. 21. – Copper grid on an aluminium block. Al^+ image (Slodzian). (Courtesy of *Ann. de Phys.*)

The strong chromatic aberration which could result from the fact that the secondary ions are emitted with various energies, and pass subsequently through a highly dispersive system, can be eliminated to the first order by positioning the image at the level of an achromatic point which is located, in this simple arrangement, inside the magnetic prism, at a distance from the pole face equal to $\frac{2}{3}$ of the radius of curvature.

The behaviour of this first experimental apparatus was quite satisfactory. The distortions of the various lenses could be made to compensate one another by a suitable choice of the excitation of the projection stage. One of the first pictures we obtained is shown on Fig. 21. It represents the image, produced by Al^+ ions, of a composite sample prepared by pressing a copper grid (25 μm) onto an aluminium block. The distortion is negligible and the resolving power is one micron.

In this first experimental arrangement, the images were obtained by simply photographing the fluorescent screen from outside with a conventional camera. Such a procedure led to a serious reduction of the detection sensitivity; now, it is easy to show that the attainable resolving power is proportional to the sensitivity of detection of the characteristic images. This is due to the fact that the image is produced from the material of the sample itself, so that some minimum volume of material must be destroyed for producing an image point.

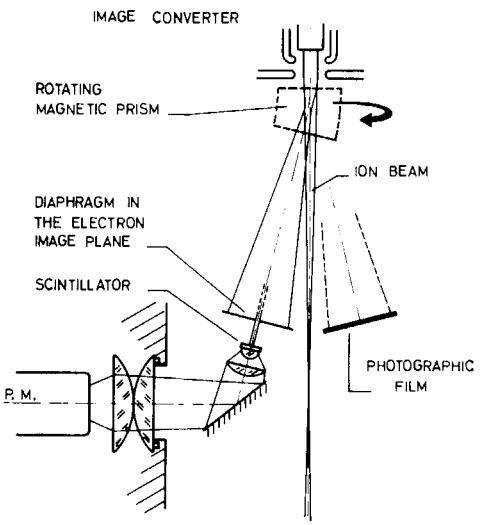


Fig. 22. – Detail of the image converter (Cameca). (Courtesy of *Proceedings 5th International Congress on X-Ray Optics and Microanalysis.*)

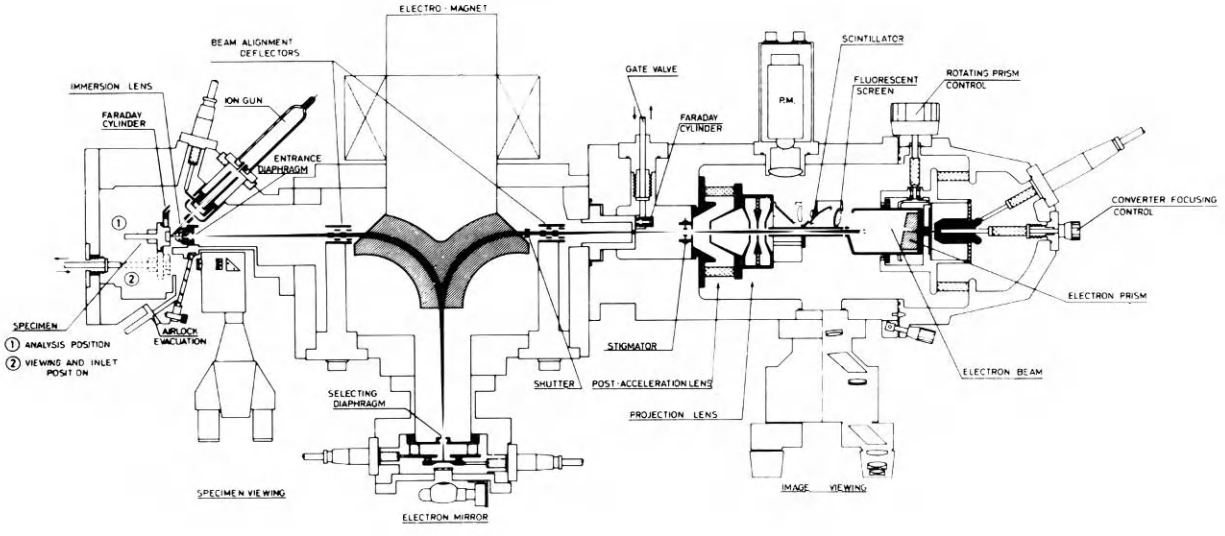


Fig. 23. – General diagram of the commercial model (Cameca) of the secondary ion microanalyzer. (Courtesy of *Proceedings 5th International Congress on X-Ray Optics and Microanalysis.*)

The minimum diameter of the analysed region may be estimated theoretically; for instance, in the case of an aluminium base sample or of an ionic compound, where secondary ion emissions are especially high, it is of the order of 500 Å units. Such a resolving power requires a direct registering of the resulting images on a camera located inside the vacuum; furthermore, the initial magnification produced by the emission lens must be large enough to ensure that the image quality will not be destroyed by the aberrations of the dispersive system. Such an increase in the initial magnification results in a reduction of the diameter of the imaged part of the sample.

However, a serious drawback of this first experimental instrument was that the magnetic sector gave momentum-discrimination, instead of the mass-discrimination which is really what is required. As a result, the separation of neighbouring masses was imperfect, especially for heavy elements. This difficulty was overcome in a more elaborate instrument, which has been commercially developed in France, which uses the more sophisticated dispersive device, including two deflections at 90° separated by a reflection at an electrostatic mirror, which is described in another lecture ([6]). As a result, neighbouring masses are perfectly separated, even for heavy elements.

Another improvement in the industrial model is the possibility of registering the characteristic images directly on a photographic film, inside the apparatus. By means of a rotating magnet (Fig. 22), the electron beam which carries the final image may be directed, during the operation, either towards the photographic film, or towards a fluorescent screen for visual observation of the images. A pinhole in the center of the fluorescent screen makes possible to isolate the emission from a very small area of the sample (2 μm for example); this area is easily chosen on the image. The corresponding electrons are received by a scintillator followed by a photomultiplier. A diagram of the instrument is shown on Fig. 23.

3·3. **Possibilities and limitations of secondary ion microanalysis.**

The range of application of this technique is extremely wide, in various fields such as metallurgy, mineralogy or even biology. It has some important advantages over the classical technique of electron probe microanalysis:

— First of all, the number of image-forming particles (secondary ions) is much higher, all things being equal, than the number of X-ray photons in a scanning microprobe. As a result, the images are devoid of statistical noise and the sensitivity to extremely low concentrations is much greater.

— Isotopic analysis is possible on a micron scale.

— The lighter elements such as hydrogen are easily detected.

— Spatial resolution in the plane of the sample surface is better and it is not limited by the diffuse penetration of the primary particles (the penetration of the primary ions is much less than that of the electrons of a microprobe).

— And, last but not least, the depth resolving power is extremely good: the analysis may be restricted to the first atomic layers if one does without a high resolving power in the plane of the surface.

In addition to all these advantages, there is a serious drawback: the relation between the intensity of emission of a characteristic ion and the local concentration of the corresponding element is complicated (except in the case of isotopic mixtures) and very sensitive to chemical bonds. As an example, Fig. 24 shows inclusions of cuprous oxide Cu_2O embedded in a matrix of pure copper. The image is produced by the ions $_{63}Cu^+$; it is seen that the oxide emits many more copper ions than does the pure copper matrix; this is due to the ionic character of the compound Cu_2O.

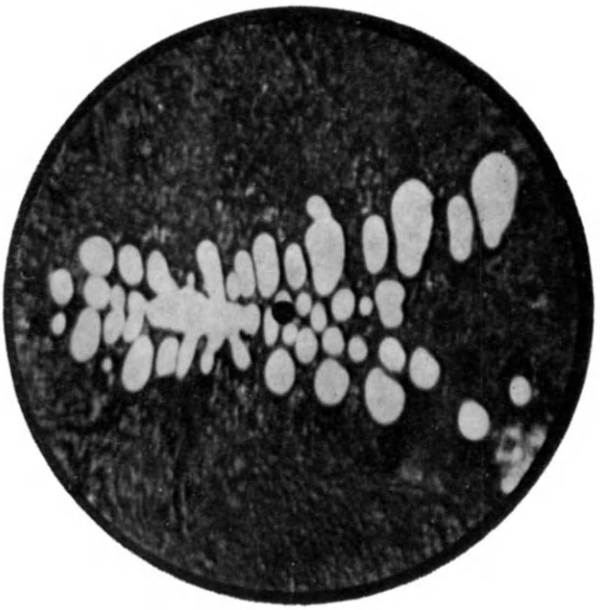


Fig. 24. – Cu_2O inclusions in copper. $_{63}Cu^+$ image (Slodzian). Magn.: 400×. (Courtesy of *Ann. de Phys.*)

Even in the case of metallic alloys, strong matrix effects are commonly observed when the component metals are very dissimilar. In the course of our first experiments, we observed that the emission of Cu^+ ions from a copper-beryllium alloy containing 2% beryllium was 50 times larger than that of pure copper, in the same conditions of primary bombardment. As a matter of fact, we discovered later that such matrix effects are due to surface phenomena; they are very much reduced if the vacuum is good enough and if the etching speed of the sample is high enough to keep the surface permanently clean.

3·4. The various processes involved in secondary ion emission.

One of the reasons why this phenomenon of secondary emission is so complex lies in the fact that various physical processes may be involved in the production of the secondary ions. In this respect, we were able to distinguish two essential processes; we shall denote them as the « kinetic » process and the « chemical » process.

The kinetic process occurs in a metallic sample when an ion core suffers a collision strong enough to eject one or more of the core electrons. The increase in ionization so produced is screened by the conduction electrons, but the metastable state (constituted by this screened ionization) may, if its lifetime is long enough, be preserved until the knocked-on particle leaves the lattice, having meanwhile suffered many collisions which result in a lowering of its velocity. The particle leaves the sample in a neutral metastable state, whose Auger de-excitation, as we shall see below, may give rise to the production *in vacuo*, very near from the sample surface, of a positive ion. The phenomenon essentially depends upon the electronic structure of the free atom and the band structure of the metal.

The chemical process depends upon the chemical bonds in the sample. In the oxides, for instance, the breaking of ionic bonds gives rise essentially to negative oxygen ions and positive metallic ions. The ionization ratio, that is the ratio between the number of secondary ions and the number of corresponding neutrals, is generally much higher in the chemical than in the kinetic process.

This is the reason why the Cu_2O compound (Fig. 24) emits much more Cu^+ ions than the pure copper, other things being equal.

It is clear, under such conditions, that it is not possible to deduce the concentration of a metallic element M from the general emission of M^+

secondary ions, if the sample contains for example inclusions of oxide whose diameter is less than the limit of resolution of the image. On the other hand, if the inclusions are large enough to be visible on the image, they will be identified easily from the spectrum of their secondary ions.

It may be noted at this point that negative secondary ions can be used as well for producing the image, by simply changing the potentials on the lenses and the sign of the magnetizing current in the prism. Figures 25 shows images obtained by Rouberol *et al.*, by using VC^- ions (top) and C_2^- ions (bottom), on a sample of pig iron containing vanadium; the light areas correspond to inclusions of vanadium carbide and of graphite in the pig iron.

Let us return now to the case of a pure metallic sample; the only process which can be involved is the kinetic one, if the sample is located in a perfect vacuum and bombarded with noble gas ions which do not interact chemically with it.

Now, such an ideal case cannot be found in practice. The sample is surrounded by a gaseous phase (the residual gas inside the instrument), some components of which may be chemically active; furthermore, the primary beam may contain chemically active impurities if it is not filtered carefully. Under such conditions, a contaminated layer is present on the target surface; it is produced by the reactions of the sample with chemically active species (such as oxygen or water) present in the gaseous phase or in the primary ion beam. If the primary beam is carefully filtered, the only source of contamination is the gaseous phase; as a consequence the thickness of the contaminated layer is an increasing function of the partial pressure of the active component in the surrounding atmosphere, and a decreasing function of the sputtering speed, which is itself proportional to the density of the primary ion beam (for a given accelerating voltage). This surface layer may lead to a « chemical » emission which is superimposed to the general kinetic emission of the metallic sample.

This phenomenon appeared with striking evidence in the series of experiments made some years ago in our laboratory by Miss Guénot [7], who measured the secondary ion emission of an aluminium sample as a function of the partial pressure of oxygen in the residual atmosphere of the apparatus. It is seen in Fig. 26 that the emission of the Al^+ ions shows a lower plateau at low pressures and an upper plateau at high pressures. The density of the primary beam (A^+ ions, 10 keV) was about 10 $\mu A/mm^2$; for such a primary density the sputtering speed is of the order of 10 atomic layers per second; as a result, when the oxygen partial pressure is lower than 10^{-6} torr (which

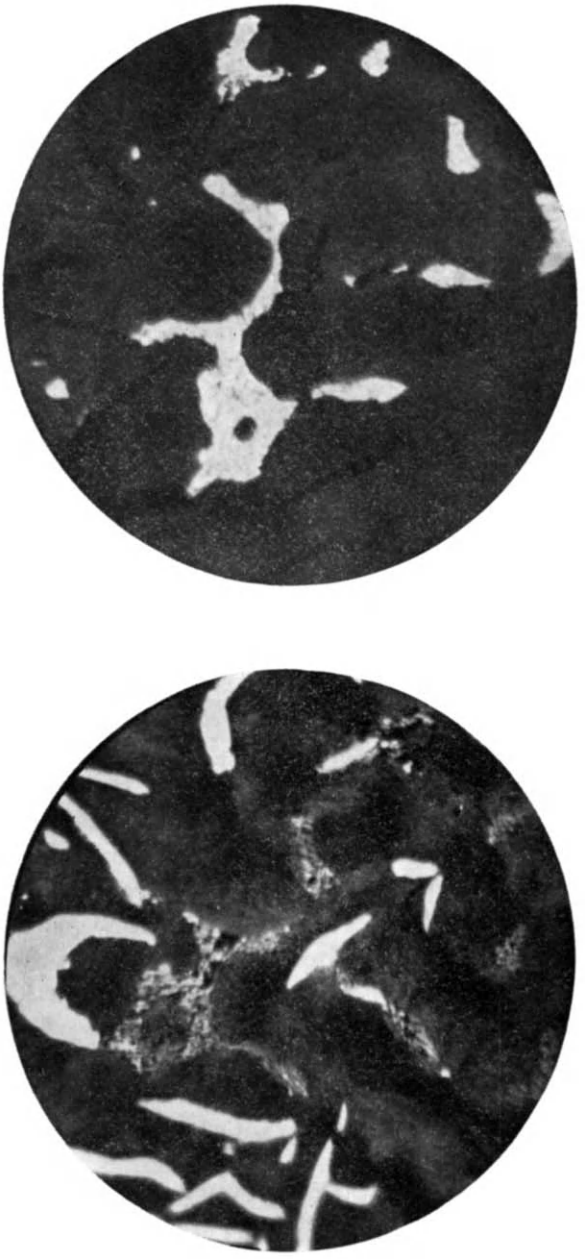


Fig. 25. – Pig iron. Top: VC^- image; bottom: C_2^- image (Rouberol). Magn.: 400×.

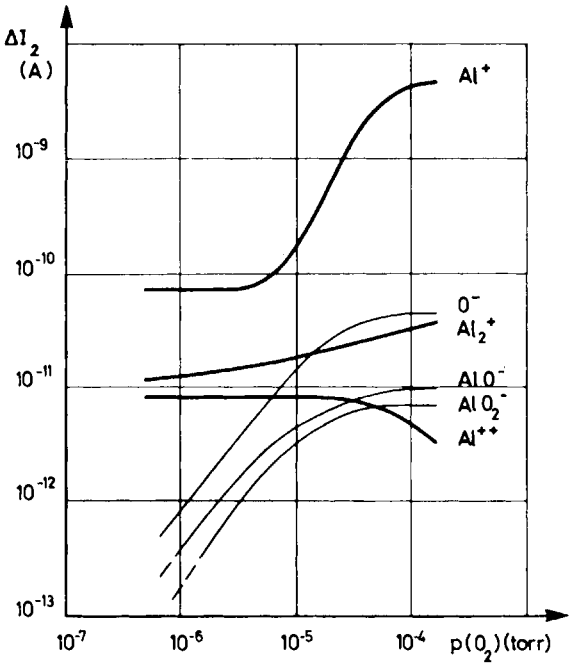


Fig. 26. – Emission of slow secondary ions as a functions of oxygen pressure. Primary beam: A^+, 10 keV (Guénot). (Courtesy of *Journ. de Phys.*)

corresponds to the deposition of one atomic layer per second), the surface is continuously cleaned off by sputtering, whereas a continuous oxidized layer is present at the free surface if the oxygen pressure is 10^{-4} torr or more, leading to an intensification of the Al^+ emission (through chemical processes) by a factor of 100. The intensification is much lower for molecular Al_2^+ ions, and the emission of the doubly ionized Al^{++} ions (which is essentially related to a kinetic process) is *decreased* by the presence of oxygen. The emission of the negative ions O^-, AlO^- and AlO_2^- saturates at high pressures; it disappears at very low pressures.

The experiments of Miss Guénot were carried out with the ionic microanalyzer, so that the measurements were restricted to the low energy component of the secondary ion emission only. More significant results were obtained a little later by Hennequin ([8]) in the course of an experimental investigation of the energetic and spatial distributions of the secondary ions; by integrating the distribution curves, Hennequin was able to plot the total ionic yield (that is, the ratio between the total emission of the secondary ions of a given type

and the number of primary ions impinging on the target sample in the same time) as a function of the residual oxygen pressure, for targets of Al, Mg, Si, Ti, Ni, and Cu. The curves so obtained for Al, Mg and Si are shown in Fig. 27; they show saturation values (on the high pressure side), which are

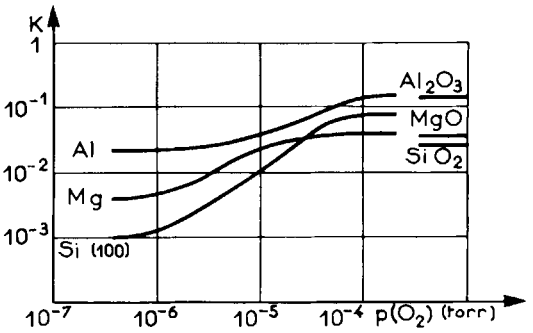


Fig. 27. – Ionic yields of Al, Mg and Si. Primary beam: A^+, 8 keV (Hennequin). (Courtesy of *Compt. Rend. Acad. Sci. Paris.*)

quite similar, in the case of Al and Mg, to the ionic yields of the pure oxides (Al_2O_3 and MgO). In the case of Si, the saturation yield is higher than the ionic yield of SiO_2, which might be due to the fact that the composition of the surface layer is similar to that of the monoxide SiO. The comparative measurements on the oxides were carried out on samples prepared by pressing together the powdered oxide and copper powder.

It is observed in general, when the residual pressure in the instrument is lowered below a limiting value which depends on the sputtering rate (all things being equal, this limiting value is proportional to the current density of the primary beam) that the emission of the various types of secondary ions is independent of the residual pressure; one can be sure in such a case that the ionic emission which is being measured is the pure « kinetic » emission of the bare metal; the parasitic ions are practically eliminated, at least after the very short time which is necessary for sputtering out any contaminated layer which could be present at the surface of the sample.

It may be noted at this point that some advantage can be taken from the high efficiency of the chemical process, by using chemically active ions, such as O^+ ions, for bombarding a metallic sample. The images so obtained are generally much brighter, and the sensitivity is higher for trace analysis; but the spectra are more complicated and strong matrix effects make the

quantitative interpretation more difficult, except in the case of low concentrations where calibration from standards is quite easy (*e.g.* doping elements in semiconductors) or isotope analysis.

3·5. The alternative procedure using an ion microprobe.

It is interesting to comment at this point on the alternative procedure which was developed later on by Long ([9]) and Liebl ([10]) for secondary ion microanalysis. These authors use a scanning technique: an ion microprobe is scanned across the specimen surface and the secondary ions are collected and mass-analysed to control the beam of an oscilloscope. The advantages lie in the simplification of the optics, a little better sensitivity due to the fact that a larger part of the energy spectrum of the secondary ions may be used for analysis, and a lower amount of redeposition on the sample of sputtered material reflected from the neighbouring surfaces. Now, the time for getting a distribution image is much larger, because the total ionic current which can be brought to focus on a one micron probe is less than 10^{-4} μA in the present state of the technique; so that, even if the efficiency of collection of the ions was 100 times higher, the time required for producing the same image as our instrument, where the primary intensity impinging on the imaged area is about 10 μA, would be 1000 times larger.

But a major disadvantage results from the contamination of the surface by the residual atmosphere. The *mean* sputtering speed of the sample is 10^5 times lower, so that the same amount of elimination of surface impurities would require vacua 10^5 times better! As a rule the specimen surface would be saturated permanently in surface impurities, and various types of composite ions would be emitted whose presence would be related more to the gaseous phase than to the sample itself.

This drawback would be reduced in the case of point analysis using a fixed probe, for the ionic primary density could then be made high enough to ensure cleaning of the surface in a good vacuum. But, even in this case, such a cleaning would occur in the central part of the probe only. The marginal parts would be contaminated and their strong chemical emission would falsify the interpretation of the analysis.

For all those reasons, we believe that, except perhaps in the case of an ionic compound *analysed with a fixed probe*, the technique based on a general bombardment of the sample, using a selection on the image of the spot to be analysed is much more suitable for quantitative work and much less sen-

sitive to artefacts arising from parasitic elements present in the residual atmosphere.

The redeposition of sputtered material onto the sample, which is disturbing only if a very low concentration must be measured at a given point while the analysed element is highly concentrated in the vicinity of that point, can be lowered at will by reducing the bombarded area, say to a diameter of 10 μm, and selecting the analysed spot by means of the pinhole in the fluorescent screen.

3'6. The main features of the « kinetic » process.

An extensive study of the « kinetic » emission has been carried out in our laboratory by Slodzian, Hennequin, Joyes, Blaise and Brochard; we will discuss briefly the results of that work, which has made it possible to understand the main features of the intricate phenomena involved in the kinetic process.

In the experimental arrangement used by Hennequin ([8]) for studying the energetic and spatial distributions of the secondary ions, the secondary beam emitted in any chosen direction goes through a grid which makes it possible to determine its energy spectrum by a counter-field method; it is then focussed into a convergent beam, filtered in mass (roughly) by a permanent magnet, and measured in a Faraday cylinder. The whole of the analysing device (exit aperture, counter-field grid, mass spectrometer) may rotate around an axis lying in the target surface; the target itself may rotate around the primary beam, so that the secondary emission may be analysed in any direction.

The energy distributions so obtained for the secondary ion emission of polycrystalline samples of copper and aluminium are shown on Fig. 28; the primary ions are impinging normally on the surface and the secondary emission is observed at 30° to the normal. It is seen that the mean energy of the Cu^+ ions is larger than that of the Al^+ ions. On the other hand, it was observed that, in the case of a normal bombardment by A^+ ions (8 keV), the mean energies of the secondaries are larger at angles close to the bombarded surface, when the direction of emergence is further from the direction of the bombardment. The same result is obtained for oblique directions of bombardment; it shows that the secondary ions are not produced by a thermal process and that a particle, when it is ejected from the surface, has suffered a small number of collisions only, insufficient for ensuring thermal equilibrium.

We are led to the same conclusion if we consider the spatial distribution of the secondary particles (ions or neutral atoms) ejected from single crystals.

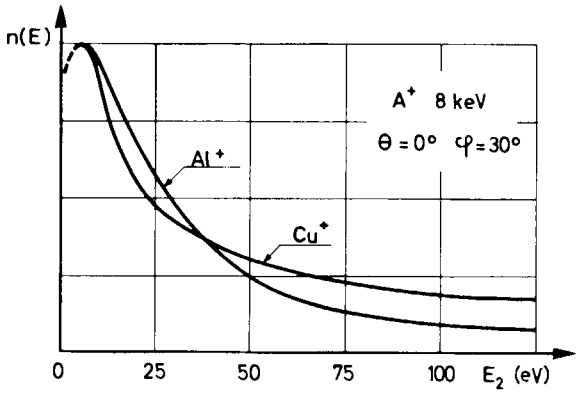


Fig. 28. – Energy distribution (arbitrary units) of secondary ions Al^+ and Cu^+ (Hennequin). (Courtesy of *Journ. de Phys.*)

Neutral atoms emitted from the sample can be collected in the same experimental instrument, either on a plastic film (in this case the distribution of the density of the deposit is measured by optical densitometry) or on a plate of pure graphite; in this last case the thickness of the deposit is measured by electron probe microanalysis. The curves represented on Fig. 29 were obtained by the second method; they show the spatial distribution of the neutral atoms ejected from the (100) face of a single crystal of Al; the strong

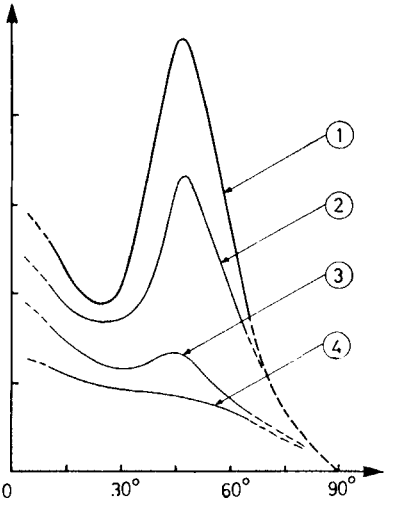


Fig. 29. – Spatial distribution (arbitrary units) of Al neutrals (Hennequin).

peak in the [110] close packed direction disappears gradually when the residual pressure of oxygen in the specimen chamber is increased (curves 1, 2, 3, 4 correspond to partial pressures of oxygen equal to $4 \cdot 10^{-7}$, $5 \cdot 10^{-6}$, $5 \cdot 10^{-5}$ and 10^{-4} torr, respectively). Figure 30 shows the spatial distribution (determined

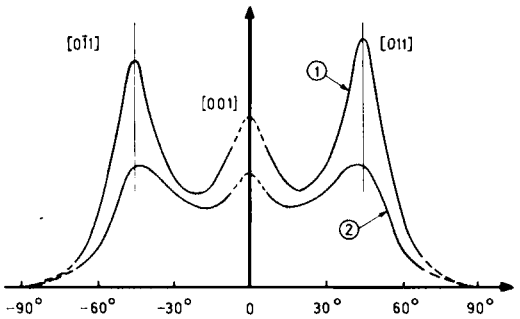


Fig. 30. – Spatial distribution of Cu neutrals (Hennequin). (Courtesy of *Compt. Rend. Acad. Sci. Paris.*)

by optical densitometry) of the neutrals ejected from a single crystal of copper ((100) orientation) bombarded at normal incidence. Curve 1) corresponds to a very low oxygen partial pressure (the presence of argon at a pressure of $2 \cdot 10^{-4}$ torr does not change the result); curve 2) shows the moderate weakening of the peaks which occurs when oxygen is introduced at a partial pressure of $2 \cdot 10^{-4}$ torr.

Let us return to the pure kinetic emission observed in good vacua. The angular distributions of the secondary ions and of the neutrals ejected from a copper target (single crystal or polycrystal) bombarded at normal incidence are shown in Fig. 31. The main result is that the maximum in the emission which is observed for the neutral atoms in the close packed directions of the crystal lattice does not appear at all for the ion emission.

All those results are consistent with the mechanism of kinetic emission we proposed some years ago ([11]) and whose quantitative theory was developed by Joyes ([12]): the origin of the secondary ion emission lies in the so-called « erratic emission » which occurs in all directions of space, whereas a large part of the sputtering arises from chains of focussed collisions ejecting neutral atoms in the close packed directions of the lattice only. When an ion core, in the metal lattice, is knocked on by a fast particle (primary ion or high energy displaced atom) an inner electron may be removed; the ion core moves inside the lattice and it may leave the surface as a sputtered particle (erratic

emission); as the ejection speed is much lower than the speed of the conduction electrons, the particle leaves the lattice as a neutral atom because electrons follow it easily to neutralize its charge. The empty level in the inner shell which had been produced initially is generally destroyed by Auger de-excitation before the atom leaves; but if the Auger lifetime is long enough, the empty level may still be present on the free atom after it has left the lattice.

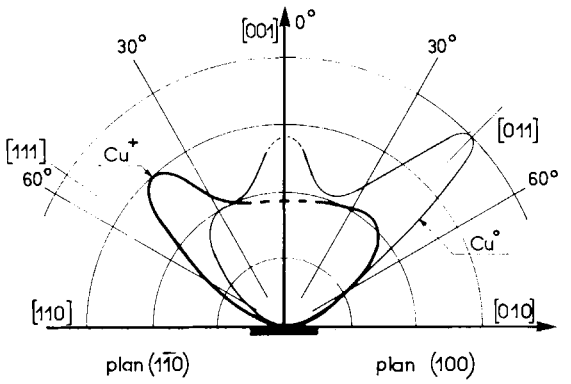


Fig. 31. – Spatial distributions of Cu^0 (neutrals) and Cu^+ (secondary ions) in the (100) and (110) planes (arbitrary units). Normal bombardment [001] (Hennequin). (Courtesy of *Journ. de Phys.*)

In that case, the Auger de-excitation occurs in vacuo and leads to the ejection of one electron (or more) and the production of a secondary positive ion. This occurs, generally, for light elements whose Auger lifetimes (of the order of 10^{-14} s for the $2p$ hole) are similar to the time which the atom needs to escape from the lattice.

Let us consider now the atoms which are sputtered in close packed directions by a mechanism of focussed collisions; the energy which is transferred along a focussed sequence of collisions is much lower than the energy which would be necessary for producing an empty inner level and the ejected atoms (the end of the chains) will be devoid of such empty levels and unable therefore to give rise to ions by Auger de-excitation.

Such a mechanism is supported by a series of experiments carried out by Hennequin ([13,14]) three years ago; the purpose of those experiments was to identify, in the general background of secondary electrons emitted from a sample when bombarded with primary ions, the presence of electrons produced by Auger de-excitations. In the experimental arrangement used by

Hennequin, the secondary electrons were accelerated and their energy spectrum was recorded by a small spectrometer which consisted of an electromagnet and an electron multiplier. The spectra represented on Fig. 32 show, in the case of aluminium, a well-pronounced Auger peak (de-excitation of a $2p$ hole) whose energy and width are in perfect agreement with the band structure of aluminium, and quite independent of the nature and energy of the primary ions used for the bombardment. Similar Auger peaks were obtained with samples of magnesium, silicon and beryllium.

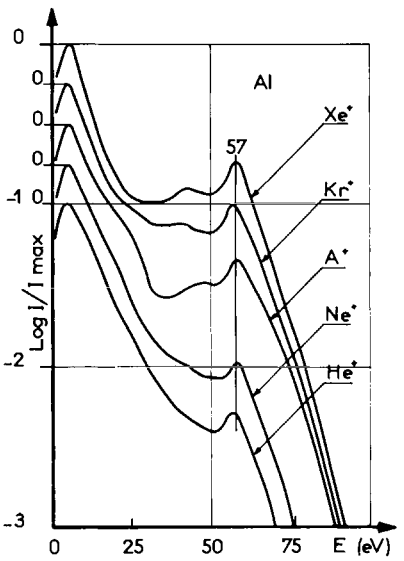


Fig. 32. – Auger peak in the secondary electron emission from Al (Hennequin). (Courtesy of *Compt. Rend. Acad. Sci. Paris.*)

On the other hand, no Auger peak is visible in the case of transition elements. Furthermore, it should be noted that the Auger peaks disappear progressively when the oxygen partial pressure in the vicinity of the sample is increased.

It is thus clear that, at least in the case of light elements, ionization of inner shells is produced in the ion cores, and that Auger de-excitation follows; but it has not been possible till now to distinguish, among the Auger electrons, those which have been produced by an Auger de-excitation in vacuo, leading to a secondary ion, and those which were produced by an Auger de-excitation

inside the metal, at the immediate neighbourhood of the free surface. We hope that new experiments using a spectrometer whose energy resolution is much better will make it possible to identify the Auger electrons produced in vacuo: the corresponding peaks are sharper since the transitions occur between atomic levels, but the sharpness will be limited, due to the fact that the lifetime of the free atom, before the Auger process occurs, is very short.

This type of process, involving the initial production of an empty inner level (2*p* for the light elements) seems to be the only one leading to multiply charged ions such as Al^{++} or Al^{+++}; but other processes, involving the excitation of outer electrons only, may occur in the production of singly charged ions. If the ionization energy of the atom is low enough, those processes, which will be considered a little later in the case of the transition elements, may be the most effective for producing singly charged ions; that explains why aluminium, whose ionization energy is especially low, leads to a secondary emission of Al^{+} ions much larger (about 10 times) than the Mg^{+} and Si^{+} emissions of the neighbouring elements, whereas the emissions of the doubly charged ions Al^{++}, Mg^{++} and Si^{++} are nearly the same.

Quite interesting results were obtained recently, in this respect, by Slodzian and Brochard (*), in the course of a study of the emission of singly charged and multiply charged aluminium ions from pure aluminium and copper-aluminium samples.

First of all, the ratio of Al^{++} and Al^{+++} emissions was found to be equal to 200, whatever the aluminium content of the sample and the energy of the primary ions. This is due to the fact that this ratio is controlled by the probabilities of the two types of de-excitation of an aluminium atom leaving the sample with a 2*p* level. The de-excitation occurs in vacuo; it leads partly to doubly ionized, partly to trebly ionized ions, the relative amounts being independent of the original metal.

Furthermore, the ratio Al^{+}/Al^{++} between singly and doubly ionized ions, when measured on Cu-Al alloys at various concentrations (namely 1.4%, 4.8% and 8.3% in mass) is found to be proportional (with a very good accuracy) to the ratio of the atomic concentrations of Cu and Al in the alloy! A tentative explanation of this curious effect is as follows: Al^{++} ions can be produced by the de-excitation of an inner shell only, such an excited level cannot be produced in a Cu-Al collision inside the sample (for reasons of electronic structure and of mass difference); the only effective collisions are the Al-Al ones, which implies that Al^{++} emission is proportional to the square

(*) Private communication.

of the aluminium concentration. On the other hand, processes involving no $2p$ level, but excitation of outer electrons only, have been seen to be the most effective for producing singly charged aluminium ions; if it is assumed that such processes are produced mainly in Cu-Al interactions (for reasons of electronic structure) we find that the Al^+ emission is proportional to the product of the copper concentration with the aluminium concentration, which explains exactly the experimental results. Note that if both types of collisions were effective for the second process, the result would not be very different since the copper concentration is close to unity in the alloys which have been studied so far; but the first process (production of a $2p$ hole) is necessarily related to the Al-Al collisions only.

A theoretical and experimental study of the special case of the transition elements was undertaken two years ago, in our laboratory, by Blaise and Slodzian ([15]). The experiments of Hennequin had shown no Auger peak in the secondary electron spectrum emitted by such elements. That is due to the fact that the $3d$ electrons protect the lower levels from the collisions quite effectively. The initial process may be shown to be the excitation of a $3d$ electron; the corresponding Auger electrons have rather low energies (a few electron-volts) so that the peak is masked by the general secondary electron emission.

In the theory developed by Blaise, the secondary ion emission is related to the de-excitation in vacuo of atoms which have been ejected in a super-excited, self-ionizing state.

The calculation is based on the determination of all the self-ionizing states which can be populated by electrons from the conduction band.

The theory may be checked by assuming that the excitation of a $3d$ electron occurs in the same way for all the elements of the series, from titanium to copper. The ion yields would be proportional, under such conditions, to the probability of populating a self-ionizing state, when the particle leaves the free surface.

Calculated and measured values are seen to agree in the limit of theoretical and experimental uncertainties (about 30%). Furthermore, the case of dilute alloys may be treated in the same way. For example, the emission of Ni^+ ions in a Cu-Ni alloy where nickel concentration is less than 30% may be estimated from the band structure of the alloy, which is assumed to be the same as that of pure copper (rigid band model) and the electronic structure of the nickel atom. It is then compared with the emission of Ni^+ ions from a pure nickel target, where the calculation involves the same structure for the atom, but a different band structure for the metal (pure nickel).

If C_{Ni} is the atomic concentration of nickel, the emitted intensity of Ni^+ ions from the alloy will be denoted Ni^+ (all) and that from pure nickel, for the same primary bombardment, Ni^+(Ni).

Let us write

$$\frac{Ni^+ (all)}{Ni^+ (Ni)} = \varrho_{CuNi} C_{Ni} .$$

ϱ_{CuNi} is the coefficient of enhancement of the nickel emission by the copper matrix, which is calculable from the band structures of copper and nickel.

Calculations give the value 1.14 for this enhancement coefficient. Measurements made on a 30% Cu-Ni alloy lead to the value 1.16 for fast secondary ions and to the value 1.1 for slow secondary ions. Such an agreement is very good indeed; it shows that our understanding of the intricate physical processes involved in kinetic ion emission has reached the point where quantitative interpretation of secondary ion microanalysis may be established on a firm base.

3·7. Some applications of secondary ion microanalysis.

The applications may be classified roughly into two essential groups: those where a strict localisation of the analysis is required in the three dimensions of space, and those where advantage may be taken from the extreme resolving power in depth which can be obtained if one does away with the optimum localisation in the plane of the sample surface.

In the first group, we find all the identifications of precipitates, segregations or inclusions in metallurgical or mineralogical samples. Figure 33 shows three pictures that Slodzian obtained many years ago from a cast Al-Mg-Si specimen. A quick examination of the three distribution images of aluminium, magnesium and silicon shows the presence of pure silicon inclusions and Mg_2Si precipitates in the aluminium matrix. Fig. 34 is due to Rouberol *et al.*: it represents the distribution of aluminium in an eutectic mixture of aluminium and calcium; the resolving power (one micron) is better than that obtainable from electron probe microanalysis.

Ion emission microanalysis is especially valuable in the field of mineralogy, for the rock samples are made essentially from ionic compounds, the secondary ion emission of which is very strong and roughly proportional to the concentration of the corresponding element. The difficulties which could arise from electrically insulating samples are overcome quite satisfactorily by

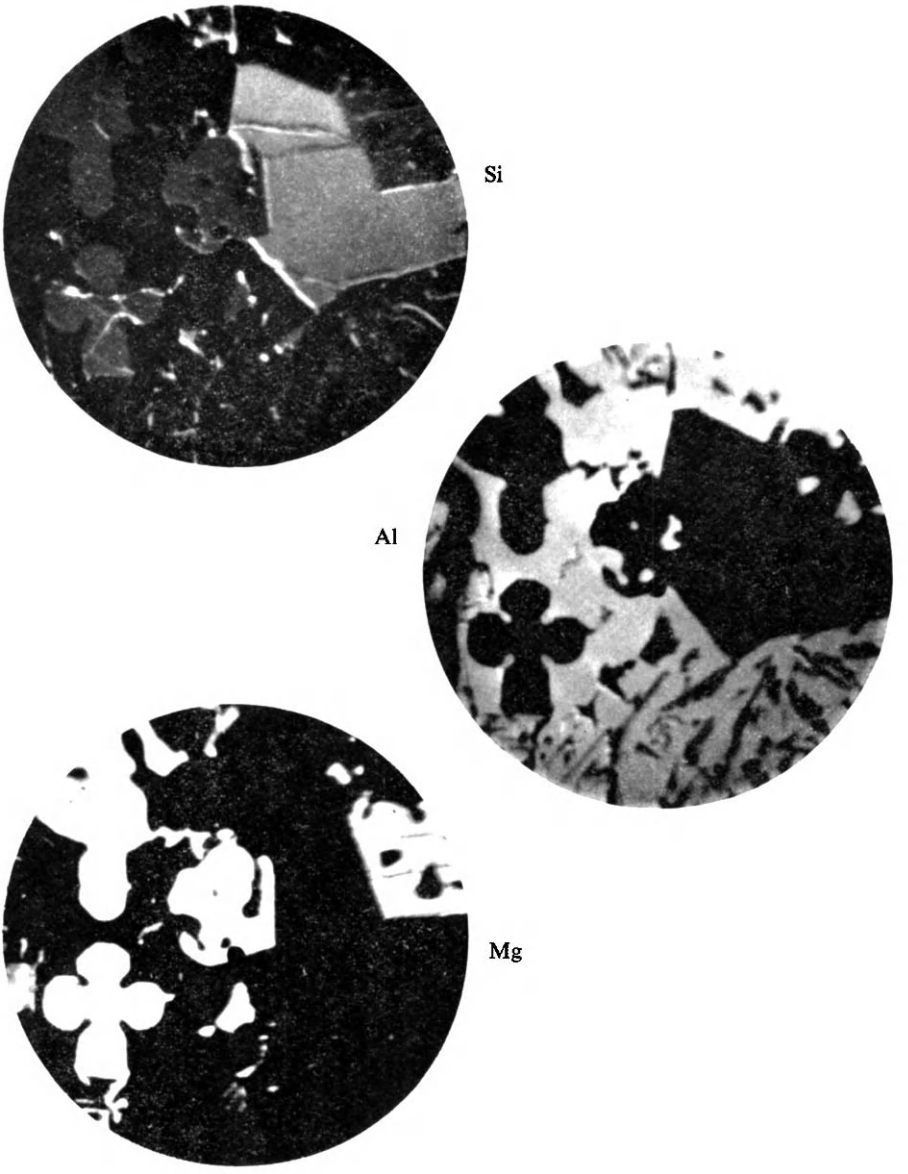


Fig. 33. – Al-Mg-Si specimen. Images $_{24}Mg^{+}$, $_{27}Al^{+}$ and $_{28}Si^{+}$ (Slodzian). Magn.: 300×
(Courtesy of *Journ. de Microscopie.*)

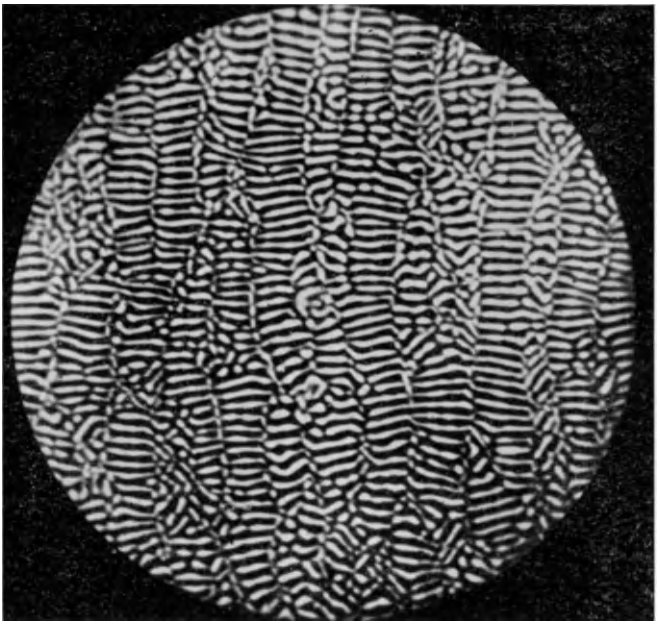


Fig. 34. – Al-Ca eutectic. Al^+ image (Rouberol). Magn.: 400×. (Courtesy of *Proceedings 5th Congress on X-Ray Optics and Microanalysis.*)

vacuum depositing onto the sample surface a metallic grid which ensures elimination of superficial charges. One of the first applications in this field is illustrated by Fig. 35, which shows the respective distributions of Li, Na, Al, Si, K and Ca in a specimen of granite containing inclusions of lepidolite (a variety of mica where sodium is partly replaced by lithium) embedded in a quartz matrix. For such semiquantitative estimations of the various components of a mineralogical sample, ion emission microanalysis is a most valuable tool, for it enables a rapid identification of the various elements or isotopes to be made; very low concentrations, of the order of the *ppm*, may be detected; on the other hand, the specimen may be moved as rapidly as in a conventional microscope for looking at the distribution of any element across an extended area.

The other group of applications covers all the situations where the chemical or isotopical constitution of the sample is varying essentially in one direction of space; such is the case in surface analysis (*e.g.* chemical reactions) and for diffusion studies. In this respect, secondary ion microanalysis is especially valuable for semiconductor analysis.

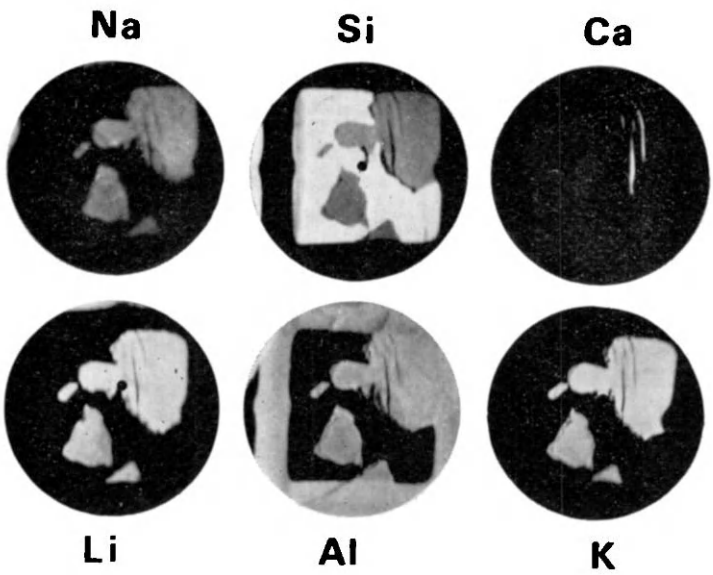


Fig. 35. – Granite specimen. Distribution images of Li, Na, Al, Si, K and Ca (Slodzian). (Courtesy of *Ann. de Phys.*)

As a matter of fact, single crystals are sputtered quite uniformly by the primary ion beam, so that deeper and deeper layers are progressively laid bare in the course of the operation; a simple recording of the ion emission from the doping element as a function of time is sufficient to give the diffusion profile. The etching speed may be calibrated at a later stage by interferometry; the emitted ion intensities are measured with the multiplier (Fig. 22) and known standards are used for calibration.

The uniformity of the etching speed, in the whole of the analysed area, may be made still better if the bombarding beam is scanned a little across the sample surface so as to ensure perfect homogeneity of the ion primary density. This technique was applied in our laboratory by Slodzian and Bernheim for measuring implantation profiles: the resolving power in depth may be estimated to some tenths of Ångström units and it could be made much better by using primary ions of low energy.

Isotopic analysis opens the way to self-diffusion studies. The diffusion of $_{18}O$ in uranium oxide has been examined recently by Contamin and Slodzian ([16]). Natural oxide was plated with a layer of oxide enriched in $_{18}O$. After the heat treatment, cuts were made by mechanical polishing at various depths. From the measurement of $_{16}O^-$ and $_{18}O^-$ emissions at the various

depths, a general diffusion profile was easily obtained; the results are shown on Fig. 36. The logarithm of the excess concentration in $_{18}O$ (over the natural concentration) is plotted as a function of the square of the depth. Furthermore, the time variation of the emitted ion intensity was recorded for each sample during the course of its sputtering by the ion beam; as a result, local

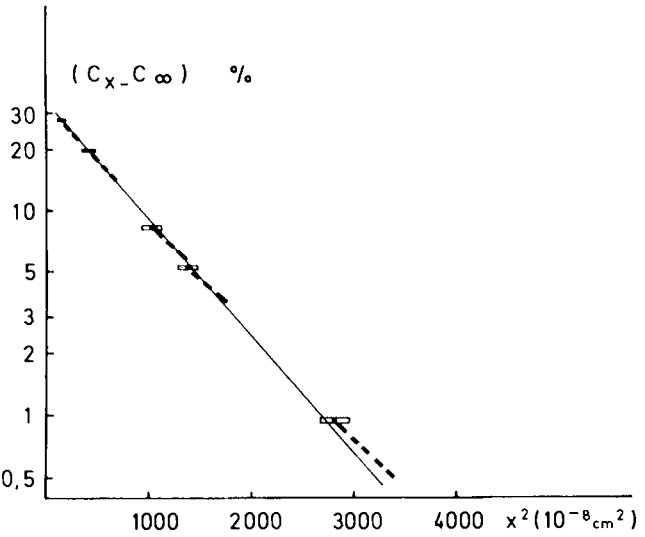


Fig. 36. – Diffusion of $_{18}O$ in uranium oxide (Contamin and Slodzian). (Courtesy of *Compt. Rend. Acad. Sci. Paris.*)

diffusion profiles were obtained (Fig. 36) which are seen to agree with the general profile satisfactorily, in spite of the fact that the etching speed is less uniform in such polycrystalline samples.

The few examples we have considered here were in the field of solid state physics; however there is no doubt that an enormous range of applications is open to this technique in other fields such as chemistry, mineralogy and biology.

REFERENCES (Section 3)

1) R. Castaing and G. Slodzian: *Proc. of the 2nd Europ. Reg. Conf. on Electron Microscopy, Delft 1960* (Amsterdam, 1960), vol. **1**, p. 169.
2) R. Castaing and G. Slodzian: *Compt. Rend. Acad. Sci. Paris*, **255**, 1893 (1962).

3) G. SLODZIAN: *Thesis*, Univ. of Paris (1963); *Ann. Phys.* **9**, 13 (1964).
4) R. CASTAING: *Thesis*, Univ. of Paris (1951), Publ. O.N.E.R.A. n° 55.
5) V. E. COSSLETT and P. DUNCUMB: *Nature*, **177**, 1172 (1956).
6) R. CASTAING: Section **1**.
7) D. GUÉNOT: Diplôme d'Etudes Supérieures, Paris (1966).
8) J.-F. HENNEQUIN: *Rev. Phys. Appl.*, **1**, 273 (1966); *Journ. Phys.*, **29**, 655, 957 (1968).
9) J. V. P. LONG: *Brit. Journ. Appl. Phys.*, **16**, 1277 (1965).
10) H. LIEBL: *Journ. Appl. Phys.*, **38**, 5277 (1967).
11) P. JOYES and R. CASTAING: *Compt. Rend. Acad. Sci. Paris*, B **263**, 384 (1966).
12) P. JOYES: *Journ. Phys.*, **29**, 774 (1968); *Thesis*, Univ. of Paris (1968).
13) J.-F. HENNEQUIN, P. JOYES and R. CASTAING: *Compt. Rend. Acad. Sci. Paris*, **265**, 312 (1967).
14) J.-F. HENNEQUIN: *Journ. Phys.*, **29**, 1053 (1968).
15) G. BLAISE and G. SLODZIAN: *Compt. Rend. Acad. Sci. Paris*, **266**, 1525 (1968).
16) P. CONTAMIN and G. SLODZIAN: *Compt. Rend. Acad. Sci. Paris*, **267**, 805 (1968).

High Intensity Electron Sources and Scanning Electron Microscopy (*)

A. V. CREWE

Department of Physics and Enrico Fermi Institute, University of Chicago - Chicago, U.S.A.

1. Field emission and an electron gun.

1·1. Introduction.

Field emission, as its name implies, is the process whereby electrons are emitted from a material by the influence of a large electric field. The process has been known for some time, and the basic equations were written down by Fowler and Nordheim in 1928 (1). The development of the technology and experimental understanding was begun by Müller in 1937 (2) and he has been the most outstanding contributor in the field since that time.

This process has found its principal use in the elucidation of the properties of metal surfaces and their interaction with other materials such as gases. Very few devices have been developed which use this effect. Apart from our own microscopes, the only other use appears to be as a source of high peak currents in pulsed electron beam generators.

For those not familiar with the field emission we will give a brief explanation. A much more complete discussion is given by Gomer (3).

A much simplified picture of a metal is that of a large number of unbound electrons which are free to move among the relatively stationary positive charge centers. The number of these electrons is very large, there being perhaps one free electron for every positive charge center.

Electrons are fermions so that there can be only two electrons in each

(*) Work reported here was supported by the U.S. Atomic Energy Commission.

defined level. This means that all levels are filled successively and the highest level (the Fermi level) may be several volts above the lowest level. Such an energy corresponds to a temperature of 10^4 °K. This remains true almost independently of the actual temperature of the metal. A metal can be considered to be a material where the available band for electrons is only partially full.

Figure 1*a*) illustrates the various parameters that are used. The diagram

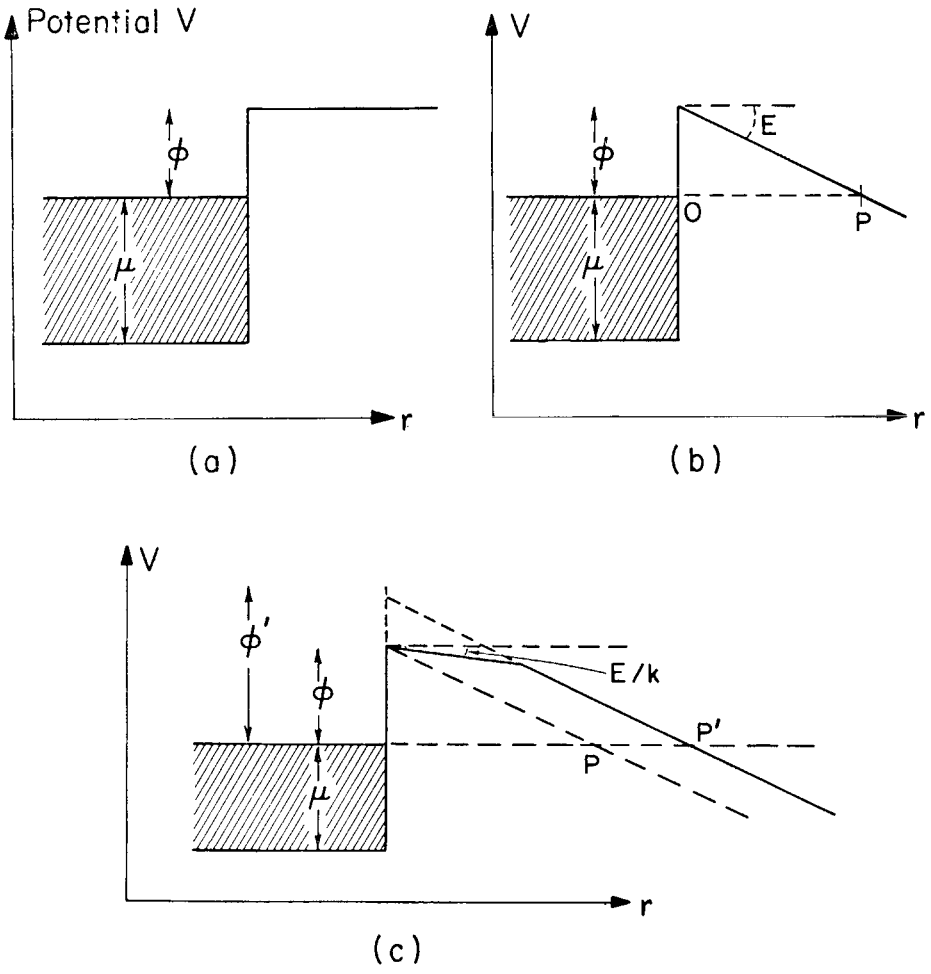


Fig. 1. – Illustration of the variation of the potential as a function of distance from the surface of a metal: *a*) with no applied field; *b*) with an applied field E; *c*) the metal is covered with a layer of gas and the field E is applied.

is a representation of the electrical potentials which exist close to the surface of a metal. Inside the metal, electrons fill the available bands up to a level μ. There remains a gap φ to reach the vacuum level. φ is the work function. That is, if we can give an energy $> e\varphi$ to an electron it can escape the metal (thermionic- or photo-emission, for example).

Now consider Fig. 1*b*) which shows the same example except that an electric field E is applied to the metal. This causes the potential to fall linearly in the region outside the metal. Under these conditions it is possible for a completely new process to occur. It can be seen that an electron does not require any energy at all to escape from the metal because it can « tunnel » through the potential barrier along the line *OP*.

The phenomenon of tunneling was first described by Gamow ([4]) who used it successfully to explain α-emission from radioactive nuclei. Briefly, the probability for tunneling to take place can be calculated using the Schrödinger equation and matching boundary conditions inside and outside the barrier. The probability depends strongly upon the height φ of the barrier and upon the width φ/E (the distance *OP* in Fig. 1*b*)).

The calculation was performed by Fowler and Nordheim who obtained a value for the emission current density i

$$i = 6.2 \cdot 10^6 \cdot \frac{(\mu/\varphi)^{\frac{1}{2}}}{(\mu + \varphi)} \cdot E^2 \exp\left[-6.8 \cdot 10^7 \frac{\varphi^{\frac{3}{2}}}{E}\right] . \tag{1}$$

We can consider the Fowler-Nordheim equation to be an accurate representation of an ideal condition, namely an atomically flat metal with no other materials (such as gas molecules) present.

The real world is considerably different from this, however, and we should now consider some perturbing effects.

Consider, for example, a monomolecular layer of gas molecules on the surface of the metal. Then these molecules will behave like a dielectric. Assuming the same potential on the metal as before and maintaining the same field E as before, we see that the field in the dielectric is reduced to E/k and so the point P moves to P' and the tunneling diagram is similar to the previous one except that the work function has apparently increased from φ to φ'. We can estimate this change because

$$\varphi' - \varphi \sim d \cdot E ,$$

where d is the thickness of the monolayer.

Assuming a rare-gas layer, $d \sim 2$ Å. Field emission usually occurs with

values of E of about 0.5 V/Å and therefore

$$\varphi' - \varphi \sim 1 \text{ V} \sim \varphi/4 .$$

This is a small change, but it can cause a profound reduction of emission current in view of the exponential dependence of i on φ.

We can also see that the width of the barrier φ/E is of the order of a few Ångströms. φ is usually about 4 V and E is about 0.5 V/Å. Consequently the barrier width is only 8 Å. In view of this, we can appreciate that the diagrams in Fig. 1 are very much idealized. The actual edge of the metal will be of the order of 1 Å thick and will vary considerably from point to point and should depend upon the crystal orientation of the surface atoms. We would therefore not be surprised if the current density i depends upon the crystal orientation.

1·2. Field emission as an electron source.

One cannot conveniently achieve electric fields of the order of 1 V/Å using a flat surface. It is possible to do so, however, if the surface is small and hemispherical. Under these conditions

$$E = \frac{V}{r}$$

and if $V \sim 10^3$ V and $r \sim 10^3$ Å, one can achieve $E \sim 1$ V/Å.

Under normal circumstances one does not have an isolated sphere but rather a hemispherical boss on the end of a cylindrical shank. Under these conditions one has

$$E = \frac{V}{kr},$$

where k is a constant and $k \sim 5$. Higher voltages are therefore required.

The current density of a high field emission tip is a complicated function of angle off the axis. Dyke and Trolan [5] have used an average technique to define the current density in terms of the emission current (1). Using this approach, they have measured continuous current densities up to 10^6 A/cm^2 compared to approximately 10 A/cm^2 for a hot filament tip [6]. This feature alone is interesting, but there is also the fact the apparent source

size is much smaller than the actual tip size. If we consider a tip with a hemispherical end and also assume that electrons will be emitted within a finite voltage range 0 to V_T, then one can show that the apparent source radius r is approximately ([7])

$$r = R(V_T/V_1)^{\frac{1}{2}},$$

where R is the radius of the tip and V_1 is the potential applied to the tip. This approximation is good for a spherical source only. The effect of the shank will be to change this value, but not by an order of magnitude. Reasonable values to insert in this equation are $R = 500$ Å, $V_T = 0.5$ V, and $V_1 = 1$ kV, which lead to $r = 11$ Å.

Most high field emission work has been done using tungsten as the tip material. The reasons for this selection are its suitable properties, such as a high melting point, a low vapor pressure, relatively high electrical and thermal conductivity, and high mechanical strength.

We fabricated our tips using the techniques outlined by Dyke *et al.* ([8]). A piece of 0.125 mm diameter tungsten wire (1÷3) mm long is spot-welded onto a preformed 0.2 mm diameter tungsten filament which is hairpin shaped. The assembly is electropolished, and the tip is etched by immersing it in a one normal sodium hydroxide (NaOH) solution and by applying 12 V dc between the tip and a remote electrode in the solution (Fig. 2).

The etched tip is mounted in an enclosure which is evacuated to about 10^{-9} torr. The tip is « formed » (rounded off) by sending a brief pulse of current through the filament. As the magnitude of the current pulse is increased, the filament begins to glow red. By this time the heat has usually driven off the contaminants so that the pressure no longer rises during the flash. The tip is then tested to see whether cold field emission is occurring. After emission is detected, a check is made to determine if the tip is properly formed by comparing an experimental voltage-current curve with that predicted by the Fowler-Nordheim equation. A typical Fowler-Nordheim plot for a tungsten tip having the plane with Miller indices (310) perpendicular to the axis is shown in Section 1·4. This crystal orientation was selected because it produces intense emission along the axis ([9]).

The subsequent performance of the tip appears to be more dependent on the local gas pressure than on any other parameter. In general, the emission current at constant voltage appears to be a curve similar to that of Fig. 3 ([10]). At first there is a small decline in emission current as the surface of the tip becomes coated with contaminants which increase the work func-

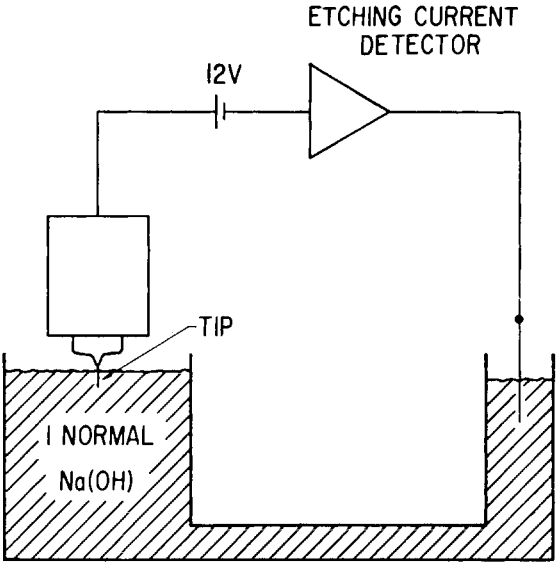


Fig. 2. – Tip etching arrangement. (Courtesy of *Rev. Sci. Instr.*)

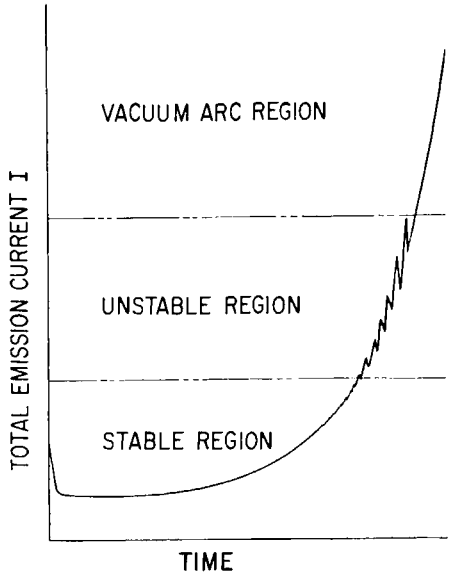


Fig. 3. – Typical dependence of emission current with time. As the tip becomes coated with contaminants, the emission first of all drops, and then begins a steady rise until the emission becomes erratic and the tip will eventually destroy itself with a vacuum arc. (Courtesy of *Rev. Sci. Instr.*)

tion. Thereafter the current rises until it becomes erratic, and the tip eventually destroys itself by a vacuum arc ([11-13]). The time scale for this process can vary from seconds to thousands of hours depending on the pressure ([10]).

When the emission current becomes erratic, the performance of the tip can easily be restored to its original condition by providing a pulse of current through the filament to evaporate the contaminants.

1·3. The electron gun.

The type of electron gun used in conventional microscopes has been developed to satisfy the needs of that instrument. Rather than attempting to modify such a gun for use with a field emission source, we decided to design a gun for this application alone.

We require a potential of a few kilovolts to provide the emission current from the tip and we would like to operate the microscope using an electron beam of a few tens of kilovolts. We must therefore use a system with at least three electrodes.

The electrons from the field emission source will be attracted towards a first anode which is held at a few kilovolts. Some of these electrons must be allowed to pass through an aperture in this anode so that they can be accelerated towards a second anode which is held at a potential of a few tens of kilovolts. A fraction of these electrons must be allowed to pass through a second aperture to form the usable electron beam.

We therefore have a requirement for two anodes, each with an aperture. It is well known that apertures in anodes such as this can act as lenses. Such lenses have large aberrations which can easily degrade the quality of the electron beam. This lens effect is produced whenever the electric field strength on the two sides of the aperture is different.

We decided that it would be better to design an electron gun in such a way that these aperture lenses were as weak as possible. In this way, any lens effect would be produced by the accelerating field itself and this field can be shaped to reduce aberrations.

The way in which this was accomplished is shown in Fig. 4. The electric potential rises sharply in the neighborhood of the tip and levels off at a distance of several tip radii. Thereafter the electric field is almost zero. Assuming that an anode must be placed at position *A*, we now insist that the field remains zero through the aperture in this anode. The potential must then be allowed to rise towards the second anode (at position *B*), but imme-

diately before the aperture in this anode we must reduce the slope to zero so that the electric field on this side of the aperture will be zero. On the other side of the aperture in the second anode the electric field is automatically zero because this anode is kept at ground potential.

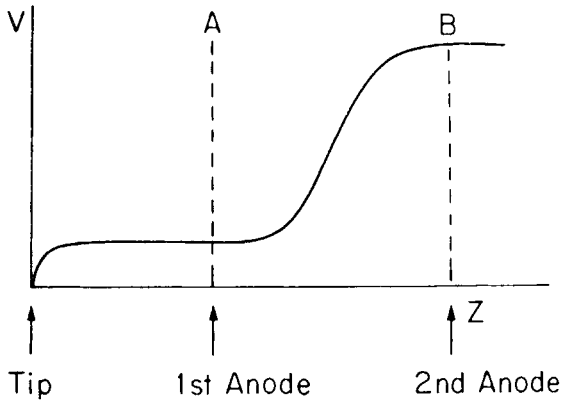


Fig. 4. – Schematic drawing of the required axial potential distribution for an electron gun. The principal requirement is that the slope of the curve be zero when passing through the anode apertures. (Courtesy of *Quart. Rev. Biophys.*)

It can be seen that we need some kind of *S*-shaped curve for the variation of potential between the two anodes. The particular shape of curve can be selected to reduce aberrations as much as possible and the required field shape can be produced by a suitable shaping of the electrodes.

The shape of the potential curve is such that there are two zeros in the slope. We therefore require at least a cubic term, and the simplest possible expression for the variation of potential with axial distance would be

$$V = az^2 + bz^3 .$$

We will take $z = 0$ at the first anode, $z = 1$ at the second anode, $V_{z=0} = 0$ and $V_{z=1} = 1$. We can therefore rewrite the expression as

$$V = az^2 + (1 - a)z^3 .$$

Differentiating, we obtain

$$\frac{\partial V}{\partial z} = 2az + 3(1 - a)z^2 .$$

$\partial V/\partial z$ is zero when $z=0$, and if we put $a=3$, then $\partial V/\partial z=0$ when $z=1$, which is the required condition. We now have

$$V=3z^2-2z^3\,.$$

V must also be a function of r, the radial distance from the beam axis. This dependence can be determined by inserting the above expression for V into Laplace's equation. We then obtain the complete expression for V.

$$V=3\left[z^2\left(1-\frac{2}{3}z\right)-\frac{r^2}{2}(1-2z)\right].$$

The shape of the electrodes can now be obtained by putting $V=0$ or 1. For $V=0$, we have

$$r^2=\frac{2}{3}z^2\cdot\left[\frac{3-2z}{1-2z}\right].$$

The electrode shape for $V=1$ is a mirror image of that for $V=0$ about the $z=\frac{1}{2}$ plane. Figure 5 shows the shape of the electrodes calculated from this expression.

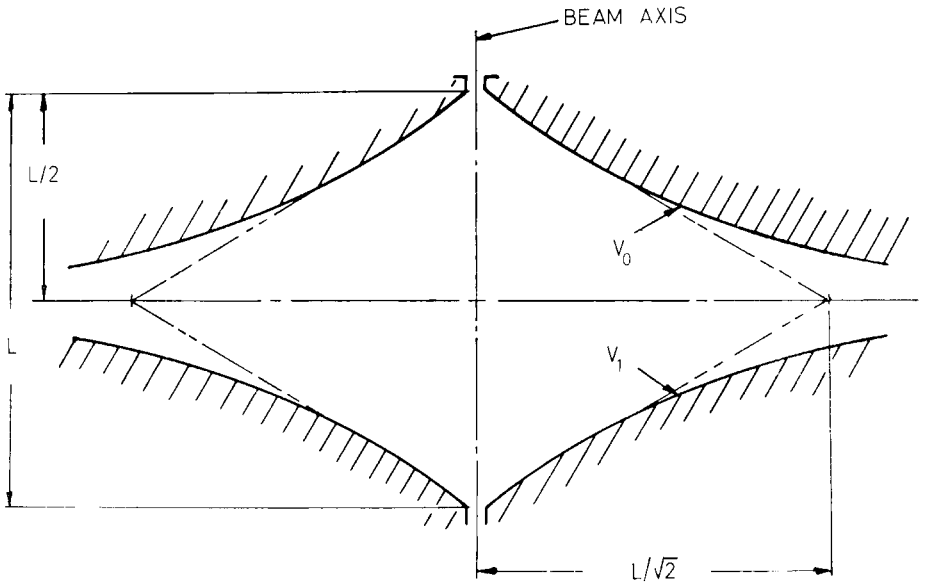


Fig. 5. – Scale drawing of the anode slopes which produce the required accelerating field. (Courtesy of *Quart. Rev. Biophys.*)

The particular potential distribution given above is not, of course, the only *S*-shaped curve which can be drawn to satisfy the prescribed condition. It is, however, the one with the greatest mathematical simplicity.

Butler investigated many such curves in an attempt to find the one which gives the smallest spherical aberrations when used as an accelerating system ([14]). He computed the first and third order optical properties of families of such curves, but concluded that the simplest solution given above was among the best. This is the basis of all the electron guns which we have used ([15-17]).

In spite of the simplicity of the electric field distribution, it is easier to compute the optical properties of the gun rather than attempt an analytical solution. In particular, both Butler and Thomson ([18]) have written programs which can compute all the essential optical properties.

1·4. Optical properties of the electron gun.

In order to use this electron gun in an electron microscope we need to know a few of the more important characteristics of the gun. In particular, we need

a) the position and magnification of the image of an electron source;

b) the chromatic aberration coefficient;

c) the spherical aberration coefficient.

Each of these quantities is a function of the position of the electron source and the voltages on the two electrodes.

This information is given in Fig. 6*a*), *b*) and *c*) for a 2 cm gun, that is, for a gun where the distance between the two anode apertures is 2 cm. Together with the characteristics of the field emission sources, such as effective source size and the energy spread of the electrons, these Figures give the information necessary to calculate all the relevant properties of the gun.

As an example of the characteristics of this gun, we can calculate the properties of the real image of the tip which can be produced at various distances from the second anode. In order to produce a good image we must include a small defining aperture. Experimentally we have concluded that the best location for this defining aperture is at the second anode. The reason for this choice is that small amounts of contamination on such an aperture tend to become electrically charged and introduce astigmatism into the beam. This effect is most pronounced when this charging process occurs in a region

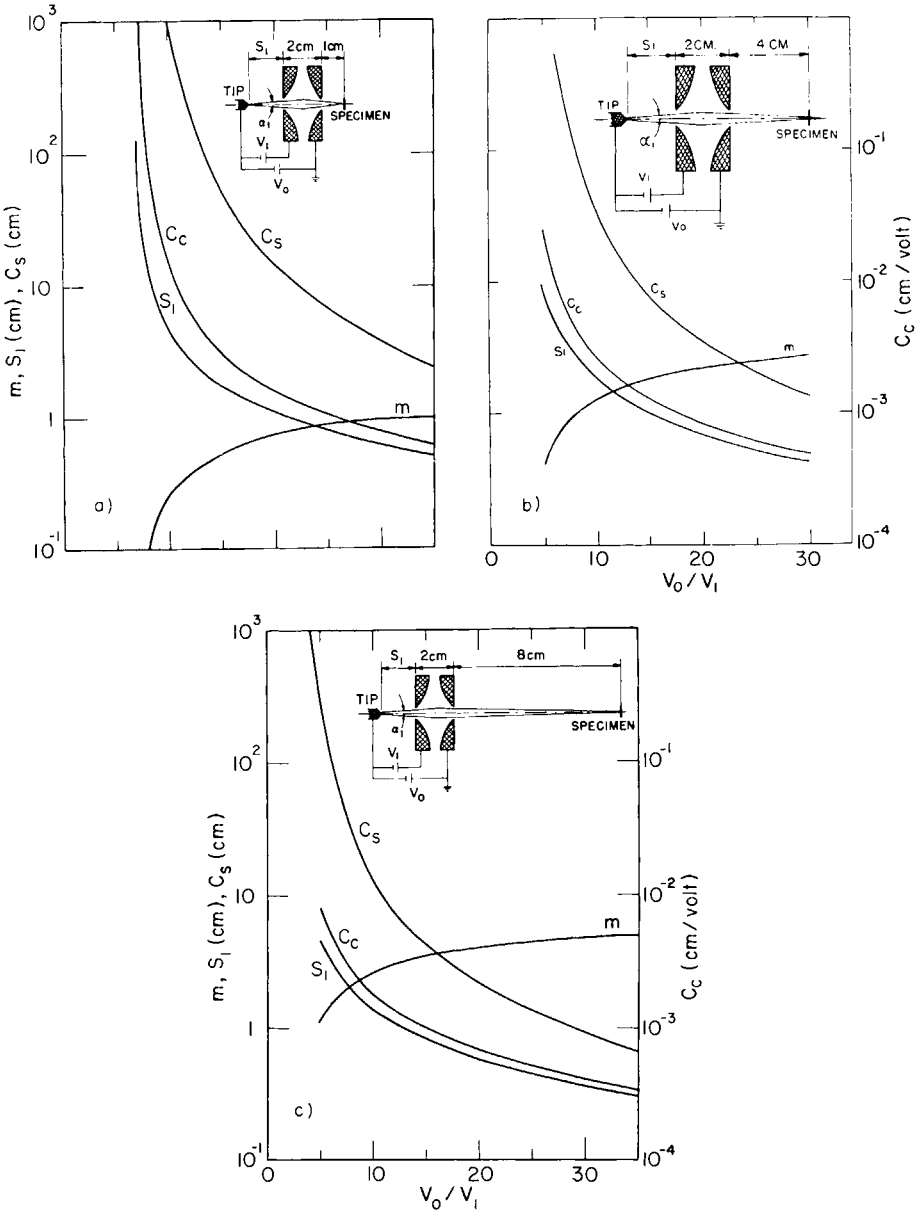


Fig. 6. – Calculated gun characteristics for three different positions of the image distance: *a*) image of the tip at 1 cm from the gun; *b*) image of the tip at 4 cm from the gun; *c*) image of the tip at 8 cm from the gun. m is the geometrical magnification, C_s the spherical aberration constant and C_c the chromatic aberration constant divided by the voltage V_1 of the first anode. (Courtesy of *Quart. Rev. Biophys.*)

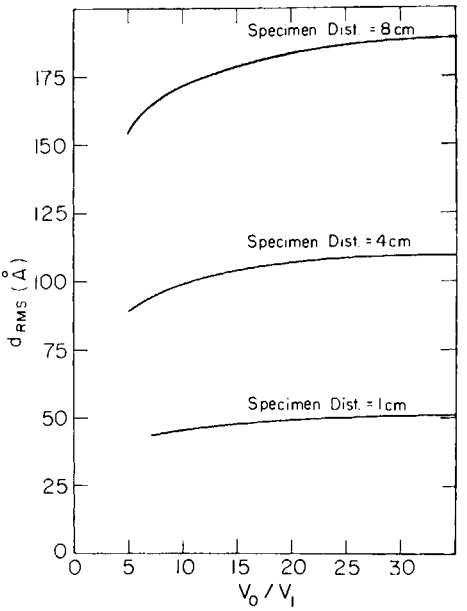


Fig. 7. – Graph of the dependence of the optimum probe diameter (d_{RMS}) upon the ratio V_0/V_1 for various fixed image distances. (Courtesy of *Quart. Rev. Biophys.*)

where the electron beam energy is the lowest and least pronounced when the beam energy is high and the electron beam current available to charge the contamination is low.

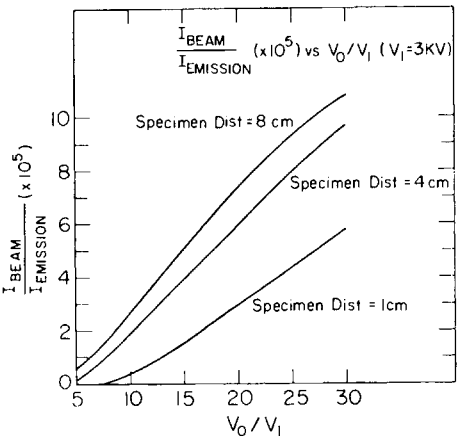


Fig. 8. – Beam intensity in the focused spot for the conditions given in Fig. 7. (Courtesy of *Quart. Rev. Biophys.*)

We therefore include an aperture at the second anode and can now calculate the following properties of a real image of the tip.

a) The beam current in the focused spot.
b) The Gaussian image size.
c) The contribution to image size from spherical aberration.
d) The contribution to image size from chromatic aberration.
e) The effects of diffraction at the aperture.

The various terms *b*)-*e*) can be combined quadratically to give an estimate of the final image size.

All these factors can be readily calculated from the information we have given previously and in Fig. 7 and 8 we present the results for image distances of 1, 4 and 8 cm.

It can be seen from these Figures that a very small image of the tip can be produced, and furthermore, the beam current available in this image is more than adequate for scanning microscopy. It is interesting to note that the spherical aberration of the gun is not a significant contributing factor to the spot size when the aperture is optimised for the smallest spot size. Chromatic aberration effects are much larger and probably cannot be reduced because such effects are inherent in any electrostatic focusing system.

REFERENCES (Section 1)

1) R. H. FOWLER and L. NORDHEIM: *Proc. Roy. Soc.*, A **119**, 173 (1928).
2) E. W. MÜLLER: *Zeits. Phys.*, **106**, 541 (1937).
3) R. GOMER: *Field Emission and Field Ionization*, Harward (1961).
4) R. GAMOW: *Zeits. Phys.*, **52**, 510 (1928).
5) W. P. DYKE and J. K. TROLAN: *Phys. Rev.*, **89**, 799 (1953).
6) V. E. COSSLETT and M. E. HAINE: *Proc. 3rd Int. Conf. on Electron Microscopy, London 1954* (Roy. Micr. Soc., London, 1956), p. 639.
7) M. DRECHSLER, V. E. COSSLETT and W. C. NIXON: *Proc. 4th Int. Conf. on Electron Microscopy, Berlin 1958* (Springer, Berlin, 1960), p. 13.
8) W. P. DYKE, J. K. TROLAN, W. W. DOLAN and G. BARNES: *Journ. Appl. Phys.*, **24**, 570 (1953).
9) 310 oriented tungsten wire may be obtained from Field Emission Corp., McMinnville, Ore., U.S.A.
10) E. E. MARTIN, J. K. TROLAN and W. P. DYKE: *Journ. Appl. Phys.*, **31**, 782 (1960).
11) W. P. DYKE, J. K. TROLAN, E. E. MARTIN and J. P. BARBOUR: *Phys. Rev.*, **91**, 1043 (1953).
12) W. W. DOLAN, W. P. DYKE and J. K. TROLAN: *Phys. Rev.*, **91**, 1054 (1953).

13) W. P. DYKE and W. W. DOLAN: *Advan. Electronics Electron Phys.*, **8**, 89 (1956).
14) J. W. BUTLER: *Proc. 6th Int. Conf. on Electron Microscopy, Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 191.
15) A. V. CREWE, E. N. EGGENBERGER, J. WALL and L. M. WELTER: *Rev. Sci. Instr.*, **39**, 576 (1968).
16) A. V. CREWE, D. JOHNSON and M. ISAACSON: *Proc. 26th Annual EMSA Meeting, New Orleans 1968*, p. 360.
17) A. V. CREWE, J. WALL and L. WELTER: *Journ. Appl. Phys.*, **39**, 5861 (1968).
18) M. G. R. THOMSON: private communication (1969).

2. Microscope design using field emission gun.

2·1. Simple scanning microscope.

It is clear from the previous Section that a field emission source together with an electron gun can produce a focused spot of electrons about 100 Å in diameter a few centimeter away from the electron gun. The addition of some other components can convert the electron gun into an electron microscope with a resolution of 100 Å. Such a microscope can be used either in transmission or can be used with secondary electrons.

2·1.1. *Description of microscope.* – A photograph of the microscope is shown in Fig. 9. The field emission electron gun has been described in detail in the previous lecture.

The focused spot produced by the gun is scanned across a specimen in a television type raster by means of an electrostatic deflection system ([1]). Eight Inconel plates are mounted on a Mycalex insulator, in such a way as to shield completely the insulator from the beam (see Fig. 10). The deflection voltages are placed on four of the plates, while two quadrupole fields, one rotated by 45° with respect to the other, are superimposed on the deflection fields using all eight plates. The two quadrupoles are excited through a sine-cosine potentiometer which results in a quadrupole field of arbitrary angle and magnitude for the correction of astigmatism ([2]).

Information about the specimen is obtained by detecting transmitted electrons using a solid-state detector. The photomultiplier signal is amplified and used to modulate the intensity of a synchronously scanned display tube to form an image of the specimen.

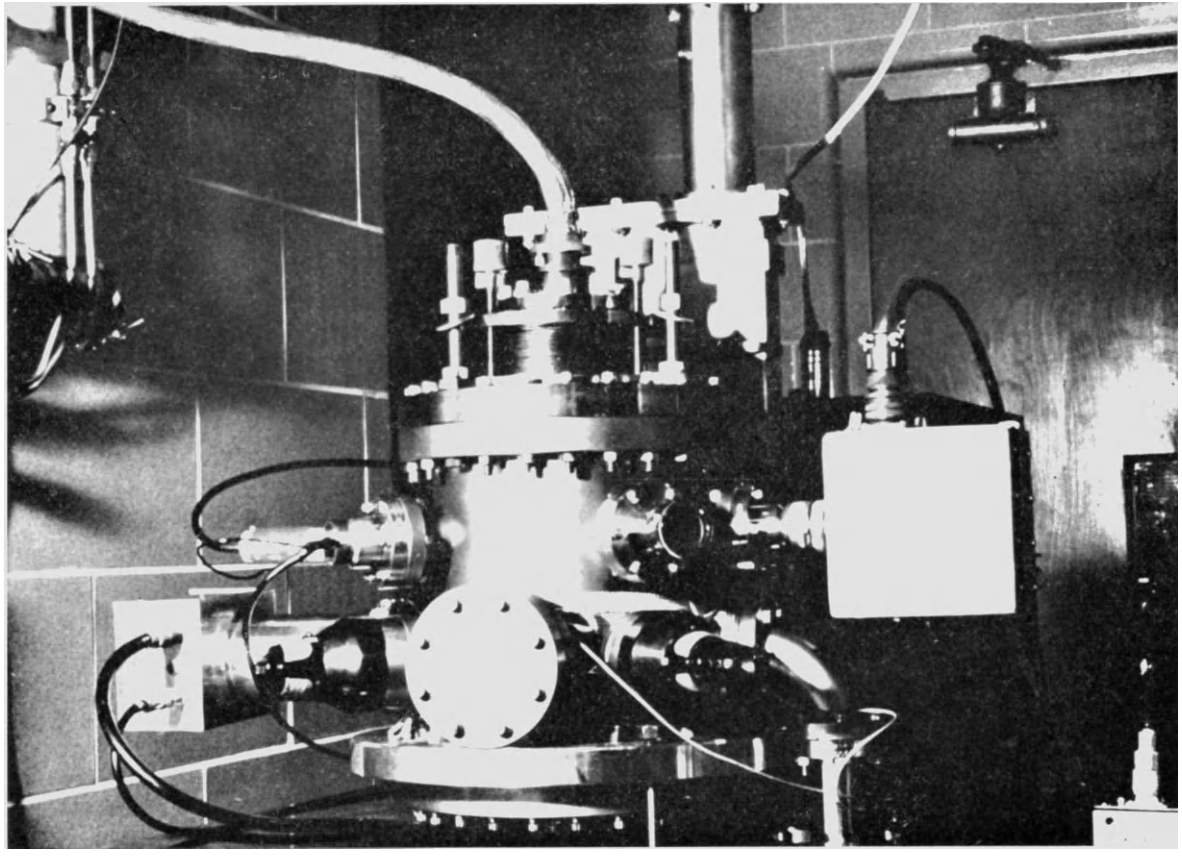


Fig. 9. – The gun microscope. The field emission tip sits approximately at the level of the top row of ports and the specimen sits slightly above the level of the bottom row of ports. Two micrometer motions on the lower ports are used to move the specimen. The photomultiplier-scintillator combination used to detect the transmitted electrons is below the bottom flange (not visible in this picture). It is similar to the device used for detecting secondaries which is seen coming out of a side port in the lower left part of the picture. (Courtesy of *Rev. Sci. Instr.*)

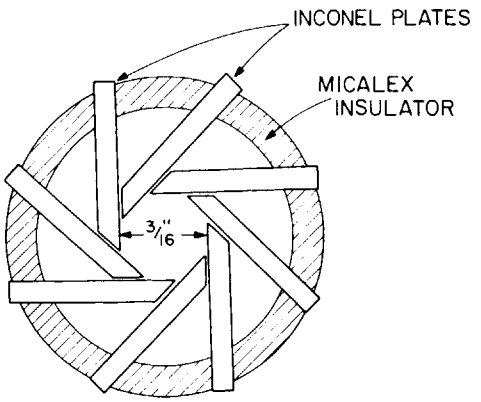


Fig. 10. – Diagram of the electrostatic stigmator and deflection system. The Inconel plates are mounted on a Mycalex insulator so that the insulator is completely shielded from the beam. The deflection voltages are placed on four of the plates, while two quadrupole fields, one rotated by 45° with respect to the other, are superimposed on the deflection fields using all eight plates. The two quadrupoles are excited through a sine-cosine potentiometer which results in a quadrupole field of arbitrary magnitude and angle for the correction of astigmatism. (Courtesy of *Rev. Sci. Instr.*)

The magnification is determined by the size of the display tube raster compared to the beam scan raster. The useful range varies from 400 to 400 000×.

To obtain stable field emission, the microscope chamber is kept at $\sim 10^{-9}$ torr by a 400 liter/s Varian Vacion pump. The accelerating voltage, the field emission voltage, and a current supply for periodic heating and cleaning the tip are provided by a stable (4 ppm/h) 30 kV supply ([3]).

2·1.2. *Operation of microscope.*

a) *Field emission tips.* We generally use (310) and (111) oriented tungsten wires for our field emission tips because they produce intense emission along the wire axis ([4]). All these tips are checked in an auxiliary tip testing system before they are put into the microscope. Using this auxiliary system we can determine the value of V_1 for a given emission current and whether or not the intense emission is centered on the axis of the tip holder ([5]).

Once in the microscope the tips are first cleaned by sending a brief pulse of current through the filament (« flashing ») so that the tip reaches about 1900 °K (white hot). After that, the tips are only periodically flashed at about 1000 °K (red hot). Because the ambient pressure is only about 10^{-9} torr,

this means that we run the microscope with the surface of the tip covered almost uniformly by a monolayer of adsorbed gases and the emission current gradually rises over a period of time until the current becomes erratic ([6]). When this happens the V_1 supply is turned off, the tip is « flashed » and the microscope is ready for operation again. The running period between flashes is usually of the order of 30 minutes to several hours depending upon the local pressure in the vicinity of the tip.

Using field emission tips in the above fashion we have experimentally measured 10^{-10} A of beam current in a 100 Å spot for 10 μA of tip current.

According to the calculations of Oatley *et al.* ([7]), a probe current of 10^{-10} A should be sufficient to record a 600 line picture in ~ 10 s using secondary electrons. Using transmitted electrons and assuming only 10% transmission, we expect that a 600 line picture taken in 10 s will provide 3% statistics per resolution point in the image.

b) *Alignment.* Alignment consists only of placing the field emission source on the electron optical axis of the gun, since the gun is prealigned ([1]). This is done by moving the source so that there is no image movement as V_0 is varied. The calculated and experimental aligment tolerance for a 100 Å probe size is ~ 25 μm. Movement of the tip is accomplished through a bellows on the top of the microscope.

2'1.3. *Description of secondary electron detector.* – A silicon surface barrier detector ([8]) was chosen for the secondary electron detection system. For this application, a semiconductor detector has definite advantages over the photomultiplier-scintillator combination generally used in conventional scanning microscopes ([9]). Ultra high vacuum problems are alleviated by eliminating unbakeable scintillators and the need for optical coupling. In addition, standard ultra high vacuum feedthroughs can be used. Finally, construction is considerably simplified because the use of a semiconductor detector obviates the need for careful polishing and optical coupling of a light pipe and scintillator ([9]). The detector was constructed with a lavite insulator and a minimum of epoxy to reduce vacuum problems. Also, the gold contact layer on the front surface of the detector, through which the electrons must pass in order to be detected, was made as thin as possible (< 200 Å). The detector was placed at the end of a beryllium-copper tube, the front of which is covered with a copper screen (see Fig. 11). The insulators holding the detector and tube are made of Mycalex.

A voltage of $\sim +200$ V is placed on the tube and screen while the entire detector is raised to a potential of $\sim +10$ kV. Secondary electrons pro-

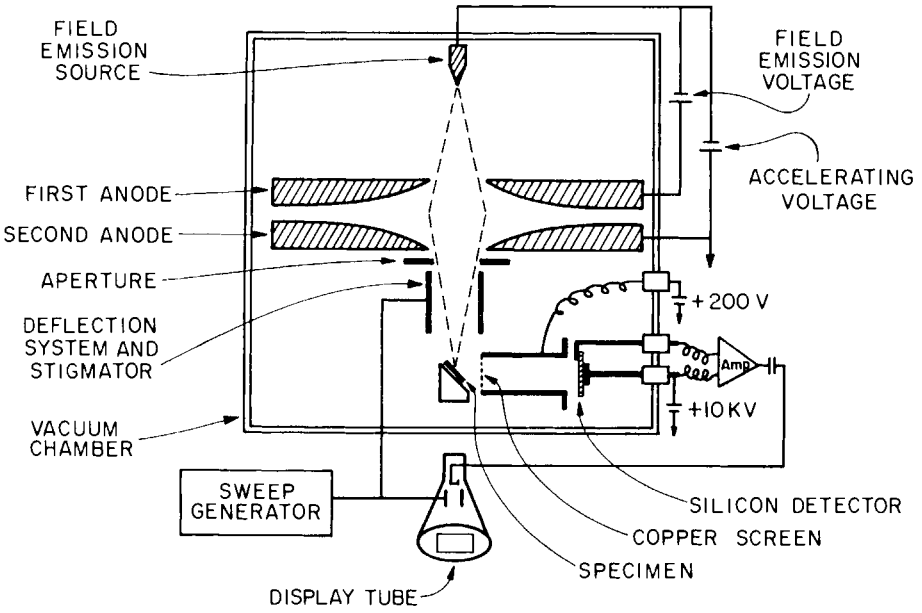


Fig. 11. – A schematic diagram of the microscope and the secondary electron detector. Between the anodes, the lens action of the electrostatic field focuses the electron beam and forms a real image of the field emission tip at the specimen level. The plane of the specimen is usually at an angle of 30° with respect to the incident beam. The secondary electron detection system is mounted at the specimen level with its axis perpendicular to the incident beam direction. (Courtesy of *Rev. Sci. Instr.*)

duced at the specimen are drawn through the screen and accelerated down the tube where they strike the detector.

The charge generated in the detector by the accelerated secondary electrons absorbed in the detector gives rise to a current which flows across a load resistor and the resulting voltage is amplified by a three-transistor amplifier ([10]).

The battery powered amplifier operates at the +10 kV of the detector with its output capacitively coupled to a video amplifier at ground. While it is possible to obtain a d.c. output ([11]), capacitive coupling was chosen mainly for its simplicity and because d.c. levels of the secondary electron signal were not required. Circuit components were chosen to give a bandpass of 20 Hz to 150 kHz which is more than sufficient to record a 500 line picture in 10 s. With an estimated detector gain of $2 \cdot 10^3$ (an 8 kV electron creates about $2 \cdot 10^3$ electron-hole pairs in silicon), a load resistor of 100 kΩ,

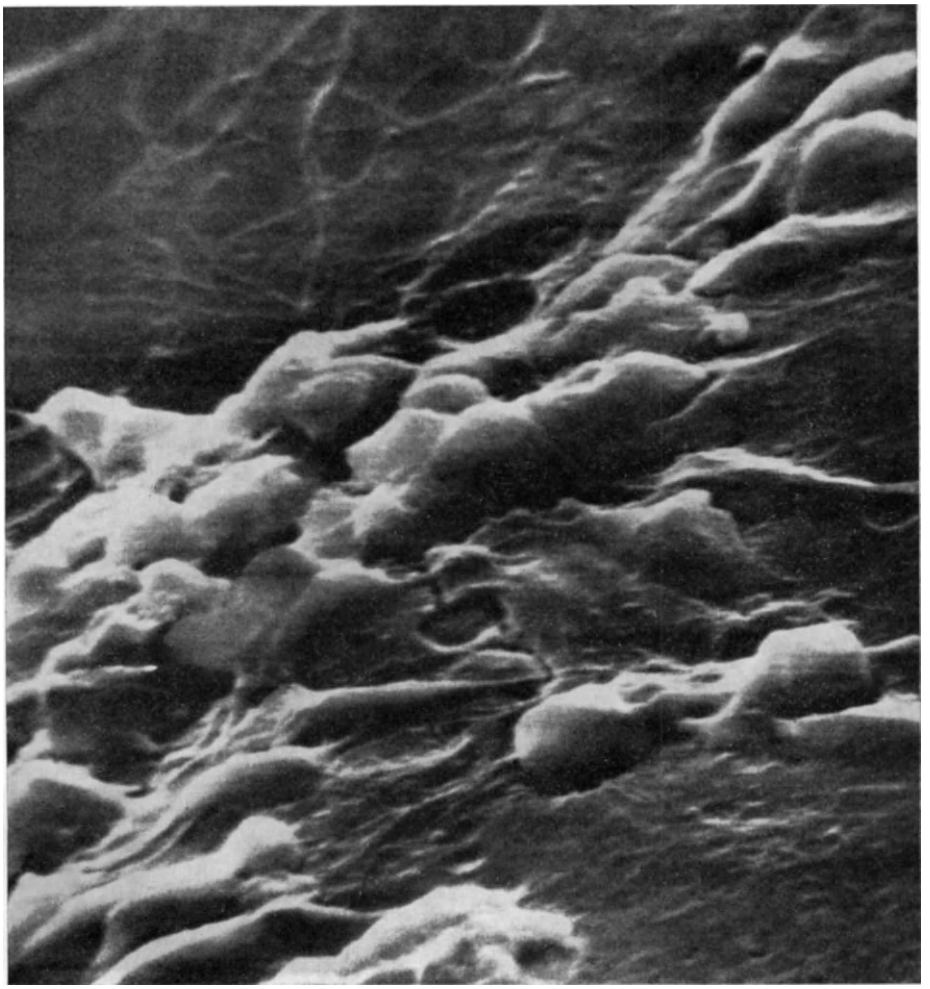


Fig. 12. – Micrograph of 10-day-old embryonic chick retinal cells. The cells were grown on small glass cover slips, fixed with gluteraldehyde and freeze-dried. The cover slips were coated with $\sim$ 100 Å of gold. Full horizontal scale is 30 μm. (Courtesy of *Rev. Sci. Instr.*)

and an amplifier gain of 500, the system gives a 10 V output signal for a detected secondary electron current of 10^{-10} A.

a) *Performance.* Resolution of (100÷200) Å has consistently been obtained on a variety of different specimens using this secondary electron detector, and is shown in the high magnification micrographs in Fig. 12

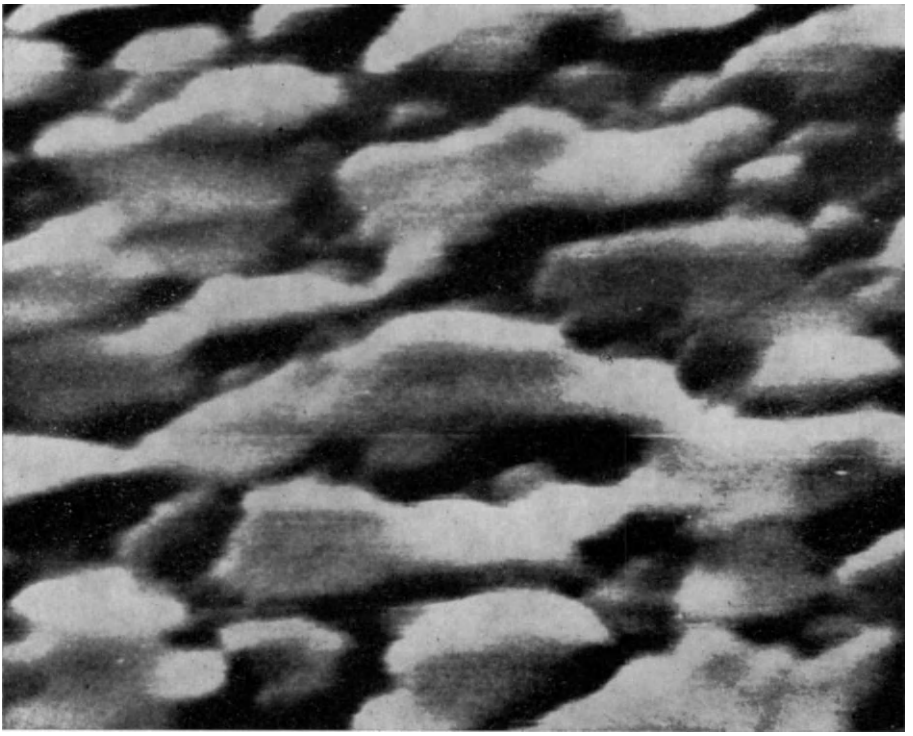


Fig. 13. – Micrograph of the surface of a thin (~ 500 Å) evaporated aluminum specimen showing the topographical structure of the crystal islands. Full horizontal scale is 1 μm. (Courtesy of *Rev. Sci. Instr.*)

through 15. All micrographs were taken with primary beam currents of (10^{-11}÷10^{-10}) A and scan times of (10÷100) s. The accelerating voltage which was used varied between 12 and 27 kV.

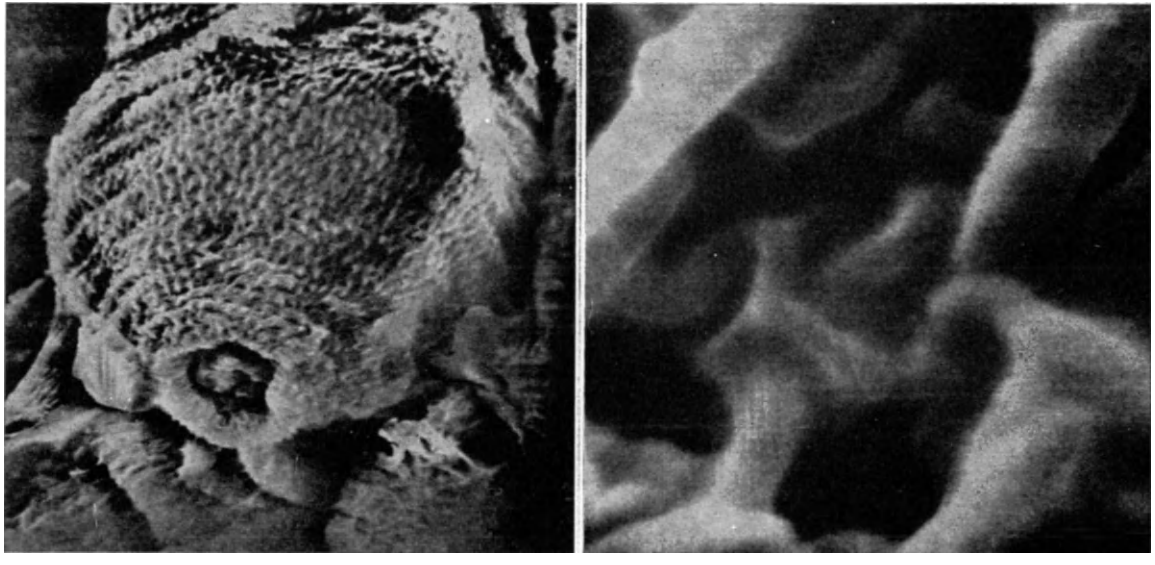


Fig. 14. – Micrographs of human blood flukes (the cercaria form of Schistosoma Mansoni). Left shows the body of one blood fluke where the tail has become detached revealing the tail plate. Full horizontal scale is 70 μm. Right is a higher magnification of left showing the spiny surface just above the tail plate region. Full horizontal scale is 1 μm. (Courtesy of *Quart. Rev. Biophys.*)

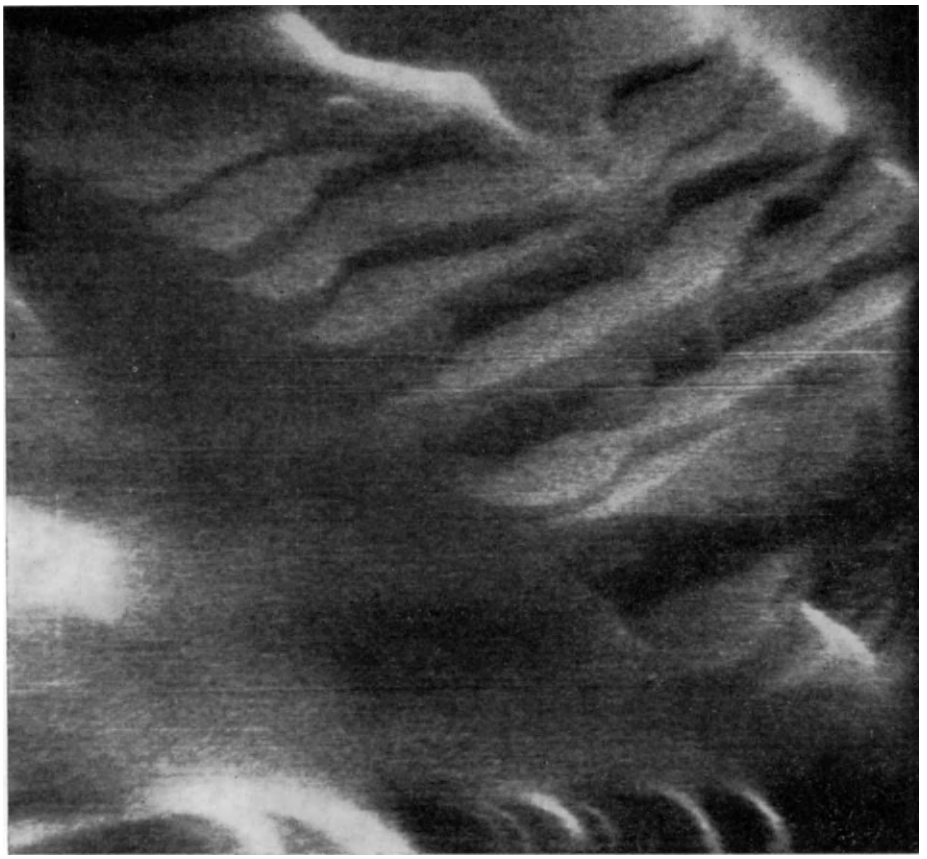


Fig. 15. – Micrograph of the surface of a piece of volcanic rock. Cleavage steps with heights less than 200 Å are discernible. Full horizontal scale is 3.0 μm. (Courtesy of *Rev. Sci. Instr.*)

2·2. **High resolution microscope.**

2·2.1. *Formation of the focused spot.* – To compete with conventional electron microscopes it would be necessary to obtain a focused spot of electrons a few Å in diameter. It does not appear possible to accomplish this in one stage using the electron gun alone. We must therefore consider a system which uses the electron gun followed by a lens of small focal length which will demagnify the image of the tip which is produced by the electron gun (Fig. 16).

If we consider such a system, we can make a list of all possible contributions to the size of the final focused spot. They are

a) The Gaussian image of the tip.

b) The effect of spherical aberration in the gun.

c) The effect of chromatic aberration in the gun.

d) The effect of spherical aberration in the lens.

e) The effect of chromatic aberration in the lens.

f) Diffraction.

The contributions *a*), *b*), *c*), are all subject to the demagnification of the magnetic lens, and we can see immediately that their total effect is small.

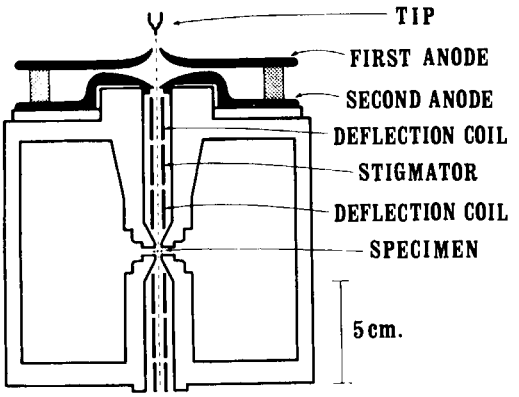


Fig. 16. – Schematic drawing of the high resolution scanning microscope. The gun is followed by a magnetic lens with 1.1 mm focal length. The specimen is placed in the center of the lens field.

Suppose the magnetic lens has a focal length of 1 mm and is 5 cm from the image of the source. Then all these effects will be demagnified by a factor of 50. Assuming a 100 Å image then this will become 2 Å. (In practice it will be smaller than this because the 100 Å estimate includes the effect of diffraction.)

Contribution *e*) will be small in a well-designed lens, particularly since the energy spread of electrons from a field emission source is small (about 1/4 V). We will neglect it here.

Contributions *d*) and *f*) are exactly the same elements which occur in the calculation of the resolving power of a conventional microscope. Combining

them leads to a resolving power

$$\delta = 0.7\, C_s^{\frac{1}{4}} \lambda^{\frac{3}{4}}.$$

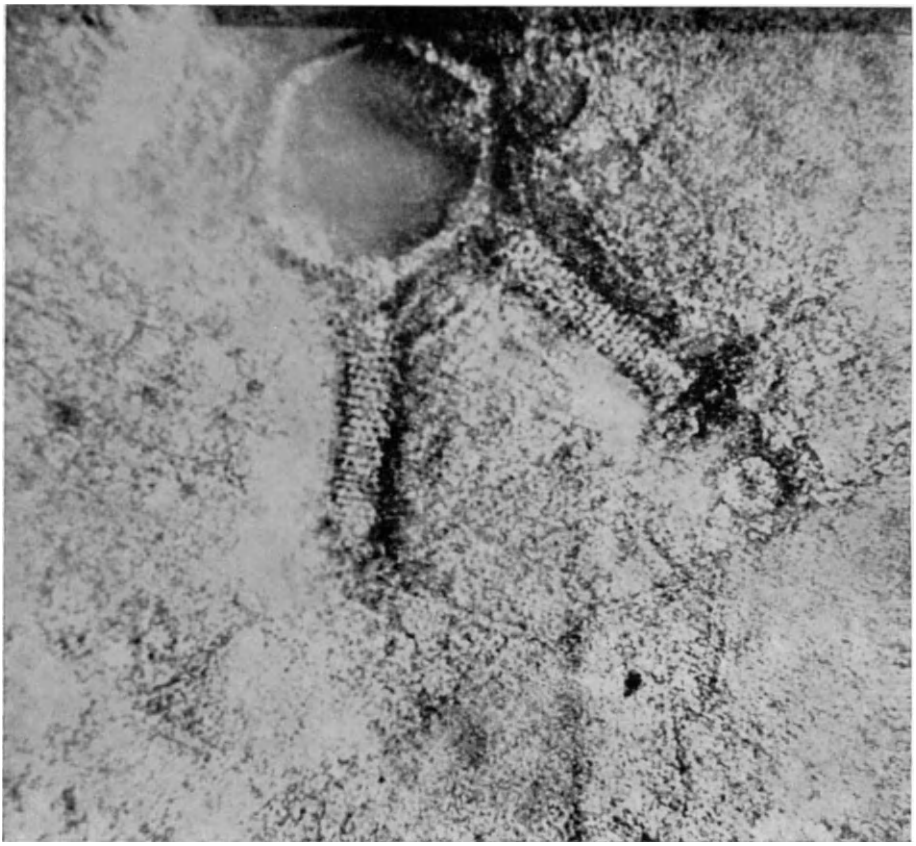


Fig. 17. – Bacteriophage T_4 negatively stained with Uranyl Acetate. 3000 Å full horizontal scale. (Courtesy of *Quart. Rev. Biophys.*)

In our particular system we designed a microscope lens with a focal length of 1 mm and $C_s \sim 0.3$ mm. This gives a minimum resolving power

$$\delta = 4.4 \text{ Å}.$$

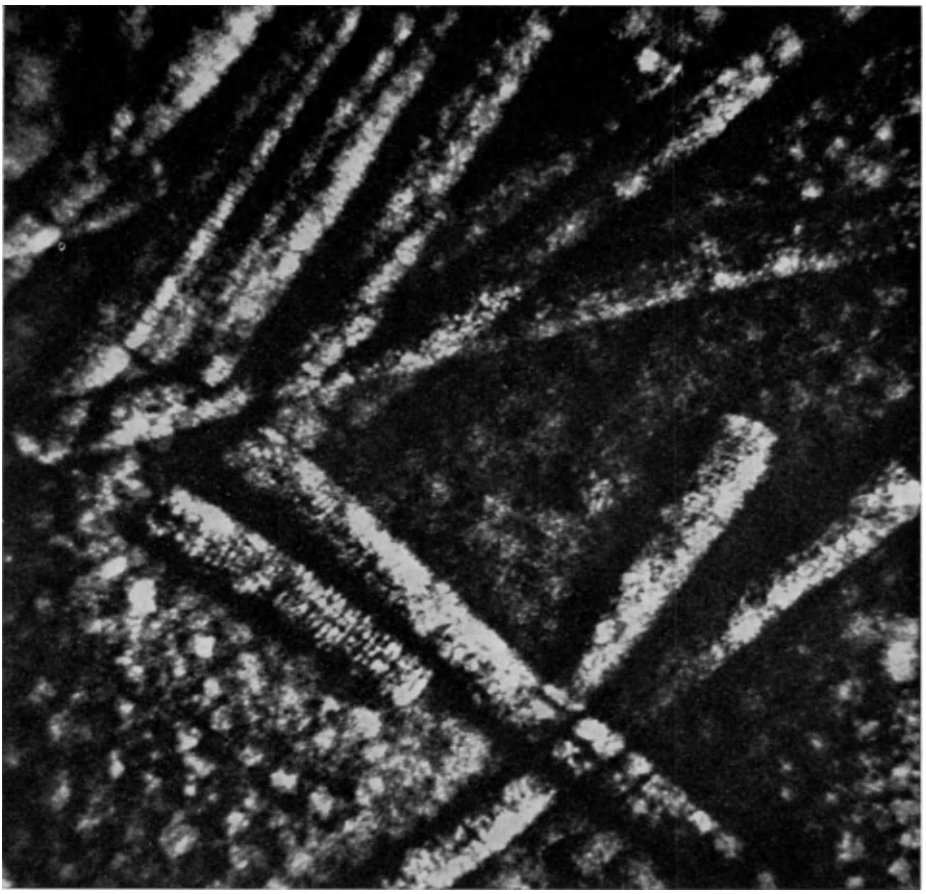


Fig. 18. – Tobacco mosaic virus negatively stained with Uranyl Acetate. Short segment is the stacked disc form with 20 Å spacing. Specimen was provided by A. Klug. 2200 Å full scale. (Courtesy of *Quart. Rev. Biophys.*)

It is perhaps not clear how the effects of the gun and the lens should be combined to allow us to estimate the final resolution. We assume here that they should be added in quadrature. In that case we would have

$$\delta = \sqrt{4.4^2 + 2^2} = 4.85 \text{ Å} .$$

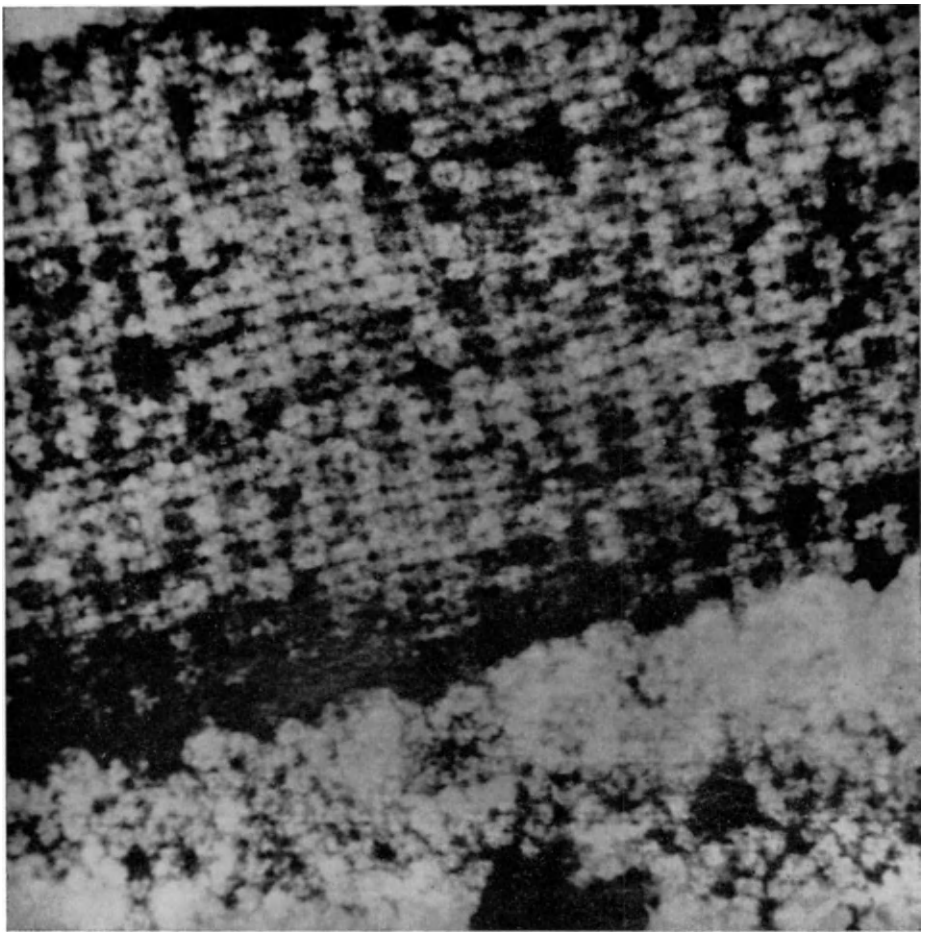


Fig. 19. – Catalase crystal negatively stained with Uranyl Acetate, showing 90 Å spacing. Specimen was provided by A. Klug. 3000 Å full scale. (Courtesy of *Quart. Rev. Biophys.*)

The effect of the gun then appears to be a small effect (10%). In that respect it is similar to the effect of lenses other than the objective in a conventional microscope.

The design of our microscope is shown in Fig. 16. The best resolution obtained to date is 5 Å, confirming the estimate for the resolving power given above.

Some representative micrographs are shown in Figs 17, 18, 19, and 20.

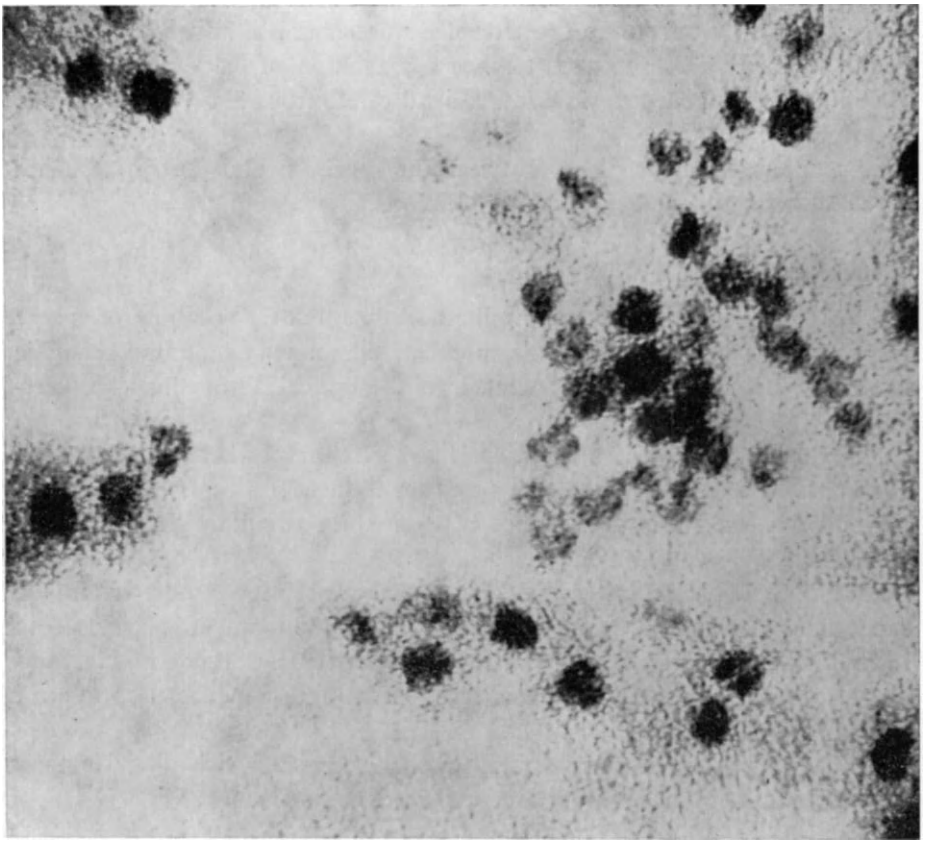


Fig. 20. – Ferritin, air dried on carbon film, showing (5 ÷ 10) Å structure in background. 1500 Å full scale.

2·2.2. *Detection system.* – In order to exploit the possibilities of a high resolution scanning microscope it is necessary to use the transmitted electrons rather than secondaries for the reasons given previously. However, the absence of optical elements below the specimen provides considerable latitude in the choice of the kind of electrons to detect. In particular it was thought that energy loss electrons might be capable of providing additional information about the specimen. Therefore a spectrometer was installed in the microscope below the specimen. In this way one can choose to display all transmitted electrons, only those electrons which have lost no energy, or electrons which have lost a specific amount of energy.

A spherical electrostatic spectrometer was chosen although other kinds could be used. This spectrometer has a resolution of 0.3 V at 25 kV and the slits can be opened up to allow a band of electrons 200 V wide into the detector.

An aperture can be placed between the specimen and the spectrometer to select a small angular range of electrons.

2'2.3. *Contrast.*

a) Energy loss. It has been well established that the spectrum of energy loss electrons which emerge from a specimen illuminated by a monochromatic source of electrons is directly related to the optical absorption characteristics of the specimen.

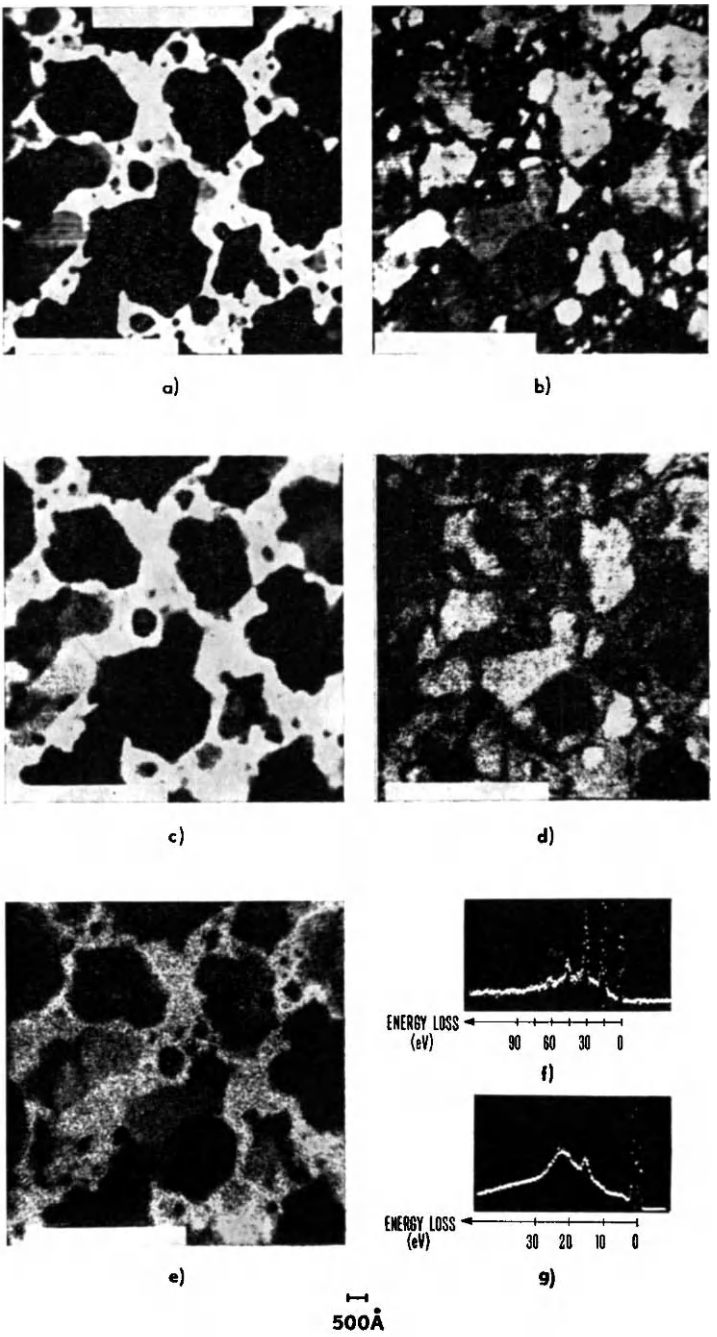
Specifically the optical absorption is given by Im (ε) where ε is the dielectric constant and the energy loss spectrum is given by Im ($1/\varepsilon$). There is, in general, no significant difference between these two unless the real part of the dielectric constant approaches zero.

Most materials exhibit significant optical absorption properties in the range from zero to 20 or 50 V and this information can often be used to identify the material. It is apparent, then, that the energy loss electrons can also convey such characteristic information and could be used at least as a partial identification method.

There is a wide range of energy loss phenomena which all have their counterparts in optical absorption, and we mention a few.

1) Characteristic X-ray lines: The incident beam of electrons can eject an electron from the *k*-shell (or other shells) producing an energy loss spectrum which consists of a sharp leading edge at the energy corresponding to the X-ray line and a long trailing edge as the electrons are ejected further into the continuum.

Fig. 21. – Thin evaporated aluminum specimen. *a)* is a micrograph taken with zero-energy loss electrons while *b)*, *c)*, *d)* and *e)* are micrographs taken with an energy loss of 12, 22, 30, and 35 V, respectively. Figure 21 *a)* corresponds to the kind of picture which would be obtained in a conventional electron microscope. *f)* and *g)* show energy-loss data taken while the electron probe was stationary. Figure 21 *f)* was taken on a black area of the specimen and the curve indicates substantially pure aluminum. The peaks are the plasma-loss peaks which occur at 15, 30, 45, and 60 V. Figure 21 *g)* was taken with the electron probe stationary on a white area and indicates that this is probably aluminum oxide. (Courtesy of *Journ. Appl. Phys.*)


a)
b)
c)
d)
e)
ENERGY LOSS (eV)
90 60 30 0
f)
ENERGY LOSS (eV)
30 20 10 0
g)
500Å

The intensity of such losses is small, but may be useful in specific applications. Using a spectrometer of high enough resolution it should be possible to detect shifts in the position of the edge corresponding to the chemical binding energy.

Such electrons (and all other energy loss phenomena) can either be used as a method of identification on selected areas or as a signal to be displayed on the screen.

2) Plasma losses: The most pronounced plasma losses occur in metals and appear when the real part of the dielectric constant goes through zero. The effect can be considered as the resonance excitation of the free electron gas. Multiple plasma losses occur and their intensities depend on the thickness of the specimen.

The most pronounced plasma losses are seen in aluminum where they occur at multiples of 15 V. This is shown in Fig. 21 where we also show some micrographs obtained at various values of energy loss.

Broad plasma losses occur even in non-conductors and are generally responsible for the gross appearance of an energy loss spectrum such as that seen in carbon.

3) Other effects: There is a large variety of energy loss phenomena which can be obtained from the literature on optical absorption; for example, exciton production in solids, particularly ionic crystals and the U.V. spectra of biological molecules. So far, very little work has appeared on the electron energy loss phenomena corresponding to these effects.

REFERENCES (Section 2)

1) A. V. Crewe, D. N. Eggenberger, J. Wall and L. M. Welter: *Rev. Sci. Instr.*, **39**, 576 (1968).
2) O. Rang: *Optik*, **5**, 518 (1949).
3) R. N. Lewis, E. A. Jung, L. M. Welter, L. S. van Loon and G. L. Chapman: *Rev. Sci. Instr.*, **39**, 1522 (1968).
4) (310) and (111) oriented tungsten wire may be obtained from Field Emission Corp., McMinnville, Ore., U.S.A.
5) A. V. Crewe and M. Isaacson: *Proc. of the 26th Annual EMSA Meeting, New Orleans 1968*, p. 359.
6) E. E. Martin, J. K. Trolan and W. P. Dyke: *Journ. Appl. Phys.*, **31**, 782 (1960).
7) C. W. Oatley, W. C. Nixon and R. F. W. Pease: *Adv. Electronics Electron Phys.*, **21**, 181 (1965).

8) G. DEARNALEY and D. C. NORTHROP: *Semiconductor Counters for Nuclear Radiations*, Wiley, New York (1963), Chap. 6.
9) T. E. EVERHART and R. F. W. THORNLEY: *Journ. Sci. Instr.*, **37**, 246 (1960).
10) T. C. PENN: *Electronics*, **41**, 58 (1968).
11) A. J. GONZALES: *I.E.E.E. 9th Ann. Symposium on Electron, Ion and Laser Beam Technology* (1967), p. 188.

3. Contrast mechanisms in a high resolution scanning microscope.

3·1. Mechanisms identical to the conventional microscope.

Figure 22 is a composite diagram of a conventional microscope (reading from left to right) and a scanning microscope (reading right to left).

It is clear from this diagram that the essential electron optics of the two types of microscope are identical except for the direction of motion of the electrons. This is an important concept because optical effects are independent of the direction of the waves.

For example, a detector on the axis of the scanning microscope would correspond to the illumination aperture of the conventional microscope.

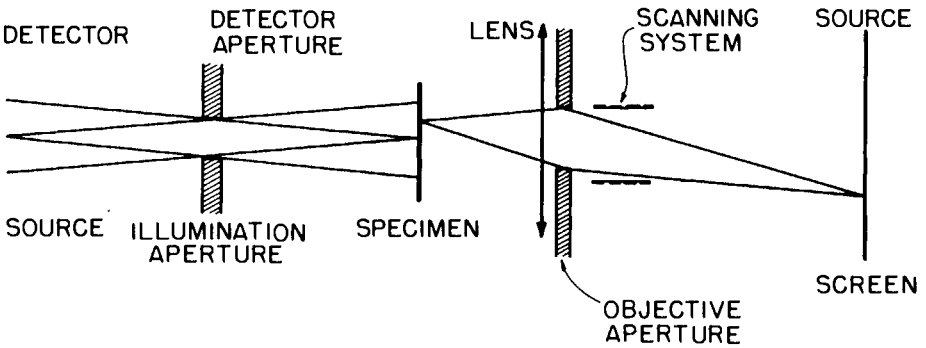


Fig. 22. – Schematic diagram of conventional and scanning electron microscopes. For the conventional microscope, read the diagram from left to right. We have an electron source providing illumination through an aperture onto the specimen. The specimen is then imaged by a lens through the objective aperture onto the screen. For the scanning microscope, we read the diagram from right to left. The electron source is imaged by a lens through an aperture onto the specimen. This image is scanned across the specimen by a scanning system. Electrons transmitted through the specimen pass through an aperture onto a detector. (Courtesy of *Quart. Rev. Biophys.*)

The use of a very small detector would correspond to a conventional microscope with a highly collimated illumination system.

It will be appreciated that these statements are qualitative ones which remain to be verified quantitatively. Indeed, we had not pursued this concept further until it became clear that some of the effects we were observing in the microscope were due to diffraction and interference effects which were exactly analogous to the effects observed in a conventional microscope. We will now examine this concept of the identity of the two types of microscope (except for the reversal of the beam direction) for the various types of conventional microscope images to determine whether or not they should be observed in a scanning microscope ([1,2]).

3'1.1. *Scattering contrast.* – We consider the specimen to consist of a number of electron scattering centers and also that the lens is focused exactly on the specimen.

Looking at the conventional microscope we see that contrast is obtained when scattered electrons are removed by the defining aperture.

If we consider the illuminating beam to be an axial plane wave, then the intensity recorded on the photographic plate will be

$$I = 2\pi \int_0^{\alpha_0} I(\alpha) \sin \alpha \, d\alpha ,$$

where $I(\alpha)$ is the angular distribution of the scattered electrons and α_0 is the half-angle of the defining aperture.

Taking the case of the scanning microscope we consider the analogous case of a very small detector. Then electrons from various positions of the illuminating cone can be scattered into the detector and the intensity recorded by the detector is

$$I = \left(\frac{\alpha_1}{\alpha_0}\right)^2 2\pi \int_0^{\alpha_0} I(\alpha) \sin \alpha \, \mathrm{d}\alpha .$$

The normalization factor $(\alpha_1/\alpha_0)^2$ is the ratio of the solid angle subtended by the detector (half-angle α_1) to the solid angle of the illumination.

The expressions for the intensities are identical except for the normalization factor. This same factor occurs when one considers the intensity in the image with no specimen in the microscope. Therefore the contrast in the two microscopes is identical.

3·1.2. *Interference and diffraction contrast.*

a) *Fresnel fringes.* In Fig. 23 we show schematically the formation of Fresnel fringes in a conventional microscope.

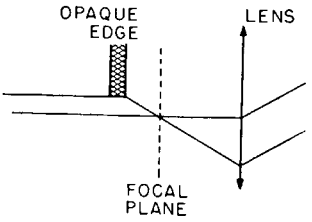


Fig. 23. – Illustration of the formation of Fresnel fringes. For a conventional microscope, read the diagram from left to right. Electrons scattered or diffracted by the opaque edge interfere with the primary beam giving an interference pattern which is placed in the focal plane of the lens so that an image appears on the screen. For a scanning microscope, read the diagram from right to left, where it is clear that if all the path differences are the same we will again observe fringes as we scan the electron beam across the focal plane. (Courtesy of *Quart. Rev. Biophys.*)

Ideally a plane wave is incident on the specimen which we take to be an opaque edge. The edge acts as a secondary source of radiation and interference takes place in the region close to the edge forming fringes. If the objective lens is focused onto a plane in this region the interference fringes can be observed. These fringes are well known to conventional microscopists and are used as a test of the instrumental resolving power and as a way to test for astigmatism. Experimentally one observes a dark fringe close to the edge when the lens is over-focused, no fringes when the lens is focused on the edge and a bright fringe when the lens is under-focused. The number of fringes which are observed depends on the degree of collimation of the illumination.

For the case of a scanning microscope we simply view the diagram from right to left instead of from left to right. Providing we duplicate the ray diagram exactly, those interference effects should be exactly the same because their existence depends only upon path differences in the various rays.

Such fringes can be observed in a scanning microscope and are shown in Fig. 24. They have exactly the same character as in a conventional microscope and appear identical in all respects.

In order to observe such fringes it is necessary to use a very small aperture

above the detector in order to duplicate the highly collimated illumination of a conventional microscope. One therefore sacrifices intensity and if one wanted to obtain a large number of fringes the intensity would be very small.

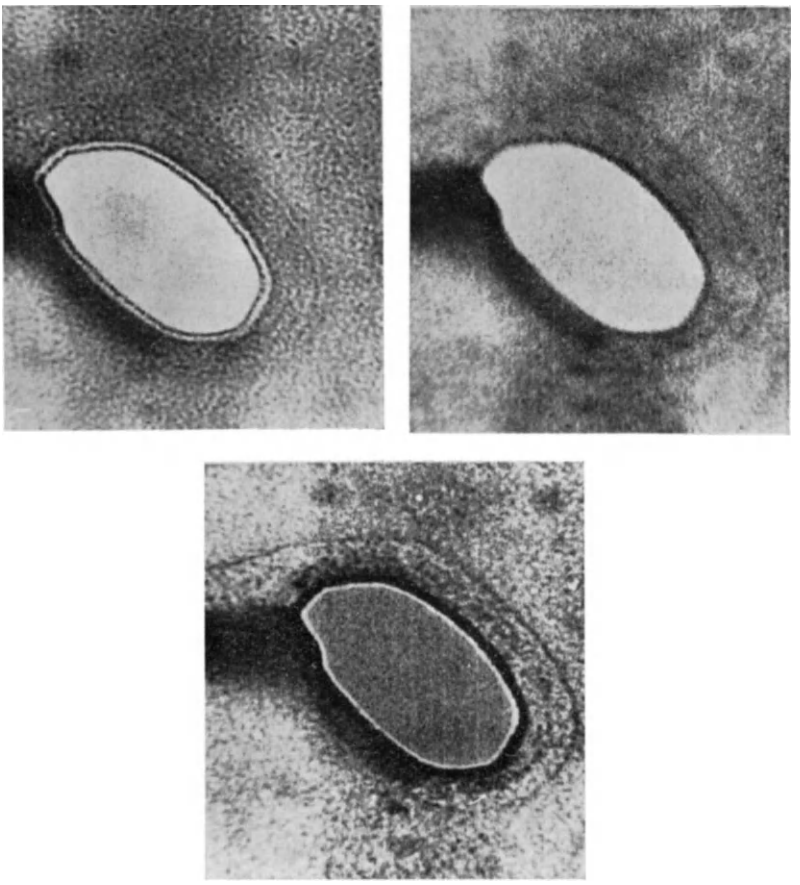


Fig. 24. – Fresnel fringes obtained in a scanning microscope. The fringes were observed around a hole in an aluminum film. Top, left: over-focus fringe. Top, right: in focus. Bottom: under-focus fringe. Scale: the micrographs show an area which is 3100 Å in the horizontal direction. (Courtesy of *Quart. Rev. Biophys.*)

b) *Phase contrast of lattice images.* Again we refer to the explanation for the observation of these effects in a conventional microscope. The ray diagram is given in Fig. 25.

A highly collimated beam of electrons is incident on the specimen which consists of a periodic phase structure such as a crystal lattice. We consider three beams which emerge from the specimen, the undisturbed beam and

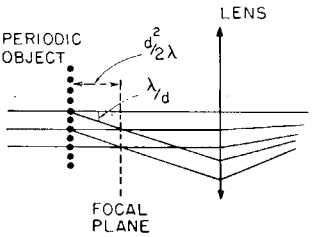


Fig. 25. – Illustration of the formation of phase contrast images of periodic lattice. For a conventional microscope, read the diagram from left to right. The first order diffracted beam interferes with the zero order beam in the focal plane of the lens so that an image is formed of the interference pattern on the screen. For the scanning microscope, read the diagram from right to left. As all the path differences are the same, we should observe interference fringes as we scan the electron beam across the focal plane. (Courtesy of *Quart. Rev. Biophys.*)

the two first order diffracted beams. These interfere in a region close to the specimen to produce interference fringes which form an intensity distribution with the same period as the original phase lattice. If the lens is focused on this region we can obtain an image of this intensity distribution.

Experimentally it is better to under-focus the lens rather than over-focus because the aberration defect produced by the defocus error tends to compensate the aberration defect produced by the spherical aberration of the lens.

In order to observe small spacings it is generally necessary to increase the size of the lens aperture. This tends to decrease the point (scattering) resolution but increase the lattice resolution. For this reason it is generally possible to observe lattice spacings which are smaller than the point resolution of the instrument.

These effects can be duplicated in a scanning microscope, and again we simply reverse the diagram. In this case we provide the zero order and two first order beams in the illumination systems (other beams are present but can be ignored). The zero order beam passes through the specimen and proceeds along the axis to the detector. The first order beams can be diffracted so that they also proceed along the axis and interfere with the zero order beam. The analysis of this situation is precisely the same as in the case

of the conventional microscope because again, the interference depends only upon path differences, not the direction.

Such fringes have been observed and are shown in Fig. 26. As in the case of the conventional microscope we can increase the lens aperture to obtain increased lattice resolution. We show a 3.4 Å spacing in graphite, whereas

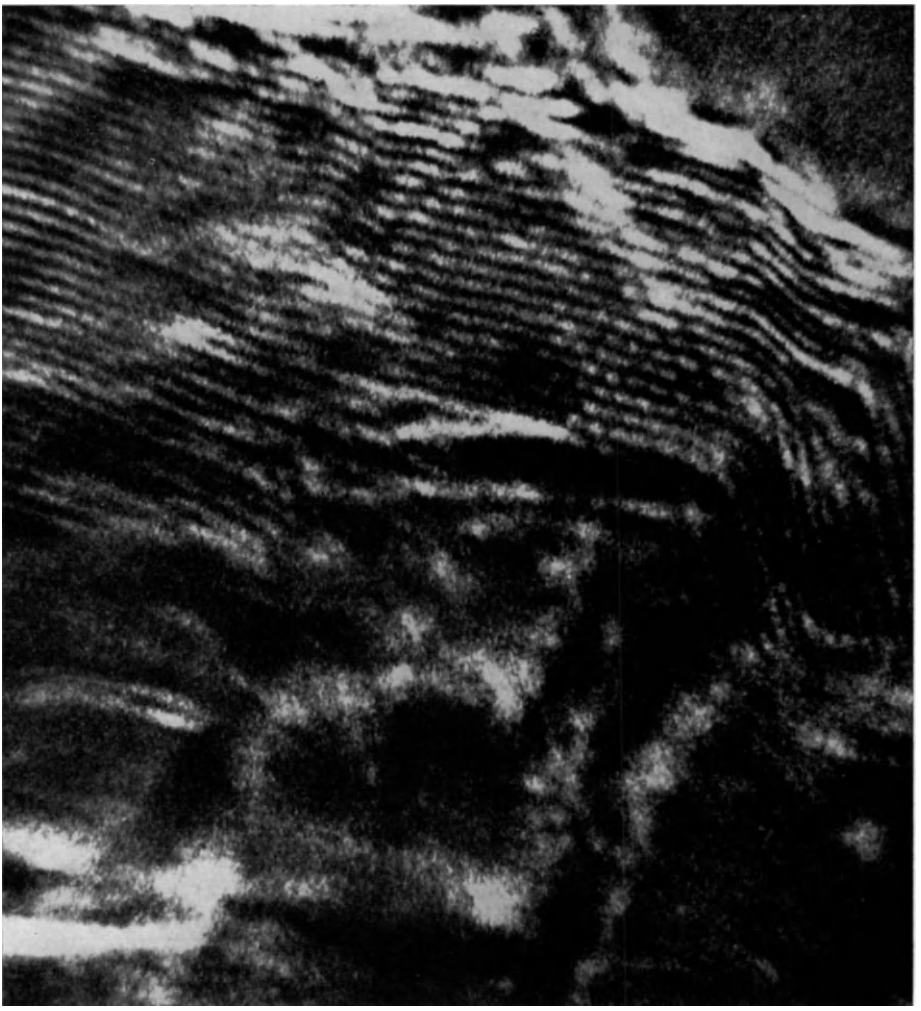


Fig. 26. – Phase contrast micrograph showing the 3.4 Å spacing in partially graphitized carbon. This picture was taken with a small aperture above the detector, a wide illumination angle and with the lens in an under-focus condition. (Courtesy of *Quart. Rev. Biophys*).

the point resolution of the instrument appears to be about 5 Å when operated at the same voltage (25 kV).

c) *Other interference and diffraction effects.* It is clear now that the scanning microscope produces an image identical to the conventional microscope when the ray diagrams are equivalent.

Any imaging condition for the conventional microscope can be duplicated for the scanning microscope by reversing the direction of the rays (as we shall see later the inverse of this statement is not true).

We therefore conclude that all the various types of image of which the conventional microscope is capable can be duplicated by the scanning microscope. So far we have observed diffraction contrast from lattice dislocations and extinction contours in addition to those previously described. There is no reason to doubt that other effects such as Kikuchi patterns could be obtained.

3·2. Mechanisms peculiar to the scanning microscope.

We will examine the fate of electrons which pass through a very thin specimen. These electrons can be divided into three categories.

a) *Elastically scattered electrons.* These electrons have almost the same energy as the incoming electrons. The difference in energy is undetectable.

The number of such electrons can be calculated from the value for the atomic elastic cross-section

$$\sigma_e = 46.5 \cdot \frac{Z^{\frac{4}{3}}}{V} \qquad (\text{Å}^2) .$$

This elastic scattering is characterized by a very wide scattering distribution which is proportional to

$$\frac{1}{(\theta^2 + \theta_0^2)^2} ,$$

where θ is the scattering angle and θ_0 a constant (the screening angle) which in our case is of the order of (50÷100) mrad.

Most of these electrons, therefore, are scattered outside the cone of the incident illumination ((10÷20) mrad). These electrons can be most readily detected by means of an annular detector of such a size that the unscattered electrons just pass through the hole in the detector.

b) *Inelastically scattered electrons.* These are the electrons which lose energy in the specimen. They can lose any amount of energy, but the vast majority of them fall into a group such that $1\,\text{eV} < \Delta E < 100\,\text{eV}$.

The total cross-section for this process is

$$\sigma_i = 868 \frac{Z^{\frac{1}{3}}}{V} \qquad (\text{Å}^2)$$

and the angular distribution is characterized by a narrow angular distribution

$$\frac{1}{\theta^2 + \theta'^2},$$

where

$$\theta' \sim \frac{\Delta V}{2V} \sim 1\ \text{mrad}.$$

Most of these electrons, therefore, will be within the cone of incident illumination and will pass through the hole in an annular detector.

c) *No-loss electrons.* By this we mean electrons which pass through the specimen without any interaction. Therefore

$$N = N_e + N_i + N_0,$$

where N is the number of incident electrons,

N_e is the number of elastically scattered electrons,

N_i is the number of inelastically scattered electrons,

N_0 is the number of no-loss electrons.

The no-loss electrons can easily be separated from the inelastically scattered electrons with the aid of the energy analyzing spectrometer. It is therefore a simple matter to arrange the scanning microscope to give three simultaneous signals as the beam scans across a specimen. These three signals correspond to the three groups of electrons. This is shown in Fig. 27. These three signals can be used in a variety of ways, but two illustrations will suffice here.

i) N_e/N_0. This is a normalized elastic scattering signal which is formally equivalent to the use of dark-field illumination in a conventional microscope with the important exception that the detector can subtend an angle of 250 mrad or more, thereby providing a large signal.

ii) N_e/N_i. We call this our « Z » contrast signal because it can be easily verified that $N_e/N_i \sim Z/19$. In other words this ratio provides a signal which is proportional to Z, the atomic number. In turn this means that a micrograph can be obtained where the intensity is proportional to Z.

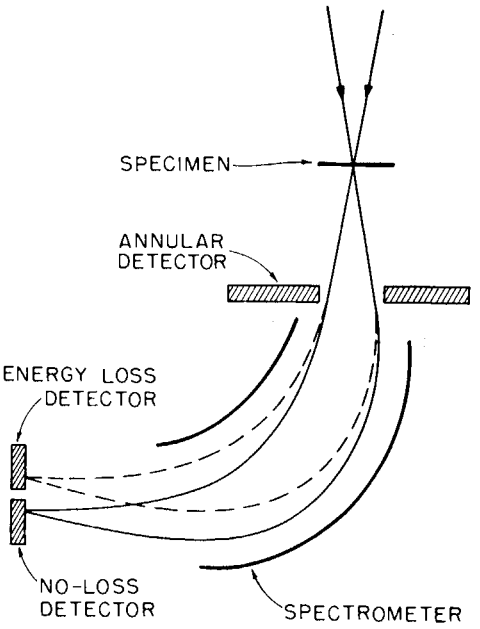


Fig. 27. – Electrons emerging from the specimen can be sorted into three separate groups by means of an annular detector which detects the elastically scattered electrons and a spectrometer which separates no-loss electrons from the inelastically scattered electrons. By this mechanism three simultaneous electrical signals can be obtained from the microscope.

This mode of contrast cannot be achieved in a conventional microscope. Very few high resolution microscopes have been fitted with energy analyzing equipment. Even where this has been done the resolution deteriorates because of the electron-optical problems.

While it may be possible to extract the two pieces of information N_0 and N_i with difficulty from a conventional microscope, it does not appear feasible to extract them simultaneously.

We show some examples of micrographs obtained with these signals in Figs 28, 29, 30.

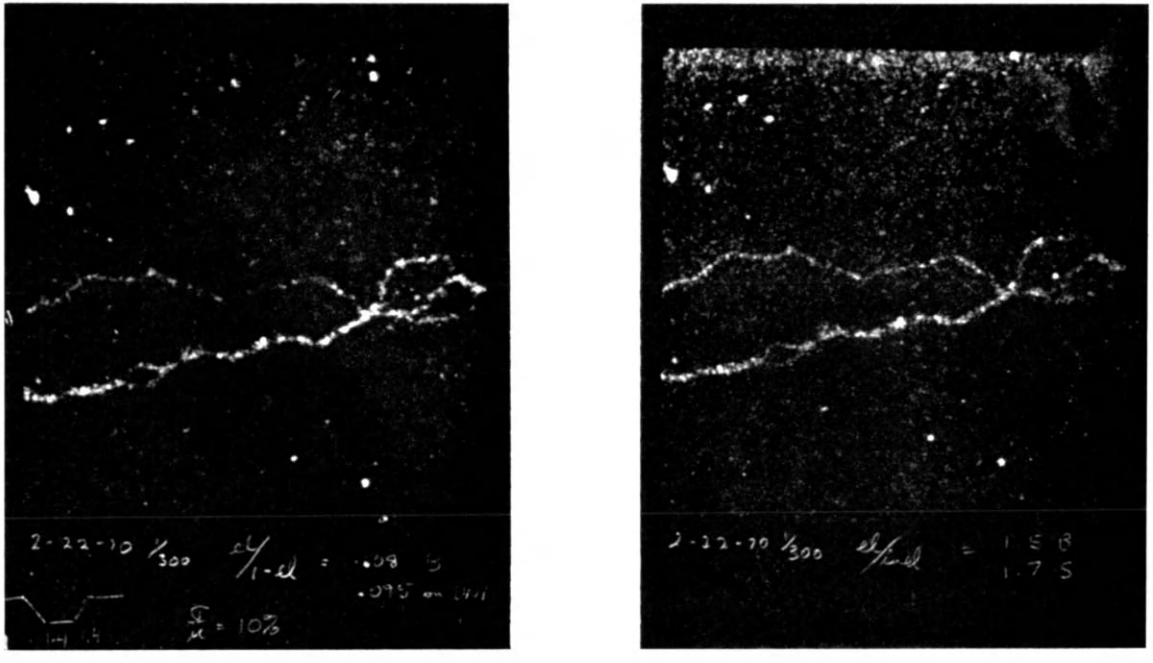


Fig. 28. – T_4 DNA stained with 10^{-3} M CsCl. 3000 Å full scale. Picture on the left is formed by elastically scattered electrons; the one on the right by the ratio of elastic to inelastic.

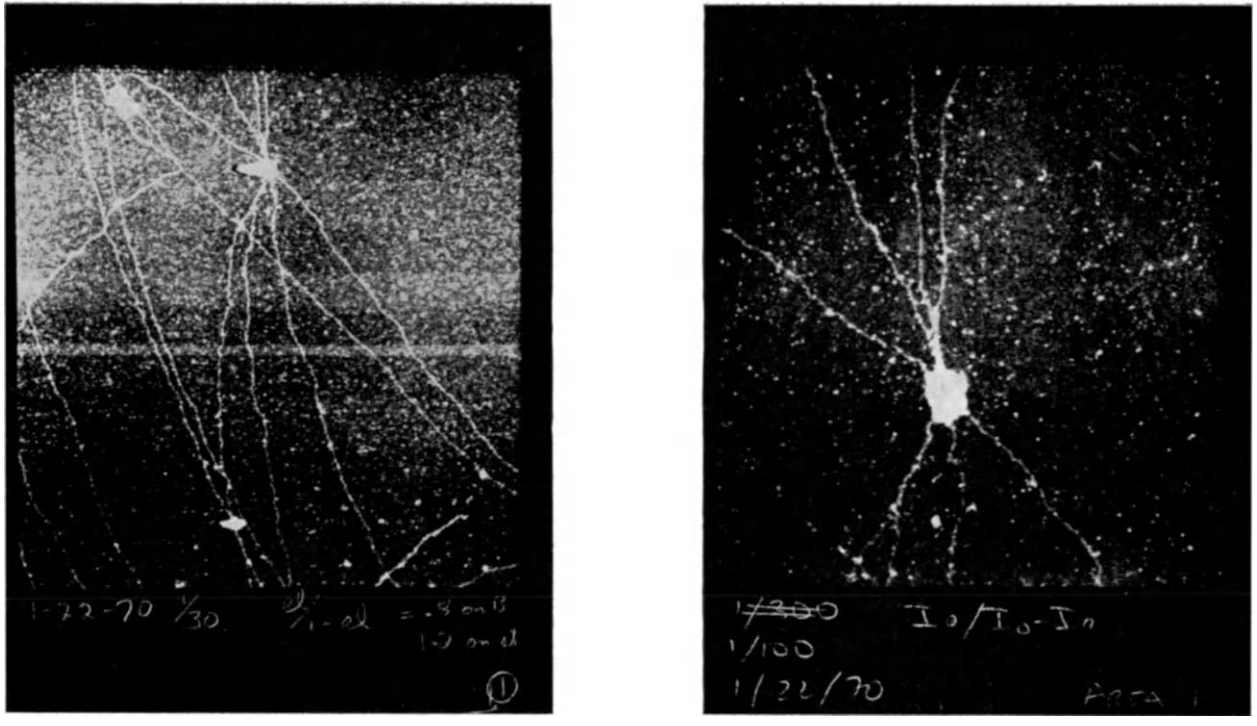


Fig. 29. – T_4 DNA in 10^{-3} M sodium. Both pictures are formed by elastically scattered electrons. Left: 3 μm full scale; right: 1 μm full scale.

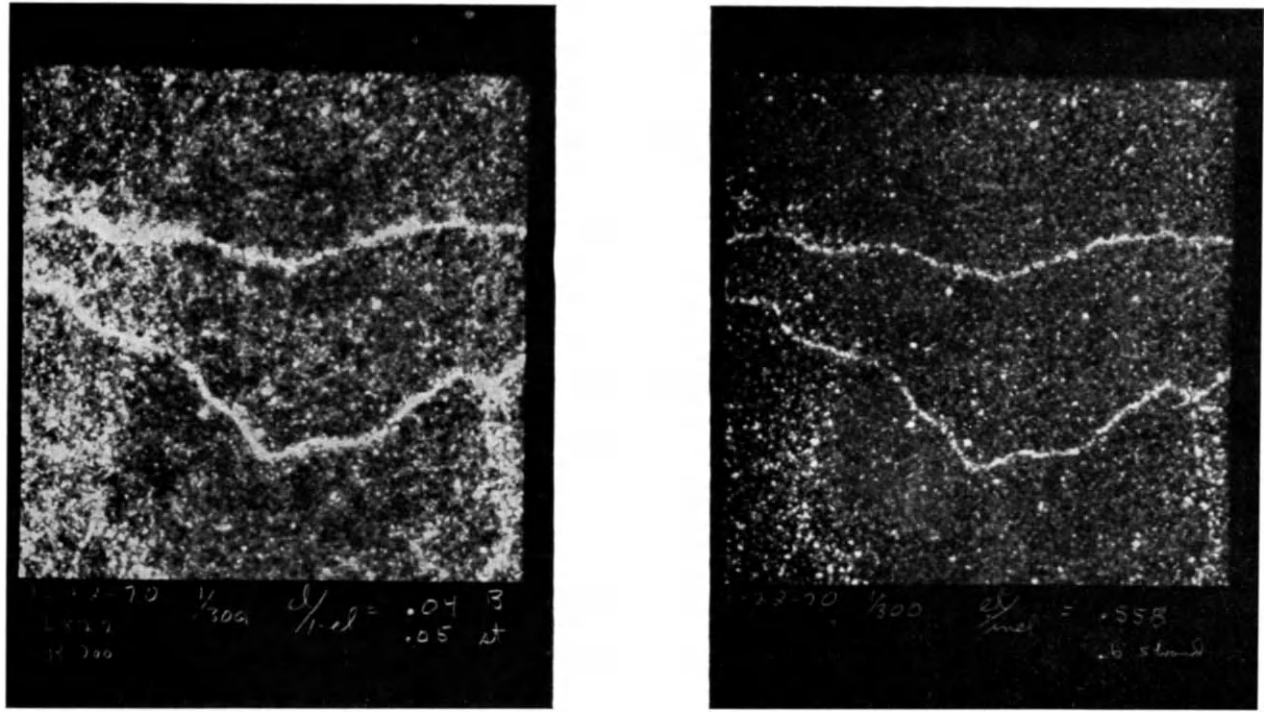


Fig. 30. – T_4 DNA in 10^{-3} M sodium, 3000 Å full scale. The picture on the left is formed by elastically scattered electrons; the one on the right by the ratio of elastic to inelastic.

3·2.1. *Single atom contrast.* – As a specific case let us consider the visibility of a single heavy atom which is placed on a substrate of carbon. This is a typical biological problem involving the visibility of specific stains. Single atom visibility has not yet been achieved, but we calculate it here as an example.

The signals from the annular detector and the energy-loss detector are

$$N_e = N \cdot \frac{\sigma_a + n_c \sigma_c}{\sigma_b} ,$$

$$N_i = N \cdot \frac{\sigma_a' + n_c \sigma_c'}{\sigma_b} ,$$

where σ_a and σ_c are the scattering cross-sections for the atom and carbon (the prime indicates the inelastic process), n_c is the number of carbon atoms intercepted by the beam, σ_b is the cross-section of the beam and N is the number of incident electrons.

Then

$$\frac{N_e}{N_i} = \frac{\sigma_a + n_c \sigma_c}{\sigma_a' + n_c \sigma_c'} .$$

But

$$\frac{\sigma}{\sigma'} = \frac{Z}{19} \qquad \text{and} \qquad \frac{\sigma_a}{\sigma_c} = \left(\frac{Z}{6}\right)^{\frac{4}{3}} .$$

We therefore obtain

$$\frac{N_e}{N_i} = \frac{Z/19 + (6/Z)^{\frac{1}{3}} \cdot (6/19) \cdot n_c}{1 + (6/Z)^{\frac{1}{3}} \cdot n_c} .$$

We can proceed one step further by noting that

$$n_c = \sigma_b \cdot t \cdot 0.11 ,$$

where t is the thickness of the carbon film in Å (there are 0.11 atoms per cubic Å in carbon).

This gives

$$\frac{N_e}{N_i} = \frac{Z/19 + 0.063(t \cdot \sigma_b)/Z^{\frac{1}{3}}}{1 + 0.2(t \cdot \sigma_b)/Z^{\frac{1}{3}}} .$$

This function is plotted in Fig. 31 for various values of σ_b and assuming $t = 20$ Å.

It appears from this figure that single atoms should be visible using a signal corresponding to N_e/N_i. However, we should investigate the statistical properties of the signal.

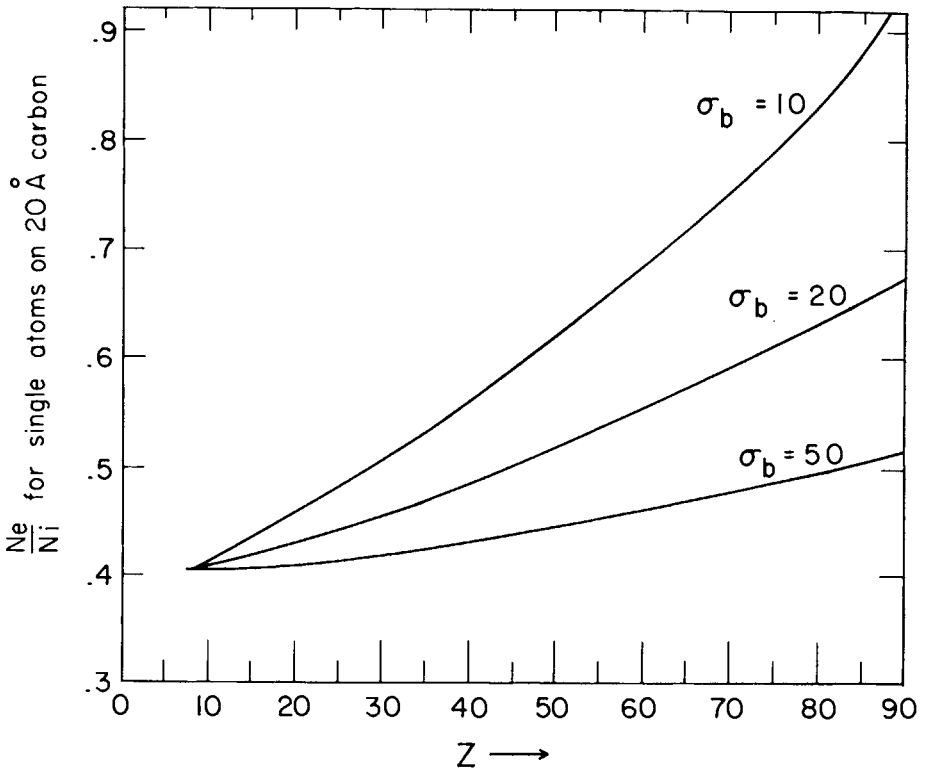


Fig. 31. – The value of the ratio of elastic to inelastic signal obtained for a single atom of the atomic number Z resting on top of a carbon film 20 Å thick. The three curves show the expected visibility of the atom as a function of the cross-sectional area of the focused beam. The cross-sectional areas are given in square Ångströms.

We therefore need numerical values for N_e/N and N_i/N. These can be obtained from Subsect. 3·2, *a*) and *b*). We will assume here a value $\sigma_b = 20$ Å^2 (approximately our current value). Then we can plot N_e/N and N_i/N as a function of Z. The values of N_e/N_i can be obtained from these curves, but we are interested in statistical variations about the mean. We take

$$\frac{N_e \pm \sqrt{N_e}}{N_i \pm \sqrt{N_i}} = \frac{N_e}{N_i}\left[1 \pm \frac{\sqrt{N_e}}{N_e} \pm \frac{\sqrt{N_i}}{N_i}\right] = \frac{N_e}{N_i}\left[1 \pm \left(\frac{1}{\sqrt{N_e}} + \frac{1}{\sqrt{N_i}}\right)\right].$$

The statistical variations can therefore be represented by

$$\frac{N_e}{N_i}\frac{\sqrt{N_e} + \sqrt{N_i}}{\sqrt{N_e N_i}} = \frac{1}{N_i}\left(\sqrt{\frac{N_e}{N_i}}\sqrt{N_e} + \sqrt{N_i}\right).$$

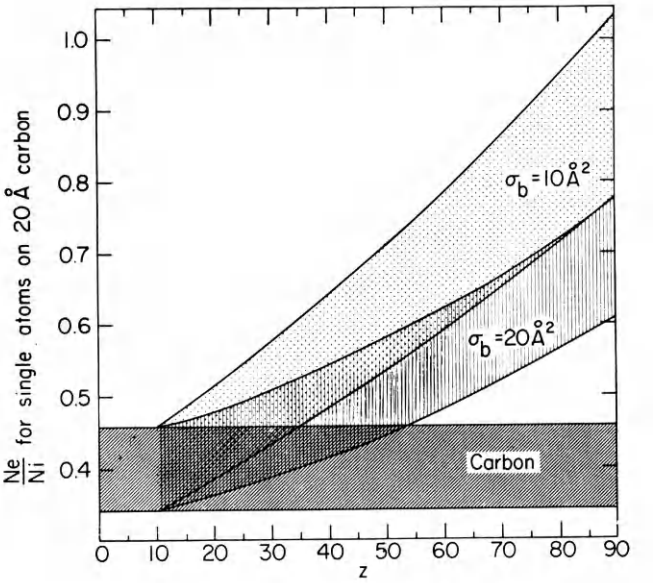


Fig. 32. – Visibility of single atoms as a function of atomic number taking into account statistical variations.

We now only need a value of N. A representative value for our present instrument is $N = 1500$ electrons per resolution element. Using this value we can calculate the probable errors in the curves of Fig. 31. We show the results in Fig. 32.

We conclude from these calculations that it may be possible to « see » single atoms with the present machine, but the task would be easier with an intrument with higher resolving power.

It should be noted that the assumptions made here mean that a uranium atom, for example, provides about 60 elastically scattered electrons and about 12 inelastically scattered electrons. It remains to be seen whether or not such an atom will remain in place on the specimen during this process.

REFERENCES (Section 3)

1) A. V. Crewe and J. Wall: *Optik*, **30**, 461 (1970).
2) A. V. Crewe and J. Wall: *Proc. 27th Annual EMSA Meeting, St. Paul 1969*, p. 172.

Special Electron Microscope Specimen Stages

U. VALDRÈ

Istituto di Fisica dell'Università - Bologna, Italy

M. J. GORINGE

Metallurgy Department, University of Oxford - Oxford, England

1. Introduction.

We shall define a *specimen stage* as any device by means of which it is possible to act on the specimen or to perform some treatment of the specimen. The treatment may be geometrical (*e.g.* displacement, orientation, etc.) or mechanical (*e.g.* straining, scratching), chemical, physical (*e.g.* evaporation, deposition, bombardment), thermal, electrical, etc. All those accessories which enable us to collect, record and analyse the signal produced by a treated specimen are not considered here as specimen stages (*e.g.* X-ray or light spectrometers, secondary electron detectors, probe forming lenses, scanning systems) and therefore will not be described.

In addition, as this course is devoted to the application of electron microscopy to material science, where the materials are essentially crystalline objects showing predominantly diffraction contrast, we shall deal here with those stages which allow, at the least, performance of diffraction contrast experiments. It is obviously not possible and also not really profitable, even with these restrictions, to cover the whole range of designs of specimen stages for any given application; therefore a full description will only be given for some selected stages although, where possible, comparison with others and technical comments will be made. Further selection criteria used are that no detailed description will be given of stages commercially available from electron microscope manufacturers and that, in the main, only recently published work will be presented.

2. Generalities and definitions.

In the early days of electron microscopy when the lack of knowledge of the effect of possible mechanical instabilities on the resolution worried the designers, the specimen was rigidly clamped to the objective pole piece.

Later, specimen stages allowing traverse (x, y) movement of the specimen were introduced, followed by the adoption of specimen air locks. It is fair to say that these represent the first specimen stages and the first specimen treatments inside the microscope.

However, the advent of transmission electron microscopy of thin crystals, where the image contrast is essentially diffraction contrast and therefore orientation dependent (see for instance Howie, this volume), called for the development of special stages, the *inclination stages*, capable of tilting the specimen in any direction. For any meaningful experiment on crystals, inclination stages are indispensable.

While the stages for traverse movement have reached a very high degree of perfection, inclination stages are still not entirely satisfactory. This is because of the obvious complications arising when combining tilting and traverse and because of the limited space available inside the objective pole piece where the specimen is usually located for high resolution work.

In the following, general principles will be outlined and only a few relevant examples will be given, selected from the large number of practical inclination stages.

We shall define (following an arbitrary, unofficial but widely used practice):

Tilting stage a device which performs tilting of the specimen and does not usually provide a direct and accurate measurement of the angular co-ordinates of the specimen (see Sect. **4**).

Goniometer stage a device for performing tilting where the specimen orientation (or more correctly, the orientation of the specimen holder) is directly measured, with an accuracy better than $\pm 0.1°$. Very few goniometer stages exist ([1-3]); their maximum tilting angle (typically 10°) is usually smaller than for tilting stages; they replace the whole specimen chamber including the air lock, and, of course, their cost is much higher (at least 10 times). They are used for very specialized work (*e.g.* where the relative orientation of various regions of the specimen has to be determined and Kikuchi lines cannot be used or are too difficult to interpret) and will not be treated here.

Polar diagram a two dimensional calibration of the inclination stage. The polar diagram for a given stage can easily be obtained optically in the following way, which is also a good macroscopic test of its performance ([4]). Referring to Fig. 1, *I* is the inclination stage which is placed above a plane *P*.

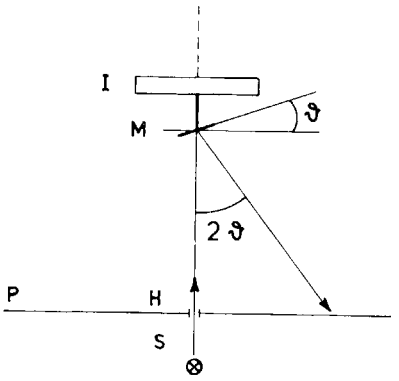


Fig. 1. – Schematic drawing of a device for testing inclination stages and for tracing their polar diagrams.

The specimen is replaced by a mirror *M* on which impinges a collimated beam of light coming from a source *S* through a hole *H* in *P*. The beam is normal to *P* and represents, for the stage *I*, the axis of the electron microscope. The light, reflected by the mirror, produces a spot on *P* corresponding to each inclination θ of the mirror. On plane *P* concentric circles of centre *H* can be drawn, each one corresponding to a given specimen inclination θ. By means of this simple device it is therefore possible to:

1) measure the maximum angle of tilt in the various directions, *i.e.* to determine the shape of the polar diagram;

2) calibrate the stage by recording the correspondence between the reading of the tilt controllers and the angles of tilt (latitude and azimuth, see Sect. **4**);

3) check the performance of the stage (reproducibility, inertia, hysterysis, continuous smooth or jerking movement, etc.).

Ideal inclination stage. An ideal inclination stage should fulfill the following requirements which one must keep in mind when selecting a stage:

a) Highest tilting angle in all directions. Medium tilting is now considered to be in the range of $\pm 20°$ to $\pm 30°$, low and high tilt below 20° and above 30° respectively.

b) High resolution. The inclination stage should not reduce the ultimate resolution of the instrument.

c) High accuracy. The tilting of the specimen should be smooth, continuous, reproducible and free from backlash and inertia over the whole range of tilting angles. In addition there should be no interference between tilting motion and other possible motions (*e.g.* traverse).

d) No specimen shift and no change of level of the observed area during tilting.

e) Constant image orientation; *i.e.* the image of the specimen should preserve its orientation during tilting with respect to a given fixed reference frame (such as the photographic plates). While requirement *c*) depends on the accuracy of the construction, *d*) and *e*) are related to the working principle of the device.

f) No restriction of the traverse movement with respect to that of standard stages.

g) Air lock for changing specimens and anticontamination device available.

h) A wide range of specimen sizes and thicknesses should be accepted by the device.

i) There should be a linear relationship between the tilting angles and the reading of the tilt controllers. Indication of the tilt angles should also be given.

j) Robust, low maintenance, low cost.

k) Electrically foot-operated in order to be able to carry out traverse and tilting simultaneously.

Such a stage does not exist; we will see later which items feature in the performance of practical stages.

Categories of inclination stages. As traverse movement is an essential feature of any specimen stage which has to be retained in the inclination stages, it turns out that inclination stages may be divided in mainly three categories:

i) Stages where the tilt action is accomplished inside the specimen (traverse) stage (Fig. 2*a*)). In this case the tilt axes *aa* and *bb* follow the traverse of the specimen and therefore only one point of the specimen (that

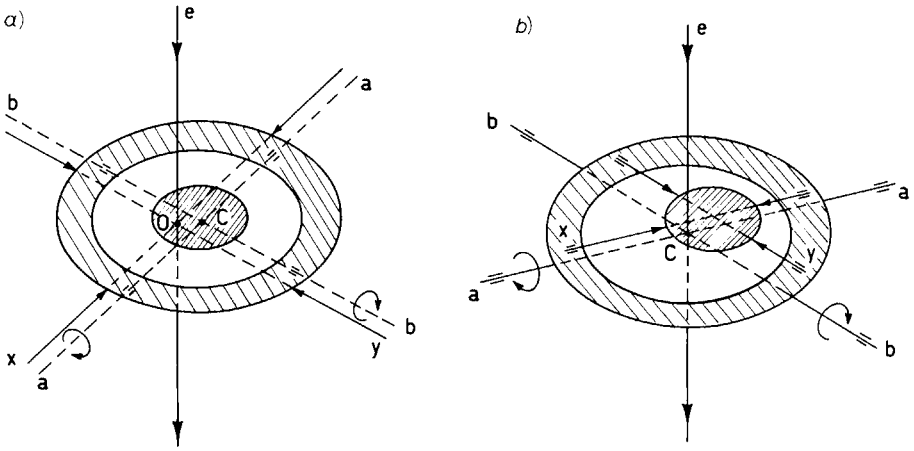


Fig. 2. – Schematic representation of tilting stages. *a*) The tilting device is built inside the traverse stage; *b*) the specimen traverse stage is inside the tilting stage.

coincident with the point *C* of intersection of the tilt axes) will keep its level constant during tilting and therefore maintain focusing condition under observation. For all other areas changes of level and side movements usually occur during tilting. Refocusing of the specimen changes the magnification of the image and the diffraction camera length. This change can be accounted for if the microscope is provided with an objective current meter by noting the deviation of the current reading from normal operating conditions. Nearly all practical stages belong to this category.

ii) Stages where the traverse action is carried out inside the inclination stage (Fig. 2*b*)). In this case, once the point *C* of intersection of the tilt axes *aa*, *bb* is made to coincide with the electron beam axis and once the specimen plane is made to coincide with the plane defined by the tilt axes, every region of a flat specimen, when it is observed, will be at the same level as *C* and the device operates at constant focus (and magnification) condition.

iii) Stages which combine (for one tilting and one traverse axis each) the features of categories i) and ii); in particular stages where one of the tilt shafts is physically coincident with one of the traverse shafts.

As for the *location* of the specimen stage, two solutions can be adopted: the stages may rest on the top plate of the objective lens (« top » stages) or may be built in the objective pole piece gap and rest practically on the top of the lower pole piece (« gap » stages). The first solution has the advantage of leaving more room in the proximity of the specimen holder for special treatments (*e.g.* for magnetization, liquid helium, evaporation stages. See Sects **5**, **6**, **7**) than the second solution. Conversely the latter is useful when narrow but ample room is required at the specimen level (*e.g.* for specimen straining. See Sects **5**, **6**, **7**).

The *specimen holder*, in the form of a short cartridge or a long rod, can be inserted in the specimen stage from above the objective lens (« top » stages) or from its side (« gap » stages). In the first case it has become common to name the holders *cartridges*, in the second case they are referred to as specimen *rods*, side entry rods or injectors.

Stages belonging to category ii) are definitely to be preferred to the others because, in principle, no shift and no change of focus is suffered by the area under observation during tilting. However, they require a complete redesign of the specimen section of the microscope and it seems that it will be extremely difficult to obtain a high degree of tilt in all directions.

The most popular stages belong to category i) because they use the existing traverse facilities of the instrument and only relatively minor alteration and additional work is required for fitting the tilting mechanism. They are normally used in connection with top entry cartridges.

Solution iii) is straightforward for those electron microscopes where the specimen holder is of the side entry rod type.

Two *basic principles* may be used to obtain a predetermined inclination of the specimen:

a) Rotation of the specimen around an axis *normal* to the specimen followed by tilting around a fixed axis; *i.e.*, combination of one axis of rotation with one axis of tilt;

b) tilting around two mutually perpendicular axes.

In case *a*) it is very simple to obtain a given specimen tilt when the rotation axis is parallel to the beam, but a complicated calculation is required before a second predetermined tilt is obtained. In such circumstances it is advisable to bring the specimen back to the horizontal position. This way of getting tilt is not ideal for stereo pictures because the images of the stereo pair are usually rotated with respect to each other. Also the different orienta-

tions of the images with respect to the operator at various tilts is sometimes confusing and it is difficult to recognize the same area for work involving the use of various reciprocal lattice vectors. The polar diagrams obtainable using this principle are always circular and the linearity condition can often be satisfied.

It is possible in theory to avoid rotation of the specimen for cartridges of this category if, following an idea developed for a goniometer stage ([1]), the tilting device can be disengaged from the specimen holder, rotated by the required amount and then recoupled for tilting.

When principle *b*) is used, direct measurement of the tilt angles is not generally possible and use should be made of relatively simple formulae (see Sect. **4**). *b*) does not suffer the inconvenience of *a*) noted above and it is ideal at least for qualitative work. Various types of polar diagrams (circular, elliptic, square, etc.) may be obtained, according to the amount of tilt available for the two axes and to the working principle of the practical device. Only in special cases is linearity obtainable.

3. Examples of practical tilting stages.

In practice the two basic principles of operation, *a*) and *b*) of Sect. **2**, can be applied in many different ways according to the mechanism or the working principle devised for the operation of the tilting device. We shall present below only a few examples of tilting stages; they have been selected according to their popularity, simplicity or outstanding features.

1) *Double tilting stages of category* iii). Practical examples exist only for side entry rods. As already mentioned, in a side entry stage the rod carrying the specimen may easily be animated by two motions: a traverse motion along the axis of the rod and a rotation (tilting) motion around the same axis.

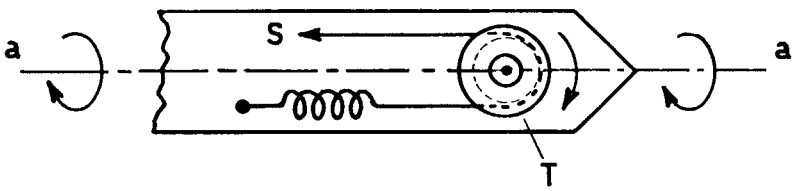


Fig. 3. - Combination of a single axis of tilt *aa* and rotation in a side entry type of tilting stage.

A tilting stage based on principle *a*) can be made by holding the specimen in a turret *T* which can rotate around an axis normal to the specimen plane (Fig. 3) and passing through the first tilt axis *aa*. Rotation may be produced by the action, for instance, of a string *S* and of a counterspring ([5]).

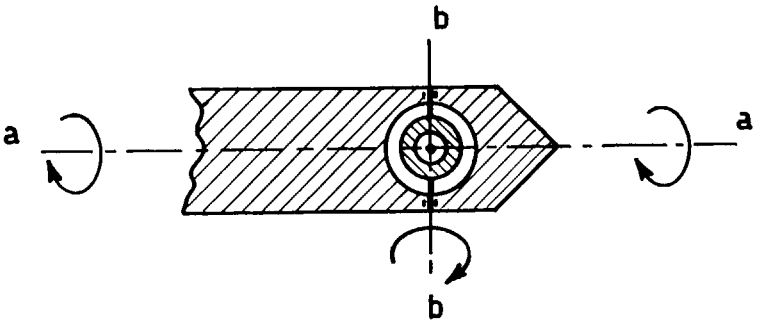


Fig. 4. – Two mutually perpendicular axes of tilt *aa* and *bb* in a rod type tilting stage.

Alternatively and better, a second, orthogonal tilt axis *bb*, may be added by simply mounting the specimen on a platform which can pivot around an axis perpendicular to the rod axis ([6]) (Fig. 4). Suitable means (levers, strings, etc.) passing through the hollow centre of the rod produce the required motion.

One special version of this category of stages has the two traverse movements (x, y) built inside the rod used for tilting ([6]), as schematically shown in Fig. 5. Sliding rod *R* inside shaft *S* provides the *x* traverse movement, rotation of *S* around point *C* produces the *y* traverse. One axis of tilt (*aa*) is provided by external tube *T*; the second axis has to be incorporated in rod *R*, as described for the case of Fig. 4. When the axis *aa* is made to intersect the axis of the microscope (at *O*) and the plane formed by the axes *aa*

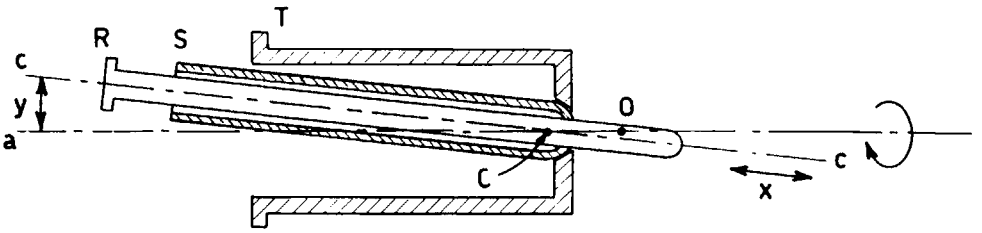


Fig. 5. – Schematic example of a practical stage having the specimen traverse movements built inside one tilt axis.

and *cc* is normal to the axis of the microscope, then the observed area of a flat specimen whose plane is coincident with the above-mentioned plane will coincide with *O* and tilting around *aa* will not alter the specimen level. The focus therefore remains unchanged and consequently the magnification of the micrographs and the effective camera length for diffraction work are not altered. This property does not apply however for the second axis of tilt.

2) *Double tilting stages of category* i). This system has been extensively utilised for top entry cartridges. The specimen is usually mounted either on the central portion of a universal suspension ([7,8]) or in a ball-like holder which seats in a spherical bearing ([9-11]). The spherical bearing is very easy and quick to build, it can be spring loaded in order to eliminate possible plays and has a good thermal stability. It may however introduce an additional, unwanted movement to the specimen (*e.g.* a small, uncontrolled, rotation).

The universal suspension does not suffer this inconvenience but the plays are difficult to eliminate, it is more delicate and requires painstaking and

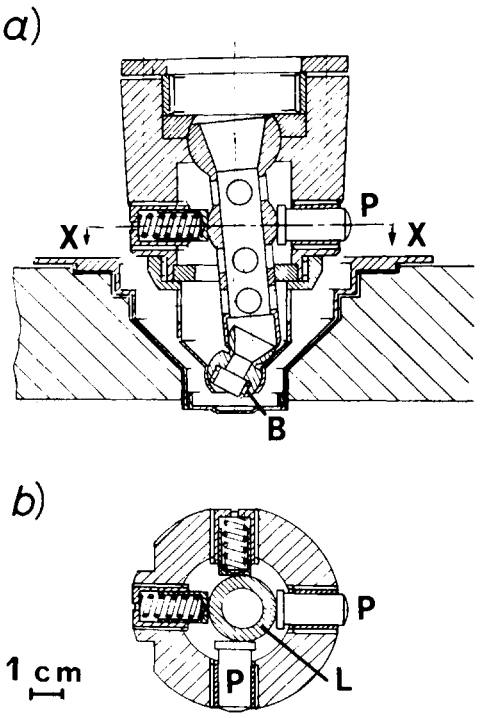


Fig. 6. – Practical example of a cartridge type tilting stage with two mutually perpendicular axes of tilt. *b*) is a section of *a*) through *XX*.

skilled workmanship. The various designs differ in the way the specimen holder is controlled.

A typical example is shown in Fig. 6 ([10]). Two micrometers operated from outside the microscope act on plungers *P* which in turn tilt the tubular lever *L*. *L* engages with the spherically shaped tail of ball *B* carrying the specimen and tilts *B* in its sprung support. Tilting angles up to 30° may be obtained in this way without loss of resolution. For a smooth and reproducible movement all sliding surfaces should be highly polished and friction minimized. An alternative solution is to replace ball *B* and its bearing by a set of gimbals. The rotational symmetry of this solution allows easy and accurate machining and the achievement of high mechanical and thermal stability. It is also easy to adapt this cartridge for use in special pole pieces or in work which requires a different specimen level (*e.g.* study of magnetic and lattice properties of magnetic materials outside the objective lens field).

A different approach to the problem of producing double tilting is to use translational movements in push-pull ([12]). Two pairs of strips or rods *RR*, *SS*, at the end of which is mounted a platform *P* (Fig. 7) carrying the

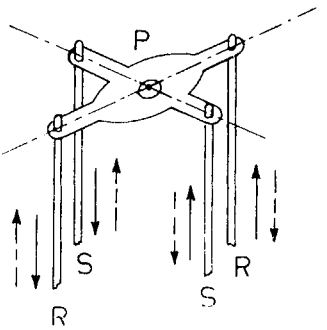


Fig. 7. – Working principle of a push-pull type of tilting cartridge.

specimen, are independently driven in push-pull. Actually the system can be simplified by using one push-pull pair of strips to provide one tilt and a third strip for operating directly the orthogonal tilt ([13]). Apart from the slight complexity of the system, lack of rotational symmetry, difficulty in obtaining control of the spring action of the strips, this device has the advantage of producing high angles of tilt (up to 40°) and a large clear solid angle to the specimen, which may be very useful, for instance in the study of X-rays or light emission from the specimen. In fact the specimen is screened only by the three or four strips of the push-pull system which can be made very thin.

4. Specimen orientation determination.

The determination of the orientation of thin crystals is usually very conveniently performed by analysing their diffraction patterns and, for thick and fairly perfect crystals, very accurate measurements can be made by using Kikuchi lines ([14]). However this method cannot always be used, and in any case, for polycristalline, amorphous, or biological materials such information cannot be derived from diffraction patterns and therefore it is necessary to provide the tilting stage with suitable meters. This specimen-independent determination of the specimen orientation in space is particularly useful in stereo microscopy ([15,16]) where, for quantitative work, it is necessary to know the tilt angle between stereo images accurately and, for best three-dimensional contrast, the pair of stereo photographs should be taken at a predetermined stereo angle depending on specimen thickness and magnification ([17]). We shall consider here the case of flat specimens and of tilting stages working on the principle *b*) of two mutually perpendicular axes.

As the determination of the specimen orientation is based on the knowledge of its position normal to the electron beam (« horizontal » position) we shall first examine some of the simplest and straightforward methods that can be used for checking the specimen horizontality.

At very low magnification (only the intermediate of the magnifying lenses excited) use can be made of a reference specimen (*e.g.* a square mesh grid) whose image will appear undistorted only when horizontal.

At low magnification (only objective lens on) the use of two, small, concentric apertures placed at a certain distance one above the other allows the horizontality of the specimen seating to be set within 0.5° by tilting the stage until a circular shape for the image (shadow) of the apertures is obtained.

At high magnification horizontality may be checked by using a flat specimen and by finding the conditions under which no change of focus occurs on traversing the specimen stage.

The operation of calibration for horizontality is usually required only once for a given tilting stage and records should be made of the readings of the meters connected to the tilt controllers. Let us call α and β the angular co-ordinates indicated by the tilt meters and suppose $\alpha = \beta = 0$ when horizontality is satisfied.

The specimen orientation in space may be determined by means of latitude φ and azimuth θ of its normal $\boldsymbol{n}$ taken from point O of intersection of

the tilt axes *aa* and *bb* (Fig. 8). φ and θ can be expressed in terms of the tilt angles α and β.

Starting from the horizontal position let us tilt the specimen until unit vector $\boldsymbol{n}$ joins point O with point P. This orientation may be obtained by an

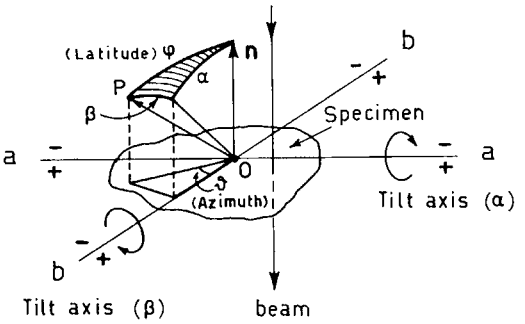


Fig. 8. – Defining the quantities used for measuring the specimen orientation.

amount α of tilt around axis *aa* followed by tilting around axis *bb* by a quantity β. It follows to a first approximation, by considering the spherical triangle shown dashed in Fig. 8 as a plane triangle:

$$\text{latitude } \varphi \simeq (\alpha^2 + \beta^2)^{\frac{1}{2}}, \qquad \text{azimuth } \theta \simeq \operatorname{arctg} \beta/\alpha . \tag{1}$$

The accuracy of these formulae decreases with increasing angle, but is still better than 7% ($\sim 1.5°$) when $\varphi = 25°$. The sign of α and β are referred to the clock or anticlockwise rotation of the tilting controllers (micrometers).

If the specimen is not flat, formulae (1) are in error, usually by a small amount; however no additional error is introduced when they are used for calculating the relative tilt angle between two specimen orientations.

5. Combined double tilting stages.

Under this heading we refer to special specimen holders which combine the basic double tilting operation with one additional specimen treatment (such as, for instance, rotation, heating, straining, etc.) and therefore

allow the use of diffraction contrast during the performance of a physical, mechanical, chemical, etc. experiment.

The increased manufacturing difficulties with respect to the straightforward double tilting stages have so far restricted their development.

5·1. Double tilting and rotation.

Two types of specimen rotation should be distinguished:

i) Rotation of the specimen around the electron beam (or, more precisely, in practice, around the microscope axis). During this rotation the specimen normal acquires a precession motion, while the diffraction pattern remains unchanged as well as the focusing condition of the imaged area of the specimen. The image of the obscrved area rotates around the microscope axis at a fixed distance.

ii) Rotation around an axis normal to the specimen (usually passing through the crossing point of the tilt axes). The image of the specimen rotates usually along an elliptic orbit around the point of intersection of the rotation axis with the specimen. This type of motion may be sometimes useful for bringing into the field of view areas of the specimen otherwise inaccessible. Focus and diffraction conditions change during rotation. Type ii) rotation is easier to achieve inside a holder than type i) and it is more conveniently obtained by rotating the seating of the specimen holder (usually a cartridge). Rotation should perform continuously in both directions through 360° with no end stops.

These special holders are useful when it is necessary to align specimen details or diffraction spots with respect to specific directions, for instance, given by electrical or magnetic fields, slits (in the case of velocity analysers), plates, serial of sections, etc.

They are very difficult to devise for a side entry rod and the practical examples refer therefore to top entry cartridges.

Rotation around the beam may be obtained by rotating the conical seating of the cartridge and by designing the tilt independent from rotation ([18,19]). In this case the cartridge is made up of two main parts: one part which is stationary, during rotation, with respect to the traverse stage and meets the tilt controllers, and one part which seats in the female cone of the stage and rotates rigidly with the cone. The tilting action is transferred from the top part to the low part through levers, the coupling being made via plane surfaces (sliding during rotation) which remain constantly perpendicular to the axis

of the microscope and whose level is controlled by the amount of tilt given to the specimen. In this way no unwanted tilting force is applied to the specimen holder during its rotation. (See Sect. 7, Fig. 12.)

In a similar stage, provided with rotation of the cartridge seating cone, rotation may also be performed around a specimen normal ([20]). The tilt controllers operate two rings which are pivoted independently of each other in the top part of the cartridge. These rings transmit their tilting motion by means of short rods and sliding contact, to the inner ring of a set of gimbals which is mounted in the top of the rotatable part of the cartridge. This inner ring is spring loaded so that it maintains contact with the short rods (and therefore maintains its inclination) during rotation. The inner ring could be the specimen holder in itself for work at long focal distance; however for high resolution work the specimen is held at the centre of a second set

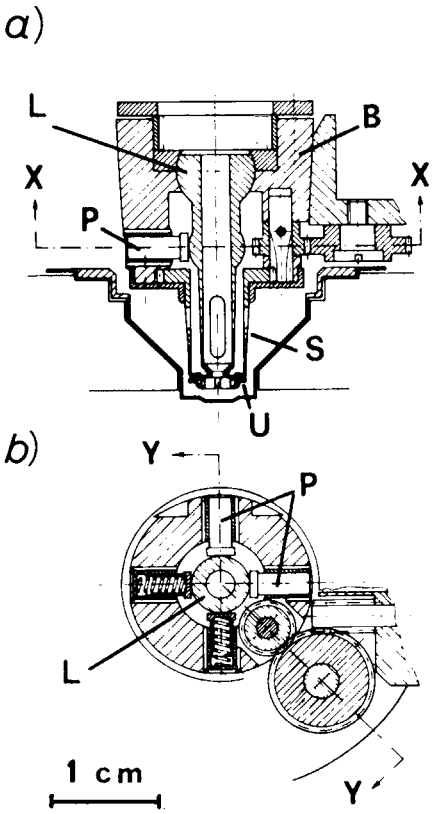


Fig. 9. – Example of double tilting cartridge with rotation around a specimen normal. *a*) and *b*) are sections through *YY* and *XX* respectively. (Courtesy of *Nuovo Cimento*.)

of gimbals which is operated, through long push rods, by the first one.

Another and much simpler way of achieving rotation around a specimen normal is shown in Fig. 9. Here the specimen is mounted at the centre of a universal suspension *U* and tilt is obtained by means of tubular lever *L*. The rotation controller is used for rotating the outer supports *S* of the gimbals (the cartridge body *B* remains fixed to the stage) and therefore the specimen. The constancy of the specimen normal is guaranteed by the position of tubular lever *L* which changes only when the tilt pushers *P* are operated. Due to the kinematical properties of the universal suspensions the rotation speed is not uniform during a 360° turn.

5·2. **Double tilting and lifting.**

In these stages the specimen level can be varied continuously with a motion in the direction of the axis of the microscope by means of a controller

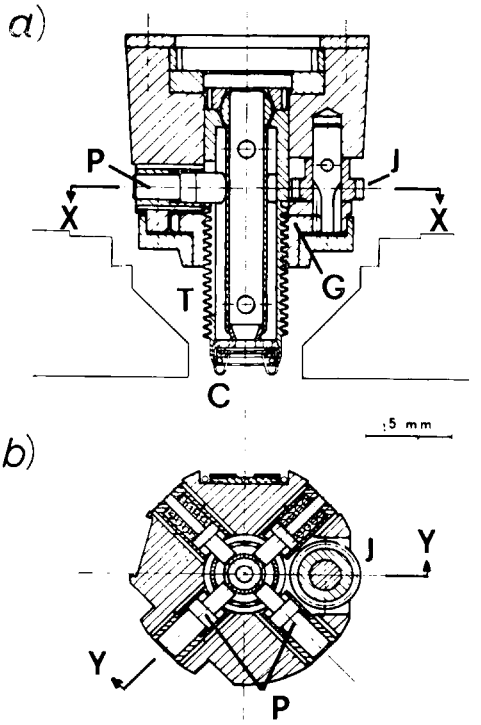


Fig. 10. – Double tilting and lifting cartridge provided with electric contacts. *a*) and *b*) are sections through *YY* and *XX* respectively.

external to the microscope. Typical applications are in the study of lattice defects of the specimens requiring high resolution microscopy and their correlations with low angle electron diffraction (for the study, for instance, of dispersed phases or of p-n junctions in semiconductors) or in connection with out of focus techniques or when the specimen has to be taken out of the magnetic field of the lens or to be given a particular treatment which cannot be carried out inside the objective lens.

Figure 10 shows a cartridge performing double tilting ($\pm 25°$) and lifting of the specimen by 7 mm ([21]). The lifting is obtained by making use of the same facilities as for rotation, through gears *J*, *G* and the threaded tube *T*. The maximum tilt angle is constant at every level of the specimen, but different excursions of pushers *P* are required. The cartridge is also provided with electric contacts *C*.

5·3. **Double tilting and deformation.**

Tensile straining is the type of deformation which has been so far almost exclusively applied to thin foils and to it we shall refer in the following.

A straining device can be *hard* or *soft* depending on whether a given strain rate or a constant load are applied respectively. Nearly all the straining stages available belong to the hard type.

The straining is *symmetrical* if both the specimen clamping jaws move apart at the same rate. The observed strained area should therefore remain in the field of view during deformation. With asymmetrical straining large shifts of the image occur.

A straining device should be able to perform coarse strain (up to a rate say of $10^{-2}\,s^{-1}$) and fine strain (down to $10^{-6}\,s^{-1}$) independently applicable and a maximum total strain of the order of a few hundreds per cent (for the study of fracture and of super-plastic materials).

It is also essential to incorporate means for quantitative studies, *i.e.* strain gauges for the measurement of stress and strain. The local strain can be very accurately measured from the separation of stationary details present in or on the specimen; the applied load (from which the stress can be derived) can be measured by means of strain gauges (based on changes of resistance or capacity), calibrated springs, or pressures.

The specimen may be glued or, preferably, clamped to the mounting jaws. In the latter case high temperature experiments are possible. It is important to shape the specimen properly in order to concentrate the strain in a very

localised area. Suitable dishing, polishing and transferring techniques have been developed ([19,22]).

Straining has been obtained mechanically (by means of levers, screws or cams), elastically (springs or elastics), hydraulically, by thermal expansion of wires or rods, by thermal deformation of bimetallic strips and by gravity.

It is amazing that despite the large variety of stages designed and constructed, comparatively little applied work has been done and very few results obtained. The reasons for this deficiency are primarily to be found in the lack of double tilting straining stages capable of quantitative measurements, and only secondarily in the fact that electron microscope specimens do not always represent the properties of the bulk material.

Straining devices to be used in goniometer stages are relatively simple ([23-26]); the tilt angle is however restricted to less than 10°. Only one ([24]) allows measurement of the applied stress.

Usually the measured load contains a systematic error in the sense that it is the sum of at least two terms: the load applied to the specimen and the load applied to overcome frictional forces acting on the moveable parts of the straining device.

A good example for quantitative tensile device is that designed by Saka *et al.* ([24]). The specimen is glued at the ends of two parallel jaws, one being thin and flexible and the other (the driving jaw) rigid. The moveable jaw is held by a flat spring, preloaded by means of a wire. On heating the wire the driving jaw moves, pulls the specimen and bends the flexible jaw, on the side of which a semiconductor strain gauge is glued. Load measurements in the range 0.05 to 50 g are possible with a linear response. In the present version this quantitative straining device is expected to suffer from large specimen drift.

The construction of double tilting straining stages is more difficult and very few practical examples exist ([19,27]); one of these will be described in Sect. 7 as a multipurpose stage.

Side entry holders seem to be more suitable than top entry cartridges in the construction of accurate straining devices, although more limited in the second tilt.

5·4. Double tilting and heating.

The performance of a heating stage may be characterized by the following factors: maximum temperature obtainable; accuracy of the temperature

measurement; temperature stability, specimen displacement and drift; rate of heating and cooling; heat input; further specimen treatment (*e.g.* oxidation, reduction, etc.).

The highest range of working temperature obviously satisfies the requirements for the study of the largest variety of phenomena and materials; however attention should be drawn to the fact that even at moderate temperatures (for instance 400 °C for Cu) thermal diffuse scattering reduces the number of electrons which pass through the objective aperture with the result of a drastic decrease in the image transparency. In addition, surface migration takes place from the thin edges of the specimen towards thicker areas and, unless suitable precautions are taken (*e.g.* reducent atmosphere), oxidation of the specimen may occur. Finally damage to delicate parts of the microscope (*e.g.* the objective pole piece) may result at high power inputs.

Two systems have so far been used for a controlled heating of the specimen: furnace heating and direct heating, although other forms of heating have been used (for instance beam heating) or can be conceived (*e.g.* electron guns).

In the first case the specimen is clamped by a screw inside a small furnace; the specimen temperature is easily controlled and fairly accurately known (within a few degrees without taking into account the heating from the beam and if the open solid angle for radiation losses from the specimen is small). If the furnace and the furnace winding are properly designed (circular symmetry along the microscope axis and use of bifilar or spiralized wire), the specimen drift as well as the beam displacement are negligible and the resolution is not marred. Typical heating and cooling rates are 10 °C/s. These devices are ideal for experiments in stationary conditions. In the case of double tilting stages a slight change of temperature (of the order of 10 °C) may occur during tilting as a result of altered thermal losses.

In the direct heating type, electric current is passed through the specimen itself and heat is produced by Joule effect proportionally to the electrical resistance of the specimen (or mounting grids). The calibration of devices of this type is therefore very difficult and unreliable; the specimen temperature can therefore be only guessed. Fast heating and cooling times are possible which may be particularly useful for experiments on structural changes of the specimen with temperature (*i.e.* annealing and quenching). The temperature stability is not good and large image displacements and drift occur. The heat input is, of course, much lower than for the furnace heating type.

In practice, a small furnace is built in the central portion of a universal suspension ([7,19]), or in the holder of two push-pull rods ([28]), or the terminals of two electrical contacts are brought in the specimen holder where the spec-

imen establishes a resistive contact ([13]). A thermocouple is usually placed on the seating of the specimen and may be calibrated by observing the occurrence of phase transformations on selected specimens. (See Sect. 7, Fig. 12.)

5·5. Double tilting and cooling.

It is commonly understood that the term cooling stage refers to devices capable of cooling the specimen down to liquid nitrogen temperatures, whereas the expression liquid helium stage means a microscope attachment covering the temperature range from about room temperature down to about 4 °K. Liquid helium stages usually require the replacement of the entire object section of the microscope.

Cooling may be produced mainly in two ways: *a*) by thermal conduction from a reservoir (or cooling device) placed either outside or inside the microscope and *b*) by direct flow of a coolant close to the specimen.

A small heater is usually incorporated in the cooling devices in order to cover the entire range of temperatures from the minimum obtainable to slightly above room temperature. Temperature control may also be achieved by acting on the temperature and/or the rate of flow of the coolant or by introducing a thermal resistance along the heat path.

The highest cooling and warming rates are obtained by means of system *b*), which may suffer from mechanical vibrations; on the other hand large drifts are experienced when the system *a*) is adopted.

The requirements listed before for the heating stages apply to cold stages except for the last one which is replaced by specimen contamination due to condensation of residual vapours in the microscope. In addition, specimen vibrations may occur if the system operates by a coolant flow.

i) *Liquid nitrogen stages.* In one practical version ([8]) the nose of the cartridge described in Fig. 6 is thermally isolated from the cartridge body and cooled either by a cold finger or by an elastically retained cold ring pressed against the nose, the ring being placed in the objective pole piece gap. The ball carrying the specimen is cooled by conduction down to — 130 °C.

One interesting solution has been recently proposed ([29]) for side entry rods. The specimen is mounted mid-way along the axis of a small cylinder which seats on two *V*-shaped grooves machined at the end of the side entry rod. Rotation of the cylinder by means of elastically loaded wires provides the second tilt and cooling is obtained by conduction through the rod itself which is cooled by liquid nitrogen. A system for compensating the thermal contraction is employed.

In order to prevent the condensation of residual vapours on the specimen, use should be made of anticontamination devices with this category of cooling stages.

ii) *Liquid helium stages.* The development of double tilting liquid helium stages has been very intense in recent years and was stimulated by the hope of observing superconductivity phenomena and to increase specimen transparency, for studying solidified gases, delicate specimens (ionic crystals and plastics), phase transformations, etc. Certainly more liquid helium stages are now available than double tilting liquid nitrogen cooled cartridges or rods.

Before using a liquid helium stage the microscope must be carefully checked for vacuum leaks (always present in a large number in instruments used for routine work); once this has been done, the cryogenic pumping action of the coolest parts of the stage (the He inlet pipe or the He reservoir) will produce a high degree of vacuum around the specimen which usually prevents contamination effects.

Usually the tilting angles are of the order of $\pm 10°$, the resolution 25 to 30 Å and the minimum temperature of the illuminated area about 8 °K. The specimen temperature is derived from the observation of the condensation in equilibrium condition on the specimen of suitable gases introduced into the microscope ([30]) or by the occurrence of phase transformations. Some stages have air lock facilities ([31-35]) which is an enormous advantage for increasing the efficiency of the observations, especially when liquid nitrogen shielding and liquid helium reservoirs are used ([31,34,36]). Stages using the principle of the coolant flow can cool the specimen down to a few degrees K in a few minutes. Typical helium consumption is of the order of a few litres per hour for maintaining the lowest temperatures. The specimens are commonly dish polished until a small hole is produced in order to present a thick rim for safe handling and clamping, and for good heat transfer.

iii) *Liquid helium and magnetization stages.* In the study of superconducting specimens or ferromagnetic materials at low temperatures it is necessary to apply a suitable magnetic field to the specimen. In the early work on superconductors use was made of the magnetic field provided by the objective lens of the microscope, the specimen being placed at an angle with a horizontal plane and/or positioned at different levels by means of a lifting stage ([32,37]).

Figure 11 shows an example of liquid-helium lifting stage. Liquid helium from a dewar and an ordinary transfer line circulates through flexible tubings T very close (a few tenths of mm) to the specimen which is mounted inside

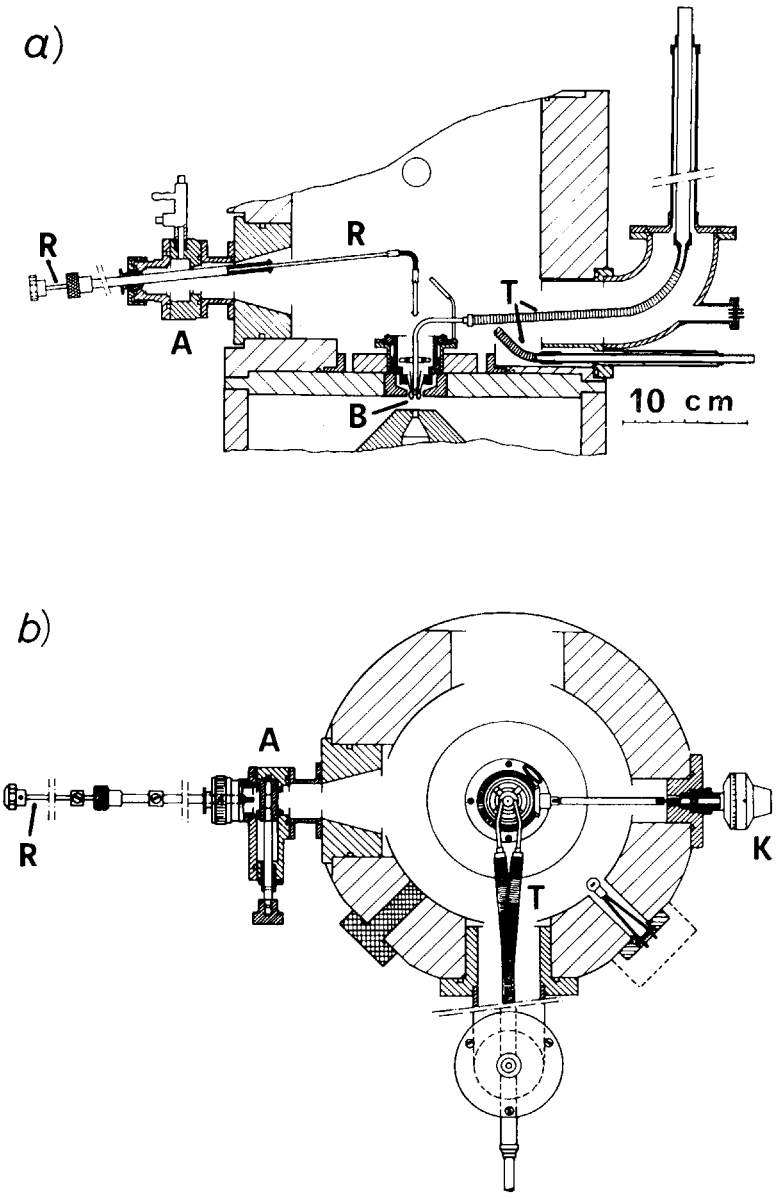


Fig. 11. – *a*) Schematic cross-section of a double tilting, lifting, liquid helium stage for a high voltage electron microscope. *b*) Plan view showing the actual position of the various components.

a holder at the centre of a spherical bearing *B*. By rotating knob *K* the specimen level can be changed and therefore different values of the objective magnetic field can be applied to the specimen. Flexible tubings and bellows *T* allow *x*, *y*, *z* displacements of the specimen and double tilting, the latter being produced by externally operated pushers acting against two counter-springs. The specimen holder can be picked up from or loaded in the spherical bearing *B* by means of rod *R* operated from outside the microscope through the air lock *A*. The specimen exchange time is only of a few minutes. It is also very easy to change cooling agent (liquid nitrogen, water, etc.).

Much greater flexibility is obtained with the use of an auxiliary magnetic field ([36,38,39]) applied to the specimen independently from the lens field. However the problem of satisfying simultaneously the various requirements of a high resolution, high magnification and versatile magnetization stages is still far from being solved.

iv) *Liquid helium stage with pressure cell.* This accessory has been developed in order to study lattice defects in condensed gases ([40]). This study is only possible if large crystals can be produced. By annealing in a pressure cell the small crystals (0.1 to 0.5 μm) present in gases condensed on various substrates, crystals up to several microns in size can be obtained which are suitable for observation. The pressure cell is in the form of a cartridge which is mounted in the tiltable holder of the liquid helium stage. The cartridge has two thin windows (made of formvar, carbon or sapphire) to allow the electron beam to pass through the cell and one pipe for the introduction of the gas. The plastic windows can stand pressures up to (10÷20) torr and the beam path through the gas is about 20 mm. A small heater is used for annealing the condensed gas. Provisions for ion bombardment or specimen evaporation are also incorporated.

6. Ultra-high vacuum stages.

They have been primarily designed for *in situ* chemical reactions and vacuum deposition studies in controlled surroundings.

Ultra-high vacuum electron microscopy may be achieved:

i) by converting the all microscope in an ultra-high vacuum system ([41]);

ii) by producing an ultra-high vacuum in the specimen region only, by means of differential pumping through various steps ([42-47]).

The latter solution seems the most popular today because it is more versatile and easier to achieve as it sets stringent requirements only on a small portion of the microscope. In a two-step system ([48]) a pressure of less than $1 \cdot 10^{-8}$ torr may be obtained by means of ion or cryogenic pumps in the specimen region, while a pressure of $\sim 10^{-7}$ torr is produced by similar pumps in the guard vacuum surrounding the ultra-high vacuum region. Here « dirty » operations may take place, like outgassing of filaments and evaporants. The remaining parts of the microscope are at a pressure of $\sim 10^{-5}$ torr obtained with the conventional pumping system of the instrument.

The vapour deposition takes place on substrates made with cleaved layer structure materials or on carbon films. The substrate can be heated up to a few hundreds of °C. For work near room temperature double tilting facilities are already available.

7. Multipurpose stages.

A microscope should ideally be equipped with a universal specimen stage capable of performing all the desired specimen treatments, if necessary, simultaneously. Such a stage would obviously be very complicated and extremely difficult (if not impossible) to construct to the required degree of accuracy for high resolution work.

This problem has been tackled gradually. Firstly by developing a system of devices (cartridges or rods) which are all compatible with a permanent objective section of the microscope, rather than developing a set of completely different units. Secondly, by designing multipurpose stages and holders.

A unique example of a multipurpose stage for a conventional 100 kV microscope is given by the stage built by Mills and Moodie ([13]). Here provisions are made for three translational degrees of freedom x, y and z (along the microscope axis), double tilting ($\pm$ 40°), cooling and (direct) heating from $-$ 140 to 1200 °C, gas inlet and decontamination.

A series of multipurpose cartridges and rods are under development for a 1 MV electron microscope ([19]) which can be equipped with either large bore or large gap objective pole pieces. Figure 12 shows a cartridge capable of

double tilting (± 25°), rotation around the microscope axis (± 45°), furnace heating (1000 °C), gas inlet (including facilities for mounting a pressure cell) and gas analysis by means of a sniffer placed close to the specimen.

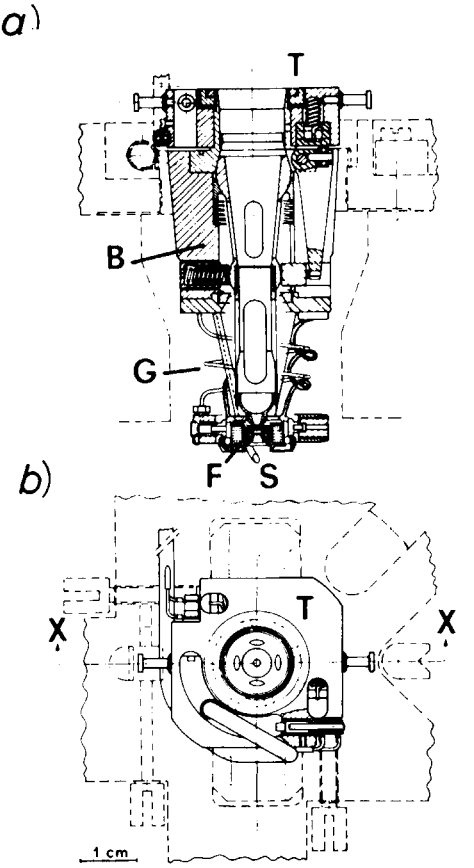


Fig. 12. – Cartridge for a high voltage electron microscope. It allows double tilting, rotation, heating, gas inlet and gas sampling. *a*) is a section of *b*) through *XX*. *F*, furnace; *G*, gas inlet pipe; *S*, sniffer; *T*, fixed part; *B*, rotatable body.

Figure 13 is a schematic drawing of a side entry rod for double tilting (± 45°, ± 5°) and furnace heating of the specimen up to 600 °C. The specimen may also be deformed with a symmetrical, hard type straining device provided with coarse (mechanical) and fine (electrical) controls. A strain gauge is mounted for quantitative measurements. The tensile device may

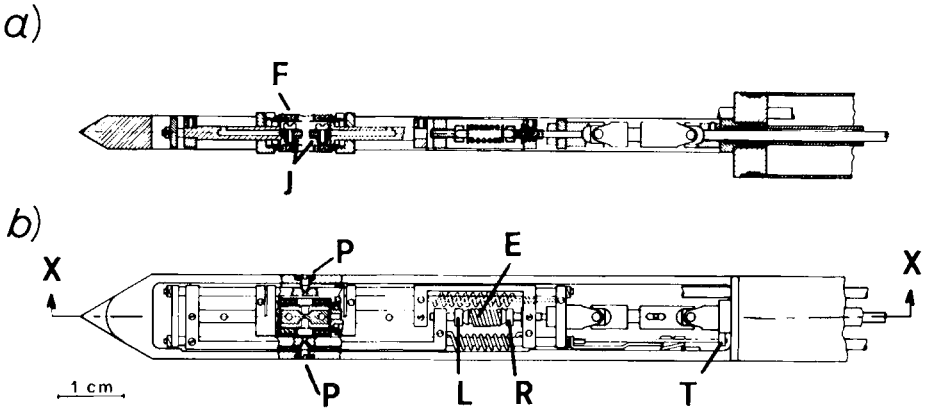


Fig. 13. – Rod for a high voltage electron microscope. It allows double tilting, straining, heating, gas inlet. *a*) is a section of *b*) through *XX*. *F*, furnace; *J*, jaws; *R*, *L*, right and left screws for mechanical straining; *E*, heater for fine strain; *P*, pivots; *T*, second tilt controller.

be operated hydraulically with little alteration and may easily be converted into a soft straining machine. Gas inlet facilities are also incorporated.

These devices are very delicate and their use should be restricted to special experiments, although it is easy, for routine work, to stop some of the degrees of freedom or to dismount unnecessary facilities.

8. Conclusions.

From the early days when specimens were clamped to the objective pole piece a lot has been learned, confidence has been acquired on the design of specimen stages and specialized workmanship has been trained. Lately more attention has been paid to the design of the object section of the microscopes and to the closely related objective lens in order to increase access, space and versatility; close contacts have also begun to take place between users of electron optical instruments and manufacturers in order to increase the efficiency of the instrumentation and to satisfy the real needs of the users. However much closer contacts are necessary and we cherish the time when, reversing the present criteria, the electron microscope will be made to fit a predesigned universal specimen stage containing the facilities of a laboratory for specimen treatments.

REFERENCES

1) G. Lucas, H. Phillips and P. W. Teare: *Journ. Sci. Instrum.*, **40**, 23 (1969).
2) J. Leteurtre: private communication.
3) H. Bollmann: private communication.
4) U. Valdrè: *Nuovo Cimento*, **53** B, 157 (1968).
5) P. H. Harris and E. L. Thomson: *Journ. Sci. Instrum.*, **40**, 111 (1963).
6) R. S. M. Revell: *Proc. 6th Int. Congr. for El. Micr., Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 179.
7) U. Valdrè: *Journ. Sci. Instrum.*, **42**, 853 (1965).
8) U. Valdrè: *Proc. 6th Int. Congr. for El. Micr., Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 165.
9) G. V. Patser and P. R. Swann: *Journ. Sci. Instrum.*, **39**, 58 (1962).
10) U. Valdrè: *Journ. Sci. Instrum.*, **39**, 278 (1962).
11) E. C. Shaffer and J. Silco: *Proc. 5th Int. Congr. for El. Micr., Philadelphia 1962* (Academic Press, New York, 1962), vol. **1**, p. E-8.
12) P. R. Ward: *Journ. Sci. Instrum.*, **42**, 767 (1965).
13) J. C. Mills and A. F. Moodie: *Rev. Sci. Instrum.*, **39**, 962 (1968).
14) P. B. Hirsch, A. Howie, R. B. Nicholson, D. W. Pashley and M. J. Whelan: *Electron Microscopy of Thin Crystals*, Butterworths, London (1965).
15) J. F. Nankivell: *Optik*, **20**, 171 (1963).
16) Z. S. Basinski: *Proc. 5th Int. Congr. for El. Micr., Philadelphia 1962* (Academic Press, New York, 1962), vol. **1**, p. B-13.
17) B. Hudson and M. J. Makin: *Journ. Phys. E (Journ. Sci. Instrum.)*, **3**, 311 (1970).
18) J. L. Williams: communicated at the meeting on High Voltage Electron Microscopy, Harwell (G. B.) (April 1970).
19) U. Valdrè: *Proc. 7th Int. Congr. for El. Micr., Grenoble 1970* (Paris, 1970), vol. **1**, p. 131.
20) G. Browning: communicated at the meeting on High Voltage Electron Microscopy, Harwell (G. B.) (April 1970).
21) P. G. Merli and U. Valdrè: *Proc. 7th Int. Congr. for El. Micr., Grenoble 1970* (Paris, 1970), vol. **2**, p. 589.
22) H. G. F. Wilsdorf: *Rev. Sci. Instrum.*, **29**, 323 (1958).
23) B. Lehtinen, E. Broberg and L. Dahnè: *Journ. Sci. Instrum.*, **44**, 289 (1967).
24) H. Saka, T. Imura and N. Yukawa: *Journ. Phys. Soc. Japan*, **25**, 906 (1968).
25) J. Leteurtre: private communication.
26) P. J. E. Forsyth and R. N. Wilson: *Journ. Sci. Instrum.*, **37**, 37 (1960).
27) U. Valdrè: *Proc. EMAG Conf., Cambridge (G. B.) 1971*, in press.
28) P. R. Ward: *Journ. Sci. Instrum.*, **44**, 681 (1967).
29) P. R. Swann: private communication.
30) G. R. Piercy, R. W. Gilbert and L. M. Howe: *Journ. Sci. Instr.*, **40**, 487 (1963).
31) J. A. Venables, D. J. Ball and G. J. Thomas: *Journ. Phys. E (Journ. Sci. Instrum.)*, **1**, 121 (1968).

32) U. VALDRÈ and M. J. GORINGE: *Journ. Phys. E (Journ. Sci. Instrum.)*, **3**, 336 (1970).
33) U. VALDRÈ and M. J. GORINGE: *Proc. 7th Congr. Ital. Soc. El. Micr., Modena 1969* (S.I.M.E., Milan, 1971), p. 179.
34) K.-J. SCHULZE and G. SCHIMMEL: *Proc. 6th Int. Congr. for El. Micr., Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 173.
35) C. COLLIEX: private communication.
36) N. KITAMURA, O. N. SRIVASTAVA and J. SILCOX: *Proc. 6th Int. Congr. for El. Micr., Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 169.
37) M. J. GORINGE and U. VALDRÈ: *Proc. 4th Eur. Reg. Conf. on El. Micr., Rome 1968* (Rome, 1968), vol. **1**, p. 41.
38) E. M. HÖRL: *Rev. Sci. Instrum.*, **39**, 1027 (1968).
39) G. POZZI and U. VALDRÈ: *Proc. 4th Eur. Reg. Conf. on El. Micr., Rome 1968* (Rome, 1968), vol. **1**, p. 355; *Phil. Mag.*, **23**, 745 (1971).
40) J. A. VENABLES and D. J. BALL: *Journ. of Crystal Growth*, **3-4**, 180 (1968).
41) R. E. HARTMAN and R. S. HARTMAN: *Proc. 6th Int. Congr. for El. Micr., Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 159.
42) U. VALDRÈ, D. W. PASHLEY, E. A. ROBINSON, M. J. STOWELL, K. J. ROUTLEDGE and R. VINCENT: *Proc. 6th Int. Congr. for El. Micr., Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 155.
43) F. C. S. M. TOTHILL, W. C. NIXON and C. W. B. GRIGSON: *Proc. 4th Eur. Reg. Conf. on El. Micr., Rome 1968* (Rome, 1968), vol. **1**, p. 229.
44) D. N. BRASKI, J. R. GIBSON and E. H. KOBISK: *Rev. Sci. Instrum.*, **39**, 1806 (1968).
45) R. D. MOORHEAD and H. POPPA: *Proc. 27th Annual EMSA Meeting, St. Paul, Minn.*, C. J. ARCENEAUX ed. (Baton Rouge, 1969), p. 116.
46) A. BARNA, P. B. BARNA and J. F. PÓCZA: *Vacuum*, **17**, 219 (1967).
47) R. K. HART, T. F. KASSNER and J. K. MAURIN: *Proc. 6th Int. Congr. for El. Micr. Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 161.
48) U. VALDRÈ, E. A. ROBINSON, D. W. PASHLEY, M. J. STOWELL and T. J. LAW: *Journ. Phys. E (Journ. Sci. Instrum.)*, **3**, 501 (1970).

Image Recording with Semiconductor Detectors and Video Amplification Devices

1. Image recording with semiconductor detectors.

K.-H. HERMANN, D. KRAHL, A. KÜBLER, and V. RINDFLEISCH

Siemens A. G. - Berlin and Karlsruhe, Germany

1'1. Introduction.

Since the beginnings of the electron microscopy, the microscopist has become used to recording the image information, given by the current density distribution in the image plane, on the photographic plate. The photographic plate is in fact an almost ideal recording means for electron images; it can resolve a large number of image elements and is at the same time so sensitive that for the blackening of image elements within the range of the limiting resolution of the plate only a few electrons are sufficient, which means that the plate resolution is partly determined by electron statistics.

The photographic plate has on the other hand some disadvantages which are disturbing in various applications.

1) The image information is not immediately available. This, for example, means that for the adjustment of the microscope (focusing, astigmatism) we must depend on the system: final image screen—tenfold binocular magnifier—eye, which in some respects is overcome by the photographic plate. For this reason also the minimisation of electron bombardment damage in sensitive specimens is difficult.

2) The quantitative measurement of current densities encounters difficulties especially when an intensity range of several magnitudes must be measured.

This latter limitation, a result of the limited blackening extent of the plate, is especially disturbing in structure analysis by means of electron diffraction.

In these lectures we will describe two recording methods, which have been developed taking the above disadvantages into account: I) the measurement of the current density at the highest image resolution by means of semiconductor detectors and II) the conversion of the electron image into a video image. We will describe the arrangement of the devices, show some possible applications and especially examine the efficiency and principal limitations of both methods. It will be interesting to compare both methods with each other. It will show that both methods favourably serve as supplements for each other.

1'2. Measuring device with semiconductor detectors.

To begin with, let us survey the historical development, which led to the use of semiconductor detectors.

When analysing uranium minerals, H. Becquerel discovered in 1896 a previously unknown radiation. With the help of electric and magnetic fields this radiation could be analysed into three components. In the following years and decades the physicists aimed at a more exact separation and discrimination of the components of the Becquerel radiation. To achieve this, numerous instruments have been developed. The first instrument of this type was the « ionisation chamber », which was used for the detection and measurement of ionising radiation. A disadvantage of that apparatus lies in the small stopping power of the filling gas for weak ionising radiation. For that reason it was tried at an early stage to fill the ionisation chamber with denser agents than gases. Thus suitable crystals such as diamond, AgCl, AgBr, KCl, LiF, NaI, were equipped with thinly evaporated electrodes, to which a voltage was applied. If now ionising radiation is directed into the counter, it ionises the atoms of the crystal. In insulating crystals, charges (*i.e.* electrons) are set free, they drift through a field of some kV/cm to an electrode and cause there a charging pulse, which can be measured afterwards.

Difficulties in the production of these crystal counters and the statistical limitation of the resolution, especially for « low energetic » radiation below 500 keV, have considerably diminished their importance.

By statistical resolution we mean the limitation of the resolution ΔE, *i.e.* the possibility of still separating ionising radiation with an energy dif-

ference ΔE. It is given by the equation

$$\Delta E/E = 1/\sqrt{N}$$

where N is the number of electron processes in the counter, *i.e.* the number of electrons produced. In NaI counters one needs for instance an energy of 700 eV in order to form a photoelectron. A 100 keV electron will therefore form about 140 electrons in the counter. Thus we obtain a statistical resolution of $\Delta E = 8.3$ keV.

The way for a change was opened when at the end of the forties the insulators were replaced by semiconductors. The advantage of these materials lies above all in the small energy of formation needed for a charge carrier pair and the resulting higher statistical resolution. The statistical resolution of a Si detector can serve as a comparison to the afore-mentioned value of 8.3 keV for the NaI counter: with an energy of formation of only 3 eV we obtain a statistical resolution of 0.66 keV when using 100 keV electrons.

With these semiconductor detectors counters became available, which allowed simple recording and analysis of radiation with high statistical resolution. McKay [1] first described in 1949 a Li-drifted germanium detector. However, still about 10 years had to pass before industry developed components of commercial size and manufactured them with the necessary reliability.

Highly pure silicon or germanium serve as base material. Compared to Si, Ge has the advantage of greater charge carrier mobility; but for the reduction of thermally produced charges, *i.e.* for the reduction of noise effects it must be cooled to liquid hydrogen temperature. For this reason attempts were made to compensate these positive free charge carriers by diffusing lithium into the detector materials. To do this Li is diffused through the surface under the influence of electric fields at temperatures between 50 °C and 250 °C. But after turning off the field, a back diffusion starts, so that these drifted Ge crystals too have to be kept below -70 °C. In this respect Li-drifted silicon, which remains stable even at room temperature, behaves more favourably.

One has to distinguish between two methods of production, which lead to two different types of counters, namely diffused counters and boundary layer counters.

a) In diffused counters a strong asymmetrical *p-n* transition is produced by diffusing a counter doping of high concentration, mostly phosphorus,

into *p*-conducting material at low depths ($\leqslant 1$ μm). When a voltage is applied this field zone is extended and forms the recording region for the incoming radiation.

b) In boundary layer counters the barrier layer is produced by a surface charge on a gold film, which has been evaporated on *n*-conducting silicon. Differential, or dE/dx counters also belong to this type of counter. They have very thin contact faces on very thin Si base material. The base material thickness is kept small in comparison with the penetration range of the radiation.

In operation, diffused counters are to be preferred to boundary layer counters. They are not affected by external influences, for example, humidity and touching. If necessary it is even possible to clean their surfaces.

Let us now see how such counters can be used as measuring elements for image recording and current density measurement in the electron microscope.

To begin with, we will consider the mode of operation of such a detector. If a voltage is applied to a diffused counter, the boundary layer is expanded proportionally to the square root of the applied voltage and to the specific resistance. For 100 V and 10 000 Ωcm *n*-silicon we obtain a field thickness of 500 μm. If a particle enters into the field, electron-hole pairs are formed. The electric field separates the electrons from the positive holes and decreases the probability of recombination. The electrons drift in the electric field and produce a charging pulse. If a radiation particle passes right through the boundary layer and enters deeper into the *p*-region, it also produces charge carrier pairs, but these recombine much quicker owing to the absence of the electric field. Only a few charges from this zone reach the boundary layer and contribute to the charging pulse.

Although the charging pulse still grows with increasing particle energy, the linear relation between pulse amplitude and particle energy is lost. For this reason the zone thickness is adjusted to the particle energy by choice of a blocking voltage. For high energy radiation a limitation is set by the breakdown voltage of the semiconductor. The charging pulse is collected from an input resistance and fed to a pulse amplifier. If the amplifier is fast enough to amplify pulses of only a few microseconds duration, individual electrons, which enter into the field, can be detected and counted. Some requirements are placed on the amplifier with regard to background noise, which however will not be discussed here.

The basic circuit shown in Fig. 1 is used for the current density measurement of the electrons. Apart from the components already mentioned, it

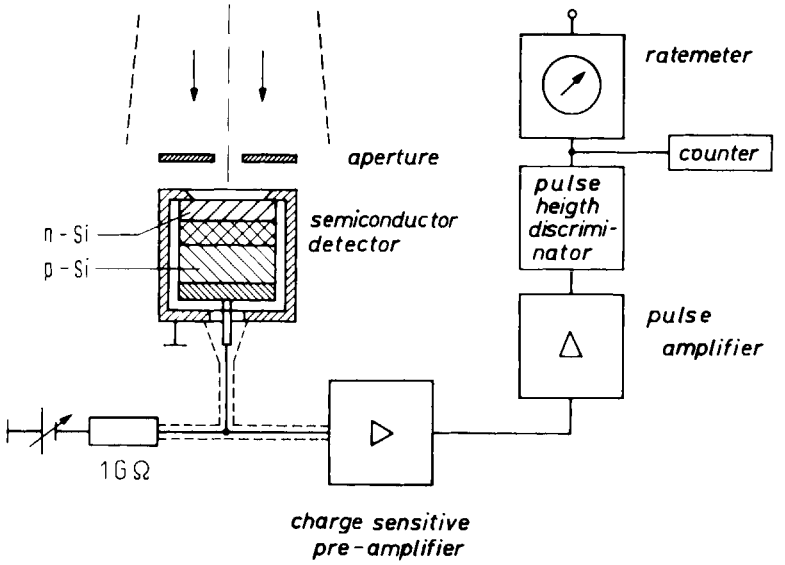


Fig. 1. – Semiconductor detector for the measurement of the electron current.

also contains a rate-meter, a counter and a pulse height discriminator, all well-known units in radiation measuring techniques. The discriminator is important for several reasons. As a differential discriminator, it enables us to pick up the complete pulse height vertical distribution. For bombardment with 80 keV electrons the result of such a measurement is shown in Fig. 2. From this it follows that, apart from the peak caused by the electrons, a noise band is to be found at small pulse heights, which would considerably adulterate the measurement. But if the discriminator is operated as integral discriminator, we can block these interference pulses. The threshold of the discriminator has been indicated in the figure with a dotted line. The semiconductor detector is thus capable of counting practically each high energy incoming electron, with which the theoretical limit of sensitivity of this measuring method has been reached. From the number N of electrons, which is counted within a measuring time T, there results a current density j according to the equation

$$j = eN/TF\,, \tag{1}$$

where F is the area of the detector which is being bombarded, N the number of electrons, e the elementary charge and T the measuring time.

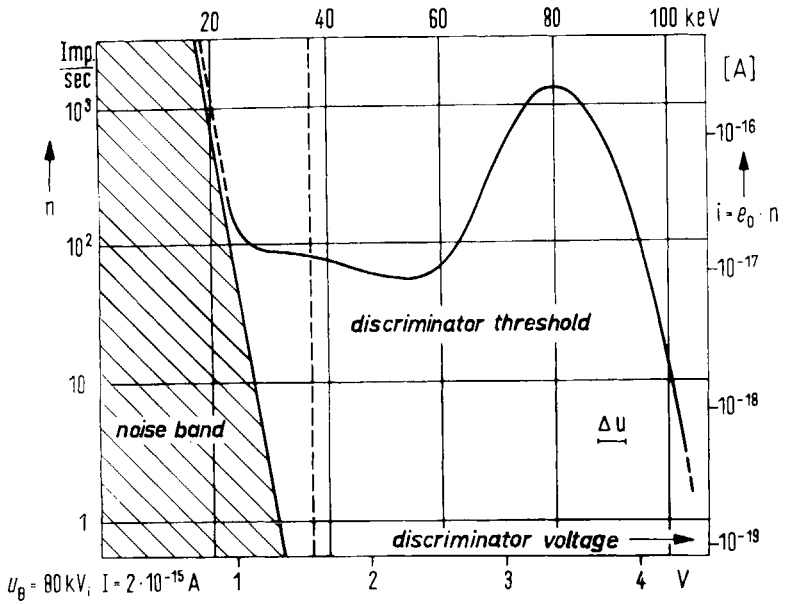


Fig. 2. – Pulse-height distribution of a semiconductor detector.

It can already be seen here that the detector is capable of measuring small current densities if only the measuring time is made long enough. But it can also be calculated that within the range of the usual current densities the measuring face F, which normally is 250 mm^2, must be highly reduced by an aperture, since the circuit, having a dead time of some μs, cannot detect electrons coming in at close time intervals. Therefore an aperture in front of the detector is necessary. In order to be able to adjust to a wide range of current densities and lateral resolutions, it is advantageous to plan an aperture changing device. Figure 3 shows the device which is installed in a viewing window of the Elmiskop 101 and can be put into the path of rays as a complete unit. A changing mechanism, which is shown disassembled on the right of the figure, permits us to bring in front of the detector by means of two controlling systems, either a number of various round apertures of diaphragm or a slit, whose width and length can be varied.

If only a few single points in an image are to be measured, the necessary shift of image can be carried out by adjusting the specimen stage. If however linear or area-like distributions in images or diffraction patterns are to be measured, it is particularly advantageous to undertake the shifting by means

of an electromagnetic deflection system. Within the intermediate lens the Elmiskop 101 possesses an electromagnetic stigmator, which accomplishes this in connection with a suitable circuit. A control system, which has been taken from our X-ray microanalyser, permits a line shifting in any direction as well as a linear shifting of the image or of the diffraction pattern.

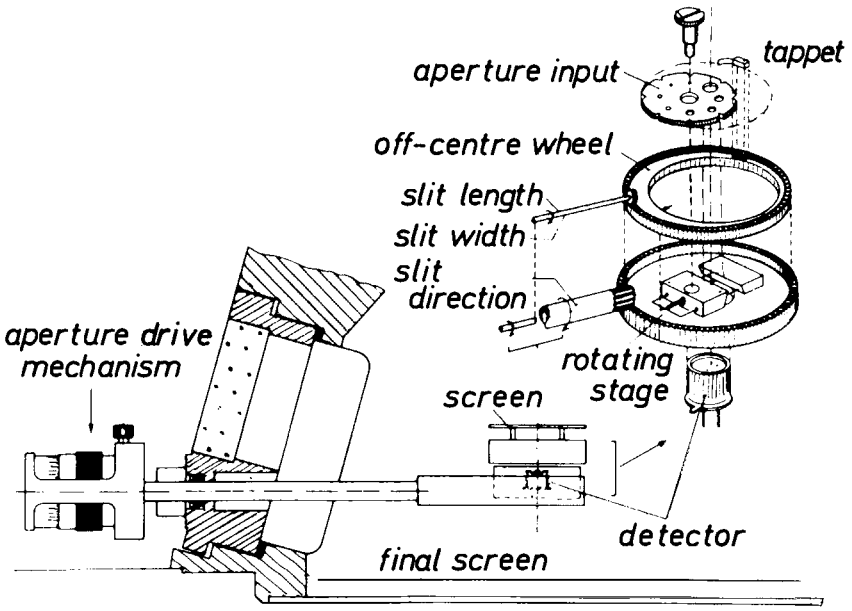


Fig. 3. – Semiconductor detector with variable aperture.

The block diagram of the whole device is shown in Fig. 4. The deflection system and the semiconductor detector are indicated with its variable aperture in the electron ray diagram of the microscope. The right-hand side of the Figure contains the various electronic components, which are of use for the electrical control of the deflection system and for the recording of the measured intensity values. Different recording systems are used, depending on whether we are concerned with the measurement of distinct elements, the measurement along a line or a two-dimensional system. In the case of element measurements, the image can be brought into the desired position with the help of x_0, y_0 adjusters and the electrons can be counted with the counter or indicated by the rate-meter.

We would like to just mention that the values can also be written down with a recorder and that even a control circuit is thinkable, which is fed by a computer via a digital-analogue-converter, so that a programmed series of

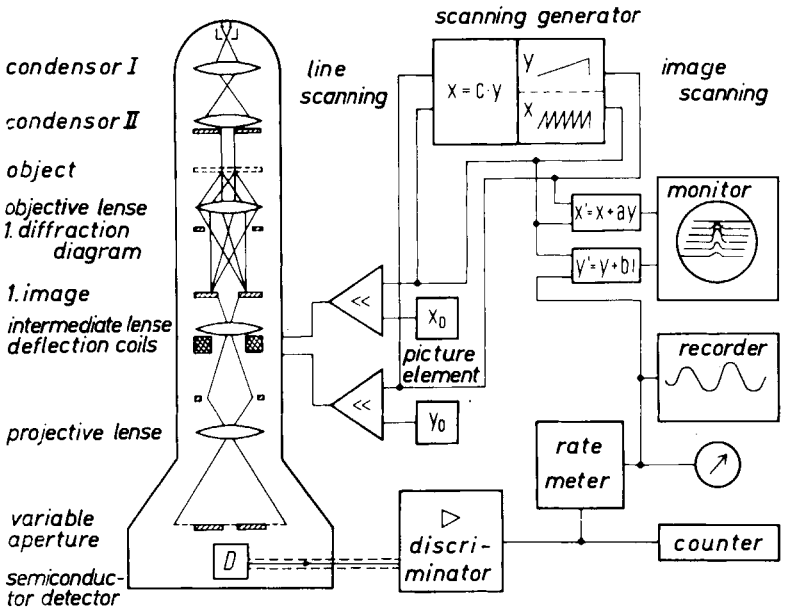


Fig. 4. – Recording method in image current density measurement.

intensity elements, for example the spots of a single crystal diffraction pattern, can be measured. These steps, however, have not been undertaken by us as yet.

In case of line measurements, the deflection system is controlled by a saw-tooth generator. The values are noted by a recorder. For the recording of the large number of measurement values, which are obtained when scanning surfaces line by line, we use an oscillograph. It records in synchronism with the control of the deflection system in the line scan. The measured mean value is superimposed on this line scan as a vertical deflection. Embossed images are obtained, in which on account of an additional horizontal displacement of the lines a perspective effect is produced. Figure 5 shows such recordings of a single crystal and a Debye-Scherrer diagram; on the right-hand side are shown segments taken from the complete diagrams reproduced on

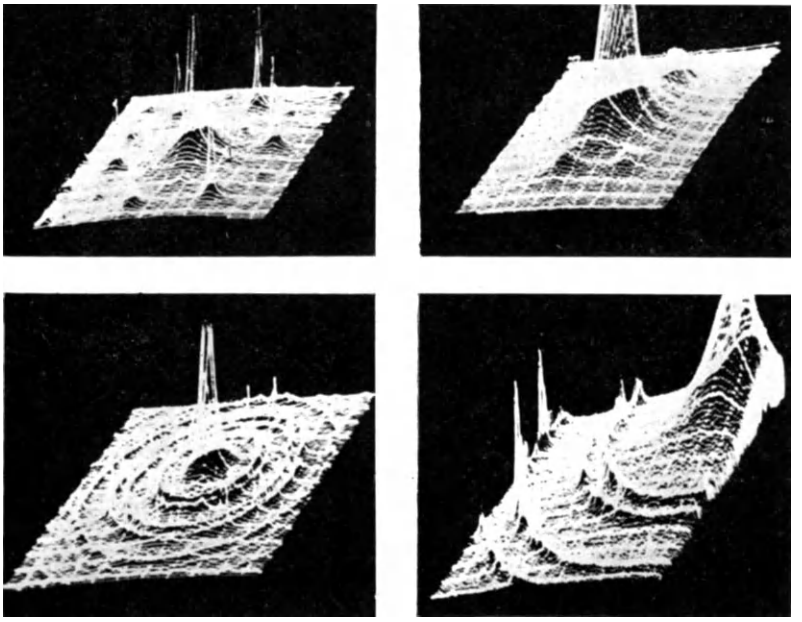


Fig. 5. – Areal intensity recording of selected-area diffraction patterns.

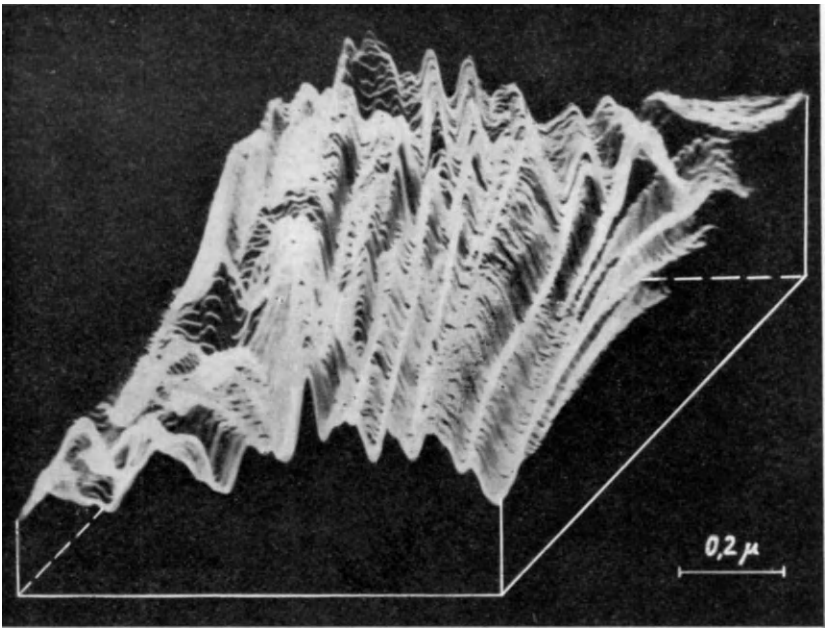


Fig. 6. – Areal intensity recording of extinction lines in gold.

the left-hand side. The images contain quantitatively the current density distribution. Figure 6 shows the recording of extinction lines in gold.

Let us now discuss the most important advantages of this method, but also its disadvantages and limitations. The possibilities of application will thus be immediately evident.

1) The measuring values are available at once, that is either as analogue or digital values.

2) The measuring range covers more than five orders of magnitudes. In this connection, which above all is of interest in the measurement of diffraction patterns, a clear superiority exists in comparison with the photographic plates. In order to profit fully from this characteristic, a logarithmic range of the rate-meter is advantageous. Figure 7 shows such a logarithmically recorded current density distribution in a diffraction pattern of a gold film.

3) If the pre-aperture has been chosen sufficiently small, the lateral resolution can be even better than the photographic plate. Figure 8, which shows the current density distribution in Fresnel fringes, does not as yet represent the optimum performance which could be reached. But it also shows that very small contrast effects can be measured.

On the other hand these favourable characteristics are associated with a disadvantage, which strongly limits the possibilities of application of this measuring method. As it is known, the electrons do not bombard the selected area of the semiconductor detector in regular time intervals, but are subject to accidental fluctuations in their bombardment frequency. According to the statistical laws, a measuring error ε has to be taken into account during each single measurement, which, on average, is given by the number N of electron processes

$$\varepsilon = \Delta N/N = 1/\sqrt{N}\,. \tag{2}$$

If the current density j_0 in the object plane, the measuring time T and the measuring surface relating to the target F_0 are inserted in (2), we obtain:

$$\varepsilon = \Delta j_0/j_0 = \sqrt{e/j_0 \cdot T \cdot F_0}\,, \tag{3}$$

or

$$T = e/j_0 \cdot F_0 \cdot \varepsilon^2\,.$$

It becomes obvious that a small statistical error can be obtained only with long measuring times. This may possibly be tolerated in image element

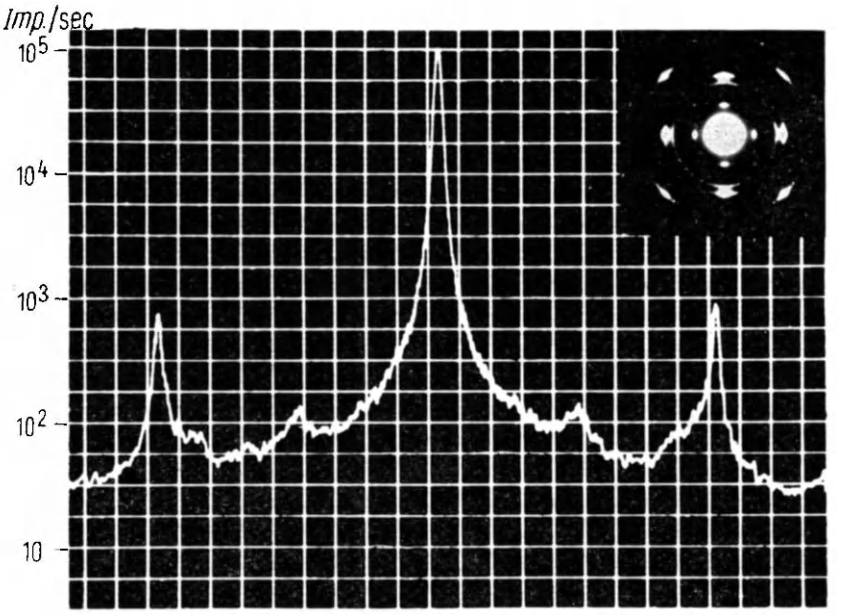


Fig. 7. – Intensity distribution for the diffraction pattern of an epitaxial gold film.

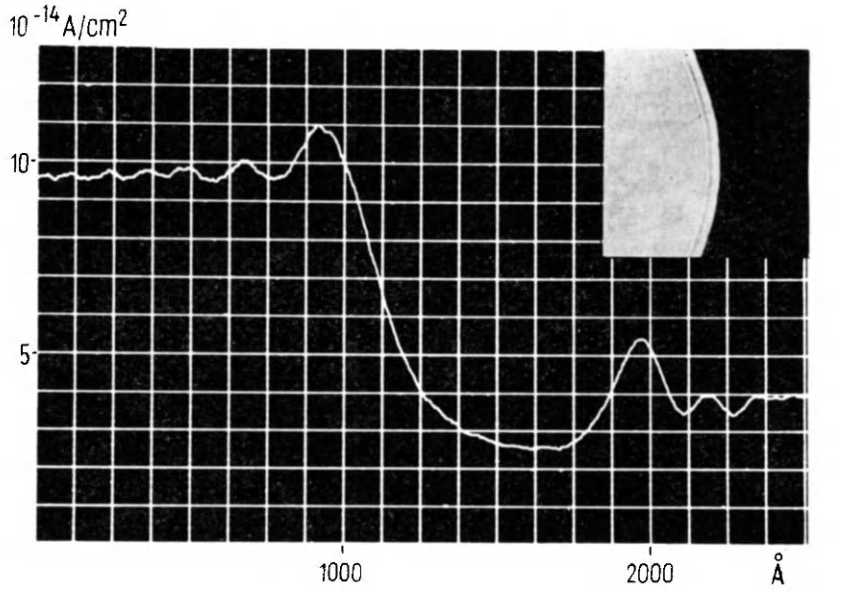


Fig. 8. – Intensity distribution through Fresnel fringes.

measurements. In line measurements already a considerable limitation of the velocity of scanning can be noticed. If, in the interest of a desired resolution δ, a dimensioned measuring area $F = \delta^2$, with a prescribed measuring error ε is to be measured, the velocity v of scanning may not exceed the value:

$$v = \delta/T = \delta \cdot j_0 \cdot \delta^2 \cdot \varepsilon^2/e = (j_0/e) \cdot \delta^3 \cdot \varepsilon^2 . \tag{4}$$

Figure 9 shows this situation quantitatively. If we wish to record with a measuring error of 1 % and a resolution of 10 Å, only a distance of 100 Å can be scanned within 1 second. A corresponding value of the damping time constants of the amplifier must be adjusted on the rate-meter. In a line-like scanning of a bigger image area the situation is even worse.

The recording time T of the whole image has been plotted in Fig. 10 in a different mode of demonstration against the resolution. The number of image elements is taken as parameter. The figure clearly shows what long times are reached if some requirements are demanded from the resolution and the image field.

We wish to note that compared to television, the measuring time for an image of comparable number of image elements is about (12÷16) h.

Here the most disagreeable limitation of the measuring method becomes obvious, from which it follows that only specimens which are not too sensitive to electron bombardment are suitable for area-like measurements. Naturally, any object contamination must be avoided by an effective cooling system of the specimen surroundings. Already here the advantages of storage systems can be seen, as offered by the photographic plate and the charge sensitive layers in TV camera tubes, with regard to the recording of complete images. In this respect the photographic plate is very superior, since it is able to record the electrons of all image elements at the same time. During the exposure time of a plate the semiconductor detector has measured only one image element.

In practice our detector will be used successfully only if the desired quantitative information can be obtained with as few measuring points as possible. This is the case in contrast measurements of single image elements or longitudinal lines, which, for instance, in the case of amorphous objects, can give us information about the mass density distribution in the specimen. In crystal specimens, extinction lines or variations in intensity in the vicinity of dislocations or small area-like defects can be measured very exactly. The intensity measurement in diffraction patterns has already been discussed. No doubt, the method can here be used successfully. As already indicated,

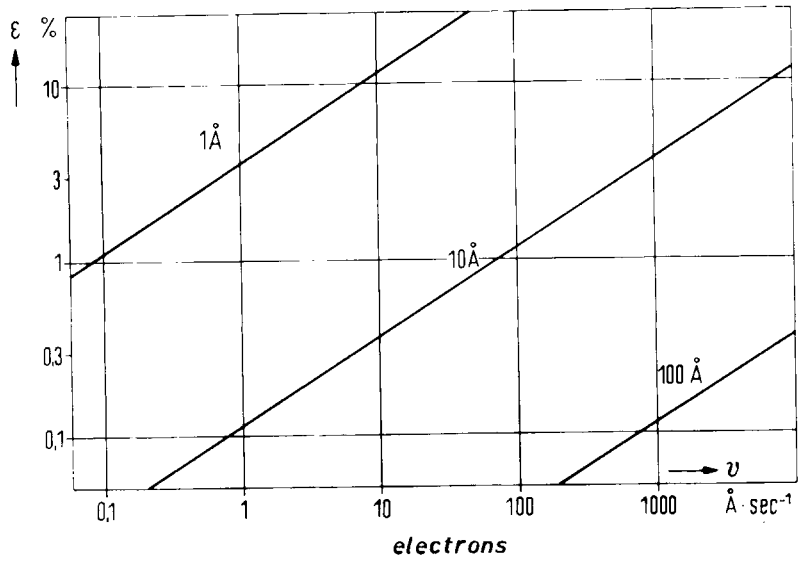


Fig. 9. – Statistical measuring error ε as a function of the recording velocity.

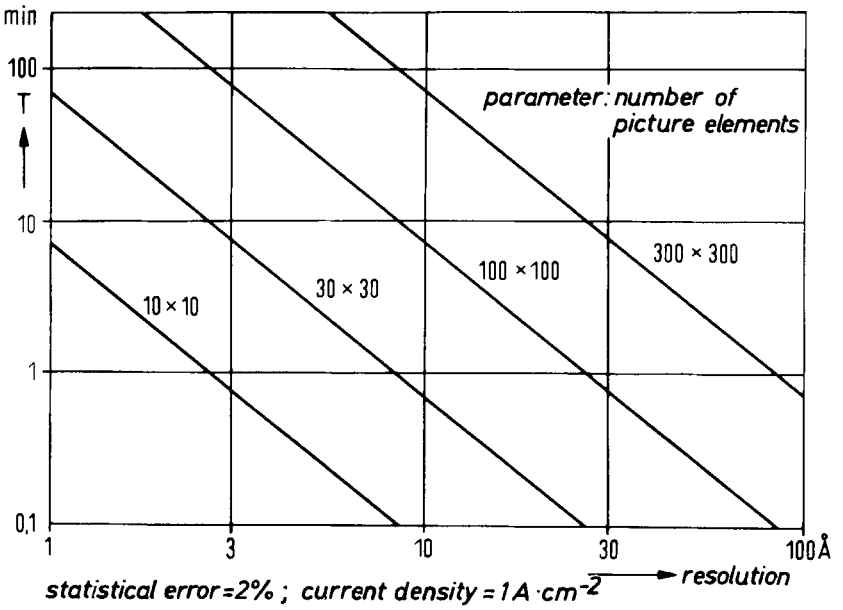


Fig. 10. – Recording time T at linear scanning of a square field.

a considerable shortening of the measuring time can be achieved if only individual diffraction spots can be selected by a programme control. In this regard the method can be extended to the immediate further processing of the digital values obtained in a computer, as has been customary already for some time in the field of X-ray analysis.

The area-like measurement and recording of the measured intensity values, which can produce quite impressive images, will most probably remain restricted to special cases on account of its long measuring times. In particular it can be maintained that this recording has *nothing* to do with the problem of image amplification. The measuring times which would be needed to record the whole image information on a video monitor are completely unrealistic. The method would correspond approximately to the tests with the Nipkow disc, carried out at the beginning of the video technique. Only if for each single image element an individual detector could be on hand (that means, if no image information would get lost during the recording time), could one count on shortest recording times. Although the technical realization of these ideas is not as utopian as originally might be thought, we shall take this only as the principle which forms the basis for the video image amplification: the simultaneous detection of all image elements without losing any electron.

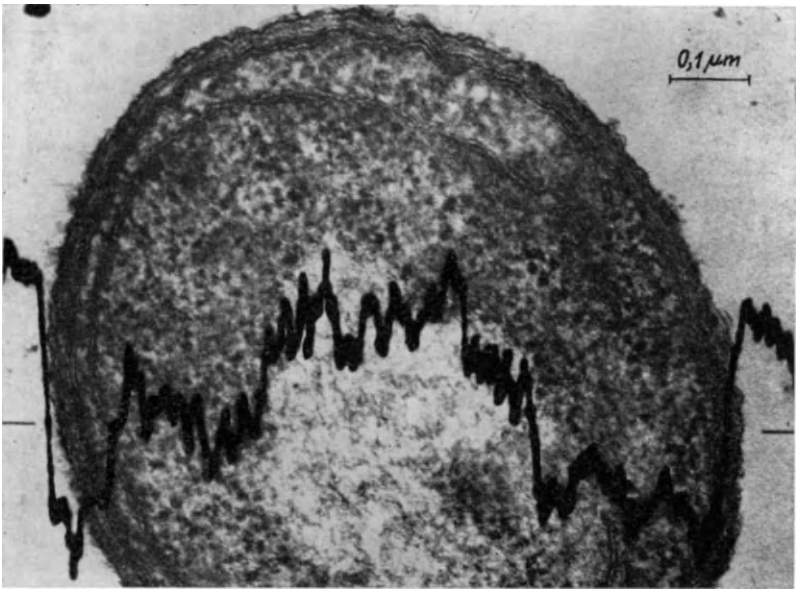


Fig. 11. – Transparency variation in a section of Bacterium coli.

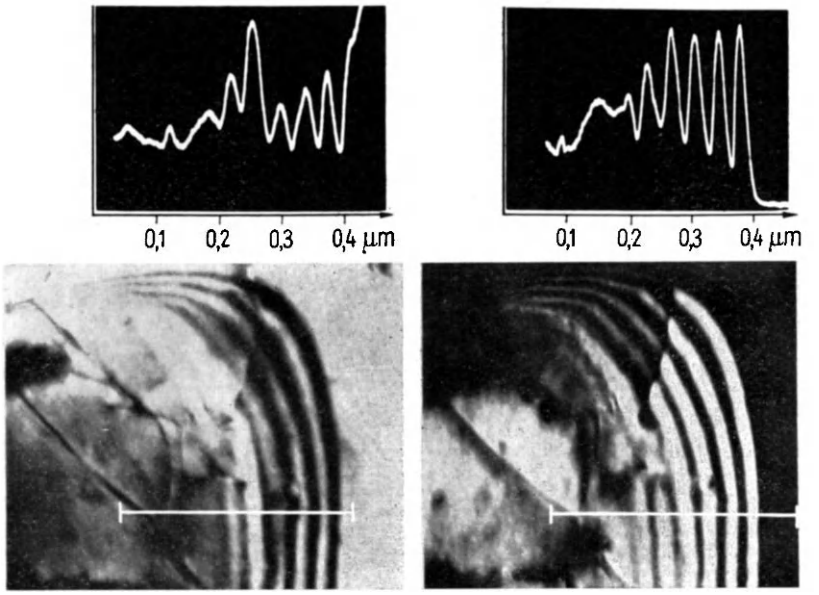


Fig. 12. – Intensity distribution in a grain boundary of a CoFe alloy. Left: bright-field image; right: dark-field image.

Finally, Fig. 11 and 12 show two examples of line recording along the directions marked on the pictures; Fig. 11 shows the variation in transparency in a biological specimen and Fig. 12 shows the variation in intensity in a bright and a dark field image of a grain boundary in a CoFe alloy.

REFERENCES (Section 1)

1) K. G. McKay: *Phys. Rev.*, **76**, 1537 (1949).

2. Image amplification with television methods.

K.-H. HERMANN, D. KRAHL, A. KÜBLER, K.-H. MÜLLER, V. RINDFLEISCH
Siemens A. G. - Berlin and Karlsruhe, Germany

2'1. Introduction.

The principal aim of image amplification with TV methods lies, first of all, less in quantitative detection of the image information but rather in amplifying the brightness of the image to the extent that it can be observed without the necessity for dark adaptation of the human eye, which under certain circumstances requires up to 30 min. This means that the eye can be used without the usual deterioration of its resolving power at small image brightnesses.

In the case of the fluorescent substances, which are used for final image screens in the electron microscope, current densities in the final image plane of about $(10^{-10} \div 10^{-9})$ A/cm^2 are necessary in order to produce images, which are observable without the disturbing dark adaptation. Already with electron optical magnifications of 50000 and upwards, the above-mentioned current densities in the final image plane can hardly be reached. The situation gets even more critical, when objects with little bounding energy, for example plastics, catalists or macromolecules are to be observed. Depending on the bond type, such objects can stand only current densities in the specimen plane of $j_0 \sim (10^{-3} \div 10^{-4})$ A/cm^2. In electron optical magnifications of 100 000, final image current densities of only $(10^{-13} \div 10^{-14})$ A/cm^2 are obtained; such images are not accessible to observation by the human eye.

With the aid of image amplifying devices, the focussing and astigmatism correction in high electron optical amplifications are facilitated and observations of electron sensitive objects made possible.

The television technique offers particularly favourable conditions, since television electronics render possible additional techniques, for example the accentuation of contrast, but especially since all these methods have one important common characteristic: they have a storage target, on which the signals of all image elements are stored at the same time, in order to be then scanned by a reading electron beam one line after the other, line by line, and to be converted into the electric signal, the so-called video signal.

Here will be shown the advantage in comparison with the electron detector, in which at any instant only one image element, which was selected by the aperture, detected the incoming current; the electrons, bombarding the other image elements, are not recorded.

2·2. **Method.**

For the amplification of the image brightness mainly three methods have become known so far, which have also been used in electron microscopes:

a) The method with direct converting layers, which makes use of the « electron bombardment conductivity » effect, that is of the conductivity in insulators and semiconductor layers produced by electron bombardment with primary electrons, *i.e.* of the direct conversion of the electron image into a charge image on the target.

b) The use of commercial television camera tubes of various types: vidicon, plumbicon, orthicon, SEC tube. Here a conversion of the electron image into a light image via a luminous screen takes place, since commercial television camera tubes can only detect light images.

c) A third method makes use of multistage intensifiers, which however do not have the advantage of storage targets.

We shall now concern ourselves in detail with the first two methods.

2·2.1. *Direct converting layers.* – It is known that for example, in a dielectric a conductivity can only be produced by photons, if the photon energy becomes equal to one of the bands of the absorption spectrum. Fast electrons, as we find them in the electron microscope, can on the contrary induce a conductivity through ionization effects. Studies in 1948 showed, however, only small effects of this kind. In 1951 Ansbacher and Ehrenberg [1] found a strong dependence of the conductivity in As_2S_3 under electron bombardment. The amplification factor, *i.e.* the relation of the current in the layer under electron bombardment to the incoming current of the primary electrons, reached maximum values of up to 40000. During the following period of time, other amorphous dielectrics (Se, Sb_2S_3, As_2Se_3, Al_2O_3, CdS) with similar characteristics were discovered and tested.

In 1958 Haine, Ennos and Einstein [2] published a paper concerning an image amplifier for an electron microscope, which was based on the principle of the EBC effect (EBC = electron bombardment conductivity). An

amorphous Se layer of 15 μm thickness was used as electron sensitive layer. The experimental arrangement was the following. A support ring was coated with a thin carrier layer of 6 μm Melinex foil, on to which was evaporated on both sides a thin Al layer, approximately 100 Å thick. The 15 μm Se layer was then applied to one side of this carrier. The free side of the Se is scanned with an electron beam and so charged to cathode potential.

The electrons induced in the layer by primary electrons are removed by an electric field which is built up with a bias voltage and leaves in the layer a positive charge distribution, the height of which is proportional to the incoming current density. When the target is read out with the scanning beam, the initial situation is restored again. To do this, the scanning beam sends electrons to the target. A charging pulse produces a voltage pulse on the input resistor, the video signal. Unfortunately this relatively simple and inexpensive method of image amplification has some considerable disadvantages, which have so far prevented such devices from being used in the microscope.

a) The amplification factor of Se (1000÷2000) is so low, that primary current densities of 10^{-10} A/cm^2 are necessary for obtaining good image quality, in order to set the image signal off against the background noise. But the amorphous structure of Se, which shows the EBC effect, starts to crystallize after a few minutes; a process which is stimulated by the incandescent light of the cathode. Thus the characteristics of the layer are greatly changed: on one hand the dark current increases considerably, which is noticeable through a worsening of the signal/noise ratio; on the other hand the transverse conductivity of the target increases remarkably. This causes the charge distribution to spread so that the storage effect of the layer worsens. The layers tested in our laboratory have had a lifetime of less than 30 min.

b) The arsenides, which possess such high amplification factors that true image amplification devices could be built with them (As_2S_3, 40000; As_2Se_3, 20000) show such a high transverse conductivity that the charge image spreads almost immediately, *i.e.* in 1/25 s, the duration for successive scannings.

c) Big residual charges, which remain in the target after the scanning, caused by incomplete discharge of the target, lead to smearing effects, which produce diffused images especially in the case of dynamic processes.

On account of all these difficulties it does not seem of any advantage to use this type of image amplification in the microscope, especially since

with commercial television tubes we have components at our disposal, where devices can be used which show much better characteristics, *i.e.* higher amplification factors and longer lifetime.

2·2.2. *Image amplification devices equipped with television camera tubes.* – There are different types of commercial camera tubes, which in principle are all more or less suitable for our purposes and have been used already. At first we will explain the principle on which these tubes work, then elaborate the fundamentals, which form the basis for the structure of the image amplification devices. The criteria, according to which we will have to estimate their efficiency and limitations, will be elucidated. Finally, we will concern ourselves in more detail with one device, which we consider to be particularly suitable.

A) *Principle of the camera tubes.* – The camera tubes always contain three necessary basic elements:

a) A photoelectric conversion layer, on which the light image is projected.

b) A storage target, which is connected to the photoelectric layer via an electron optical system and on which a charge image builds up, the latter being proportional to the light image.

c) A scanning device in the form of an electron probe, which with a scanning movement evaluates the load distribution in the storage target.

Vidicon tubes work on the principle of the outer photoelectric effect. The photoelectric conversion layer and storage target coincide in this type of tube. The storage target is built up of thin homogeneous semiconductor layers with a high electrical resistance (Se, Sb_2S_3), which have been coated on a light transmissive but conducting supporting layer, called the signal plate. The resistance of the semiconductor layer is proportional to the local brightness. The negative charge supplied during one scanning process partly flows off, before the next scan, according to the resistance of the image element. In comparison with the cathode the image element will have a positive potential. When scanning anew, the charge which has leaked away is replaced. A current pulse produces the video signal on the input resistor.

The photoelectric layer and storage layer are separated in the orthicon. Due to the inner photoelectric effect, photoelectrons are emitted from the photocathode when irradiated with light; the electron image is projected on

the storage layer as a result of electron optical projection. The electrons, which have been accelerated up to 2 keV, produce secondary electrons on the storage plate, which in standard orthicons consists of a special glass a few μm thick (in special types of orthicons Al_2O_3 is used); these secondary electrons are drained off by a fine net in front of the storage target. As in the vidicon, the target is discharged again by a scanning beam. However, here the signal is not the charging pulse, but that fraction of the electrons in the scanning beam, which, depending on the electron optics of the system, return on the same path and enters a secondary electron multiplier, being amplified there and used as signal.

A new type of tube, the SEC (secondary electron camera) tube developed by the Westinghouse Company ([3]), will now be described. It will be interesting to learn that the EBC effect, already discussed, is applied in this tube which as a commercial component shows its superiority in comparison with all methods of image conversion so far known.

The target of the tube contains an Al_2O_3 support film of thickness about 700 Å, a signal plate, which is coated on the support film (700 Å). The actual storage target with a density of only $(1 \div 2)\,\%$ of the normal density consists of highly insulating KCl. Its thickness is approximately $(10 \div 20)$ μm. As in other tubes, the scanning beam lifts the target to cathode potential, the electric field is built up by a bias voltage on the target. The target is then bombarded with these photoelectrons, which have been accelerated to a maximum of 7 keV in the converter of the tube. Secondary electrons produced in the KCl layer are drained off by the field and leave positive charges behind, owing to the small drift velocity. The amplification factors are naturally not as favourable as in the layers using the EBC effect mentioned before. For the production of a SE in the KCl, approximately 30 eV are needed, so that about 300 secondary electrons are produced by one 7 keV electron. About 30 % of the SE are lost on account of recombination, so that amplification factors of ~ 200 are to be expected.

As far as sensitivity is concerned the orthicon and SEC tubes are better by the factor 30 than the tubes of the vidicon type.

B) Basic structure of an image amplification device. – The basic structure of such a device is shown in Fig. 13. A transparent luminous screen converts the electron image into a light image, whereby an Al coating layer of $(100 \div 150)$ Å thickness ensures that the light originating in the screen is radiated only downwards, *i.e.* in the direction of the camera tube. This coupling link shown here contains mainly a light optical system with the help

of which the luminous screen image is projected onto the photocathode of the camera tube. But it can also contain active, *i.e.* brightness amplifying elements, for example intensifiers, as will be seen later on in our image ampli-

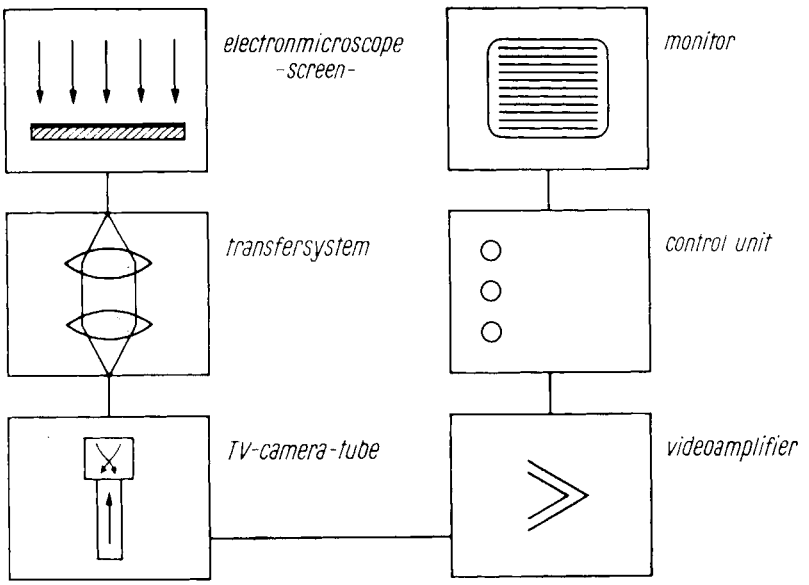


Fig. 13. – Schematic drawing of the TV system connected to the electron microscope.

fying device. As described earlier, the video signal which is produced in the camera tube, and is subsequently amplified in a wide-band amplifier, together with the corresponding synchronous signals, controls the recording beam of the monitor tube.

Which requirements have now to be demanded from such a device?

C) Basic requirements. – The general requirement, which we have to establish, is that, on its way through the television system, no image detail should get lost and that the image will be transmitted without distortion. In other words, no loss of information and no adulteration of the image information should occur.

A first basic requirement, which has to be met, is that the resolving power does not worsen. Especially the scale of reproduction between the luminous screen and the camera tube must be chosen in such a manner that the banded

structure of the television device (we use the 625 line system, standardised in Germany) is fine enough in comparison with the luminous screen. This requirement can only be met by a limitation of the transmitted image field. This limitation is smallest when the resolving power of the luminous screen is approximately adjusted to the resolving power of the camera tube, given by the banded structure. Since different commercial types of camera tubes have different photocathodes and target sizes, different optical magnifications are therefore necessary for optimum adjustment, and this can be reached by choice of the photooptical coupling links.

The next basic requirement concerns the capability of the device to project images at as small a final image current densities as possible. This capability may be impaired by background noise. A more exact analysis, which the theorists of television techniques undertook some time ago, shows, that television devices possess a background noise which originates from the input stage of the video amplifier, that is from the electron multiplier in image orthicon tubes. It is understandable that the image perceived on the monitor will be better the more the video signal dominates over the noise. For this reason it has proved to be useful to denote the image quality by the signal/noise ratio (*i.e.* $S =$ signal voltage to noise voltage). An image with $S = 30$ is to be considered as excellent, with $S = 10$ as good, and with $S = 2$ as still just usable.

In order to make the image signal as big as possible, it is now our task to undertake the coupling of the television device to the luminous screen by means of coupling links of high light transmissivity. For this purpose so-called tandem optics have proved to be useful. These are objectives of high aperture ratio which are used in pairs, and which have been focussed for infinity individually and which produce an image scale proportional to the focal length. By coupling the camera tube to the transparent luminous screen by means of fiberplates, which have a transmission of about (70÷80)% compared with 15% of the highest intensity tandem objectives, the sensitivity may be increased by a further factor 5. Assuming the use of such fiberplates, we have compared various types of devices by calculating their characteristic curve of signal to noise ratio as a function of the final image current density in the microscope and also by measuring them to a great extent. The results shown in Fig. 14 make it clear that the devices become more sensitive in the following order: vidicon, plumbicon, orthicon (tube for studio television), SEC tube (discussed later on), MgO orthicon (a high sensitive special type) and SEC tube with image intensifier connected in series. With the last combination we will concern ourselves in detail later on. The other tubes can

naturally all benefit from the 200 fold increase in sensitivity given by an image amplifier.

We will now try to answer the question to what extent an increase in sensitivity on an image amplifying device is convenient at all. It will seem

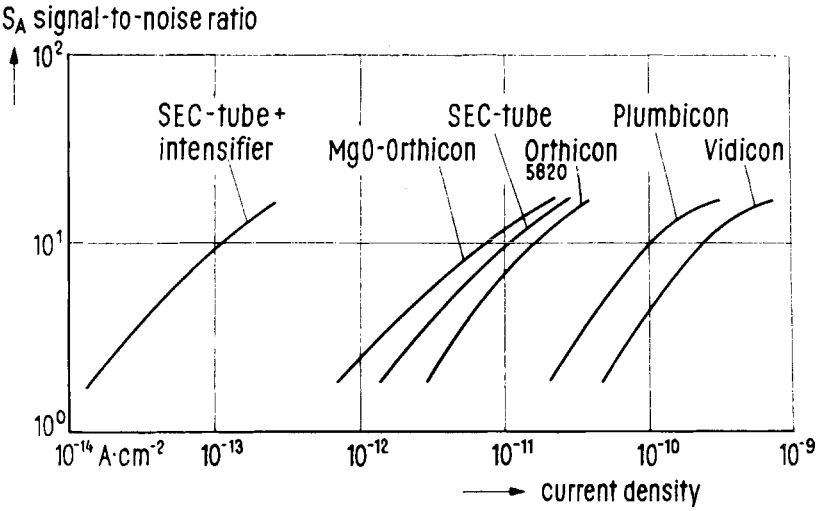


Fig. 14. – Signal-to-noise ratio of TV systems for electron microscopes.

plausible after the explanations in the first part of these lectures, that a natural limit will then be reached when the device is capable of making visible the signals of individual electrons per image element. Then the image quality is determined not any longer by the noise of the device but by the noise of the electrons of the microscope. For this we can also calculate a signal to noise ratio with the equation of statistics used before.

$$\Delta j_E/j_E = 1/\sqrt{N} = \sqrt{e/j_E T\delta^2},$$

or

$$S_Q = j_E/\Delta j_E = \sqrt{j_E}\, T\delta^2/e\,. \tag{5}$$

Here the symbols used are:

S_Q = signal to noise ratio subject to the electrons of the microscope (quantum noise),

T = the integration time, in our case 1/25 s. When observing the monitor with the human eye, the storage time of the eye must be inserted, which is about 0.2 s,

j_E = the final image current density,

δ = the resolving power relative to the final image plane, about 50 μm.

The equation given in (5) has been inserted as the dotted curve in Fig. 15. We can now say that the overall signal to noise ratio is determined by both noise levels, the amplifier (or video noise) and the quantum noise according

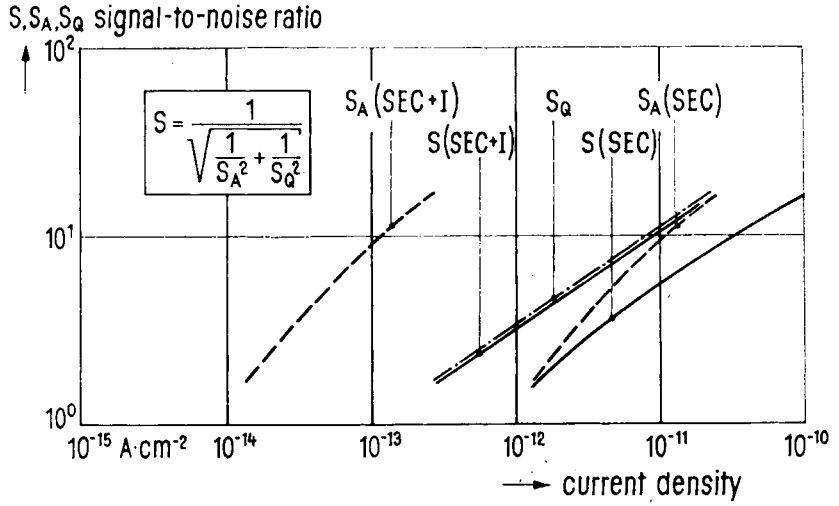


Fig. 15. – Signal-to-noise ratio of SEC systems connected to Elmiskop 101.

to the equation given in the diagram. Here the characteristic curves of the SEC tube with and without image amplifier are reproduced once more, as we have seen in the preceding figure. The over-all signal to noise ratio for the device with series-connected intensifier approaches closely the natural limit, which is given by the quantum noise of the electron beam of the microscope. The image quality is thus mainly determined by the quantum structure of the image forming electrons. A further increase in sensitivity would be meaningless, since no further information would be obtained. We can see from this, that a television image amplifying device can be looked at as being the more efficient, the closer its total characteristic curve is to the limiting curve of the electron noise.

For the device without image amplifier, the quantum noise and the amplifying noise are of the same magnitude, the image quality is thus determined by both components. This can be noticed in the total characteristic curve through the fact that it does not coincide with the natural limit, but lies below.

At Siemens we have decided to use the system of the SEC tube with image amplifier as an accessory instrument for the Elmiskop 101, since apart from the high sensitivity this system offers still other advantages. We use fibreplates for the photo optical coupling, which, as already mentioned, have the advantage of high light transmissivity, of short over-all length and of better transfer characteristics in comparison with tandem optics. Apart from the noise, the information to be transmitted is determined by the MTF of the television system (MTF = modulation transfer function). It is well known from the information theory of transmission systems, that small distances are not as good as bigger ones as far as the contrast is concerned, that is, that they are transferred with impaired contrast. If now the contrast drops below the physiological threshold, which should be assessed at about 10% contrast, depending somewhat on the background brightness, these structures are not resolved by the human eye anymore and the information to be transmitted gets lost.

Each of our transmission links has such a MTF. The facts, which form the basis for these characteristic curves, are different in the various components of the image amplifying system. In the case of the luminous screen the diffusion halo will mainly be responsible for the decrease in contrast. The MTF of the camera tubes is caused by the scanning mechanism. Structures with distances which are bigger than the scanning spot are modulated fully. If, however, the distances come within the area of the scanning spot, the modulation factors get smaller on account of edge smearing. The background brightness increases so that structures of this distance are not set off against the background anymore, and get lost as information. One assumes, that the electron optical magnification will be chosen in such a manner that the patterns to be transferred will not disappear in the transmitted image on account of an underevaluation of its contrast.

The modulation transfer function of the total system is formed by multiplying the individual MTF's. A decrease in the contrast at small distances has to be taken into account with each additional component in the transfer channel. Particular attention has to be paid to this fact, since images with a modulation of 100% are only very rarely offered to the image amplifying device. In the case of biological or highly resolved objects the contrast offered will hardly exceed 20%.

We have now covered the basic problems in detail and should turn to the practical design of such an image amplification device.

Figure 16 is a schematical drawing of the device we have adopted. The bottom plate of the electron microscope camera has been tightly vacuum

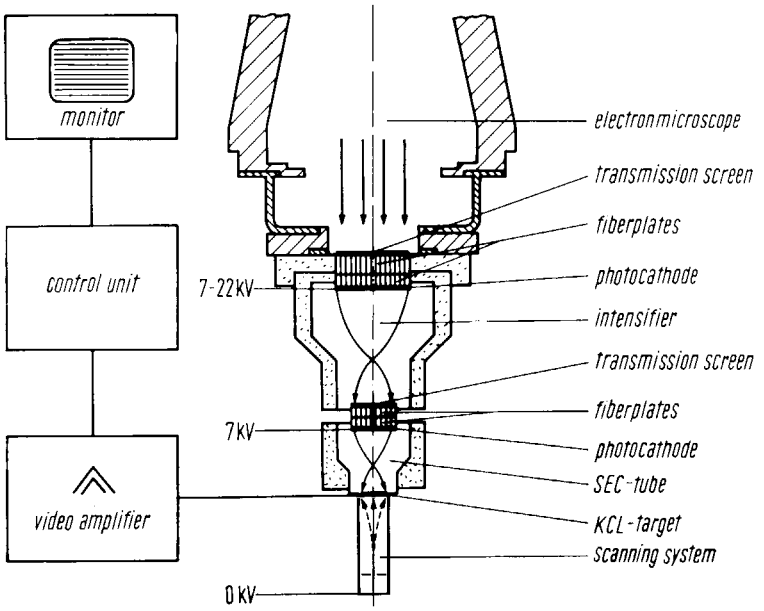


Fig. 16. – TV system with SEC tube and intensifier connected to Elmiskop 101.

sealed by a fiberplate, on the topside of which a transmission screen with coating layer has been installed. A single stage electrostatic image intensifier with fiberplate input has been flanged to the fiberplate. By means of its photocathode it converts the incoming light image into an electron image. On account of an acceleration of the electrons to a maximum of 15 kV and an electron optical reduction by a factor 2 (with which we fulfil the first basic requirement: the adjustment of the resolving power of the luminous screen and the camera tube) a light image is again produced at the output on the luminous screen, the brightness of which is already 200 times higher than that of the microscope screen. This light image is again transmitted to the photocathode of the SEC tube via fiberplates, which are component parts of the image amplifier and the camera tube. The electron image created is then accelerated to a maximum of 7 kV in the image converter and is

reproduced on the actual storage target at a magnification of about 1:1. Figure 17 is a picture of this device.

The high voltages, which are connected to the SEC tube and the image amplifier are of some importance. The 15 kV of the image amplifier has to

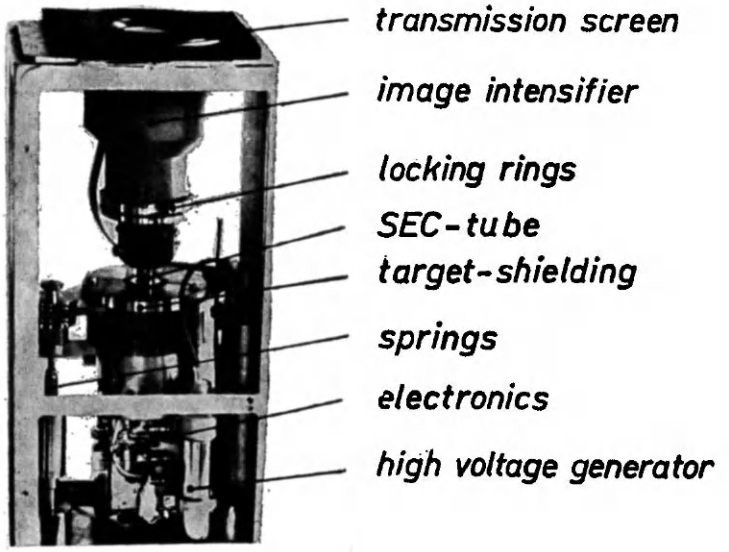


Fig. 17. – Assembly of the image intensifier.

be added to the 7 kV of the SEC tube, so that on the bottom side of the fiber plate of the microscope we obtain a voltage of 22 kV relative to ground, *i.e.* relative to the transparent luminous screen of the microscope. The breakdown strength is then a problem due to the high voltage, but can be mastered by a sufficiently thick fiberplate and an insulating ring. Conducting layers on the fiberplate ensure that a uniform distribution of the potential always exists. Both voltages can also be reduced during operation, which causes a decrease in the image amplification and image conversion effect. This procedure is important in order to deal with higher electron current densities, *i.e.* with higher image brightness, without overloading the device. The conversion characteristics of the device remain largely unaffected. As we have seen in Fig. 14, the characteristic curve of a television device covers only about one order of magnitude of current density. But since we want to cover several orders of magnitude, the shifting of the characteristic curve by means of the image intensifier voltage offers a valuable facility. One then operates

on the straight line part of the characteristic curve at all image brightnesses. We undertake this adjustment automatically in our device by forming out of the video signal a control signal for the high voltage of the image intensifier and for the image conversion part of the SEC tube.

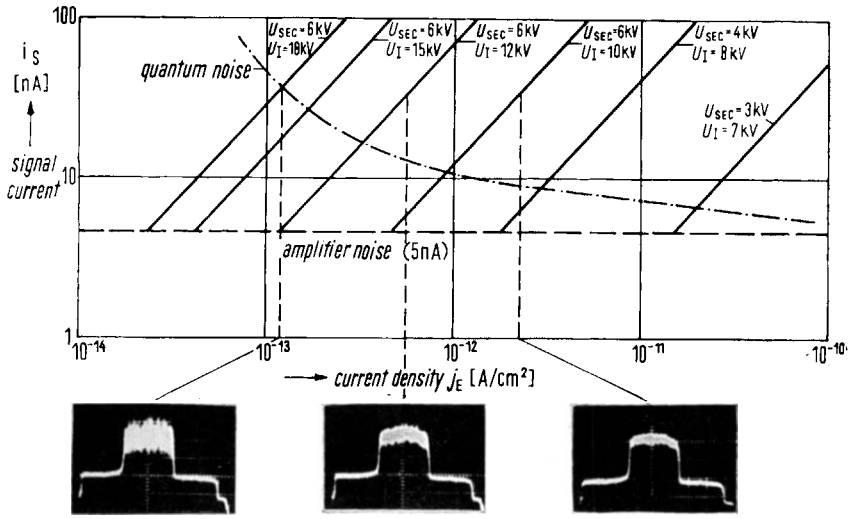


Fig. 18. – Signal current and quantum noise of the SEC-TV system with intensifier.

This possibility is illustrated in Fig. 18 by measurements in the SEC device. So far we demonstrated the signal to noise ratio; here the signal current itself is plotted against the final image current density. By varying the tube voltage, the characteristic signal curve can obviously be shifted over a wide range. In addition we have plotted the amplifier noise as a constant value (5 nA) and the quantum noise, which increases with decreasing current densities. The inserted oscillograms show the video signal of an individual video scanning line, which passes through an illuminated part of the image area. In all three cases the same image signal height was produced at different current densities and correspondingly different image amplification. At bigger current densities only the noise of the television device can be noticed, *i.e.* the amplification noise; at smaller densities, however, the electron noise becomes visible to a growing extent as a widening of the noise band. It can also be seen that below 10^{-13} A/cm² the electron noise even exceeds the signal.

We will see now what this means for the observation of the image. We prefer to use defocussed images of holes in thin foils as test objects for our

device. It is known that only with smallest illumination apertures and hence smallest current densities, the multiple Fresnel fringes become visible. Our image amplifying device makes possible the observation of such images, as it is demonstrated by Fig. 19, photographed from the monitor. We have

Fig. 19. – Defocussed image of a hole in a carbon film. $U_B = 100\,\mathrm{kV}$; $V_{el} = 80000\times$; $\Delta f/f = 3.8\cdot10^{-2}$; $\alpha_B = 1\cdot10^{-5}$; $j_E = 5\cdot10^{-15}\,\mathrm{A/cm^2}$; $j_0 > 3.2\cdot10^{-5}\,\mathrm{A/cm^2}$; $t_B = 25\,\mathrm{s}$.

photographed the monitor picture with different exposure times t_B at various very small current densities (which all completely exclude normal observation on the final screen), and we have represented the result in Fig. 20. Here an exposure time of about 0.2 s corresponds approximately to the impression, which the eye will have when observing the monitor. At $10^{-15}\,\mathrm{A/cm^2}$ only the signals of individual electrons are visible and only at $4\cdot10^{-14}\,\mathrm{A/cm^2}$ the image starts to fill up. The effect of a longer exposure time can be realised by a longer after glow time when looking at the monitor. Our devices have thus been equipped with such monitors.

There are, it must be mentioned, still other, much more effective possibilities to advance into the field of smallest current densities, for which the best present possibilities are offered by the SEC tube. On one hand the SEC target has the characteristic that a picture load stored on it is «cleared

off » except for a few per cent of its initial load. This means that a SEC tube can reproduce quick changing processes, it does not show any smearing effects, which prevent the observation of dynamic processes, especially in the

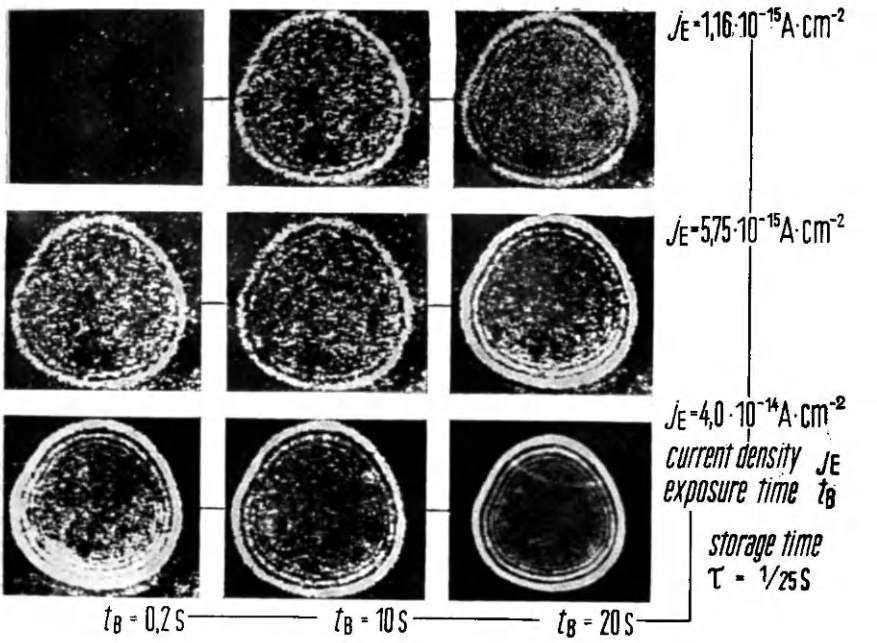


Fig. 20. – Monitor pictures of Fresnel fringes.

vidicon. But as soon as the electron beam is switched off, the image information is held for several hours on account of the high insulating ability of the SEC target. It is just this characteristic which distinguishes the SEC tube from all other camera tubes. On account of the high electric resistance, it is not possible for a charge distribution to flow off. As far as the other camera tubes are concerned (vidicon, plumbicon, orthicon), the electrical target characteristics permit only storage times up to 0.5 s, without a considerable loss of information happening, which causes a decrease in signal height, subsequently a worsening of the signal to noise ratio, edge smearing at the black-and-white borders, *i.e.* a lack of focus and an increase in background brightness. In the case of the SEC target, however, we can wait at small illumination intensities until enough image information has been accumulated on the target, and then can start a reading process. But it requires

further technical developments to observe this image stationary, which at first would only appear on the monitor as a flash. Magnetic tape recorders are one of the means, which would be taken into consideration at the moment. Let us now consider once more the eq. (5), which represents the signal to noise

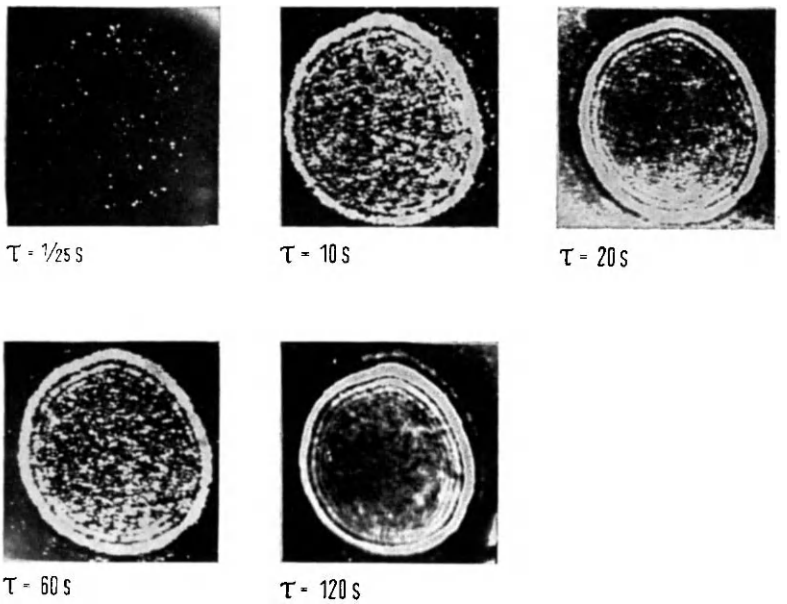


Fig. 21. – Monitor pictures of Fresnel fringes for various storage times τ. $j_E = 1.16 \cdot 10^{-15}$ A/cm^2, $j_0 = 1.85 \cdot 10^{-4}$ A/cm^2.

ratio for the electron noise. According to the television standard we have so far calculated with storage times of about 1/25 s. At longer storage times τ the limitation curve defined in (5) shifts to smaller current densities, where $j \propto 1/\tau$. In this manner one can advance to image current densities of below 10^{-15} A/cm^2 at still tolerable storage times. Figure 21 demonstrates how an image, which is still invisible on the monitor with a normal storage time of 1/25 s, fills up with increasing storage time. In our television device we have provided for a timer, on which various storage times can be dialed. The flash-like appearing image can either be recorded via synchronized monitor photography or magnetic tape recorder.

With its storage capability the television device has a characteristic, which also is shared by the photographic plate, that is the *most important* characteristic, which makes the plate so superior to the visual observation of the final

screen image by the human eye. The device is capable of building up an image of electrons only occasionally falling, on the image plane as a result of a correctly adjusted storage time. After recording this image on magnetic tape, it can be regarded as a stationary image on the monitor. From the question of how small current densities one can reach by this method, immediately results the question of the resolving power of the stored images. An incompletely « filled » image naturally has an impaired resolution. By means of Fresnel fringes we therefore measured the resolving power during half image operation and different storage times, and plotted it against the final image current density in Fig. 22. One can see that for current densities below 10^{-15} A/cm^2 storage times of the magnitude of minutes are already needed, provided the resolution can be maintained.

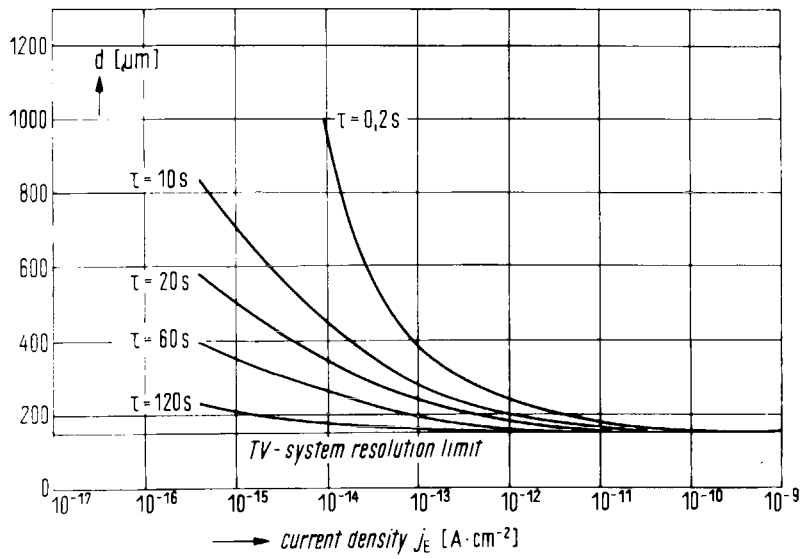


Fig. 22. – Resolvable distances d of Fresnel fringes for various storage times τ.

Something has to be added to the resolving power during storage operation. As already known, we work in such commercial television devices with the interlacing system, that is at first the even numbered lines of the television image (0, 2, 4, ... etc.) are scanned, and after that the scanning beam jumps back in order to scan the second half image with the lines 1, 3, 5, ... etc. Our image is thus composed of two half images, which succeed each other with

a time interval of 1/50 s. One difficulty which results from this is for our resolution during storage operation. During the discharge of the first half image, also the second half image is discharged to a great extent (up to about 50%). This means that the signal to noise ratio worsens accordingly, so that the signal will disappear among the noise of the television device. At storage times of several seconds, this 1/50 s between the half image is not sufficient any more to raise the second half image up to the full signal height. In these cases we have to take into account a resolution loss in vertical direction, *i.e.* our image is practically composed of half of the number of lines. By combining a special scanning technique with a specially controlled recording technique, this difficulty may be overcome: If during the scanning of the target, after one storage period has elapsed, only the first half image is recorded (for example with a polaroid camera). If after a second storage period has elapsed, during which the target has been charged anew, only the second half image is scanned (this can be achieved by electronic techniques). The storage times must only be integral multiples of 1/50 s, and registered on the recording device, on which the first half image has already been stored. A complete image with an optimum resolution is thus obtained.

So far we have only considered the influence of amplifier and quantum noise on the resolution, and will now consider the effects of the modulation transfer function.

In Fig. 23 is shown a test pattern familiar to entertainment television. It has been projected on the image amplifier input and photographed from the monitor. Immediately one can notice that in a big image area hardly any image distortions occur; small distortions can be noticed only at the outer edge of the image. The MTF of the system can now be determined with the aid of the test pattern. The test pattern offers to the device the same contrast for all distances. As can be seen from the video signal of a line, which has been drawn through this scanning pattern in the upper right corner, the smaller distances are underemphasized in the contrast which is associated with a decrease in signal amplitude and at the same time increase in background. If these measurements are evaluated, the curve shown at the bottom left of the figure is obtained, which demonstrates, to what degree the contrast decreases as a function of the local frequencies specified in MHz as is customary in television techniques. From such curves one can then calculate the object distances which will still be transferred by such a device. However, the MTF of the luminous screen, which has so far been disregarded here, has to be taken into account. This situation is demonstrated in the lower right part of the figure. With an electron optical magnification of 250000 and an available

contrast of 20%, object distances of about 7 Å (assuming the physiological contrast threshold at 10%) can be transmitted, which means that they should become visible on the monitor. With a contrast of 50%, distances of about (4÷5) Å should still be transmissible; here we already approach the limitation given by the bandwidth of the amplifier with $\Delta f = 8$ MHz.

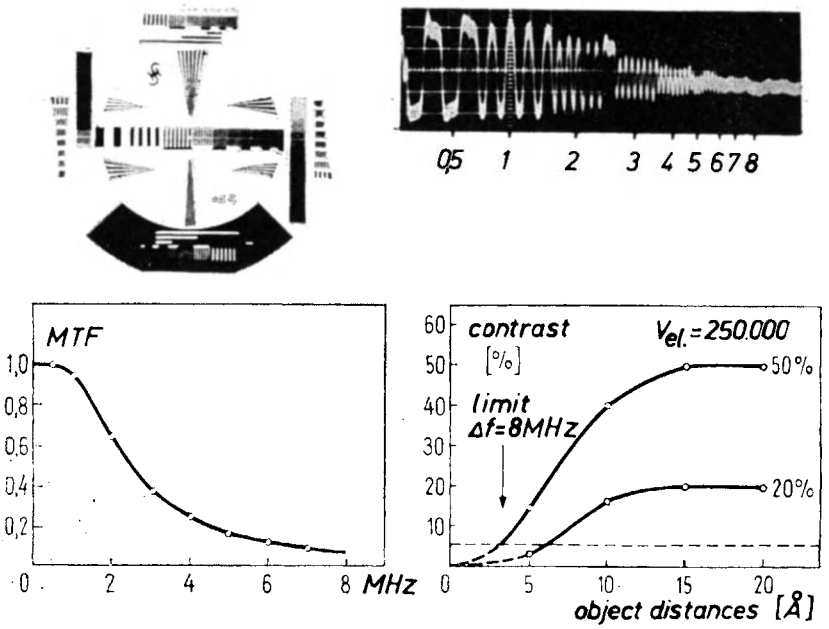


Fig. 23. – Contrast transfer of the SEC image intensifier system.

As a demonstration Fig. 24 shows a monitor picture of objects from the field of biology and metallurgy. It is not claimed that these are particularly characteristic examples of such applications. Figure 25 shows more examples of biological applications. The following can be said about the main fields of application of the image amplification device:

1) For high resolution microscopy the image amplification will be invaluable, since the correction and focussing situation can be adequately recognized before the picture is taken.

2) Objects which are sensitive to electron radiation, such as are found for example in organic chemistry and biology, can be tested within the range of 10^{-4} A/cm^2 in the object plane.

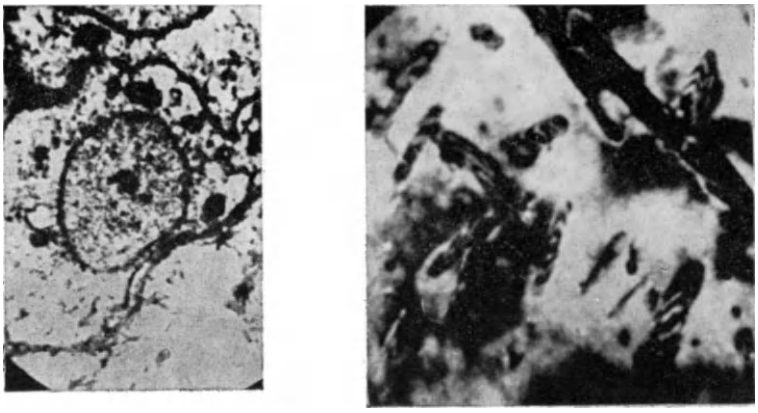


Fig. 24. – Monitor pictures of biological and metallurgical specimens. Left: $j_E = 1 \cdot 10^{-12}$ A·cm^{-2}, $t_B = 20$ s, 5200×; right: $j_E = 1 \cdot 10^{-12}$ A·cm^{-2}, $t_B = 20$ s, 40000×.

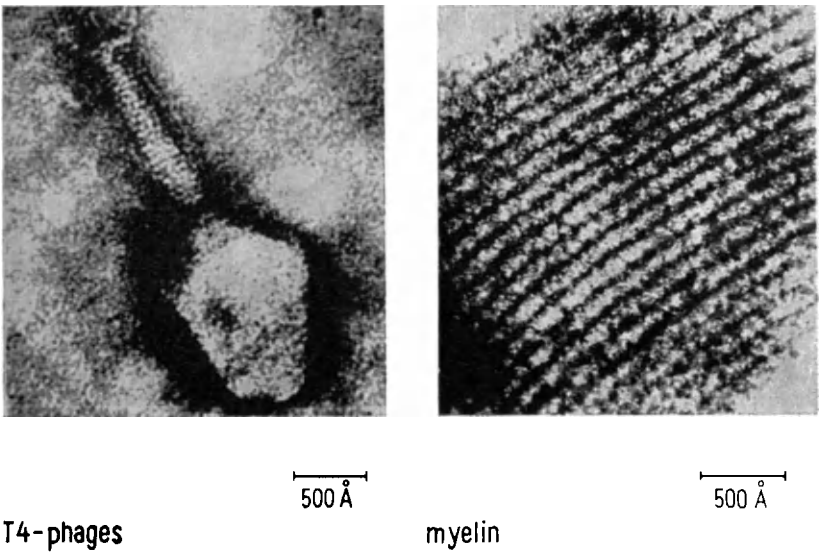


Fig. 25. – Monitor pictures of biological specimens.

3) A complete replacement of the plate by television is not to be expected at the moment on account of the limited image area. In special cases, however, one would be satisfied with video recording. This, for example, applies to motion picture recording of dynamic processes.

4) In images of poor contrast (for example thin specimens), electronic technique offers valuable possibilities of contrast enhancement. The biologist will certainly make use of this.

5) For quantitative image contrast measurements, television does not offer the exactness and the range of orders of magnitude measured by semiconductor detectors. But observations of the image signal in the television device on an oscilloscope screen can now and then be valuable on account of the rapidity with which it follows changes.

6) In future the television techniques could offer favourable conditions for an image analysis by electronics. An automatic focussing aid, for example, could possibly use an image amplification device as basic instrument.

We hope that we were able to make clear the uses and limitations of two measuring methods, which, it must be said, cannot replace the photographic plate, but which for the microscopist have such profitable characteristics, that in future they will attain increasing importance as accessory instruments, and may favourably serve as supplements to each other.

REFERENCES (Section 2)

1) F. Ansbacher and W. Ehrenberg: *Proc. Phys. Soc.*, **64** A, 362 (1951).
2) M. E. Haine, A. E. Ennos and P. A. Einstein: *Journ. Sci. Instrum.*, **35**, 466 (1958).
3) A. M. Boerio, R. R. Beyer and G. W. Goetze: *Adv. Electronics Electron Phys.*, **22** A, 229 (1966).

The Theory of Electron Diffraction Image Contrast

A. HOWIE

Cavendish Laboratory - Cambridge, England

1. Introduction.

Considerations of time and space do not allow the presentation in these lecture notes of more than a fraction of the theory of electron diffraction image contrast. Most attention has therefore been given to the introductory sections which assume no more than an elementary knowledge of electron waves. Later sections dealing with more sophisticated topics which have already been described in some detail elsewhere [1] are presented in a more condensed form. More up to date references have been quoted where possible but no pretence is made at completeness in this respect.

2. Foundations of diffraction theory.

Electrons propagating in free space after acceleration through a potential E_0 may be described by a plane wave $\exp[2\pi i\boldsymbol{\chi}\cdot\boldsymbol{r}]$ where the wave vector $\boldsymbol{\chi}$ describes the direction of motion and has magnitude $\chi = \lambda^{-1} = (2m_0eE_0)^{\frac{1}{2}}/h$. A more accurate formula for relativistic electrons is given later (see eq. (28)). When the wave strikes a crystal, elastic scattering (Bragg reflection) occurs and the amplitude of the wave $\exp[2\pi i\boldsymbol{\chi}'\cdot\boldsymbol{r}]$ emerging in the direction $\boldsymbol{\chi}'(\chi' = \chi)$ contains an interference factor A taking account of the different path lengths involved for scattering by different atoms.

$$A = \sum_{\boldsymbol{r}_{aj}} f_j \exp[2\pi i(\boldsymbol{\chi} - \boldsymbol{\chi}')\cdot\boldsymbol{r}_{aj}], \tag{1}$$

where f_j is the atomic scattering factor of the j-th atom at the angle 2θ in question and depends only on $|\boldsymbol{\chi}'-\boldsymbol{\chi}| = 2\sin(\theta)/\lambda$ (see Hall's lecture). In a perfect crystal the position of the j-th atom in the n-th unit cell is given by $\boldsymbol{r}_{nj} = \boldsymbol{r}_n + \boldsymbol{\rho}_j$, where $\boldsymbol{r}_n = n_1\boldsymbol{a} + n_2\boldsymbol{b} + n_3\boldsymbol{c}$. $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{c}$ are then the translation vectors of the lattice. We then have

$$A = \sum_j f_j \exp\left[2\pi i(\boldsymbol{\chi}-\boldsymbol{\chi}')\cdot\boldsymbol{\rho}_j\right] \sum_n \exp\left[2\pi i(\boldsymbol{\chi}-\boldsymbol{\chi}')\cdot\boldsymbol{r}_n\right] \tag{2}$$

$$= F\sum_n \exp\left[2\pi i(\boldsymbol{\chi}-\boldsymbol{\chi}')\cdot\boldsymbol{r}_n\right], \tag{3}$$

where $F(|\boldsymbol{\chi}'-\boldsymbol{\chi}|)$ is the scattering amplitude of the unit cell. The behaviour of the interference term depends on $\boldsymbol{\chi}'-\boldsymbol{\chi}$, a vector in reciprocal space and which it is convenient to describe relative to the reciprocal lattice.

The vectors $\boldsymbol{g}$ of the *reciprocal lattice* are defined by

$$\boldsymbol{g} = h\boldsymbol{a}^* + k\boldsymbol{b}^* + l\boldsymbol{c}^*, \tag{4}$$

where the reciprocal lattice translation vectors have the property $\boldsymbol{a}\cdot\boldsymbol{a}^* = \boldsymbol{b}\cdot\boldsymbol{b}^* = \boldsymbol{c}\cdot\boldsymbol{c}^* = 1$, $\boldsymbol{a}\cdot\boldsymbol{b}^* = \boldsymbol{a}\cdot\boldsymbol{c}^* = \boldsymbol{b}\cdot\boldsymbol{c}^* = 0$ and can easily be shown to be given by

$$\boldsymbol{a}^* = \frac{\boldsymbol{b}\wedge\boldsymbol{c}}{V_c}, \qquad \boldsymbol{b}^* = \frac{\boldsymbol{c}\wedge\boldsymbol{a}}{V_c}, \qquad \boldsymbol{c}^* = \frac{\boldsymbol{a}\wedge\boldsymbol{b}}{V_c}, \tag{5}$$

where V_c is the volume of the unit cell. (See Goringe and Hall, *Problem 1.*)

As a consequence of these definitions the vector $\boldsymbol{g}$ is normal to the crystal lattice plane with Miller indices h, k, l and $\boldsymbol{g}\cdot\boldsymbol{r}_n = $ integer for any lattice vector $\boldsymbol{r}_n$. Reference to eq. (3) then shows that all of the waves scattered by the different unit cells will have the same phase, leading therefore to a maximum in the scattered amplitude A, if and only if

$$\boldsymbol{\chi}'-\boldsymbol{\chi} = \boldsymbol{g}. \tag{6}$$

It is left as an exercise for the student to show that this condition is equivalent to Bragg's law $\lambda = 2d\sin\theta$. When the Bragg condition is not exactly fulfilled it is convenient to write

$$\boldsymbol{\chi}'-\boldsymbol{\chi} = \boldsymbol{g} + \boldsymbol{s}, \tag{7}$$

where $\boldsymbol{s}$ is a small vector in reciprocal space denoting deviation from the Bragg condition. Figure 1 giving the Ewald sphere construction used to locate the important Bragg reflections, shows the relation between these various vectors.

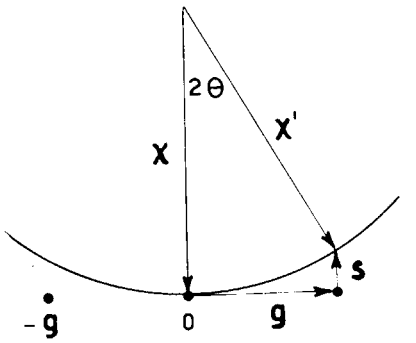


Fig. 1. – The Ewald sphere construction.

The diffracted intensity $|A|^2$ from a perfect thin crystal of thickness t in the z direction and lateral dimensions L_x and L_y is obtained by substituting from eq. (7) and replacing the sum in eq. (3) by an integral.

$$|A^2| = \frac{|F|^2}{V_c^2} \frac{\sin^2(\pi t s_z)}{(\pi s_z)^2} \frac{\sin^2(\pi L_x s_x)}{(\pi s_x)^2} \frac{\sin^2(\pi L_y s_y)}{(\pi s_y)^2} . \tag{8}$$

We can see from this expression that $|A|^2$ will be very small unless $s_z t$, $s_x L_x$ and $s_y L_y$ are all small in magnitude. Since L_x and L_y are much greater than t we therefore usually regard s_x and s_y as being negligible compared with s_z and simply denote s_z by s.

In the electron diffraction case the small values of λ lead to small Bragg angles $\theta_B \simeq \lambda/2d \simeq 0.01$ rad and the Ewald sphere is very large. Moreover the atomic scattering amplitudes $f(\theta)$ fall off fairly rapidly with increasing $\sin\theta/\lambda$ so that the important reciprocal lattice points giving rise to Bragg reflection in a given case lie on a plane of the reciprocal lattice passing through the origin and lying approximately normal to the incident beam direction. The ability to recognise and index the spots in these *cross grating diffraction patterns* is an important preliminary in the interpretation of electron micrographs of crystals. (See Goringe and Hall, *Problem 2.*)

Equation (8) shows that a given cross grating pattern will be observed over a certain range ($\sim 5^{o}$) of incident angles (the value of s and hence the intensity, but not the position, of each spot may change). Fortunately in the case of thicker crystals one observes in the diffraction pattern *Kikuchi lines* due to inelastic scattering and as the crystal is tilted these lines move as if rigidly attached to it (see Fig. 2 for a description of the origin of the lines).

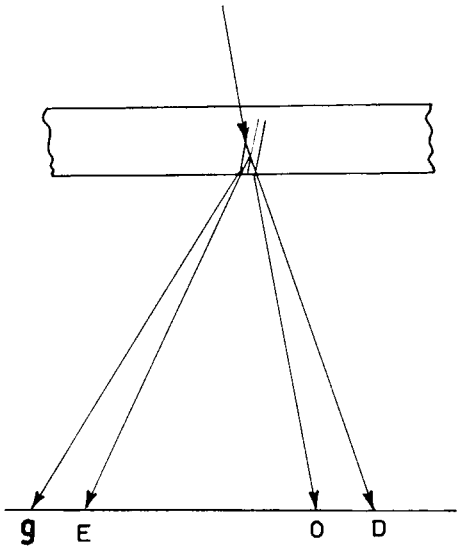


Fig. 2. – Formation of Kikuchi lines. Elastically scattered electrons form diffraction spots at 0 and g. Inelastically scattered electrons travel in various directions but those travelling towards D and falling at a Bragg angle on a particular set of planes are diffracted towards E or *vice versa*.

Kikuchi lines occur in parallel pairs separated by g (the magnitude of the reciprocal lattice vector corresponding to the Bragg reflection involved). At the exact Bragg reflecting position the two lines pass through 0 and g as shown in Fig. 3. The precise incident beam orientation can readily be worked out when the lines are in different positions. In very thick crystals where the Bragg spots may not be clearly visible against the background it may be easier to identify prominent crystal orientations from the complex but characteristic intersecting patterns of Kikuchi lines or bands (the dark or bright strip often observed in the region between two lines). By comparing the angular width of these bands with the purely geometrical angle between two such prominent orientations, the Bragg angle and hence the accelerating

voltage of the instrument can be measured. (See Goringe and Hall, this volume, *Problems 3* and *4*.)

Returning to the question of elastic scattering we note that eqs (1), (2) and (3) may be modified to deal with the case of imperfect crystals. The

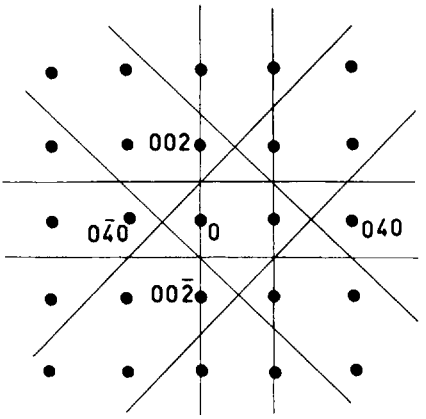


Fig. 3. – Kikuchi lines near 100 orientation, exact Bragg position for (020) and symmetry position for (002) and (002) reflections.

effect of the imperfection may be characterised by a change in F (due to impurity atoms for instance) or to an elastic strain which causes a displacement of the unit cell from the perfect crystal position $\boldsymbol{r}_n$ to $\boldsymbol{r}_n + \boldsymbol{R}(\boldsymbol{r}_n)$. Using eqs (3) and (7) we then find in the latter case

$$A = F\sum_n \exp\left[-2\pi i s z - 2\pi i \boldsymbol{g}\cdot\boldsymbol{R}\right], \tag{9}$$

where we have used the fact that $\boldsymbol{g}\cdot\boldsymbol{r}_n = \text{integer}$, $\boldsymbol{s}\cdot\boldsymbol{r}_n = sz$ and have neglected a term $\boldsymbol{s}\cdot\boldsymbol{R}$ which is very small since $s \ll g$ and $R \ll r_n$. The presence of defects thus gives rise to changes in the diffraction pattern and these can sometimes be used to give information about the nature of the displacement function $\boldsymbol{R}$. It can be seen for instance that if the direction of $\boldsymbol{R}$ lies in a plane, there will be Bragg reflections (from planes of atoms parallel to this) which are unaffected by the imperfection, since $\boldsymbol{g}\cdot\boldsymbol{R} = 0$. In general however, with the development of the methods of electron microscopy and also of X-ray topography it has been found much easier to get direct information about defects by studying the intensity leaving the exit surface of the crystal rather than the intensity in the diffraction pattern.

Figure 4 shows schematically the way in which the objective lens in an electron microscope forms an image (in the plane *CD*) of the intensity at the exit face of a crystalline object and a diffraction pattern in the plane *AB*. Subsequent lenses can thus be focussed on either *CD* or *AB* to obtain micrographs or diffraction patterns respectively. In the diffraction mode an inter-

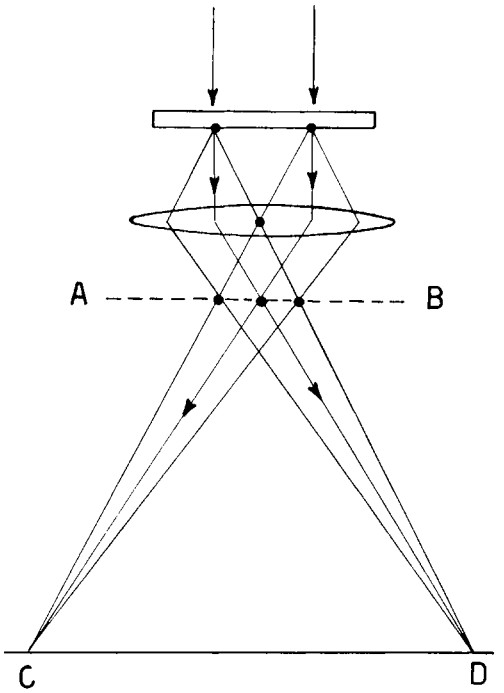


Fig. 4. – Ray paths in the objective lens.

mediate or *selecting area aperture* is placed in the image plane *CD* to define approximately the area of specimen contributing to the observed diffraction pattern. Similarly the parts of the diffraction pattern contributing to the micrograph can be controlled by inserting an *objective aperture* in the diffraction plane *AB*. When the objective aperture encloses only the central spot corresponding to the direct beam a *bright field* image is obtained. With the aperture enclosing one of the other diffraction spots a *dark field* image is obtained. Since a particular plane wave emerging from the object plane is brought to a point in the diffraction plane it can be seen (neglecting lens aberrations) that the amplitudes in these two planes are related by a Fourier

transform. Consequently the size of the objective aperture used can limit the fineness of the detail which can be observed on the micrograph. In particular, the lattice planes in the crystal will not be resolved unless the aperture encloses at least two Bragg spots simultaneously. These *direct resolution* images are usually obtained from rather thin crystals using the zero order (direct beam) spot and a low-order Bragg spot.

3. Simplified theories of wave propagation in crystals.

The origin of the diffraction effects just described is made clearer by considering electron propagation in the crystal potential $V(\boldsymbol{r})$ with the periodic property $V(\boldsymbol{r}+\boldsymbol{r}_n)=V(\boldsymbol{r})$ guaranteed by writing

$$V(\boldsymbol{r})=\sum_g V_g \exp\left[2\pi i\boldsymbol{g}\cdot\boldsymbol{r}\right]. \tag{10}$$

The constants V_g depend on the form of the potential in a given case (see eq. (30)) and have the property $V_g=V_{-g}^*$ since $V(\boldsymbol{r})$ is real. The effect of a potential V is to change the electron wave vector χ to a new local value χ_l where

$$\chi_l=\{2m_0e(E_0+V)\}^{\frac{1}{2}}/h\simeq\chi\left(1+\frac{1}{2}\frac{V}{E_0}\right)=\chi+\frac{m_0eV}{h^2\chi}. \tag{11}$$

When a plane wave $\varphi_0\exp\left[2\pi i\boldsymbol{\chi}\cdot\boldsymbol{r}\right]$ falls on a thin crystal slab $\mathrm{d}z$ an extra phase shift will occur in the slab so that the emerging wave becomes

$$\begin{aligned}\varphi_0\exp\left[2\pi i\boldsymbol{\chi}\cdot\boldsymbol{r}\right]\exp\left[2\pi i m_0eV(\boldsymbol{r})\,\mathrm{d}z/h^2\chi\right]&\simeq\\ &\simeq\varphi_0\exp\left[2\pi i\boldsymbol{\chi}\cdot\boldsymbol{r}\right]\{1+2\pi i m_0eV(\boldsymbol{r})\,\mathrm{d}z/h^2\chi\}\simeq\\ &\simeq\varphi_0\exp\left[2\pi i\boldsymbol{\chi}\cdot\boldsymbol{r}\right]\{1+(2\pi i m_0e\,\mathrm{d}z/h^2\chi)\sum_g V_g\exp\left[2\pi i\boldsymbol{g}\cdot\boldsymbol{r}\right]\}.\end{aligned} \tag{12}$$

From the right-hand side of (12) we see that in addition to the directly transmitted wave a number of diffracted waves leave the slab with the wave vectors $\boldsymbol{\chi}'$ and amplitudes $\mathrm{d}\varphi_g$ where

$$\mathrm{d}\varphi_g\exp\left[2\pi i\boldsymbol{\chi}'\cdot\boldsymbol{r}\right]=(2\pi i m_0e\,\mathrm{d}z/h^2\chi)\varphi_0V_g\exp\left[2\pi i(\boldsymbol{\chi}+\boldsymbol{g})\cdot\boldsymbol{r}\right].$$

Using eq. (7) and considering as before only the z component of s, we have

$$\mathrm{d}\varphi_g=\frac{\pi i}{\xi_g}\varphi_0\exp\left[-2\pi isz\right]\mathrm{d}z, \tag{13}$$

where the quantity ξ_g, with the dimensions of length, is the *extinction distance* given nonrelativistically by

$$\xi_g = h^2\chi/2m_0eV_g\,. \tag{14}$$

Typical values of ξ_g in electron diffraction lie in the range 100 Å to 1000 Å. (See Goringe and Hall, *Problem 5.*) Equation (13) is the basis of the *kinematical theory of diffraction* in perfect crystals and can be integrated directly (assuming that φ_0 is constant $= 1$) to give

$$\varphi_g(t) = \frac{\pi i}{\xi_g}\exp\left[-\pi its\right]\frac{\sin(\pi st)}{\pi s}, \tag{15}$$

$$|\varphi_g(t)|^2 = \frac{\pi^2\sin^2\pi ts}{\xi_g^2(\pi s)^2}\,. \tag{16}$$

Equation (16), with the same interference factor which appeared in eq. (8), can be used to discuss some of the features observed in dark field electron micrographs of perfect crystals. If the crystal thickness t varies, *thickness fringes* will be observed with a spacing $\Delta t = 1/s$. When t is constant but s varies due to local bending of the crystal, *extinction contours* or *bend contours* appear following the locus of points where the crystal is at the Bragg position. Equation (16) shows that these contours are symmetrical in $\pm s$ and have a bright central maximum flanked by subsidiary maxima of decreasing intensity but separated by constant amounts of $\Delta s = 1/t$. Contours of this type are in fact observed in very thin crystals. When only one Bragg reflection is important the intensity in the bright field image is complementary to that in the dark field image, *i.e.* $|\varphi_0|^2 = 1 - |\varphi_g|^2$.

Equations (10), (12) and (13) can readily be generalised to give the *kinematical theory for imperfect crystals.* As a result of displacements $\boldsymbol{R}$ due to defects the potential at the point $\boldsymbol{r}$ in the imperfect crystal is the same as that at $\boldsymbol{r} - \boldsymbol{R}(\boldsymbol{r})$ in the perfect crystal, *i.e.*

$$V(\boldsymbol{r}) = \sum_g V_g \exp\left[2\pi i\boldsymbol{g}\cdot\boldsymbol{r}\right]\exp\left[-2\pi i\boldsymbol{g}\cdot\boldsymbol{R}\right]. \tag{17}$$

Proceeding as before we then find

$$\mathrm{d}\varphi_g = \frac{\pi i}{\xi_g}\varphi_0\exp\left[-2\pi isz - 2\pi i\boldsymbol{g}\cdot\boldsymbol{R}\right]\mathrm{d}z\,, \tag{18}$$

$$\varphi_g(t) = \frac{\pi i}{\xi_g}\varphi_0\int_0^t \exp\left[-2\pi isz - 2\pi i\boldsymbol{g}\cdot\boldsymbol{R}\right]\mathrm{d}z\,. \tag{19}$$

Using these equations, in which the phase factor previously appearing in eq. (9) may be recognised, the bright and dark field images of various defects with known displacement functions $\boldsymbol{R}(\boldsymbol{r})$ can be calculated by numerical evaluation of the integral. It should be noted that for each point x, y, t on the exit surface $\varphi_g(t)$ is obtained by integrating down a column of crystal at the position x, y. Any change in the interference effects from neighbouring columns due to different values of $\boldsymbol{R}$ in these columns is ignored. This is known as the *column approximation.*

The kinematical theory is qualitatively quite successful in describing a number of the features observed in electron micrographs of both perfect and imperfect crystals but is not quantitatively reliable since it assumes that the diffracted waves are always so weak that the incident wave amplitude is constant with depth z in the crystal. As a consequence unphysical results are sometimes obtained. For instance eq. (16) implies that $|\varphi_g|^2$ may (for small s and large t) exceed the incident intensity.

Some of these difficulties can be overcome by using the *two-beam dynamical theory* in which it is assumed that the incident wave amplitude φ_0 and the diffracted wave amplitude φ_g both vary with depth z in the crystal. Only one diffracted wave is considered. It is then clear that eqs (13) and (18) will still be valid but must be combined with a similar equation for $\mathrm{d}\varphi_0$, the change in φ_0 due to diffraction from φ_g. Since the wave vector change $\boldsymbol{\chi}' \rightarrow \boldsymbol{\chi}$ is now reversed, the phase term is also changed in sign so that the coupled equations for an imperfect crystal become

$$\left.\begin{aligned} \frac{\mathrm{d}\varphi_g(z)}{\mathrm{d}z} &= \frac{\pi i}{\xi_g}\varphi_0(z)\exp\left[-2\pi i s z - 2\pi i \boldsymbol{g}\cdot\boldsymbol{R}\right], \\ \frac{\mathrm{d}\varphi_0(z)}{\mathrm{d}z} &= \frac{\pi i}{\xi_g}\varphi_g(z)\exp\left[2\pi i s z + 2\pi i \boldsymbol{g}\cdot\boldsymbol{R}\right]. \end{aligned}\right\} \tag{20}$$

Contrast calculations for specific defects can be carried out (again using the column approximation) by integrating this pair of equations from $z = 0$ to $z = t$ with the starting condition $\varphi_0(0) = 1$, $\varphi_g(0) = 0$. Details are given in Brown's lectures. As an exercise for the student it is left to show that the bright and dark field intensities are complementary, *i.e.* that $|\varphi_0|^2 + |\varphi_g|^2$ is constant. It may also be shown that in the case of a perfect crystal

$$|\varphi_g(t)|^2 = \frac{\pi^2}{\xi_g^2}\,\frac{\sin^2\left(\pi t\sqrt{s^2 + 1/\xi_g^2}\right)}{s^2 + 1/\xi_g^2}. \tag{21}$$

This formula is directly comparable with the kinematical theory eq. (16) to which it reduces when $t \ll \xi_g$ or when $s^2 \gg 1/\xi_g^2$, *i.e.* when the crystal is not too close to the Bragg position. Once again thickness fringes and bend contours are predicted but in contrast to the previous results the thickness fringes have a spacing $\Delta t = \xi_g/\sqrt{1+w}$ where $w = s\xi_g$ (with a maximum value of ξ_g at $s = 0$) and the bend contours, though still symmetrical, need not have a central maximum and the subsidiary fringes are not evenly spaced. All of these differences are confirmed experimentally. (See Goringe and Hall, *Problem 6.*)

The two-beam dynamical theory just outlined represents a great improvement on the kinematical theory but still has some limitations. In the first place several important absorption effects are not included. It is more convenient to deal with these later after introduction of the concept of Bloch waves, however it may be noted that eq. (20) can be modified to take account of absorption if the quantity $1/\xi_g$ is replaced by $1/\xi_g + i/\xi_g'$ where ξ_g' is an absorption parameter usually of the order of $10\xi_g$ or $20\xi_g$ in magnitude. A second limitation of the theory is the neglect of other Bragg reflections. It will be evident on consideration that the method used to derive eq. (20) could be extended to cover *n-beam dynamical theory* involving n coupled equations in n wave amplitudes (see eq. (49)). Such a theory is necessary for incident beam orientations where a number of reciprocal lattice points lie on or near the Ewald sphere. To some extent these orientations can be avoided but reference to Fig. 1 shows that, if the point $\boldsymbol{g}$ is on the sphere, the point $-\boldsymbol{g}$ will be deviated by only a small amount $s = g^2/\chi$. For the validity of two beam theory in this case we thus obtain the (approximate) condition

$$g^2\xi_g/\chi \gg 1\,. \tag{22}$$

The condition simply expresses the fact that if the deviation of the point $-\boldsymbol{g}$ from the sphere is less than $1/\xi_g$ it too will be in the dynamical region and a three beam theory will be required. In practice the two-beam theory is a very useful approximation in many cases but breaks down for strongly scattering crystals where small values of ξ_g can occur for low values of g.

Finally we note another useful approximation to the scattering problem which takes some account of many-beam dynamical effects at the expense of other disadvantages. This is the *phase grating approximation* in which the total phase shift of the wave in passing through the crystal is simply obtained by multiplying together the factors for successive slabs dz given on the left of eq. (12). We thus obtain for the wave function on the exit face

the expression

$$\psi(x, y, t) = \varphi_0 \exp[2\pi i \boldsymbol{\chi} \cdot \boldsymbol{r}] \exp[(2\pi i m_0 e/h^2 \chi) \int_0^t V(x, y, z)\, dz]\,. \tag{23}$$

This formula, which is not restricted to crystalline specimens, gives $|\psi|^2 = 1$, *i.e.* no amplitude contrast but only phase contrast (unless we effectively introduce some absorption by using a complex potential V). Amplitude contrast can be obtained in out-of-focus pictures (see pp. 540-623). The phase grating theory is a very useful one particularly for thin biological specimens but its basic assumption that the optical path is to be computed along a straight line trajectory parallel to z (equivalent to ignoring defocusing effects in a distance t) is not always justified. In the case of a potential $V = V_0 + 2V_g \cos(2\pi g x)$ for instance it can readily be seen that the lowest diffracted orders will travel at an angle $\theta_B = \pm g/2\chi$ and will eventually explore quite different regions of potential if $\theta_B t > 1/g$. The phase grating theory therefore requires the condition

$$g^2 t/2\chi \ll 1\,. \tag{24}$$

Comparison with eq. (22) then shows that it represents in a sense an alternative to the-two beam theory at the opposite extreme of approximation. (See Goringe and Hall, *Problems 10* and *11*.)

4. Formal theory of elastic scattering in perfect crystals.

The simplified treatment of electron diffraction just presented can be given a more rigorous basis combined with the deeper insight necessary for further extensions of the theory by a study of the perfect crystal solutions of the Schrödinger equation

$$\nabla^2 \psi(\boldsymbol{r}) + (8\pi^2 m e/h^2)\{E + V(\boldsymbol{r})\}\psi(\boldsymbol{r}) = 0\,. \tag{25}$$

Apart from possible spin effects this equation describes the diffraction of relativistic electrons provided that E and m are given in terms of the accelerating potential E_0 and the electron rest mass m_0 by the equations

$$m = m_0(1 - v^2/c^2)^{-\frac{1}{2}} = m_0(1 + eE_0/m_0c^2)\,, \tag{26}$$

$$E = E_0(1 + eE_0/2m_0c^2)/(1 + eE_0/m_0c^2)\,. \tag{27}$$

The magnitude of the wave vector χ is given by

$$\chi = \lambda^{-1} = 2m_0 eE_0\{1 + eE_0/2m_0c^2\}^{\frac{1}{2}}/h\,. \tag{28}$$

As before the potential $V(\boldsymbol{r})$ is given by the Fourier series

$$V(\boldsymbol{r}) = \sum_g V_g \exp\,[2\pi i\boldsymbol{g}\cdot\boldsymbol{r}] = \frac{h^2}{2me}\sum_g U_g \exp\,[2\pi i\boldsymbol{g}\cdot\boldsymbol{r}]\,, \tag{29}$$

where the constants U_g can be related to the scattering amplitude for electrons f_j of the j-th atom in the unit cell (at $\sin\theta/\lambda = g/2$).

$$U_g = \frac{m}{m_0}\frac{\exp\,[-M_g]}{\pi V_c}\sum_{j=1}^{l}\exp\,[-2\pi i\boldsymbol{g}\cdot\boldsymbol{\rho}_j]f_j(\sin\theta/\lambda)\,. \tag{30}$$

U_g then decreases slightly with increasing temperature because of the Debye-Waller factor $\exp\,[-M_g]$ and increases with increasing energy E because of the factor m/m_0. We look for a solution of eq. (25) in the form of a linear combination of plane waves linked by the Bragg reflection process.

$$\psi(\boldsymbol{r}) = b(\boldsymbol{r}) = \sum_g C_g \exp\,[2\pi i(\boldsymbol{k}+\boldsymbol{g})\cdot\boldsymbol{r}]\,. \tag{31}$$

It may be noted that this has the Bloch form of a plane wave $\exp\,[2\pi i\boldsymbol{k}\cdot\boldsymbol{r}]$ multiplied by a periodic function. By substituting into (25) and equating to zero the coefficient of each term in $\exp\,[2\pi i\boldsymbol{g}\cdot\boldsymbol{r}]$ we obtain a set of equations for the wave amplitudes C_g.

$$\{K^2 - (\boldsymbol{k}+\boldsymbol{g})^2\}C_g + \sum_{g'\neq 0} U_{g'}C_{g-g'} = 0\,, \tag{32}$$

where $K^2 = \chi^2 + U_0$ is a constant depending essentially on the electron energy. This set of N equations (N depends on the number of reciprocal lattice points or beams considered) constitutes the fundamental equations of dynamical theory. The method of solution can be demonstrated most easily in the *two-beam approximation* where only the wave amplitudes C_0 and C_g are considered. We then have a pair of equations

$$(K^2 - k^2)C_0 + U_{-g}C_g = 0\,, \qquad U_gC_0 + (K^2 - (\boldsymbol{k}+\boldsymbol{g})^2)C_g = 0\,. \tag{33}$$

For a nontrivial solution to occur we require

$$\begin{vmatrix} K^2-k^2 & U_{-g} \\ U_g & K^2-(\boldsymbol{k}+\boldsymbol{g})^2 \end{vmatrix} = (K^2-k^2)(K^2-(\boldsymbol{k}+\boldsymbol{g})^2) - |U_g|^2 = 0 . \quad (34)$$

This relation between the energy K^2 and the wave vector $\boldsymbol{k}$ constitutes a dispersion equation and enables us to plot a dispersion surface in reciprocal space for a given energy. In our case since $|U_g| \simeq K^2/500$ the dispersion surface consists of two spheres of radius K centred on 0 and $\boldsymbol{g}$ in the reciprocal lattice but splits slightly apart near the points where they touch on the Brillouin zone boundary. Part of the surface (not to scale) is shown in Fig. 5.

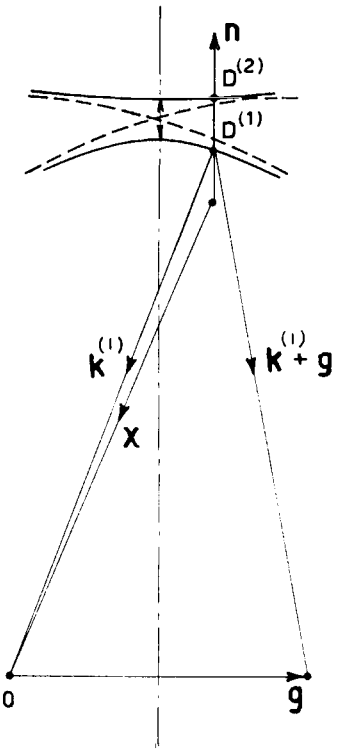


Fig. 5. – Two-beam dispersion surface. The vertical broken line is the Brillouin zone boundary.

Any point on this surface of two branches $j = 1, 2$ corresponds to a wave vector $\boldsymbol{k}^{(j)}$ and a solution $C_0^{(j)}$, $C_g^{(j)}$ of the eqs (33). The actual solutions which will be excited in a given case must have the same tangential components of

wave vector as the incident wave exp $[2\pi i\boldsymbol{\chi}\cdot\boldsymbol{r}]$ falling on the crystal entrance surface, *i.e.* $k_x^{(j)} = \chi_x$, $k_y^{(j)} = \chi_y$. The two points $D^{(1)}$ and $D^{(2)}$ on the dispersion surface can then be found by drawing through the end of the vector $\boldsymbol{\chi}$ a line in the direction $\boldsymbol{n}$ of the entrance surface normal and finding its intersections with the dispersion surface (see Fig. 5). For the case most frequently considered when $\boldsymbol{n}$ is parallel to the Brillouin zone boundary it can be seen that Δk, the difference between the z components of the two excited Bloch wave vectors reaches a minimum at the zone boundary, where $k^2 = (\boldsymbol{k}+\boldsymbol{g})^2$ (the exact Bragg position). Using eqs (34) and (14) we then find

$$\Delta k = |k_z^{(1)} - k_z^{(2)}| = K\cos\theta_B/|U_g| = 1/\xi_g\,, \tag{35}$$

where ξ_g is the extinction distance. Equations (33) then show that the waves at the zone boundary take the simple form $C_0^{(1)} = -C_g^{(1)} = 1/\sqrt{2}$; $C_0^{(2)} = C_g^{(2)} = 1/\sqrt{2}$. It is conventional to choose the wave amplitudes of a given Bloch wave so that $\sum_g |C_g|^2 = 1$. The intensity $|b^{(j)}|^2$ or current distribution in these two Bloch waves is shown relative to the crystal planes in Fig. 6. In wave (1)

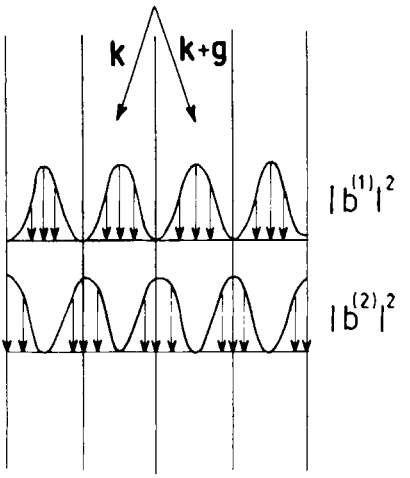


Fig. 6. – Current distribution relative to the Bragg planes in the two-beam case ($s = 0$).

the electrons mainly explore the regions of high potential energy between the atoms, whereas in wave (2) they travel in the regions of low potential energy near the atoms. Because the total energy E of the two waves is identical they

must have different kinetic energies and hence different values of k_z as seen in eq. (35). The difference $\Delta \boldsymbol{k}$ in k_z between the two waves results in a beating or interference effect as they travel through the crystal with a characteristic periodicity $\Delta k^{-1} = \xi_g$ at the reflecting position. This is the origin of the thickness fringes already noted and it can readily be demonstrated from eq. (33) that, for crystals deviated from the Bragg position, leading to an increase in Δk (see Fig. 5) the decrease in the thickness fringe spacing agrees with eq. (21).

Most of these simple ideas carry over to the case where N beams are considered in the construction of the Bloch wave in eq. (31). The dispersion eq. (34) then involves the vanishing of an $N \times N$ determinant defining a dispersion equation of N branches corresponding to spheres of radius K centred on the relevant reciprocal lattice points. There are thus N different Bloch waves $b^{(j)}(\boldsymbol{r})$ with the correct energy and tangential components of wave vector and the complete wave function may be written

$$\psi(\boldsymbol{r}) = \sum_{j=1}^{N} \psi^{(j)} \sum_{g} C_g^{(j)} \exp\left[2\pi i(\boldsymbol{k}^{(j)} + \boldsymbol{g})\cdot \boldsymbol{r}\right]. \tag{36}$$

The Bloch wave excitation amplitudes are determined from the condition that at the top of the crystal ($\boldsymbol{r} = 0$) only the direct wave amplitude should occur, *i.e.*

$$\sum \psi^{(j)} C_g^{(j)} = \delta_{0g}. \tag{37}$$

The general solution of this equation can be shown to be

$$\psi^{(j)} = C_0^{(j)}. \tag{38}$$

In practice the Bloch wave elements $C_g^{(j)}$ and wave vectors $k_z^{(j)}$ appear basically as eigenvectors and eigenvalues in the dynamical eq. (33) and can easily be computed by standard methods. Evidently however there are a vast number of situations to explore. So far most work has been done on the case of only *systematic reflections* $n\boldsymbol{g}$ along a line since the excitation of these reflections is controlled by the excitation of the low order reflection $\boldsymbol{g}$. *Accidental reflections* not along this line can often be avoided if desired. In certain symmetry situations analytical solutions of the dynamical equations can be obtained. For instance at the exact Bragg position for $2\boldsymbol{g}$ the equations (taking account of C_0, C_g, C_{2g}) simplify since $(\boldsymbol{k} + 2\boldsymbol{g})^2 = k` = k_z^2 + g^2, U_g =$

$= U_{-g} = U_1, U_{2g} = U_{-2g} = U_2$

$$\left.\begin{aligned}(K^2 - k_z^2 - g^2)C_0 + U_1 C_g + U_2 C_{2g} &= 0\,,\\ U_1 C_0 + (K^2 - k_z^2)C_g + U_1 C_{2g} &= 0\,,\\ U_2 C_0 + U_1 C_g + (K^2 - k_z^2 - g^2)C_{2g} &= 0\,.\end{aligned}\right\} \tag{39}$$

The appropriate dispersion surface is shown in Fig. 7. By inspection (or by symmetry arguments) we see that there are symmetric solutions with $C_0 = C_{2g}$

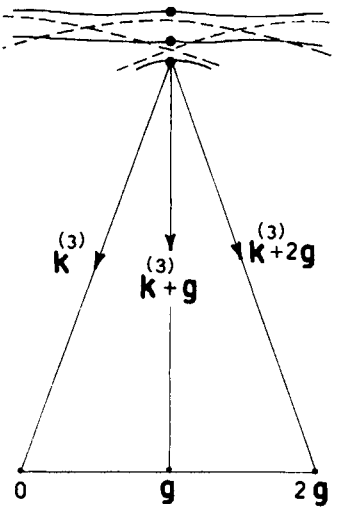


Fig. 7. – Three-beam dispersion surface at the reflecting position for $2g$.

or antisymmetric solutions with $C_0 = -C_{2g}$. In the latter case $C_g = 0$ and $k_z^2 - K^2 \simeq 2K(k_z - K) = -(U_2 + g^2)$. For the symmetric case it is easily shown that $k_z^2 - K^2 = -\frac{1}{2}(g^2 - U_2) \pm \{((g^2 - U_2)/2)^2 + 2U_1^2\}^{\frac{1}{2}}$. It may be noted that the smaller two of the three values of k_z will be equal provided that

$$U_2^2 + U_2 g^2 = U_1^2\,. \tag{40}$$

This condition (when two branches of the dispersion surface touch) can sometimes be fulfilled at a given energy since, as shown by eq. (30), U_1 and U_2 vary like m/m_0. At the critical energy a minimum is observed in the intensity of the second order diffracted beam ([2,3,4]) (the extinction contour correspond-

ing to $2\boldsymbol{g}$ vanishes). Measurement of the critical energy gives an accurate method of determining U_1 in terms of U_2 or *vice versa*. Such measurements are likely to be useful in, for instance, the study of ordering effects in alloys. It is suggested that the student should solve for himself various other many-beam problems in cases of symmetry, *e.g.* the four-beam case $-\boldsymbol{g}$, 0, $\boldsymbol{g}$, $2\boldsymbol{g}$ at the Bragg position for $\boldsymbol{g}$ ([5,6]); cases where several reciprocal lattice points all lie on the Ewald sphere such as $2\bar{2}0$, 202 (a three-beam case) or 200, 020, 220 (a four-beam case). In all cases the wave vectors $k_z^{(j)}$, the wave elements $C_g^{(j)}$ and the Bloch wave intensities $|b^{(j)}(\boldsymbol{r})|^2$ in the crystal can be found. (See Goringe and Hall, *Problems 7* and *9*.)

5. Anomalous absorption effects.

Various observations such as the noncomplementary nature of bright- and dark-field images, the disappearance of thickness fringes in thick crystals where there is still appreciable transmission and asymmetrical intensity distribution on either side of bright-field extinction contours suggest that some attenuation of Bloch waves occurs as they pass through the crystal. Moreover different Bloch waves are differently attenuated because they explore different regions of the unit cell. This attenuation can be included in the dynamical theory by allowing the wave vector components $k_z^{(j)}$ to have a small imaginary part $iq_z^{(j)}$ given by

$$q_z^{(j)} = \frac{me}{\hbar^2 k_z} \int b^{(j)*}(\boldsymbol{r}) V'(\boldsymbol{r}) b^{(j)}(\boldsymbol{r}) \, \mathrm{d}\tau \, . \tag{41}$$

$iV'(\boldsymbol{r})$ is an additional imaginary potential defining the parts of the unit cell where strong « absorption » of the waves occurs.

In analogy with the relativistic equation for the extinction distance ξ_g, we can define absorption distances ξ_g' in terms of the Fourier expansion coefficients of $V'(\boldsymbol{r})$

$$V'(\boldsymbol{r}) = \sum_g V_g' \exp\left[2\pi i \boldsymbol{g}\cdot\boldsymbol{r}\right] - \frac{\hbar^2}{2me} \sum_g U_g' \exp\left[2\pi i \boldsymbol{g}\cdot\boldsymbol{r}\right], \tag{42}$$

$$\xi_g = \chi/U_g \, , \qquad \xi_g' = \chi/U_g' \, . \tag{43}$$

Using the full solution for the two beam wave amplitudes $C_g^{(j)}$ from eq. (33) namely

$$C_0^{(1)} = C_g^{(2)} = \cos(\beta/2) \, , \qquad C_0^{(2)} = -C_g^{(1)} = \sin(\beta/2) \, , \qquad w = \operatorname{ctg}\beta \, , \tag{44}$$

we find from eqs (41), (42) and (43)

$$q_z^{(j)} = \frac{1}{2}\left(\frac{1}{\xi_0'} \pm \frac{1}{\xi_g'\sqrt{1+w^2}}\right), \tag{45}$$

where the lower sign refers to branch (1) which, as expected, is subject to less attenuation since it avoids the atoms. ξ_0' describes orientation-independent background absorption, ξ_g' describes the anomalous absorption which is especially noticeable near the Bragg position. Since ξ_g' can approach ξ_0' in magnitude the anomalous transmission of wave (1) is very important in practice and is responsible for the bright high transmission regions seen on the $w > 0$ side of low order extinction contours (see Fig. 8*a*)). Equations (44) show that for $w < 0$ mainly wave (2) is excited and this is heavily attenuated. (See Goringe and Hall, *Problem 8.*)

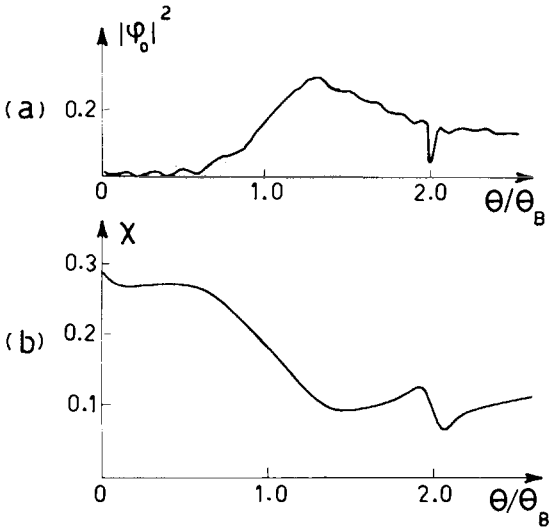


Fig. 8. – Computed curves for *a*) electron transmission (bright field: 111 reflections in Al); *b*) X-ray production (200 reflections in Ni).

Physically the « absorption » is due to scattering of the beam outside the objective aperture as a result of *a*) thermal diffuse scattering, *b*) inelastic collisions with atomic core electrons, *c*) high-order weak Bragg beams. Recent calculations ([7,8]) (where earlier references are quoted) of V_g' in terms of these processes have shown that the main contribution comes from *a*) but *b*) con-

tributes appreciably to the background term V_0'. The dependence of ξ_g' on temperature is complicated but to a rough approximation $\xi_g' \propto T^{-1}$ at high temperatures and increases by a factor of about 2 or 3 on cooling to 4 °K. As a function of electron velocity v, ξ_g' to a first approximation varies like v^2 whereas ξ_g varies like v.

6. Experimental verification of perfect crystal dynamical theory.

The theory just described taking account of anomalous absorption effects has been confirmed both qualitatively and to a fair extent quantitatively by numerous transmission electron microscope studies of thickness fringes and extinction contours in bent crystals. In flat crystals the information available from a bend contour can conveniently be obtained by the convergent beam technique ([9,10]). The advent of high voltage electron microscopy has stimulated more detailed studies of various critical voltage effects in the elastic (Bragg) scattering ([2-4]) (as in eq. (40)) and also in the anomalous transmission in many-beam situations ([11,12]). The close similarity between the anomalous transmission effect and the channelling effects observed in the last few years with protons, α particles and other ions ([13]) is apparent and can be developed in detail. The dependence of various scattering processes on the Bloch wave structure, and hence on the incident beam direction, have been directly observed and agree with calculation ([14]) in a number of cases, *e.g.* inner shell X-ray production ([15,16]) and electron back scattering ([15]) reach a maximum when Bloch wave (2) is excited (Fig. 8*b*)). Recently the electron backscattering or secondary emission anomaly has been observed in much more detail in scanning electron microscopy ([17,18]) and also in Leed studies ([19]). In the case of the X-rays or of the high-energy component of the secondary emission it appears that the anomaly only arises from the region near the entrance surface where the incident electron Bloch waves mainly responsible for the effect have not yet been inelastically scattered. Deeper regions of the crystal may be important in the case of secondaries whose energy is much less than that of the incident beam since the effect then is more strongly influenced by the difficulty these electrons experience in travelling back to the surface especially if the incident wave was anomalously transmitted so that it penetrates a long way before it is inelastically scattered and begins to produce secondaries ([19]).

The applications of the theory can be further extended by use of the *reciprocity theorem* (the principle of reversibility of the waves). Even in the

presence of inelastic scattering the principle can still be applied and states that if the positions of the source (of given intensity) and the collector are interchanged the *intensity* received by the collector will not be affected ([20]). In the case of elastic scattering only, the principle refers to amplitudes rather than intensities. We can apply the reciprocity principle to the case of activated crystals where electrons are emitted by atoms on lattice sites to show that the electrons will emerge from the crystal mainly along the directions along which an incident electron beam directed into the crystal would produce waves with a high probability of striking the atoms ([21]) (Fig. 8*b*)). The principle can also be used ([22]) to explain various diffraction contrast effects observed in the scanning electron microscope ([23]).

7. Formal theory of elastic scattering in imperfect crystals.

Before discussing the details of diffraction theory for imperfect crystals we can point out that the reciprocity theorem (RT) just outlined can be used (subject to the column approximation) in conjunction with possible symmetry operations such as inversion in a centre (P) or mirror reflection (M) in a plane midway between the crystal surfaces, to demonstrate two general and very useful symmetry principles for defect images ([20]).

a) Bright field intensities from two columns 1 and 2 in a centrosymmetric crystal will be identical if the displacement functions $\boldsymbol{R}_1(z)$ and $\boldsymbol{R}_2(z)$ (p. 279) satisfy the relation

$$\boldsymbol{R}_1(z) = \boldsymbol{R}_0 - \boldsymbol{R}_2(t-z)\,, \tag{46}$$

where $\boldsymbol{R}_0$ is any constant ([24]) (see Fig. 9*a*)).

b) Dark-field intensities from two columns will be identical (see Fig. 9*b*)) provided that ([25])

$$\boldsymbol{R}_1(z) = \boldsymbol{R}_0 + \boldsymbol{R}_2(t-z)\,. \tag{47}$$

These principles of image symmetry were first observed experimentally ([24,26,27]) and are of considerable value in defect studies (see Brown, this volume).

There are a number of ways in which solutions can be developed for the Schrödinger equation in an imperfect crystal where the potential is given by eq. (17). The simple theory previously used is quickly obtained by taking

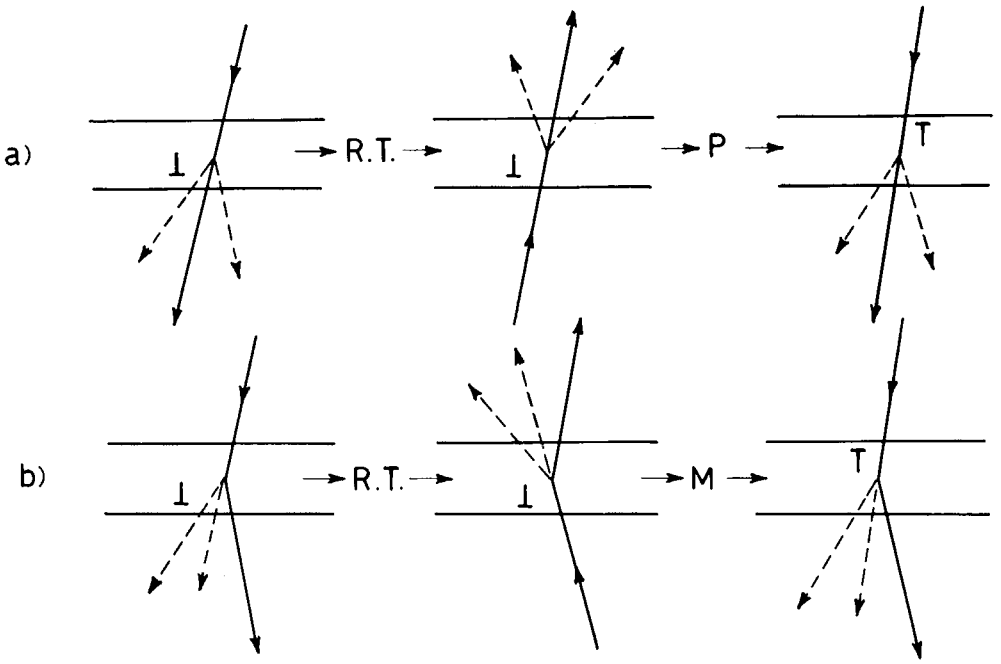


Fig. 9. – Image symmetry principles for *a*) bright field and *b*) dark field images generated from the reciprocity theorem (RT) combined with either inversion in a centre (*P*) or mirror reflection (*M*).

the wave function $\psi(\boldsymbol{r})$ in the form

$$\psi(\boldsymbol{r}) = \sum_g \varphi_g(\boldsymbol{r}) \exp\left[2\pi i(\boldsymbol{\chi} + \boldsymbol{g} + \boldsymbol{s}_g)\cdot\boldsymbol{r}\right]. \tag{48}$$

Substituting into the Schrödinger equation, and assuming that the functions $\varphi_g(\boldsymbol{r})$ and $\boldsymbol{R}(\boldsymbol{r})$ vary only slowly over a lattice distance so that the coefficient of the different terms in $\exp\left[2\pi i\boldsymbol{g}\cdot\boldsymbol{r}\right]$ can be separately equated to zero, we obtain the equations

$$(\boldsymbol{\chi} + \boldsymbol{g} + \boldsymbol{s}_g)_z \frac{\partial \varphi_g}{\partial z} = \sum_h \pi i U_{g-h} \exp\left[2\pi i((s_h - s_g)z + (\boldsymbol{h} - \boldsymbol{g})\cdot\boldsymbol{R})\right]\varphi_h(\boldsymbol{r}) - \\ - \left[(\boldsymbol{\chi} + \boldsymbol{g})_x \frac{\partial \varphi_g}{\partial x} + (\boldsymbol{\chi} + \boldsymbol{g})_y \frac{\partial \varphi_g}{\partial y} - \frac{i}{4\pi}\nabla^2\varphi_g\right]. \tag{49}$$

These equations can easily be recognised as a generalisation to the many-beam case of the two-beam eq. (20) provided the terms in square brackets can be neglected. This corresponds to the column approximation (see below). Many-beam images of defects can readily be computed with these equations but for small defects with localised strain fields in thick crystals they become a little cumbersome since the wave amplitudes φ_g change continuously even in the perfect crystal region. In these cases it is better to write ψ in the form

$$\psi(\boldsymbol{r}) = \sum_j \psi^{(j)}(z) \sum_g C_g^{(j)} \exp\left[2\pi i(\boldsymbol{k}^{(j)} + \boldsymbol{g})\cdot\boldsymbol{r}\right] \exp\left[-2\pi i \boldsymbol{g}\cdot\boldsymbol{R}\right]. \tag{50}$$

This differs from eq. (36) for the perfect crystal in that the Bloch wave amplitudes $\psi^{(j)}$ now vary with depth z in the crystal (still starting however at the value $\psi^{(j)}(0) = C_0^{(j)}$ given by eq. (38)) and that the Bloch waves contain the factor $\exp[-2\pi i \boldsymbol{g}\cdot\boldsymbol{R}]$ allowing these to follow automatically the local lattice displacement. Substituting into the Schrödinger equation, using the orthogonality of different Bloch waves and ignoring the x and y derivatives of $\boldsymbol{g}\cdot\boldsymbol{R}$ and all second derivatives of $\boldsymbol{g}\cdot\boldsymbol{R}$ and $\psi^{(j)}$ (a kind of column approximation) we readily obtain

$$\frac{\mathrm{d}\psi^{(j)}}{\mathrm{d}z} = 2\pi i \left\{ \sum_l \psi^{(l)}(z) \exp\left[2\pi i(k_z^{(l)} - k_z^{(j)})\, z\right] \left[\sum_g C_g^{(j)*} C_g^{(l)} \boldsymbol{g}\cdot\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}z} \right] \right\}. \tag{51}$$

These equations offer not only a second useful alternative for the numerical computation of defect images but also give a number of valuable insights into the basic principles of image contrast.

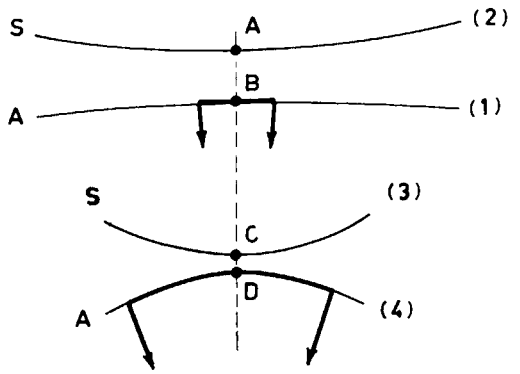


Fig. 10. – Four-beam theory dispersion surfaces.

a) The defect causes elastic scattering between different points on the dispersion surface, *e.g.* the points *A*, *B*, *C*, *D* in Fig. 10. These scattering transitions can be classed as *intraband* $l = j$ or *interband transitions* $l \neq j$ and strong diffraction contrast (at least in the case of localised defects) is due to the interband transitions ([28]). Reference to the equations shows that the interband terms can in fact be completely removed by redefining $k_z^{(j)}$.

b) The term in square brackets gives rise to certain selection rules and certain transitions can be forbidden, *e.g.* if $\boldsymbol{g} \cdot \boldsymbol{R} = 0$ for all the relevant reciprocal lattice vectors. At special points in the zone, *e.g.* Bragg reflecting positions on the zone boundary or at the centre of the zone $k_x = k_y = 0$, the Bloch waves themselves, as discussed previously, have certain symmetry properties and it can be seen from eq. (51) that only transitions where the Bloch wave symmetry is changed from symmetric (S) to antisymmetric (A) will be allowed. In particular, intraband transitions do not occur at these points.

c) The actual transition probability and hence the defect visibility depends on the rate at which $\beta_g' = \boldsymbol{g} \cdot \mathrm{d}\boldsymbol{R}/\mathrm{d}z$ varies. More precisely, using a weak scattering approximation, where we replace $\psi^{(l)}(z)$ on the right by $\psi^{(l)}(0) = C_0^{(l)} = \text{constant}$, we find

$$\psi^{(j)}(t) = \psi^{(j)}(0) + 2\pi i \sum_l \sum_g C_0^{(l)} C_g^{(j)*} C_g^{(l)} \int_0^t \boldsymbol{g} \cdot \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}z} \exp\left[2\pi i (k_z^{(l)} - k_z^{(j)}) z\right] \mathrm{d}z \,. \qquad (52)$$

Thus the transition probability depends on the Fourier transform of β_g' (for small localised defects not too close to the crystal surfaces the limits of integration may be extended to $\pm \infty$). The image contrast will be weak if β_g' varies either too rapidly or too slowly over a distance of the order of the relevant extinction distance.

It is now clear why many defects have image widths of the order of the extinction distance because it is only within this distance of the defect that sufficiently rapid variations of β_g' occur. Moreover it has been shown ([29,30]) that the image widths and visibility of small precipitate strain fields can be quantitatively assessed in terms of Δ_{lj} the integral appearing in eq. (52). For precipitates near the surface the image intensity (in bright field or strong dark field images) is expected to depend on Δ_{lj} but for precipitates near the middle of thick crystals a dependence on $|\Delta_{lj}|^2$ is expected.

These arguments strongly suggest the possibility that the effective resolution of very small defects could be improved if transitions involving large

values of $|k_z^{(i)} - k_z^{(j)}|$ corresponding to short extinction distances could be used ([1]). Such transitions can be obtained in the four-beam case shown in Fig. 10 if the crystal is set to the Bragg position so that the main waves excited are A and B. Transitions to C and D can then be detected by taking tilted dark field images in the weak beams $-\boldsymbol{g}$ or $+2\boldsymbol{g}$ (to which these waves mainly contribute). From symmetry, only the transitions AD or BC are allowed and the latter will be more important in all cases where the defect does not lie close to the entrance surface. Calculations ([30]) and more recently experiments ([31]) have shown that high resolution dark field images showing bright against a low intensity background can be obtained in this way. Various alternative schemes can be used, for instance, the reverse transitions DA and CB could be studied by setting the crystal to the Bragg position for $3\boldsymbol{g}$ and taking the dark field image in $\boldsymbol{g}$ or $2\boldsymbol{g}$. These two situations can be related by the reciprocity theorem. Further discussion of these techniques is given in Goringe's lecture and in *Problem 16*.

These possibilities of very high resolution studies of defects draw attention to the need for critical consideration of a number of approximations in dynamical theory, in particular the column approximation. Calculations ([30]) using eq. (49) showed that for strong beam images of dislocations the column approximation is extremely good but effects would be expected for more localised defects. In terms of Fig. 10 the column approximation consists of restricting the Bloch waves considered to the points A, B, C, D whereas in reality a range Δk_x of scattered wave points should be considered inversely proportional to the image width. For the usual dislocation images $|\Delta k_x| \ll g$ and it can be seen that the scattered waves all propagate in a narrow fan of semi-angle $\theta_D \ll \theta_B$. For weak beam images however $|\Delta k_x|$ is a good deal larger and θ_D may be more comparable with θ_B. It seems likely that the effects of the column approximation may be more severe when weak beam images using the transitions AD, BC are used rather than the inverse transitions DA, CB.

8. Inelastic scattering.

Some of the important aspects of inelastic scattering have already been discussed in terms of absorption effects, *i.e. loss of intensity* due to inelastic scattering outside the aperture. The dependence of these effects on both temperature and accelerating voltage are questions of immediate interest.

Many inelastically scattered electrons pass through the aperture however and contribute directly to the image, particularly in thick crystals ($t \gtrsim 2000$ Å)

where typically nearly all the electrons have lost energy $\Delta E \simeq (10 \div 20)$ eV at least once to the excitation of plasma oscillations (the predominant loss mechanism in the low-angle region). The mechanism by which the elastic scattering contrast effects just calculated can be preserved after these inelastic scattering events is of interest and is shown in Fig. 11.

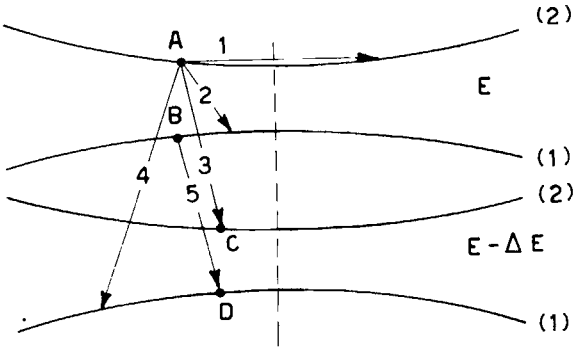


Fig. 11. – Elastic and inelastic scattering transitions.

Kikuchi lines and bands demonstrate directly the Bragg reflection and anomalous absorption effects experienced by inelastically scattered electrons. We therefore describe these electrons by a dispersion surface for energy $E - \Delta E$, shown in the two-beam approximation, relative to the dispersion surface for the zero loss electrons of energy E. Scattering transitions 1, 2, 3, 4 can then be classed as elastic intraband and interband, inelastic intraband and interband. Preservation of contrast is explained firstly by the fact that contrast-destroying transitions of the inelastic interband type 4 are forbidden or very weak, secondly by the fact that inelastic intraband transitions 3 and 5 can proceed using the same momentum change, *i.e.* the *same* plasma oscillation ([28]). The phase relation between the waves A and B is thus preserved at C and D. Thickness fringes, bend contours and dislocation images can thus be clearly observed using inelastically scattered electrons ([32]). This has now been demonstrated clearly for plasmon excitation ([33]) and with rather less certainty for single electron excitations ([33-35]). Contrast preservation after thermal-diffuse scattering is certainly much poorer but there is still some controversy on the point ([34.36.37]). Such electrons make only a very small contribution to the image in practice however. The extent to which contrast is not perfectly preserved after inelastic scattering may be of significance in determining the maximum thickness of crystal which can be studied at a given energy.

Contrast preservation may no longer hold in out-of-focus images because the inelastically scattered electrons have a spread of wave vectors $\Delta k_x = \theta_E k_z$ where $\theta_E = \Delta E/2E \simeq 10^{-4}$ is the angular width of the scattering distribution after plasmon excitation. The inelastically scattered electron image thus behaves as if it had a lateral coherence length $\Delta x = (\Delta k_x)^{-1} \simeq 300$ Å. The effect of this has recently been detected ([38]) using pointed filaments to increase the lateral coherence length of the incident illumination and hence the zero loss image. The in-focus images of the zero loss and first loss electrons were closely similar but on going out of focus, loss of contrast occurred much more quickly in the first loss image. This effect could be of importance in any situation where inelastically scattered electrons contribute to out-of-focus images.

9. Conclusions.

It is clear that the dynamical theory of diffraction contrast is now at a very interesting stage of development. Many observations can be explained in great detail so that there is no doubt about the essential correctness of most of the basic ideas. However, there is a continuing stream of new investigations for instance in high voltage electron microscopy, scanning microscopy, reflection diffraction, and in high resolution dark field microscopy which stimulate further theoretical efforts and test more rigorously the various approximations made. Questions of electron optical performance and design will evidently become of more importance in studies of lattice defects, in particular the convergence and coherence properties of the illumination and the imaging of the inelastically scattered electrons. It may be hoped that some day we will be able to describe the motion of each electron wave packet from the filament, down the column to the specimen and from there through the imaging lenses to the final screen. Only then will the wave-like properties of the electron upon which electron microscopy is based be made fully manifest.

REFERENCES

1) P. B. Hirsch, A. Howie, R. B. Nicholson, D. W. Pashley and M. J. Whelan: *Electron Microscopy of Thin Crystals*, Butterworths, London (1965).
2) D. Watanabe, R. Uyeda and M. Kogiso: *Acta Cryst.*, A **24**, 249 (1968).
3) D. Watanabe, R. Uyeda and A. Fukuhara: *Acta Cryst.*, A **25**, 138 (1969).
4) A. J. F. Metherell and R. M. Fisher: *Phys. Stat. Sol.*, **32**, 551 (1969).

5) A. HOWIE: *Phil. Mag.*, **14**, 223 (1966).
6) R. SERNEELS and R. GEVERS: *Phys. Stat. Sol.*, **33**, 703 (1969).
7) C. J. HUMPHREYS and P. B. HIRSCH: *Phil. Mag.*, **18**, 115 (1968).
8) G. RADI: *Acta Cryst.*, A **26**, 41 (1970).
9) M. HORSTMANN and G. MEYER: *Zeits. Phys.*, **159**, 563 (1960).
10) P. GOODMAN and G. LEMPFUHL: *Zeits. Naturfor.*, **19***a*, 818 (1964).
11) J. STEEDS and U. VALDRÈ: *Proc. 4th Eur. Conference on Electron Microscopy, Rome 1968* (Rome 1968), vol. **1**, p. 43.
12) C. J. HUMPHREYS and J. S. LALLY: *Journ. Appl. Phys.*, **41**, 232 (1970).
13) M. W. THOMPSON: *Contemp. Phys.*, **9**, 375 (1968).
14) P. B. HIRSCH, A. HOWIE and M. J. WHELAN: *Phil. Mag.*, **7**, 2095 (1962).
15) P. DUNCUMB: *Phil. Mag.*, **7**, 2101 (1962).
16) C. R. HALL: *Proc. Roy. Soc.*, A **295**, 140 (1966).
17) D. G. COATES: *Phil. Mag.*, **16**, 1179 (1967).
18) G. R. BOOKER, A. M. B. SHAW, M. J. WHELAN and P. B. HIRSCH: *Phil. Mag.*, **16**, 1185 (1967).
19) H. TAUB, R. A. STERN and V. F. DVORYANKIN: *Phys. Stat. Sol.*, **33**, 573 (1969).
20) A. P. POGANY and P. S. TURNER: *Acta Cryst.*, A **24**, 103 (1968).
21) P. N. TOMLINSON and A. HOWIE: *Phys. Lett.*, **27** A, 491 (1968).
22) J. M. COWLEY: *Appl. Phys. Lett.*, **15**, 58 (1969).
23) A. V. CREWE: *Quart. Rev. Biophysics*, **3**, 137 (1970).
24) A. HOWIE and M. J. WHELAN: *Proc. Roy. Soc.*, A **263**, 217 (1961).
25) C. J. BALL: *Phil. Mag.*, **9**, 541 (1964).
26) H. HASHIMOTO, A. HOWIE and M. J. WHELAN: *Proc. Roy. Soc.*, A **269**, 80 (1962).
27) M. F. ASHBY and L. M. BROWN: *Phil. Mag.*, **8**, 1083 (1963).
28) A. HOWIE: *Proc. Roy. Soc.*, A **271**, 268 (1963).
29) M. WILKENS: *Phys. Stat. Sol.*, **6**, 939 (1964).
30) A. HOWIE and Z. S. BASINSKI: *Phil. Mag.*, **17**, 1039 (1968).
31) D. J. H. COCKAYNE, I. L. F. RAY and M. J. WHELAN: *Phil. Mag.*, **20**, 1265 (1969).
32) Y. KAMIYA and R. UYEDA: *Journ. Phys. Soc. Japan*, **16**, 1361 (1961).
33) R. CASTAING, A. EL HILI and L. HENRY: *Compt. Rend.*, **262**, 169 (1966).
34) S. L. CUNDY, A. HOWIE and U. VALDRÈ: *Phil. Mag.*, **20**, 147 (1969).
35) C. J. HUMPHREYS and M. J. WHELAN: *Phil. Mag.*, **20**, 164 (1969).
36) R. CASTAING, P. HENOC, L. HENRY and M. NATTA: *Compt. Rend.*, **265**, 1293 (1967).
37) S. L. CUNDY, A. J. F. METHERELL and M. J. WHELAN: *Phil. Mag.*, **15**, 623 (1967).
38) D. R. SPALDING: *Ph.D. Thesis*, University of Cambridge (1969).

Application of Electron Diffraction

R. GEVERS

Rijksuniversitair Centrum Antwerpen - Antwerpen
Solid State Physics Department, S.C.K./C.E.N. - Mol, Belgium

1. Images of planar defects in electron transmission microscopy.

1·1. Model of planar defect.

One considers (see Fig. 1) a plate-shaped crystal foil (thickness z_0) formed by the superposition of two plane-parallel perfect crystals I and II (thicknesses z_1 and z_2, $z_0 = z_1 + z_2$).

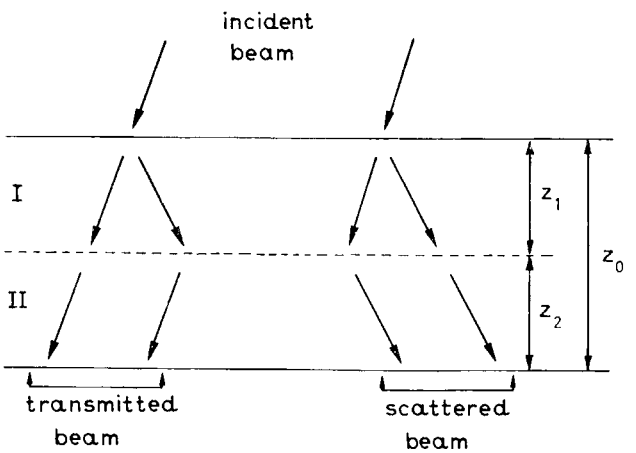


Fig. 1. – Schematic drawing showing the different beams in a crystal formed by the superposition of two plate-shape foils.

The two parts may have either the same lattice (stacking faults, antiphase boundaries, microtwins), or slightly different lattices (domain boundaries).

It is assumed that the foil orientation corresponds to a reasonable good « two-beam » situation, *i.e.* there is in both parts only one strongly scattered beam. The reflecting planes are defined by the reciprocal lattice vectors $\boldsymbol{g}_1$ in I, and $\boldsymbol{g}_2$ in II, and we notice:

$$\Delta \boldsymbol{g} = \boldsymbol{g}_2 - \boldsymbol{g}_1 . \tag{1}$$

Since $\Delta \boldsymbol{g} \neq 0$ means that the orientation and (or) the lattice parameter of the reflecting planes is slightly different in I and II, one has also to introduce a different deviation parameter from the exact Bragg orientation s_1 and s_2 in I and II.

It is easily seen from the reflection sphere construction of Fig. 2 that:

$$s_1 - s_2 = \Delta s = \Delta \boldsymbol{g} \cdot \boldsymbol{e}_z , \tag{2}$$

where $\boldsymbol{e}_z$ is the unit vector normal to the foil surfaces in the sense of propagation of the electrons.

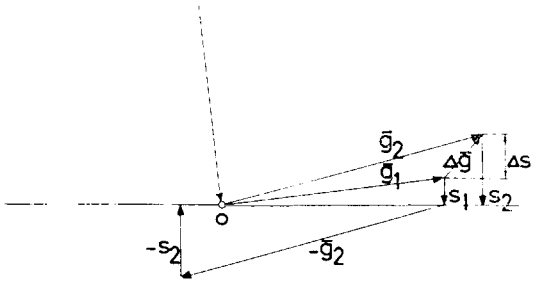


Fig. 2. – Reflection sphere construction showing in particular Δs. One has replaced, in good approximation, the sphere by its tangent plane. (Courtesy of *Phys. Stat. Sol.* **18**, 325 (1966).)

The amplitudes of the beams scattered and transmitted by a plate-shaped perfect crystal foil depend on the Fourier coefficients of the $\boldsymbol{g}$ term in the Fourier series of the crystal potential. Let ξ_1 and ξ_2 be the corresponding extinction distances in I and II, and ξ_1' and ξ_2' the two absorption lengths. Furthermore we notice θ_1 and θ_2 for the phase angles of these coefficients, for the same choice of origin.

It will be shown how one can calculate the intensities of the transmitted beam, $I_T(z_1, z_2)$ (bright-field image), and of the scattered beam $I_S(z_1, z_2)$ (dark field image).

These expressions apply also in the case of an interface inclined with respect to the surface (slope angle ψ, projected width a), assuming the column approximation to be valid. The depth z_1 of the interface under the entrance surface is then, however, a function of the position x of the column under consideration (see Fig. 3).

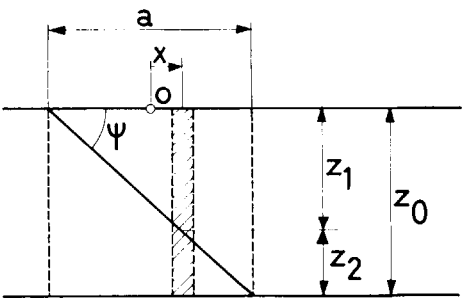


Fig. 3. – Scheme illustrating notations for interface inclined with respect to foil surface.

One has then to substitute:

$$z_1 = \left(\frac{a}{2} + x\right) \operatorname{tg} \psi, \qquad z_2 = \left(\frac{a}{2} - x\right) \operatorname{tg} \psi, \qquad -\frac{a}{2} \leqslant x \leqslant \frac{a}{2} \tag{3}$$

into the expressions for I_T and I_S, and one finds:

$$I_T = I_T(x), \qquad I_S = I_S(x), \tag{4}$$

where (4) depends on the illumination condition, *i.e.* s_1 and s_2, and on the parameters describing the planar defect. It is hoped that one can deduce from the observation of (4), significant informations about these parameters.

It will turn out that (4) are functions with nearly equidistant maxima and minima. One observes, in bright and dark field, a « fringe system » contrast image.

1'2. α and δ fringes.

1'2.1. *α-fringes.* – The most simple planar defect occurs if I and II are identical but displaced with respect to each other (stacking fault, anti-phase boundary).

One has then:

$$s_1 = s_2 = s \ (\Delta s = 0), \quad \xi_1 = \xi_2 = \xi, \quad \xi_1' = \xi_2' = \xi'. \tag{5}$$

The planar defect is completely characterized by the displacement vector $\boldsymbol{R}$ of part II with respect to part I. It has then be proposed as a good approximation to put:

$$V_2(\boldsymbol{r} + \boldsymbol{R}) = V_1(\boldsymbol{r}), \tag{6}$$

for the crystal potentials in both parts.

For the phase factors of the $\boldsymbol{g}$ Fourier coefficient, one obtains then:

$$\theta_2 + 2\pi \boldsymbol{g} \cdot \boldsymbol{R} = \theta_1,$$

or

$$\alpha = \theta_1 - \theta_2 = 2\pi \boldsymbol{g} \cdot \boldsymbol{R}. \tag{7}$$

The images (4) depend thus only on the phase angle α, and will be called « α-fringe » images.

1'2.2. *δ-domain boundaries.* – Ordering effects introduce mostly slight homogeneous deformations in a matrix crystal, *e.g.* ordering of impurities in Nb, of spins in NiO, of electrical dipoles in $BaTiO_3$. The symmetry of the resulting new phase is mostly lower than the symmetry of the matrix crystal. As a result, domains are formed, the deformation at different sides of the domain boundary being related by a symmetry operation of the matrix. The plane of the boundary is then a common lattice plane for the lattices of the adjacent domains. One can then choose the base vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ of the unit cell in the boundary, in part I and part II. The third base vector is $\boldsymbol{c}$ in I and $\boldsymbol{c} + \Delta\boldsymbol{c}$ in II.

One has then:

$$0 = \Delta(\boldsymbol{g} \cdot \boldsymbol{a}) = \Delta \boldsymbol{g} \cdot \boldsymbol{a}, \qquad \text{since } \Delta \boldsymbol{a} = 0$$

and also:

$$0 = \Delta(\boldsymbol{g} \cdot \boldsymbol{b}) = \Delta \boldsymbol{g} \cdot \boldsymbol{b}.$$

From $\Delta \boldsymbol{g} \cdot \boldsymbol{a} = \Delta \boldsymbol{g} \cdot \boldsymbol{b} = 0$ follows then that $\Delta \boldsymbol{g}$ will be normal to the interface. If $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are the base vectors of the reciprocal lattice corresponding to $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{c}$, one has:

$$\Delta \boldsymbol{g} = \Gamma \boldsymbol{C} \qquad (\Gamma \ll 1). \tag{8}$$

Furthermore, it follows from:

$$0 = \Delta(\boldsymbol{g} \cdot \boldsymbol{c}) = \Delta \boldsymbol{g} \cdot \boldsymbol{c} + \boldsymbol{g} \cdot \Delta \boldsymbol{c}$$

if one notes for $\Delta \boldsymbol{c}$

$$\Delta \boldsymbol{c} = \alpha \boldsymbol{a} + \beta \boldsymbol{b} + \gamma \boldsymbol{c}\,, \tag{9}$$

that

$$\Gamma = -(\boldsymbol{g} \cdot \Delta \boldsymbol{c}) = -(\alpha h + \beta k + \gamma l)\,, \tag{10}$$

if

$$\boldsymbol{g} = h\boldsymbol{A} + k\boldsymbol{B} + l\boldsymbol{C}\,. \tag{11}$$

One can always orient $\boldsymbol{C}$ such that it points from I to II. One has then (see Fig. 4):

$$\Delta s = \Delta \boldsymbol{g} \cdot \boldsymbol{e}_z = \Gamma \cos \psi C = -(\boldsymbol{g} \cdot \Delta \boldsymbol{c}) C \cos \psi\,. \tag{12}$$

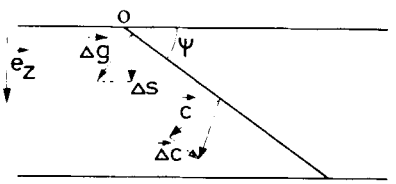


Fig. 4. – Schematic drawing illustrating the relation between $\Delta \boldsymbol{g}$ and Δs, for a boundary for which $\Delta \boldsymbol{g}$ is normal to the boundary plane.

In the case under consideration, one has, in very good approximation $\xi_1 = \xi_2$, $\xi_1' = \xi_2'$ and mostly $\theta_1 = \theta_2$, if one chooses the origin in the interface.

The characteristics of the fringe images will now depend on the parameter:

$$\delta = \xi \Delta s \tag{13}$$

and therefore will be called « δ-fringe » images.

If $\Delta \boldsymbol{c}$ is parallel to the interface, the two domains at opposite side of the boundary are « twin » related. The twin vector is then, however, very small as compared to a lattice vector.

It is possible that a domain boundary must be described by a α and a δ. These « α-δ » fringe patterns are more difficult to interpret than the pure α, or pure δ types.

If, moreover, $\xi_1 \neq \xi_2$ (and $\xi_1' \neq \xi_2'$) the images can become very complicated, *e.g.* images of twins in quartz.

1·3. Calculation of the amplitudes.

The amplitudes of the transmitted and scattered beam, leaving a plate-shaped crystal with thickness z, and deviation parameter s, are given by:

$$\Psi_T = \exp[i\pi s z] T(s, z), \tag{14a}$$

$$\Psi_S = \exp[-i\pi s(z + 2z')] S(s, z) \exp[i\theta], \tag{14b}$$

if z' is the distance from origin to entrance surface in the z-direction normal to the surfaces.

1·3.1. *Transmitted beam.* – The beam transmitted through the bi-crystal of Fig. 1, is the superposition of two beams.

1) The doubly transmitted beam $T_1 - T_2$. Its amplitude is given by:

$$\Psi_T^{(1)} = \exp[i\pi s_1 z_1] T_1 \exp[i\pi s_2 z_2] T_2 \exp[i 2\pi \boldsymbol{k}_0 \cdot \boldsymbol{r}],$$

or

$$\Psi_T^{(1)} = \exp[i\pi(s_1 z_1 + s_2 z_2)] T_1 T_2 \exp[i 2\pi \boldsymbol{k}_0 \cdot \boldsymbol{r}] \tag{15}$$

and its wave vector is $\boldsymbol{k}_0$, the wave vector of the incident beam.

2) The doubly scattered beam $S_1 - S_2$. In order to calculate its amplitude, one has to take the following remarks into account.

The beam scattered by the first part is considered as the beam incident on the second part. It can be scattered back into the original direction. However, the scattering vector is now $-\boldsymbol{g}_2$ and the corresponding deviation parameter is $-s_2$, as can be seen on the reflection sphere construction of Fig. 2. The phase angle θ to be introduced is then the one corresponding to $-\boldsymbol{g}_2$,

i.e. $-\theta_2$. One takes the origin in the entrance surface of the foil, this means that one has to put $z'=0$ in part I, but $z'=z_1$, in part II, in the expression (14*b*). At last, the wave vector in this partial beam is:

$$\boldsymbol{k}_0+(\boldsymbol{g}_1+s_1\boldsymbol{e}_z)-(\boldsymbol{g}_2+s_2\boldsymbol{e}_z)=\boldsymbol{k}_0-\Delta\boldsymbol{g}+(s_1-s_2)\boldsymbol{e}_z\,.$$

One has thus for the second transmitted beam:

$$\Psi_T^{(2)}=\exp\left[-i\pi s_1 z_1\right]\exp\left[i\theta_1\right]S_1\cdot$$
$$\cdot\{\exp\left[i\pi s_2(z_2+2z_1)\right]\exp\left[-i\theta_2\right]S_2^-\}\exp\left[i2\pi(\boldsymbol{k}_0-\Delta\boldsymbol{g}+(s_1-s_2)\boldsymbol{e}_z)\cdot\boldsymbol{r}\right], \qquad (16)$$

if one introduces the notations

$$T^-,\,S^-(s)=T,\,S(-s)\,. \qquad (17)$$

If one takes the origin on the intersection of the interface and the entrance surface, one has:

$$\boldsymbol{r}=\boldsymbol{r}_1+z_2\boldsymbol{e}_z$$

(for the meaning of $\boldsymbol{r}_1$ we refer to Fig. 5).

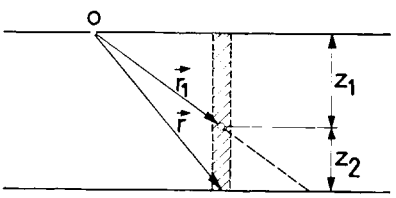


Fig. 5. – Scheme showing the meaning of the notation $\boldsymbol{r}_1$ used in text.

We calculate the phase factors occurring in (16). Namely (we leave out the factor π):

$$\varphi=-s_1z_1+s_2(z_2+2z_1)-2\Delta\boldsymbol{g}\cdot(\boldsymbol{r}_1+z_2\boldsymbol{e}_z)+2\Delta s(z_1+z_2)\,,$$

or, taking (2) into account:

$$\varphi=(s_1z_1+s_2z_2)-2(\Delta s)z_1-2(\Delta s)z_2+2\Delta s(z_1+z_2)-2\Delta\boldsymbol{g}\cdot\boldsymbol{r}_1\,,$$

or

$$\varphi = (s_1 z_1 + s_2 z_2) - 2\Delta \boldsymbol{g} \cdot \boldsymbol{r}_1 . \tag{18}$$

Taking (18) and the definition (7) into account, one obtains:

$$\Psi_T^{(2)} = \exp [i\pi(s_1 z_1 + s_2 z_2)] \exp [i(\alpha - 2\pi \Delta \boldsymbol{g} \cdot \boldsymbol{r}_1)](S_1 S_2^-) \exp [i2\pi \boldsymbol{k}_0 \cdot \boldsymbol{r}] . \tag{19}$$

From (15) and (19) one deduces finally for the amplitude of the transmitted beam, leaving out irrelevant phase factors:

$$T = T_1 T_2 + S_1 S_2^- \exp [i(\alpha - 2\pi \Delta \boldsymbol{g} \cdot \boldsymbol{r}_1)] . \tag{20}$$

1˙3.2. *Scattered beam.* – For the beam $T_1 - S_2$, *i.e.* transmitted by I and scattered by II, one has now:

$$\Psi_S^{(1)} = \exp [i\pi s_1 z_1] T_1 \{\exp [-i\pi s_2 (z_2 + 2z_1)] \exp [i\theta_2] S_2\} \cdot \cdot \exp [i2\pi(\boldsymbol{k}_0 + \boldsymbol{g}_2 + s_2 \boldsymbol{e}_z) \cdot \boldsymbol{r}] , \tag{21}$$

while the beam $S_1 - T_2$, *i.e.* scattered by I and transmitted by II, is given by:

$$\Psi_S^{(2)} = \exp [-i\pi s_1 z_1] S_1 \exp [i\theta_1] \exp [-i\pi s_2 z_2] T_2^- \cdot \cdot \exp [i2\pi(\boldsymbol{k}_0 + \boldsymbol{g}_1 + s_1 \boldsymbol{e}_z) \cdot \boldsymbol{r}] , \tag{22}$$

or, after calculating again first the phase factors:

$$\Psi_S^{(1)} = \exp [i2\pi(\Delta s) z_1] \exp [-i\pi(s_1 z_1 + s_2 z_2)] \exp [i\theta_2](T_1 S_2) \cdot \cdot \exp [i2\pi(\boldsymbol{k}_0 + \boldsymbol{g}_2 + s_2) \cdot \boldsymbol{r}] , \tag{23}$$

$$\Psi_S^{(2)} = \exp [i2\pi(\Delta s) z_1] \exp [-i\pi(s_1 z_1 + s_2 z_2)] \exp [i\theta_2] \cdot \cdot (S_1 T_2^-) \exp [i(\alpha - 2\pi \Delta \boldsymbol{g} \cdot \boldsymbol{r}_1)] \exp [i2\pi(\boldsymbol{k}_0 + \boldsymbol{g}_2 + s_2) \cdot \boldsymbol{r}] . \tag{24}$$

From (23) and (24) follows then for the total scattered beam, leaving out not significant phase factors:

$$S = T_1 S_2 + S_1 T_2^- \exp [i(\alpha - 2\pi \Delta \boldsymbol{g} \cdot \boldsymbol{r}_1)] . \tag{25}$$

1·3.3. *Matrix notation.* – The formulae (20) and (25) can be summarized as follows:

$$\begin{pmatrix} T \\ S \end{pmatrix} = \begin{pmatrix} T_2 & S_2^- \\ S_2 & T_2^- \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \exp\left[i(\alpha - 2\pi\Delta\boldsymbol{g}\cdot\boldsymbol{r}_1)\right] \end{pmatrix} \begin{pmatrix} T_1 & S_1^- \\ S_1 & T_1^- \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{26}$$

The physical meaning of (26) is straightforward. The first matrix from the right describes the boundary conditions at the entrance surface. The next matrix corresponds to the transmission through the first perfect part. The following diagonal matrix represents the planar defect, introducing a phase shift α and an orientation difference $\Delta\boldsymbol{g}$.

The last matrix corresponds again to a perfect plate-shaped crystal.

1·4. **Influence of absorption.**

The well-known expressions for T and S are:

$$T = \frac{1}{2}\left(1 - \frac{s}{\sigma_r}\right) \exp\left[i\pi\sigma_r z\right] \exp\left[-\pi\sigma_i z\right] + \\ + \frac{1}{2}\left(1 + \frac{s}{\sigma_r}\right) \exp\left[-i\pi\sigma_r z\right] \exp\left[\pi\sigma_i z\right], \tag{27}$$

$$S = \frac{1}{2(\sigma_r\xi)} \exp\left[i\pi\sigma_r z\right] \exp\left[-\pi\sigma_i z\right] - \frac{1}{2(\sigma_r\xi)} \exp\left[-i\pi\sigma_r z\right] \exp\left[\pi\sigma_i z\right], \tag{28}$$

where

$$\sigma_r = \frac{1}{\xi}(1 + \omega^2)^{\frac{1}{2}}, \tag{29a}$$

$$\omega = s\xi, \tag{29b}$$

$$\sigma_i = \frac{1}{\xi'}(1 + \omega^2)^{-\frac{1}{2}}. \tag{29c}$$

The first term in (27) and (28) counts the electrons which are in the strongly absorbed wave field, while the second term corresponds to the electrons in the easily transmitted wave field (one has left out in (27), (28) the exponential factor describing normal absorption).

For a thickness z not very small compared to $1/\sigma_i$, the second term is much more important then the first one. It evenb ecomes then a good approximation to neglect, in first approximation, the first term. This means that one does not count the relative few electrons which have survided at the back surface in the strongly absorbed wave field. The physical picture becomes than more simple, and so do the calculations.

For a thick crystal, one has, in first approximation:

$$\begin{pmatrix} T & S^- \\ S & T^- \end{pmatrix} = \frac{1}{2} \exp\left[-i\pi\sigma_r z\right] \exp\left[\pi\sigma_i z\right] \begin{pmatrix} 1+\dfrac{s}{\sigma_r} & -\dfrac{1}{\sigma_r \xi} \\ -\dfrac{1}{\sigma_r \xi} & 1-\dfrac{s}{\sigma_r} \end{pmatrix}. \tag{30}$$

As an example we consider the case of a stacking fault ($\Delta \boldsymbol{g} = 0$), and calculate the amplitudes near the center of the fault, assuming that $z_0/2$ is sufficiently large to accept (30) as a good approximation for part I and II. We assume moreover, for simplicity, that $s = 0$.

One finds then, from (26) and (30), with $s = 0$:

$$\begin{pmatrix} T \\ S \end{pmatrix} = \frac{1}{4} \exp\left[-i\pi \frac{z_0}{\xi}\right] \exp\left[\pi \frac{z_0}{\xi'}\right] \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \exp\left[i\alpha\right] \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

or

$$\begin{pmatrix} T \\ S \end{pmatrix} = \frac{1}{4} \exp\left[-i\pi \frac{z_0}{\xi}\right] \exp\left[\pi \frac{z_0}{\xi'}\right] \left(1 + \exp\left[i\alpha\right]\right) \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Noting $\begin{pmatrix} T_0 \\ S_0 \end{pmatrix}$ for $\begin{pmatrix} T \\ S \end{pmatrix}$ in the absence of the fault ($\alpha = 0$), one finds:

$$\begin{pmatrix} T \\ S \end{pmatrix} = \frac{1 + \exp\left[i\alpha\right]}{2} \begin{pmatrix} T_0 \\ S_0 \end{pmatrix}.$$

For the intensities:

$$\begin{pmatrix} I_T \\ I_S \end{pmatrix} = \cos^2 \frac{\alpha}{2} \begin{pmatrix} I_T^{(0)} \\ I_S^{(0)} \end{pmatrix}. \tag{31}$$

The mean intensity near the middle of the image of a stacking fault is strongly reduced with respect to the background intensity, by a factor $\cos^2 \alpha/2$, if $s = 0$ (in f.c.c. metals: $\alpha = \pm 2\pi/3$, and thus $\cos^2 \alpha/2 = \frac{1}{4}$).

The physical reason is that electrons, which cannot change from one wave field to another in a perfect crystal, can do so at the stacking fault. One has assumed that the electrons which arrive at the interface are those of the easily transmitted field. Many of these will, however, be transferred at the fault into the strongly absorbed field, and will not reach the back surface, *i.e.* the intensity will be lower than in the absence of the fault. The result (31) estimates how important the effect is.

The assumptions are too drastic. In reality the fluctuating terms will give rise to a fringe image near the center of the image, however, with poor contrast and a low background.

The reasoning is not valid for the two extremes of the images, since part I or part II become now too thin for the approximation (30) to be of any value for that part.

It is, however, still possible to use (30) for the thicker part, leading immediately to correct predictions about very significant properties of the images.

Let us first suppose that z_1 is small and z_2 is large. Using the approximation for part II, one finds then (for $s = 0$):

$$\begin{pmatrix} T \\ S \end{pmatrix} = \frac{1}{2} \exp\left[-i\pi \frac{z_2}{\xi}\right] \exp\left[\pi \frac{z_2}{\xi'}\right] \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \exp[i\alpha] \end{pmatrix} \begin{pmatrix} T_1 \\ S_1 \end{pmatrix},$$

or

$$\begin{pmatrix} T \\ S \end{pmatrix} = (\ldots) \begin{pmatrix} T_1 - S_1 \exp[i\alpha] \\ -(T_1 - S_1 \exp[i\alpha]) \end{pmatrix}.$$

For the intensities:

$$\begin{pmatrix} I_T \\ I_S \end{pmatrix} = \frac{1}{4} \exp\left[2\pi \frac{z_2}{\xi'}\right] |T_1 - S_1 \exp[i\alpha]|^2 \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{32}$$

One has immediately as a first conclusion: *bright* and *dark field are the same near the front surface* ($z_1 \ll z_0$) *in a thick crystal.*

Near the front surface, it becomes a good approximation to neglect absorption for the first part.

One can then substitute in (32):

$$|T_1 - S_1 \exp[i\alpha]|^2 = |\cos \pi \frac{z}{\xi} - i \sin \pi \frac{z}{\xi} \exp[i\alpha]|^2 = 1 + \sin\alpha \sin 2\pi \frac{z_1}{\xi}. \tag{33}$$

From (33) follow the further conclusions: *the first fringe at the front surface* in a thick crystal will be:

$$\textit{bright if} \quad \sin\alpha > 0\,,$$

$$\textit{dark} \quad \textit{if} \quad \sin\alpha < 0\,.$$

The depth position of the first fringe is $z_1 = \frac{1}{4}\xi$. In a column intersecting the fault close to the entrance surface, it is possible that electrons moving in the strongly absorbed field are saved from absorption by changing wave field at the fault plane. The intensity will then be higher than in the absence of the fault (bright field). The formula (33) shows where this will happen, and also how important the effect can be (for $s = 0$).

Let us consider now the other extreme of the image, *i.e.* $z_2 \ll z_0$.

One can now use the approximation (30) for the first part, leading to (for $s = 0$):

$$\begin{pmatrix} T \\ S \end{pmatrix} = \frac{1}{2}\exp\left[-i\pi\frac{z_1}{\xi}\right]\exp\left[\pi\frac{z_1}{\xi'}\right]\begin{pmatrix} T_2 & S_2 \\ S_2 & T_2 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & \exp[i\alpha] \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix},$$

or

$$\begin{pmatrix} T \\ S \end{pmatrix} = (\ldots)\begin{pmatrix} T_2 - S_2\exp[i\alpha] \\ -(T_2 - S_2\exp[-i\alpha])\exp[i\alpha] \end{pmatrix}.$$

For the intensities:

$$\begin{pmatrix} I_T \\ I_S \end{pmatrix} = \frac{1}{4}\exp\left[2\pi\frac{z_1}{\xi'}\right]\begin{pmatrix} |T_2 - S_2\exp[i\alpha]|^2 \\ |T_2 - S_2\exp[-i\alpha]|^2 \end{pmatrix}, \tag{34}$$

or taking (33) into account:

$$\begin{pmatrix} I_T \\ I_S \end{pmatrix} = \frac{1}{4}\exp\left[2\pi\frac{z_1}{\xi'}\right]\begin{pmatrix} 1 + (\sin\alpha)\sin 2\pi\dfrac{z_2}{\xi} \\ 1 - (\sin\alpha)\sin 2\pi\dfrac{z_2}{\xi} \end{pmatrix}. \tag{35}$$

From (35) follow the conclusions: *the bright field is symmetrical*; *the dark field is asymmetrical*: the nature of the first and last fringe are different at the back surface, where bright and dark field are pseudo-complementary. The first property is not only approximately true. It follows in fact from:

$$T = T_1 T_2 + S_1 S_2^- \exp[i\alpha] \quad \text{and} \quad S^- = S\,.$$

1·5. Properties of stacking fault images in thick crystal.

The properties deduced in Subsect. 1·4 for $s = 0$, can be proved by more precise calculations for $s \neq 0$, if z_0 is sufficiently large and $|s|$ sufficiently small.

For the nature of the first fringe at the front surface and the last fringe at the back surface, one has thus found:

TABLE I.

	Front surface		*Back surface*	
	B. F.	D. F.	B. F.	D. F.
$\sin\alpha > 0$	*B*	*B*	*B*	*D*
$\sin\alpha < 0$	*D*	*D*	*D*	*B*

The depth period of the fringe pattern is of course the period of $T_{1,2}$ and $S_{1,2}$, *i.e.* $\Delta z = \xi(1+\omega^2)^{-\frac{1}{2}}$.

1·6. Properties of δ-fringe pattern.

For the domain boundaries described in Subsect. 1·2.2, one has:

$$\alpha = 0\,, \qquad \Delta\boldsymbol{g}\cdot\boldsymbol{r}_1 = 0$$

and (26) becomes for this case:

$$\begin{pmatrix} T \\ S \end{pmatrix} = \begin{pmatrix} T_2 & \bar{S_2} \\ S_2 & \bar{T_2} \end{pmatrix} \begin{pmatrix} T_1 \\ S_1 \end{pmatrix}. \tag{35}$$

Let us consider the region close to the front surface ($z_1 \ll z_0$). Introducing the approximation (30) gives now for the bright field image:

$$T = \frac{1}{2} \exp\left[-i\pi\sigma_{2r} z_2\right] \exp\left[\pi\sigma_{2i} z_2\right] \left\{ \left(1 + \frac{s_2}{\sigma_{2r}}\right) T_1 - \frac{1}{\sigma_{2r}\xi_2} S_1 \right\}. \tag{36}$$

If one neglects the absorption for the thin part, one obtains for the expression between brackets:

$$\{...\} = \left(1 + \frac{s_2}{\sigma_{2r}}\right)\left[\cos \pi\sigma_{1r} z_1 - i \frac{s_1}{\sigma_{1r}} \sin \pi\sigma_{1r} z_1\right] - i \frac{1}{\sigma_{2r}\xi_2} \frac{1}{\sigma_{1r}\xi_1} \sin \pi\sigma_{1r} z_1 .$$

From the latter follows that:

$$|\{...\}|^2 = \left(1 + \frac{s_2}{\sigma_{2r}}\right)^2 \cos^2 \pi\sigma_{1r} z_1 + \left[\frac{s_1}{\sigma_{1r}}\left(1 + \frac{s_2}{\sigma_{2r}}\right) + \frac{1}{\sigma_{1r}\xi_1} \frac{1}{\sigma_{2r}\xi_2}\right]^2 \sin^2 \pi\sigma_{1r} z_1 ,$$

or

$$|\{...\}|^2 = \left(1 + \frac{s_2}{\sigma_{2r}}\right)^2 + \left[\left(\frac{s_1(\sigma_{2r} + s_2)}{\sigma_{1r}\sigma_{2r}} + \frac{1}{(\sigma_{1r}\xi_1)(\sigma_{2r}\xi_2)}\right)^2 - \frac{(\sigma_{2r} + s_2)^2}{(\sigma_{2r})^2}\right] \sin^2 \pi\sigma_{1r} z_1 .$$

We calculate now the coefficient A of $\sin^2 \pi\sigma_{1r} z_1$, using the notations (29*a*). One has:

$$A = \frac{[\omega_1[(1 + \omega_2^2)^{\frac{1}{2}} + \omega_2] + 1]^2 - [(1 + \omega_2^2) + \omega_2]^2 (1 + \omega_1^2)}{(1 + \omega_1^2)(1 + \omega_2^2)} =$$

$$= \frac{2(\omega_1 - \omega_2)[(1 + \omega_2^2)^{\frac{1}{2}} + \omega_2]}{(1 + \omega_1^2)(1 + \omega_2^2)},$$

or, introducing the notation

$$\delta = \omega_1 - \omega_2 = s_1\xi_1 - s_2\xi_2 , \tag{37a}$$

or

$$\delta = (\Delta s)\xi , \quad \text{if} \quad \xi_1 = \xi_2 , \tag{37b}$$

$$A = 2\delta \frac{1 + s_2/\sigma_{2r}}{(\sigma_{1r}\xi_1)^2 \sigma_{2r}\xi_2} , \tag{38a}$$

or

$$A = \left(1 + \frac{s_2}{\sigma_{2r}}\right)^2 2\delta \frac{\sigma_{2r}\xi_2}{(\sigma_{1r}\xi_1)^2} \left(1 - \frac{s_2}{\sigma_{2r}}\right) . \tag{38b}$$

One has then close to the entrance surface, in good approximation:

$$I_T = \frac{1}{4} \exp [2\pi\sigma_{2i} z_2] \left(1 + \frac{s_2}{\sigma_{2r}}\right)^2 \left\{1 + 2\delta \frac{(\sigma_{2r}\xi_2)(1 - s_2/\sigma_{2r})}{(\sigma_{1r}\xi_1)^2} \sin^2 \pi\sigma_{1r} z_1\right\} . \tag{39}$$

From (39) one concludes:

1) The fringe depth periodicity is $1/\sigma_{1r}$, corresponding to the effective extinction distance of the thin part.

2) The position of the first fringe is $z_1 = 1/\sigma_{1r}$.

3) The nature of the first fringe is solely determined by δ and does not depend on the diffraction condition $s_1 + s_2$.

The first fringe at the entrance surface is bright if $\delta > 0$, and dark if $\delta < 0$.

4) The contrast depends on s_1 and s_2 separately. For values of δ of the order of magnitude of one, the contrast will be important.

Let us consider now the dark field close to the front surface.

From (35) follows, making again use of the approximation (30):

$$S = \frac{1}{2} \exp\left[-i\pi\sigma_{2r} z_2\right] \exp\left[\pi\sigma_{2i} z_2\right] \left\{ -\frac{1}{\sigma_{2r}\xi_2} T_1 + \left(1 - \frac{s_2}{\sigma_{2r}}\right) S_1 \right\}. \tag{40}$$

Comparing (36) and (40), taking (29) into account, one obtains:

$$S = -\frac{1}{\sigma_{2r}\xi_2} \frac{1}{1 + s_2/\sigma_{2r}} T,$$

or

$$I_T = [(\sigma_{2r} + s_{2r})\xi_2]^2 I_S . \tag{41}$$

From (41) one concludes that bright and dark field are similar near the entrance surface.

The expression

$$T(z_1, z_2) = T_1 T_2 + S_1 S_2 ,$$

where one has taken into account that $S_2^- = S_2$ is invariant for the interchange $(z_1 \rightleftarrows z_2)$, provided one substitutes also $s_1 \to s_2$, $s_2 \to s_1$, $\xi_1 \to \xi_2$, $\xi_2 \to \xi_1$, $\xi_1' \to \xi_2'$, $\xi_2' \to \xi_1'$.

As a consequence, one has also to substitute $\delta \to -\delta$.

The bright field is asymmetrical. In particular the nature of first and last fringe are different.

In the same way, one remarks that the expression

$$S(z_1, z_2) = T_1 S_2 + S_1 T_2^-$$

does not change if z_1 and z_2, are inverted, provided one substitutes:

$$\xi_1 \leftrightarrows \xi_2, \quad \xi_1' \rightleftarrows \xi_2', \quad s_1 \rightarrow -s_2, \quad s_2 \rightarrow -s_1.$$

The important parameter δ remains then unchanged. One concludes: the dark field image is pseudo-symmetrical, *i.e.* the nature of first and last fringe is the same for the dark field image.

We summarize the results about the nature of the outer fringes in Table II.

TABLE II.

	Front surface		*Back surface*	
	B. F.	D. F.	B. F.	D. F.
$\Delta s > 0$	*B*	*B*	*D*	*B*
$\Delta s < 0$	*D*	*D*	*B*	*D*

1·7. Examples of application.

1·7.1. *Determination of type of stacking faults in f.c.c. metals.* – The stacking faults present in f.c.c. metals can be obtained either by removing a layer and closing the gap (intrinsic fault), or by inserting a layer (extrinsic fault).

The displacement vector $\boldsymbol{R}$ of the second part II with respect to the first, can always be noted as:

$$\boldsymbol{R} = \tfrac{1}{3}[111] \tag{42}$$

and is normal to the stacking fault.

For an intrinsic fault $\boldsymbol{R}$ must point from II to I, since $\boldsymbol{R}$ closes the gap after the removal of the layer. For an extrinsic fault $\boldsymbol{R}$ must point in the opposite sense, from I to II, since II must first be displaced to make room for the new layer to be inserted.

The determination of the type of the fault corresponds then to the determination of the sense of $\boldsymbol{R}$. If $\boldsymbol{R}$ changes sign, so does α, and, as a consequence, the nature of the outer fringes. One concludes that it must be possible to deduce the type of the fault from the nature of the outer fringes. One can always suppose, without loss of generality that $\boldsymbol{g}$ points to the right of the fringes which are parallel to the intersection of surface and fault plane. From the diffraction pattern one deduces the type $\{hkl\}$ of the diffraction spot.

There are, however, three problems:

1) is the fault plane sloping up to the right or to the left;

2) what is the sign combination to be taken for $\{hkl\}$;

3) what is the type of the fault.

Let us consider the situation (L, I) (sloping up to the left, intrinsic fault) (see Fig. 6). The cosinus of the angle β between $\boldsymbol{R}$ and $\boldsymbol{g}_p$, the projection of $\boldsymbol{g}$ on the plane normal to the fringes, is then positive, and thus also the cosinus of the angle between $\boldsymbol{R}$ and $\boldsymbol{g}$.

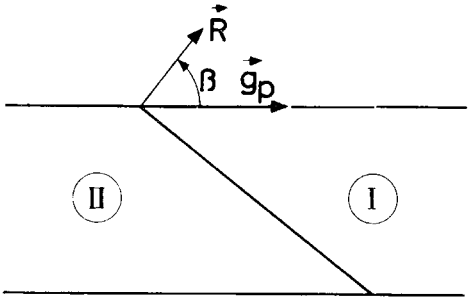


Fig. 6. – Scheme illustrating the notations $\boldsymbol{R}$, $\boldsymbol{g}_p$ and β used in text for stacking fault.

This means that one must satisfy the condition:

$$\boldsymbol{g}\cdot\boldsymbol{R} = \tfrac{1}{3}(h + k + l) > 0\,. \tag{43}$$

If $\cos\beta > 0$, one must take the sign combination leading to (43) and $h + k + l$ no multiple of three. For $\cos\beta < 0$ the sum must be negative.

If (43) is satisfied the sign of $\sin\alpha$ will depend on the type of diffraction spot. There are two classes:

$$B)\ \text{leading to } \sin\alpha > 0\,,$$

$$A)\ \text{leading to } \sin\alpha < 0\,.$$

To class B) belongs $\{111\}$, $\{220\}$, ... since:

$$\alpha = 2\pi\boldsymbol{g}\cdot\boldsymbol{R} = \frac{2\pi}{3}(1 + 1 - 1) = \frac{2\pi}{3} \qquad \text{for } (11\bar{1})\,,$$

$$\alpha = \frac{2\pi}{3}(2 + 2 + 0) = \frac{8\pi}{3} = \frac{2\pi}{3} + 2\pi \qquad \text{for } (220)\,.$$

To class *A*) belongs {200}, {222}, ... since:

$$\alpha = \frac{2\pi}{3}(2+0+0) = \frac{4\pi}{3} \qquad \text{for } (200),$$

$$\alpha = \frac{2\pi}{3}(2+2-2) = \frac{4\pi}{3} \qquad \text{for } (22\bar{2}).$$

Summarizing:

TABLE III.

	class *B*)	class *A*)
$\cos\beta > 0$ $\cos\beta < 0$	$\sin\alpha > 0$ $\sin\alpha < 0$	$\sin\alpha < 0$ $\sin\alpha > 0$

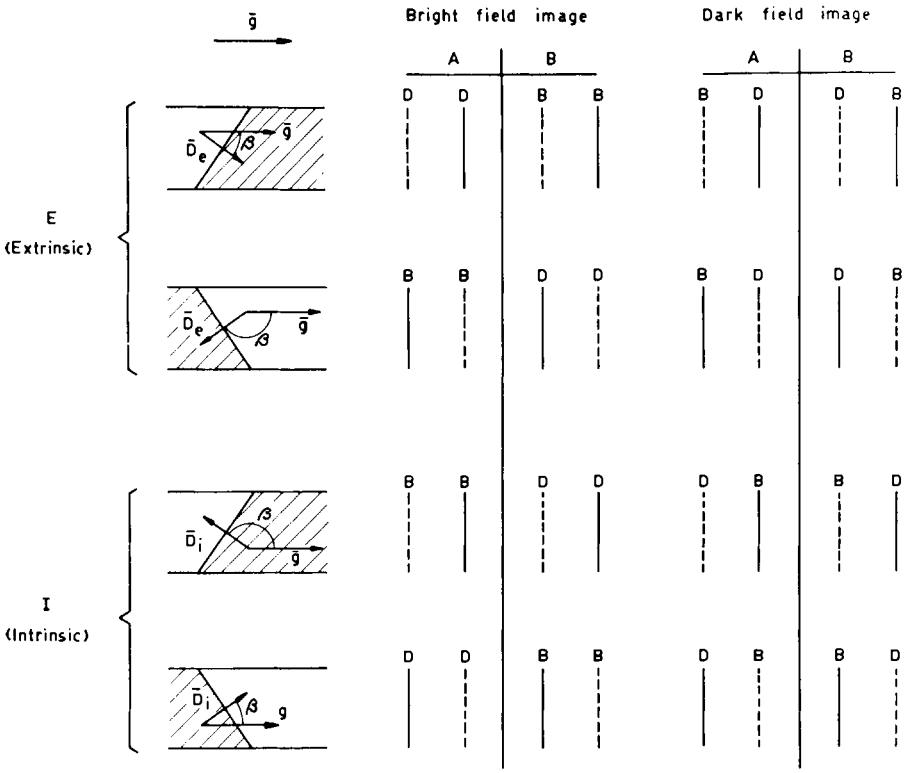


Fig. 7. – Table showing the nature of the first and last fringe of a stacking fault image for all different possibilities. (*B*: bright; *D*: dark; diffraction vector to the right). (Courtesy of *Phys. Stat. Sol.*, **3**, 1563 (1963))

Considering Table I and III, one can now construct Fig. 7. From Fig. 7 follows then the simple rule:

1) Orient the diffraction pattern with respect to the dark field image such that the diffraction vector points to the right of the fringes.

2) If the $\boldsymbol{g}$ vector points from the bright to the dark outer fringes in the dark field, the fault is intrinsic if $\boldsymbol{g}$ belongs to B; extrinsic if $\boldsymbol{g}$ belongs to A.

If the diffraction vector points from dark to bright the conclusions are reversed.

1.7.2. *Microtwins and stacking faults.* – We consider a microtwin as in Fig. 8, with thickness Δ, and we assume that the foil is oriented such that the twinned region is not diffracting.

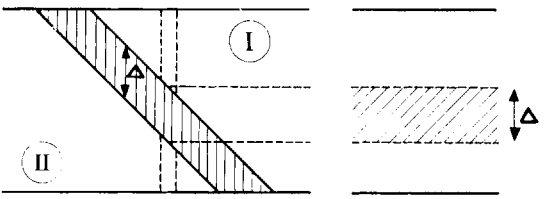


Fig. 8. – Schematic representation of a microtwin with thickness Δ: left: inclined, right: parallel to surface.

Part II is now displaced with respect to part I by $n\boldsymbol{R}$, if $\boldsymbol{R}$ is the twin vector and n the number of planes in the microtwin.

As for a stacking fault, this displacement will introduce a phase angle:

$$n\alpha\,, \qquad \alpha = 2\pi \boldsymbol{g}\cdot\boldsymbol{R}\,. \tag{44}$$

For f.c.c. metals, *e.g.* $\alpha = \pm 2\pi/3$, and thus $n\alpha = 0,\ 2\pi/3,\ 4\pi/3$ (mod. 2π). One can recommence the reasoning of Subsect. **1**'3, with the following modification.

For the amplitude of a beam scattered by the part II, one has now to replace z' in (14*b*) not by z_1 but by $z_1 + \Delta$.

This introduces a supplementary phase shift:

$$2\pi s\Delta\,.$$

The microtwin can then be considered as a planar defect with the same image of a stacking fault, with phase shift angle:

$$\beta = n\alpha + 2\pi s\Delta \,. \tag{45}$$

The image of the overlapping part of a microtwin is similar to the image of a stacking fault. Misinterpretations are possible if Δ is so small that there are no wedge fringes in the nonoverlapping part.

Tilting experiments, however, make it possible to differentiate. During tilting the second term of (45) varies, resulting in a more drastic variation of the contrast of the image than for a stacking fault. It is even possible that $\sin\beta$ changes sign during tilting, resulting in a change of nature of the outer fringes. Moreover, it can be possible to achieve orientations for which β becomes zero. Nearly extinction of the image is then to be expected.

1·7.3. *Domain boundaries in barium titanate.* – When cooling below a transition temperature, ferroelectric domains are formed in barium titanate. In a domain one of the three cubic axis becomes somewhat larger than the other two (tetragonal distortion). In two adjacent domains, the tetragonal c-axes are different. The lattices are then rotated relative to each other over a small angle 2θ in order to bring the two domains in a strict twin relationship, the twin boundary being of the type $\{011\}$.

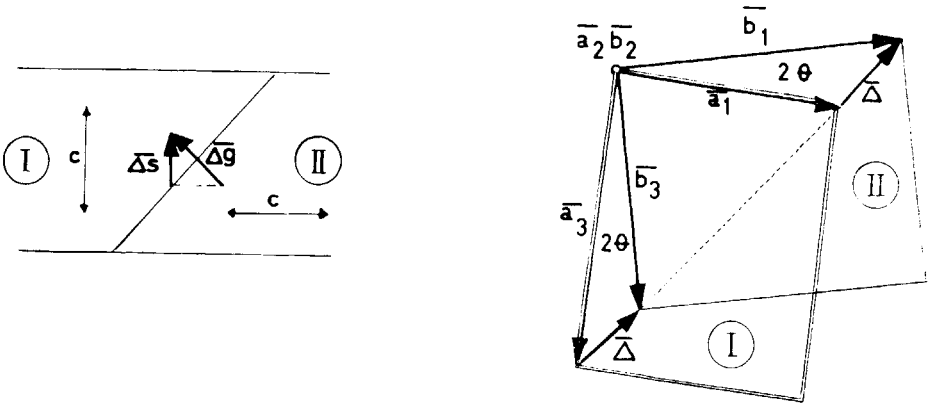
Assume one observes the geometrical situation of Fig. 9*a*), in a foil with surfaces parallel to a cubic plane.

The boundary plane I/II which is a (011) plane is sloping upwards to the right. The tetragonal c-axis is almost horizontal in II and almost vertical in I. We call the base vector of the lattice in crystal I $\boldsymbol{a}_1$, $\boldsymbol{a}_2$ and $\boldsymbol{a}_3$ where $\boldsymbol{a}_3$ is along the c-axis; in crystal II the crystal axes are similarly called $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ and $\boldsymbol{b}_3$ where now $\boldsymbol{b}_1$ is along the c-axis. It is clear that $\boldsymbol{a}_1$ and $\boldsymbol{b}_1$ and also $\boldsymbol{a}_3$ and $\boldsymbol{b}_3$ enclose angles 2θ, whereas $\boldsymbol{a}_2$ and $\boldsymbol{b}_2$ coincide (Fig. 9*b*)). The lower part of the crystal is derived from the upper part by a shear, which we will call $\boldsymbol{\Delta}$. In this particular case it is clear that:

$$\boldsymbol{\Delta} = \alpha(\boldsymbol{a}_1 - \boldsymbol{a}_3)\,, \tag{46}$$

with $\alpha > 0$. We have hereby used the system $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, $\boldsymbol{a}_3$ as the reference system. In this same reference system we can write:

$$\Delta\boldsymbol{a}_1 = \boldsymbol{\Delta}\,, \quad \Delta\boldsymbol{a}_2 = 0\,, \quad \Delta\boldsymbol{a}_3 = \boldsymbol{\Delta} \quad (\Delta\boldsymbol{a}_i = \boldsymbol{b}_i - \boldsymbol{a}_i)\,. \tag{47}$$



(a) (b)

Fig. 9. – *a*) Cross-section through a foil of barium titanate normal to boundary and surface plane. Δg and Δs are indicated. *b*) Reference system used in discussing the geometry of the domain boundary structure in barium titanate. (Courtesy of *Phys. Stat. Sol.*, **5**, 595 (1964).)

We now calculate Δg for a given diffraction vector g. Let the base vectors of the reciprocal lattice of a_1, a_2, a_3 be A_1, A_2 and A_3. The vector g can always be written as:

$$g = hA_1 + kA_2 + lA_3 , \tag{48}$$

where $l = 0$ for the reflections of interest here (the foil plane is (001) in the reference system used). The components of Δg are found by projection on the crystal axes. Since $g \cdot a_1 =$ integer we can write:

$$\Delta g \cdot a_1 = -g \cdot \Delta a_1 = -g \cdot \Delta = \alpha(l - h)$$

and similarly:

$$\Delta g \cdot a_2 = -g \cdot \Delta a_2 = 0 ,$$

$$\Delta g \cdot a_3 = -g \cdot \Delta a_3 = \alpha(l - h) .$$

One finds as a result:

$$\Delta g = \alpha(l - h)(A_1 + A_3) \tag{49a}$$

and hence for the operating reflections ($l = 0$):

$$\Delta g = -\alpha h(A_1 + A_3) . \tag{49b}$$

For a diffraction vector to the right of the intersection line of the boundary plane and the foil plane the angle between $\boldsymbol{a}_1$ and $\boldsymbol{g}$ is acute and $h = \boldsymbol{g} \cdot \boldsymbol{a}_1$ is positive. We conclude that in this case $\Delta \boldsymbol{g}$ has the direction and sense of $-(A_1 + A_3)$, *i.e.* it is as shown in Fig. 9*a*). Since $\Delta s = s_1 - s_2$ is the pro-

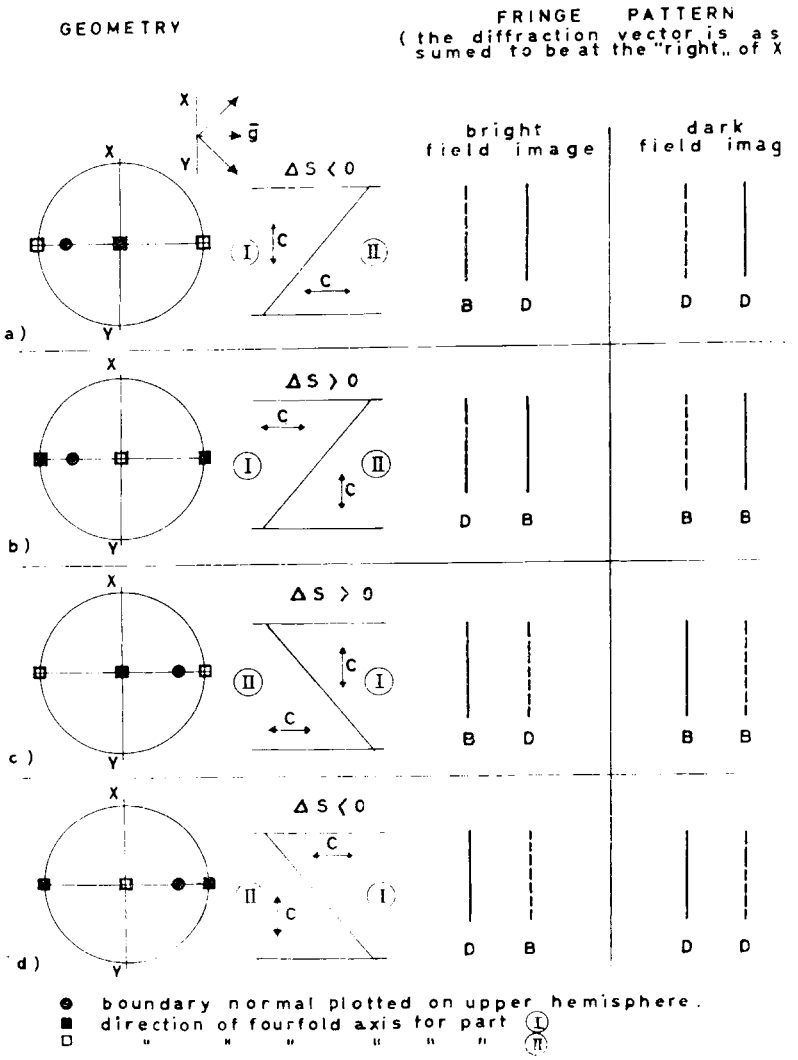


Fig. 10. – Table showing the nature of first and last fringe of the image of a domain boundary in barium titanate for all different possibilities. (*B*: bright; *D*: dark; diffraction vector to the right). (Courtesy of *Phys. Stat. Sol.*, **5**, 595 (1964).)

jection of $\Delta \boldsymbol{g}$ on the normal to the foil plane, we find that Δs is negative (the positive sense of s is downward). From Table II we conclude that the first fringe in the bright field image is dark and the last fringe bright. In the dark field image both first and the last fringe are dark.

If the orientation of the tetragonal axis in the two crystal parts are interchanged, *i.e.* for the configuration of Fig. 10*b*), it is clear that Δ and α change sign and consequently that $\Delta \boldsymbol{g}$ and hence also Δs change sign. A similar reasoning can be made when the contact plane is sloping upwards to the left as in Fig. 10*c*) and *d*). The results for a diffraction vector pointing towards the right of XY are summarized in the table of Fig. 10 where the nature of the first and last fringes is also represented schematically. The geometry is represented by means of a cut perpendicular to the foil and to the boundary and also by means of a stereographic projection on the foil plane. The boundary plane is represented by a full dot, assuming that the pole is in the upper hemisphere. The c-axes are represented by small squares; a full one for part I, an empty one for part II.

If the sense of the sloping of the boundary plane is not required the direction of the c-axis can be deduced from the dark field image only. The following rule becomes immediately obvious when consulting Fig. 10. If in the dark field image the first (and last) fringe is a dark fringe the $\boldsymbol{g}$ vector points towards the region where the tetragonal axis is horizontal. It points towards the region where the c-axis is vertical if the first and last fringe is a bright one.

1·8. Moiré fringes.

If the interface is not a common lattice plane for both parts, one has:

$$\Delta \boldsymbol{g} \cdot \boldsymbol{r}_1 \neq 0$$

in (26).

This term introduces a further strictly periodic intensity modulation. A moiré fringe system appears, superposed on the wedge fringes. The absence of this moiré proves that it was correct to assume that the interface was a common lattice plane.

1·9. Observations.

In the Figs 11 to 27 a few examples are given of observations of stacking faults, antiphase boundaries, microtwins, domain boundaries, in order to illustrate the significant properties of the images. For details we refer to the captions.

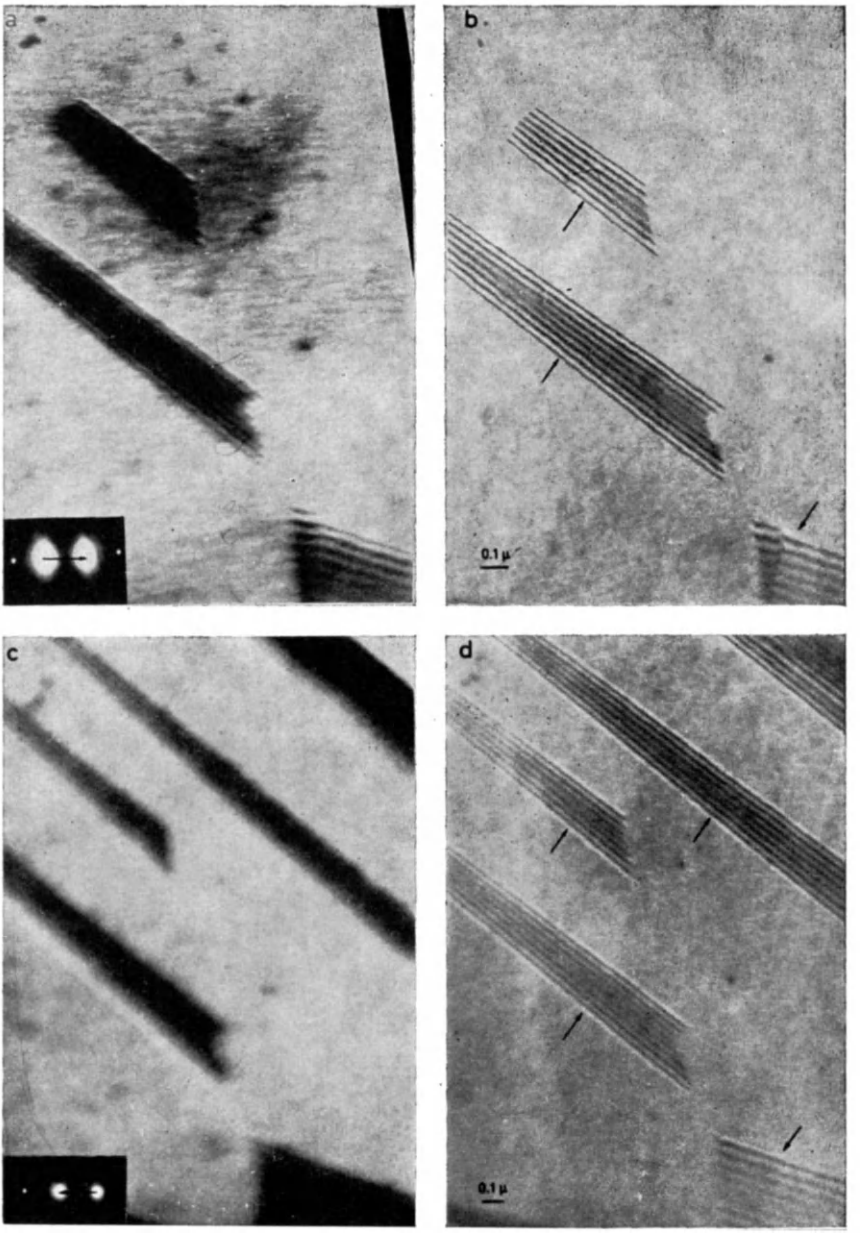


Fig. 11. – Dark and bright field image of a stacking fault in a Cu-Ga foil for two opposite diffraction vectors.

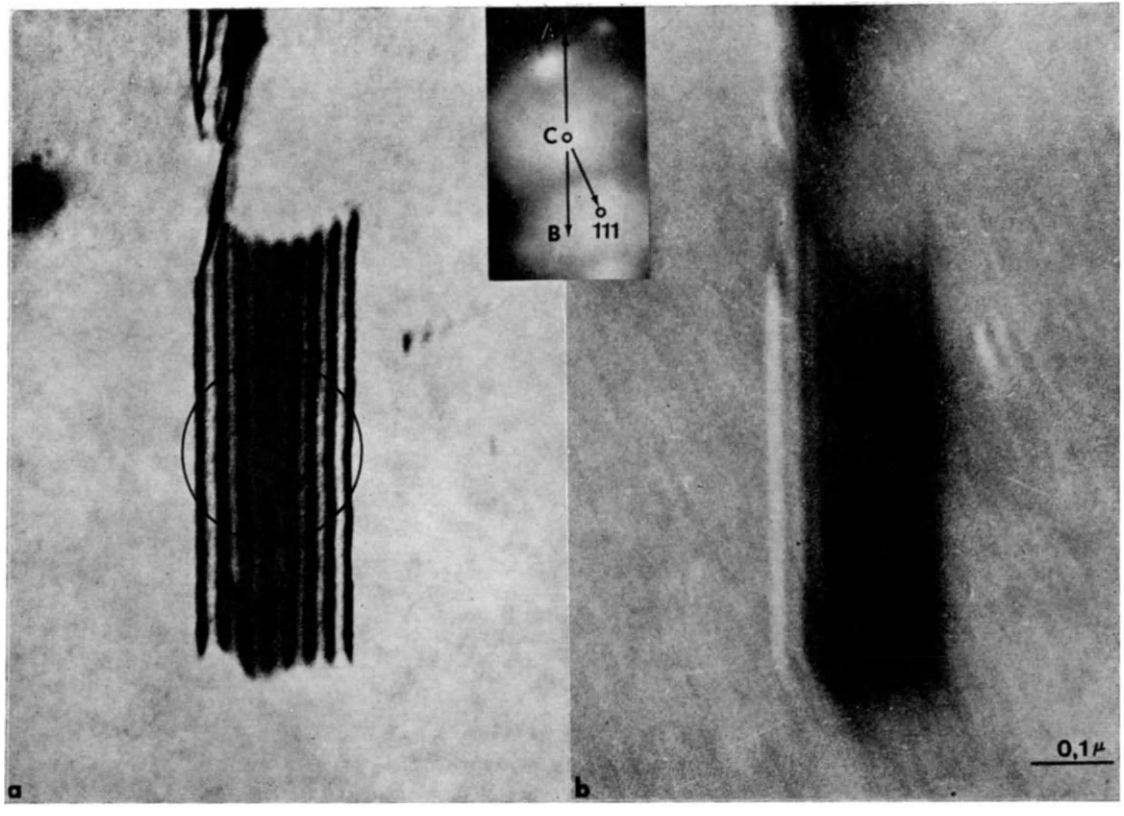


Fig. 12. – Bright and dark field image of a stacking fault in a Cu-Ga foil, illustrating the use of the rule for determining the type of fault (intrinsic). (Courtesy of *Phys. Stat. Sol.*, **3**, 1563 (1963).)

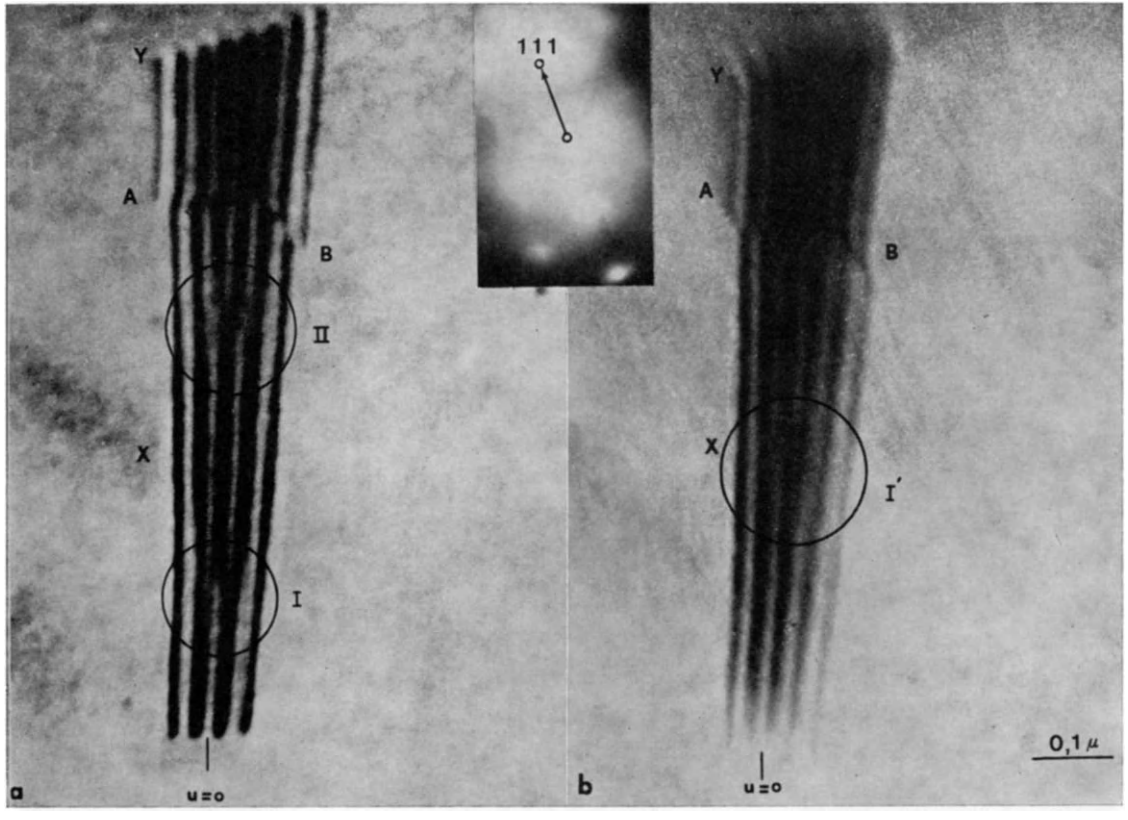


Fig. 13. – Same as for Fig. 12. At both sides of the partial dislocation *AB* the fault is of different nature. (Courtesy of *Phys. Stat. Sol.*, **3**, 1563 (1963).)

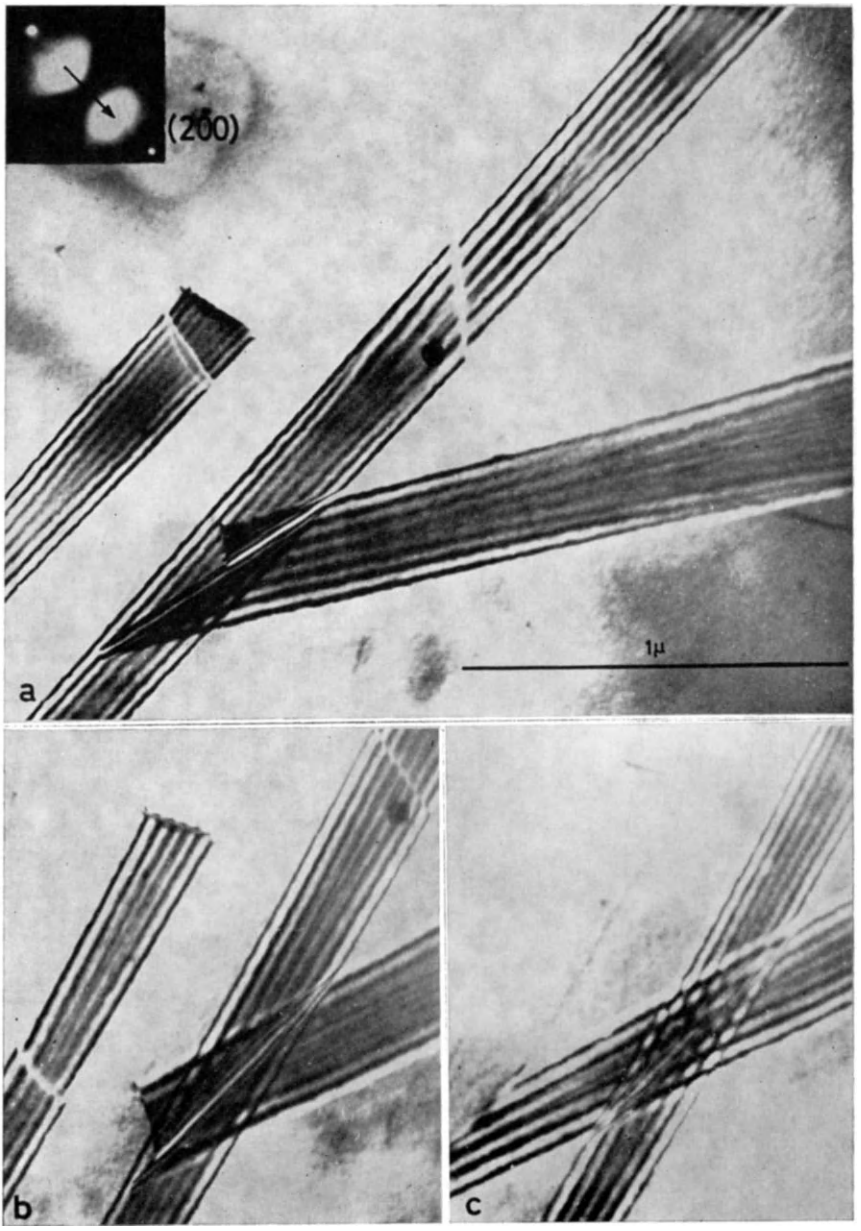


Fig. 14. – Bright field images of stacking faults. Note the poor contrast near the centre in the thicker part. Notice also the images of the intersecting faults.

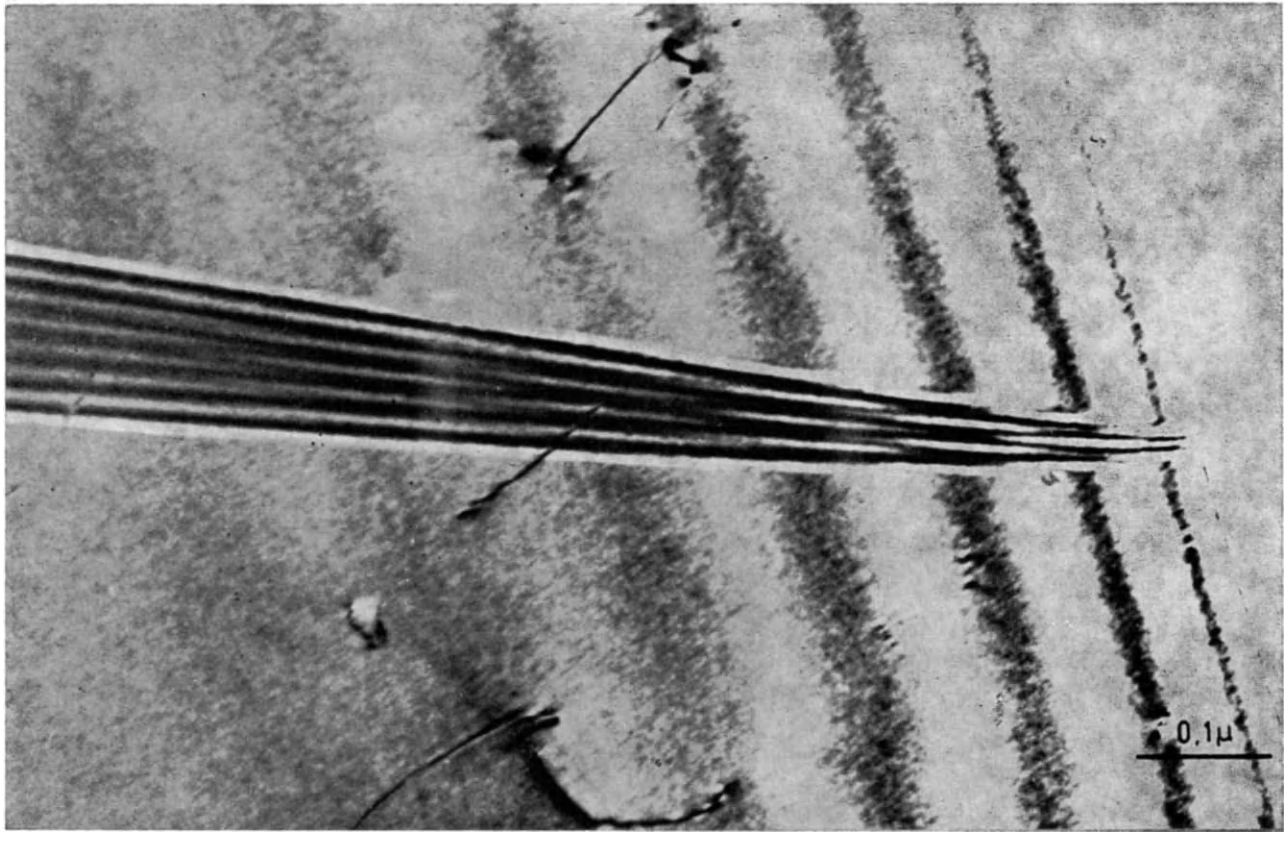


Fig. 15. – Bright field image of a stacking fault in a wedge-shaped foil of stainless steel. Notice the different characteristics of the image in the thinner and the thicker part of the foil.

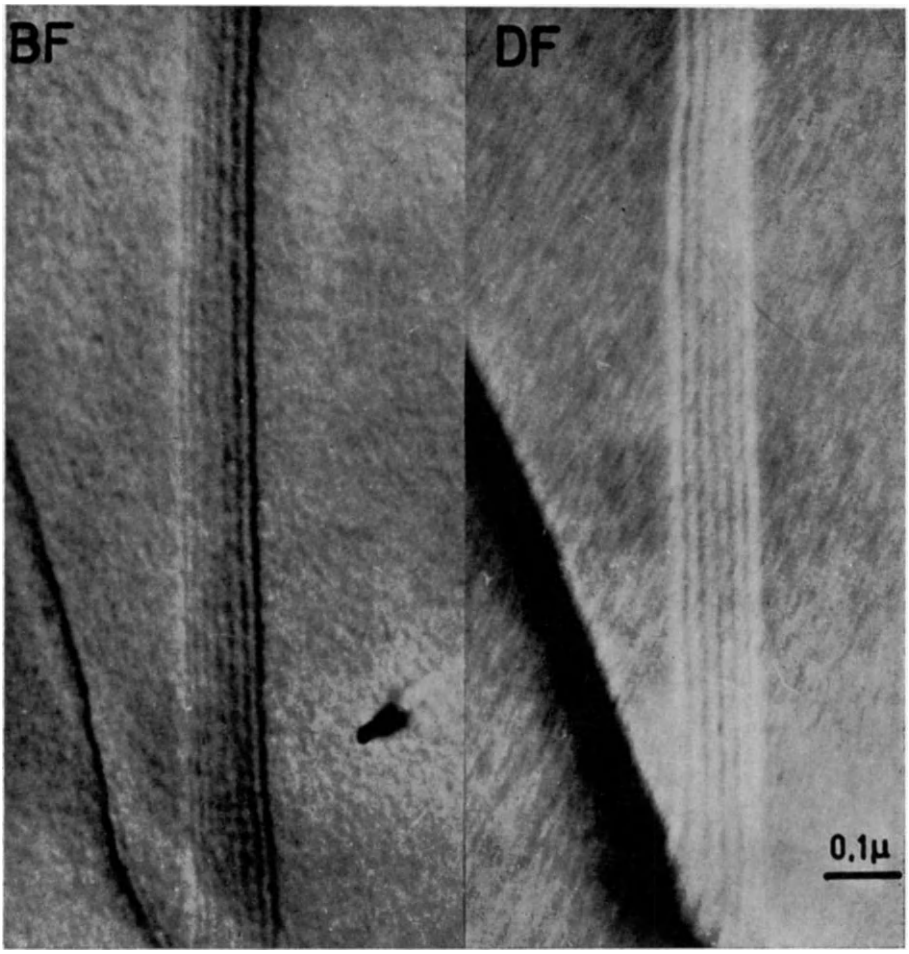


Fig. 16. – Bright and dark field image of a boundary between ordered domains due to impurities in niobium. The BF is asymmetrical and the DF symmetrical. There is only a slight difference in background in the two adjacent domains.

Fig. 17. – Ordered domain boundaries in Nb containing impurities. Note the difference in background in the different domains (bright field image). (Courtesy of *Phys. Stat. Sol.*, **5**, 595 (1964).)

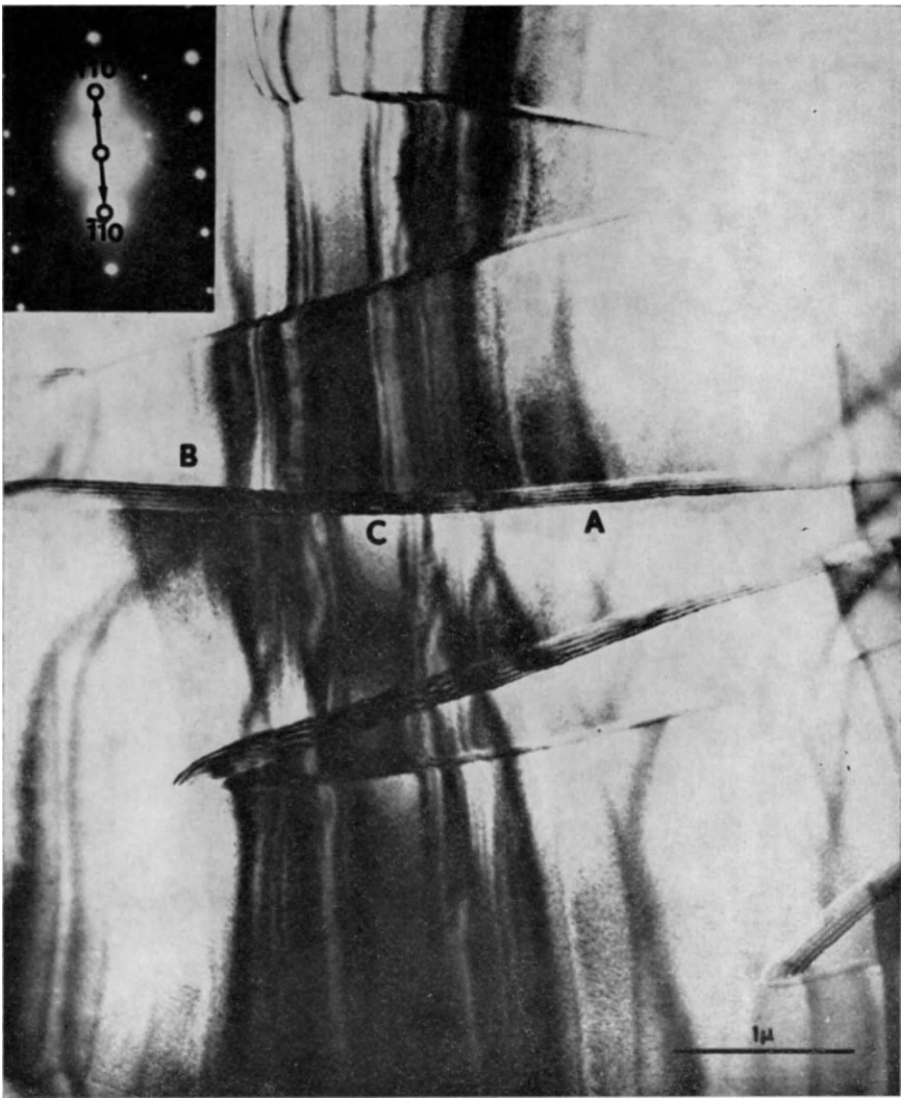


Fig. 18. – Bright field image of a domain wall in niobium. Diffraction pattern of the total area is given as an inset. Notice that the side of the bright fringe is different in *A* and *B*, the active diffraction vector is different in *A* and *B*. (Courtesy of *Phys. Stat. Sol.*, **5**, 595 (1964).)

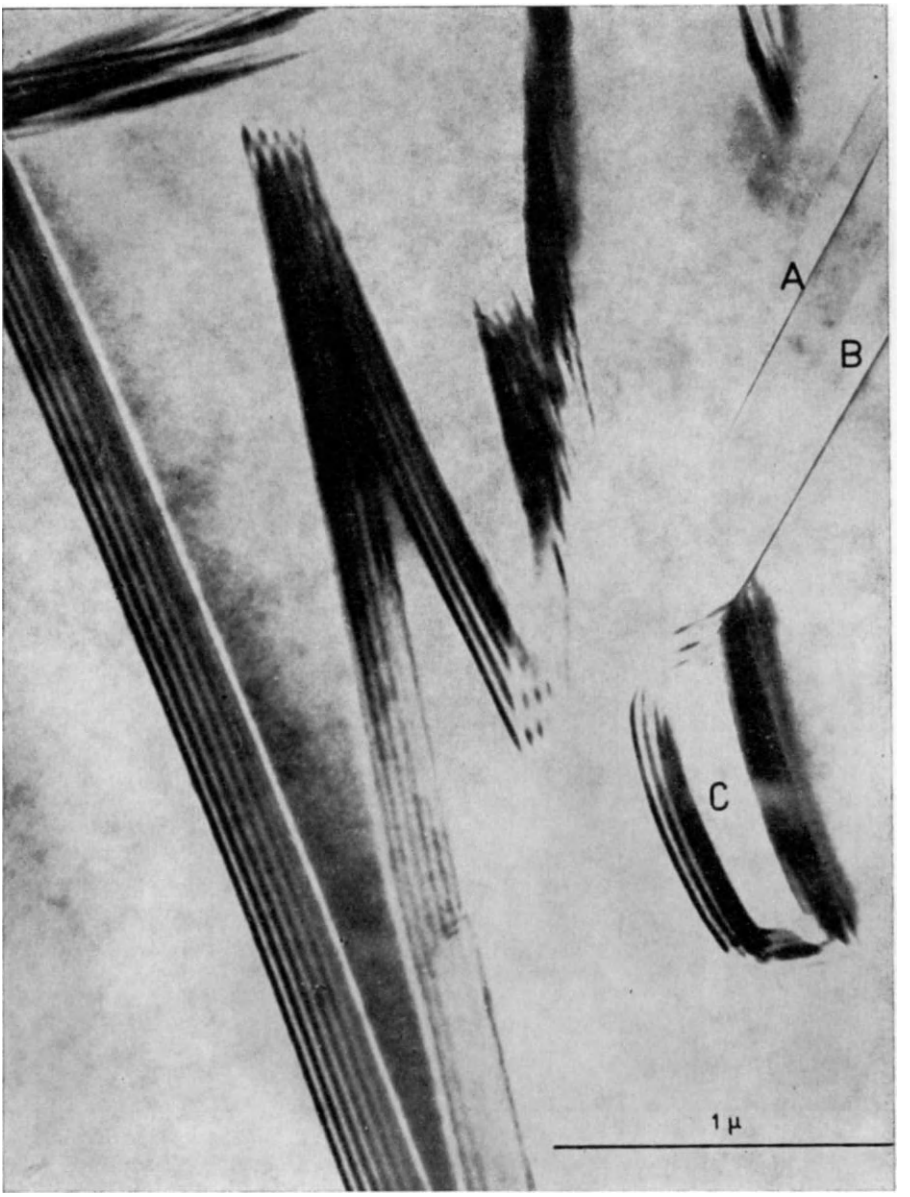


Fig. 19. – Ferroelectric domain boundaries in barium titanate. The first and last fringe are opposite in nature. The two walls are parallel since the bright fringes are at opposite sides (bright field image). (Courtesy of *Phys. Stat. Sol.*, **5**, 595 (1964).)

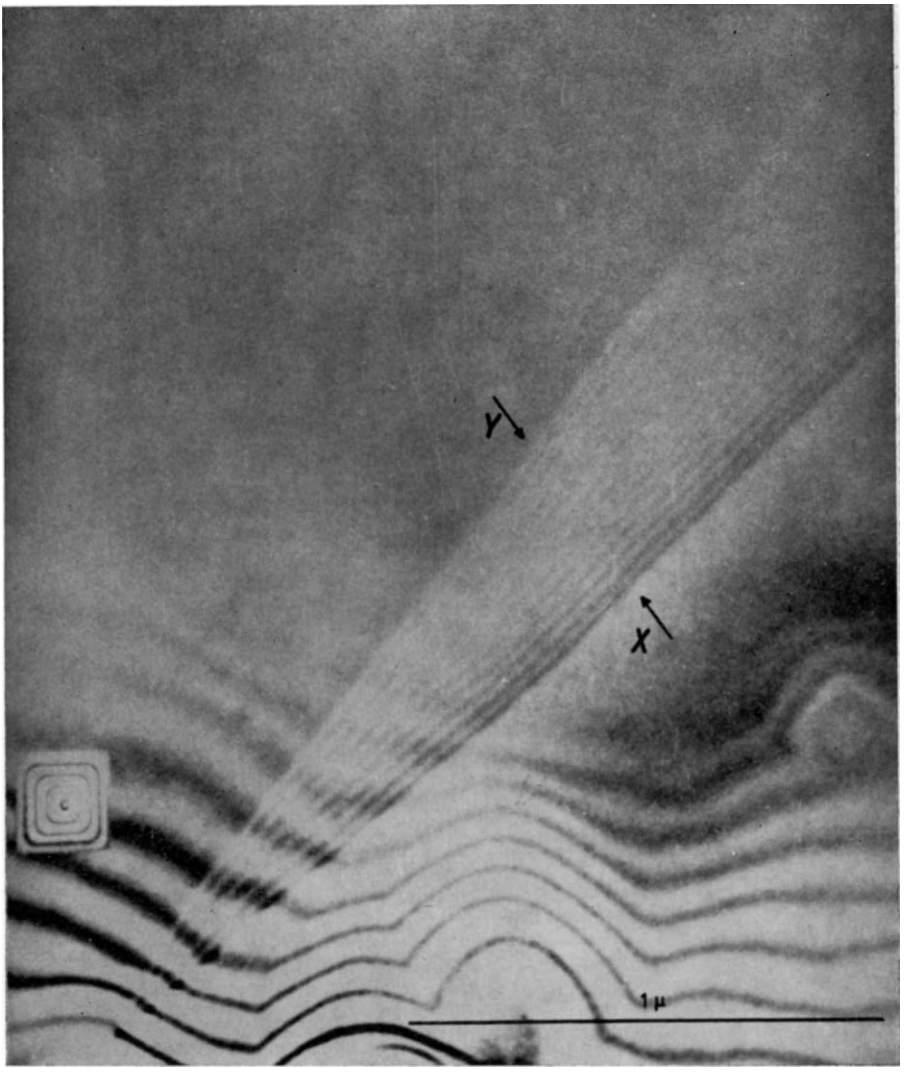


Fig. 20. – Bright field image of domain boundary in nickel oxyde. The fringe spacing is nearly the same at both ends of the pattern (*i.e.* $|s_1| \simeq |s_2|$), but the contrast is different. The contrast changes periodically with foil thickness, it is maximum in the dark thickness contours. (Courtesy of *Phys. Stat. Sol.*, **5**, 595 (1964).)

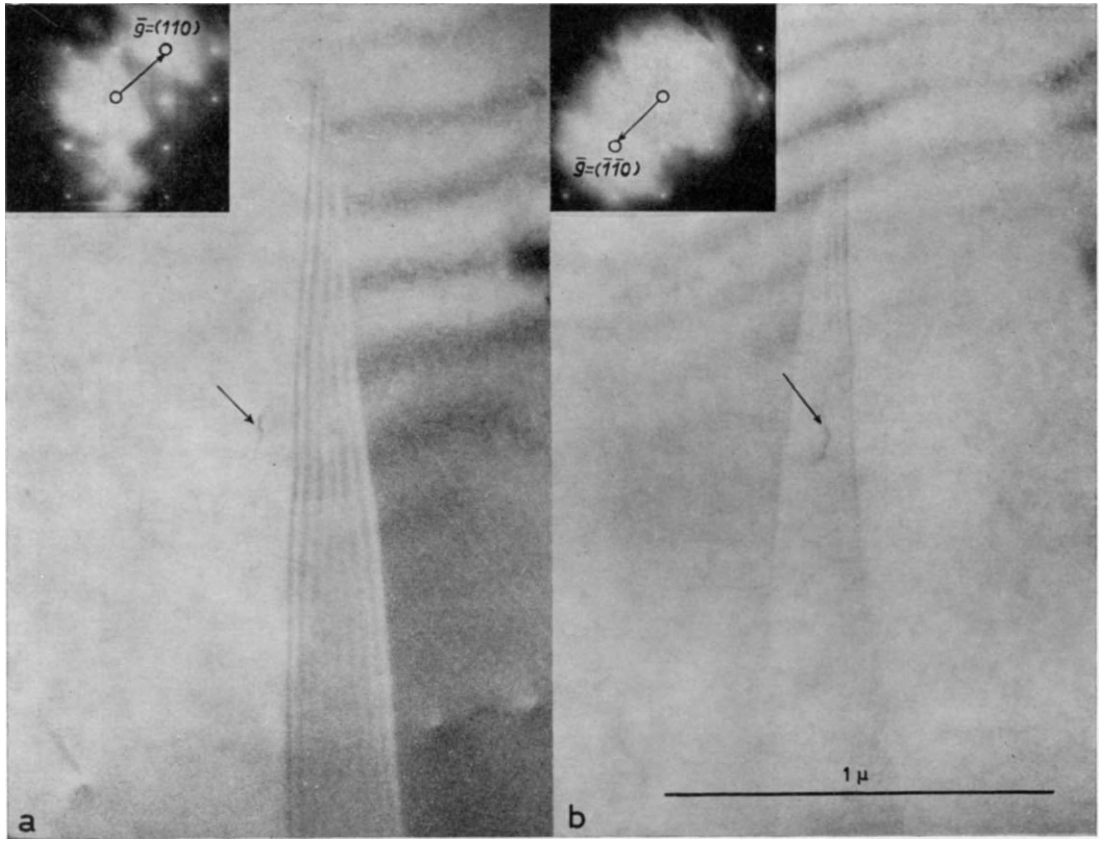


Fig. 21. – Bright field images of the same domain boundary in nickel oxyde for g and $-g$. Notice that the nature of the outer fringes changes. (Courtesy of *Phys. Stat. Sol.* **5**, 595 (1694).)

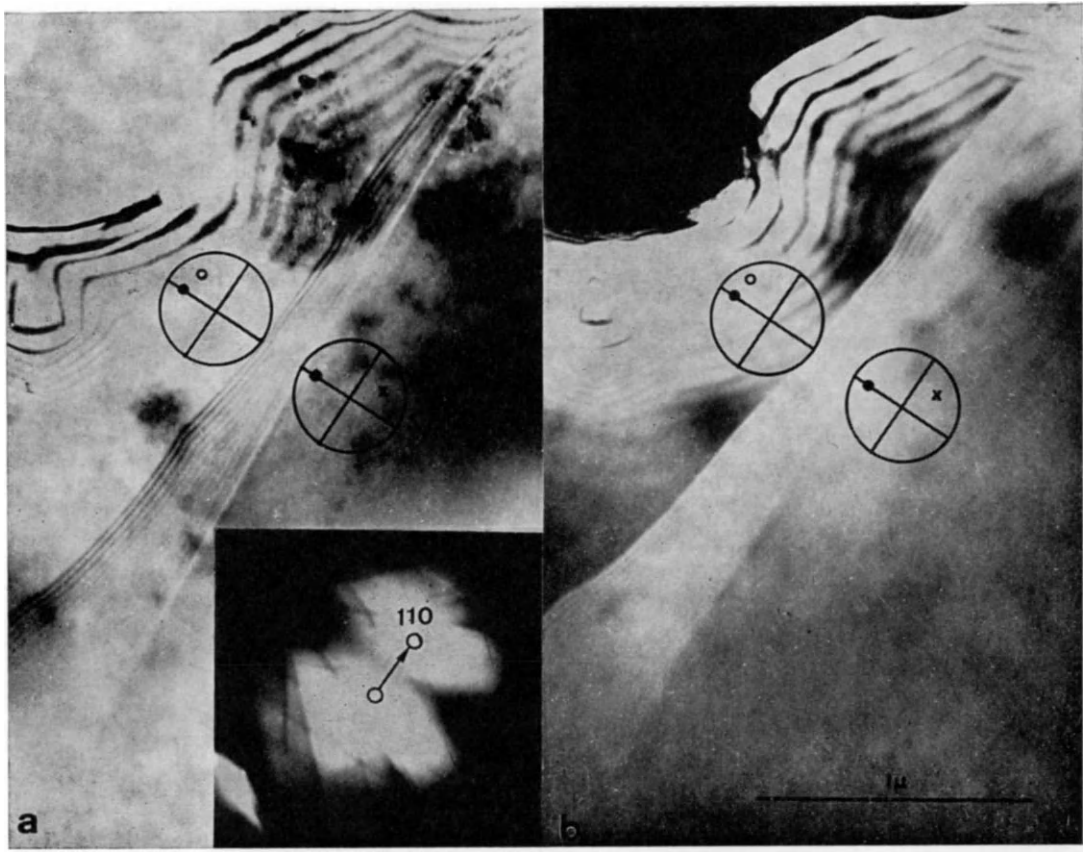


Fig. 22. – Bright and dark field image of ferroelectric domain boundary in nickel oxype. (Courtesy of *Phys. Stat. Sol.*, **5**, 595 (1964).)

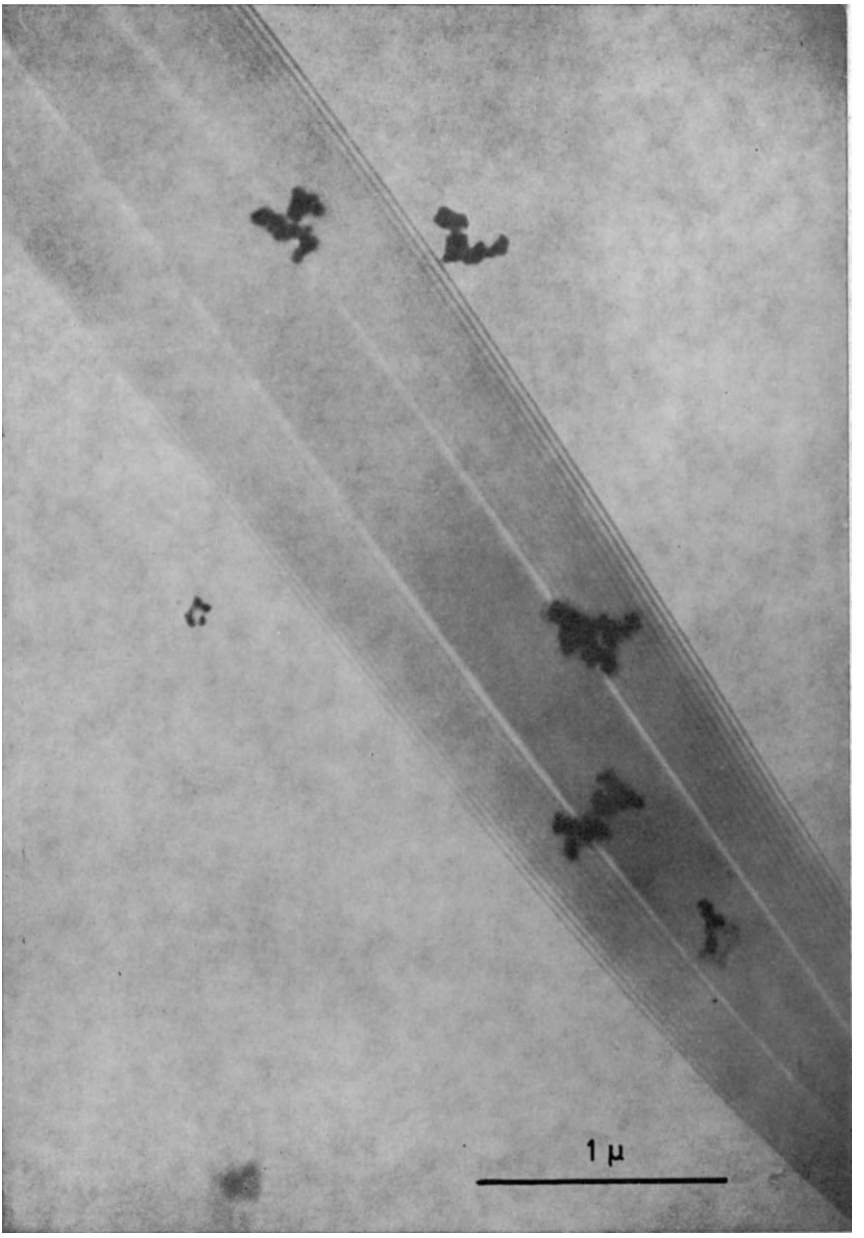


Fig. 23. – Pair of {110} domain boundaries producing tilts of opposite signs (barium titanate, bright field). The complete pattern has a line of symmetry. The two contact planes are parallel. (Courtesy of *Phys. Stat. Sol.*, **5**, 595 (1964).)

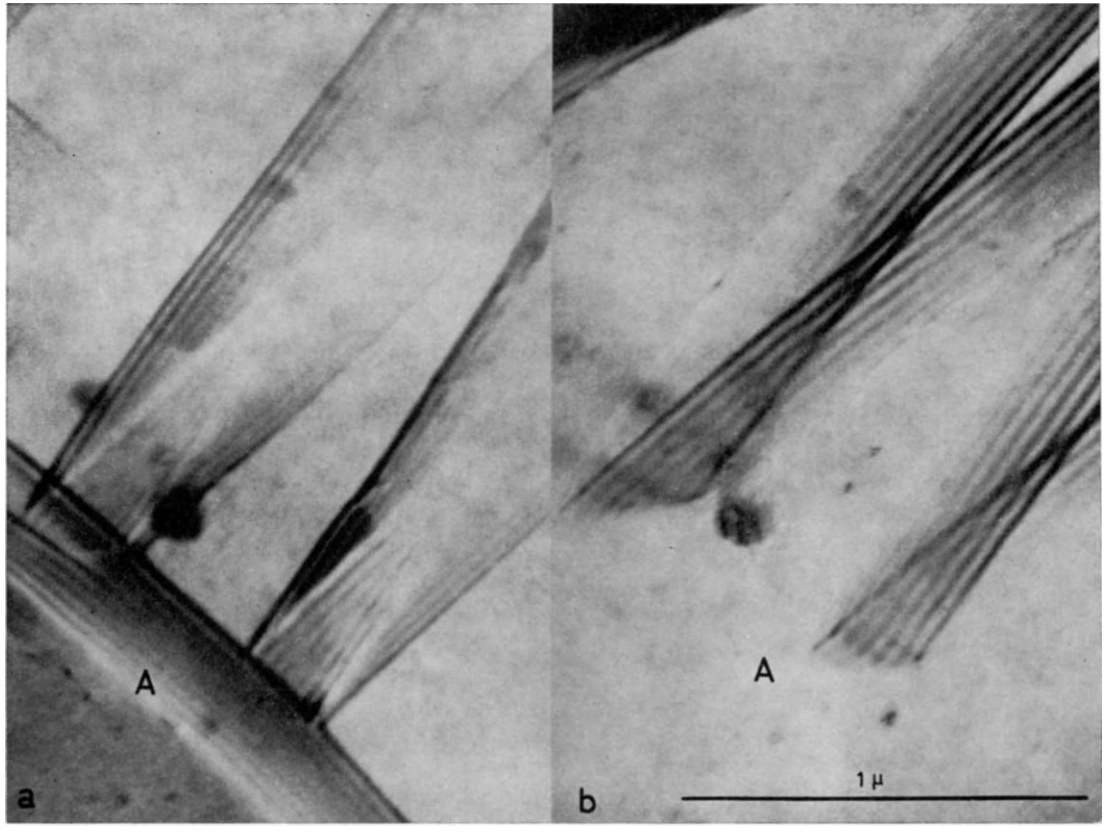


Fig. 24. – Bright field images of domain walls in barium titanate in the same area under two different diffraction conditions: *a*) the wall marked *A* is in contrast; *b*) the wall marked *A* is out of contrast (extinction). In *b*) $\Delta \boldsymbol{g} = 0$, or $\boldsymbol{g} \cdot \boldsymbol{\Delta} = 0$. (Courtesy of *Phys. Stat. Sol.*, **5**, 595 (1964).)

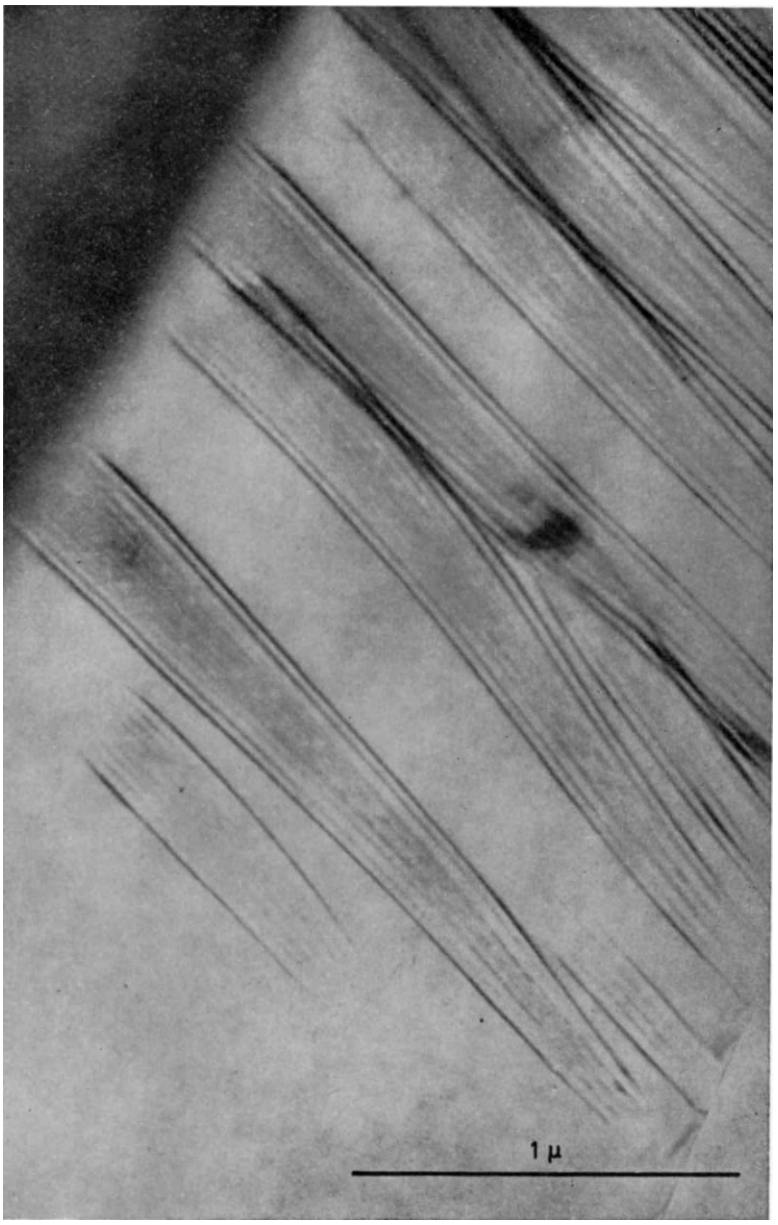


Fig. 25. – Bright field image of close pairs of parallel domain boundaries, producing tilts of opposite sign, in barium titanate. The individual fringe patterns overlap causing a pattern which is effectively symmetrical. The complications in the inner part of the images can be explained if one assumes that a third boundary of different type is present between the boundaries of the pairs. (Courtesy of *Phys. Stat. Sol.*, **5**, 595 (1964).)

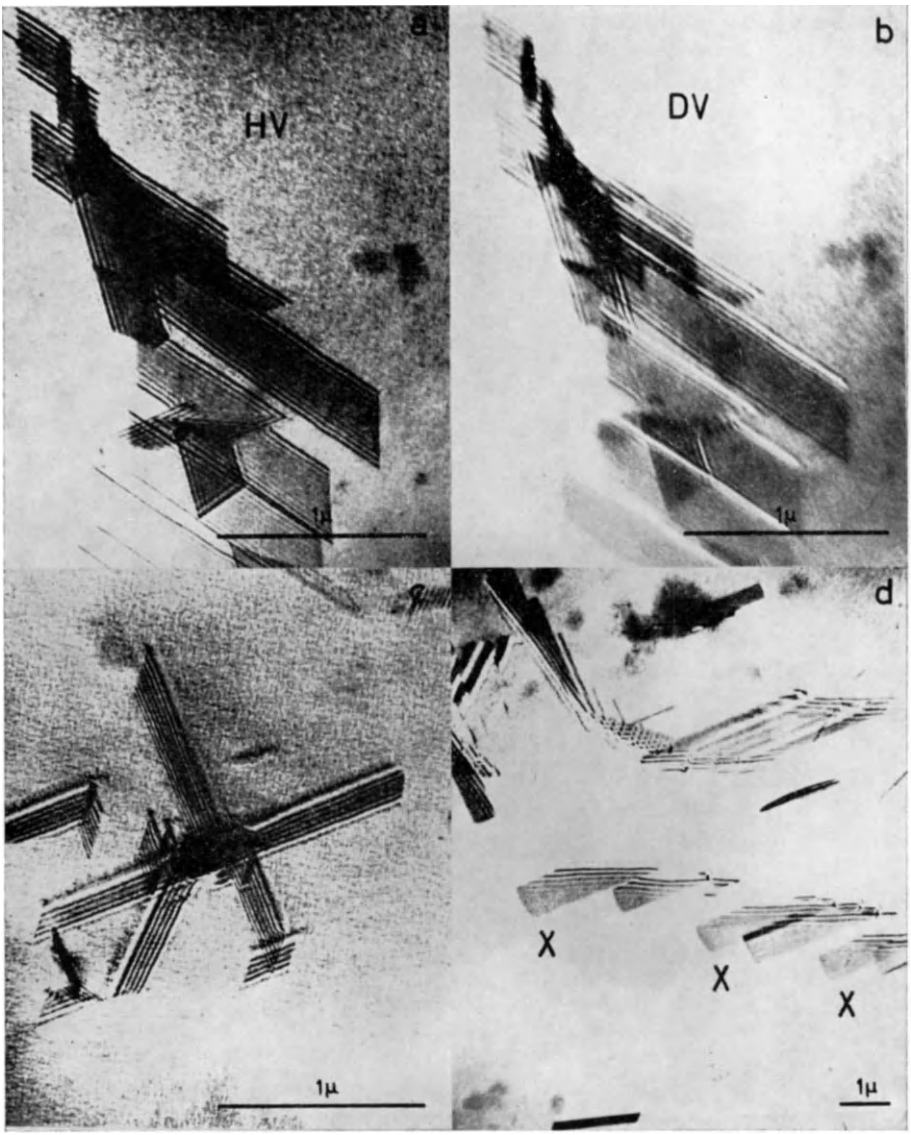


Fig. 26. – Bright (*a*) and dark field (*b*) images of plate-shaped thin precipitates in niobium with impurities. The fringe pattern resulting from the overlap of the individual images of the two boundaries is symmetrical in the bright field and asymmetrical in the dark field.

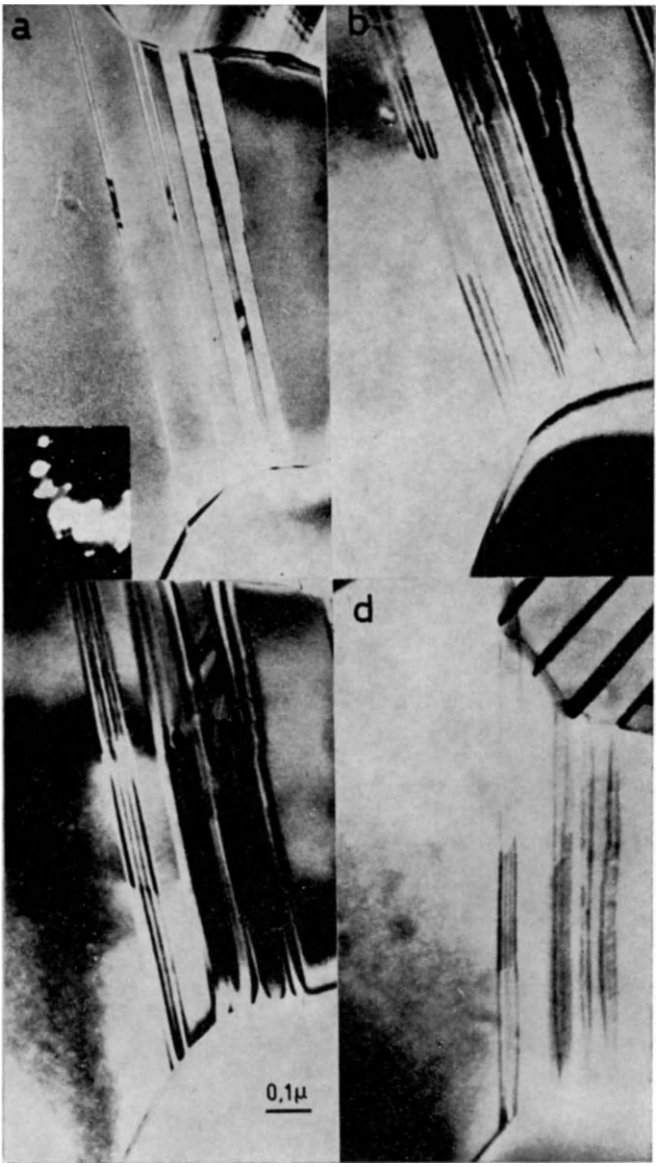


Fig. 27. – Bright field images of micro-twins in rutile, for different inclinations and diffraction conditions. In *a*) the boundaries are seen edge on, in *b*), *c*), *d*), which show the twin in inclined position, discontinuous changes in the pattern associated with sudden variations of the thickness can be seen (*n* varies). One notes that, depending on t hedifffraction condition *s*, the pattern at the overlapping part varies, and may even disappear [*b*) and *d*); $\beta \simeq 0$]. (Courtesy of *Phys. Stat. Sol.*, **9**, 135 (1965).)

REFERENCES (Section 1)

A. ART, R. GEVERS and S. AMELINCKX: *Phys. Stat. Sol.*, **3**, 697 (1963).
A. FOURDEUX, R. GEVERS and S. AMELINCKX: *Phys. Stat. Sol.*, **24**, 195 (1967).
R. GEVERS: *Phil. Mag.*, **7**, 1681 (1962).
R. GEVERS: *Phys. Stat. Sol.*, **3**, 1214 (1963).
R. GEVERS: *Phys. Stat. Sol.*, **3**, 1672 (1963).
R. GEVERS: *Phys. Stat. Sol.*, **3**, 2289 (1963).
R. GEVERS, A. ART and S. AMELINCKX: *Phys. Stat. Sol.*, **3**, 1563 (1963).
R. GEVERS, A. ART and S. AMELINCKX: *Phys. Stat. Sol.*, **7**, 605 (1964).
R. GEVERS, P. DELAVIGNETTE, H. BLANK and S. AMELINCKX: *Phys. Stat. Sol.*, **4**, 383 (1964).
R. GEVERS, P. DELAVIGNETTE, H. BLANK, J. VAN LANDUYT and S. AMELINCKX: *Phys. Stat. Sol.*, **5**, 595 (1964).
R. GEVERS, J. VAN LANDUYT and S. AMELINCKX: *Phys. Stat. Sol.*, **11**, 689 (1965).
R. GEVERS, J. VAN LANDUYT and S. AMELINCKX: *Phys. Stat. Sol.*, **18**, 325 (1966).
H. HASHIMOTO, A. HOWIE and M. J. WHELAN: *Proc. 2nd Eur. Reg. Conf. on Electron Microscopy, Delft 1960* (de Nederlandse Vereniging voor Elektronenmikroskopie, 1960), vol. **1**, p. 207.
H. HASHIMOTO, A. HOWIE and M. J. WHELAN: *Phil. Mag.*, **5**, 967 (1960).
H. HASHIMOTO, A. HOWIE and M. J. WHELAN: *Proc. Roy. Soc.*, A **269**, 80 (1962).
H. HASHIMOTO, M. MANNAMI and T. NAIKI: *Phil. Trans.*, A **253**, 459 (1961).
A. HOWIE: *Met. Rev.*, **6**, 467 (1961).
A. HOWIE and M. J. WHELAN: *Proc. 2nd Eur. Reg. Conf. on Electron Microscopy, Delft 1960* (de Nederlandse Vereniging voor Elektronenmikroskopie, 1960), vol. **1**, p. 181.
A. HOWIE and M. J. WHELAN: *Proc. 2nd Eur. Reg. Conf. on Electron Microscopy,Delft 1960* (de Nederlandse Vereniging voor Elektronenmikroskopie, 1960), vol. **1**, p. 194.
A. HOWIE and M. J. WHELAN: *Proc. Roy. Soc.*, A **263**, 217 (1961).
G. REMAUT, R. GEVERS and S. AMELINCKX: *Phys. Stat. Sol.*, **20**, 613 (1967).
G. REMAUT, R. GEVERS, A. LAGASSE and S. AMELINCKX: *Phys. Stat. Sol.*, **10**, 121 (1965).
G. REMAUT, R. GEVERS, A. LAGASSE and S. AMELINCKX: *Phys. Stat. Sol.*, **13**, 125 (1966).
F. SECCO D'ARAGONA, P. DELAVIGNETTE, R. GEVERS and S. AMELINCKX: *Phys. Stat. Sol.*, **31**, 739 (1969).
J. VAN LANDUYT: *Phys. Stat. Sol.*, **16**, 585 (1966).
J. VAN LANDUYT, R. GEVERS and S. AMELINCKX: *Phys. Stat. Sol.*, **7**, 519 (1964).
J. VAN LANDUYT, R. GEVERS and S. AMELINCKX: *Phys. Stat. Sol.*, **9**, 135 (1965).
J. VAN LANDUYT, R. GEVERS and S. AMELINCKX: *Phys. Stat. Sol.*, **18**, 167 (1966).
M. J. WHELAN and P. B. HIRSCH: *Phil. Mag.*, **2**, 1121 (1957).
M. J. WHELAN and P. B. HIRSCH: *Phil. Mag.*, **2**, 1303 (1957).

2. Fine structure of diffraction spots.

2·1. General formulation.

Suppose that one has been able to calculate the wave function of the electron beams passing through a plate-shaped foil, with thickness z. One can always take the origin in the back surface, and let $(\boldsymbol{e}_x, \boldsymbol{e}_y)$ be orthonormal base vectors in that plane. The unit vector $\boldsymbol{e}_z$ is taken perpendicular to the surface, in the sense of propagation of the electrons. A point of the back surface is given by:

$$\boldsymbol{r}_0 = x\boldsymbol{e}_x + y\boldsymbol{e}_y. \tag{50}$$

We note for the wave function inside the crystal:

$$\Psi(\boldsymbol{r}) \exp\left[i2\pi \boldsymbol{k}_0 \cdot \boldsymbol{r}\right], \tag{51}$$

where $\boldsymbol{k}_0$ is the wave vector of the incident beam.

After the back surface, the wave function is a superposition of plane waves with wave vectors $\boldsymbol{k}$, satisfying:

$$k^2 = k_0^2 \tag{52}$$

(condition for elastic scattering).

The condition (52) means that the endpoint of $\boldsymbol{k}$ must lie on the reflection sphere.

One can always note:

$$\boldsymbol{k} = \boldsymbol{k}_0 + \boldsymbol{\omega}, \tag{53}$$

where $\boldsymbol{\omega}$ is any vector joining the origin of the reciprocal space with a point of the reflection sphere.

It is always possible to decompose $\boldsymbol{\omega}$ as follows:

$$\boldsymbol{\omega} = \boldsymbol{\omega}_{\parallel} + \omega_{\perp} \boldsymbol{e}_z, \tag{54}$$

where $\boldsymbol{\omega}_{\parallel}$ is the projection of $\boldsymbol{\omega}$ in the back surface. From (54) follows that it sufficies to give $\boldsymbol{\omega}_{\parallel}$.

For the wave function after the crystal, one can write:

$$\Psi_*(\boldsymbol{r}) = \int\!\!\int_{\Sigma} A(\boldsymbol{\omega}_{\parallel}) \exp\left[i2\pi(\boldsymbol{k}_0 + \boldsymbol{\omega}_{\parallel} + \omega_{\perp}\boldsymbol{e}_z)\cdot\boldsymbol{r}\right] \mathrm{d}^2\boldsymbol{\omega}_{\parallel}, \tag{55}$$

where the integral is extended over the back surface Σ. One must express now that the wave function is continuous at the back surface, *i.e.*

$$\Psi(\boldsymbol{r}_0)\exp[i2\pi\boldsymbol{k}_0\cdot\boldsymbol{r}_0]=\Psi_*(\boldsymbol{r}_0)\,. \tag{56}$$

From (56), (51) and (55) follows then, taking into account that:

$$\boldsymbol{r}_0\cdot\boldsymbol{e}_z=0\,,$$

$$\Psi(\boldsymbol{r}_0)=\iint_{\Sigma}A(\boldsymbol{\omega}_{\parallel})\exp[i2\pi\boldsymbol{\omega}_{\parallel}\cdot\boldsymbol{r}_0]\,\mathrm{d}^2\boldsymbol{\omega}_{\parallel}\,. \tag{57}$$

The expression (57) signifies that Ψ is the bi-dimensional Fourier transform of A.

One finds then from (57) by inverse transformation:

$$A(\boldsymbol{\omega}_{\parallel})=\iint_{\Sigma}\Psi(\boldsymbol{r}_0)\exp[-i2\pi\boldsymbol{\omega}_{\parallel}\cdot\boldsymbol{r}_0]\,\mathrm{d}^2\boldsymbol{r}_0\,. \tag{58}$$

The formula (58) means that the diffracted beams are formed by the interference of spherical wavelets emitted by the points $\boldsymbol{r}_0$ of the back surface, with amplitude $\Psi(\boldsymbol{r}_0)$.

2·2. The different beams.

The wave function Ψ is a superposition of Bloch waves, *i.e.*

$$\Psi(\boldsymbol{r}_0)=\Psi_0(\boldsymbol{r}_0)+\sum_g\Psi_g(\boldsymbol{r}_0)\exp[i2\pi\boldsymbol{g}\cdot\boldsymbol{r}_0]\,. \tag{59}$$

The first term corresponds to the transmitted beams, whereas the terms Ψ_g represent the diffracted beams.

Introducing (59) into (58), one obtains:

$$A(\boldsymbol{\omega}_{\parallel})=A_0(\boldsymbol{u})+\sum_g A_g(\boldsymbol{u})\,, \tag{60}$$

if

$$A_0(\boldsymbol{u})=\iint\Psi_0(\boldsymbol{r}_0)\exp[-i2\pi\boldsymbol{u}\cdot\boldsymbol{r}_0]\,\mathrm{d}^2\boldsymbol{r}_0\,, \tag{61a}$$

$$A_g(\boldsymbol{u})=\iint\Psi_g(\boldsymbol{r}_0)\exp[-i2\boldsymbol{u}\cdot\boldsymbol{r}_0]\,\mathrm{d}^2\boldsymbol{r}_0\,, \tag{61b}$$

with

$$\boldsymbol{\omega}_{\parallel}=\boldsymbol{g}_{\parallel}+\boldsymbol{u}\,. \tag{61c}$$

For a perfect plate-shaped foil, Ψ_0 and Ψ_g are constants, and it follows then from (61)

$$\boldsymbol{u} = 0,$$

giving rise to the sharp diffraction spots.

The functions Ψ_0 and Ψ_g are no longer constant if the foil contains defects. There are two possibilities:

1) $A_g(\boldsymbol{u})$ is a single broadened function. The diffraction spot is broadened and has a certain structure.

2) $A_g(\boldsymbol{u})$ consists of several separated sharp peak functions. The diffracted beam $\boldsymbol{g}$ has then several components.

We shall concentrate on the second possibility.

2·3. The diffraction pattern of a fringe pattern.

One can choose now the y-axis along the fringes, and let:

$$\boldsymbol{u} = u\boldsymbol{e}_x + v\boldsymbol{e}_y . \tag{62}$$

The functions Ψ_0 and Ψ_g do not depend on y, and one finds then from (61):

$$A_0(u, v) = A_0(u)\,\delta(v), \qquad A_0(u) = \int \Psi_0(u) \exp[-i2\pi ux]\,\mathrm{d}x , \tag{63a}$$

$$A_g(u, v) = A_g(u)\,\delta(v), \qquad A_g(u) = \int \Psi_g(u) \exp[-i2\pi ux]\,\mathrm{d}x , \tag{63b}$$

where

$$\delta(v) = \int \exp[-i2\pi vy]\,\mathrm{d}y . \tag{63c}$$

If the dimension in the y-direction is taken sufficiently large, $\delta(v)$ is the delta function.

The condition $v = 0$ leads to the following conclusion: the different components of the diffracted beam $\boldsymbol{g}$ lie on a line through $\boldsymbol{g}$ normal to the fringes. The spots are elongated in the x-direction.

In most cases, Ψ_0 and Ψ_g are of the form:

$$\Psi_0(x) = \sum_i C_0^{(i)} \exp[i2\pi u_i x], \tag{64a}$$

$$\Psi_g(x) = \sum_i C_g^{(i)} \exp[i2\pi u_i x]. \tag{64b}$$

Introducing (64) into (63) gives:

$$A_0(u) = \sum_i C_0^{(i)} P(u-u_i)\,, \tag{65a}$$

$$A_g(u) = \sum_i C_g^{(i)} P(u-u_i)\,, \tag{65b}$$

where

$$P(u) = \int_{-a/2}^{+a/2} \exp\,[i2\pi ux]\,\mathrm{d}x = \frac{\sin \pi ua}{\pi u} \tag{65c}$$

is a sharp function centered around $u = 0$ (a: projected width of the fault).

The different components of a spot, are situated at positions $u = u_i$, for the transmitted beam as well as for the different diffracted beams. The geometrical configuration of the satellites is the same for all spots. The relative intensities do, however, differ from spot to spot.

2·4. Stacking fault: two-beam case. (*)

2·4.1. *Transmitted beam.* – For the amplitude of the beam transmitted by a foil containing a stacking fault, one has:

$$T = T_1 T_2 + S_1 S_2 \exp\,[i\alpha]\,,$$

or if T_0 is the amplitude in the absence of the fault:

$$T = T_0 - (1 - \exp\,[i\alpha])\, S_1 S_2\,.$$

Explicitly:

$$T = T_0 + \frac{1-\exp\,[i\alpha]}{(\sigma\xi)^2} \sin \pi\sigma z_1 \sin \pi\sigma z_2\,,$$

or

$$T = \left(T_0 - \frac{1-\exp\,[i\alpha]}{2(\sigma\xi)^2} \cos \pi\sigma z_0\right) + \frac{1-\exp\,[i\alpha]}{2(\sigma\xi)^2} \cos \pi\sigma(z_1 - z_2)\,.$$

The first term gives the contribution of the fault area to the main transmitted spot at $u = 0$. The second term can be rewritten, if one introduces

$$z_1 = \left(\frac{a}{2} + x\right) \mathrm{tg}\,\psi\,, \qquad z_2 = \left(\frac{a}{2} - x\right) \mathrm{tg}\,\psi \qquad (\psi\text{: slope angle}) \text{ (**)}, \tag{66}$$

(*) For notations see Section 1.

(**) The x-axis points from intersection of stacking fault plane with front surface to the intersection with the back surface.

$$\frac{1-\exp[i\alpha]}{4(\sigma\xi)^2}\exp[i2\pi\sigma\,\mathrm{tg}\,\psi x]+\exp[-i2\pi\sigma\,\mathrm{tg}\,\psi x]\,. \tag{67}$$

One concludes:

1) The two satellite spots of the transmitted beam are situated at:

$$u=\sigma\,\mathrm{tg}\,\psi \quad \text{and} \quad u=-\sigma\,\mathrm{tg}\,\psi\,, \tag{68a}$$

$$\sigma=\frac{1}{\xi}(1+\omega^2)^{\frac{1}{2}}, \qquad \omega=s\xi\,. \tag{68b}$$

They lie at symmetrical positions with respect to the main transmitted beam.

The positions depend on the extinction distance, the slope angle and the exact orientation, but not on α.

If one tilts away from the exact Bragg orientation, the distances of satellite spots to the main spot increase. This distance does not depend on the sense of inclination, since $\sigma(s)=\sigma(-s)$.

2) the intensities of the satellite beams are proportional to:

$$\frac{1}{4}\,\frac{\sin^2\alpha/2}{(\sigma\xi)^4}$$

and they, consequently, decrease rapidly if $|s|$ increases. The two satellite beams have some intensity, whatever the crystal orientation.

2·4.2. *Scattered beam.* – For the scattered beam, one has:

$$S=T_1S_2+S_1T_2^{-}\exp[i\alpha]\,,$$

or

$$S=S_0-(1-\exp[i\alpha])S_1T_2^{-}=$$

$$=S_0-(1-\exp[i\alpha])\frac{i}{\sigma\xi}\sin\pi\sigma z_1\left(\cos\pi\sigma z_2+i\frac{s}{\sigma}\sin\pi\sigma z_2\right).$$

The terms corresponding to the satellite beams are:

$$-\frac{(1-\exp[i\alpha])}{4(\sigma\xi)}\left(1-\frac{s}{\sigma}\right)\exp[i\pi\sigma(z_1-z_2)]$$

and

$$\frac{(1-\exp[i\alpha])}{4(\sigma\xi)}\left(1+\frac{s}{\sigma}\right)\exp[-i\pi\sigma(z_1-z_2)]\,.$$

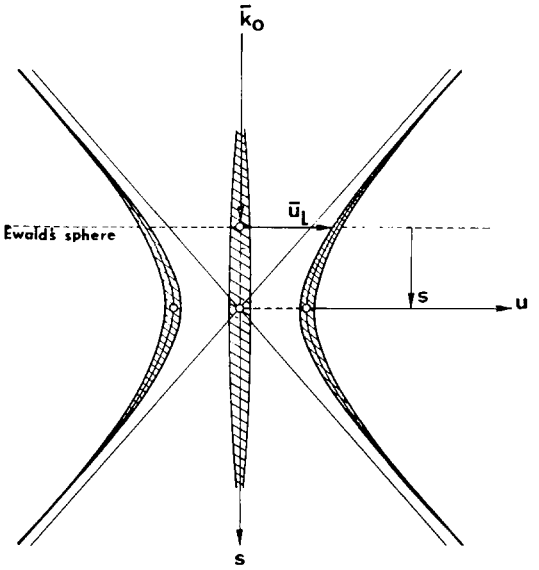


Fig. 28. – Graphical representation for the positions of the satellites around the transmitted beam. The widths of the cross-hatched strip is a measure for the intensity variation. (Courtesy of *Phys. Stat. Sol.*, **18**, 343 (1966).)

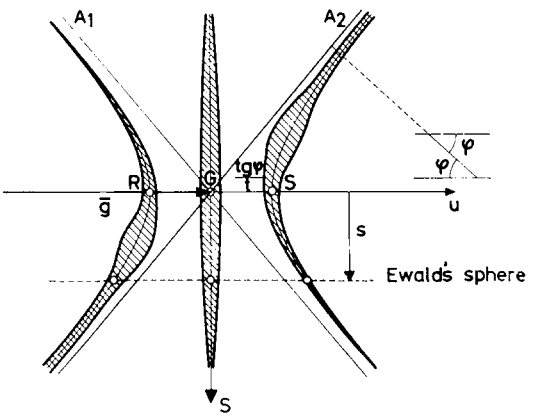


Fig. 29. – Geometry and intensity of the satellites in the case of stacking faults. From this drawing the position and intensity of the satellites can be deduced as a function of s. The width of the cross hatched part along the hyperbola is a measure of the intensity variation as a function of s, as it is also represented in Fig. 30. The dotted line parallel to g represents the Ewald sphere. (Courtesy of *Phys. Stat. Sol.*, **18**, 343 (1966).)

The geometry is the same as for the transmitted beam. However, the intensities are now proportional to:

$$\frac{1}{4}\frac{\sin^2\alpha/2}{(\sigma\xi)^2}\left(1-\frac{s}{\sigma}\right)^2 \quad \text{and} \quad \frac{1}{4}\frac{\sin^2\alpha/2}{(\sigma\xi)^2}\left(1+\frac{s}{\sigma}\right)^2 \tag{69}$$

and are different unless $s = 0$.

One obtains for the ratio of the intensities:

$$\frac{I(u=\sigma\,\mathrm{tg}\,\psi)}{I(u=-\sigma\,\mathrm{tg}\,\psi)} = \left(\frac{1-s/\sigma}{1+s/\sigma}\right)^2.$$

Reflection sphere constructions showing the positions of the satellite beams for varying s are given in Fig. 28 and 29, while Fig. 30 gives a schematic representation of the variation of the relative intensities for different s values.

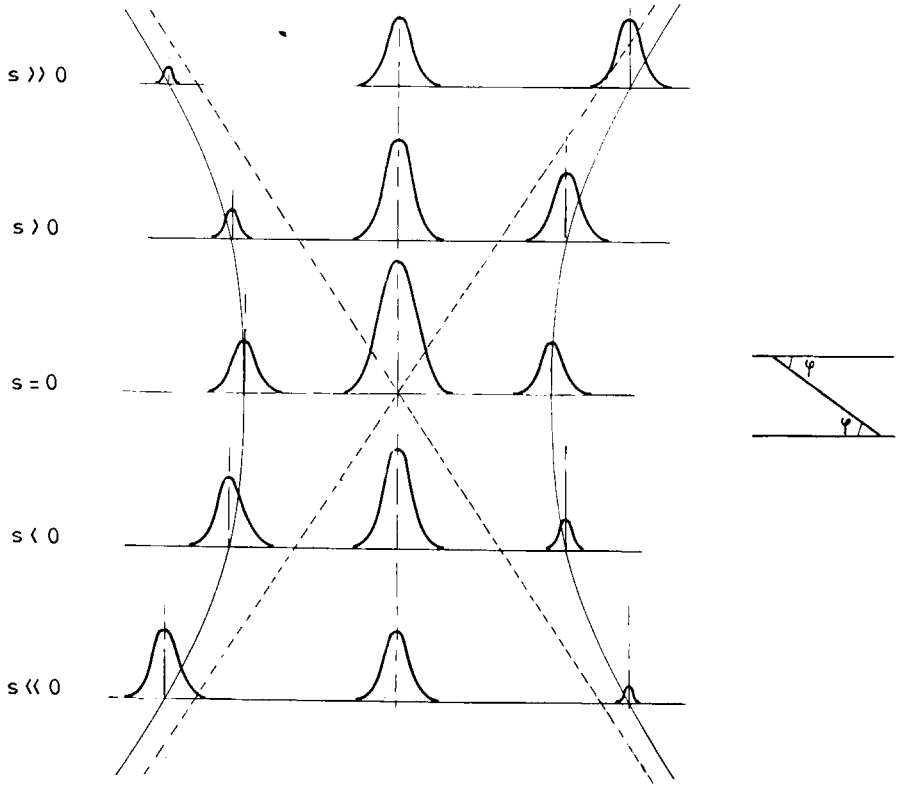


Fig. 30. – Schematic representation of the satellite position and intensity for the scattered beam. (Courtesy of *Phys. Stat. Sol.*, **18**, 343 (1966).)

2·5. One-beam kinematical approximation.

In the kinematical limit $|s|/\sigma \to 1$, it follows from (69) that one of the two components becomes too weak to be observed.

In a kinematical spot, *i.e.* the corresponding reciprocal lattice point is very far from the reflection sphere, one expects only one weak satellite.

The position of this spot will be determined by the sign of s and the slope orientation of the fault plane.

If $s > 0$, the satellite will be $u = -s \operatorname{tg} \psi$, since now $1 - s/\sigma \simeq 0$, $\sigma \simeq s$. For a weak spot it is easy to determine the sign of s from the diffraction pattern. The condition $s > 0$ means in fact that the reciprocal lattice point is inside the reflection sphere.

As can be seen from Fig. 29 the component $u = -s \operatorname{tg} \psi$ is given by the intersection of the reflection sphere and the line perpendicular to the fault plane. From the knowledge of the sign of s, and the position of the satellite with respect to the main spot, the orientation of the fault plane can thus be deduced.

For $s > 0$, $\sigma = |s|$, one observes $u = |s| \operatorname{tg} \psi > 0$, again situated at intersection of reflection sphere and rod normal to the fault plane. The conclusion is the same as in the former case.

2·6. Influence of anomalous absorption.

The anomalous absorption is accounted for if one considers σ as a complex quantity:

$$\sigma = \sigma_r + i\sigma_i \,. \tag{70}$$

The u values become then also complex, *i.e.*

$$u = u_r + iu_i \,, \quad u_r = \pm \sigma_r \operatorname{tg} \psi \,, \quad u_i = \pm \sigma_i \operatorname{tg} \psi \,. \tag{71}$$

The effect is that the peak function $P(u)$, given by (65*c*) has to be replaced by the Lorentzian curve:

$$P(u - u_n) = \frac{1}{\pi^2 [(u - u_r)^2 + u_i^2]} [\sin^2 \pi (u - u_r) a + \sinh^2 \pi u_i a] \,. \tag{72}$$

The peak height and shape are influenced by absorption.

The electrons in a satellite beam are those which changed from wave field at the fault plane. In a « two-beam » situation there are two wave fields I and II. One satellite contains electrons which have made the change I $\to$ II, the other those which made the change II $\to$ I. The electrons which cross the fault plane remaining in their wave field are found back in the main spot.

The absorption coefficients for the two wave fields are very different. However the ratio of the intensities of the two satellites is not influenced

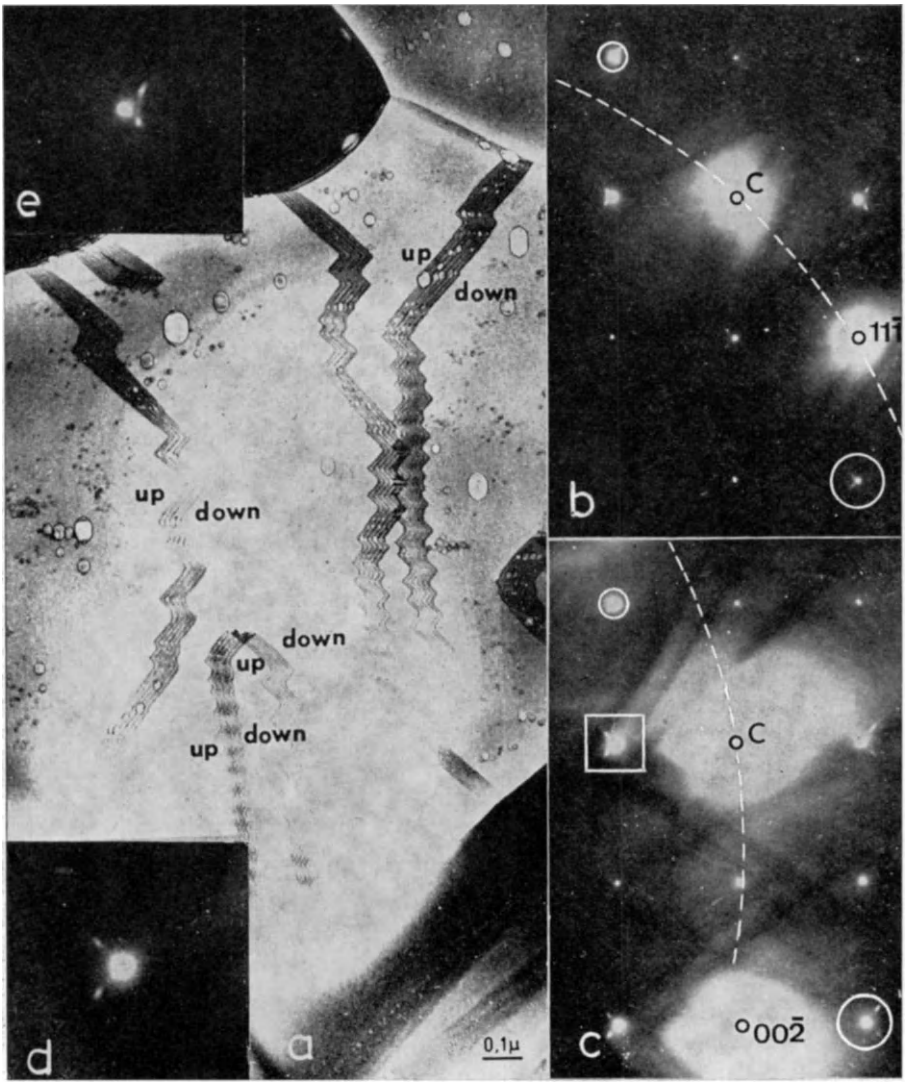


Fig. 31. – Bright-field image of anti-phase boundaries in rutile. *b*) and *c*) are diffraction patterns of this grain in different orientations. The patterns are in correct orientation with respect to the image. Attention is drawn on the position of the satellites at the spots marked by a circle. The spot marked by a square is shown enlarged in *d*). *e*) is an enlargement of the upper spot in pattern *b*) (weak kinematical spots). (Courtesy of *Phys. Stat. Sol.*, **18**, 363 (1966).)

by absorption. This could be expected since both satellites contain electrons which moved in one wave field before the fault and in the other one after the fault has been crossed.

In a multiple beam situation complications arise, and anomalous absorption can influence the relative intensities of the different satellites.

27. Observations.

Observations of anti-phase boundaries in TiO_2 and stacking faults in stainless steel are shown in Figs 31, 32, 33, 34, 35, 36. They confirm plainly the theoretical predictions.

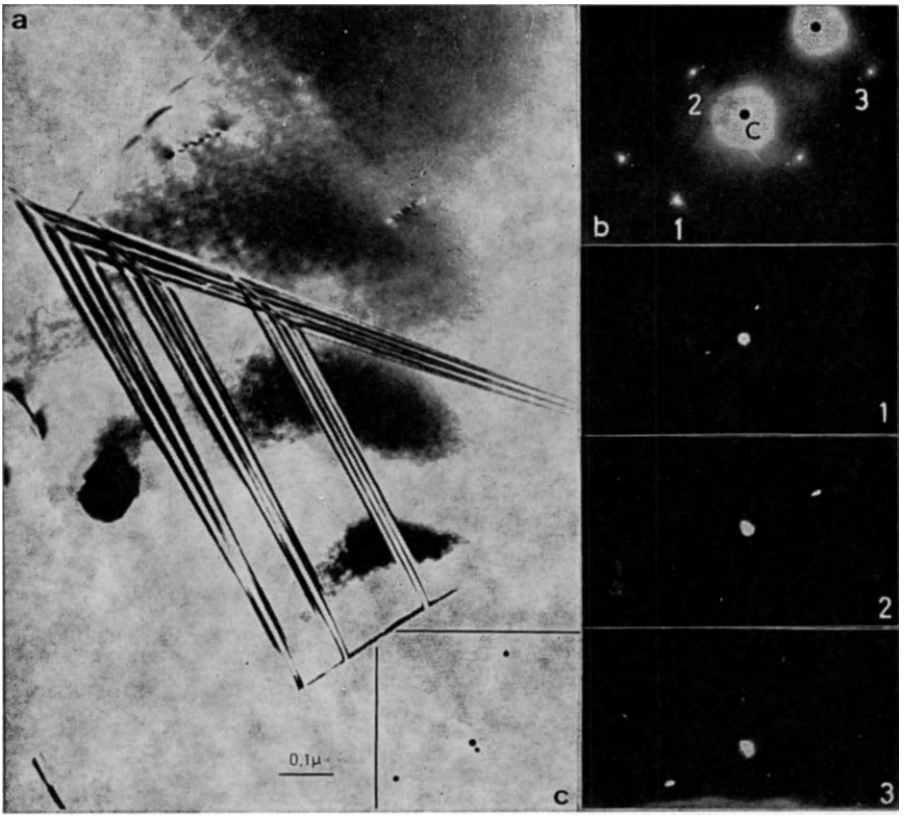


Fig. 32. – Bright field image of stacking faults in stainless steel. The faults, on two sets of planes, cause satellites in diffraction pattern (*b*). Enlargements of spots marked 1, 2 and 3 are shown as insets. Configuration 3 is also represented schematically in *c*). Notice also the splitting of the main spot due to the wedge shape of this crystal (weak kinematical spots). (Courtesy of *Phys. Stat. Sol.*, **18**, 363 (1966).)

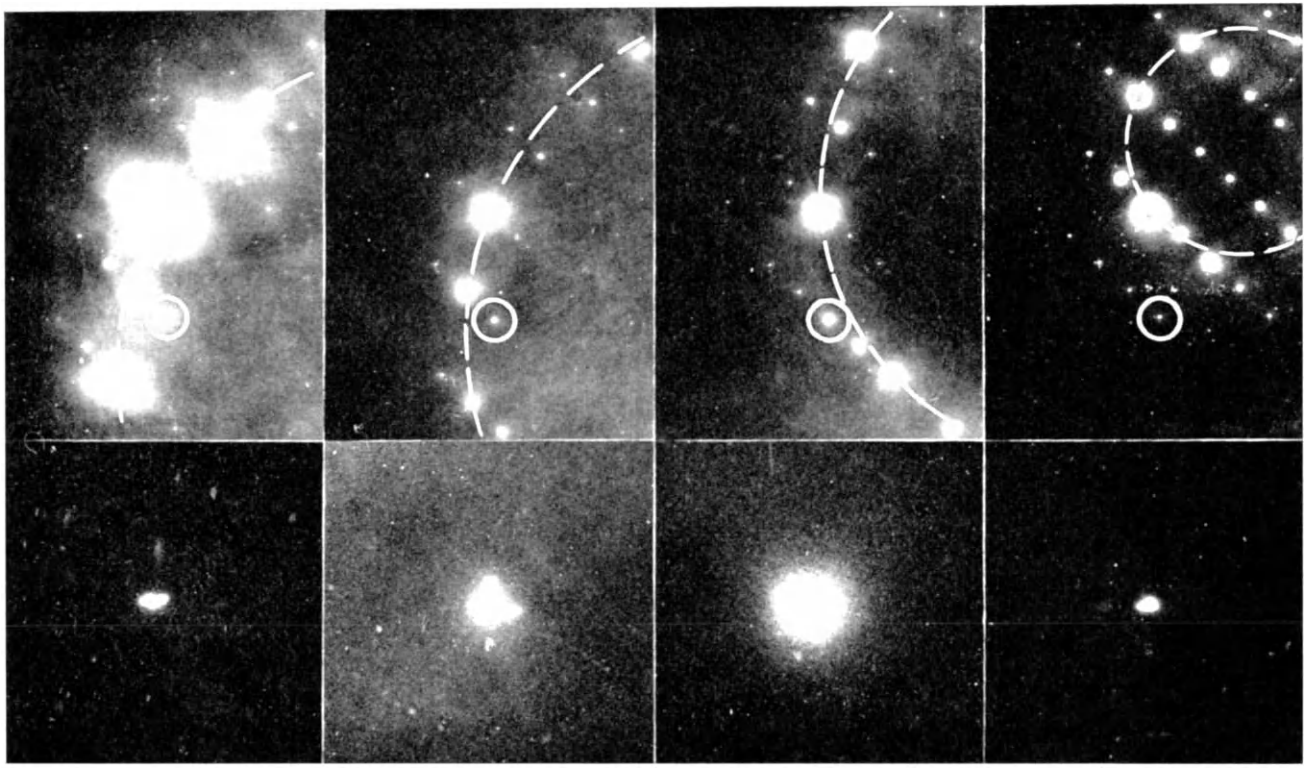


Fig. 33. – Effect of tilting through the Ewald sphere. The behaviour of the satellites at the spot marked by a circle can be followed on the enlargements shown as insets under the respective diffraction patterns. The dashed lines represent the intersection circles with the Ewald sphere. (Courtesy of *Phys. Stat. Sol.*, **18**, 363 (1966).)

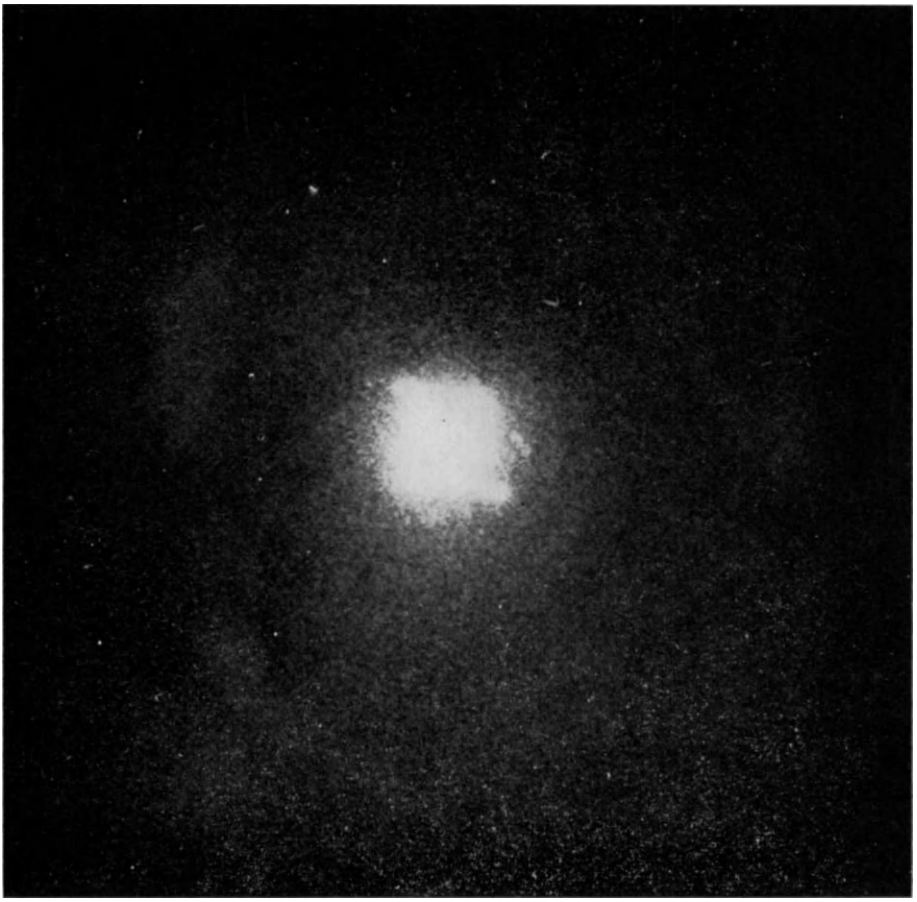


Fig. 34. – « Dynamical » spots of a diffraction pattern from the same region as Fig. 31. Notice the two pairs of satellite spots. (Courtesy of *Phys. Stat. Sol.*, **18**, 363 (1966).)

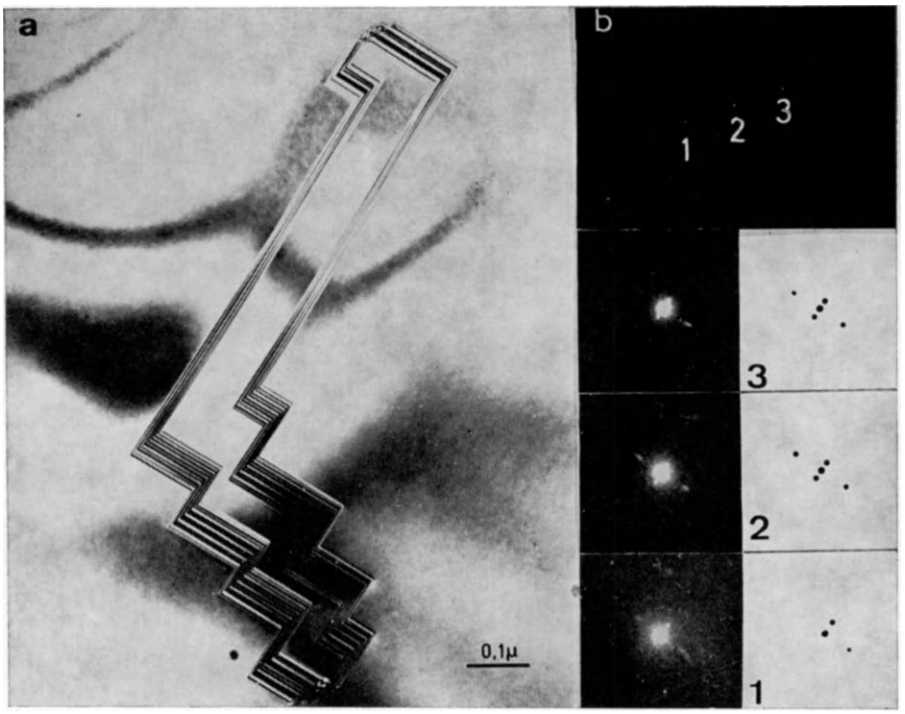


Fig. 35. – Bright field image of anti-phase boundary in rutile. The transmitted beam, as shown highly enlarged in inset 2, shows four satellite beams. Insets 1 and 3 show enlargements of the two other spots. A schematic representation is given next to the insets. The pattern is due to the two families of faults inclined at different angles to the foil plane (transmitted, dynamical and kinematical spots). (Courtesy of *Phys. Stat. Sol.*, **18**, 363 (1966).)

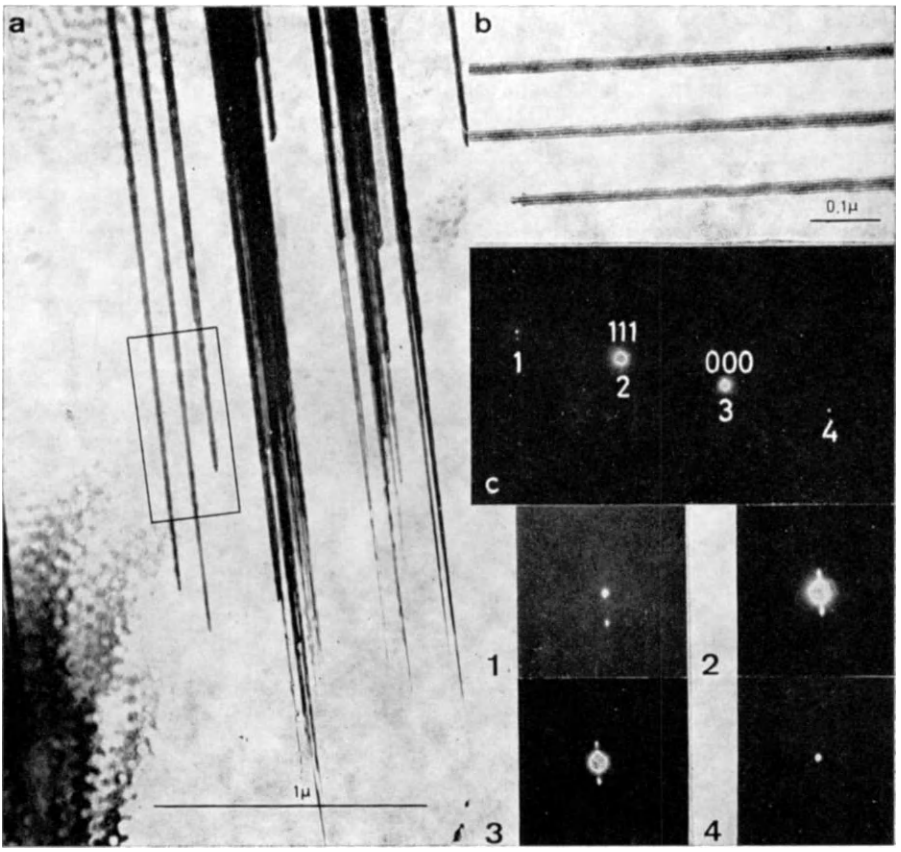


Fig. 36. – Bright field image of stacking faults in steel, *a*) and the associated diffraction pattern *c*). An enlargement of the selected area is given in *b*). The diffraction pattern *c*) is in correct orientation with respect to *b*). Insets 1, 2, 3 and 4 show enlargements of the four spots. The two satellites at the trasmitted and « dynamical » spots are clearly seen in insets 2 and 3. In spot 1 and 4 only one satellite is visible. (Courtesy of *Phys. Stat. Sol.*, **18**, 363 (1966).)

2'8. Two-beam kinematical approximation.

The « one-beam » kinematical approach is not realistic, since in a « two-beam » illumination condition there are two strong beams. The electrons in a weak beam arise then from scattering out of these beams.

If one makes the calculations for this case, one finds that the « single » satellite at $-s \operatorname{tg} \psi$ is replaced by two satellites, their mutual distances being $\sigma \operatorname{tg} \psi$, where σ refers to the dynamical beam. However, their intensity ratio is strongly affected by the anomalous absorption. This becomes evident if one takes into account that each satellite contains electrons scattered away into the weak beam direction from a different wave field. The satellite corresponding to the strongly absorbed wave field will be very weak. Mostly

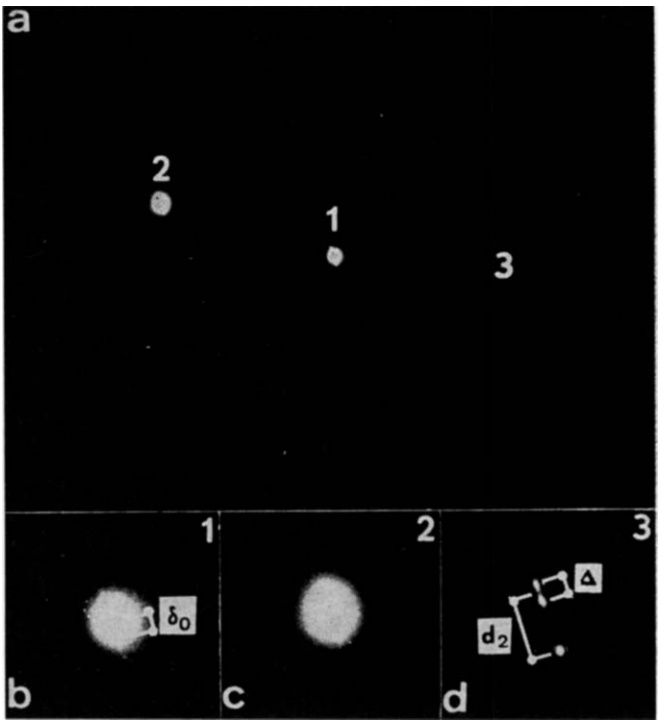


Fig. 37. – The diffraction pattern of a stainless steel foil containing stacking faults. The fine structures of the transmitted and the different scattered beams are shown in the enlargements *b*), *c*) and *d*). Notice that $\Delta = \delta_0$. (Courtesy of *Phys. Stat. Sol.*, **23**, 549 (1967).)

it will be too weak to be observed. One has then a single satellite, in accord with the « one-beam » kinematical theory.

By careful tilting the crystal into a favourable orientation, the second, more weak, satellite can also be revealed. An example is shown in Fig. 37.

REFERENCES (Section 2)

R. De Ridder, J. Van Landuyt, R. Gevers and S. Amelinckx: *Phys. Stat. Sol.*, **30** 797 (1968).

A. G. Fitzgerald and M. Mannami: *Proc. Roy. Soc.*, A **293**, 469 (1966).

R. Gevers, R. Serneels, J. Van Landuyt and S. Amelinckx: *Phys. Stat. Sol.*, **31**, 681 (1969).

R. Gevers, J. Van Landuyt and S. Amelinckx: *Phys. Stat. Sol.*, **18**, 325 (1966).

R. Gevers, J. Van Landuyt and S. Amelinckx: *Phys. Stat. Sol.*, **18**, 343 (1966).

R. Gevers, J. Van Landuyt and S. Amelinckx: *Phys. Stat. Sol.*, **23**, 549 (1967).

R. Gevers, J. Van Landuyt and S. Amelinckx: *Phys. Stat. Sol.*, **26**, 577 (1968).

H. Hashimoto, A. Howie and M. J. Whelan: *Phil. Mag.*, **5**, 967 (1960).

J. Van Landuyt, R. Gevers and S. Amelinckx: *Phys. Stat. Sol.*, **18**, 363 (1966).

M. J. Whelan and P. B. Hirsch: *Phil. Mag.*, **2**, 1303 (1957); **2**, 1121 (1957).

C. Willaime, P. Delavignette, R. Gevers and S. Amelinckx: *Phys. Stat. Sol.*, **17**, K 173 (1966).

Metallurgical Information from Electron Micrographs

L. M. BROWN

Cavendish Laboratory, University of Cambridge - Cambridge, England

1. Introduction.

The title of this set of lectures is perhaps a misnomer, because they are concerned with only a limited type of metallurgical information, namely that concerning the configuration of planar, linear, and point defects. Such problems as the identification of precipitates by electron diffraction or the calculation of dislocation density are ignored altogether. We are concerned with information which can be obtained by observation of the relatively fine structure of electron images, and which requires analysis on the basis of the dynamical theory of electron diffraction and image formation (see Howie, this volume).

The material in the lectures follows a set pattern. First, a simplified treatment of the contrast to be expected from planar, linear and point defects is given. That part of the subject which is discussed in the book « Electron microscopy of thin crystals » by P. B. Hirsch, A. Howie, R. B. Nicholson, D. W. Pashley and M. J. Whelan (1) and hereafter referred to as HHNPW is mentioned very briefly; whereas developments that have occurred since the book was written are discussed more fully, and references are given to the later work. Thus it is hoped that the lectures will be both self-explanatory and up-to-date. However it has been impossible to discuss the ramifications of the application of these developments to particular problems in radiation damage, work hardening, the study of the theoretical strength, the study of stacking-fault energy, and so on. The reader will find only allusions to these subjects, some of which will be dealt with in other lectures.

2. Contrast from planar defects.

First we consider contrast from stacking faults, giving a simplified treatment of the features to be expected from a stacking fault in a thick, absorbing crystal. We then try to give a survey of the types of fringes to be expected at interfaces generally.

2·1. Stacking faults.

The stacking fault is the simplest kind of planar defect, and the first whose appearance in the electron microscope was understood (Whelan and Hirsch ([2a,b])).

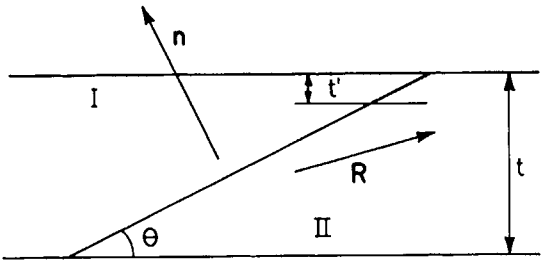


Fig. 1. – A planar fault, on a plane of normal $\boldsymbol{n}$, defined so that crystal II is displaced by a vector $\boldsymbol{R}$ with respect to crystal I.

A stacking fault arises when the crystal on one side of a plane is displaced by a vector $\boldsymbol{R}$ with respect to the crystal on the other side (Fig. 1). The vector $\boldsymbol{R}$ need not necessarily be parallel to the plane, although it often is; if $\boldsymbol{R}$ has a component perpendicular to the plane it is assumed that material has been added to or taken away from the cut, so that holes do not open up nor is there interpenetration of crystal. On either side of the fault, the crystal is perfect and the solutions of the dynamical theory for perfect crystals discussed in Howie's lectures may be used. For simplicity, let us consider the

case in which $s = 0$; in that case, Howie's eqs (20) become

$$\left.\begin{aligned} \frac{d\phi_g}{dz} &= \frac{i\pi}{\xi_g}\phi_0 \exp\left[-2\pi i \boldsymbol{g}\cdot\boldsymbol{R}\right], \\ \frac{d\phi_0}{dz} &= \frac{i\pi}{\xi_g}\phi_g \exp\left[2\pi i \boldsymbol{g}\cdot\boldsymbol{R}\right]. \end{aligned}\right\} \tag{1}$$

In a crystal for which $\boldsymbol{R} = \text{constant}$ (*i.e.* a perfect crystal) the general solution of eqs (1) is

$$\phi_0 = a\cos\frac{\pi z}{\xi_g} + b\sin\frac{\pi z}{\xi_g}, \qquad \phi_g = i\left(a\sin\frac{\pi z}{\xi_g} - b\cos\frac{\pi z}{\xi_g}\right), \tag{2}$$

where the unknown constants are determined from the boundary conditions at the entrance surface of the crystal. Now in the undisplaced crystal, crystal I, at whose top ($z = 0$) surface $\phi_0 = 1$ and $\phi_g = 0$, the solutions are

$$\phi_0^{\mathrm{I}} = \cos\frac{\pi z}{\xi_g}, \qquad \phi_g^{\mathrm{I}} = i\sin\frac{\pi z}{\xi_g}. \tag{3}$$

These solutions describe the transmitted and diffracted amplitudes in a column down to $z = t'$, where the displacement $\boldsymbol{R}$ occurs. If we call $\alpha = 2\pi \boldsymbol{g}\cdot\boldsymbol{R}$, eqs (1) tell us that there must be a discontinuity in the z-derivatives of ϕ_0 and ϕ_g so that

$$\left.\begin{aligned} \left[\frac{d\phi_0^{\mathrm{I}}}{dz} - \frac{d\phi_0^{\mathrm{II}}}{dz}\right]_{z=t'} &= \frac{i\pi}{\xi_g}\left[\phi_g^{\mathrm{I}} - \phi_g^{\mathrm{II}}\exp\left[i\alpha\right]\right]_{t'}, \\ \left[\frac{d\phi_g^{\mathrm{I}}}{dz} - \frac{d\phi_g^{\mathrm{II}}}{dz}\right]_{z=t'} &= \frac{i\pi}{\xi_g}\left[\phi_0^{\mathrm{I}} - \phi_0^{\mathrm{II}}\exp\left[-i\alpha\right]\right]_{t'}. \end{aligned}\right\} \tag{4}$$

If there is a step discontinuity in the derivatives, the amplitudes themselves must be continuous. We can now do a wave-matching calculation for the unknown constants of the general solution in the displaced crystal. Since the amplitudes themselves must be continuous, we write

$$\phi_0^{\mathrm{II}} = \cos\frac{\pi z}{\xi_g} + a\sin\frac{\pi(z-t')}{\xi_g},$$

$$\phi_g^{\mathrm{II}} = i\sin\frac{\pi z}{\xi_g} + b\sin\frac{\pi(z-t')}{\xi_g},$$

and use eqs (4) to determine the constants a and b. We easily find for the transmitted and diffracted amplitudes at depth z

$$\left.\begin{aligned}\phi_0^{\mathrm{II}} &= \cos\frac{\pi z}{\xi_g} - 2i\exp[i\alpha/2]\sin\alpha/2\sin\frac{\pi t'}{\xi_g}\sin\frac{\pi(z-t)}{\xi_g},\\ \phi_g^{\mathrm{II}} &= i\sin\frac{\pi z}{\xi_g} + 2\exp[-i\alpha/2]\sin\alpha/2\cos\frac{\pi t'}{\xi_g}\sin\frac{\pi(z-t')}{\xi_g}.\end{aligned}\right\} \tag{5}$$

Equations (5) can be used to discuss the contrast at stacking faults in the absence of absorption; we refer the reader to the discussion in HHNPW p. 229. The fringes in this case are rather complicated, consisting of alternating strong and weak fringes. However, a much simpler situation, and one more commonly met with, is the case of a thick absorbing crystal. To deal with this case we multiply the amplitudes of eqs (5) by $\exp[-\pi z/\xi_0']$ to take account of the « background » or « mean » absorption, and we replace $1/\xi_g$ by $1/\xi_g + i/\xi_g'$ wherever this occurs (see Howie, this volume). A thick absorbing crystal will be taken to be one in which the amplitudes have been reduced by e on traversing the crystal: for typical cases in which the absorption lengths are ten times the extinction distances, this means a crystal about $3\xi_g$ thick or thicker. Under these circumstances it becomes a good approximation to write

$$\left.\begin{aligned}\sin\left(\frac{\pi t}{\xi_g}+\frac{i\pi t}{\xi_g'}\right) &= -\frac{i}{2}\exp[\pi t/\xi_g']\exp[-i\pi t/\xi_g],\\ \cos\left(\frac{\pi t}{\xi_g}+\frac{i\pi t}{\xi_g'}\right) &= \frac{1}{2}\exp[\pi t/\xi_g']\exp[-i\pi t/\xi_g].\end{aligned}\right\} \tag{6}$$

We now look at eqs (5) under two circumstances: the first for a column near the top of the fault ($t' \ll t$, see Fig. 1) and the second for a column near the bottom ($t' \simeq t$). It is a simple matter to use eqs (6) in eqs (5) to find for the transmitted and diffracted intensities the following:

$$\left.\begin{aligned}&\text{fault near top of foil, } t' \ll t\\ &|\phi_0^{\mathrm{II}}|^2 = |\phi_g^{\mathrm{II}}|^2 = \frac{1}{4}\exp\left[-2\pi t\left[\frac{1}{\xi_0'}-\frac{1}{\xi_g'}\right]\right]\left\{1+\sin\alpha\sin\frac{2\pi t'}{\xi_g}\right\};\\ &\text{fault near bottom of foil, } t' \simeq t\\ &|\phi_0^{\mathrm{II}}|^2 = \frac{1}{4}\exp\left[-2\pi t\left[\frac{1}{\xi_0'}-\frac{1}{\xi_g'}\right]\right]\left\{1+\sin\alpha\sin\frac{2\pi(t-t')}{\xi_g}\right\},\\ &|\phi_g^{\mathrm{II}}|^2 = \frac{1}{4}\exp\left[-2\pi t\left[\frac{1}{\xi_0'}-\frac{1}{\xi_g'}\right]\right]\left\{1-\sin\alpha\sin\frac{2\pi(t-t')}{\xi_g}\right\}.\end{aligned}\right\} \tag{7}$$

When the fault is in the centre of the foil, $t' \simeq t/2$, no very simple approximation can be made, but one can see that the oscillating terms in eqs (5) will contribute only weakly to the total intensity.

Equations (7) demonstrate a number of important features of stacking fault contrast.

Firstly, when the fault is at the top of the foil, the transmitted and diffracted intensities are *similar*; in the approximations made here, they are identical. Thus bright-field and dark-field micrographs will show similar images of faults near the top (electron entrance) surface of the foil. But when the fault is at the bottom of the foil, the transmitted and diffracted intensities are *complementary*. These are very general characteristics of diffraction contrast. We may understand how they arise by noting that the effect of anomalous absorption is always to make the diffracted amplitude approximately equal to the transmitted amplitude after the wave has traversed about $3\xi_g$. This is true regardless of the boundary conditions at the top of the crystal. If the total incident intensity is unity, then the trasmitted and diffracted intensities both approach $\frac{1}{4}\exp[-2\pi t(1/\xi_0' - 1/\xi_g')]$ (see eqs (7) with $\alpha = 0$) and the phases of both also become equal. Thus whatever the amplitudes are when the wave enters crystal II near the top of the foil, the amplitudes will be equal near the exist surface and the bright- and dark-field images will be similar. Also, since the amplitudes incident upon crystal II near the bottom of the foil are equal, and since in the absence of absorption intensity is conserved, the two intensities will be complementary when the fault is near the bottom of the crystal.

Secondly, the bright-field image is *symmetrical* and the dark-field image is not; the dark-field image is approximately antisymmetrical. This situation is peculiar to the type of displacement generated by the fault; it is not a general feature of images. In general, if the displacement function $\boldsymbol{R}(z)$ by a suitable choice of origin can be made an *odd* function of z, the bright-field image will be symmetrical; if the displacement function can be made an *even* function of z the dark-field image will be symmetrical (see the symmetry rules proved by Howie, his eqs (46) and (47)).

Thirdly, the image is associated with fringes whose period in t' is one extinction distance. This is a feature of images which is generally true in thick absorbing crystals, but not in very thin ones. It is also not true if the crystal is deviated from the Bragg position: the depth periodicity of the fringes is then $\xi_g/\sqrt{1+w^2}$ ($w = s_g\xi_g$ and is proportional to the angular deviation from the Bragg angle).

Fourthly, the fringes are such that whether they are black or white depends

upon the sign of α. From eqs (7) one can formulate a rule: for bright-field images, the top and bottom fringes (the outermost fringes on a micrograph) are brighter than background if α is positive. This feature of the image can be used to get valuable information about the nature of the fault. The reader should refer to Gevers' lectures for details; the point is that if the fault can be regarded as made by the removal of a plane of atoms (an intrinsic fault) then $\boldsymbol{R}$ in Fig. 1 is pointing upwards, whereas if the fault is to be regarded as made by the insertion of a plane of atoms (an extrinsic fault) $\boldsymbol{R}$ will have opposite sign. One can thus distinguish faults which have formed by the condensation of vacancies from these which have formed by the condensation of interstitials; recently use has been made of this to verify that the loops observed as a result of electron damage in the high voltage microscope are composed of interstitial atoms (Ipohorski and Spring [3]).

Fifthly, it is clear from eqs (5) that when $\alpha = 0, 2\pi, \ldots 2n\pi$, the contrast from the fault vanishes completely. This condition for invisibility enables one to determine experimentally the vector $\boldsymbol{R}$ for a given fault. The physical significance of this vanishing criterion is purely geometrical; it does not depend upon any detailed treatment of the dynamical theory. For $\boldsymbol{g}\cdot\boldsymbol{R} = 0$ implies that $\boldsymbol{R}$ is parallel to the Bragg planes, and if only two beams are excited, displacement of a Bragg plane parallel to itself does not affect the diffraction. When $\alpha = \pi, 3\pi, \ldots (2n+1)\pi$, eqs (7) show that there will also be no fringes; however eqs (5) show that the contrast does not vanish, the image is just a dark band. Contrast from these so-called π-faults is discussed in detail in HHNPW p. 241.

Sixthly, and finally, it is of interest to ask how small a value of α can be detected. If one can just detect a small fractional deviation from background of f, then eqs (7) show that the minimum observable value of α is also f. Howie and Jouffrey [4] estimate that $f \simeq 0.02$, and hence they set limits on the displacement of the atomic planes perpendicular to a shear fault in cadmium. However, this visibility limit corresponds to « normal viewing » conditions, that is observation under dynamical conditions in a thick absorbing crystal. Undoubtedly application of dark-field techniques will improve the visibility of stacking-faults.

Before we leave this Section, we should point out that although the calculation on which the discussion is based is rather simple and instructive, when a number of fault planes overlap it is useful to have a routine method to carry through the wave-matching. The reader will appreciate that the quickest (and most error-free) method of plotting intensities from the amplitudes of eqs (5) is to use a computer to evaluate the formulae. Thus there

is no point in avoiding the use of a computer early on in the calculation; the most efficient way of doing the calculation is to set up matrices which relate the wave amplitudes on the exit surface of a slab to those on the entrance surface, and to use the standard routines for matrix multiplication to calculate the images. The interested reader should consult HHNPW, Chapter 10 and Goringe, this volume.

2·2. **More general discussion of contrast from planar defects.**

In electron microscopy, fringes arise whenever two (or more) plane waves which are coherent but are travelling in slightly different directions are made to interfere. If we assume that the z-component of the wave-vectors is the same, but the wave vectors differ by a small amount in their x-components, so that

$$\boldsymbol{k}_1 = (-\delta k, 0, k), \quad \boldsymbol{k}_2 = (\delta k, 0, k),$$

then

$$A = \exp[2\pi i \boldsymbol{k}_1 \cdot \boldsymbol{r}] + \exp[2\pi i \boldsymbol{k}_2 \cdot \boldsymbol{r}] = (\text{a phase factor})(\cos 2\pi \delta k x).$$

Thus the intensity ($= |A|^2$) has a periodicity of $(2\delta k)^{-1}$ in the x-direction. The fringes appear as stripes perpendicular to the plane containing the two wave vectors, and the spacing of the fringes is the wavelength divided by the angle between the wave vectors.

It is extremely instructive to see how stacking fault fringes arise in the dispersion surface construction. We remember that in this construction $\boldsymbol{k}$ vectors lie only on the surface shown in Fig. 2; this surface (*) arises from the solution of the Schrödinger equation (see Howie, this volume). When waves pass from one crystal into another, new $\boldsymbol{k}$ vectors are required to carry out the wave-matching; since the tangential components of the $\boldsymbol{k}$ vectors must be equal, the new vectors are found by drawing a normal to the surface of the crystal and finding its intersection with the other branches of the dispersion surface. Let us suppose, in Fig. 2, that the incident wave gives rise to the wave points P_1 and P_2, so that waves with these wave vectors propagate in crystal I. When they reach crystal II, the wave-matching will

(*) Strictly speaking, Fig. 2 shows a section of the dispersion surface defined by the plane containing $\boldsymbol{g}$ and a vector normal to the surface of the specimen.

give rise to two further wave points, P_3 and P_4. Now in bright-field viewing conditions, four plane waves with wave vectors joining the origin of reciprocal space to the four wave points will pass through the objective aperture

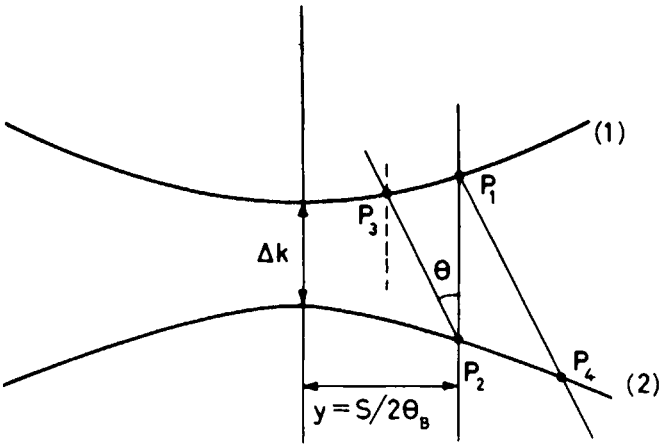


Fig. 2. – Wave-matching in the dispersion surface. The lines P_1P_4 and P_2P_3 are both parallel to the normal $\boldsymbol{n}$ of Fig. 1. In the upper crystal, wave points P_1 and P_2 are excited, corresponding to a deviation parameter s. The separation of the branches of the dispersion surface is $\Delta k = (\xi_g/\sqrt{1+w^2})^{-1}$, equal to the inverse of extinction distance for $w=0$.

and give rise to interference fringes. If we consider a column which intersects the bottom of the fault, the waves associated with the wave-point P_1 will be very weak, having been attenuated by anomalous absorption. Thus, at the exit surface of the crystal, two-beam or cosine fringes will be observed due to interference between waves associated with P_2 and P_3. The angle between the wave vectors is $(P_3P_2)\sin\theta/|\boldsymbol{k}|$ so that the spacing of the fringes in real space is $(P_3P_2)^{-1}\operatorname{cosec}\theta$. But $(P_3P_2)^{-1}$ is to a good approximation $(P_1P_2)^{-1}\cos\theta$ so that the depth periodicity of the fringes (the depth periodicity is, by definition, the periodicity in the x-direction times $\operatorname{tg}\theta$) is $(P_1P_2)^{-1} = \xi_g/\sqrt{1+w^2}$ as given earlier. Similarly, for a column which intersects the top of the fault, cosine fringes due to interference between waves associated with wave points P_2 and P_4 are observed. For columns intersecting the middle of the fault, weak four-beam fringes are observed.

Further analysis is required to find the symmetry of the images and the rules governing the appearance of the bright-field and dark-field fringes. But this construction shows that every planar discontinuity in the foil will give rise to fringes; and the fringes will have a spacing tied to the extinction distance. Fringes of this type are: 1) thickness fringes (see Howie, this volume); 2) structure-factor fringes, which arise when the foil contains a thin slab of material of different scattering power from the matrix; 3) δ-fringes (see Gevers, this volume). These fringes arise when the Bragg planes of crystal II are rotated slightly from the Bragg planes of crystal I. The wave-matching is now from one dispersion surface to another in a slightly different position, but the fringe spacing is still given approximately by the extinction distance; however, the fringes are somewhat irregularly spaced. By an appropriate choice of origin, $\boldsymbol{R}$ can be made an *even* function of z, so that the dark-field images of these fringes are symmetrical. Observations of δ-fringes have been made in barium titanate, where the ferro-electric anti-phase boundaries are seen in this way (Gevers *et al.* [5]); also in V_3Si which undergoes a martensitic transformation at low temperatures to a tetragonal form; the structure is twinned and the twin boundaries display δ-fringes (Goringe and Valdrè [6]); another interesting example is a planar coherent interface between a precipitate and matrix (Ardell [7]).

In addition to these types of fringes, one can have fringes formed from two beams which do not arise because of the dispersion surface. If two overlapping crystals give rise to different Bragg reflections $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, and both of these are allowed through the objective aperture to form an image, cosine fringes will be observed perpendicular to $\boldsymbol{g}_2-\boldsymbol{g}_1$ and with spacing $|\boldsymbol{g}_2-\boldsymbol{g}_1|^{-1}$. These are so-called moiré fringes, or if $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ arise from the same crystal, they are lattice fringes; evidently their spacing is entirely geometrically determined and has nothing to do with the extinction distances in the crystals. The interested reader should consult ref. [1] Chapter 15; also try problem 14, this volume.

The observation of moiré fringes permits an accurate comparison of the spacing of atomic planes in two lattices. An interesting recent example is provided by the work of Vincent [8] who observed the variation of strain in islands of tin deposited on tin telluride. As the islands grow, they become progressively less strained by the introduction of misfit dislocations in the interface, an effect first predicted by Franck and van der Merwe [9]. Figure 3 shows some of these moiré finges. The fringe spacing is inversely proportional to the mismatch between the Sn and the SnTe, and it can be seen that smaller islands have fringes of larger spacing than large islands. Thus the

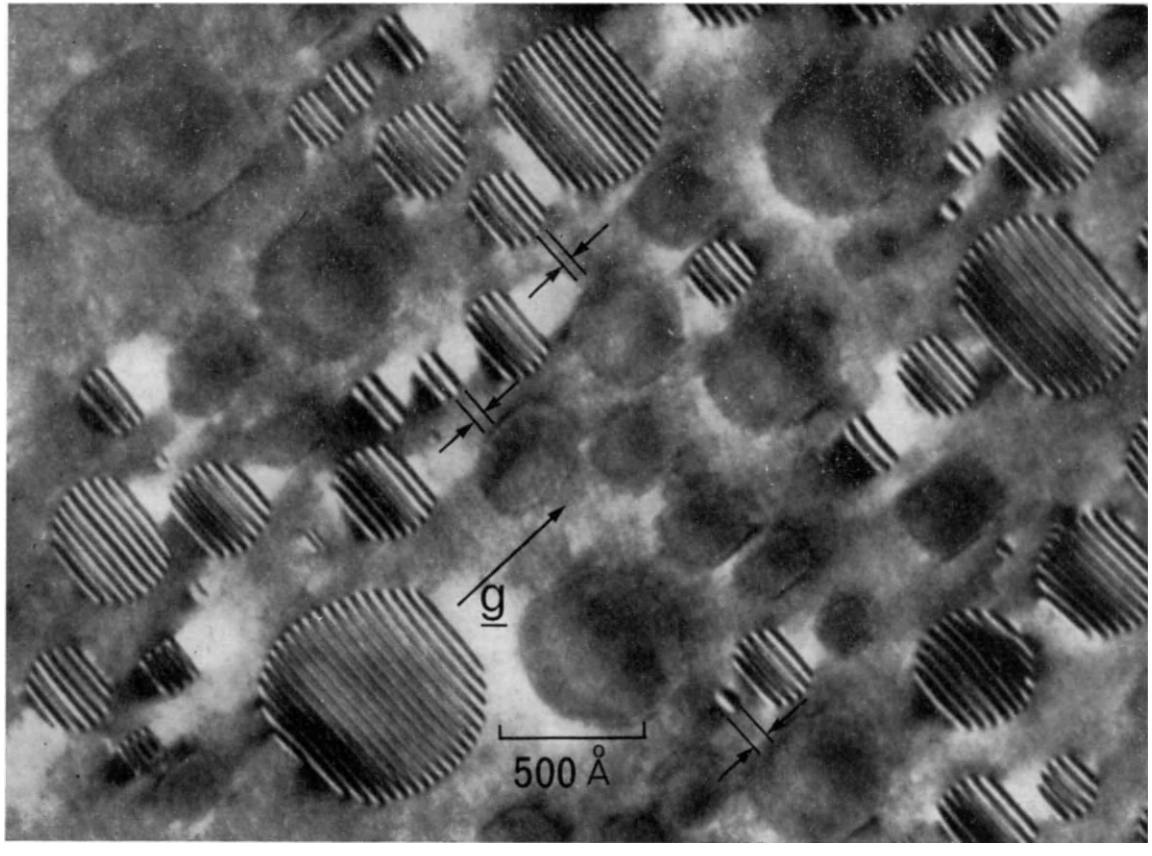


Fig. 3. – Dark-field micrograph of epitaxial tin islands grown by evaporation onto SnTe; the tensile strain within the smaller islands causes an increase in the moiré fringe spacing and also black-white strain contrast in the adjacent areas of the substrate. (Courtesy of R. Vincent.)

mismatch between the two lattices increases as the islands grow, and less elastic strain is necessary to accommodate the mismatch.

All these different types of fringes are summarised in the following Table.

Type of fringe	Spacing	Visibility	Sign rules	References
Stacking-fault. Displacement $\boldsymbol{R}$ of II with respect to I	Depth periodicity $\xi_g/\sqrt{1+w^2}$	Best at $s=0$	If $\boldsymbol{g}\cdot\boldsymbol{R}>0$, outermost fringes of *bright* field are white; bright-field symm.	(1), Ch. 10
δ-fringe. Crystal II has s_2, crystal I has s_1	Depth periodicity $\sim\xi_g$ but irregular, depending on s	Best for $s_1=-s_2$	If $s_1>s_2$, outermost fringes of *dark* field are white; dark-field symm.	(5), (10)
Thickness fringes	Depth periodicity $\xi_g/\sqrt{1+w^2}$	Best at $s=0$	—	(1), Ch. 8
Structure-factor fringes	Depth periodicity $\xi_g/\sqrt{1+w^2}$	*Invisible* for $s=0$	—	(1), Ch. 10.6 (cavity)
Moiré fringes	$\lvert\boldsymbol{g}_1-\boldsymbol{g}_2\rvert^{-1}$	Whenever there is intensity in both beams	—	(1), Ch. 15

3. Contrast from dislocations.

Most treatments of contrast from dislocations, and indeed from defects in which $\boldsymbol{R}$ is a continuous function of z, use a digital computer to solve Howie's eqs (20). Each case really has to be calculated afresh, although certain general principles can be used to see what is happening, and we will try to emphasise these. For a discussion of the computational techniques see Goringe, this volume.

The problem of contrast from dislocations was first studied by Howie and Whelan ([11]). As most later workers have also done, they used the elastic continuum theory of dislocations in an isotropic medium to derive the dis-

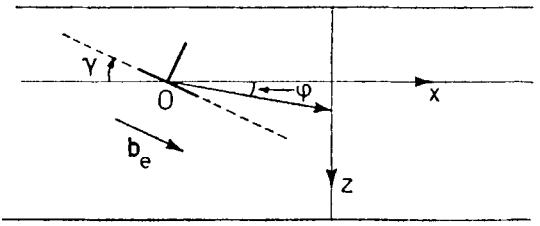


Fig. 4. – The co-ordinate system used to describe dislocation contrast.

placement $\boldsymbol{R}$. For instance, for a screw dislocation the displacements are everywhere parallel to the Burgers vector, and can be shown to be

$$\boldsymbol{R} = \frac{\boldsymbol{b}}{2\pi} \operatorname{tg}^{-1} \varphi \,. \tag{8}$$

(See Fig. 4 for the co-ordinate system.) The contrast depends upon the quantity $\boldsymbol{g} \cdot \boldsymbol{R}$, so that if this is zero, no contrast will be observed. It follows that for a screw dislocation in an isotropic medium that if $\boldsymbol{g} \cdot \boldsymbol{b} = 0$ there will be no contrast. This criterion for the vanishing of a dislocation depends entirely upon geometry, and does not depend upon the detailed application of the dynamical theory: it follows because Bragg planes parallel to the dislocation are undistorted, so any image formed using a $\boldsymbol{g}$-vector perpendicular to the dislocation will be unaffected by the presence of the dislocation.

For a more general dislocation,

$$\boldsymbol{R} = \frac{1}{2\pi} \left\{ \boldsymbol{b}\Phi + \boldsymbol{b}_e \frac{\sin 2\Phi}{4(1-\nu)} + \boldsymbol{b} \times \boldsymbol{u} \left(\frac{1-2\nu}{2(1-\nu)} \log |r| + \frac{\cos 2\Phi}{4(1-\nu)} \right) \right\} \tag{9}$$

(see Read ([12])). In this expression, ν is Poisson's ratio, $\Phi = \varphi - \gamma$ (Fig. 4), $\boldsymbol{b}$ is the total Burgers vector, $\boldsymbol{b}_e$ is the edge component, and $\boldsymbol{u}$ is a unit vector parallel to the dislocation line. It will be seen that there are components of the displacement perpendicular to the slip plane, arising because elastic compression or extension always leads to displacements perpendicular to the axis

of compression or extension. If the dislocation has pure edge character, the vanishing condition is now $\boldsymbol{g}\cdot\boldsymbol{b}=0$ and $\boldsymbol{g}\cdot(\boldsymbol{b}\times\boldsymbol{u})=0$. This means that $\boldsymbol{g}$ is parallel to the dislocation line, and once more has a simple geometrical meaning.

Howie and Whelan noted the following general features of dislocation contrast. Firstly, when a dislocation threads the foil, the bright-field image is symmetrical about the mid-point of the foil and the dark-field image is not. This is another example of the symmetry rules proved by Howie in this volume. The dark-field and bright-field images are similar for the part of the dislocation near the top of the foil, and complementary for the part of the dislocation near the bottom.

Secondly, the image of the dislocation in a thick absorbing crystal is associated with depth oscillations in appearance which can be quite complicated. The oscillations have a period of $\xi_g/\sqrt{1+w^2}$, and are most pronounced near either surface of the foil under dynamical conditions. Near the centre of the foil, a dislocation appears as a dark line.

Thirdly, the images of edge dislocations tend to be wider by about a factor of two than those for screw dislocations. Near the middle of the foil, a screw dislocation appears as a dark line of width about $\xi_g/3$.

Fourthly, in face-centered crystals, where partial dislocations can be found for which $\boldsymbol{g}\cdot\boldsymbol{b}=\pm\frac{1}{3}$, $\pm\frac{2}{3}$, etc., Howie and Whelan found by numerical calculation for reasonable visibility criteria that partial dislocations with $\boldsymbol{g}\cdot\boldsymbol{b}=\frac{1}{3}$ were invisible.

Since this work, and since HHNPW was written, a great deal of work has been done. One should mention the work of Silcock and Tunstall ([13]) on the contrast from partial dislocations. They extend Howie and Whelan's treatment, and are able to demonstrate a new form of precipitation in austenitic stainless steels.

The difficulties which may be encountered in applying the vanishing conditions to determine Burgers vectors are illustrated by the work of Dingley and Hale ([14]). These authors used dark-field techniques to investigate the types of dislocation occuring in iron. They found that in addition to $\frac{1}{2}a\langle 111\rangle$, Burgers vectors of the $a\langle 100\rangle$ and $a\langle 110\rangle$ type were present. However, calculations by France and Loretto ([15]) and by Dingley ([16]) show that great care must be taken in this kind of experiment, and that almost certainly no dislocations of the unusual $a\langle 100\rangle$ and $a\langle 110\rangle$ types are present. The reason for this difficulty is that large values of $s_g\xi_g=w$ can cause the image to vanish, and if reflections with large extinction distances are used (these are higher-order reflections in monatomic cubic lattices) the images will vanish

for rather small values of s_g: In practice, micrographs are often taken with $s_g > 0$, in order to achieve good transmission and to minimise dynamical oscillations and contrast from artefacts, so this effect will make interpretation difficult. In their most recent publication, Loretto and France ([17]) present very complete data, taking into account possible effects of elastic anisotropy.

The best way to overcome these difficulties has been developed by Head ([18]). He programmed the computer to generate simulated micrographs. The intensities which are derived from the Howie-Whelan equations are displayed as « points » of varying density. Thus a full-stop (·) may represent an intensity of x units, and a colon (:) an intensity of $2x$, and so on. Head points out that a very large number of integrations of the equations are required to generate one picture of a dislocation threading the foil; he estimates something like 65. It is necessary to use the computer as economically as possible, and Head indicates how this may be done; it turns out that images corresponding to all depths of dislocation in the foil can be generated by just two integrations of the equations (see Goringe, this volume, for details).

At the same time, the Australian workers have performed calculations which do not make the assumption of isotropic elasticity contained in eqs (8) and (9). In a series of papers (Head, Loretto and Humble ([19]); Humble ([20])) dislocations in β-brass are analysed. β-brass has a very large elastic anisotropy, and screw dislocations for which $\boldsymbol{g}\cdot\boldsymbol{b} = 0$ are visible as a pair of lines on either side of the dislocation. Of the various theoretical possibilities, $\boldsymbol{b} = a[\bar{1}11]$ corresponds most closely with the experimental data. None of the other theoretical possibilities correspond closely with observation. It is thus possible to identify Burgers vectors in a most direct and striking way. One should mention here another effect of elastic anisotropy on dislocations which can be observed in β-brass. This is that certain orientations of the dislocation line become forbidden, and the dislocations assume a characteristic bent appearance. The reader should consult the references for details.

So far, the identification of dislocations (and defects generally) has been done by guesswork. One knows what is a likely candidate for the defect, and then one attempts to demonstrate by computing that the guess is consistent with observation. A recent paper by Head ([21]) shows that the reverse process is in principle possible: one can construct the displacement field of a defect uniquely from electron micrographs. It would be a great tour-de-force if this could be done, although there are a number of numerical difficulties.

A most promising development has been announced by Cockayne, Ray and Whelan ([22]). As is known (see Howie, this volume), the width of a

dislocation image is controlled by the extinction distance ξ_g. This is because, as his eq. (52) shows, Fourier components of the strains with this wavelength contribute most to the image. It has been recognised for some time that the images could be made narrower by decreasing the extinction distance, and a convenient way to do this is to make use of kinematical conditions, in which the extinction distance becomes s_g^{-1} and can, in a two-beam model, become as short as one would like. Of course, the limitation is that in the bright-field image the contrast is lost, and the image has usually disappeared by the time $w = s_g \xi_g$ is about two or three, for typical low-order reflections. However, in dark-field, although the total intensity scattered into the image will be small, it will nonetheless be very much greater than the background intensity, so there is the possibility of achieving sharp images with good contrast by taking pictures in kinematical dark-field conditions (see [1], p. 192). Cockayne, Ray and Whelan have done this, and have been able to resolve partial dislocations only 120 Å apart (see Goringe and Hall, *Problem 16.*)

Cockayne, Ray and Whelan also give an ingenious method for calculating the position of the maximum in intensity in a kinematical image. The problem is to find the column for which the kinematical integral

$$\int_{\text{column}} \exp\,[2\pi i[sz + \boldsymbol{g}\cdot\boldsymbol{R}]]\,\mathrm{d}z$$

gives a maximum amplitude. One is accustomed to the stationary phase method of estimating the integral. If there is some value of z, z_0 for which

$$\frac{\mathrm{d}}{\mathrm{d}z}\,[sz + \boldsymbol{g}\cdot\boldsymbol{R}] = 0\,,$$

then the integral becomes

$$\int_{\text{column}} \exp\left[2\pi i\left[\frac{1}{2}z^2\left(\frac{\mathrm{d}^2(\boldsymbol{g}\cdot\boldsymbol{R})}{\mathrm{d}z^2}\right)_{z=z_0} + \frac{z^3}{6}\left(\frac{\mathrm{d}^3(\boldsymbol{g}\cdot\boldsymbol{R})}{\mathrm{d}z^3}\right)_{z=z_0} + \ldots\right]\right]\mathrm{d}z$$

and can be approximated. However, if there is a column for which

$$\left(\frac{\mathrm{d}^2(\boldsymbol{g}\cdot\boldsymbol{R})}{\mathrm{d}z^2}\right)_{z=z_0} = 0\,,$$

then the integral will evidently give a maximum amplitude—this might be called the principle of the most stationary phase! Cockayne, Ray and Whelan show that calculations based on this principle predict positions of images which agree very well with those obtained for many-beam dynamical calculations; problem 16 is intended for those who want to follow this up. The use of « high resolution dark-field » techniques could well resolve a number of outstanding problems; in particular, it should prove possible to get information on the stacking-fault energy of materials of higher stacking-fault energy than have hitherto been studied.

4. Contrast from inclusions.

Inclusions can give rise to contrast by a large number of different mechanisms, and although we shall list some of these, we cannot treat them all in very great detail.

4·1. Structure factor contrast.

For a small inclusion, the reader is recommended to try problem 19; see also ([1]) p. 336. The treatment has been recently extended by Gleiter ([23]).

The principle is very simple ([24b]). A small inclusion of size Δt changes the effective foil thickness by $\Delta t(1/\xi_g^i - 1/\xi_g)$. Thus by differentiating eqs (3) one finds for the intensity change due to the inclusion

$$\Delta I = -\pi \Delta t \left(\frac{1}{\xi_g^i} - \frac{1}{\xi_g}\right) \sin \frac{2\pi t}{\xi_g} .$$

Maximum visibility occurs where $t = (2n + 1)\pi/4$, and the contrast will be alternately bright and dark, depending on whether $\xi_g^i > \xi_g$ or $\xi_g^i < \xi_g$.

4·2. Interface contrast.

A great variety of mechanisms operate: if the inclusion is semi-coherent, so that a regular dislocation array exists in the interface, the dislocations may

be made visible. This case has not been treated theoretically, but a recent paper by Weatherly and Nicholson ([25a]) reports a number of observations. They find that the conditions for visibility of the dislocations are very stringent. Matrix and precipitate reflections must co-incide or nearly co-incide, and the misfit must not be too large. They have obtained many beautiful and striking pictures in a number of systems.

All the other types of fringes discussed earlier can provide contrast. In particular, one may mention again the occurrence of δ-fringes when a fully coherent interface extends through the foil. This case has been treated by Ardell ([7]) and by Weatherly ([25c]). Outside a planar interface separating a uniformly expanded inclusion from the matrix, the matrix is stretched into a tetragonal form. Thus any Bragg plane neither parallel nor perpendicular to the interface will undergo slight rotations across the interface, and will give rise to δ-fringes.

4·3. Strain contrast.

We now turn our attention to strain contrast, in which a number of developments have occurred since HHNPW was written. Let us consider at the outset contrast from the misfitting sphere treated dynamically using the two-beam Howie-Whelan equations. The displacement vector $\boldsymbol{R}$ is always radially directed and is given by

$$R = \varepsilon r_0^3/r^2, \qquad r \geqslant r_0,$$

$$R = \varepsilon r, \qquad r \leqslant r_0.$$

where r_0 is the radius of the sphere, and ε is the misfit parameter, related to the unconstrained fractional difference in lattice parameter δ by the relationship ([26])

$$\varepsilon = \frac{3K\delta}{3K + 2E/(1+\nu)} \sim \frac{2}{3}\delta .$$

Here, K is the bulk modulus of the precipitate, and E and ν are the Young's modulus and Poisson's ratio respectively of the matrix. Figure 5 shows the situation. If one set of the atomic planes shown there are the Bragg planes, then it is clear that the Bragg plane passing through the centre will transmit background intensity. It follows that the image will be characterised by

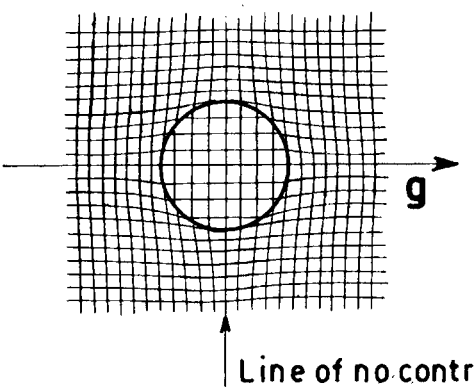


Fig. 5. – Figure showing the displacements around a misfitting sphere.

a line of no contrast perpendicular to g and passing through the centre of the precipitate. The line of no contrast is observed to swing around so that it is always perpendicular to the locally operative g.

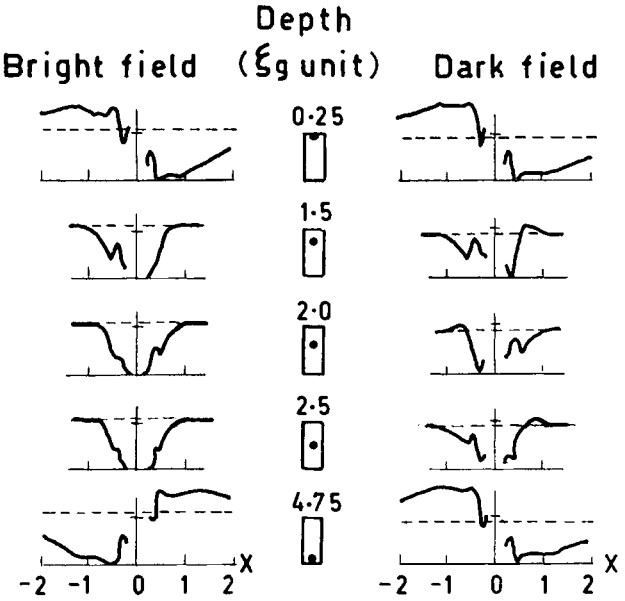


Fig. 6. – The variation of the images of a misfitting sphere with the depth of the sphere in the foil. Note that near either surface of the foil, very large, asymmetrical images are found, termed « anomalous images »; otherwise, the image width does not depend drastically on depth. x is the distance from the centre of the precipitate measured in extinction distances.

Ashby and Brown ([24a]) showed that for this model the width of the image is predominantly controlled by the parameter $\varepsilon g r_0^3/\xi_g^2$, and if r_0 were known, an estimate of the misfit parameter could be made. Furthermore, they concluded that the sign of the misfit parameter could be determined from the dark-field image. For displacements of the above type, the dark-field images from defects placed in varying positions along the column will be symmetrical. An example of this is shown in Fig. 6, where the image from a defect at depth $\xi_g/4$ is the same as the image from a defect at depth 19 $\xi_g/4$ in a foil of thickness 20 $\xi_g/4$. Furthermore, images from defects close to either surface of the foil are anomalously wide and characteristically black on one side and white on the other: if $\varepsilon > 0$, then the images are dark in the direction of positive $\boldsymbol{g}$, whereas if $\varepsilon < 0$ the reverse is true. This rule has come to be called the Ashby-Brown rule.

Further work has shown that this particularly simple picture must be modified. First we discuss the modifications to the image width as a result of many-beam effects, particle-shape effects, and elastic anisotropy. Then we discuss the sign-measuring methods which have been developed.

Howie and Basinski ([27]) have given a careful discussion of the effects of the approximations of the dynamical theory on the image width. Figure 7 shows their results. It will be seen that the effect of taking into account four beams instead of two changes the image width only slightly (at most 15 %) and Howie and Basinski conclude that the use of two-beam theory with a two-beam extinction distance will provide quite an accurate estimate of the image width even when quite strong systematic reflections are excited.

The effect of nonspherical shape of the inclusion has been discussed by Sass, Mura and Cohen ([28]). They conclude that for situations in which $\boldsymbol{g}$ is not perpendicular to a symmetry plane of the precipitate, the images are characteristically unsymmetrical, and they present images which compare well with their computed shape. Similar results have been observed in Cu-Co ([29]). However, Sass, Mura and Cohen suggest that the error in the measured value of the mismatch for a 220 reflection and a cubic precipitate may still be only 25 %, decreasing for small values of the mismatch.

The problem in the discussion of these effects is that the effects due to the « shape » of the precipitate and effects due to the elastic anisotropy of the matrix are inextricably linked. If the outline of the precipitate cannot be observed, which is commonly the case in systems with coherent precipitates where both the size and valence of the atoms in the precipitate tend to be similar to those of the matrix, the cubic symmetry of the images may be ascribed either to a cubically shaped precipitate or to the cubic anisotropy

of the matrix. Woolhouse and Brown ([30a,b]) have attempted to separate the two effects by making use of the principle that when the image is weak it tends to come from the projection of the periphery of the particle. This is true for spherical particles, and one might expect that the same would hold

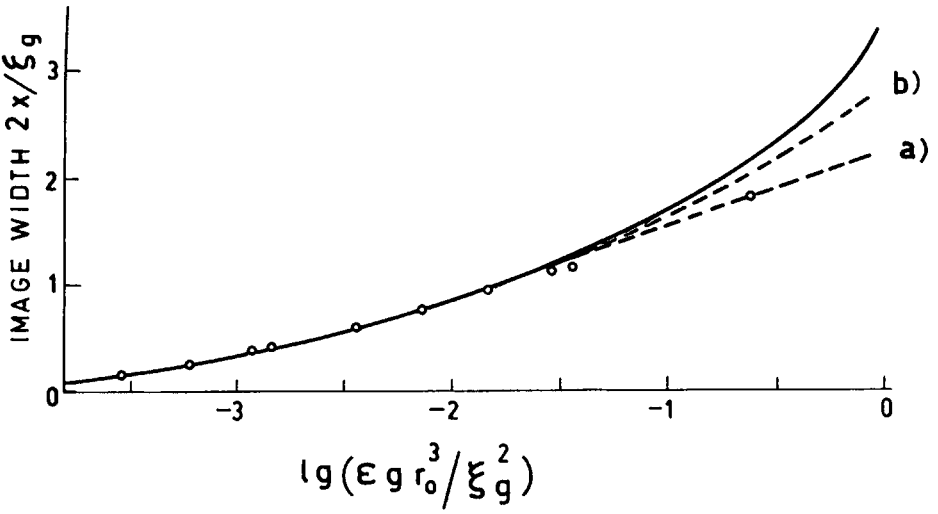


Fig. 7. – A curve of image width as a function of $\varepsilon g r_0^3/\xi_g^2$ (the parameter giving the « strength » of the coherency strain). The full line is from the two-beam dynamical theory; the points are from the four-beam dynamical theory; and the dotted lines are calculated on the basis of approximate formulae. (From ([27]) by courtesy of the *Phil. Mag.*)

for particles whose shape did not differ drastically from a sphere. To weaken the images, Woolhouse and Brown made observations under « quasi-kinematical » conditions, *i.e.* large values of s were used, corresponding to the Kikuchi band being displaced by $g/4$ from the Bragg condition (*). It appeared that the precipitates in Cu-Co were not spherical, but had the shape of a cubo-octahedron. However, much further (and rather intricate) work is required before this conclusion can be considered certain. Evidently the dark-field techniques of Cockayne, Ray and Whelan ([22]) can cast light on the problem.

(*) Application of the method of the most stationary phase shows that the image will have a peak at the periphery of the particle provided $s \geqslant g\varepsilon$.

The effect of elastic anisotropy can be judged from recent work by Yoffe ([31]). Her calculation makes use of the fact that the image shape depends not upon a full knowledge of the elastic displacements, but upon the Fourier transform of one of them. The calculation of the elastic displacement field in an anisotropic medium is a major numerical undertaking, but approximation methods exist for finding the Fourier transform. The result is that the expected image shape for any combination of elastic forces can be found simply and analytically, to an approximation whose validity has not yet been fully tested. Figure 8 shows an example of Yoffe's results: the figure is appropriate to a centre of pressure in copper, and the importance of elastic anisotropy on the image can be appreciated.

From a practical point of view, the most difficult measurement to make when estimating the mismatch is the particle size. If the outline of the particle can be clearly seen, as in the Cu-SiO_2 system ([24b]) the measurement will be most accurate. In other cases, the measurement may be quite wildly in error if the particle size is not known. As an example, we may take the Cu-Al_2O_3 system, which was judged by Ashby and Brown to have a very large mismatch. If the alumina particles are imaged in a reflection which is perpendicular to a symmetry plane of the precipitate, they show clear lines of no contrast, and Ashby and Brown assumed that the precipitate diameter was given by the length of the line of no contrast. However, other reflections yield very complicated images; and extraction replicas show that the precipitates are at least twice as large as estimated by Ashby and Brown. The mismatch is thus about ten times smaller than their estimate; the particles are not coherent. This conclusion is confirmed by the observation of spots from the precipitates in the diffraction pattern, and by other less direct evidence ([30b]).

Nonetheless, in situations where the particle size is known the strain measuring method can be applied and useful information extracted. The measurement of strains in plate-shaped precipitates and in dislocation loops is likely to be accurate for the same reasons. One might cite as examples of useful reliable measurements Brown and Mazey's ([32]) study of strains around gas bubbles in irradiated copper and stainless steel; and Weatherly's recent work ([25b]), in which he shows that at the surface of silica particles in copper, dislocations may be generated by stresses of about one-hundredth of the shear modulus.

We turn now to a discussion of the sign-measuring technique. It has been

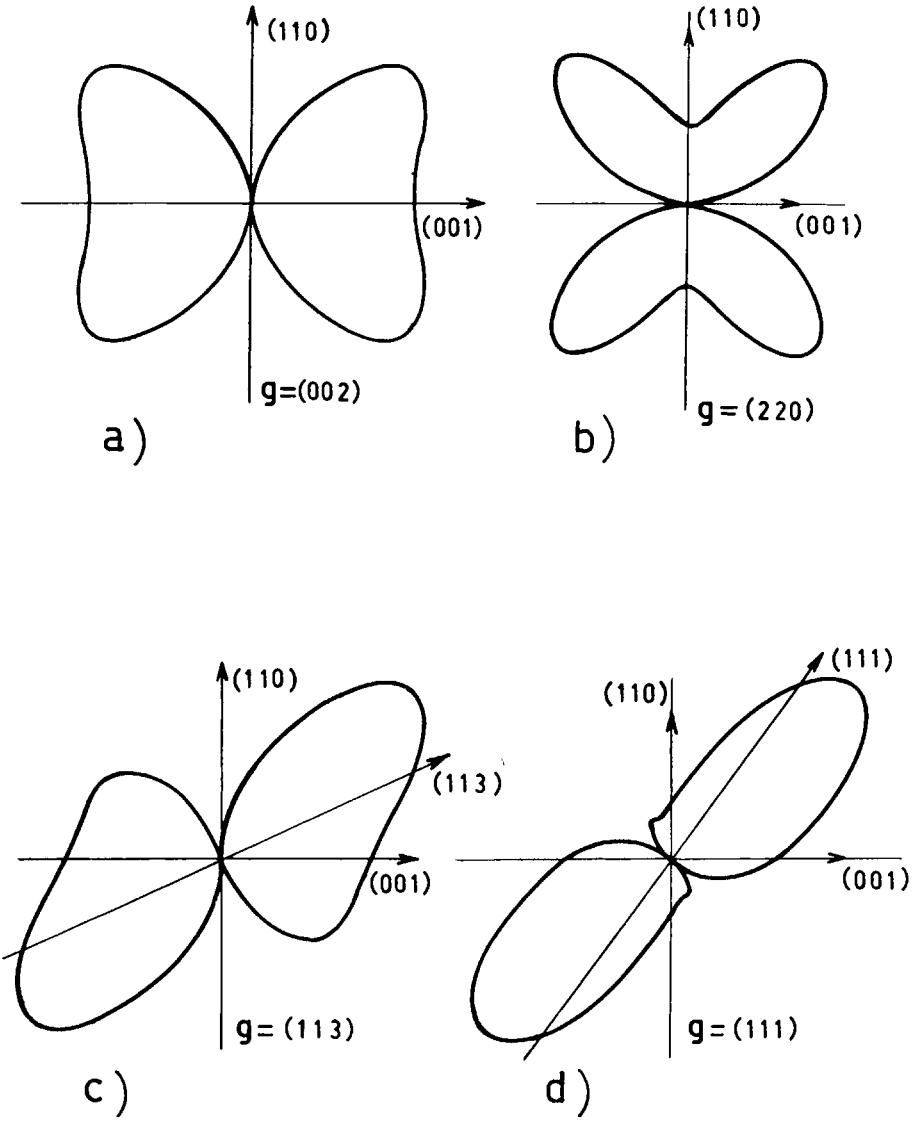


Fig. 8. – Figure showing the effect of elastic anisotropy on the image from a « centre of pressure » in copper. (A centre of pressure has the same external strain-field as a misfitting sphere). *a*) $g = (002)$; *b*) $g = (220)$; *c*) $g = (113)$; *d*) $g = (111)$.

applied successfully to many precipitation systems by a number of authors. However, when attempts were made to apply the method to small defect-clusters, it was found that the dark-field images did not display a unique black-white direction. (See, for instance, ([33,34])). In order to appreciate the problem, a certain amount of history is necessary. In irradiated metals, vacancies and interstitials are produced which are observable in the form of rather indistinct black dots, sometimes containing structure. Essmann and Wilkens ([33]) observed that when the dots were viewed under dynamical conditions, they showed a characteristic black-white contrast and it seemed that it might be possible to tell which dots were composed of vacancies and which of interstitials, and hence to discover something about the process of formation of the damage.

Before assessing the sign of these defects, it is necessary to decide on their geometry. Essmann and Wilkens, observing that the images were streaked always parallel to the trace of $\langle 111 \rangle$ directions, concluded that the defects were in the form of small Frank loops, of Burgers vector $\frac{1}{3}a\langle 111 \rangle$. This conclusion has been confirmed by Ruhle, Wilkens and Essmann ([35]) who performed machine calculations based on the two-beam dynamical theory and isotropic elasticity. Yoffe's theory ([31]), based on anisotropic elasticity has also been compared with observation by McIntyre, Brown and Eades ([36]) who find that the majority of the loops are indeed Frank loops.

However, the machine calculations showed that the image of a defect depends strongly on its depth; if we consider the dark-field case, an interstitial loop will have an image *dark* in the direction of positive $\boldsymbol{g}$ if it lies within $\xi_g/4$ of either foil surface; *bright* in the direction of $\boldsymbol{g}$ if it lies between $3\xi_g/4$ and $5\xi_g/4$ of either foil surface, and so on. These variations in contrast die out as the defect approaches the centre of the foil, and one can expect to see at most three or four reversals. The reader will appreciate that these oscillations are very similar to the oscillations associated with stacking-fault contrast, except that the *dark-field* oscillations are symmetrical and the phase of the oscillations is different; reversals occur at depths of $(2n + 1)\xi_g/4$ instead of at $\xi_g/2$. These differences are due to the different symmetry of the displacement function, and the proof of this forms the basis of problem 20. The predicted contrast from an interstitial loop whose Burgers vector and loop normal are parallel to $\boldsymbol{g}$ is shown in Fig. 9.

Now the elastic displacements due to a misfitting sphere, and those due to a prismatic loop are very similar; and it might be expected (and indeed it is true) that the contrast from precipitates and loops would be similar.

How then can the Ashby-Brown rule be reconciled with the present results? The answer lies in the effect of the stress-free surface of the foil. If the image of the defect is wide, a column near the edge of the image contains displace-

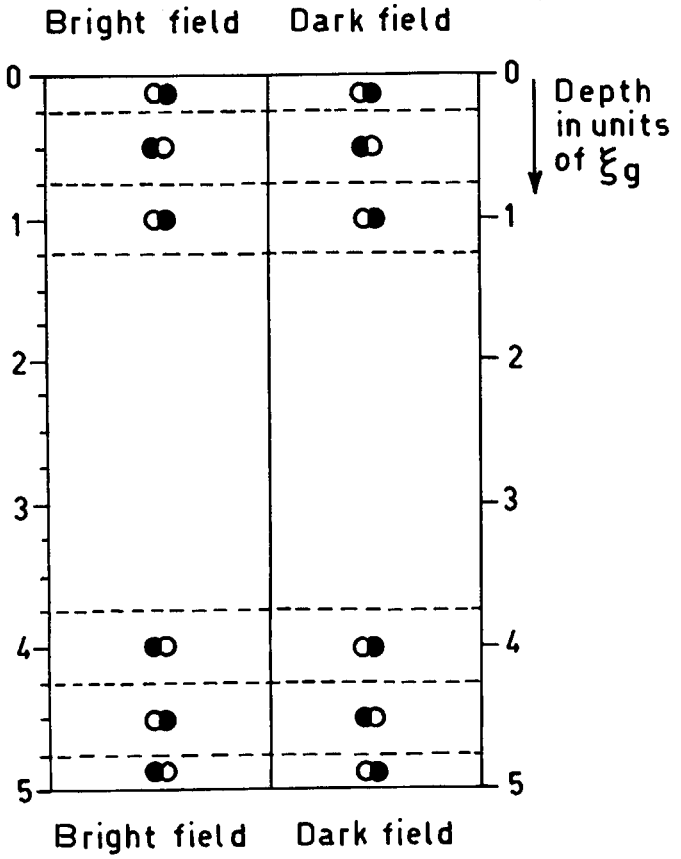


Fig. 9. – The predicted contrast from an interstitial loop as a function of depth in the foil.

ments that are strongly affected by the presence of the elastic « image » defect which creates (approximately) the stress-free surface. This displacement has the effect of reversing the sense of the black-white image in the second layer, and thus of making all the layers (except the scarcely visible fourth) obey the Ashby-Brown rule. McIntyre and Brown ([37]) show that if $\varepsilon g r_0^3/\xi_g^2$ is

greater than about 0.2, the Ashby-Brown rule will work; if it is less, it will not. Similarly for prismatic loops, the reversals in contrast will prevent the unique assignation of the sign of the Burgers vector unless $\boldsymbol{g}\cdot\boldsymbol{b}r_l^2/\xi_g^2 > 1$ (here $\boldsymbol{b}$ is the Burgers vector of the loop and r_l is its radius). Very similar conclusions were reached by Chik, Wilkens and Ruhle ([38]).

In the case of prismatic loops, the sign determination can never be done by application of the Ashby-Brown rule to dark-field micrographs, unless $\boldsymbol{g}\cdot\boldsymbol{b}$ is greater than about 4. Other methods must be used, and although various possibilities were tried by various authors, the best one is the stereo technique of Diepers and Diehl ([39]). In this technique, three pictures are taken of the same area: a stereo pair in kinematical conditions, and one in dynamical conditions. The depths of the defects are measured from the stereo pair, using a parallax bar, and then the measured depth is correlated with the observed black-white sense of the image in the dynamical photograph.

Unfortunately, application of this technique to radiation damage in neutron irradiated copper at first yielded opposite answers for the sign of the damage. Ruhle and Wilkens ([40]) found vacancy loops, whereas McIntyre ([41]) found interstitial loops. However, a later study by Ruhle, Haussermann, Huber and Wilkens ([42]) showed that a number of large interstitial loops are present, together with smaller vacancy loops; similarly Ipohorski and Brown ([43]) have observed, in addition to the large interstitial loops, a a few smaller vacancy loops. Both English and German work now shows that copper, irradiated at room temperature, contains large interstitial loops and smaller vacancy loops. A discrepancy between the two schools still exists concerning the estimated number of vacancy loops; however, one might expect the observation of these very small defects to be dependent upon both the resolution of the electron microscope and the details of the specimen preparation; certainly, very small clusters of vacancies must be present in both countries! The significance of these and related observations is discussed by Makin in this volume.

REFERENCES

1) P. B. HIRSCH, A. HOWIE, R. B. NICHOLSON, D. W. PASHLEY and M. J. WHELAN: *Electron Microscopy of Thin Crystals*, Butterworths, London (1965).

2*a*) M. J. WHELAN and P. B. HIRSCH: *Phil. Mag.*, **2**, 1121 (1957).

2*b*) M. J. WHELAN and P. B. HIRSCH: *Phil. Mag.*, **2**, 1303 (1957).
3) M. IPOHORSKI and M. S. SPRING: *Phil. Mag.*, **20**, 937 (1969).
4) A. HOWIE and B. JOUFFREY: *Phil. Mag.*, **14**, 201 (1966).
5) R. GEVERS, P. DELAVIGNETTE, H. BLANK and S. AMELINCKX: *Phys. Stat. Sol.*, **4**, 383 (1964).
6) M. J. GORINGE and U. VALDRÈ: *Proc. Roy. Soc.*, A **295**, 192 (1966).
7) A. J. ARDELL: *Phil. Mag.*, **16**, 147 (1967).
8) R. VINCENT: *Phil. Mag.*, **19**, 1127 (1969).
9) F. C. FRANK and J. H. VAN DER MERWE: *Proc. Roy. Soc.*, **A 198**, 216 (1949).
10) R. GEVERS, P. DELAVIGNETTE, H. BLANK, J. VAN LANDUYT and S. AMELINCKX: *Phys. Stat. Sol.*, **5**, 595 (1964).
11) A. HOWIE and M. J. WHELAN: *Proc. Roy. Soc.*, A **267**, 206 (1962).
12) W. T. READ: *Dislocations in Crystals*, McGraw-Hill, New York (1953), p. 116.
13) J. M. SILCOCK and W. J. TUNSTALL: *Phil. Mag.*, **10**, 361 (1964).
14) D. J. DINGLEY and K. F. HALE: *Proc. Roy. Soc.*, A **295**, 55 (1966).
15) L. K. FRANCE and M. H. LORETTO: *Proc. 4th Eur. Conf. on Electron Microscopy, Rome 1968* (Rome, 1968), vol. **1**, p. 301.
16) D. J. DINGLEY: *Proc. 4th Eur. Conf. on Electron Microscopy, Rome 1968* (Rome 1968), vol. **1**, p. 303.
17) M. H. LORETTO and L. K. FRANCE: *Phys. Stat. Sol.*, **35**, 167 (1969).
18) A. K. HEAD: *Austr. Journ. Phys.*, **20**, 557 (1967).
19) A. K. HEAD, M. H. LORETTO and P. HUMBLE: *Phys. Stat. Sol.*, **20**, 505, 521 (1967).
20) P. HUMBLE: *Phys. Stat. Sol.*, **21**, 733 (1967).
21) A. K. HEAD: *Austr. Journ. Phys.*, **22**, 43 (1969).
22) D. J. H. COCKAYNE, I. L. F. RAY and M. J. WHELAN: *Phil. Mag.*, **20**, 1265 (1969).
23) H. GLEITER: *Phil. Mag.*, **18**, 847 (1968).
24*a*) M. F. ASHBY and L. M. BROWN: *Phil. Mag.*, **8**, 1083 (1963).
24*b*) M. F. ASHBY and L. M. BROWN: *Phil. Mag.*, **8**, 1649 (1963).
25*a*) G. C. WEATHERLY and R. B. NICHOLSON: *Phil. Mag.*, **17**, 801 (1968).
25*b*) G. C. WEATHERLY: *Metal Sci. Journal*, **2**, 237 (1968).
25*c*) G. C. WEATHERLY: *Phil. Mag.*, **17**, 791, (1968).
26) N. F. MOTT and F. R. N. NABARRO: *Proc. Phys. Soc.*, **52**, 86 (1940).
27) A. HOWIE and Z. S. BASINSKI: *Phil. Mag.*, **17**, 1039 (1968).
28) S. L. SASS, T. MURA and J. B. COHEN: *Phil. Mag.*, **16**, 679 (1967).
29) H. P. DEGISCHER: unpublished work.
30*a*) G. R. WOOLHOUSE and L. M. BROWN: *Proc. 4th Eur. Conf. on Electron Microscopy, Rome 1968* (Rome, 1968), vol. **1**, p. 297.
30*b*) G. R. WOOLHOUSE and L. M. BROWN: *Journ. Inst. Met.*, **98**, 106 (1970).
31) E. H. YOFFE: *Phil. Mag.*, **21**, 833 (1970).
32) L. M. BROWN and D. J. MAZEY: *Phil. Mag.*, **10**, 1081 (1964).
33) U. ESSMANN and M. WILKENS: *Phys. Stat. Sol.*, **4**, K 53 (1964).
34) W. BELL, D. M. MAHER and G. THOMAS: *Lattice Defects in Quenched Metals*, Academic Press, New York (1965), p. 739.
35) M. RUHLE, M. WILKENS and U. ESSMANN: *Phys. Stat. Sol.*, **11**, 819 (1965).

36) K. G. McIntyre, L. M. Brown and J. A. Eades: *Phil. Mag.*, **21**, 853 (1970).
37) K. G. McIntyre and L. M. Brown: *Journ. Phys.*, **27**, C3-178 (1966).
38) K. P. Chik, M. Wilkens and M. Ruhle: *Phys. Stat. Sol.*, **23**, 113 (1967).
39) H. Diepers and J. Diehl: *Phys. Stat. Sol.*, **16**, K109 (1966).
40) M. Ruhle and M. Wilkens: *Phil. Mag.*, **15**, 1075 (1967).
41) K. G. McIntyre: *Phil. Mag.*, **15**, 205 (1967).
42) M. Ruhle, F. Haussermann, P. Huber and M. Wilkens: *Proc. 4th Eur. Conf. on Electron Microscopy, Rome 1968* (Rome, 1968), vol. **1**, p. 397.
43) M. Ipohorski and L. M. Brown: *Phil. Mag.* **22**, 931 (1970).

The Application of Electron Microscopy to Radiation Damage Studies

M. J. MAKIN

Metallurgy Division, A.E.R.E. - Harwell, England

1. The damage process.

1·1. The relevance of radiation damage studies.

The need for radiation damage studies really began when it was realized that it was possible to make the uranium fission reaction self-sustaining, *i.e.* to use the neutrons emitted in the fission process to produce further fissions:

$$^{235}\mathrm{U} + \mathrm{n} \rightarrow \mathrm{A} + \mathrm{B} + (2 \div 3)\mathrm{n} + 160 \text{ MeV} .$$

Furthermore, the fission process is *controllable*, so that it is possible to release energy slowly and over a long period. This can be done because some of the neutrons are emitted after delays of up to one second, so increasing the doubling time, *i.e.* the time required for the neutron flux to double in density, and giving the control apparatus sufficient time to move neutron absorbing material either into or out of the reactor so as to depress or increase the neutron flux, and hence the rate of heat generation. It was discovered that there were substantial practical advantages in keeping the density of fissile material low. Only about 0.7% of natural uranium is ^{235}U, and increasing this concentration by removing ^{238}U atoms is expensive. To obtain a self-sustaining reaction in natural uranium, however, it is necessary to slow the neutrons down to thermal energies ($\sim$ 1/40 eV) from their fission energy of (1÷10) MeV. This is accomplished by arranging for the neutrons to diffuse through a mass containing a large number of light atoms, since the energy

loss per collision is greatest when the incident and struck particles are of equal mass. The choice of this material however is also governed by its capture cross-section for neutrons. If this is too high then too many neutrons will be lost for the fission process to be self-sustaining. The practical choice of moderators, as such substances are known, is hence very limited, and in practice is restricted to virtually two: graphite and heavy water. In addition to the moderator, arrangements have to be made for cooling the uranium billets, which will of course get hot from the heat generated by the fission process. Similar neutron capture considerations apply to the structure of the reactor as to the moderator *i.e.* materials with high capture cross-sections can be used only sparingly. The general outlines of a nuclear reactor thus emerge. Many different types of reactor have now been studied, and a wide range of different designs built. In general, as time has progressed the degree of ^{235}U enrichment has tended to increase, so enabling the reactor to be smaller (since the critical mass is smaller) and the operating temperatures have increased, so raising the thermodynamic efficiency of the heat engine coupled to the coolant outlet. The power output per reactor has also greatly increased. Reactors are now the heat sources in a substantial number of power stations. In recent years interest has increased in the so-called « fast » reactors, in which the moderator is dispensed with, and the nuclear reaction sustained by using the much smaller capture cross-section of fissile materials for fission energy neutrons. Although these reactors are technologically difficult, since they are small and hence the heat and radiation fluxes are high, and expensive to fuel, since they require a very high enrichment, they have the over-riding advantage that they can be made to « breed », *i.e.* to generate more fissile material than they consume. By the use of such reactors the available heat content of the world supply of uranium is of course multiplied many times, since it is now theoretically possible to fission 100% of the uranium, instead of only 0.7%.

The great advantages of nuclear power, *i.e.* the independence from the need to transport large quantities of fuel to the power station, the complete absence of atmospheric pollution, and the very large reserves of power available, are sufficient to ensure that in time a high proportion of the world's power will be generated from nuclear sources. Already in the U.K. $\sim 20\%$ of the total quantity of electricity generated comes from the nuclear stations, and this percentage will rise steadily.

The development of the nuclear power industry has been very rapid, occurring almost entirely within the last twenty five years, and it has required the development of many new materials and technologies. For example,

the need for low capture cross-section structural materials led to the development of new magnesium and zirconium alloy systems with the required nuclear cross-section, corrosion and fabrication properties. The new coolants required, such as CO_2 and liquid sodium, posed many problems not before encountered and the unusual mechanical and physical properties of such materials as uranium and graphite posed many problems. In addition to the problems which occur when any new material, or new type of structure, are used there is also one completely new parameter, radiation damage. For the first time large and complex structures were required to operate at high temperatures while being constantly irradiated by energetic neutrons, fission fragments, knocked-on atoms and γ-rays. It was soon found that the effects of these radiations were substantial, and highly novel, and much work was, and still is, necessary to determine the nature of these effects and how they may endanger the continued safe operation of the reactor. The problem is not made any easier by the fact that unlike almost any other engineering structure, is it not possible to repair, or even inspect, much of the structure once it has operated, because of the high radiation levels which build up, and which would take many years to decay away to safe levels. Hence, once a reactor has operated, the possibility of changes in the design or repairs are limited.

It is the purpose of these lectures to try to demonstrate how electron microscopy has been used to study both the practical radiation damage problems, and also the nature of the damage itself in different materials. (See also Goringe and Hall, this volume.) As it happens, the radiation produced defects responsible for many of the macroscopic effects are in the right size range to be observable by transmission electron microscopy, and the technique has hence proved to be extremely valuable in establishing the *mechanisms* by which the macroscopic effects occur. The impact has been so great that it is difficult now to remember the air of mystery which used to surround many of the effects, such as the growth and hardening phenomena for example, before the advent of electron microscopy. Many examples can also be quoted of where the use of electron microscopy has greatly reduced the effort and expense involved in studying a particular radiation effect. This occurs because when the defect responsible for the macroscopic effect has been identified in the microscope it is possible to carry out experiments under a wide range of experimental conditions using very small specimens which occupy much less reactor space and absorb many fewer neutrons than the large samples required for macroscopic measurements. An extreme example of this is the simulation of very high neutron dose effects in an accelerator. In this case it is possible to simulate the effect

of many years irradiation in a few hours, so enabling studies to be made which would otherwise be impossible.

It should not be supposed, however, that all the problems have been solved. In fact as higher flux reactors are designed new radiation damage problems are being encountered, and it is unfortunately a fact that the economics of many of the existing reactors could be considerably improved if the radiation damage problems could be solved. It is likely, therefore, that this subject will be actively studied for many years yet.

1·2. The primary event.

1·2.1. *Basic processes.* – When solids are bombarded by radiation there are three general types of damage: the removal of electrons from their normal orbits, the displacement of atoms from their normal sites and the introduction of impurities, either by nuclear transmutations or by the bombarding ions stopping within the solid. In these lectures we shall be primarily concerned with the last two effects. The displacement of electrons produces no effects in good conducting materials, where the displaced electrons are rapidly replaced by conduction electrons. Effects are produced in insulators, notably the colouration effects in alkali halides. Also the displacement of electrons can produce « displaced » atoms by the destruction of the chemical bonding in the material, and a striking example is the large effect produced in plastics. Although such effects can be observed in the electron microscope, microscopy has not been extensively utilised to study them as yet and I shall not deal with this subject.

1·2.2. *The displacement process.* – The simplest primary event is the collision between a charged particle and the atomic nucleus. This can be treated as a two-body collision provided that the mean free path between collisions is much greater than the interatomic spacing. If we initially assume that the collisions are elastic and that the velocities are sufficiently low for nonrelativistic mechanics to apply then from the laws of conservation of energy and momentum it is easy to show that

$$E_2 = \Lambda E_1 \sin^2 \varphi/2 ,$$

where

$$\Lambda = \frac{4M_1 M_2}{(M_1 + M_2)^2} ,$$

E_2 is the energy transferred to the struck atom, E_1 is the energy of the incident

particle φ is the angle of deflection of the incident particle as a result of the collision and M_1 and M_2 are the masses of the incident and struck atoms. A has special significance when $\varphi = \pi$, which is a head-on collision since the maximum energy transfer occurs for such a collision, $E_2(\max) = AE_1$. In any solid there is obviously a minimum value of E_2 for the production of damage. This is known as the *displacement energy* E_d and may be expected to depend on the crystal structure, the interatomic forces between the atoms, the direction of displacement, the temperature etc. Theoretically, therefore, it is incorrect to assume a constant value for E_d, but the difficulties encountered in calculating E_d are such that this assumption is necessary in practice for most materials, and the value of 25 eV is widely used. Unfortunately E_d is not an easy quantity to measure experimentally and although many such measurements have been made, with the results quoted in Table I,

TABLE I. – *Measured values of displacement energy.*

Element	E_d (eV)	Reference
Cu	22 ± 3	[1]
	$19 \div 20$	[2]
	22	[3]
	$19 \div 22$	[4]
Al	16	[4]
	19	[5]
	32	[3]
Au	$33 \div 36$	[6]
Pt	37	[7]
	36	[8]
Fe	24	[3]
Mo	37	[3]
W	> 35	[3]
Ni	24	[3]
Ti	29	[3]
Ag	28	[3]

in very few cases has the orientation dependence been measured. Where this has been done the results (Table II) show that there does not appear to be any single direction with a substantially lower energy. If this were the case then when the values obtained in the other directions are multiplied by (cosine)^2 of the angle between that direction and the lowest energy direction ($\langle 110 \rangle$ in face centred crystals) then a constant value would be obtained. It can be seen from Table II that this is not the case.

TABLE II. – *Displacement energies in the principal directions in some f.c.c. crystals.*

	⟨110⟩	⟨100⟩	⟨111⟩	$(\cos^2 45°)E_{100}$	$(\cos^2 35°)E_{111}$
Au [9]	16	23	20	11.4	13.3
Ag [9]	15	22	19	10.9	12.7
Cu [9]	16	—	—	—	—
Cu [10]	25	24	80	11.9	—
Cu [11]	19	19	—	9.4	—
Cu [12]	19.2	21.6	23.6	10.8	15.7

In the case of electrons, relativistic quantum mechanics must be used to calculate the energy transfer to struck atoms. Because of the disparity of masses it can be assumed that the electron velocity is unaltered by the collision, in which case the momentum transfer ΔP in a collision in which the electron is scattered through an angle θ is $2mv \sin \theta/2$, where m and v are the relativistic mass and velocity of the electron. The energy transfer is $(\Delta P/2M_2)^2$ and hence

$$E_2 = \frac{2m^2v^2}{M_2} \sin^2 \theta/2 .$$

Since

$$m = m_0 \left(1 - \frac{v^2}{c^2}\right)^{-\frac{1}{2}} \quad \text{and} \quad E = (m - m_0)c^2 ,$$

we get

$$E_2 = \frac{2E(E + 2m_0c^2)}{M_2c^2} \sin^2 \theta/2 .$$

Using this formula and the one given previously, the order of magnitude of the particle energies required to transfer the displacement energy is given in Table III.

TABLE III.

Particle	Target	E_1 (eV)
Electron	Li	10^4
Electron	U	10^6
Proton	Li	10^1
Proton	U	10^3
Heavy ion ($M = 100$)	Li	10^1
Heavy ion ($M = 100$)	U	10^3

1'2.3. *The collision cross-section.*

a) *Light ions.* The cross-section for displacement collisions when bombarding with charged particles is calculated by assuming that the atoms are displaced by the Coulomb interaction between the nuclei, and that the effect of the screening electrons is to cut off this interaction sharply at a radius of about $a_0/Z^{\frac{1}{3}}$, where a_0 is the Bohr radius and Z the atomic number. When the energy of the moving particle is smaller than that required to give an impact parameter equal to the screening radius then the collision approximates to a hard sphere collision, there being no interaction until the electron cloud is penetrated, and then a relatively large force exists for a short time. There is thus a definite lower particle energy below which the simple Rutherford collision formula must not be used. This energy limit is given approximately by:

$$L_A = 2E_R Z_1 Z_2 (Z_1^{\frac{2}{3}} + Z_2^{\frac{2}{3}})^{\frac{1}{2}} \cdot \frac{(M_1 + M_2)}{M_2},$$

where E_R is the Rydberg energy (13.6 eV). As can be seen in Table IV this energy limit increases rapidly as the weight of the moving ion increases and hence for heavy ions the simple Rutherford formula is not very relevant. For light energetic charged particles however the formula is believed to be

TABLE IV.

Particle	Target	L_A (eV)	L_B (eV)
D^+	C	$4 \cdot 10^2$	$8 \cdot 10^2$
	Al	$1 \cdot 10^3$	$2 \cdot 10^3$
	Cu	$3 \cdot 10^3$	$8 \cdot 10^3$
	Au	$1 \cdot 10^4$	$4 \cdot 10^4$
C	C	$5 \cdot 10^3$	$3 \cdot 10^5$
Al	Al	$3 \cdot 10^4$	$9 \cdot 10^6$
Cu	Cu	$2 \cdot 10^5$	$4 \cdot 10^8$
Au	Au	$2 \cdot 10^6$	$4 \cdot 10^{10}$

fairly accurate. Above L_A the effect of the screening is to cut off the Rutherford collisions at values of the screening radius equal to the impact parameter. This limits the minimum energy which can be transferred to:

$$E^* = \frac{4E_R^2 Z_1^2 Z_2^2 (Z_1^{\frac{2}{3}} + Z_2^{\frac{2}{3}}) M_1}{M_2 E}.$$

As long as E^* is more than the displacement energy all Rutherford collisions displace atoms, but when the energy of the moving atom exceeds L_B, where:

$$L_B = \frac{4E_R^2 Z_1^2 Z_2^2 (Z_1^{\frac{2}{3}} + Z_2^{\frac{2}{3}}) M_1}{M_2 E_d},$$

then only some collisions do so. When $E \gg L_B$ then only half the energy lost in Rutherford collisions is in the displaced atoms.

The cross-section σ_p for displacement by Rutherford collisions is given by:

$$\sigma_p = 4M_1 Z_1^2 Z_2^2 E_R^2 \left(1 - \frac{E_d}{\Lambda E}\right) \frac{\pi a_0^2}{M_2 E_d E}, \qquad (\pi a_0^2 = 8.8 \cdot 10^{-17}\ \mathrm{cm}^2).$$

The term $(1 - E_d/\Lambda E)$ is nearly unity in all cases of practical interest and hence can be omitted. The cross-section is therefore inversely proportional to E. The mean energy $\bar{E}$ of the knock-ons is

$$\bar{E} = E_d \ln\left(\frac{\Lambda E}{E_d}\right).$$

It is a characteristic of Rutherford collisions that they have a relatively high cross-section σ_p but that the mean energy of the knock-ons is low (only a few times the displacement energy).

b) *Heavy ions.* For low-energy heavy ions and primary knock-ons, where the energy is below L_A in Table IV, considerable uncertainty exists as to the correct treatment. The simplest model is the hard sphere approximation in which the atoms are treated as billiard balls, the energy being shared randomly between the colliding atoms (assuming they have equal masses). Since an atom can only create a *new* displacement when it has an energy greater than $2E_d$ the average number of displaced atoms is

$$N_d = \bar{E}/2E_d,$$

where $\bar{E}$ is the initial energy. It is emphasized that this is only the *average* figure and the number of displacements can vary widely from this value in individual cases. An alternate approach is the use of the inverse square approximation for the potential and in this case the cross-section and mean

energy are:

$$\sigma_p = \frac{\pi^2 a^2 L_A \sqrt{\Lambda}}{4\sqrt{E E_d}} \quad \text{and} \quad \bar{E} = \sqrt{\Lambda E E_d}\,.$$

c) *Fission fragments.* Fission fragments are high energy heavy ions (typically $M_1 = 96$, $E_1 = 95$ MeV and $M_2 = 137$, $E_2 = 55$ MeV). At such high energies the Rutherford collision model is adequate and shows that primary recoils are produced at intervals of only a few Å along the track, with a recoil energy of (500÷1000) eV. The vast majority of a fission fragment's energy is lost by ionisation however, and this gives rise to some unusual features which will be described later.

d) *Electrons.* As mentioned previously electrons cannot be treated with classical formulae. An approximate solution of the Dirac equation for light elements has been given by McKinley and Feshbach ([14]):

$$\sigma_p = \frac{4\pi a_0^2 Z_2^2 E_R^2 E_m}{m_0^2 c^4 E_d}\left(\frac{1-\beta^2}{\beta^4}\right) \cdot \left[1 + 2\pi\alpha\beta\left(\frac{E_d}{E_m}\right)^{\frac{1}{2}} - \frac{E_d}{E_m}\left\{1 + 2\pi\alpha\beta + (\beta^2 + \pi\alpha\beta)\ln\frac{E_m}{E_d}\right\}\right],$$

where E_m is the maximum recoil energy and $\alpha = Z_2/137$.

This formula is reasonably accurate for the light elements, but seriously underestimates σ_p for heavy elements. Oen ([15]) has computed the cross-section by a method which avoids the McKinley-Feshbach approximation and has found that in gold the cross-section at 1.7 MeV is over four times the McKinley-Feshbach value.

e) *Neutrons.* Collisions between neutrons and atoms is a special case, since having no charge a neutron does not interact with a nucleus except at very short ranges. The available neutron scattering data suggests that the collision is then essentially hard sphere, with a cross-section of the order $3 \cdot 10^{-24}$ cm^2 at an energy of (1÷2) MeV. The mean energy of the knock-ons is hence

$$\bar{E} = \tfrac{1}{2}\Lambda E$$

and, because E is typically (1÷2) MeV for fission neutrons $\bar{E}$ is large (10^5 eV). An important complication of neutron damage is that in a thermal reactor neutrons of all energies are present, and the calculation of the number of

primary knock-ons demands a knowledge of the spectrum in the irradiation position (the spectrum changes rapidly as a fuel element is approached). In a typical irradiation position close to fuel element in a thermal reactor the mean knock-on energy may be $\sim AE_{\text{fiss}}/10$.

f) *Transmutations.* A further complication with neutrons is that nuclear reactions are possible with neutrons of all energies, and the cross-sections for some of these reactions are very large for thermal neutrons. The cross-sections tend to increase at lower neutron energies because of the low velocity of the neutron and the consequently long time it spends near a nucleus. Typical of such reactions are the gas producing reactions:

$$^{6}\text{Li}+\text{n} \xrightarrow{950\text{b}} {}^{3}\text{H}+{}^{4}\text{He}$$

and

$$^{10}\text{B}+\text{n} \xrightarrow{3990\text{b}} {}^{7}\text{Li}+{}^{4}\text{He}\,.$$

Some of the effects of such reactions will be described later.

g) (nγ) *reactions.* Another complication is the possibility of damage production by (nγ) reactions. These are only of importance in a well thermalized flux; in such cases however they can account for the majority of the displacement damage. The recoil energy which an atom acquires as a result of such an event is typically (100÷500) eV but the cross section may be as large as 2500 b (cadmium). In aluminium however the cross-section is only 0.23 b, and in copper 3.8 b.

h) *Summary.* Table V is a useful summary of the properties of the recoil atoms produced by the main types of radiation. The mass of the struck atoms M_2 is assumed to be 50.

TABLE V.

Particle	Spacing between events (cm)	Track length of particle (cm)	Mean recoil energy (keV)
1 MeV proton	10^{-3}	10^{-3}	0.2
100 MeV fission fragment	10^{-7}	10^{-4}	1
50 keV heavy ion	10^{-6}	10^{-5}	7
1 MeV electron	0.1	0.1	0.05
2 MeV neutron	5	100	160
Thermal reactor neutron	5	100	10

1·3. Collision cascades.

In many of the preceding cases, and particularly for neutrons, the primary knock-ons are sufficiently energetic to produce many more displacements. These primary knock-ons are heavy ions and hence have a short range and a high collision cross-section. Also, being heavy, they transfer large amounts of energy at each collision, especially as they near the end of their range, so producing further energetic knock-ons which produce more displacements. The calculation of the number of displacements in such a cascade has attracted considerable attention. The simplest possible model is to assume hard sphere collisions, when the number of displacements is simply $\bar{E}/2E_d$, where $\bar{E}$ is the mean primary knock-on energy so that the concentration of point defects is

$$C_d = t\varphi\sigma_p \frac{E}{2\bar{E}_d}, \quad \text{for } \Lambda E \ll E_d,$$

where t is the irradiation time and φ the irradiation flux.

Alternative estimates are, for light ions:

$$C_d = t\varphi \frac{2\pi a_0^2 M_1 Z_1^2 Z_2^2 E_R^2}{M_2 E E_d} \ln \frac{\Lambda E}{E_d},$$

for heavy ions:

$$C_d = \frac{t\varphi\pi^2 a^2 L_A \Lambda}{8E_d}$$

and fast neutrons:

$$C_d = t\varphi\sigma_{\text{total}} \frac{\Lambda E}{4E_d}.$$

Computer simulations of cascades have been made ([16]), and a feature of these is that they enable the spatial distribution of the displacement to be seen. In general it is found that the vacancies tend to lie in a compact central volume whose centre of gravity is behind that of the interstitials, which tend to lie in a larger volume.

1·4. Crystal lattice effects.

In all the preceding the solid has been assumed to be amorphous. The crystal has several effects on the radiation damage produced in it.

1'4.1. *Channelling.* – Heavy particles fired in specific directions at crystals penetrate very much larger distances than similar particles fired in random directions. This was strikingly demonstrated by Piercey *et al.* ([17]) who showed that the penetration of 40 keV ^{85}Kr ions in aluminium was strongly dependent on the orientation, and much greater in low index directions (maximum in 110). It was later found by Nelson and Thompson ([18]) that the effect diminished and eventually disappeared as the bombardment continued. These effects were interpreted in terms of the focusing of the bombarding ion into the centre of the channel by glancing collisions with the closely packed rows of atoms forming the edges of the channel. The disappearance of the effect was interpreted as the blocking of the channels by accumulated defects in the lattice. Such effects are now well established, both experimentally and by computer simulations. Much weaker planar channelling has also been discovered for heavy ions ([18]).

Channelling clearly affects the radiation damage produced. Firstly it reduces the average number of displaced atoms because a channelled atom is less likely to make displacement collisions. Secondly the long range of channelled particles spreads the damage over a larger crystal volume. Thirdly the rate of damage will be dose dependent, since blocking of the channels will return the crystal towards the amorphous behaviour. Fourthly, it will introduce a temperature dependence into the damage function as at high temperatures the channels will be less perfect.

Channelling also occurs during electron irradiation, and the effect of this on radiation damage will be demonstrated in Sect. **4**'3.2. In this case, of course, the behaviour must be described in terms of the wave nature of the electron beam. The effect of electron channelling on the inelastic scattering has been known for many years in electron microscopy, being the reason for the well known anomalous transmission effect.

1'4.2. *Focusing.* – A further effect of the crystal structure is the possibility of the focusing of momentum along close packed directions. This was first analysed by Silsbee ([19]). This occurs only at low energies (up to a few hundred eV). Computer simulations of collision cascades have shown that assisted focusing can also occur at energies above the limit for simple focusing. These occur by the focusing effect of the rings of atoms surrounding the row along which the sequence is passing. It is assisted focusing which enables sequences to occur in the ⟨100⟩ as well as the ⟨110⟩ directions in the *fcc* and *bcc* lattices.

Focusing has only a small effect on the production of damage in perfect

crystals. In defected crystals however, there is a possibility of an enhanced damage rate at discontinuities, due to de-focusing. The range of focused sequences is short (of the order of 100 Å in gold ([20])). Focused replacement sequences can also occur, in which each atom replaces the atom ahead of it in the sequence. Such a sequence produces an interstitial atom when it is de-focused. The maximum energy at which a replacement sequence can occur, however, is only about half that of a focusing sequence.

1·5. **Thermal spikes.**

The last question to be discussed in this lecture is the question of « thermal spikes », *i.e.* the local heating which occurs in a collision cascade. As the cascade proceeds the energy is distributed between larger and larger numbers of atoms within the cascade volume and hence it becomes reasonable to think of the local « temperature ». There is now little doubt that high local temperatures are attained in the cascades ([21]), (see Table VI).

TABLE VI.

Metal	Recoil energy (keV)	T (°K)	Radius (Å)	Duration (10^{-12} s)
Au	43	910	110	3
Ag	45	530	134	6
Cu	38	49	130	5
Zn	39	150	250	1
Ni	42	600	128	9
Ge	41	1060	95	3

The effect of these spikes in promoting point defect recombinations or clustering is difficult to establish quantitatively and is relevant only to very low temperature irradiations (where the interstitials would otherwise be immobile). There is plenty of evidence that at room temperature, for example, there is both clustering and recombination occurring between the defects within each cascade. Some of the evidence for this will be discussed later.

REFERENCES (Section 1)

1) J. W. CORBETT, J. M. DENNY, M. D. FISKE and R. M. WALKER: *Phys. Rev.*, **108**, 954 (1957).
2) A. SOSIN: *Phys. Rev.*, **126**, 1968 (1962).

3) P. G. LUCASSON and R. M. WALKER: *Phys. Rev.*, **127**, 485 (1962).
4) G. W. ISELER, H. I. DAWSON, A. S. MEHNER and J. W. KAUFFMAN: *Phys. Rev.*, **146**, 468 (1966).
5) H. M. NEELY and W. BAUER: *Phys. Rev.*, **149**, 535 (1966).
6) W. BAUER and A. SOSIN: *Phys. Rev.*, **135**, 521 (1964).
7) E. A. BURKE, C. M. JIMENEZ and L. F. HOWE: *Phys. Rev.*, **141**, 629 (1966).
8) W. BAUER and W. F. GROEPPINGER: *Phys. Rev.*, **154**, 584 (1967).
9) G. J. OGILVIE: *Defects and Radiation Damage in Metals*, Cambridge University Press (1969), p. 327.
10) J. B. GIBSON, A. N. GOLAND, M. MILGRAM and G. H. VINEYARD: *Phys. Rev.*, **120**, 1229 (1960).
11) A. SOSIN and K. GARR: *Phys. Stat. Sol.*, **8**, 481 (1965).
12) M. J. MAKIN: unpublished results.
13) G. H. KINCHIN and R. S. PEASE: *Rep. Progr. Phys.*, **18**, 1 (1955).
14) W. A. MCKINLEY and H. FESHBACH: *Phys. Rev.*, **74**, 1759 (1948).
15) O. S. OEN: USAEC report, ORNL-3813 (1965).
16) J. R. BEELER: *Journ. Appl. Phys.*, **35**, 2226 (1964).
17) G. R. PIERCEY, M. MCCARGO, F. BROWN and J. A. DAVIES: *Canad. Journ. Phys.*, **42**, 1116 (1964).
18) R. S. NELSON and M. W. THOMPSON: *Phil. Mag.*, **8**, 94 (1963).
19) R. H. SILSBEE: *Journ. Appl. Phys.*, **28**, 1246 (1957).
20) B. W. FARMERY and M. W. THOMPSON: *Phil. Mag.*, **18**, 415 (1968).
21) G. H. VINEYARD: *Radiation Damage in Solids*, Academic Press (1962).
22) R. S. NELSON: *Phil. Mag.*, **11**, 291 (1965).

2. The nature of the damage: Basic effects.

2·1. General.

In the first lecture I described how atoms can be displaced from their normal sites by irradiation with energetic particles, and how the concentration of point defects produced can be estimated. To progress further we must investigate briefly the nature and properties of these « point defects », and then, most important of all, how they cluster into groups, for it has been found that most of the important radiation damage effects are the result of *clustered* defects, rather than individual defects. It is the ability of the electron microscope to reveal the defect clusters which has made it so useful in radiation damage studies, and this applies equally to the point defects and the impurity atoms produced. It is at this stage therefore that the correlation can be made between the macroscopic effects of radiation and the electron microscope results. This subject will occupy the next two lectures. The final lecture

will be devoted to a related, but more specialised topic, the actual *production* of radiation damage in the electron microscope. (See also Goringe and Hall, this volume.) In the majority of materials, displacement effects do not occur during electron irradiation until the electron energy substantially exceeds 100 keV, and hence little has hitherto been heard of this subject in electron microscopy. With the advent of high voltage microscopes, however, this effect will become of considerable practical importance to electron microscopists.

2·2. Point defects.

2·2.1. *Energy and properties.*

a) *Formation energy.* Although the experimental values of E_d, the energy required to displace an atom, so creating an interstitial and a vacancy, are experimentally found to be of the order of 25 eV, this does not represent the sum of the formation energies of the two defects, because to produce a *stable* interstitial-vacancy pair it is necessary to separate the two defects by several lattice spacings. If this is not done then mutual annihilation will occur immediately, even at 0 °K, due to the elastic interaction between the defects. The sum of the formation energies of an interstitial and a vacancy is in general only about one quarter of the displacement energy, for this reason. It is important to realise that the equilibrium concentration of point defects in a crystal is not zero, even in the absence of radiation. The reason for this is that defects can be created due to the thermal vibration of the atoms. The concentration C of defects so produced is:

$$C = \exp\left[\frac{\Delta S_f}{K}\right] \cdot \exp\left[-\frac{E_f}{KT}\right],$$

where E_f is the formation energy, T the absolute temperature and ΔS_f the entropy of formation. From this formula it can be seen that *a*) the concentrations will be very dependent on temperature, increasing rapidly as the temperature is raised and *b*) they will only be significant ($> 10^{-5}$) when E_f is of the order of 1 eV or less ($KT = 0.086$ eV at 1000 °K).

b) *Migration energy.* In addition to being formed thermally, point defects can also migrate due to thermal fluctuations in the crystal lattice. This migration appears to require a well defined energy E_m, and hence the

rate of migration between stable sites, assuming Maxwell-Boltzmann statistics, is:

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \nu \exp\left[\frac{\Delta S_m}{K}\right] \cdot \exp\left[-\frac{E_m}{KT}\right],$$

where ν is the oscillation frequency of the defect, E_m is the migration energy and ΔS_m the entropy of migration. It is clear that the rate at which an excess of defects decays to the equilibrium concentration at any temperature depends not only on the rate of jumping $\mathrm{d}n/\mathrm{d}t$, but also on the number of jumps n required to reach a sink. The value of n depends greatly on the annihilation process, *i.e.* whether the defects are mutually combining, or migrating to fixed sinks.

c) Effect on physical properties. Point defects affect the physical properties of crystals. Clearly they must change the volume, there being an increase in volume of Ω, where Ω is the volume per lattice site, for a vacancy and similar decrease in volume per interstitial. This is modified by the lattice relaxation around the defect however, which will be observable as a change in lattice parameter a. By a combination of lattice parameter and volume measurements it is possible to derive not only the defect concentration and the relaxation per defect, but also the state of aggregation of the defects, since Δa decreases rapidly as the defects cluster. These measurements are, of course, much easier to interpret when only one species of point defect is present.

Another physical property which is affected is the electrical resistivity, due to scattering of the conduction electrons. Calculations which take into account the wave nature of the electrons show that the resistivity of vacancies in copper is $(1 \div 2) \cdot 10^{-6}$ Ω cm per 1 % defects ([1,2]). The resistivity increase due to interstitials is expected to be higher than that for vacancies, due to the greater strain energy of interstitials, but the situation is complicated by the possibility of several stable configurations for the interstitial.

Another parameter of considerable importance is the stored energy associated with the defects. If we assume that the total energy in an interstitial-vacancy pair is ~ 5 eV then with a defect concentration of 10^{-5} the stored energy is ~ 0.03 cal g^{-1}, and such a value is typical of a metal irradiated to a low dose at ~ 4 °K. In a material like graphite, however, where due to the high vacancy migration energy and the layer structure the defects do not readily recombine during irradiation at room temperature the stored energy due to the defects can be very large (100 cal g^{-1}) when the irradiation

is carried out at low temperatures (< 100 °C). A large part of this energy is released on heating to only 200 °C. It can be seen that this situation is potentially dangerous, since once the rate of energy release exceeds the specific heat in a large mass of graphite, such as the moderator block of a reactor, the temperature can no longer be controlled, and may « run away » to very high values. This was the source of much trouble in the early days of nuclear reactors, when the operating temperatures were low and the reactors were air-cooled. Periodic annealing of the graphite was necessary to release the stored energy before it became too large, and such operations were somewhat hazardous in the large, poorly instrumented reactors of those days. In fact during one such operation at Windscale irreparable damage to the reactor occurred due to overheating.

2·2.2. *Vacancies.* – The formation energy of vacancies can be estimated theoretically in various ways. For example in covalent crystals the removal of an atom requires the breaking of a given number of bonds, and the energy required is hence an estimate of the formation energy. Another method is to treat the vacancy as a cavity, and calculate the difference in surface energy of small volumes containing equal numbers of atoms *a*) with a vacancy and *b*) without. This method gives an E_f^v of about 2 eV, which is an overestimate since the inward relaxation of the lattice around the vacancy has been neglected. When this is included an E_f^v of $\sim$ 1 eV is found. This result is closer to the rigid cavity value than might be expected, because the inward relaxation introduces strain in the crystal, the energy of which must be included. An important result of these models is that the energy of two adjacent vacancies is considerably less ($\sim 80\,\%$) of twice E_f^v. Thus there will be a fairly strong binding energy between vacancies, and this energy rapidly increases as the size of the cavity increases, so that vacancies will show off strong tendency to cluster. The most accurate method of calculating E_f^v is of course an atomic model, rather than a continuum model, and the calculation is made by allowing an assembly of atoms containing a vacancy to relax according to an interatomic force law. This method can be applied analytically only to small numbers of atoms and for the most accurate results numerical computation methods are required. Values of E_f^v have been obtained for a large number of metals (for example Fumi [3] gives E_f^v as: Li, 0.55 eV; Na, 0.53 eV; K, 0.36 eV; Rb, 0.31 eV; Cs, 0.26 eV; and Tewordt [4] finds Cu, 0.9 eV). Values for the common *fcc* metals are given in Table VII.

The migration energy can also be calculated from the atomic model by comparing the energy of the vacancy at a lattice site and at the saddle point

between two sites. Many such calculations have been made, with results of about 1 eV. The divacancy migration energy has also been calculated, and shown to be lower than E_m^v, (Table VII).

TABLE VII. – *Properties of vacancies in f.c.c. metals.*

		Al	Cu	Ag	Au
E_f^v (eV)	*C*	—	1.2	1.1	1.0
	E	0.75÷0.77	1.1÷1.2	1.08÷1.09	0.94÷0.98
E_f^{2v} (eV)	*C*	1.3	$2E_f^v$÷0.15	2.1	1.85
	E	$2E_f^v$÷0.17	$2E_f^v$÷0.12	1.8	1.86
E_m^v (eV)	*C*	—	1.0	0.86	—
	E	0.63	1.08	0.85	0.83
$E_{sd} - E_f^v$ (eV)		0.72	0.94	0.82	0.85
E_{sd} (eV)		1.48	2.11	1.19	1.81
E_m^{2v} (eV)	*C*	—	0.6	0.52	—
	E	0.46	0.67	0.57	0.67
$\Delta\varrho$ per 1% ($\mu\Omega\cdot$cm)	*C*	—	1.6	1.7	1.7
	E	2.2	—	1.3	1.5

The reader is referred to *Defects and Radiation Damage in Metals*, M. W. THOMPSON (Camb. Univ. Press, 1969) for a full list of references. *C* = calculated values, *E* = experimental values.

A very large amount of work has been carried out in an attempt to determine E_f^v and E_m^v experimentally. A quenching experiments is in principle very simple; the sample being heated to a steady temperature T_q for a few minutes until it contains the equilibrium concentration of vacancies, and then cooled rapidly ($\sim 10^4$ °C s^{-1}) to a temperature T_0 where the vacancies are immobile. The properties of the sample can then be studied either directly or during annealing treatments. For example electrical resistance measurements (at 4 °K to eliminate the thermal resistance) as a function of quenching temperature enables E_f^v to be determined, $E_f^v = -KT_q \log \Delta\varrho_0$, so that E_f^v can be determined from the slope of a plot of $\log \Delta\varrho_0$ against $1/T_q$. Similarly, by measuring the resistance at 4 °K as a function of annealing treatment it is possible to calculate E_m^v. Although in principle these experiments are simple, there are considerable practical difficulties, mainly due to the

difficulty of obtaining high quenching rates (into liquids) without chemical attack. To quench a wire rapidly a liquid with a high latent heat of vaporization is required (water is very efficient, but attacks many specimens at high temperatures). The classic results are those of Bauerle and Koehler [5] on gold, and they found E_f^v to be 0.96 eV and E_m^v 0.8 eV. More results are listed in Tables VII and VIII.

TABLE VIII. – *Vacancies in various materials.*

	E_f^v (eV)	E_m^v (eV)
Graphite	3.5	3.2
Diamond	4.2	2.0
Silicon	2.0	1.1
Molybdenum	5.4	—
Tungsten	3.3	1.9
Sodium	~0.5	0.02

Another classic experiment in which the lattice parameter and length change of Al, Cu, Ag and Au were measured at temperature and used to derive E_f^v and ΔS_f are due to Simmons and Balluffi [6,7]. In the absence of vacancies $\Delta l/l = \Delta a/a$ as the temperature is changed, where l is the length and a the lattice parameter. When vacancies are present $\Delta l/l$ becomes greater than $\Delta a/a$ by:

$$\frac{1}{3} C_v = \frac{\Delta l}{l} - \frac{\Delta a}{a} .$$

An advantage of this method is that C_v is determined absolutely, whereas in the resistance experiments an absolute determination requires a knowledge of the resistivity of vacancies. Hence, in addition to E_f^v Simmons and Balluffi were also able to determine the entropy ΔS_f^v, which they found to be between K and $2K$, so that the entropy factor, $\exp[\Delta S_f^v/K]$, is between 3 and 8.

A summary of the properties of vacancies and divacancies is given in Tables VII and VIII.

2·2.3. *Interstitial atom.* – The calculation of the energy E_f^i is more difficult in the case of an interstitial than a vacancy because the energy is dominated by the strain energy (for a vacancy the strain energy is only ~0.1 eV). This large strain energy results in the possibility of several configurations in *fcc* metals with roughly the same energy. These are:

i) The body centred interstitial, where the extra atom occupies the largest open volume in the unit cell, the neighbouring atoms relaxing outwards in all directions.

ii) The dumb-bell interstitial, where two atoms share a lattice site, the axis of the pair lying along a ⟨100⟩ direction, and the relaxation occurring mainly along this axis.

iii) The crowdion, in which the extra atom is accommodated over several interatomic distances along a ⟨110⟩ direction, there being almost no relaxation in any other direction.

The energies of all these configurations in copper are calculated by the atomic model to be about 4 eV per atom, with the crowdion being slightly greater (Table IX).

TABLE IX. – *Calculated properties of an interstitial in copper.*

	Body-centred	Dumb-bell	Crowdion
E_f^i (eV)	4.0 ± 0.05	3.9 ± 0.05	4.7 ± 0.1
E_m^i (eV)	0.05	0.05	0.25

The migration energies of the three types of interstitial configuration have been calculated, and in the body centred and dumb-bell cases are found to be very low (~ 0.05 eV, ([8,9])). The crowdion migration energy is considerably higher (0.25 eV, ([9])).

Because of the considerable difficulties experienced in calculating the interstitial migration energies recourse must be done to experiment, and since interstitials have such a high energy, radiation damage is the only practical way of introducing interstitials. Since the calculated migration energy is so low, however, the irradiations must be carried out at a very low temperature in order to preserve isolated interstitials. In copper this temperature is below 20 °K, and there are considerable technical difficulties to be overcome before this can be achieved. An example of the type of apparatus required to carry out such irradiations is the cryostat of Sosin and Neely ([10]), (Fig. 1). The essential principles of the apparatus are:

1) The area of the electron beam is restricted to just that required, and the beam tube is cooled to liquid nitrogen temperature.

2) A mechanical valve enables the supply of liquid helium to the cooling block to be controlled. For annealing experiments the valve is closed and the cooling block is heated electrically. When the helium in the tube has evaporated the specimen is thermally isolated from the helium reservoir. Opening the valve and switching off the heater enables the specimen to be rapidly cooled to 4.2 °K for resistance measurements after an anneal.

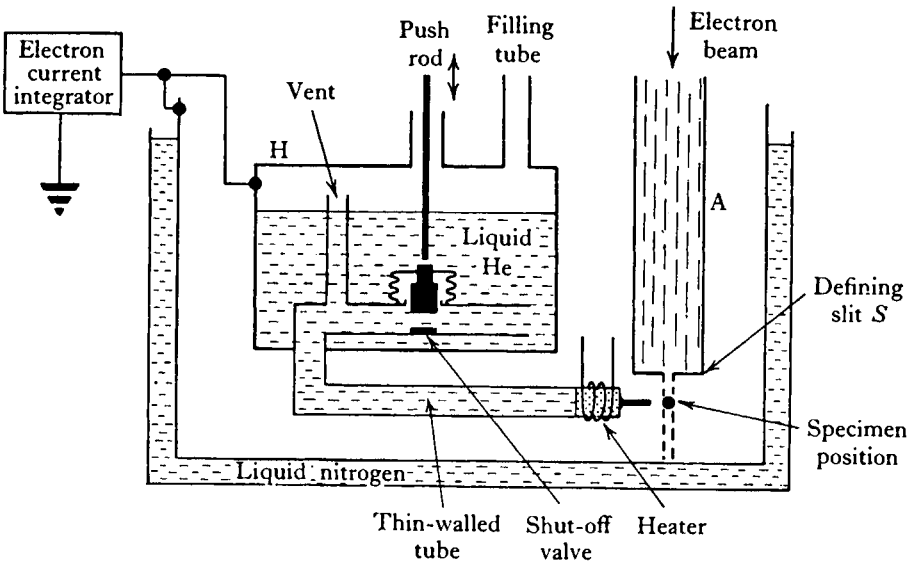


Fig. 1. – Schematic view of liquid helium electron irradiation cryostat. (Courtesy of *Cambridge University Press.*)

Such equipment enables the measurement of the displacement energy, the resistivity of interstitial-vacancy pairs, and the temperatures and magnitudes of the recovery stages which occur on annealing. It was soon established that the recovery spectrum in copper after irradiation was complex, occurring in five main stages, labelled I÷V, Fig. 2. After electron irradiation stages I and III are most important, accounting for ~ 90 and $\sim 10\%$ recovery of the resistivity respectively. It is tempting to attribute I to interstitial migration and III to vacancy migration. Detailed examination, however, shows that stage I consists of five substages, I*a*÷I*e* ([11]), whose properties are summarised in Table X. The observation of first order kinetics for I*a*, *b*, *c* and only one jump for annealing, strongly suggested that these are close interstitial-

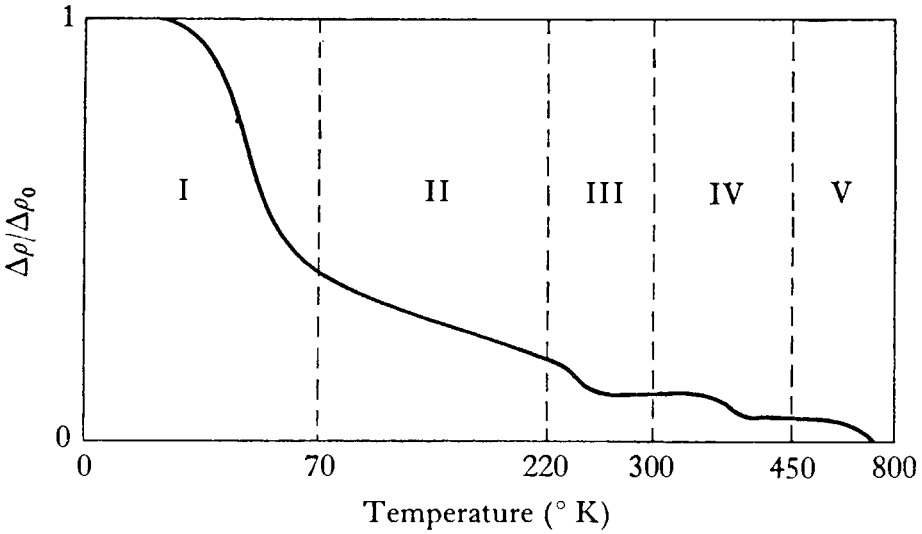


Fig. 2. – Schematic isochronal annealing curve of resistivity in copper. (Courtesy of *Cambridge University Press.*)

TABLE X. – *Stage I substages.*

	I *a*	I *b*	I *c*	I *d*	I *e*
Approximate temperature (°K)	16	28	32	39	53
Activation energy (eV)	0.050	0.085	0.095	0.12	0.12
Order of reaction	1	1	1	2÷3	2
Number of jumps	1	1	1	10	10^4

vacancy pairs. Stage I*d* is consistent with the migration of an interstitial back to its own vacancy (correlated recombination) and I*e* is consistent with long range interstitial migration. The second order kinetics of I*e* is expected since the concentration of interstitials and vacancies are equal at all times. Stage II is highly impurity dependent ([12]), and on detailed examination is found to consist of many separate peaks, each of which is dependent on the concentration of an impurity. It is tempting to attribute Stage III to vacancy migration, since the interstitials responsible for I*e* will migrate to fixed sinks, as well as vacancies, so leaving an excess of vacancies. The energy of Stage III, however, (0.6 eV) is rather low. An alternative scheme originally suggested

by Seeger ([13]) postulates two kinds of interstitial, one of which migrates in I*e* and the other in III (the crowdion and the dumb-bell). Stage IV then becomes vacancy migration.

It has proved extraordinarily difficult to decide between these two interpretations. Many so called « critical » experiments have been devised and performed, only to be promptly explained by the proponents of the opposite view. The weight of evidence, however, now seems to be on the side of the first interpretation, but considerable difficulties remain in explaining the low temperature of Stage III, and also the variation in this temperature when different bombarding particles are used. Measurements of stored energy, length or lattice parameter have not resolved the controversy.

During irradiations with heavier particles it has been found ([14]) that the observed number of defects is much less than the calculated values, and this discrepancy increased as the recoil energy increases. Neutron irradiations suffer from particular difficulties, since not only are low temperature experiments particularly difficult to carry out, but also a very wide range of primary recoil energies is present, from a few tens of eV for the (nγ) reactions up to $\sim 10^5$ eV from 10 MeV fission neutrons. This makes the results much more difficult to interpret. For example, after neutron irradiation at < 10 °K the first three substages in Stage I merge into a continuum which extends down to 7 °K ([15]). If the fast neutrons are filtered out, however, the proportion of recovery in Stage I increases and the substages emerge. The interpretation of this is that there is substantial lattice strain in a cascade which affects the migration energies of particular configurations of close-pairs. The reduction in Stage I annealing after neutron irradiation is probably due to the athermal or thermal spike clustering of both interstitials and vacancies in the cascade volume, so resulting in a smaller proportion of isolated defects and unannealed close-pairs.

2·3. **The formation of clusters during irradiation.**

2·3.1. *Introduction.* – Since electron microscopes have not yet advanced to the stage where individual point defects can be observed, it is in the study of clustered defects that the technique has had most impact. In fact this has not been much of a disadvantage, since almost all the macroscopically interesting irradiation effects are due to clustered, rather than point, defects. Also, the majority of the important irradiation effects are produced during relatively high temperature irradiations.

When the irradiation is carried out at a temperature at which the point defects are mobile there are several ways in which clusters may be nucleated. The first of these is by nucleation within a cascade. When the primary recoil energy is high we have seen that the damage is formed in cascades, within which there is a central zone of high vacancy concentration, and an outer and more diffuse zone of high interstitial concentration. Thermal migration of the defects within these zones may nucleate clusters. In this type of nucleation the density of nuclei is, of course, proportional to the dose φt.

Secondly there may be imperfections in the crystal, such as impurity atoms or precipitates, which have a high binding energy for the migrating defects. Hence, once a defect has become associated with such an imperfection, it cannot escape and if there is any excess of diffusion of one type of defect to the imperfection a cluster will form.

Thirdly it is possible for clusters to nucleate homogeneously by the chance of meeting two or more migrating point defects. In the absence of preferential absorption, this process can only occur when there is excess diffusion of one type of point defect, as otherwise there is an excellent chance that any nuclei so formed will disappear as the result of the absorption of one or more defects of the opposite sign. This type of nucleation does not therefore occur during « steady-state » conditions, which are defined as $\gamma_i i = \gamma_v v$, where γ_i and γ_v are the diffusion rates of interstitials and vacancies, and i and v are the defect concentrations. In many practical cases the conditions are non steady state, however, either because of the wide disparity between the diffusion rates of interstitials and vacancies, or because a large proportion of the vacancies cluster within the cascades, so leaving a permanent excess of interstitial diffusion. It is a characteristic of this process that it nearly always results in the formation of interstitial clusters.

2'3.2. *The morphology and energy of defect clusters.* – It is theoretically possible for defects to cluster into several different configurations: these will be briefly discussed and calculations of the relative energies described.

For vacancies the simplest possible configuration is the void, or spherical cluster. The energy of this can be readily calculated when it is large from the surface energy:

$$E_v = 4\pi r^2 \gamma = N^{\frac{2}{3}} a^2 \left(\frac{9\pi}{4}\right)^{\frac{1}{3}} \gamma ,$$

where N is the number of vacancies and γ the surface energy. This simple relation breaks down when the void becomes small because of the effect of the curvature on the surface energy.

Another configuration is the dislocation loop, which is basically a disc of vacancies in which the centre has collapsed to remove the exposed surfaces. Loops can exist in several forms in *fcc* metals, for example a Frank loop is formed when the defects condense on a {111} plane and collapse occurs in the ⟨111⟩ direction perpendicular to the loop plane, and hence a stacking fault is formed since the normal ABC ABC stacking becomes of the type ABABC. The equilibrium shape of a Frank loop is hexagonal, with sides along ⟨110⟩ directions ([16]), although other shapes also occur in practice. To a first approximation, the energy of a Frank loop is:

$$E_F = \frac{\mu b^2 r}{3(1-\sigma)}\left\{\ln\frac{r}{b}+\frac{5}{3}\right\}+\pi r^2\gamma',$$

where μ is the shear modulus, σ is Poisson ratio, b is the Burgers vector and γ' is the stacking fault energy. A more accurate expression has been given by Sigler and Kuhlmann-Wilsdorf ([17]).

The stacking fault in a Frank loop can be removed by a shear of vector $\frac{1}{6}a[112]$ across the loop plane to form a perfect prismatic loop. The Burgers vector of the resulting loop is $\frac{1}{2}a[110]$ by the reaction:

$$\frac{1}{3}a[111]+\frac{1}{6}a[11\bar{2}]\rightarrow\frac{1}{2}a[110].$$

The energy of such a loop is approximately:

$$E_p = \frac{\mu b^2 r}{2(1-\sigma)}\left\{\ln\frac{r}{b}+\frac{5}{3}\right\},$$

and a more accurate value has again been given by ([17]).

Another configuration for vacancy clusters is the stacking fault tetrahedron. This consists of four equal triangular intrinsic stacking faults lying on four intersecting {111} planes so that they form a tetrahedron, edged by stair-rod dislocations with a $\frac{1}{6}a\langle 110\rangle$ Burgers vector. Several mechanisms have been proposed whereby stacking fault tetrahedra can be formed. For example, the dislocations bounding a triangular Frank loop may dissociate into stair-rod dislocations and Shockley partials, which then glide up the three intersecting {111} planes to complete the tetrahedron ([18]). Alternatively a small cluster, probably a tetravacancy, may grow by the nucleation and migration of jog lines over the tetrahedron faces ([19]). Jøssang and Hirth ([20])

have estimated the energy to be:

$$E_T = \frac{\mu b^2 l}{6\pi(1-\sigma)}\left\{\ln\frac{4l}{6} + B_2\right\} + \sqrt{3}l^2\gamma',$$

where $B_2 = 1.017 + 0.972\nu$ and l is the length of the edges.

Although interstitials can theoreticallly form equivalent clusters to all the above, except the spherical shape (because the strain energy of interstitials is so large) in practice only perfect prismatic loops and, very rarely, stacking fault loops, have in fact been observed. The absence of the interstitial stacking fault tetrahedron is probably because of the higher extrinsic than intrinsic stacking fault energy. The common configuration for interstitial clusters is the perfect prismatic loop. This can be formed from a stacking fault loop by the passage of two partial dislocations across the loop, one above and one below the extra plane:

$$\frac{1}{3}a[111] + \frac{1}{6}a[2\bar{1}\bar{1}] + \frac{1}{6}a[\bar{1}2\bar{1}] \rightarrow \frac{1}{2}a[110].$$

Perfect prismatic loops (Burgers vector $\frac{1}{2}a[110]$) are not often found lying on $\{111\}$ planes. This may be because either they nucleate directly on $\{110\}$, or because they can lower their energy by rotating from $\{111\}$ to $\{110\}$ so as to be perpendicular to their Burgers vector. The latter mechanism is suggested by the frequent observation of four sided rhombus loops ([21]) lying on or near $\{012\}$ planes. These can form from hexagonal $\{111\}$ loops by the redistribution of defects so as to eliminate the two pure edge sides leaving a diamond shape. The dislocations forming the diamond lie on a slip cylinder composed of four $\{111\}$ planes, all of which contain the loop Burgers vector. The loop can then rotate so as to minimise its energy. Bullough and Foreman ([22]) have calculated that they can reduce their elastic energy by rotating to approximately $\{012\}$.

The results of the calculations ([17,23]) indicate that although many of the observed features can be explained, for example the tendency for voids to form in aluminium, and stacking fault tetrahedra in gold, there are many anomalies. In particular it would appear that clusters can readily exist and grow in metastable states, presumably due to the difficulty of initiating the transformation.

2·3.3. *Experimental results.* – There is now a vast literature on the clusters produced by irradiating different materials with different particles, and instead

of attempting to review it all I shall discuss only selected experiments to demonstrate the main effects before describing how the defects affect the macroscopic properties.

The basic facts and their interpretation were established in the early work at Harwell ([24-27]), and it is interesting to see how the subsequent very large effort devoted to this subject has confirmed the original interpretation.

Following the first observation of defect clusters in neutron irradiated copper by Silcox and Hirsch ([18]), who interpreted them as vacancy in character, a detailed investigation was made by Makin, Whapham and Minter ([26]) of the size distribution as a function of neutron dose during irradiation at ~ 27 °C. Considerable care was taken to determine the foil thickness of the area being photographed, and it was found that this was essential for accurate results, since as the density of defects increased the eye automatically selected thinner and thinner areas for photographing. The results are shown in Fig. 3 and 4. The density of the smallest defects < 50 Å was virtually linear with dose, whereas an approach to saturation was apparent in the larger

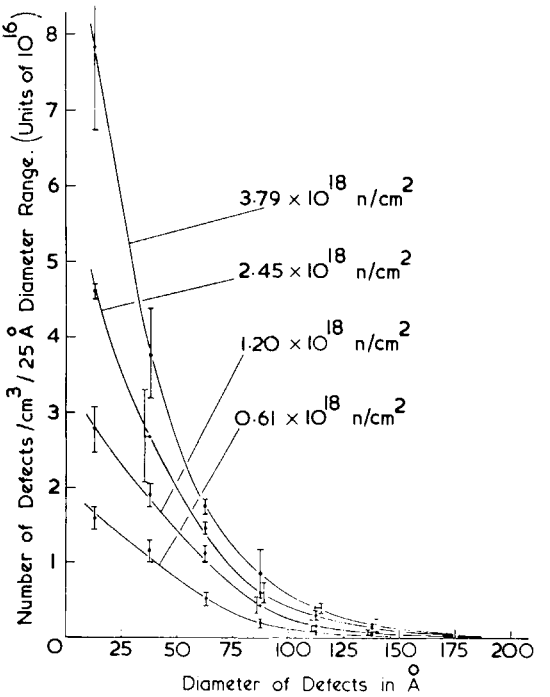


Fig. 3. – The density of defect clusters of various sizes in neutron irradiated copper. (Courtesy of *Phil. Mag.*)

(> 50 Å) defects. Further evidence which suggested that two types of defect were present was the clear grain boundary zone ~ 1000 Å wide observed for the > 50 Å defects, whereas no such zone existed for the < 50 Å defects.

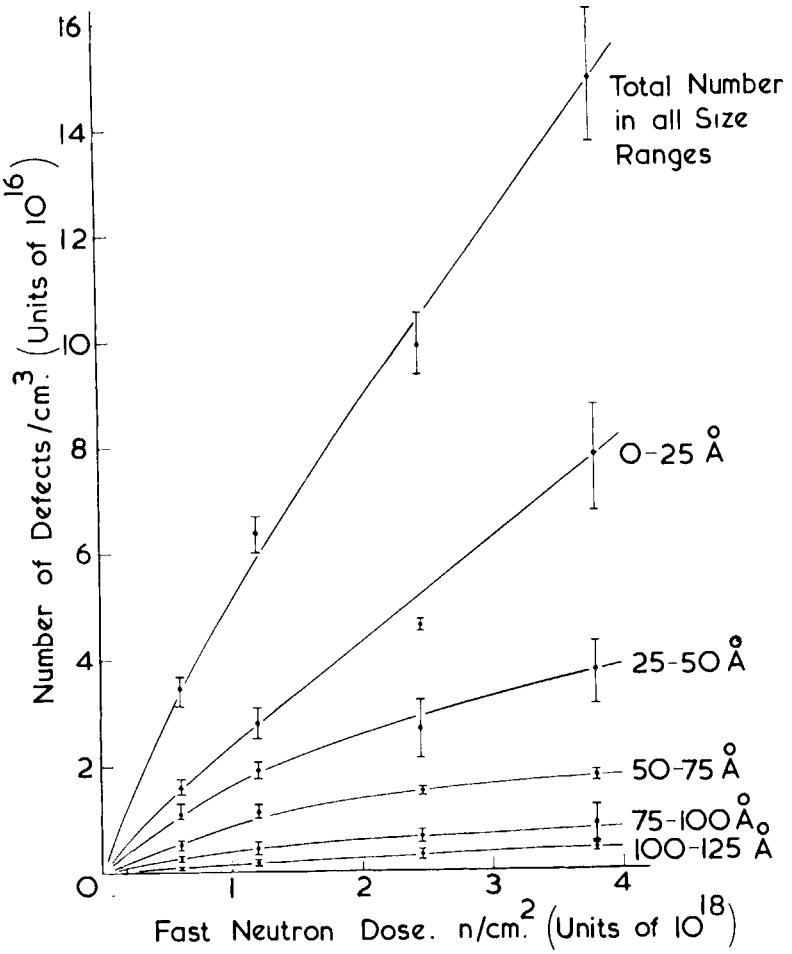


Fig. 4. – The dose dependence of the density of clusters in neutron irradiated copper. (Courtesy of *Phil. Mag.*)

Short anneals (*i.e.* 306 °C for less than 300 min) were found to enhance the distinction between the two types of cluster (Fig. 5). It was clear from the results that the density of small clusters < 50 Å was approximately proportional to dose and hence it was likely that they were nucleated directly in the

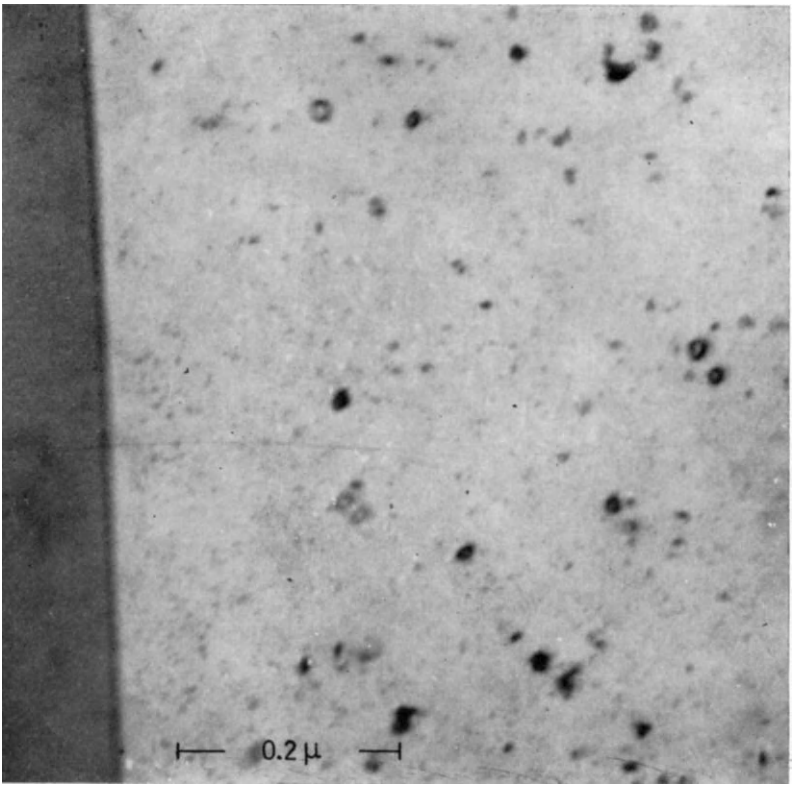


Fig. 5. – The two types of defect cluster and their grain boundary denuded zones in copper irradiated by $2 \cdot 10^{18}$ fast neutrons cm^{-2} and annealed at 306 °C for 63 min.

primary knock-on cascades and hence should be vacancy in character. If so, then one cluster was nucleated for ~ 10 such cascades. The behaviour of the large clusters, moreover, was consistent with that expected from a much more mobile defect, such as the interstitial. It is *impossible* for the large clusters to be formed by vacancy diffusion, since the jump rate of vacancies at the irradiation temperature is much too low (~ 1 per s). This interpretation was consistent with quantitative estimates of the relative numbers of vacancies and interstitials (~ 50 per primary knock-on), from which it was further deduced that the small vacancy clusters were essentially two-dimensional, *i.e.* loops or tetrahedra, rather than voids. The behaviour on annealing ([27]) also supported the interpretation, the annealing occurring in stages, the first of which has a low activation energy, $(1 \div 1.5)$ eV, during

which $\sim 70\%$ of the interstitials present in the loops > 50 Å in size disappear. The second stage of recovery has an activation energy of ~ 2 eV and corresponds to the disappearance of the small clusters and the remaining interstitial loops. The interpretation of the sense of the clusters was supported by experiments on alpha irradiated copper ([24]), in which it was found that during annealing the large loops grew, rather than disappeared, simultaneously with the appearance of helium bubbles. When helium atoms cluster they become centres of very large compressive strain, which can only be relieved by the absorption of vacancies. In this experiment therefore, not only were the vacancies being emitted from the vacancy clusters going to the gas bubbles but the vacancy concentration was so reduced that the interstitial loops were emitting vacancies and hence growing in size.

The formation of the small clusters in copper and gold has been extensively studied by Merkle ([28]) using protons, deuterons, alpha particles and fission fragments, of various energies. The experiments showed that in copper only about 10 % of the cascades formed from primary recoils with an energy of greater than $\sim 10^4$ eV produced a visible cluster, and that the size of the cluster did not increase with recoil energy. It is likely, therefore, that the visible clusters are the result of the overlapping of sub-cascades. The behaviour of gold is somewhat simpler, in that a visible cluster was formed by every recoil with an energy of greater than $\sim 3 \cdot 10^4$ eV, and the cluster increases in size with recoil energy up to a maximum of ~ 150 Å. Very energetic recoils from fission fragments produce groups of clusters, in which each subcascade produces a visible cluster. The difference between gold and copper is due to the greater « density » of the cascade in the heavier element.

Analysis of the nature of the small clusters from their contrast has utilised the black-white contrast effect which is observed under dynamical two-beam conditions. Detailed work using stereo-microscopy showed that only these defects close to the foil surfaces showed black-white contrast ([29]). Examination of the clusters showing black-white contrast in several reflections showed that the symmetry line was independent of the g vector, and nearly always lay along the projection of a $\langle 111 \rangle$ direction, thus showing that the clusters were Frank sessile loops ([30]). Calculations of the contrast expected from such loops ([29]) indicated that the direction of the black-white contrast depended not only on the sense of the defect but also on its depth in the foil, reversals of direction occurring at $\sim \xi_g/4$ and $3\xi_g/4$ (Fig. 6). Hence to determine the sense it is essential to know the position of the defect quite accurately. Unfortunately determinations of the sense by different workers did not always give the same result. Wilkens and Rühle ([31]) found that they were vacancy,

as did Merkle ([28]), who deduced the depth by varying the energy of the incident ion. McIntyre and Brown ([32]) however concluded that the small clusters in neutron irradiated copper were predominantly interstitial in nature.

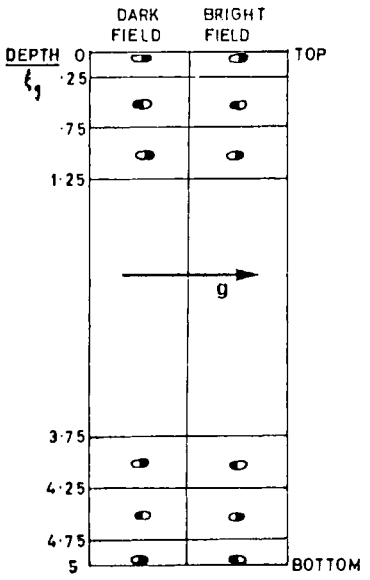


Fig. 6. – Schematic diagram of black-white image contrast from small interstitial defects as a function of depth in the foil.

The experiments is of course a difficult one to carry out, since it involves the measurements of the distance between the defect and the surface (decorated by gold islands) to an accuracy of ~ 20 Å. (See also Brown, this volume.)

The effect of irradiation temperature on the nature of the damage is also very informative on the nature of the clusters. For example, since it is generally believed that vacancies are not mobile below room temperature in copper it is clear that loops which show a gradual change in density as the irradiation temperature is reduced below room temperature *cannot* be vacancy loops. Such an effect is found for the large clusters observed during room temperature neutron irradiations in copper ([33]), hence confirming their interstitial nature. Similarly the observation of the small vacancy clusters during low temperature irradiations ([34]) confirms that these are not produced by migration, but are formed directly in the collision cascades. Although the effect of irradiation temperature on the formation of the interstitial loops strongly suggests that the interstitial is mobile at very low temperatures, there are

some experiments which were designed to test this but apparently give a conflicting result. For example Venables and Balluffi ([35]) bombarded gold containing tetrahedra formed by quenching with 200 eV A^+ ions at 140 °K. These ions displace atoms within ~ 100 Å of the surface, and loops were observed in this layer. Tetrahedra deeper in the foil, however, showed no effect until the foil was annealed at ~ 0 °C. This experiment, however, does not prove that interstitials migrate at 0 °C in gold, since the arrangement of the sinks (*i.e.* the surface and the loops) for the mobile interstitials produced during irradiation was such that very few interstitials would have been expected to reach the tetrahedra during the irradiation. Some further important clustering effects will now be described before passing to the interpretation of the macroscopic irradiation effects in terms of clusters.

2·3.4. *Effect of knock-on energy.* – The first of these concerns the energy of the primary recoil required to produce a visible cluster in different materials. We have seen how in gold a visible cluster is produced per knock-on above a critical energy, and in copper how only one knock-on in ten produces a visible cluster. In aluminium the situation is quite different since not only is the cascade very diffuse for low and medium energy knock-ons, but also vacancies are readily mobile at room temperature. Hence in aluminium no visible damage is formed during neutron irradiation at room temperature, because the density of vacancies within the cascades is insufficient to produce clusters and also the dose rate is not high enough for interstitial clusters to nucleate during the rather short non-steady state stage in this metal at room temperature. Clusters can be produced by very high dose rate experiments ([36]) and as expected, the density is very sensitive to the dose rate. Also, if a gaseous impurity is injected, such as helium, then loops are also observed ([37]). In this case aggregates of helium atoms act as a strong sink for vacancies and so leave a surplus of interstitial atoms, which form loops. It is possible to produce the same type of damage in aluminium at room temperature as in neutron irradiated copper, but it requires very energetic primary knock-ons, such as are produced during fission fragment irradiation ([38]). Loops of both types form and apparently grow despite the presence of loops of the opposite sign. The loops are clearly very mobile, however, and frequently slip or climb together to produce both larger loops and mutual annihilation. The stress field around loops is substantial ([39]) and many cases of loop movement due to mutual interaction forces have been observed.

2·3.5. *Effect of purity.* – There is considerable evidence that the purity of the metal has a marked effect on the distribution of the interstitial clusters,

a coarser distribution of larger loops being found in purer materials. This behaviour has been observed in copper ([40]) molybdenum ([41]) and graphite ([42]). The effect of boron in graphite has been interpreted by Brown, Kelly and Mayer ([43]) in terms of a chemical reaction rate theory in which the effect of the trapping of interstitials on boron atoms is to reduce the average velocity of the interstitials. The theory predicts that the loop density is proportional to the square root of the impurity concentration, and that the loop radius is proportional to $1/C^{\frac{1}{2}}$, where C is the impurity concentration. Both of these predictions are in good accord with the experimental results. In such a model, increasing the impurity concentration is akin to reducing the temperature. The results obtained on molybdenum ([41]) show one further effect, in that in high purity molybdenum it is possible to obtain distributions of predominantly vacancy loops by annealing at 900 °C after irradiations at 200 °C, whereas after irradiations at 60 °C only interstitial loops are found. Presumably the excess density of vacancies arises by the diffusion of a much larger number of interstitials than vacancies to sinks during the irradiation. At 200 °C the interstitials will be much more mobile than at 60 °C and hence there will be a greater zone denuded of interstitials around sinks.

Other effects of inert gas impurity atoms will be discussed later.

REFERENCES (Section 2)

1) F. J. BLATT: *Solid State Physics*, **4**, 322 (1957).
2) A. SEEGER and D. SCHUMAKER: *Lattice Defects in Quenched Metals*, Academic Press (1965), p. 15.
3) F. G. FUMI: *Phil. Mag.*, **46**, 1007 (1955).
4) L. TEWORDT: *Phys. Rev.*, **109**, 61 (1958).
5) J. E. BAUERLE and J. S. KOEHLER: *Phys. Rev.*, **107**, 1493 (1957).
6) R. O. SIMMONS and R. W. BALLUFFI: *Phys. Rev.*, **125**, 862 (1962).
7) R. O. SIMMONS and R. W. BALLUFFI: *Bull. Amer. Phys. Soc.*, **7**, 233 (1962).
8) H. B. HUNTINGTON and F. SEITZ: *Phys. Rev.*, **61**, 324 (1942).
9) R. A. JOHNSON and E. BROWN: *Phys. Rev.*, **127**, 446 (1962).
10) A. SOSIN and H. H. NEELY: *Rev. Sci. Instr.*, **32**, 922 (1961).
11) R. M. WALKER: *Radiation Damage in Solids*, Academic Press (1962), p. 594.
12) D. G. MARTIN: *Phil. Mag.*, **6**, 67, 839 (1961).
13) A. SEEGER: *Proc. 2nd Geneva Conf. on Peaceful Uses of Atomic Energy.*, **6**, 250 (1958).
14) H. G. COOPER, J. S. KOEHLER and J. W. MARX: *Phys. Rev.*, **97**, 599 (1955).
15) R. R. COLTMAN, C. E. KLABUNDE, D. L. MCDONALD and J. K. REDMAN: *Journ. Appl. Phys.*, **33**, 3509 (1962).

16) D. KUHLMANN-WILSDORF: *Phil. Mag.*, **3**, 125 (1958).
17) J. A. SIGLER and D. KUHLMANN-WILSDORF: *The Nature of Small Defect Clusters*, H.M.S.O. (1966), p. 125.
18) J. SILCOX and P. B. HIRSCH: *Phil. Mag.*, **4**, 72 (1959).
19) M. DE JONG and J. S. KOEHLER: *Phys. Rev.*, **129**, 49 (1963).
20) T. JØSSANG and J. P. HIRTH: *Phil. Mag.*, **13**, 657 (1966).
21) M. J. MAKIN and B. HUDSON: *Phil. Mag.*, **8**, 447 (1963).
22) R. BULLOUGH and A. J. E. FOREMAN: *Phil. Mag.*, **9**, 315 (1964).
23) R. M. J. COTTERILL: *The Nature of Small Defect Clusters*, H.M.S.O. (1966), p. 144.
24) R. S. BARNES and D. J. MAZEY: *Disc. Faraday Society*, **31**, 38 (1961).
25) M. J. MAKIN, A. D. WHAPHAM and F. J. MINTER: *Phil. Mag.*, **6**, 465 (1961).
26) M. J. MAKIN, A. D. WHAPHAM and F. J. MINTER: *Phil. Mag.*, **7**, 285 (1962).
27) M. J. MAKIN and S. A. MANTHORPE: *Phil. Mag.*, **8**, 1725 (1963).
28) K. L. MERKLE: *The Nature of Small Defect Clusters*, H.M.S.O. (1966), p. 8.
29) M. RÜHLE, M. WILKENS and U. ESSMANN: *Phys. Stat. Sol.*, **11**, 819 (1965).
30) U. ESSMANN and M. WILKENS: *Phys. Stat. Sol.*, **4**, K53 (1964).
31) M. WILKENS and M. RÜHLE: *The Nature of Small Defect Clusters*, H.M.S.O. (1966), p. 365.
32) K. G. MCINTYRE and L. M. BROWN: *The Nature of Small Defect Clusters*, H.M.S.O. (1966), p. 351.
33) G. P. SCHEIDLER, M. J. MAKIN, F. J. MINTER and W. F. SCHILLING: *The Nature of Small Defect Clusters*, H.M.S.O. (1966), p. 405.
34) L. M. HOWE, R. W. GILBERT and G. R. PIERCY: *Appl. Phys. Lett.*, **3**, 125 (1963).
35) J. A. VENABLES and R. W. BALLUFFI: *Phil. Mag.*, **11**, 1021, 1039 (1965).
36) C. J. BEEVERS and R. S. NELSON: *Phil. Mag.*, **8**, 1189 (1963).
37) D. J. MAZEY, R. S. BARNES and A. HOWIE: *Phil. Mag.*, **7**, 1861 (1963).
38) K. H. WESTMACOTT, A. C. ROBERTS and R. S. BARNES: *Phil. Mag.*, **7**, 2035 (1962).
39) A. J. E. FOREMAN and J. D. ESHELBY: A.E.R.E., R-4170 (1962).
40) M. J. MAKIN: unpublished work.
41) B. L. EYRE and M. E. DOWNEY: *The Nature of Small Defect Clusters*, H.M.S.O. (1966), p. 384.
42) A. KELLY and R. M. MAYER: *Phil. Mag.*, **19**, 701 (1969).
43) L. M. BROWN, A. KELLY and R. M. MAYER: *Phil. Mag.*, **19**, 721 (1969).

3. The nature of the damage: Technological effects.

3·1. Introduction.

In this Section attention is directed away from the more academic aspects of radiation damage and towards the interpretation of some of the important effects which occur in practice.

3·2. Radiation growth in uranium.

In the early reactors the uranium fuel was incorporated as metal rods, usually prepared by casting. It was soon found that rods with initially smooth machined surfaces became very rough and wrinkled during radiation. The surface roughness was found on about the same scale as the grain size of the metal, and it was clear that some unusual changes were taking place in which the shape of the grains was changing during irradiation at low temperatures ((100÷200) °C). Single crystal irradiations ([1]) showed that the crystals were growing in the [010] direction, shrinking in the [100] direction, and were not changing in the [001] direction. Moreover the effect was very large, the fission of 1% of the atoms producing a final length of 400% of the initial length (along [010]). In the temperature range in which the effect occurs uranium exists in the α-phase with a very anisotropic orthorhombic crystal structure ($a = 2.85$, $b = 5.86$, $c = 4.96$ Å). The rate of growth at constant temperature was found to depend only on the burnup B (fraction of uranium atoms fissioned; $B = \sigma\varphi t$), and the instantaneous length l. Hence:

$$\frac{\Delta l}{l} = G\,\mathrm{d}B \quad \text{and} \quad l = l_0 \exp\left[GBt\right],$$

where G is known as the growth factor. G is a decreasing function of temperature varying between $10^4 \div 10^5$ at 20 °K ([2]) and zero at (400÷500) °C. The effect was found in other anisotropic metals but of much smaller magnitude (G for Cd, Zr, Zn and Ti at 77 °K is 70, 60, 20 and 15 respectively ([3])). It does not occur in any cubic metals, including cubic alloys of uranium, and is almost negligible with particle irradiations in which fission does not occur, even when comparable point defect concentrations are formed ([4]). The effect is therefore associated with very high energy knock-ons, such as occur in fission fragment irradiation. The effect was not recoverable by annealing, and showed no tendency to saturate.

For many years the effect remained a mystery. Several theories were put forward ([5-7]), none of which very satisfactory. With the discovery of the formation of loops in other metals by irradiation, however, it was immediately obvious that growth might result from the accumulation of clusters of point defects in particular orientations, the interstitials forming loops on (010) planes and the vacancies loops on (100) planes. Since the total area of the two types of loop must be equal, since the vacancy and interstitial concentrations must be approximately equal in the bulk of the material,

the expansion and contraction must be equal. If N atoms per fission event per cm^3 condense on (010) planes, and the N vacancies left condense on (100) planes then there are Na^2 extra atomic planes in the [010] direction, where a is the diameter of an atom and hence a length change of

$$\Delta l = Na^3 .$$

Now

$$\Delta l = GB$$

and $B = a^3$ when one atom is fissioned, so that $G = N$. Hence G has the direct significance of being the number of point defects which precipitate per fission event. It should be noted that several rather unusual effects must occur for this model to be correct. Firstly, the interstitial and vacancy loops must precipitate on different planes. Secondly, all the interstitials and vacancies formed subsequently must not diffuse randomly to all the loops, since growth then cease. Thirdly, the loops must be able to coalesce to form complete planes, since otherwise growth would cease when the loop density reached saturation.

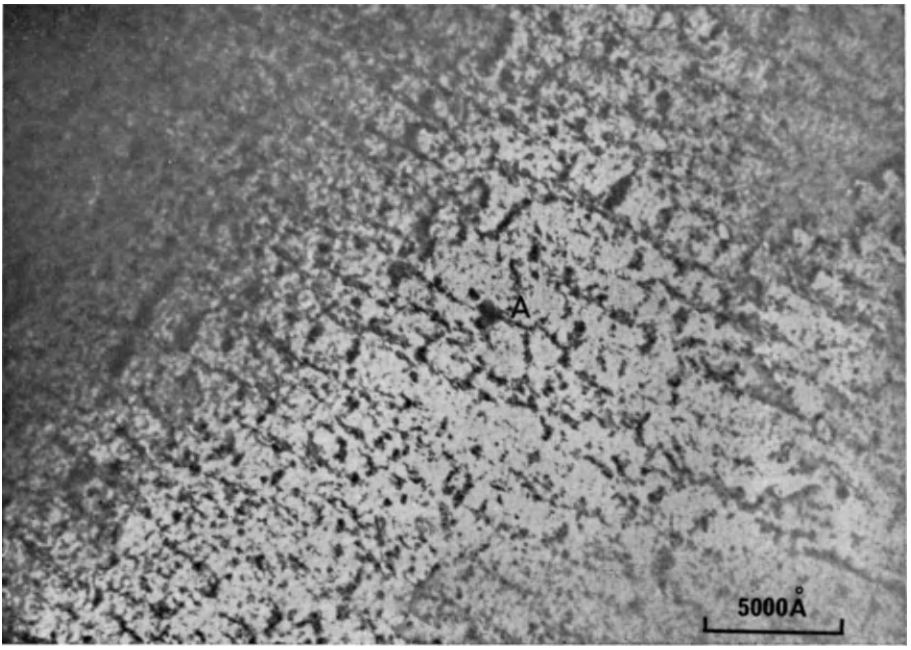


Fig. 7. – Typical brickwork pattern when both types of loop are in contrast in irradiated alpha-uranium, after $1.5 \cdot 10^{17}$ n·cm^{-2} at 80 °C. (Courtesy of *Phil. Mag.*)

Clearly there was a need for transmission microscopy of irradiated uranium, to check whether the loops followed the required arrangement and were of the correct integrated area to explain the phenomenon. Unfortunately the work was held up for a long time by the difficulty of preparing thin films free from surface oxide films (uranium oxidises readily in the atmosphere). This problem was eventually solved by electrolytic cleaning of the electropolished foils in a sulphuric acid-glycerol bath ([8]). It was found ([8,9]) that there were two sets of loops (Fig. 7), one lying on (010) planes with a $a[010]$ Burgers vector, and another on approximately (100) planes with Burgers vectors of $\frac{1}{2}\sqrt{a^2+b^2}\langle 110\rangle$. It was not possible to unambiguously determine the sign of the loops, but if it is assumed that the (010) loops are interstitial and the (100) vacancy, then the growth predicted agrees well with the macroscopically determined values. Furthermore, it was found that at temperatures of 80 °C and above, the loops lay in well defined sheets, separated by regions containing very few loops (Fig. 8), the loop planes being approximately

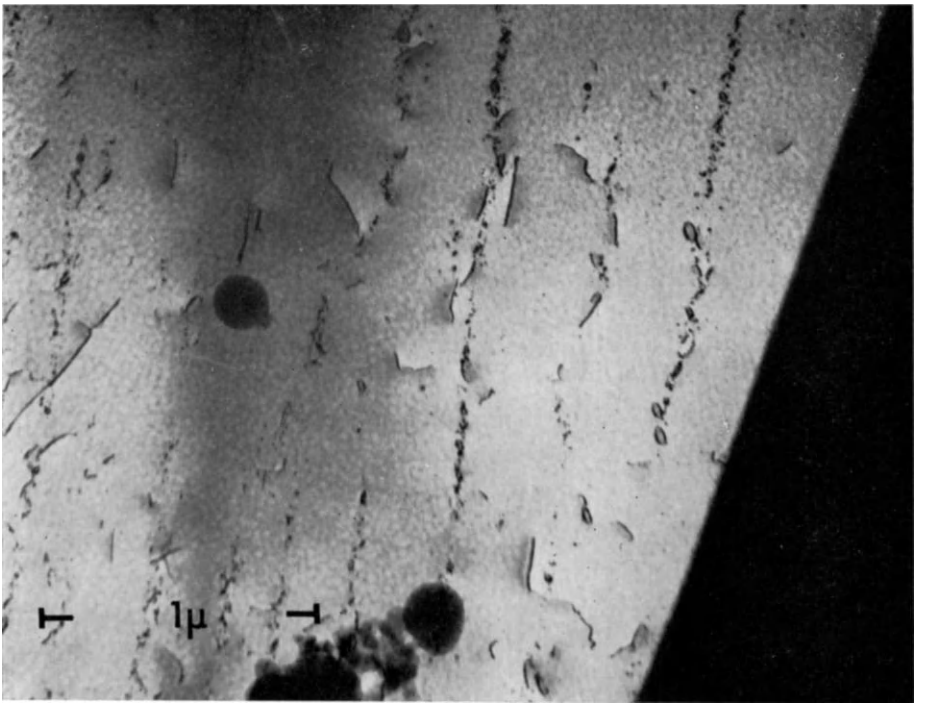


Fig. 8. – Loops aligned in sheets after irradiation at 350 °C to $1.5 \cdot 10^{17}$ n·cm^{-2}. (Courtesy of *Phil. Mag.*)

coplanar with the sheets. The observation that a high density of loops formed during irradiation at — 196 °C indicates that the loops are nucleated by fission fragment spikes, since this temperature is too low for vacancy migration. It was clear that there must be a strong attraction between the sheets of loops and new loops nucleated between the sheets. Hudson ([9]) using the calculations of Foreman and Eshelby ([10]), was able to demostrate that similar loops should be attracted into sheets, particularly when loop movement occurred by slip. At high temperatures, where climb is possible, the separation is less complete and this may be one reason why the growth rate decreases with increasing temperature. A second, and more probable, reason is that the chance of loop nucleation within a cascade increases as the temperature is reduced. At high temperatures, for instance, there may be a strong probability of a nucleus dissociating before it has a chance of growing into a cluster. This is consistent with the fact that radiation growth occurs only with fission fragment irradiation, since it is inferred that only a fission fragment can produce a high enough defect density to form nuclei.

Although many questions remain on the growth of uranium it is clear that the *mechanism* has now been established, and that this is due to the application of electron microscopy to the problem.

This is a good example of the application of the technique, since while the results do not enable anyone to solve any specific reactor problems (these were effectively solved long before the mechanism was known) the element of mystery has been removed and the reactor designer can now discuss the growth problem in a rational way and is able to think logically about probable growth rates in different alloys.

3·3. **Radiation hardening.**

The second example of the application of electron microscopy to technological reactor problems concerns the hardening effect which occurs in nearly all material during irradiation. Technologically the important parameter is not so much the yield stress as the *ductility* of the metal, and this becomes even more important in metals which have a ductile-brittle transition, since irradiation increases the transition temperature.

Considerable attention has been given to irradiation hardening, particularly in copper from the point of view of the basic mechanism, and also in many steels of use in reactors.

Experiments on polycrystalline metals shows that the yield strength and ultimate strength both increase with neutron dose, the yield strength increas-

ing considerably more rapidly however, so that the total amount of work hardening which can occur before fracture is reduced. Hence, although the work hardening rate is smaller after irradiation the total elongation before fracture is reduced. This can cause concern when the dose is high, since when the yield and ultimate strengths become almost equal the metal becomes plastically unstable and if deformation commences in an area it continues locally until fracture occurs, with no increase in stress. It is fortunate that radiation hardening generally anneals at a fairly low temperature, so that with the general increase in reactor temperatures radiation hardening due solely to interstitials and vacancies is less of a problem than hitherto.

Experiments on the mechanism of the hardening have generally concentrated on measurements of the deformation characteristics of copper single crystals, and in particular of the critical shear stress, *i.e.* the stress across the slip plane in the slip direction. The general effect of neutron irradiation on the stress-strain curves of copper single crystals is shown in Fig. 9 ([11]).

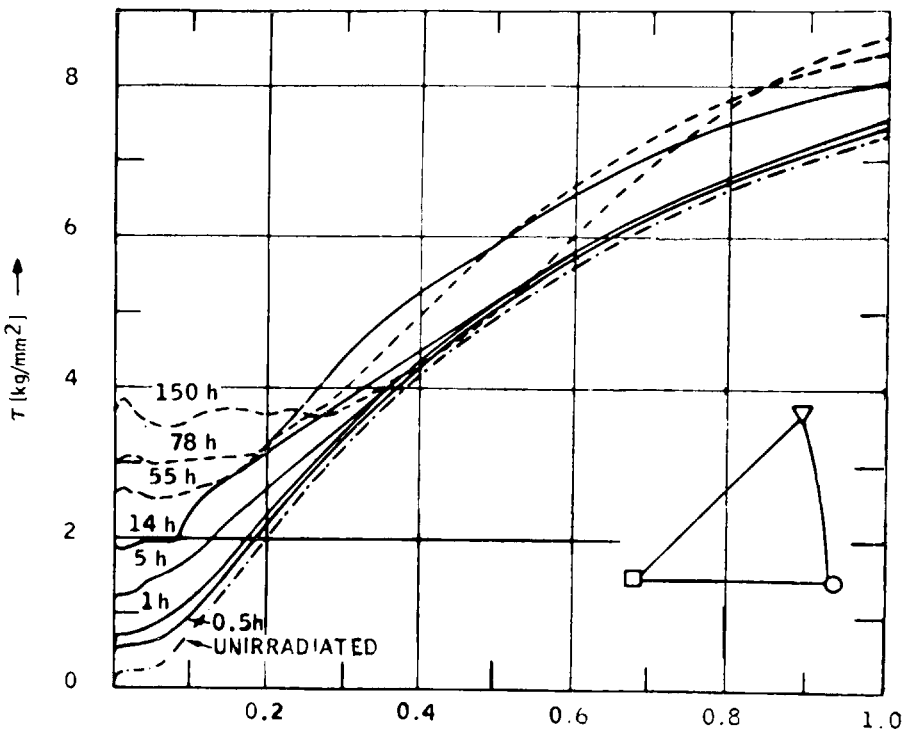


Fig. 9. – The stress-strain curves of copper single crystals as a function of irradiation time. (From J. Diehl, courtesy of *International Atomic Energy Agency*.)

The main effects of irradiation are:

a) An increase in critical shear stress.

b) A yield point.

c) A region of low work hardening following the yield during which deformation spreads along the crystal *i.e.* a Lüders band (Stage I). The extent of Stage I strain increases with dose.

d) A region of normal work hardening (Stage II) the slope of which is not appreciably altered by low neutron doses ($< 10^{19}$ n·cm^{-2}).

The dose dependence of the critical shear stress σ has been the subject of much study and controversy. The early work ([12]) showed that σ was proportional to $\varphi^{\frac{1}{2}}$. This was difficult to understand theoretically, since dislocations move on a plane a dependence on $\varphi^{\frac{1}{2}}$ is expected, assuming that the number of obstacles is proportional to the neutron dose. The $\varphi^{\frac{1}{2}}$ relation was later supported by various other workers using internal friction breakaway measurements and microstraining techniques ([13-15]). On the other hand it was suggested by Makin and Minter ([16]) that the correct expression may be $\varphi^{\frac{1}{2}}$ plus exponential saturation term to account for the saturation in the cluster density:

$$\sigma = A[1 - \exp[-B\varphi]]^{\frac{1}{2}} .$$

This type of interpretation was supported by Diehl and his co-workers ([11]). Recently it would appear that the controversy has been at least partially resolved by modifying the $\varphi^{\frac{1}{2}}$ dependence to allow for the variation in dislocation line tension with length ([17]). The increase in σ with dose is not very dependent on the irradiation temperature between 4 °K and 300 °K (to within about (10÷20)%) ([18]) and recovers only in the temperature range (300÷400) °C (Stage V) with an activation energy of about 2 eV. It is clear, therefore, that the phenomenon is not due primarily to point defects, which anneal out much below the observed recovery temperature, and it has long been thought that defect clusters were responsible. The fact that the hardening is almost independent of irradiation temperature further suggested that it was the small vacancy clusters found in the cascades which were primarily responsible for the effect. The density of these clusters is of course proportional to dose at low doses, and hence a $\sigma \propto (\varphi t)^{\frac{1}{2}}$ relationship should be found.

This interpretation is completely consistent with the result obtained from the observations of defect clusters in irradiated copper ([19]), but since the

density of vacancy clusters visible is only $\sim(1/10)$-th of the number of primary recoils it is obvious that electron microscope observations in as-irradiated copper are not going to establish the true correlation, since a part of the hardening must be due to the sub-microscopic clusters. As discussed previously, however, it is possible to anneal the majority of the submicroscopic clusters very early in the annealing process ([20]). Detailed work on the correlation between the density of vacancy clusters and the critical shear stress measured at 4.2, 77 and 293 °K ([21]) is shown in Fig. 10. These results

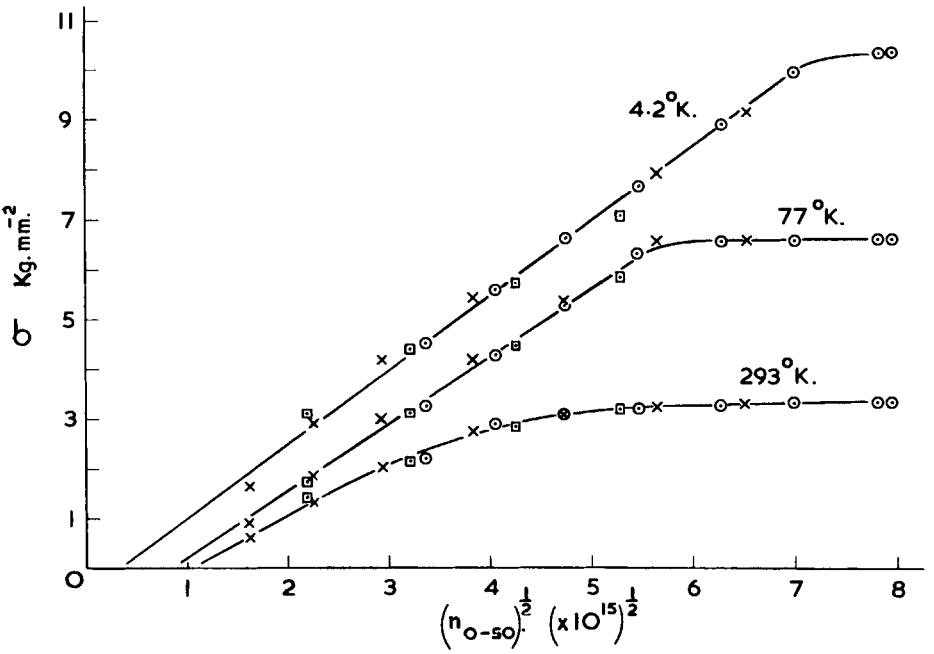


Fig. 10. – The correlation between the critical shear stress and the density of small clusters after various annealing treatments. (Courtesy of *Phil. Mag.*)

show that $\sigma_{4.2}$ is directly proportioned to $n^{\frac{1}{2}}$, where n is the density of clusters below 50 Å in diameter. There is no correlation at all between the critical shear stress and the density of interstitial loops. Hence the assumption made previously that it was the small vacancy clusters which controlled the critical shear stress is shown to be almost certainly correct. Two other points of considerable importance emerge from Fig. 10. First the value of the maximum force which can be exerted between a dislocation and a cluster, $F_{\max}$, is found to be $\approx \mu b^2$, where μ is the rigidity modulus and b the Burgers

vector. Secondly σ_{77} and σ_{293} are not initially sensitive to annealing. The significance of both these points will be discussed later.

Any attempt to calculate the critical shear stress produced by a given density of clusters must start with an assumption as to the force-distance curve for the interaction between a glissile dislocation and a cluster. Several such attempts have been made, notably by Seeger ([22]) and Fleischer ([23,24]). Seeger assumed rather a « soft » interaction, in which the potential energy of a dislocation varied with position x, as:

$$U(x) = U_0 \left[1 - \left(\frac{1}{1 + \exp [x/x_0]}\right)\right].$$

Such an interaction gives rise to the relationship:

$$\sigma^{\frac{3}{2}} = A(\phi t)^{\frac{1}{2}} [1 - BT]^{\frac{3}{2}},$$

i.e. there should be a linear relationship between σ and $(\varphi t)^{\frac{1}{2}}$, and also between $\sigma^{\frac{3}{2}}$ and $T^{\frac{3}{2}}$. The second of these relationship is not in very good agreement with experiment, since $\sigma^{\frac{3}{2}}$ is only approximately proportional to $T^{\frac{3}{2}}$ after a given anneal.

Fleischer calculated the long range elastic interaction between a loop and a dislocation and assumed that $F_{\max}$ occurred at a distance of $y = d/2$ from the loop, where d is the loop diameter. He neglected the variation in the stress tensor across the loop however, a procedure which gives only a very approximate result at the $y = d/2$ position. Fleischer's theory therefore gives rather a low value of $F_{\max}$ which is well below μb^2. Information on the nature of the interaction between a dislocation and an obstacle can be obtained by measuring the temperature and strain rate dependence of the critical shear stress, a procedure known as thermal activation analysis ([25-27]). Three different groups have carried out such measurements on irradiated copper, with rather different results. Makin ([28]) concluded that neither Seeger's theory nor Fleischers accurately represented the experimental behaviour, but that this was hardly surprising since on other grounds it was clear that a spectrum of obstacle sizes and strengths existed in as-irradiated copper. Diehl *et al.* ([29-31]) concluded that Seeger's theory was applicable if it was assumed that an obstacle spectrum existed. Koppenaal and Arsenault ([32,33]) came to the conclusion that Fleischer's theory accurately described the behaviour with a unique obstacle *i.e.* no spectrum. Controversy on this point has continued up to the present day. In my opinion, the presence of

a spectrum of obstacles is proved by the behaviour on annealing referred to previously, *viz.* the insensitivity of σ_{77} and σ_{293} to anneals which substantially reduce $\sigma_{4.2}$. If a single type of obstacle is present then this behaviour is virtually impossible ([34]), since it implies that the obstacles responsible for the high temperature critical shear stress are created during the anneal and are not present in as-irradiated material. If this were so then the temperature dependence of the high temperature critical shear stress would inevitably be altered, which is contrary to the experimental evidence. The alternative is that as-irradiated crystals contain obstacles which are *effective* only at low temperatures. These are removed by mild annealing with the result that only the low temperature critical shear stress is affected. Furthermore, it is found that after substantial anneals the low temperature critical shear stress actually become athermal, and the characteristics of the deformation change, the slip bands becoming fine, in contrast to the very coarse slip found in the thermal region. In the athermal region the obstacles are acting as if they were by-passed by the dislocations, rather than being cut by them, and the interaction force at which this occurs is $\sim \mu b^2$. This is the situation which exists for *all* the obstacles during deformation at 4.2 °K. This model, which explains the observed behaviour of both as-irradiated and annealed crystals,

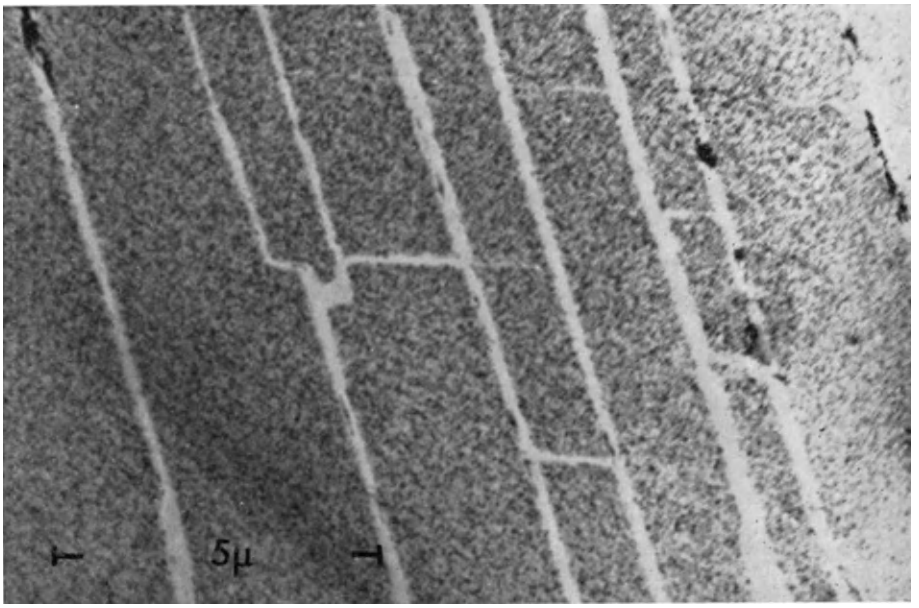


Fig. 11. – Slip bands in irradiated copper deformed at 20 °C.

predicts theoretical values of F_{max} well above μb^2, and hence is completely contrary to the Fleischer theory.

Another fact of considerable importance which has been established by electron microscopy in irradiated crystals is that during the Lüders band phenomenon, which occurs in the thermal region of deformation, all the defect clusters within the coarse slip bands are removed by the deformation Fig. 11 ([35]). This implies that during the deformation of each band there is an initial softening due to the removal, followed by a hardening up, due to the formation of a cold worked dislocation structure. This has a considerable effect on the deformation of irradiated material, since it follows that each slip band initially forms very rapidly and also that there is no true pre-yield microstrain, such as occurs readily in unirradiated metals. Both of these effects have been verified experimentally ([36 37]). The observation that the clusters are removed by deformation suggests that F_{max} is actually the force required to produce « sweeping up » or removal of the loop. There is as yet no calculation of the force required to sweep up stacking faulted loops.

3·4. Impurity effects.

Of the many elements which can be formed by nuclear transmutation processes the ones which have caused the largest effects in metals are undoubtedly the inert gases. The reasons for this are the inability of these elements to combine chemically with other elements in the material, and also their extreme insolubility. The inert gases most often result from either $^{10}B(n,\alpha)$ reactions in nonfissile materials or ^{235}U(n, fission) in fissile materials. Reactions can also occur between high energy neutrons and many other elements to produce α-particles which decay into helium atoms, and hence the production of small amounts of helium is common in most reactor irradiations. Although the concentrations are small ($< 10^{-5}$) it will be seen that they can cause large changes in the macroscopic properties.

3·4.1. *Gas bubbles.* – Initially it was thought that inert gas bubbles were of importance only in fuels, where the quantity of gas produced is quite large (~ 5 cm^3 NTP per cm^3 per 1 % burn-up). Due to the difficulty of handling highly radioactive irradiated fuel materials most of the basic work on bubbles has been done on model systems, such as α-particle irradiated copper. During production at low temperatures, the inert gas atoms are distributed randomly

and in the case of helium atoms, probably lie in interstitial positions in the crystal lattice. As the temperature is raised there is a considerable release of energy associated with the absorption of a vacancy, thereby enabling the gas atom to occupy a substitutional site, and a further release associated with the formation of gas bubbles. In order to become a gas bubble however, the gas atoms not only have to cluster but also have to absorb a large number of vacancies to reach the equilibrium size. It is this absorption of vacancies which is of course responsible for the swelling which occurs at high temperatures in irradiated fuels. This basic process was soon verified both by observing that during annealing the gas bubbles first became visible near to discontinuities in the crystal which could act as sources of vacancies ([38]), and by measurements of the lattice parameter of copper during annealing ([39]). In the latter experiment a large increase in parameter occurred during irradiation, due to the presence of interstitial helium atoms and, as these acquired vacancies during annealing, the lattice parameter decreased and became less than normal, due to the relaxation around gas atom-vacancy complexes. On further annealing the parameter returned to normal as gas bubbles were formed.

Since during irradiation at high temperatures there is always a plentiful supply of vacancies, and hence it is impossible to prevent the formation of gas bubbles, interest shifted to the properties of the bubbles.

In a solid containing gas bubbles and mobile vacancies the bubbles will reach an equilibrium size when the work done by the gas pressure in expanding the bubble becomes equal to the increase in surface energy. If the bubble has a radius r and contains gas at pressure p then the volume $v = \frac{4}{3}\pi r^3$ and the surface area $a = 4\pi r^2$. If the bubble increases in radius by dr then:

$$\mathrm{d}v = 4\pi r^2 \mathrm{d}r \quad \text{and} \quad \mathrm{d}a = 8\pi r \cdot \mathrm{d}r \,.$$

The work done by the gas in expanding is $p\mathrm{d}v$ and the increase in surface energy is $\gamma \mathrm{d}a$. When these are equal:

$$p \cdot 4\pi r^2 \cdot \mathrm{d}r = \gamma 8\pi r \,\mathrm{d}r \qquad \text{or} \qquad p = \frac{2\gamma}{r} \,.$$

It is obvious that the swelling can be minimised by containing the gas in a large number of small bubbles, where the pressure is high, rather than in a few large bubbles. The ratio of the number of vacancies to the number of gas atoms is easily derived, since for a perfect gas $pv = \frac{3}{2}mkT$, where m is

the number of gas atoms. With $p = 2\gamma/r$ and $v = \frac{4}{3}\pi r^3$ we get:

$$m = \frac{8\pi\gamma r^2}{3kT}.$$

The number of vacancies in a bubble of radius r is

$$n = \frac{4}{3}\frac{\pi r^3}{\Omega},$$

where Ω is the atomic volume. Hence the ratio of vacancies to gas atoms is $n/m = (kT/2\Omega\gamma)r$, and hence increases linearly with r. The scale of nucleation of the gas bubbles is strongly dependent on the irradiation temperature, being much finer at low relative temperatures. There are therefore two basic solutions to the technological problem of gas bubble swelling. The first of these is to run the fuel at a high relative temperature, so that although the bubbles nucleate on a very coarse scale and the fuel swells rapidly it is mechanically so weak that it imposes little stress on the fuel cans. Under these conditions a large proportion of the gas escapes from the fuel and this can be a problem if the fuel cans are not vented. In general the difficulties of this approach are such that it is not a very feasible solution. The second solution is to choose a fuel in which the operating temperature is relatively low, so that the bubbles nucleate on a fine scale and the gas is restrained at a high pressure. Such fuels swell very little and are feasible in practice. The choice of suitable fuel materials is very limited however, and one problem is that high temperature fuels such as UO_2 or UC have a lower thermal conductivity than metallic uranium, so that the fuel pins have to be of much smaller diameter in order to keep the maximum fuel temperature down. This of course adds to the complexity of the design. In addition to nucleating the bubbles on a fine scale attention is required to the feasibility of retaining the fine dispersion. There are two processes whereby bubbles can coarsen, re-solution and bubble migration. Re-solution of an inert gas is clearly very unlikely in the absence of radiation, since the heat of solution is so large. Under irradiation, however, there are several ways in which gas atoms can be redissolved, the simplest example of which is, of course, being knocked back as the result of atomic collisions. The evidence on the magnitude of the re-solution process is not yet very certain, although it is claimed that the process has been demonstrated experimentally in UO_2 ([40]), using transmission electron microscopy. Calculations of the numbers of gas atoms knocked

through the bubble surface are not very helpful, since it is certain that most of these will diffuse back to the bubble, and the effect of the local damage produced on such an event is difficult to allow for. Bubble migration has been demonstrated experimentally ([41]), especially in high temperature gradients, in which the bubble always migrates *up* the temperature gradient, at a velocity proportional to $1/r$ when the mechanism of migration is the surface diffusion of atoms from the hot to the cold side of the bubble. Bubbles can also migrate under the influence of other forces, such as a stress gradient, the motion of dislocation lines, grain boundary migration ([40]) or simply by Brownian motion. The latter is very small however, and can generally be neglected.

It was long believed that one of the most important applications of the gas bubble work was in the irradiation behaviour of α-uranium, where experimentally it was known that the poor swelling resistance of pure α-uranium could be greatly improved by small additions of iron and aluminium. This was interpreted as being due to the « anchoring » of small bubbles by the precipitates, so preventing migration and coalescence ([42]). The relative temperature at which swelling occurred in α-uranium, however, is very low compared with the model experiments, and it has since been shown by electron microscopy that the swelling does not occur by the formation of dispersed gas bubbles, but by grain boundary cavitation ([43]). Due to the irradiation growth process the grains in α-uranium are subject to constant plastic deformation during irradiation, and in pure uranium this opens up grain boundary and twin boundary cavities. Uranium containing Fe and Al additions does not show this effect until substantially higher burn-ups have been achieved, although in the same conditions the dispersed bubble densities are virtually identical ([43]). It has recently been shown that the grain boundary cavitation in pure uranium is not formed simply by slow straining, as out of reactor creep experiments do not produce it ([44]). Hence it must be an irradiation effect which can be inhibited by the addition of Fe and Al. The mechanism whereby this occurs is not known. In all of this work electron microscopy has, and is, playing a vital role.

3·4.2. *Irradiation embrittlement.* – In addition to the swelling which can result when large quantities of inert gases are produced there are also two effects of irradiation which require only very small quantities of gas. The significance of this is that effects due to gas are no longer confined to the fuel, but also occur in the structural materials used in the reactor. The high temperature embrittlement effect occurs in steels and other metals containing

trace amounts of boron, in which thermal neutrons produce the reaction $^{10}B(n\alpha)^{7}Li$ ([45]). The effect is a severe reduction in the elongation to fracture (Fig. 12) and in the stress rupture life at temperatures above those at which

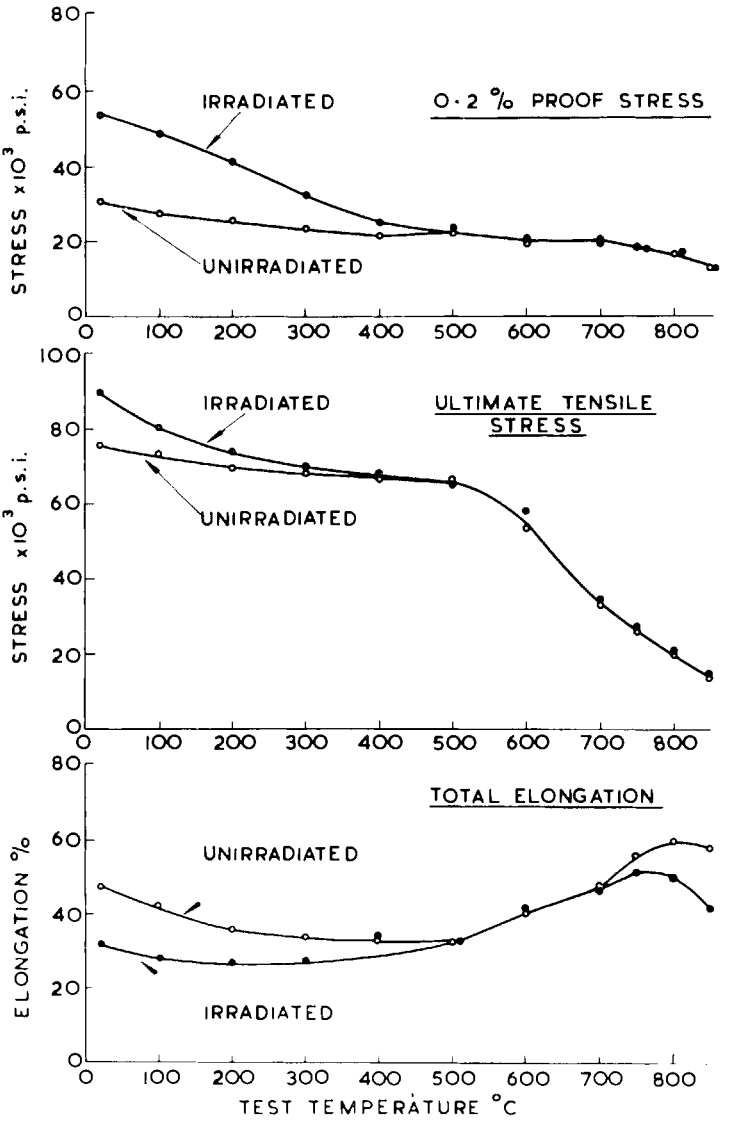


Fig. 12. – The change in properties of 20% Cr, 25% Ni:Nb stabilised austenitic steel produced by $3.8 \cdot 10^{18}$ fast and $7.6 \cdot 10^{18}$ thermal $n \cdot cm^{-2}$.

the displacement damage anneals, and occurs whether the irradiation is carried out at high temperature or the specimens are tested at high temperatures following room temperature irradiation. Very low concentrations of helium are required to produce the effect, in Nimonic PE16 for example, an observable effect occurs when the helium concentration is only $3 \cdot 10^{-9}$. The mechanism of the effect is not understood, but it is clearly associated with the behaviour of grain boundaries, since the boron in steels is known to lie primarily on the grain boundaries. Electron microscopy has been of limited use in this problem since the gas concentrations are so low. Bubbles have been seen at grain boundaries, however, and it is probable that they act as the nuclei of the cracks which lead to the fracture. The detailed mechanism is not known.

3'4.3. *Void formation.* – Another effect which appears to depend upon small quantities of gas is the formation of voids in structural metals during high dose irradiations of fast neutrons. The effect is clearly of considerable importance to the technology of fast breeder reactors, and is currently the subject of much study. The effect was first observed in stainless steel ([46]), but has since been found in several other metals, notably nickel ([47]), aluminium and copper ([48]). During high dose ($> (10^{20} \div 10^{21})$ n·cm^{-2}) fast neutron irradiations at temperatures of $\sim 0.4\, T_m$, where T_m is the absolute temperature of melting, the volume of the metal starts to increase and the density to decrease. Electron microscopy has shown that this is due to the formation of a high density of voids (Fig. 13). Although it is likely that the voids are initiated by gas, the gas pressure soon falls far below the equilibrium pressure given by $p = 2\gamma/r$, and this is shown by the fact that on annealing at 900 °C the voids shrink to the equilibrium size for the gas contained within them. The total void volume increases with neutron dose at a rate proportional to $(\varphi t)^{\sim 1.5}$. The void volume at a given dose is also dependent on the irradiation temperature, being zero until the temperature exceeds the displacement damage recovery temperature and rising to a maximum at about $0.4\, T_m$ and then falling to zero again at about $(0.6 \div 0.7)\, T_m$. The voids are observed to be distributed randomly throughout the material, except for a denuded zone typically (1000÷3000) Å wide, around the grain boundaries. The void density and average size increase with dose at constant temperature. Increasing the irradiation temperature results in fewer, larger voids. During annealing the smaller voids disappear first, presumably by the evaporation of vacancies and the larger voids tend to increase in volume, until they too begin to shrink when all the smaller voids have disappeared.

Fig. 13. – Voids in heavy ion irradiated face centred cobalt. The void faces are {111} planes.

The mechanism of the effect must involve some prefential annealing of either vacancies or interstitials, since the irradiation temperatures are sufficiently high to prevent the displacement damage clusters produced at low temperatures from being stable, and in the absence of some extra factors complete mutual annihilation of the vacancies and interstitials would occur. The extra factors are likely to be:

a) The effect of helium atoms on stabilising some of the irradiation produced vacancies to produce gas bubbles. This leaves some excess interstitials which precipitate as interstitial clusters (interstitial clusters are always observed associated with the voids).

b) The slight preferential absorption of interstitial atoms at dislocations and interstitial loops ([49]). This will leave a slight excess concentration of vacancies, a proportion of which will condense on the gas bubbles so increasing their size beyond the equilibrium size.

The efficiency of the process varies considerably with material, for example in stainless steel only about 0.1% of the point defects produced remain in the voids whereas in copper and nickel the corresponding figure is up to 10%.

3·5. Summary.

We have seen how studies of radiation growth, radiation hardening, gas bubbles and voids have depended very heavily on the transmission electron microscope technique for information on the nature of the effect and how it varies with the experimental parameters. In addition to providing information which can be obtained in no other way, however, electron microscopy is frequently used to speed up the experimental work. An excellent example of this is the use of accelerators to simulate fast neutron irradiations in void studies. To quickly reproduce the effect of very high neutron doses a heavy ion is usually used as a particle and the penetration of these in metals in only a few microns. This poses some difficulties for electron microscopists, who have to preparate a specimen from the right depth in the specimen, but it would be almost impossible to use such specimens for macroscopic measurements. In a similar way the small volume of material which is required for microscope studies enables very small specimens of materials which become highly radioactive to be irradiated. It is frequently possible to handle such specimens in the open laboratory with the minimum of shielding whereas the use of macroscopic specimens for physical measurements would be very expensive in remote handling requirements and experimental effort. It is important to realise that invaluable as microscopy is, it can never completely replace the physical measurements, and it is wise to arrange selected comparisons with bulk properties to check that no gross differences in behaviour are occurring *because* of the use of small specimens, or because of the limitations of the microscope technique. For example defects which are too small to be visible in the microscope can produce large macroscopic effects, as for instance in irradiation hardening and in the stored energy work on irradiated graphite. It is true to say, however, that without electron microscopy our knowledge of the nature of irradiation effects would be very much smaller and less certain than it is today. Electron microscopy has become, and is likely to remain, the major experimental technique in the study of the effect of irradiation on the physical and mechanical properties of materials. In other areas of radiation damage work, such as the electronic effects in semiconductors for example, it has made almost no contribution.

REFERENCES (Section 3)

1) J. H. KITTEL and S. H. PAINE: *Proc. 1st Geneva Conf. on Peaceful Uses of Atomic Energy* (1955), paper 745.
2) Y. QUÉRÉ: *Journ. Nucl. Mat.*, **9**, 3, 290 (1963).
3) S. N. BUCKLEY: private communication.
4) M. W. THOMPSON: *Journ. Nucl. Mat.* (1961).
5) S. F. PUGH: *Progr. Nucl. Energy*, **1**, 652 (1956).
6) L. L. SEIGLE and L. S. OPINSKI: *Nucl. Sci. Eng.*, **2**, 38 (1957).
7) A. M. COTTRELL: *46th Thomas Hawksley Lecture*, Inst. Mech. Eng. (1959).
8) B. HUDSON, K. M. WESTMACOTT and M. J. MAKIN: *Phil. Mag.*, **7**, 377 (1962).
9) B. HUDSON: *Phil. Mag.*, **10**, 949 (1964).
10) A. J. E. FOREMAN and J. D. ESHELBY: A.E.R.E., R-4170 (1962).
11) J. DIEHL: *Radiation Damage in Solids*, I.A.E.A. (1962), p. 129.
12) T. H. BLEWITT, R. R. COLTMAN, R. E. JAMISON and J. K. REDMAN: *Journ. Nucl. Mat.*, **2**, 277 (1960).
13) F. W. YOUNG: *Journ. Appl. Phys.*, **33**, 3553 (1962).
14) D. O. THOMPSON and V. K. PARÉ: *Journ. Appl. Phys.*, **36**, 243 (1965).
15) T. J. KOPPENAAL: *Journ. Appl. Phys.*, **35**, 2750 (1964).
16) M. J. MAKIN and F. J. MINTER: *Acta Met.*, **8**, 691 (1960).
17) T. H. BLEWITT and C. A. ARENBERG: *Proc. of Int. Conf. on Strength of Metals and Alloys*, Tokyo, Japan Institute of Metals (1968), p. 266.
18) T. H. BLEWITT: *Radiation Damage in Solids*, Academic Press, (1962).
19) M. J. MAKIN, A. D. WHAPHAM and F. J. MINTER: *Phil. Mag.*, **7**, 285 (1962).
20) M. J. MAKIN and S. A. MANTHORPE: *Phil. Mag.*, **8**, 1725 (1963).
21) M. J. MAKIN, F. J. MINTER and S. A. MANTHORPE: *Phil. Mag.*, **13**, 729 (1966).
22) A. SEEGER: *Proc. 2nd Int. Conf. on Peaceful Uses of Atomic Energy*, **8**, 250 (1958).
23) R. L. FLEISCHER: *Acta Met.*, **10**, 835 (1962).
24) R. L. FLEISCHER: *Journ. Appl. Phys.*, **33**, 3504 (1962).
25) H. CONRAD and H. WIEDERSICH: *Acta Met.*, **8**, 128 (1960).
26) G. SCHOECK: *Phys. Stat. Sol.*, **8**, 499 (1965).
27) G. B. GIBBS: *Phil. Mag.*, **16**, 97 (1967).
28) M. J. MAKIN: *Phil. Mag.*, **9**, 81 (1965).
29) J. DIEHL, G. P. SEIDEL and L. NIEMANN: *Phys. Stat. Sol.*, **11**, 339 (1965).
30) J. DIEHL, G. P. SEIDEL and L. NIEMANN: *Phys. Stat. Sol.*, **12**, 405 (1965).
31) J. DIEHL and G. P. SEIDEL: *Phys. Stat. Sol.*, **17**, 43 (1966).
32) T. J. KOPPENAAL and R. J. ARSENAULT: *Phil. Mag.*, **12**, 951 (1965).
33) T. J. KOPPENAAL: *Acta Met.*, **16**, 89 (1968).
34) M. J. MAKIN: *Phil. Mag.*, **18**, 1245 (1968).
35) J. V. SHARP: *Phil. Mag.*, **16**, 77 (1967).
36) M. J. MAKIN and J. V. SHARP: *Phys. Stat. Sol.*, **9**, 109 (1965).
37) J. V. SHARP and M. J. MAKIN: *Phil. Mag.*, **12**, 427 (1965).
38) R. S. BARNES, G. B. REDDING and A. H. COTTRELL: *Phil. Mag.*, **3**, 25, 97 (1958).
39) B. RUSSELL and I. J. HASTINGS: *Journ. Nucl. Mat.*, **17**, 30 (1965).
40) A. D. WHAPHAM and B. E. SHELDON: A.E.R.E., R.-4970.

41) R. S. BARNES and D. J. MAZEY: *Proc. Roy. Soc.*, A **275**, 47 (1963).
42) R. S. BARNES: *A.S.T.M. Technical Publication*, **380**, 40 (1965).
43) B. HUDSON: A.E.R.E., R.-5706.
44) B. HUDSON: private communication.
45) D. R. HARRIES, K. Q. BAGLEY, I. P. BELL, W. S. GIBSON, J. GILLIES, P. C. L. PFEIL and S. B. WRIGHT: *Third U. N. Conf. on Peaceful Uses of Atomic Energy*, **28**, 162 (1964).
46) C. CAWTHORNE and E. J. FULTON: *The Nature of Small Defect Clusters*, H.M.S.O. (1966), p. 446.
47) J. L. BRIMHALL and B. MASTEL: *Journ. Nucl. Mat.*, **33**, 186 (1969).
48) J. L. BRIMHALL and B. MASTEL: *Journ. Nucl. Mat.*, **29**, 123 (1969).
49) G. W. GREENWOOD, A. J. E. FOREMAN and D. E. RIMMER: *Journ. Nucl. Mat.*, **4** 305 (1959).

4. Radiation damage studies in high voltage microscopes.

4·1. Introduction.

It has hitherto been customary to consider the interaction between the electron beam and a crystalline specimen only in terms of firstly, elastic scattering of the beam, resulting in Bragg reflections, and secondly, inelastic scattering in so far as this results in « absorption », *i.e.* loss of contrast, brightness and resolution in the image. At 100 kV no permanent changes occur in metal specimen as a result of inelastic scattering. Because of this the term « radiation damage » in electron microscopy has come to mean the permanent changes in a particular class of specimens which are sensitive to ionisation effects, *i.e.* polymers, biological materials, etc. These effects arise because of the sensitivity of these materials to electron displacement or excitation events. In a metal, of course, these effects are very transient ($\sim 10^{-15}$ s) because the free electron gas rapidly removes any local perturbations in electric charge. With the advent of high voltage microscopes, however, a completely different type of radiation damage phenomena due to displacement of atoms becomes important. The range of materials in which this type of damage occurs at 100 kV is very limited, because a 100 keV electron can displace atoms of only a few light elements. This arises because the maximum energy which can be transferred is proportional to $1/M_2$, where M_2 is the mass of the target material. Hence, as M_2 is increased the maximum energy transferred decreases, and since there is a well defined displacement energy threshold for an atom to be ejected from its normal lattice position a positive effect can be seen in only a very few materials.

Because the « atom displacement » process is relatively new to electron microscopists there is, I feel, a danger of some confusion if the old term « radiation damage » is applied indiscriminately to both electronic and atomic displacement processes. The distinction between the two effects is of course, well known to workers in the field of radiation effects. The complication arises mainly in the biological and polymer fields since in these materials there is at high voltages the possibility of both electronic and atomic effects, and care should be used to distinguish between the two. In metals we need consider only the atomic displacements, since as before, the electronic effects have no observable effect on the specimen.

4·2. The displacement process.

It is generally permissible to neglect the effect of the screening electrons in calculating the scattering of electrons by atoms. This is an imperfect approximation at low energies but for energies of above about 500 keV the error involved is small and a simple Coulomb potential can be used. As described earlier (Subsect. 1·2.2.) relativistic quantum mechanics are necessary since the electron velocity in the energy range of interest is a substantial fraction of the velocity of light. An electron of velocity v has a momentum $P = mv$, where m is the mass of the electron at velocity v. Because of the disparity in masses the electron velocity is hardly altered by the collision so that the momentum transfer to the struck atom is:

$$\Delta P = 2mv \sin \theta/2 ,$$

where θ is the angle through which the electron is scattered. An atom of mass M_2 and momentum ΔP will have a velocity of v_2, where

$$M_2 v_2 = \Delta P$$

and an energy of

$$\frac{1}{2} M_2 v_2^2 = \frac{(\Delta P)^2}{2M_2} = E_2 .$$

Hence

$$E_2 = \frac{2m^2 v^2}{M_2} \sin^2 \theta/2 .$$

The electron energy E and mass m are given by:

$$E = (m - m_0)c^2 \quad \text{and} \quad m = \frac{m_0}{(1 - v^2/c^2)^{\frac{1}{2}}} .$$

Hence

$$E_2 = \frac{2(m^2 - m_0^2)c^2}{M_2} \sin^2 \theta/2 = \frac{2(m + m_0)(m - m_0)c^2}{M_2} \sin^2 \theta/2 ,$$

so that

$$E_2 = \frac{2E}{M_2 c^2}(E + 2m_0 c^2) \sin^2 \theta/2 .$$

The maximum value of E_2 occurs, of course, when $\sin \theta/2 = 1$, *i.e.* $\theta = 180°$.

Using the values of the displacement energy quoted in Table I of Section **1** the threshold electron energies are given in Table XI. It will be seen that in a 1 MeV microscope it is possible to displace atoms in a wide range of elements. In Fig. 14 the value of $E_m M$, where E_m is the maximum knock-on

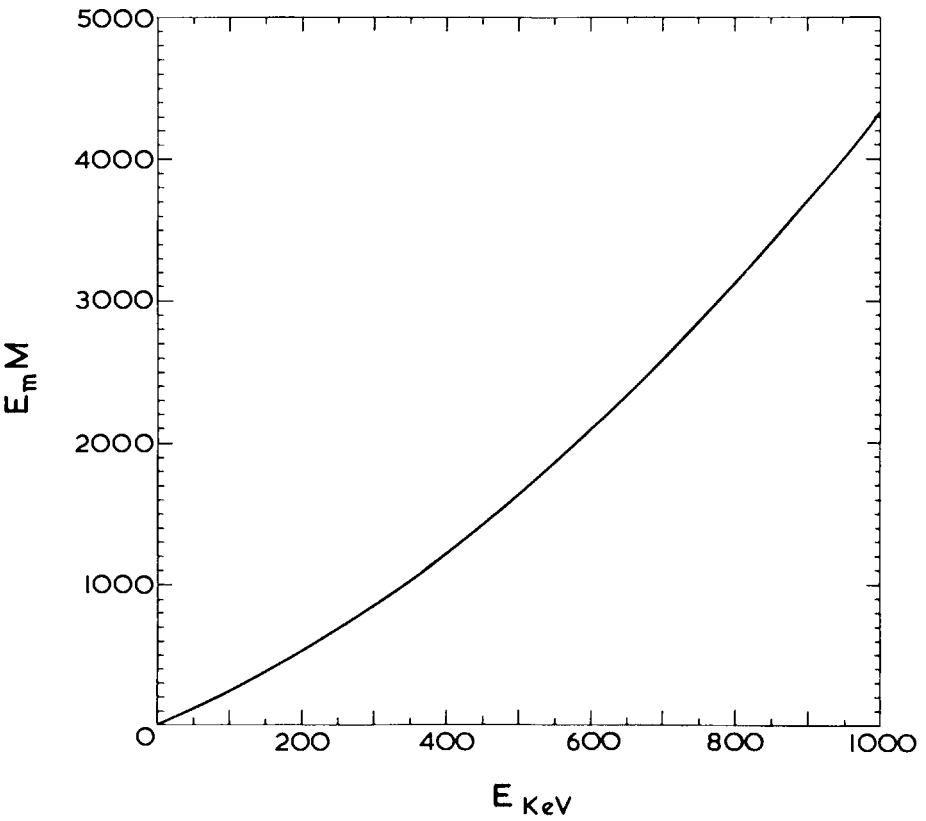


Fig. 14. – The relation between the maximum knock-on energy E_m and the electron energy E. (Courtesy of *Phil. Mag.*)

energy and M the atomic weight, is given as a function of accelerating voltage for electron irradiation ([1]). From this graph the value of E_m can be determined for any element at any voltage up to 1 MV or alternatively, for any given value of E_m, for example the displacement energy, the threshold electron accelerating voltage can be determined for any element.

TABLE XI. – *The threshold electron energy for atom displacement.*

Element	Displacement energy E_d (eV)	Electron energy (keV)
Cu	19	400
Al	16	166
Au	34	> 1000
Pt	37	> 1000
Fe	24	430
Mo	37	870
W	35	> 1000
Ti	29	440
Ni	24	450
Ag	28	790

4·2.1. *The electron current density.* – A major difference between the electron beam in a microscope and in a conventional electron accelerator is the very high current density in the former. In a well aligned microscope adjusted for a maximum beam current it is possible to obtain a current of (0.5÷0.6) μA in a 2 μ spot at the specimen. This corresponds to a current density of ~ 15 A/cm^2, or $9 \cdot 10^{19}$ electrons/cm^2/s, which is ~ 10 000 times greater than in a conventional electron accelerator.

4·2.2. *Displacement cross-section.* - For light elements the displacement cross-section is given by the McKinley-Feshbach approximation ([2]) as:

$$\sigma_d = 2.48 \cdot 10^{-25} Z_2^2 \left(\frac{1-\beta^2}{\beta^4}\right)\left(\frac{E_m}{E_d}\right) \cdot \left\{1 + 2\pi\alpha\beta\left(\frac{E_d}{E_m}\right)^{\frac{1}{2}} - \left(\frac{E_d}{E_m}\right)\left[1 + 2\pi\alpha\beta + (\beta^2 + \pi\alpha\beta)\ln\left(\frac{E_m}{E_d}\right)\right]\right\},$$

where $\alpha = Z_2/137$. As stated before (Subsect. **1**·2.3*d*), this formula is reasonably accurate for the light elements, but underestimates σ_d for heavy elements. For more accurate values in these elements the reader is referred to Oen's graphs

and tables ([3]), which were computed by a method which avoids the McKinley-Feshbach approximation. The discrepancy between the computed values and those given by the formula increases as the accelerating voltage is increased.

The total number of displacements produced per primary knock-on is:

$$N_d = 1 + \ln\left(\frac{E_m}{2E_d}\right), \quad \text{when} \quad E_m > 2E_d,$$

and $N_d = 1$, when $E_m < 2E_d$. Using this formula and the McKinley-Feshbach approximation the values of σ_d and $\sigma_d N_d$ as a function of accelerating voltage are given in Fig. 15. The concentration of displacements produced per second is, of course,

$$C = \varphi \sigma_d N_d t .$$

For example in aluminium at 1 MV, with an electron current density φ of

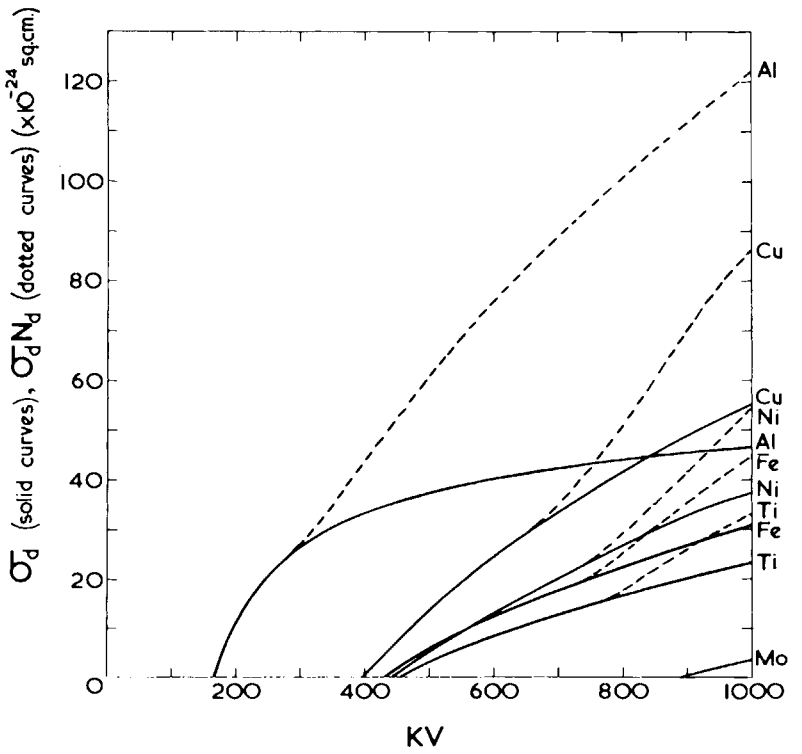


Fig. 15. – The displacement cross-sections σ_d (solid curves) and total cross-sections $\sigma_d N_d$ (dashed curves), as a function of accelerating voltage. (Courtesy of *Phil. Mag.*)

$9 \cdot 10^{19}$ electrons/cm^2/s the concentration of defects produced per second is:

$$C = 1.1 \times 10^{-2}\ \mathrm{cm^{-3}\ s^{-1}}\,,$$

and the corresponding figure for iron is $\sim 3.5 \cdot 10^{-3}$. These are extremely large figures; for example, using Nelson ([4]) estimate for iron for the number of defects produced per knock-on, a flux of 10^{13} 1 MeV neutrons produces a defect concentration of $5.4 \cdot 10^{-9}$ per second. Hence in iron a 1 MV microscope is capable of producing defects at a $6.4 \cdot 10^5$ times faster than the reactor. This can best be appreciated by considering that 1 minute in the microscope corresponds to 445 days in the reactor in terms of the point defect production.

4·3. The effect of electron irradiation at high voltages.

4·3.1. *Theoretical.* – Before describing any experimental results it is instructive to consider what might be expected in a metal specimen as the result of exposure to a high voltage electron beam in a microscope. Because of the low primary knock-on energy there are no collision cascades and the defects are produced as isolated interstitial vacancy pairs. Dienes and Damask ([5]) considered this, and gave the differential equations which, under certain assumptions, control the densities of the interstitials and vacancies:

$$\frac{\mathrm{d}v}{\mathrm{d}t} = K - K_v v - \nu_i Z(v + v_0) i$$

$$\frac{\mathrm{d}i}{\mathrm{d}t} = K - K_i\, i - \nu_i Z(v + v_0) i\,,$$

where K is the defect production rate, $K_v v$ and $K_i i$ are the loss of vacancies and interstitials to sinks, ν_i is the interstitial jump rate, Z is the number of sites around a defect from which mutual recombination can occur spontaneously, v and v_0 are the actual and thermal equilibrium vacancy concentrations, and i the interstitial concentration. It is important to note two assumptions made in these equations. The first is that the sinks to which the defects migrate are smeared out throughout the crystal, and secondly that interstitials do not cluster. We assume that $K_v = \alpha_v \nu_v \lambda^2$ and $K_i = \alpha_i \nu_i \lambda^2$ where α_v and α_i are the sink densities and λ is the jump distance ($\lambda^2 \simeq 10^{-15}$ cm^2). Analytic solutions of these equations are possible only in steady state conditions, *i.e.* when $\mathrm{d}v/\mathrm{d}t = \mathrm{d}i/\mathrm{d}t = 0$. The time required to reach steady state

can be very long at low temperatures, when the vacancy mobility is low and the steady state solutions are of little relevance in these conditions. Despite this fault however, the steady state solutions are instructive. In the steady state the vacancy and interstitial concentrations are:

$$v = \frac{(\alpha_i \lambda^2 + Zv_0)}{2Z} \cdot A$$

and

$$i = \frac{\alpha_v \nu_v (\alpha_i \lambda^2 + Zv_0)}{2\alpha_i Z \nu_i} \cdot A ,$$

where

$$A = \left\{ -1 + \sqrt{1 + \frac{4K\alpha_i Z}{\alpha_v \nu_v (\alpha_i \lambda^2 + Zv_0)^2}} \right\} .$$

At temperatures where the vacancies are immobile there is no true steady state solution, since there can be no vacancy flow into sinks, and hence the vacancy and interstitial flows cannot become truly equal (this is a condition of the steady state). When the vacancy mobility is finite but low the term $\alpha_v \nu_v (\alpha_i \lambda^2 + Zv_0)^2 \ll 4K\alpha_i Z$ and, when $\alpha_i = \alpha_v$ we get

$$v \to \left(\frac{K}{Z\nu_v} \right)^{\frac{1}{2}} \qquad \text{and} \qquad i \to \frac{1}{\nu_i} \left(\frac{K\nu_v}{Z} \right)^{\frac{1}{2}} .$$

In these conditions the vacancy concentration is very large and the mutual recombination term dominates.

At relatively high temperatures ν_i is large and hence

$$\alpha_v \nu_v (\alpha_i \lambda^2 + Zv_0)^2 \gg 4K\alpha_i Z$$

and

$$v \to \frac{K\alpha_i}{\alpha_v \nu_v (\alpha_i \lambda^2 + Zv_0)} \to \frac{K}{\alpha_v \nu_v \lambda^2} , \qquad \text{when} \qquad \alpha_i \lambda^2 \gg Zv_0 ,$$

and

$$i \to \frac{K}{\nu_i (\alpha_i \lambda^2 + Zv_0)} \to \frac{K}{\alpha_i \nu_i \lambda^2} , \qquad \text{when} \qquad \alpha_i \lambda^2 \gg Zv_0 .$$

The mutual recombination term is now unimportant, the vacancy and interstitial concentrations are low and the majority of the defects migrate to the fixed sinks.

Some enhancement of the diffusion rate may be expected in these conditions. The diffusion coefficients are:

$$D_v = \frac{K}{\alpha_v} \quad \text{and} \quad D_i = \frac{K}{\alpha_i},$$

and the diffusion lengths are:

$$L_v = \left(\frac{K}{\alpha_v}\right)^{\frac{1}{2}} \cdot t^{\frac{1}{2}} \quad \text{and} \quad L_i = \left(\frac{K}{\alpha_i}\right)^{\frac{1}{2}} \cdot t^{\frac{1}{2}}.$$

For example in aluminium at 1 MV with a production rate K of $1.1 \cdot 10^{-2}\,\mathrm{s}^{-1}$ and a sink density of 10^8 dislocation lines per cm^3 ($\alpha \simeq 10^8$), the diffusion distance is of the order of 3.3 μm after a time of 10^3 s. The effect of this should be experimentally observable at the right temperature.

At low temperatures the Dienes and Damask model may be unrealistic, since it neglects the formation of defect clusters. In the absence of cascades, clusters must nucleate by the chance collision of migrating point defects. Since the interstitial is the more mobile defect there is a greater probability of two interstitials meeting than two vacancies in the early stages of irradiation, when the interstitial and vacancy concentrations are roughly equal. If we assume that a di-interstitial is immobile and fairly stable at low irradiation temperatures then there is a finite probability that a cluster will develop from the di-interstitial nucleus by the condensation of more interstitials ([1]). A similar model has been developed by Brown, Kelly and Mayer ([6]) in an analysis of the homogeneous nucleation of loops in boronated graphite during neutron irradiation, and also briefly applied by Brown ([7]) to electron irradiated copper. The density of clusters formed can be simply estimated by the following argument ([1]). If N is the saturated cluster density, then associated with each cluster there is a volume V, where $V = 1/N$. The point defect production rate within the volume V is KV, where K is the number of defects per cm^3 per s. In the time interval between the formation of individual defects in a volume V an interstitial samples a volume $\nu_i v'/KV$, where v' is the atomic volume and $1/KV$ is the time interval between defect production. Hence in the first $1/KV$ s the probability of interstitial-vacancy recombination is $\nu_i Z v'/KV^2$. Assuming that the defects do not mutually annihilate then during the second $1/KV$ s there are two interstitials and two vacancies in the volume V. The probability of interstitial-vacancy recombination is $4\nu_i Z v'/KV^2$ and $\sqrt{2}\nu_i Z' v'/KV^2$ for interstitial-interstitial combination. Note that Z' is not necessarily equal to Z. The assumption is then made that the sum of these

probabilities is unity:

$$\frac{\nu_i v'}{KV^2}(4Z + 1.4Z') = 1\,,$$

or if $Z = Z'$ then

$$N = \frac{1}{V} = \sqrt{\frac{K}{5.4\nu_i \cdot Zv'}}\,.$$

The cluster density is hence proportional to $K^{\frac{1}{2}}$ and $\nu_i^{-\frac{1}{2}}$. The same proportionabilities, with a different factor, is also given by (6). Experimentally, higher production rates K and lower temperatures (smaller ν_i), give higher values of N, in qualitative agreement with the theory. The model gives predicted values of the cluster density which are in good agreement with experiment, bearing in mind the uncertainty in ν_i and Z. The theoretical determination of ν_i is particularly difficult since not only is this dependent on an accurate knowledge of E_m^i but it also depends on the impurity concentration, since impurity trapping of interstitials will decrease ν_i (6). Note that this is a homogeneous nucleation model, and that N is independent of time. The nucleation time will be very short in practical cases ($\sim 10^{-5}$ s). The model can also predict the rate of growth of clusters with time by calculating the probability of an interstitial in the volume V combining with either the cluster containing n interstitials, or with the n vacancies distributed throughout the volume. This leads to the expression:

$$n = 4{\cdot}77\left(\frac{K\nu_i v' Z''^2}{Z}\right)^{\frac{3}{5}} \cdot t^{\frac{3}{5}}\,,$$

where Z'' is the recombination number between a cluster site and an interstitial. The loop radius is therefore:

$$r = 0.733a\left(\frac{K\nu_i v' Z''^2}{Z}\right)^{\frac{1}{5}} \cdot t^{\frac{1}{5}}\,.$$

Clearly there should be a denuded surface layer in the foil since in the nucleation time $1/KV$ the interstitials will migrate a mean distance

$$x = 0.66\lambda\left(\frac{\nu_i}{KZv'}\right)^{\frac{1}{4}}\,.$$

Hence, within x of a surface or grain boundary the interstitial concentration during the nucleation time will be lowered by the proximity of the interfacial sink, and no cluster will nucleate.

It can be seen that the surface denuded layers will affect the threshold voltage for cluster nucleation since as the voltage is reduced towards threshold the production rate K decreases and so x increases until the denuded zones from each surface meet in the centre of the foil. Substituting $K = \sigma_d N_d \varphi / v'$, where φ is the electron dose we get:

$$h = 1.32 \left(\frac{\nu_i}{\sigma_d N_d Z \varphi} \right)^{\frac{1}{4}}$$

Foils of thickness less than h will contain no loops. Using the values of $\sigma_d N_d$ given for copper in Fig. 15, $\lambda = 2.55 \cdot 10^{-8}$ cm, $\nu_i = 10^{13} \exp [0.2/KT]$,

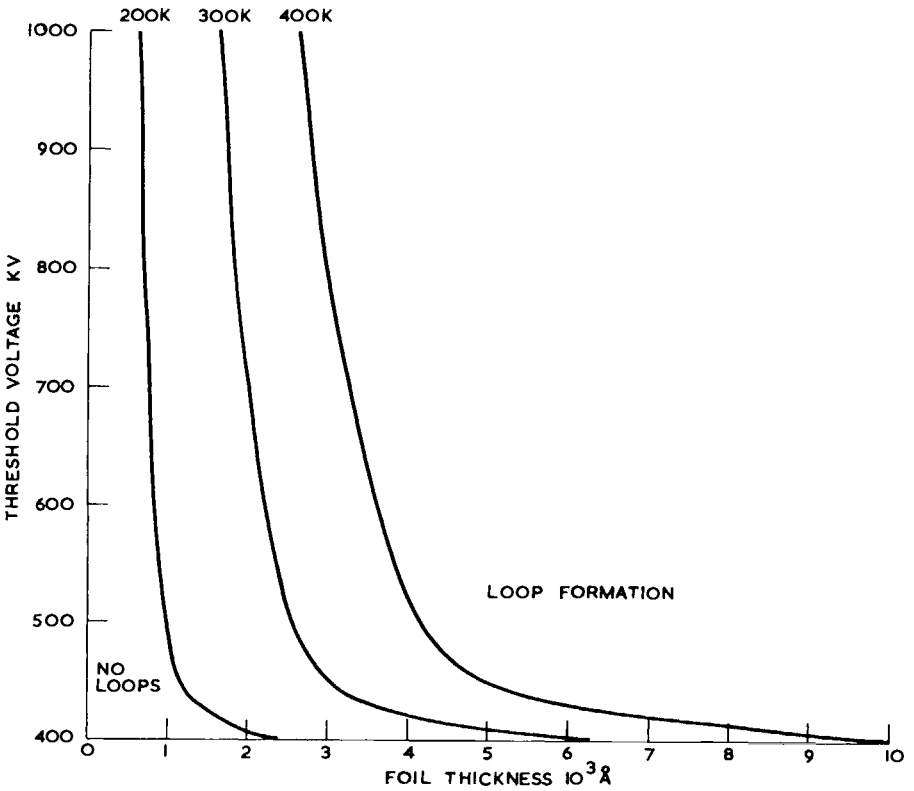


Fig. 16. – The apparent threshold voltage to produce loops in copper as a function of foil thickness, assuming a true threshold of 394 kV. (Courtesy of *Phil. Mag.*)

$\varphi = 10^{19}\,\mathrm{cm^{-2}\,s^{-1}}$ and $Z = 100$ we obtain the curves in Fig. 16. Hence in determining the displacement energy from the threshold voltage the effect of the foil thickness should be allowed for.

The model of loop formation given here can also be used to predict the vacancy and interstitial concentrations as a function of time and hence the enhanced diffusion coefficients from the expressions:

$$D_v = v\nu_v\lambda^2 \quad \text{and} \quad D_i = i\nu_i\lambda^2 .$$

The vacancy concentration is:

$$v = \frac{2.05\nu'^{\frac{5}{6}} Z''^{\frac{5}{6}} K^{\frac{5}{6}}}{Z^{\frac{5}{6}}\nu_i^{\frac{1}{6}}} \cdot t^{\frac{5}{6}},$$

and the interstitial concentration:

$$i = 0.488\left(\frac{K\nu'}{Z}\right)^{\frac{1}{6}} \cdot \frac{1}{\nu_i^{\frac{5}{6}} Z''^{\frac{5}{6}}} \cdot \frac{1}{t^{\frac{5}{6}}} .$$

Inserting the relevant numbers into these formulae we find that in copper at 300 °K, $\varphi = 10^{19}\,\mathrm{cm^{-2}\,s^{-1}}$, etc., we get:

$$v = 4.6 \cdot 10^{-7} \cdot t^{\frac{5}{6}}$$

and

$$i = 1.21 \cdot 10^{-9} \cdot t^{-\frac{5}{6}} .$$

The resulting diffusion distances ($x = \sqrt{Dt}$) are very small (~ 10 Å) and hence unless an exceptional sensitive technique is used the effect will not be observable.

4·3.2. *Experimental results.* – The earliest work which established that dislocation loops were formed as a result of exposure to a high voltage beam in a microscope was the work of Makin ([8]) on copper and aluminium in the Cambridge 750 kV microscope. It was found that dislocation loops were formed only in the electron irradiated spot (Fig. 17) and it was shown by stereo-microscopy that they extended throughout the thickness of the foil, except for the surface denuded layers. The rate of change in the loops with time depended on the beam current and there was a clearly defined threshold voltage below which damage was not produced. It is clear from these observa-

tions that the loop formation is the result of the electron irradiation, and not from ion bombardment. Ions produced in the gun and accelerated in the accelerator will be distributed fairly uniformly across the whole specimen,

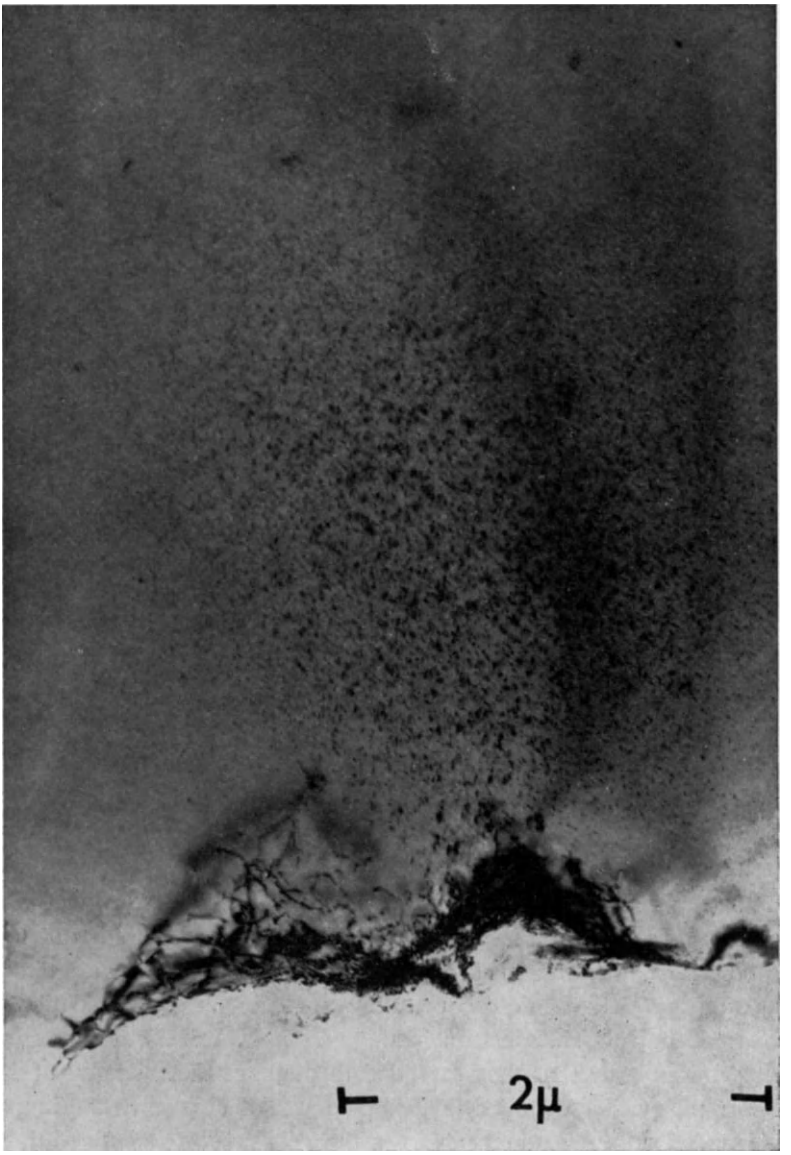


Fig. 17. – The distribution of loops within the electron irradiated spot.

since because of their large mass they are hardly affected by the lens magnetic fields. It has been found that ion beams emerge from the accelerator only if it is poorly conditioned and hence contains a relatively high gas pressure. Once the accelerator is properly conditioned the ion current is negligible. Similarly the observation of loops throughout the thickness of the foil eliminates the possibility that the damage is due to ions knocked into the foil as a result of electron collisions in the neighbourhood of the specimen. Such ions would have a maximum energy of only ~ 100 eV and hence would penetrate only a few lattice spacings. The observation of a well defined threshold voltage ([8]) is also very strong evidence that the damage is due solely to the electrons.

The loop formation characteristics showed the behaviour expected of a homogeneous nucleation process *i.e.* the full density of clusters appeared very early in the irradiation and thereafter no new clusters were nucleated (Fig. 18). In fact the density of clusters tended to decrease slightly with further

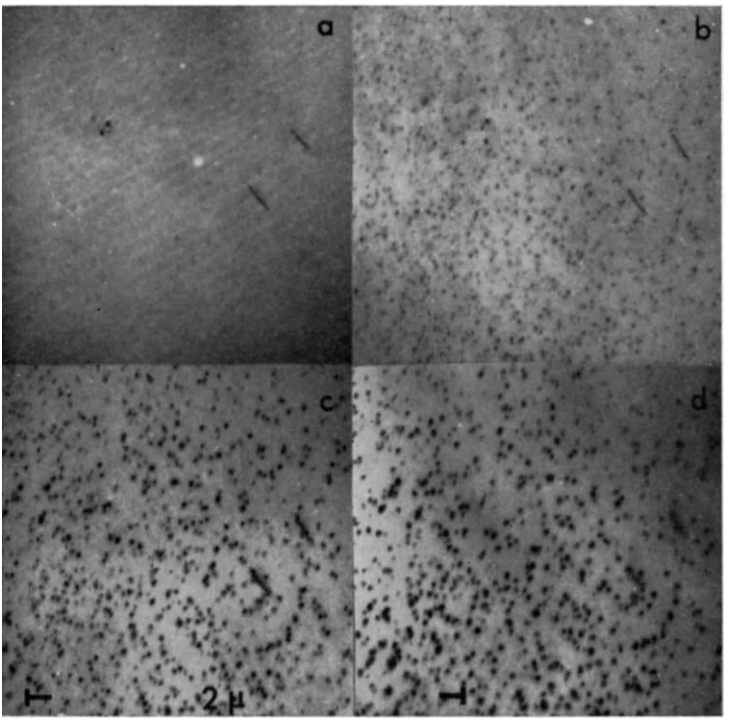


Fig. 18. – Cluster formation in copper exposed to the microscope beam for *a*) $\frac{1}{2}$, *b*) 5, *c*) 19, *d*) 32 min at 600 kV. (Courtesy of *Phil. Mag.*)

irradiation (Fig. 19) and direct observation showed that this occurred by a process of coalescence by slip. The clusters were observed to be mainly perfect prismatic dislocation loops and hence the interaction forces will

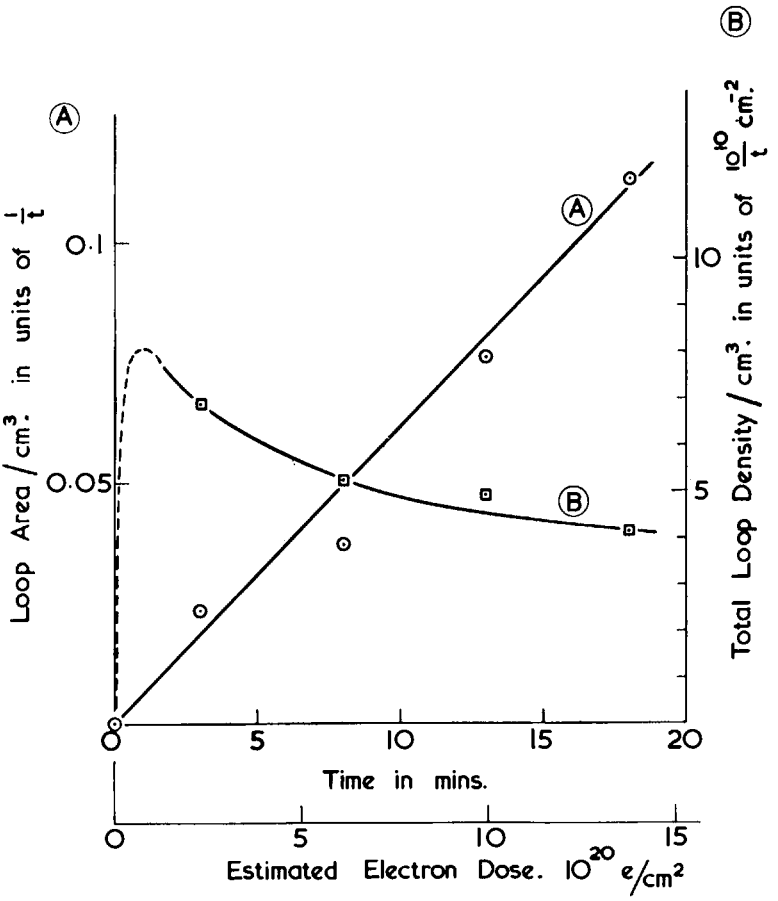


Fig. 19. – The total loop area *A* and the loop density *B*, as a function of electron dose. t is the thickness of the foil. (Courtesy of *Phil. Mag.*)

increase as the loop area, whereas the resisting force is proportional to the loop perimeter. Hence as the loops grow during irradiation there is an increasing probability of slip.

A denuded zone was observed around all grain boundaries and free surfaces (Fig. 20) but not around twin boundaries, and in copper at 300 °K with an electron flux of $\sim 10^{19}$ cm^{-2} s^{-1} this was observed to be about 1000 Å in width, in agreement with the theoretical value. The loop density was also in agreement with the value expected from the model described

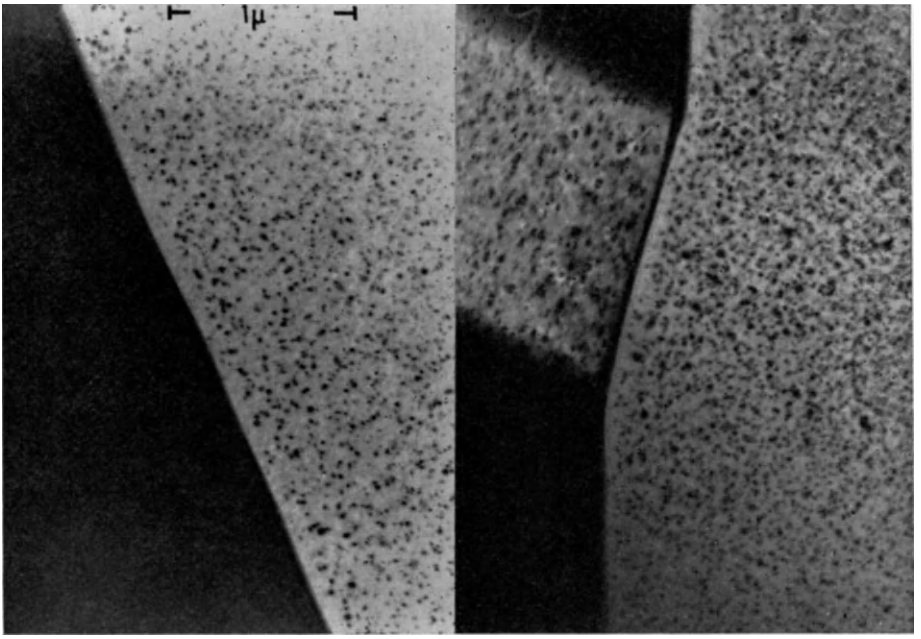


Fig. 20. – Grain boundary denuded zones in electron irradiated copper.

in 3·2. We have already seen how it is virtually impossible for the loops to be other than interstitial loops, from considerations of the point defect mobilities. This interpretation was further strengthened, (Makin ([8])), by the observations on the effect of electron irradiations on the known interstitial loops in neutron irradiated copper. It was observed that in regions containing a high density of neutron produced loops there was no nucleation of new loops, but a growth of the existing loops, so confirming that the effect was due to the condensation of interstitials. Subsequently, diffraction contrast

experiments ([9]) have confirmed the original interpretation of the sign of the loops.

Attempts have been made ([10]) to use the effect to determine the threshold energy for displacement in the principal crystal directions. An example of the results obtained in copper in the ⟨100⟩ direction are shown in Fig. 21, from which it is clear that in this experiment the threshold was ~ 495 keV, which corresponds to a displacement energy of 25 eV. To minimise the error

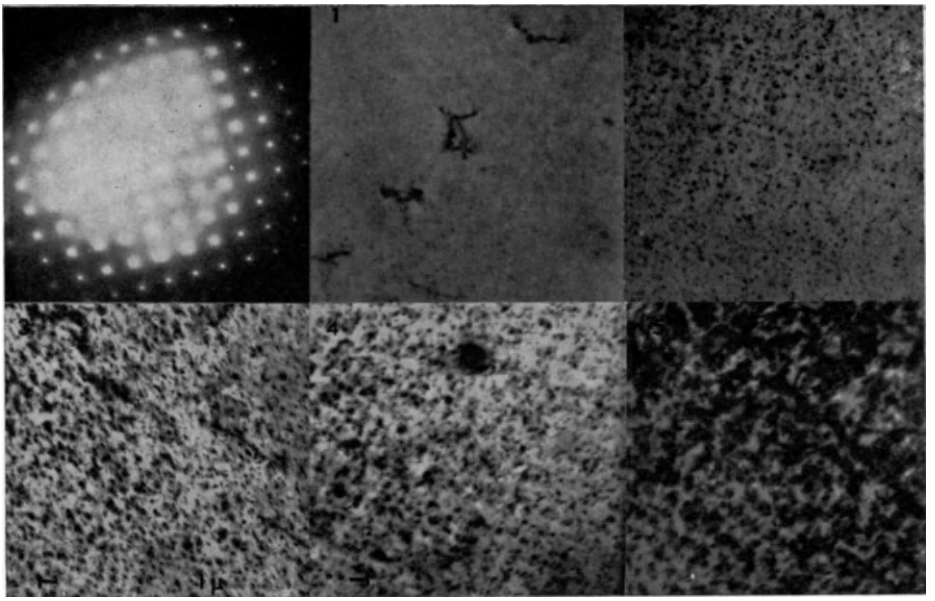


Fig. 21. – The apparent threshold energy in the [100] direction. The micrographs are after 15 min irradiation at 1) 490, 2) 500, 3) 510, 4) 520 and 5) 600 kV.

due to the finite foil thickness, experiments have been made in copper using very thick foils, with the following results: E_d⟨100⟩, 21.6 eV; ⟨110⟩, 19.2 eV and ⟨111⟩, 23.6 eV. These results have been previously quoted and commented on in Table II of Section **1**.

The behaviour of the loops after very high doses is interesting and unusual. Figure 22 shows a typical example of how the initially planar loops become complex as the irradiation is continued. The interpretation of this effect

is not clear at present. It is possible that the effect is associated with the presence of the vacancies, since in some cases loops are found which contain reentrant segments which suggest this, but these may well result from loop coalescence. The most probable interpretation is that at a particular temperature the loop remains planar only up to a certain size (this size will increase with temperature), and when this size is exceeded condensation of interstitials commences to occur on several {110} planes with the eventual formation of a very complex and tangled dislocation structure. When the loops are very small presumably the stress due to the loop is too small for this to occur. As the loop becomes larger, however, the magnitude of the stress field increases

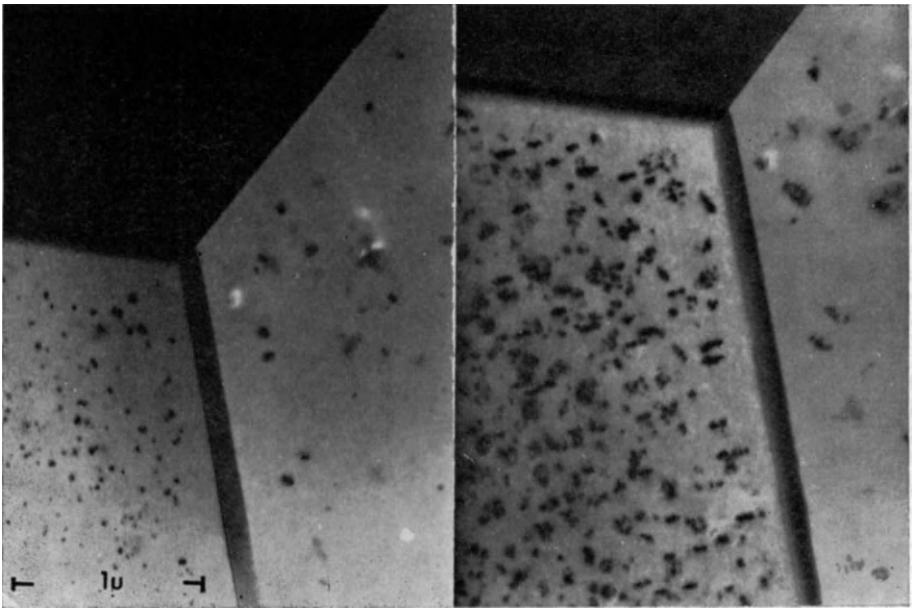


Fig. 22. – The growth of simple into complex loops at 600 kV after 10 (left) and 45 min (right) irradiation.

as the loop area, and, due to the Poissons ratio effect there are loop planes around the original loop and inclined to it on which precipitation of interstitials would relieve the stresses generated by the original loop. The effect probably occurs only under conditions of low temperature and rapid condensation rate,

since at high temperatures the rate of jog movement and climb is sufficient to ensure that all the defects assume the ultimately most stable configuration, *viz.* the planar loop.

The loop density changes with v_i in broadly the manner indicated by the theoretical model. The interstitial migration rate can be changed either by reducing the temperature ([11]) or by adding an impurity to trap the interstitials.

One further effect remains to be described in copper, namely the effect of electron channeling on the damage rate. From the dynamical theory of diffraction contrast it is predicted that the amount of inelastic scattering (atom displacement is one of the inelastic scattering mechanisms) will depend on the Bloch wave excitation. As a simple example let us consider the two beam case. From the dynamical theory the coefficients of the first and second Bloch waves are:

$$b^{(1)} \propto \sin \pi g r \quad \text{and} \quad b^{(2)} \propto \cos \pi g r \qquad (\text{at } s = 0)\,.$$

Since in reciprocal space $g = 1/d$, where d is the interplanar spacing of the reflecting planes:

$$b^{(1)} \propto \sin \frac{\pi r}{d} \quad \text{and} \quad b^{(2)} \propto \cos \frac{\pi r}{d}\,.$$

Hence wave 1) peaks when $r = d/2$ *i.e.* midway between the atom planes, and wave 2) peaks when $r = 0$ or d, *i.e.* at the atom planes ([12]). Wave 2) will therefore be more heavily scattered than wave 1), and hence should produce a larger number of displaced atoms.

The intensity of the electron current at the atom planes for the two waves is:

wave 1): $A^2 = \cos^2 \frac{\beta}{2} (1 - \sin \beta)\,,$

and

wave 2): $B^2 = \sin^2 \frac{\beta}{2} (1 + \sin \beta)\,.$

The total current at the planes is hence $A^2 + B^2 = 1 - \sin \beta \cos \beta$ where $\operatorname{ctg} \beta = w = s\xi_g$. The electron current at the planes is small when β is less than $\pi/2$ *i.e.* $s > 0$ and large when $\beta > \pi/2$, $s < 0$. Hence across a bend con-

tour one would expect a variation in damage rate because atomic displacement depends on the close approach of an electron to an atom, and hence the number of displacements should be considerably greater when wave 2) is strong. This effect has been observed in copper ([10]) for both the 220 and the 111 contours, and an example is shown in Fig. 23. It will be seen that the damage rate in the good transmission region is not zero, since dislocations in this region show clear signs of the absorption of point defects. The ratio of the expected damage rates in the symmetry position (parallel to the reflecting planes) and at an angle of 2θ from the symmetry position, where θ is the Bragg angle, is on the two-beam theory a factor of approxi-

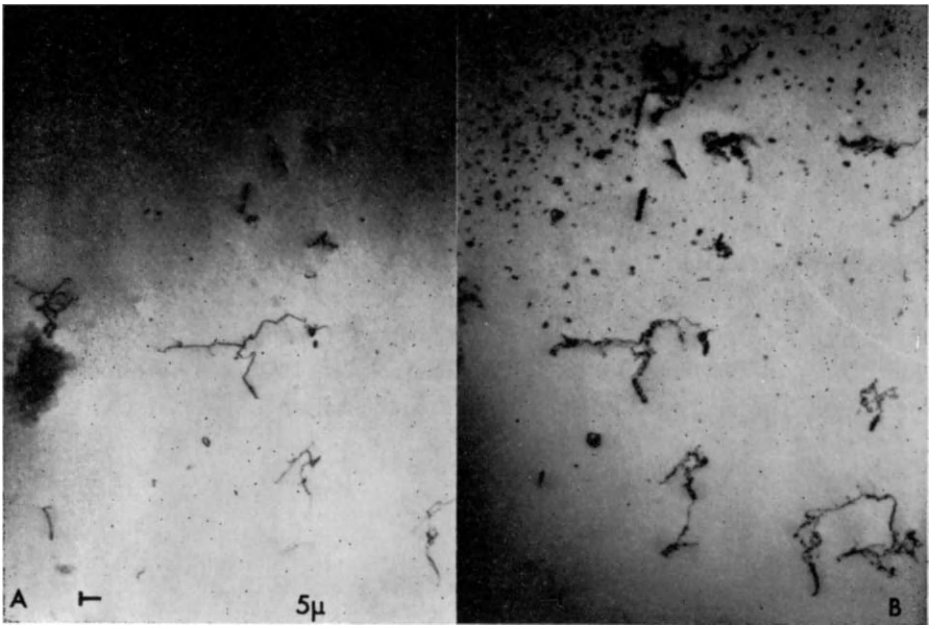


Fig. 23. – The variation in damage rate with orientation in copper. Area *A* shows the irradiation orientation and area *B* that loops are formed only in the 220 contour. Note, however, that the dislocations in the good transmission region become joggy.

mately two, using the parameters relevant to the 220 planes in copper at 600 kV. This is, of course, only a very approximate figure since a multiple beam calculation should be made and then the atom displacement process

is only one of several possible inelastic scattering processes. The effect of absorption *i.e.* the attenuation of the wave due to inelastic scattering, should also be included. Such calculations are in progress.

The effect of defects on the electron channeling should also be considered. If, for example, the channeling properties of a wave are significantly disturbed by a dislocation, such as a loop, then once loops are formed the difference in damage rate between the good channelling and poor channelling regions may become much less. Further work is required on all these effects.

Damage has been observed in various other materials, notably aluminium ([8]), iron ([13]) and graphite ([14]). In aluminium there is a threshold current below which damage is not observed at room temperature, presumably because the denuded zones occupy the whole foil thickness. In iron-carbon alloys it has been observed that small electron doses retard the precipitation of carbon in quenched alloys during subsequent ageing. Damage is observed in graphite only if the irradiation temperature is greater than about 350 °C. Presumably at lower temperatures the clusters are forming on so fine a scale that they are not visible in the microscope. Below the interstitial migration temperature clusters will not form at all, of course, except by « statistical » clustering at very high doses.

Electron irradiation in a high voltage microscope has also been observed to result in a loss of precipitate coherency, by the formation of dislocation loops around the precipitates ([15]). Presumably in this case the coherent precipitate strain can be relieved by the condensation of interstitials.

REFERENCES (Section 4)

1) M. J. Makin: *Phil. Mag.*, **20**, 1133 (1969).
2) W. A. McKinley and H. Feshbach: *Phys. Rev.*, **74**, 12 (1948).
3) O. S. Oen: ORNL-3813 (1965).
4) R. S. Nelson: AERE, R-6092 (1969).
5) G. J. Dienes and A. C. Damask: *Journ. Appl. Phys.*, **29**, 1713 (1958).
6) L. M. Brown, A. Kelly and R. M. Mayer: *Phil. Mag.*, **19**, 721 (1969).
7) L. M. Brown: *Phil. Mag.*, **19**, 869 (1969).
8) M. J. Makin: *Phil. Mag.*, **18**, 637 (1968).
9) M. Ipohorski and M. S. Spring: *Phil. Mag.*, **20**, 937 (1969).

10) M. J. MAKIN: *Proc. Int. Conf. on Atomic Collision in Solids, Brighton 1969*, North Holland (1970).
11) M. J. GORINGE and U. VALDRÈ: *Radiation Effects*, **1**, 133 (1969).
12) P. B. HIRSCH, A. HOWIE, R. B. NICHOLSON, D. W. PASHLEY, and M. J. WHELAN *Electron Microscopy of Thin Crystals*, Butterworths (1965).
13) L. E. THOMAS: *First National Conference on HVEM*, Monroeville (1969), and *Micron*, **1**, 251, (1969).
14) S. M. OHR: *First National Conference on HVEM*, Monroeville (1969).
15) G. R. WOOLHOUSE: *Nature*, **220**, 573 (1968).

Computing Methods

M. J. GORINGE

Department of Metallurgy, University of Oxford - Oxford, England

1. Introduction.

The problem facing the electron microscopist in the interpretation o micrographs of crystalline materials is, at the present time, similar to that faced by scientists in any field where diffraction plays an important role in the observations, namely that half the information available in the electron beam is usually lost in the recording process; phase relationships are removed as only intensities are recorded. Thus, as in X-ray crystallography, one of the most fruitful approaches over the last few years has been to work forward from « models » of defects, etc., present in the specimen to calculate the expected images; by comparison with the images actually obtained parameters in the model may be optimised to obtain a « best fit ». Provided the values of parameters obtained by this method are not physically unrealistic there is hope that the conclusions drawn regarding the nature of the defect are correct. Recent analysis (to be discussed in Sect. 7 below) has confirmed the uniqueness of images obtained in certain circumstances and indicated that in the future it may be possible to work « backwards » from a set of micrographs to a full description of the state of deformation of the specimen. Other possible ways of obtaining more information from micrographs than at present include holography and phase contrast microscopy, the latter of which is discussed in some detail by other contributors to this school (see this volume, Lenz's, Thon's and van Dorsten's lectures).

The main part of the discussion of computing methods which follows will be concerned with the calculation of images from models and their com-

parison with experiment. For completeness a number of formulae which have been dealt with by other contributors (see Howie's and Brown's lectures) will be quoted without proof, the notation used being predominantly that of Hirsch *et al.* [1], which should also be consulted for a fuller list of references to original papers.

2. Perfect crystals and faults: 2-beam approximation.

2·1. Perfect crystals.

In the 2-beam approximation we are concerned only with a beam emerging in the direction of the incident beam, amplitude φ_0, and a diffracted beam, amplitude φ_g, while inside the crystal only two Bloch waves, amplitudes $\psi^{(1)}$ and $\psi^{(2)}$ are excited. Under these conditions it is found that as a function of depth z in a perfect crystal

$$\left.\begin{aligned}\varphi_0(z) &= C_0^{(1)}\psi^{(1)}\exp\left[2\pi i\gamma^{(1)}z\right] + C_0^{(2)}\psi^{(2)}\exp\left[2\pi i\gamma^{(2)}z\right],\\ \varphi_g(z) &= C_g^{(1)}\psi^{(1)}\exp\left[2\pi i\gamma^{(1)}z\right] + C_g^{(2)}\psi^{(2)}\exp\left[2\pi i\gamma^{(2)}z\right],\end{aligned}\right\} \tag{1}$$

where $\gamma^{(i)} = k_z - k_z^{(i)}$ is the z-component of the wave vector difference, as shown in the dispersion surface of Fig. 1. $C_0^{(i)}$, $C_g^{(i)}$ and $\psi^{(i)}$ are related constants depending on the orientation of the crystal with respect to the incoming beam such that

$$\left.\begin{aligned} C_0^{(1)} &= C_g^{(2)} = \cos(\beta/2)\,, & C_0^{(2)} &= -C_g^{(1)} = \sin(\beta/2),\\ \psi^{(1)} &= \cos(\beta/2)\,, & \psi^{(2)} &= \sin(\beta/2),\end{aligned}\right\} \tag{2}$$

and

$$w = s\xi_g = \operatorname{ctg}\beta \tag{3}$$

where ξ_g is the extinction distance $(= 1/|\gamma^{(2)} - \gamma^{(1)}|)$ and s the excitation parameter (Fig. 1).

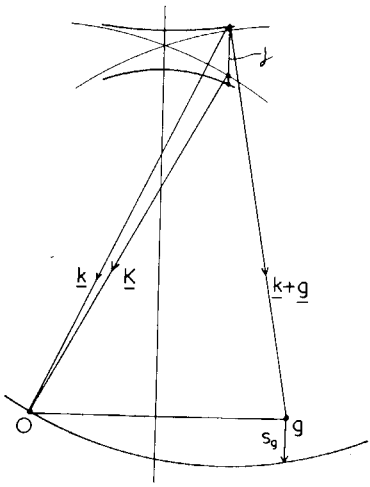


Fig. 1. – Reflecting sphere construction and dispersion surfaces.

In general $\gamma^{(i)}$ is complex, and thus eqs (1) are complex and the amplitudes of the beams at the bottom of a crystal of thickness z are

$$\left.\begin{aligned}\varphi_0(z) &= \{\cos^2(\beta/2)\exp[-iXz] + \sin^2(\beta/2)\exp[iXz]\}\exp[-\pi z/\xi_0'],\\ \varphi_g(z) &= -\cos(\beta/2)\sin(\beta/2)\{\exp[-iXz] - \exp[iXz]\}\exp[-\pi z/\xi_0'],\end{aligned}\right\} \quad (4)$$

where

$$X = \frac{\pi\sqrt{1+w^2}}{\xi_g} + \frac{\pi i}{\xi_g'\sqrt{1+w^2}},$$

ξ_0' and ξ_g' being the absorption parameters.

Calculation of bright field $(\varphi_0^*(z)\varphi_0(z))$ and dark field $(\varphi_g^*(z)\varphi_g(z))$ intensities for various values of ξ_g, ξ_0', ξ_g', w (hence β) and z enable comparison

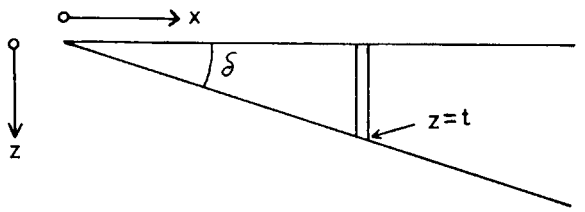


Fig. 2. – Co-ordinates used to describe a tapered foil.

to be made with experimental micrographs to deduce suitable values for these basic properties of the material. Thus, in principle, it is possible from measurements on a tapered crystal of known, constant orientation (w constant) to deduce values for ξ_g, ξ_0' and ξ_g', provided that the wedge angle, δ, is known (Fig. 2), and that the column approximation is assumed to be valid. Similar information is available from a bent crystal of known uniform thickness ($z = t$ constant), and even if absolute intensities are not available in this case information is only lost on the value of ξ_0'. (An example of a combination of the two situations will be discussed in Subsect. 5·2 below.)

2·2. Scattering matrix for perfect crystals.

For computational convenience, which will become more apparent later, eqs (1) are often written in matrix form

$$\begin{pmatrix}\varphi_0(z)\\ \varphi_g(z)\end{pmatrix} = \begin{pmatrix}C_0^{(1)} & C_0^{(2)}\\ C_g^{(1)} & C_g^{(2)}\end{pmatrix}\begin{pmatrix}\exp\left[2\pi i\gamma^{(1)}z\right] & 0\\ 0 & \exp\left[2\pi i\gamma^{(2)}z\right]\end{pmatrix}\begin{pmatrix}\psi^{(1)}\\ \psi^{(2)}\end{pmatrix}. \tag{5}$$

The « wave-matching » which defines the C's and ψ's may be written

$$\begin{pmatrix}C_0^{(1)} & C_0^{(2)}\\ C_g^{(1)} & C_g^{(2)}\end{pmatrix}\begin{pmatrix}\psi^{(1)}\\ \psi^{(2)}\end{pmatrix} = \begin{pmatrix}\varphi_0(0)\\ \varphi_g(0)\end{pmatrix}. \tag{6}$$

In matrix notation premultiplying eq. (6) by $\mathbf{C}^{-1}$ and substituting for $\boldsymbol{\psi}$ in eq. (5) yields the complex matrix equation

$$\boldsymbol{\varphi}(z) = \mathbf{C}\{\exp\left[2\pi i\gamma z\right]\}\mathbf{C}^{-1}\boldsymbol{\varphi}(0), \tag{7}$$

where $\{\}$ indicates a diagonal matrix. The matrix

$$\mathbf{P} = \mathbf{C}\{\exp\left[2\pi i\gamma z\right]\}\mathbf{C}^{-1}, \tag{8}$$

which relates the incident amplitudes $\boldsymbol{\varphi}(0)$ and the resultant amplitudes $\boldsymbol{\varphi}(z)$ is termed the scattering matrix, and in this case is 2×2 complex.

As a trivial example it may be confirmed (after some algebra) that if a single slab of perfect crystal is split into two slabs by an imaginary fault plane, which does not affect the beams (Fig. 3) then

$$\boldsymbol{\varphi}(t_1 + t_2) = \mathbf{P}_2\mathbf{P}_1\boldsymbol{\varphi}(0). \tag{9}$$

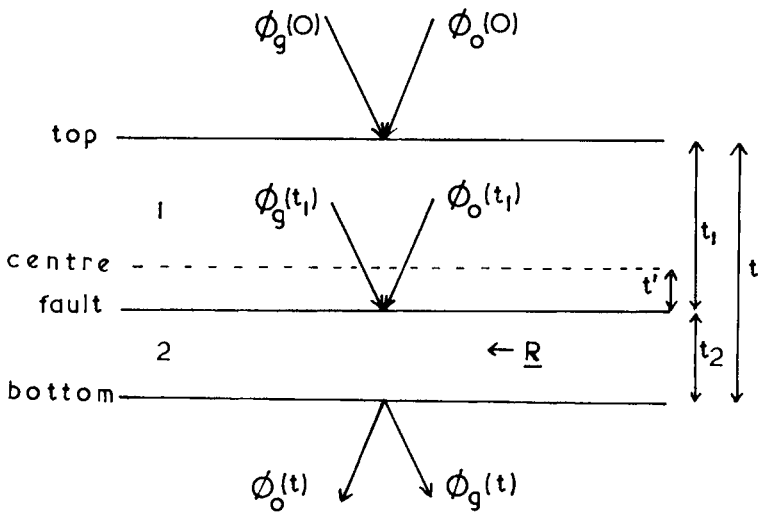


Fig. 3. – Waves φ_0, φ_g propagating through a composite slab crystal.

2·3. **Faulted crystals by scattering matrices.**

In this case the fault plane of Fig. 3 is assumed to introduce a phase change $\alpha = 2\pi \boldsymbol{g}\cdot\boldsymbol{R}$, where $\boldsymbol{R}$ is the displacement of the lower slab of crystal. In scattering matrix terms the effect of the fault may be summarised as follows. Defining

$$\boldsymbol{F}^{+} = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\alpha} \end{pmatrix}, \tag{10}$$

the effect of the two slabs of crystal and the interfacial fault is given by

$$\boldsymbol{\varphi}(t_1 + t_2) = \boldsymbol{F}^{-}\boldsymbol{P}_2\boldsymbol{F}^{+}\boldsymbol{P}_1\boldsymbol{\varphi}(0), \tag{11}$$

where $\boldsymbol{F}^{-}$ is similar to $\boldsymbol{F}^{+}$ except that α is replaced by $-\alpha$.

For a number of slabs of crystal (see Fig. 4) the result is

$$\boldsymbol{\varphi}(t) = \boldsymbol{F}_n^{-}\boldsymbol{P}_n\boldsymbol{F}_n^{+}\boldsymbol{F}_{n-1}^{-}\boldsymbol{P}_{n-1}\ldots\boldsymbol{F}_3^{+}\boldsymbol{F}_2^{-}\boldsymbol{P}_2\boldsymbol{F}_2^{+}\boldsymbol{P}_1\boldsymbol{\varphi}(0), \tag{12}$$

where $\boldsymbol{F}_j$ is defined as in eq. (10), α_j being the phase change as measured with respect to the first slab. For computational convenience the product matrices

$\boldsymbol{F}_j^+ \boldsymbol{F}_{j-1}^-$ may be written as

$$\boldsymbol{F}_j^+ \boldsymbol{F}_{j-1}^- = \boldsymbol{F}_{jk}, \qquad (k = j-1), \tag{13}$$

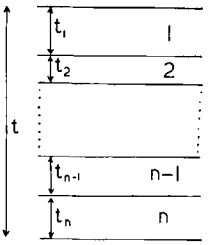


Fig. 4. – Diagram of foil made up of slabs of perfect material separated by faults.

where $\boldsymbol{F}_{jk}$ is of the same form as eq. (10), α_{jk} being the phase difference between slabs j and $j-1$. Thus

$$\boldsymbol{\varphi}(t) = \boldsymbol{F}_n^- \boldsymbol{P}_n \boldsymbol{F}_{n,n-1} \boldsymbol{P}_{n-1} \ldots \boldsymbol{F}_{32} \boldsymbol{P}_2 \boldsymbol{F}_{21} \boldsymbol{P}_1 \boldsymbol{\varphi}(0) \tag{14}$$

and if the phase relationship between φ_0 and φ_g is unimportant (*e.g.* if only intensities are required) the matrix $\boldsymbol{F}_n^-$ may be omitted. Neglecting $\boldsymbol{F}_n^-$ in eq. (14) it may be seen that the effect of each fault is to introduce a phase change in the diffracted beam only. It should be noted that eq. (14) applies to two situations of considerable interest i) stacking fault separated by slabs of perfect crystal all with the same orientation and ii) coherent twin boundaries where there is no phase change at the « fault » plane (*i.e.* $\boldsymbol{F}_{jk} = \boldsymbol{I}$ all j, where $\boldsymbol{I}$ is the identity matrix), but where the slabs of perfect crystal are differently oriented (*i.e.* the values of $\boldsymbol{C}_j$ comprising $\boldsymbol{P}_j$ are different; eq. (8)), or, of course, to any combination of i) and ii).

The situation for which it is usually required to calculate image intensities is that of faults which are inclined in the foil (Fig. 5), when the components

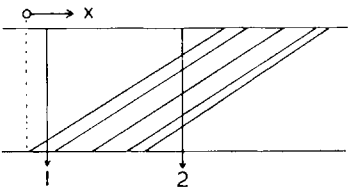


Fig. 5. – Diagram of foil containing a number of overlapping faults on inclined planes. Note that columns 1 and 2 pass through different numbers of faults.

of the matrices $\boldsymbol{P}_j$ depend on the position x of the column considered, as does the number of faults encountered. In general, in electronic computation, a subroutine is set up to calculate $\boldsymbol{\varphi}'$ from $\boldsymbol{A}$ and $\boldsymbol{\varphi}$, where

$$\boldsymbol{\varphi}' = \boldsymbol{A}\boldsymbol{\varphi} . \tag{15}$$

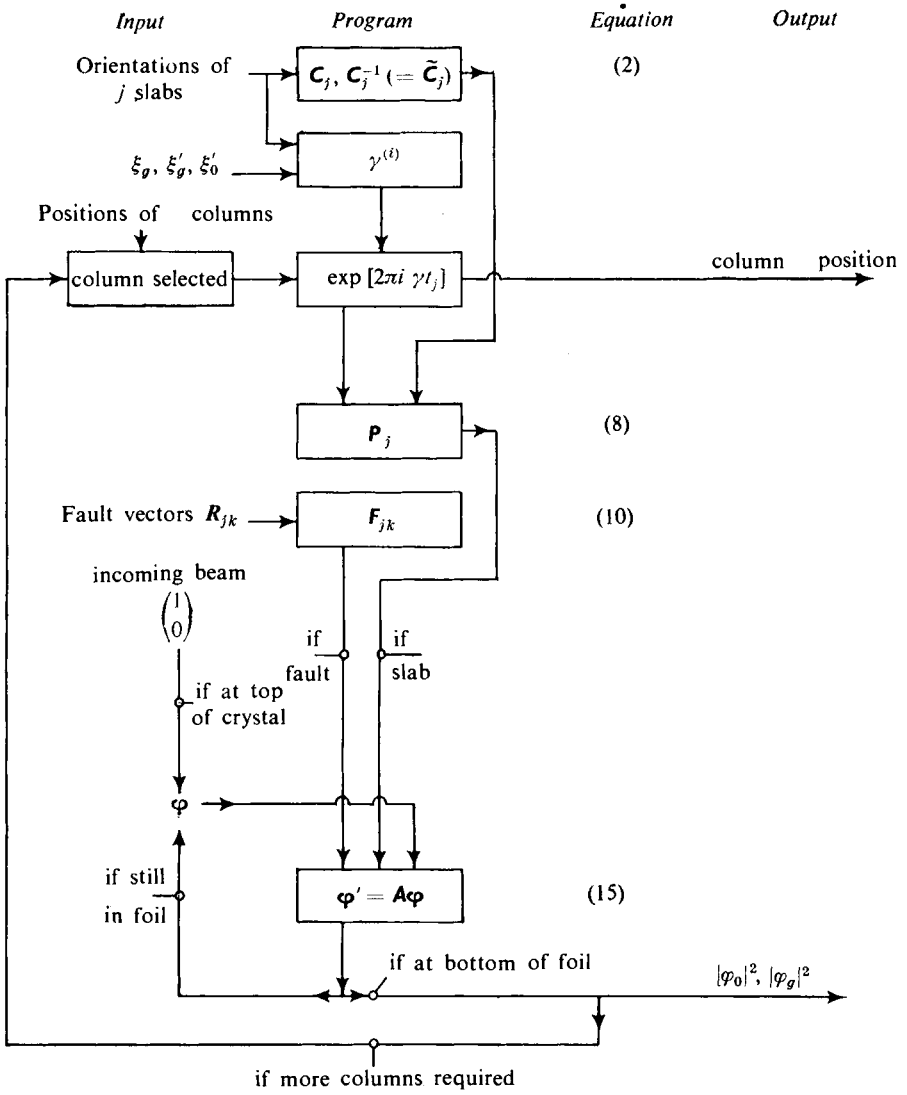


Fig. 6. – Schematic program to calculate 2-beam intensities from columns passing through a number of slabs of perfect crystal separated by faults.

$\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}'$ are 2×1 complex vectors describing the beams in the crystal before and after the operation respectively and the 2×2 complex matrix $\boldsymbol{A}$ is alternately of type $\boldsymbol{P}$ and $\boldsymbol{F}$. The subroutine is used until the bottom of the crystal is reached, when $\boldsymbol{\varphi}'$ is the required solution for the amplitudes. An outline diagram of the computational method is shown in Fig. 6. (In certain computing languages, *e.g.* Fortran, complex algebra is available as a standard facility, but in most cases it is necessary to write the matrix multiplication explicitly in terms of its real and imaginary parts.)

In simple cases it is often preferable to calculate from algebraic expressions for φ_0 and φ_g (such as may be obtained by multiplying out eq. (14) by hand) *e.g.* a single stacking fault (Hirsch *et al.* ([1]) eq. (10.10)), a single coherent twin boundary (Gevers *et al.* ([2,3])), a pair of stacking faults (Gevers ([4])) or a pair of coherent twin boundaries (Remaut *et al.* ([5])). However, in the last two cases other authors prefer digital calculation from matrix notation (*e.g.* Hirsch *et al.*, p. 226 ([1]), Goringe and Valdrè ([6])).

2·4. Moiré fringes.

Moiré fringes occur (in the simplest case) when two slabs of perfect crystal (*e.g.* as in Fig. 3) have slightly different lattice parameters (parallel moiré), or have the same lattice parameter but are rotated with respect to each other about an axis perpendicular to the « fault » plane (rotation moiré). Under these conditions eq. (11) holds but now the phase angle α is a function of position. An example of a moiré pattern is given in *Problem 14* of Goringe and Hall: « Typical Problems ... » in this volume.

2·5. Lattice fringes.

So-called « lattice fringes » occur when more than one beam is allowed to contribute to the image and recently considerable experimental effort has been expended by manufacturers to achieve fringes with the closest possible spacing to prove the superiority of resolution of their particular microscope. There is also considerable interest in seeing very fine detail in lattice images and to relate this to atomic positions near defects, *e.g.* around the core of a dislocation. However, as the image is strictly an interference pattern the correspondence between image fringes and lattice planes is not direct (see *Problem 15* of « Typical Problems ... », this volume) and considerable care must be taken in interpreting extra fringes as «extra half planes », etc ... (Cockayne ([7])).

3. Perfect crystals and faults: *n*-beam.

3·1. Wave matching.

The *n*-beam form of the wave matching calculation summarised by eqs (2) in the 2-beam case now becomes

$$\mathbf{A}C^{(i)} = \gamma^{(i)} C^{(i)}, \qquad i = 1, 2 \ldots n, \tag{16}$$

where $\mathbf{A}$ is an $n \times n$ complex matrix, $C^{(i)}$ is a column vector whose elements $C_g^{(i)}$ are the amplitudes of the i-th Bloch wave, and $\gamma^{(i)}$ is the corresponding value of γ (see Fig. 1). The matrix $\mathbf{A}$ contains the information on the periodic lattice potential (expressed here in terms of extinction and absorption distances) in its off-diagonal elements and the crystal orientation and mean absorption on its diagonal

$$A_{00} = \frac{i}{2\xi_0'}, \qquad A_{gg} = s_g + \frac{i}{2\xi_0'}, \qquad A_{gh} = \frac{1}{2\xi_{g-h}} + \frac{i}{2\xi_{g-h}'}. \tag{17}$$

However, complex matrix equations of the form of eq. (16) are time-consuming to calculate even on an electronic computer, and it is usually sufficiently accurate to compute only the real parts of eq. (16), *i.e.*

$$\mathbf{A}_r C^{(i)} = \gamma_r^{(i)} C^{(i)}, \qquad i = 1, 2 \ldots n, \tag{18}$$

and use 1st order perturbation theory to calculate $\gamma_{im}^{(i)}$ from

$$\gamma_{im}^{(i)} = \tilde{C}^{(i)} \mathbf{A}_{im} C^{(i)}, \qquad i = 1, 2 \ldots n, \tag{19}$$

where $\tilde{C}^{(i)}$ is the transpose of $C^{(i)}$.

If the excitations of the Bloch waves at the top of the crystal are denoted by the vector ψ and are produced by incoming plane waves φ then the matching condition is

$$\mathbf{C}\psi = \varphi. \tag{20}$$

However, as in the 2-beam case (eqs (2)) it can be shown that $\psi^{(i)} = C_0^{(i)}$. An example of a 4-beam wave matching calculation is given in *Problem 7* of « Typical Problems ... » in this volume.

3·2. Perfect crystals.

The i-th Bloch wave of initial amplitude $\psi^{(i)}$ propagates through the crystal with wave vector $\gamma(=\gamma_r + i\gamma_{im})$ in the same way as in the 2-beam case. Thus the scattering matrix formulation of eqs (7) and (8) holds; in

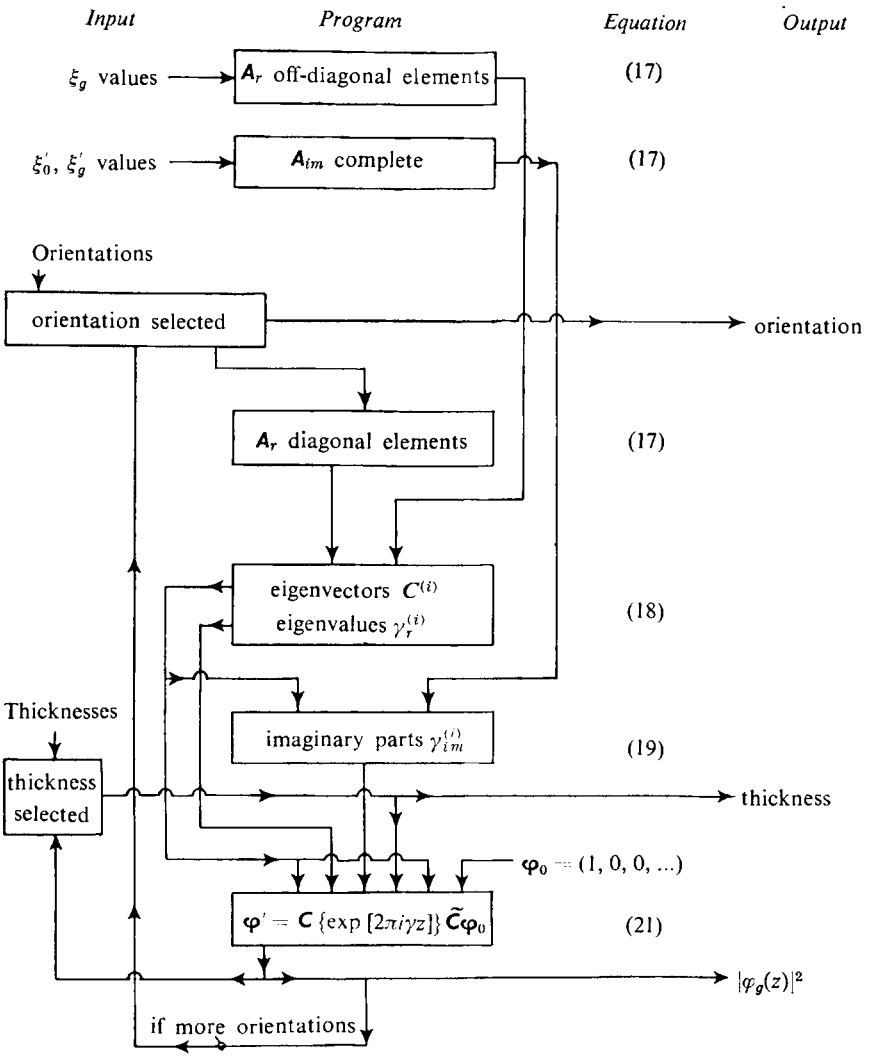


Fig. 7. – Schematic program to calculate n-beam intensities from perfect crystals.

particular the wave amplitudes $\varphi_g(z)$ at depth z for a single incident wave of unit amplitude are given by

$$\varphi_g(z) = \sum_i C_0^{(i)} C_g^{(i)} \exp [2\pi i \gamma^{(i)} z] \tag{21}$$

(This result follows immediately from eq. (7) using the fact that $\mathbf{C}^{-1} = \tilde{\mathbf{C}}$.)

Thus the problem of calculating the wave amplitudes leaving the bottom of a slab of perfect crystal consists of i) setting up the matrices $\mathbf{A}_r$ and $\mathbf{A}_{tm}$ for the particular orientation and reflexions of interest, ii) calculating the eigenvectors $\boldsymbol{C}^{(i)}$ and corresponding eigenvalues $\gamma_r^{(i)}$ of $\mathbf{A}_r$ by one of the standard subtroutines usually available in computer libraries (*e.g.* Householder's and Jacobi's method) iii) calculating the imaginary parts of γ by matrix multiplication (eq. (19)) iv) calculating the set of emergent beams φ_g for the various values of z which are of interest, and hence beam intensities, $|\varphi_g|^2$. The same set of calculations may then be carried out for the next orientation of interest except that in i) only the diagonal elements of $\mathbf{A}_r$ need to be recalculated. Stage ii) occupies by far the largest amount of computer time for large values of n, hence the necessity to change orientation the smallest possible number of times during a set of calculations. An outline of the calculation process is shown in Fig. 7.

3·3. **Planar faults.**

The equations used in the 2-beam situation may be carried over without modification (except increase in size). The beams at the lower face of the crystal are described by eq. (14) with the fault matrices of eq. (10) being generalised to

$$\mathbf{F}_{jk} = \{\exp [i\alpha]\}_{jk} . \tag{22}$$

The method of calculation is similar to that shown in Fig. 2 for the 2-beam case, with the calculation of $\mathbf{C}_j$ and $\gamma^{(i)}$ being replaced by the equivalent sections of Fig. 7.

Other cases (*e.g.* lattice fringes, moiré patterns, etc.) follow in a similar way by analogy with the 2-beam situation.

3·4. Modified extinction distances.

In many cases, however, (*e.g.* in the case of systematic reflections) we are concerned only with the perturbing effect of weakly excited beams on the two principal beams φ_0 and φ_g. Under these conditions only two Bloch waves (i and j) are strongly excited and the main features of the contrast are governed by their interference with each other. Thus we have a « principal » extinction distance $\xi(=1/(\gamma^{(i)}-\gamma^{(j)}))$ which may be used in place of ξ_g in 2-*beam* calculations with a consequent saving in computing time. The value of ξ may be calculated by the n-beam theory of eq. (18) and Fig. 7, or under certain circumstances analytically. An example of such an analytical calculation is given by Howie ([8]) for the 4-beam symmetrical systematic situation with the incident beam at the Bragg angle for the lowest order reflection g (see Fig. 8), the result being

$$2k\Delta k = U_1 + U_3 + \{(g^2 + (U_1 - U_3)/2)^2 + (U_1 + U_2)^2\}^{\frac{1}{2}} - \\ - \{(g^2 - (U_1 - U_3)/2)^2 + (U_1 - U_2)^2\}^{\frac{1}{2}}, \qquad (23)$$

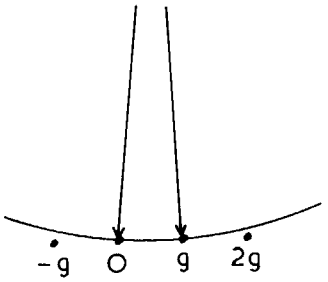


Fig. 8. – Reflecting sphere construction for Howie's 4-beam formula for extinction distance ξ.

where $\xi = 1/\Delta k$, $\xi_{g-h} = K/U_j$, $(j = |g-h|)$, U_j and K being the modified lattice potential and incident electron wave vector respectively. Equation (23) is expected to hold reasonably well when $U_2 > U_3$ and $g^2 \gtrsim |U_g|$, *i.e.* structures with uniform structure factors at not too high energies (U_g increases with energy), *e.g.* for 100 keV electrons with g=111 equation (23) gives values of $\xi = 507$ Å and 127 Å for Al and Au respectively, while ten-beam theory from eq. (18) (see Fig. 9 for typical graph) yields $\xi = 503$ Å and 117 Å,

using ξ_{111} values of 556 Å and 159 Å (and suitable values for the higher orders required). It should be noted that the diffraction conditions envisaged here are the nearest possible to 2-beam; the systematic reflections cannot be avoided by suitable tilting.

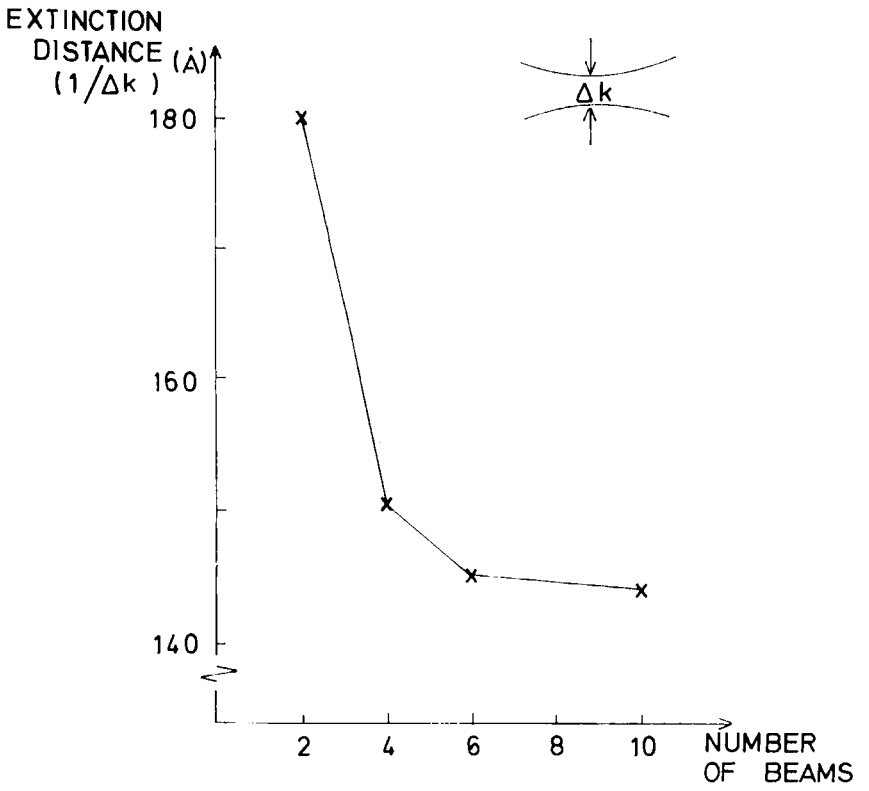


Fig. 9. – Graph showing the variation in the 200 extinction distance ($1/\Delta K$) of gold as a function of the number of systematic beams included in the calculation. 100 keV electrons. (D. J. H. Cockayne.)

The graph shown in Fig. 9 is a typical result for the effect of systematic reflexions on strong low-order principal reflections. The decrease in ξ with increasing number of beams is not, however, perfectly general; the reverse may occur for higher order reflections or in structures with varying structure factors, *e.g.* silicon $g = 111$, 2-beam $\xi_g = 605$ Å, while the many-beam systematic value is 617 Å (Booker ([9])).

4. Imperfect crystals.

4·1. φ_0, φ_g formulation (2-beam).

Here it is convenient to turn to the historically earlier description of the dynamical theory: the wave-optical formulation. Here the basic 2-beam equations for the imperfect crystal are expressed as a pair of complex first order differential equations for propagation of waves in the z-direction of the form

$$\left.\begin{aligned}\frac{\mathrm{d}\varphi_0}{\mathrm{d}z} &= -\frac{\pi}{\xi_0'}\varphi_0 + \pi\left(\frac{i}{\xi_g} - \frac{1}{\xi_g'}\right)\varphi_g,\\ \frac{\mathrm{d}\varphi_g}{\mathrm{d}z} &= \pi\left(\frac{i}{\xi_g} - \frac{1}{\xi_g'}\right)\varphi_0 + \left(-\frac{\pi}{\xi_0'} + 2\pi i(s+\beta_g')\right)\varphi_g,\end{aligned}\right\} \tag{24}$$

where

$$\beta_g' = \boldsymbol{g}\cdot\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}z}\,. \tag{25}$$

$\boldsymbol{R}\ (x, y, z)$ is the local displacement from the ideal position. If β_g' is everywhere zero we have perfect crystal, and the analytical solutions of eqs (4) are produced. Thus, provided β_g' is known for all points in the foil eqs (24) can, in principle, be integrated numerically from the initial conditions at the top of the foil ($\varphi_0 = 1$, $\varphi_g = 0$) through to the bottom of the foil. Methods of calculation of β_g' for several cases are discussed elsewhere (Brown, this volume).

Now numerical integration of this kind is only feasible using a computer, the eqs (24) being of the standard form

$$\frac{\mathrm{d}\boldsymbol{\varphi}}{\mathrm{d}z} = f(\boldsymbol{\varphi})\,, \tag{26}$$

suitable for integration by processes such as Runge-Kutta or Nordsieck ([10]) which are usually available as standard library routines. In all cases the integration is carried out by an integration routine which calculates $\boldsymbol{\varphi}(z+h)$ given $\boldsymbol{\varphi}(z)$, where h is termed the external step length. This routine (see Fig. 10) must have available to it a subroutine for calculating $f(\boldsymbol{\varphi})$, which, in turn,

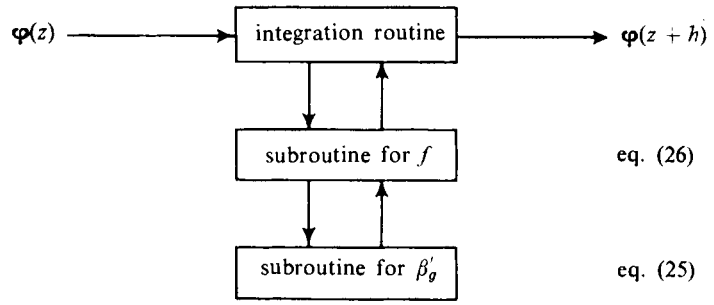


Fig. 10. – Schematic program showing operation of integration routine.

includes calculation of β'_g, best carried out in another subroutine. This second subroutine will be particular to individual problems while the former two are applicable to problems in general.

The integration routine often integrates numerically to a specified accuracy, subdividing the external steplength h into internal steplengths h_i in order to achieve this. An example of the action of such a self-adjusting routine is shown in Fig. 11 where the regions of foil integrated (by a Nordsieck routine,

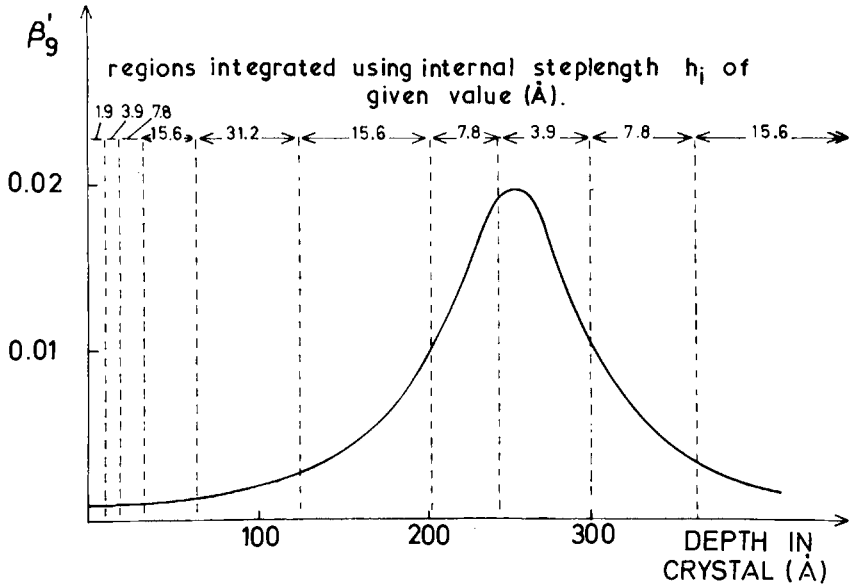
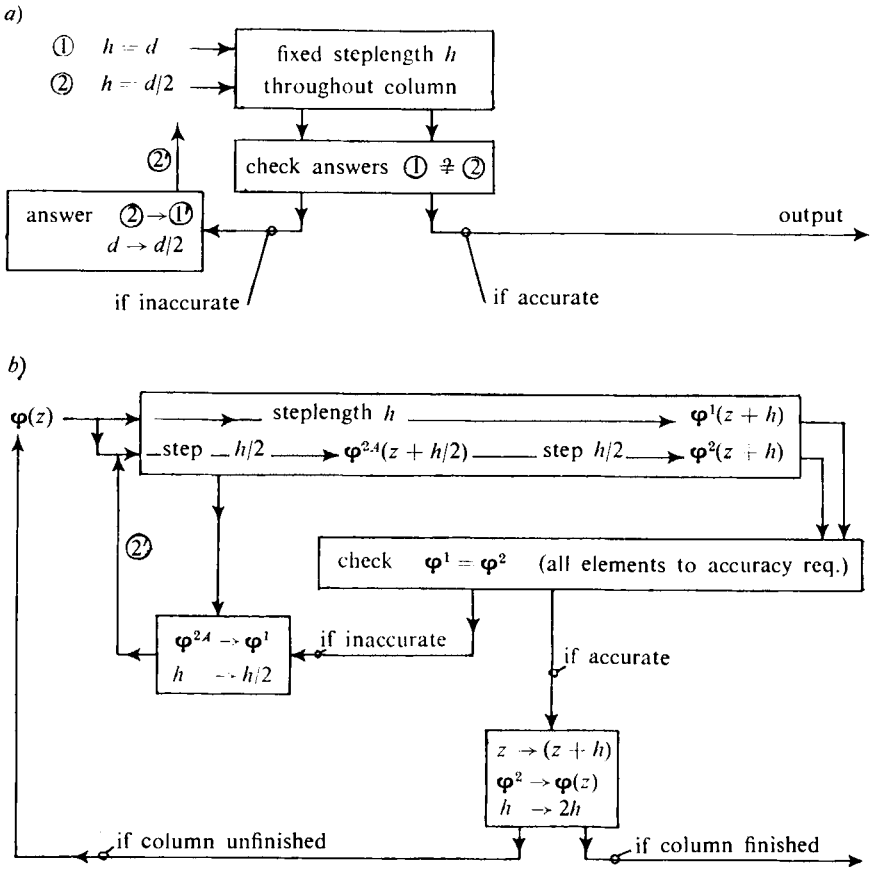


Fig. 11. – Example of steplength taken by integration routine of Nordsieck for a screw dislocation. (D. J. H. Cockayne.)

actually integrating equations in the Bloch wave formulation; see Subsect. 4·3 below) using h_i of the given value are delineated by dotted lines and β_g' by the full line. During the initiation of the routine h_i is very small, but after a few steps the numerical process is changed and while β_g' is small h_i increases. As β_g' increases h_i must be decreased to maintain accuracy, increasing again after the peak in β_g' has been passed. As the computer time used is directly proportional to the number of internal steps required it is obviously in the interests of efficiency for h_i to be adjusted in some manner, and before self-adjusting routines become available many ingenious devices were used ex-



c) Calculate $(\beta' + s)$ in advance of each use of routine for $\varphi(z) \to \varphi(z + h)$ and adjust h according to « rules » determined by experience (*e.g.* by method *a*)) to be appropriate.

Fig. 12. – Schematic « nonautomatic » accuracy checks on numerical integration routines.

ternally to achieve the same effect (*e.g.* those outlined in Fig. 12). It is also apparent that these numerical methods are only feasible where β_g' is continuous and a reasonably small quantity, certainly not for columns very close to the core of a dislocation, for example, where β_g' might become infinite; nor for faults (see Subsect. **2**·3).

A basic disadvantage of the φ_0, φ_g formulation closely related to the variation in computer time with β_g' is that eqs (24) predict appreciable values of $f(\boldsymbol{\varphi})$ (eq. (26)) even in a perfect crystal, particularly if the crystal is not exactly at the reflecting position ($s = 0$). Thus computer time is wasted changing φ_0 into φ_g and *vice versa* in a manner which has previously been discussed analytically (eqs (4)). This disadvantage may be overcome if Bloch waves are used, (as discussed in Subsect. **4**·3 below) the time required to form the Bloch waves at the top of the crystal from the incoming beam and to calculate the emitted beams from them at the bottom being more than compensated for by the quicker integration, particularly in thick foils or when there are large deviations from the reflecting position.

4·2. $\boldsymbol{\varphi}$ formulation (n-beam).

It can be shown that eqs (24) may be generalised to the form

$$\frac{\mathrm{d}}{\mathrm{d}z}\boldsymbol{\varphi}(z) = 2\pi i[\mathbf{A}(z) + \{\beta_g'(z)\}]\boldsymbol{\varphi}(z)\,, \tag{27}$$

where $\mathbf{A}(z) = \mathbf{A} + \Delta\mathbf{A}(z)$, $\mathbf{A}$ and $\{\beta_g'(z)\}$ being defined by eqs (17) and (25) respectively. $\Delta\mathbf{A}(z)$ is a matrix with zero diagonal elements, its off-diagonal elements allowing the possibility of variations in ξ_g with depth. Thus the techniques described in Subsect. **4**·1 may be carried over directly to the n-beam case. However, the disadvantages of the time wasted by integrating unnecessary changes are more acute than in the 2-beam case, all the beams oscillating with depth and s_g unlikely to be small for all g. Hence the Bloch wave formulation is definitely to be preferred in this case.

4·3. Bloch wave formulation.

If the crystal with continuously varying strain field is imagined as an assembly of very thin slabs then the scattering matrix approach of Subsect. **2**·3 (eq. (14)) and the basic relation between Bloch waves and diffracted beams

(eq. (5)) generalised to the n-beam case combine to give

$$\boldsymbol{\varphi}(z) = \{Q(z)\}^{-1}\mathbf{C}\{\exp[2\pi i\gamma z]\}\boldsymbol{\psi}(z), \tag{28}$$

where $\{Q(z)\}^{-1}$ is equivalent to $\boldsymbol{F}_n^-$ in eq. (14). Differentiating eq. (28), substituting for $\boldsymbol{\varphi}$ and $\mathrm{d}\boldsymbol{\varphi}/\mathrm{d}z$ in eq. (27), using the fact that

$$\boldsymbol{A}\mathbf{C} = \mathbf{C}\{\gamma\} \tag{29}$$

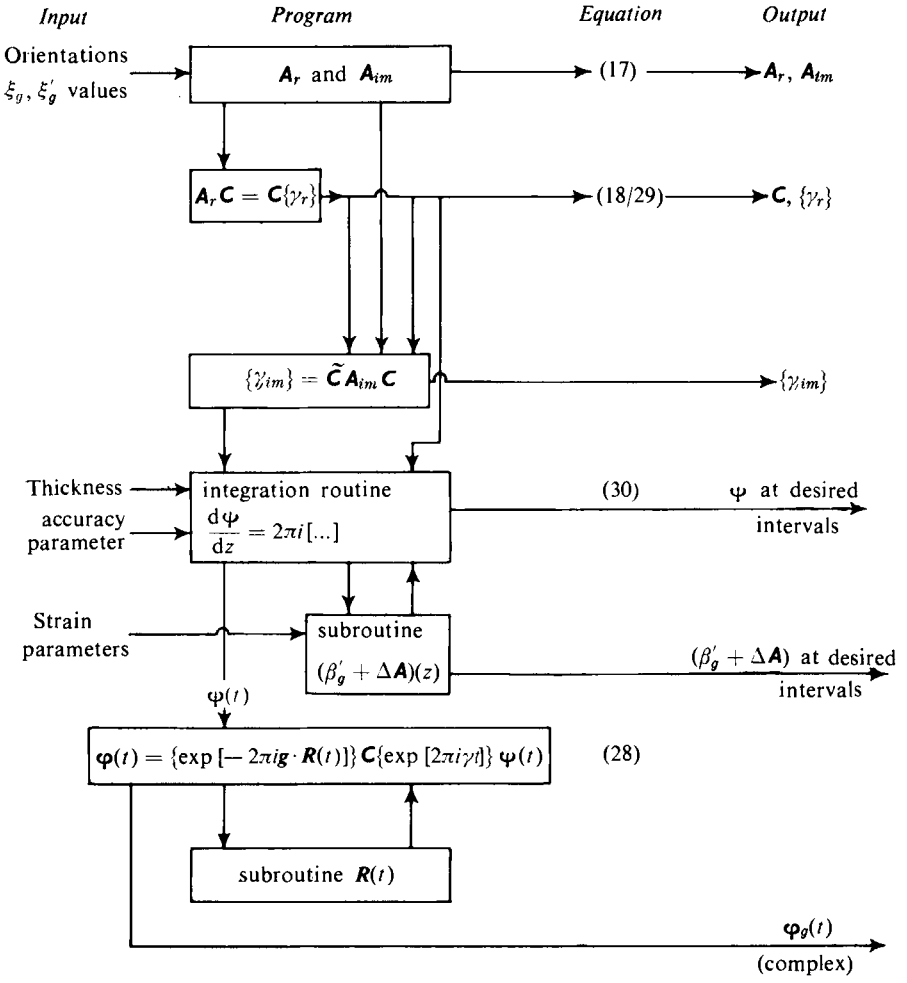


Fig. 13. – Schematic program to calculate n-beam solution for continuously varying strain fields. (D. J. H. Cockayne.)

(equation (16) in full matrix form) and rearranging yields

$$\frac{\mathrm{d}\Psi}{\mathrm{d}z} = 2\pi i \{\exp[-2\pi i\gamma z]\}\mathbf{C}^{-1}[\Delta\boldsymbol{A}(z) + \{\beta'_g(z)\}]\mathbf{C}\{\exp[2\pi i\gamma z]\}\Psi(z). \qquad (30)$$

Equation (30), given by Cockayne (7), is a more general equation for Bloch wave scattering than that normally derived (Hirsch *et al.* (1) p. 291), including in addition to $\{\beta'_g(z)\}$, the possibility of changes of extinction distance (caused by replacement of parts of the crystal by different (but similar) material) in the term $\Delta\boldsymbol{A}(z)$.

The method of calculation in this case is a combination of the n-beam perfect crystal calculations of Subsect. 3·1 and the numerical integration of eq. (30), which is of the same form as eq. (26), by the techniques described for the φ_0, φ_g case in Subsect. 4·1. If the relative phases of the waves are unimportant then the factor $\{Q(z)\}^{-1}$ may be omitted from eq. (28) for simplicity. However, in the flow diagram of a general Bloch wave computer programme shown in Fig. 13, it has been included, which enables relative phases as well as amplitudes to be obtained as output, *e.g.* for lattice fringe calculations.

5. Comparison of model calculations with micrographs.

5·1. Line profiles.

The traditional method of comparison of the results of calculations on a proposed model with actual micrographs is by comparing line profiles (*i.e.* calculations along straight lines in suitable orientations; for typical examples see Hirsch *et al.* (1)) with microdensitometer traces from experimental micrographs. For many purposes this is quite adequate, particularly if only qualitative comparison is required. In principle, line traces contain sufficient information in a quantitative sense as well, and in certain cases where intensity variations only occur in one direction (*e.g.* inclined planar faults, wedge-shaped crystals, dislocations lying parallel to the foil) this information is complete. Where there are intensity variations in two dimensions, *e.g.* dislocations threading the specimen from top to bottom, a number of traces is required to detect the characteristic « zig-zag » or « spotty » contrast (see Fig. 14 and 15).

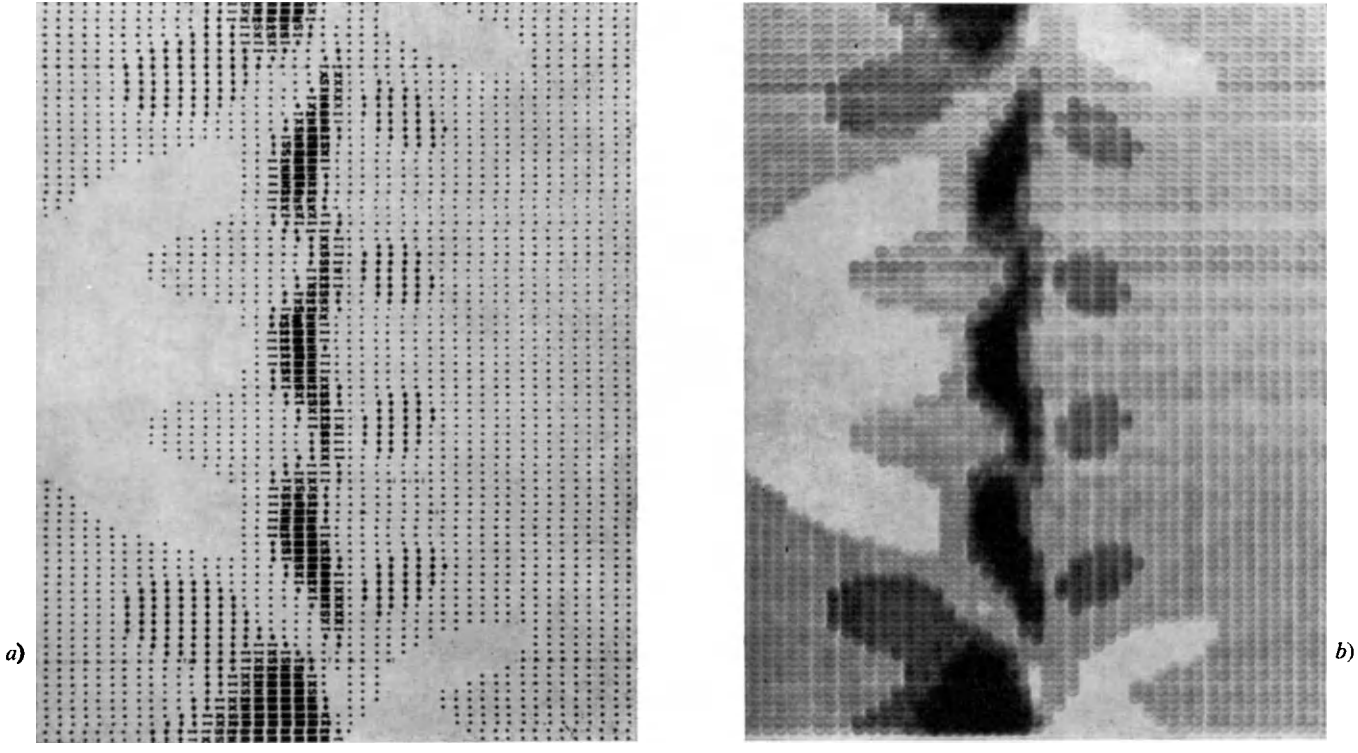


Fig. 14. – Overprinted line printer output simulating image of a dislocation in Fe-34Al calculated using anisotropic elasticity theory; *a*) print with characters used to produce grey scale visible, *b*) print out of focus to simulate a micrograph more closely. (R. C. Crawford.)

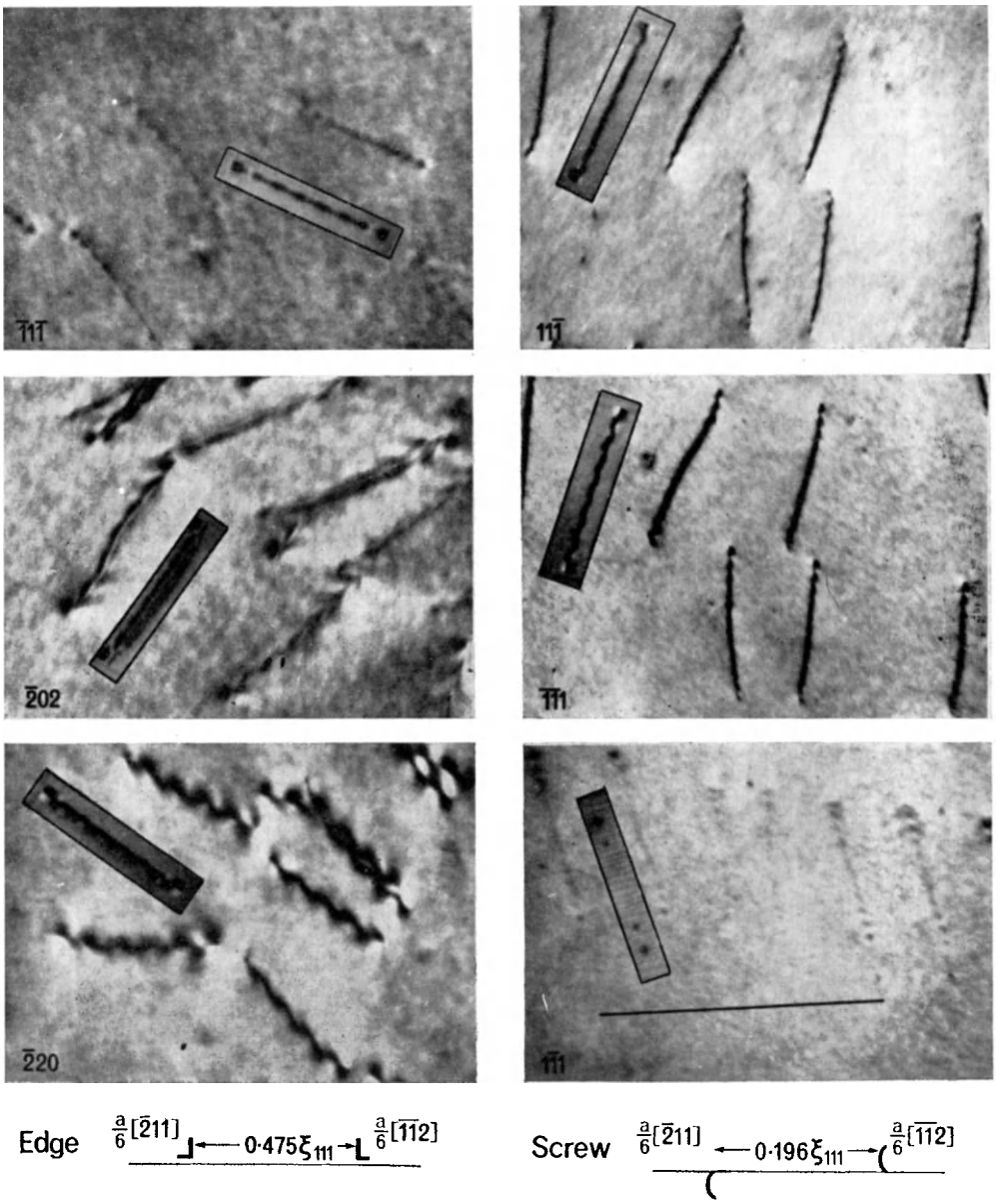


Fig. 15. – Micrographs of dislocations in Cu 10 at % Al with dislayed calculations (outlined). The figures on the micrographs give the operating reflection. (P. M. Hazzledine, H. P. Karnthaler and M. S. Spring.)

5·2. **Two-dimensional displays.**

However the errors involved in microdensitometry caused by slight variation in local crystal thickness, for example, and the fact that micrographs are « pictures » has led to the use of two-dimensional displays of calculations for comparison with micrographs. Early display methods included pen and ink contour maps (*e.g.* Hashimoto, Howie and Whelan ([11])) and pseudo-micrographs with a strictly limited grey scale formed by photographic superposition (*e.g.* Goringe and Valdrè ([6])). With the increase in computing facilities in recent years display methods have been developed considerably, two different techniques being now commonly used to obtain pseudo-micrographs. In the first method a two-dimensional display is produced by converting intensities into characters to be printed by the conventional line printer used as an output device for most computers. To achieve an adequate grey scale it is necessary to overprint certain characters (Head ([12])), the choice of characters being dictated by the particular type-face used on the printer and the results of preliminary densitometry of photographs of test output strips. An example of a calculation for a dislocation threading a foil is shown in Fig. 14*a*) where the characters used in the printing may be distinguished. Figure 14*b*) shows the same output printed out of focus (so that the characters cover the paper more fully) in an attempt to simulate a microscope dislocation image more closely. However, in this case the grey scale has not been corrected for the effects of defocus and thus the gain in image quality is not as marked as might be expected. For many high quality examples reference should be made to Head ([12]). In the second technique the calculated intensities are converted to spot patterns of variable density and brightness on a cathode ray tube display (see *e.g.* Spring ([13])) which is part of the buffered output chain of the computer. Photography of this display, usually out of focus by some standard amount, yields exceptionally high quality pseudo-micrographs. Examples of computed dislocations superimposed on micrographs to which they correspond are shown in Fig. 15. The second method (CRT display) is considerably more expensive in operation and for many purposes adequate « micrographs » may be obtained by suitable photoreduction of line printer output.

6. Time-saving techniques.

As computer calculations are expensive, consideration must obviously be given to methods of minimising the amount of unnecessary calculation undertaken (in addition to the filtering-out of worthless calculations!). This point has already been touched on in Subsect. **4**·3 where the advantages of the Bloch wave formulation were discussed. Reference to Fig. 16*a*) shows how time is

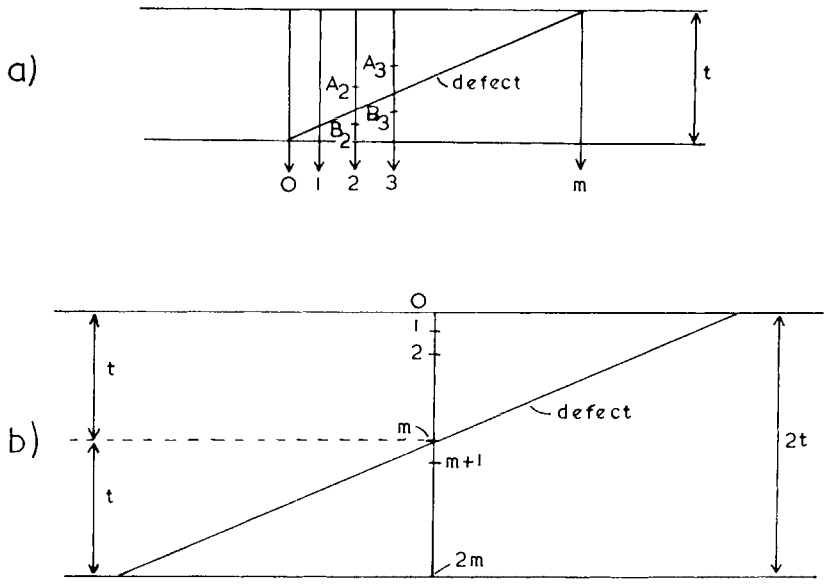


Fig. 16. – *a*) Conventional arrangement of $(m+1)$ columns to calculate contrast from a defect in a foil of thickness t. *b*) Head's single column in a foil of thickness $2t$ used to synthesize the $(m+1)$ solutions of *a*).

wasted in conventional calculations on defects (*e.g.* dislocations) threading the foil. If calculations are made down columns denoted 2 or 3 then the strain field experienced over the section A_iB_i $(i = 2, 3)$ is identical if surface effects are ignored. One method of avoiding this wastage would be to calculate the scattering matrices (see Subsect. **2**·2 and **2**·3) for the distorted region A_iB_i and perfect crystal regions of varying thicknesses and multiply out as required for the various columns. A variant of this principle was,

in fact, adopted by Head ([12]) for the 2-beam case using φ_0, φ_g notation (Subsect. 4·1).

Equations (24) are first order differential equations for φ_0 and φ_g and thus there are only two independent solutions, all other solutions being linear combinations. Thus (if surface effects may be ignored) integration of eqs (24) through a foil $2t$ thick with the dislocation at its centre (Fig. 16*b*)) in $2m$ steps using starting conditions « 0 » ($\varphi_o^{o,o} = 1$, $\varphi_g^{o,o} = 0$) and « g » ($\varphi_o^{g,o} = 0$, $\varphi_g^{g,o} = 1$) and storing the solutions ($\varphi_o^{o,j}$, $\varphi_g^{o,j}$ and $\varphi_o^{g,j}$, $\varphi_g^{g,j}$) after each step enables solutions φ_o^q, φ_g^q for the q-th column ($0 \leqslant q \leqslant m$) of Fig. 16*a*) to be found from

$$\varphi_o^q = a_o^q \varphi_o^{o,m+q} + a_g^q \varphi_o^{g,m+q}, \qquad \varphi_g^q = a_o^q \varphi_g^{o,m+q} + a_g^q \varphi_g^{g,m+q}, \tag{31}$$

such that

$$a_o^q \varphi_o^{o,q} + a_g^q \varphi_o^{g,q} = 1, \qquad a_o^q \varphi_g^{o,q} + a_g^q \varphi_g^{g,q} = 0, \tag{32}$$

or, in matrix notation,

$$\boldsymbol{\varphi}^q = \boldsymbol{\varphi}^{m+q} \boldsymbol{a}^q, \tag{31a}$$

such that

$$\boldsymbol{\varphi}_{\text{incident}} = \boldsymbol{\varphi}^q \boldsymbol{a}^q, \tag{32a}$$

$_o$ and $_g$ being dropped for clarity. Operations in the form of eq. (31) and (32) are fast compared with repeated use of the integration routine and, assuming that storage of solutions is also rapid, then the time taken to calculate $(m + 1)$ columns in a foil of thickness t by this method is only $2 \times 2t \times T$ (*i.e.* two integrations, $_o$ and $_g$, through a foil of thickness $2t$, where T is the time taken to integrate unit thickness, assumed constant) compared with $(m + 1) \times \times t \times T$ for integration of individual columns, *i.e.* less by a factor $4/(m + 1)$. Of course T is only approximately constant, depending on the strain field, etc. and is much increased in the former case if the external step length (t/m) is less than the internal step length which would give the required accuracy (see Fig. 11). Thus the optimum improvement is achieved by making t/m approximately equal to an average internal step length.

In principle the Head technique is immediately extendable to the n-beam case in the $\boldsymbol{\varphi}$ formulation, eqs (31*a*) and (32*a*) now relating $n \times 1$ vectors and $n \times n$ matrices. The necessary n independent solutions are found from starting conditions $\boldsymbol{\varphi}_{\text{initial}} = (1, 0, 0, 0 \ldots)$, $(0, 1, 0 \ldots)$, etc. On the same

basis as before the ratio of computing times now becomes $2n/(m+1)$, *i.e.* less favourable than in the 2-beam case. Also, as discussed in Subsect. 4·3, the $\boldsymbol{\varphi}$ formulation is not really suitable for n-beam calculations because of the oscillatory form of the components of $\boldsymbol{\varphi}$. However, there seems no reason why the Head technique should not be applied to the Bloch wave formulation (with the same favourable time factor $2n/(m+1)$) by the straightforward replacement of $\boldsymbol{\varphi}$ by $\boldsymbol{\psi}$ in eqs (31*a*) and (32*a*) and the use of initial conditions of the form $\boldsymbol{\psi}_{\text{initial}} = (1,0,0\ldots)$, $(0,1,0\ldots)$, etc. The only additional requirements are that $\boldsymbol{\psi}_{\text{incident}}$ must be defined through an equation of the form of eq. (20), *i.e.*

$$\mathbf{C}\boldsymbol{\psi}_{\text{incident}} = \boldsymbol{\varphi}_{\text{incident}} \tag{33}$$

and the final result $\boldsymbol{\varphi}^q$ through an equation of the form of eq. (28)

$$\boldsymbol{\varphi}^q = \mathbf{C}\{\exp[2\pi i\gamma t]\}\boldsymbol{\psi}^q . \tag{34}$$

To the author's knowledge no work has yet been carried out along these lines, but it does appear that the resulting improvement in efficiency accruing from the combined use of the Bloch wave formulation and the Head technique may become important in situations where many beams must be considered, *e.g.* in high voltage electron microscopy image calculations.

7. Uniqueness of computed results.

Recently Head ([12a]) has shown that for certain analytic displacement fields (with zero derivative at infinity and such that there is a direction in the object along which displacements are constant), *e.g.* dislocation strain fields, there is usually a unique reconstruction of the component of the displacement field in the direction of the g-vector from measurements of intensity on one micrograph. It follows that three micrographs taken with noncoplanar g-vectors uniquely identify the defect. In a second paper by Head ([12b]) the analysis is extended to the n-beam case without reservation and to the case of discontinuous displacement fields (*e.g.* stacking faults) with the proviso that the reconstruction may not necessarily be unique. Actual reconstructions from experimental micrographs have, as yet, not been calculated and it remains

to be seen whether experimental errors can be overcome sufficiently well to enable reconstruction calculations to be undertaken with any confidence. However the uniqueness proofs do confirm that the process of calculation from defect models for comparison with experimental micrographs is a very reasonable one; if a good fit between the two is found then it is most likely that the model is correct, being the « unique » solution.

8. Sources of useful parameters.

In all the preceding discussion it has been assumed that the values of ξ_g, ξ_g' suitable for a particular application have been known (or that the model, *e.g.* perfect crystal, was set up to measure them). Early on it was found that ratios of ξ_g'/ξ_g of approximately 10 were suitable for metals such as copper and stainless steel, on which much work on dislocation image contrast was carried out. The absolute values of ξ_g were calculated from the tables of atomic scattering factors, f_e, for electrons then available. Hirsch *et al.* [1] quote values of f_e for the whole periodic table from Ibers and Vainshtein [14] and values of ξ_g for various elements using the analytic approximation to f_e of Smith and Burge [15]. For more up-to-date data on f_e reference should be made to Doyle and Turner [16] and on absorption parameters to Humphreys and Hirsch [17] and Radi [18].

9. References to alternative formulations.

In the space available it has not been possible to review computing methods based on more than one basic approach (that of Howie and Whelan) to the problem of diffraction contrast; in particular no mention has been made of the Cowley-Moodie approach (see *e.g.* Cowley [19]), or to calculations which do not involve the column approximation (see *e.g.* Howie and Basinski [20]), or to the modified Bloch wave approaches of Wilkens [21] or Yoffe [22] (see also *Problem 20* of « Typical Problems ... » in this volume). The recent development of interest in high-resolution « weak-beam » experiments is also omitted, being dealt with in some detail in *Problem 16* of «Typical Problems... » in this volume.

Acknowledgements.

The author is most grateful to many colleagues for their invaluable assistance in the preparation of this manuscript. In particular he would like to thank the following for providing material for many of the Figures; Dr. D. J. H. Cockayne (Fig. 9, 11 and 13), Dr. R. C. Crawford (Fig. 14) and Drs. P. M. Hazzledine, H. P. Karnthaler and M. S. Spring (Fig. 15), and Dr. Cockayne for comments on the manuscript.

REFERENCES

1) P. B. Hirsch, A. Howie, R. B. Nicholson, D. W. Pashley and M. J. Whelan: *Electron Microscopy of Thin Crystals*, Butterworths (1965).
2) R. Gevers, P. Delavignette, H. Blank and S. Amelinckx: *Phys. Stat. Sol.*, **4**, 383 (1964).
3) R. Gevers, P. Delavignette, H. Blank, J. van Landuyt and S. Amelinckx: *Phys. Stat. Sol.*, **5**, 595 (1964).
4) R. Gevers: *Phil. Mag.*, **7**, 1681 (1962).
5) G. Remaut, R. Gevers, A. Lagasse and S. Amelinckx: *Phys. Stat. Sol.*, **10**, 121 (1965); **13**, 125 (1966).
6) M. J. Goringe and U. Valdrè: *Proc. Roy. Soc.*, A **295**, 192 (1966).
7) D. J. H. Cockayne: *D. Phil. Thesis*, University of Oxford (1970).
8) A. Howie: *Phil. Mag.*, **14**, 223 (1966).
9) G. R. Booker: *Ph. D. Thesis*, University of Cambridge (1966).
10) A. Nordsieck: *Math. Comput.*, **16**, 22 (1962).
11) H. Hashimoto, A. Howie and M. J. Whelan: *Proc. Roy. Soc.*, A **269**, 80 (1962).
12) A. K. Head: *Austr. Journ. Phys.*, **20**, 557 (1967); **22**, 43 (1969); **22**, 345 (1969).
12*a*) A. K. Head: *Austr. Journ. Phys.*, **22**, 43 (1969).
12*b*) A. K. Head: *Austr. Journ. Phys.*, **22**, 345 (1969).
13) M. S. Spring: *Ph. D. Thesis*, University of Cambridge (1970).
14) J. A. Ibers and B. K. Vainshtein: *International Crystallographic Tables*, vol. III, Kynoch Press (1962).
15) G. H. Smith and R. E. Burge: *Acta Cryst.*, **15**, 182 (1962).
16) P. A. Doyle and P. S. Turner: *Acta Cryst.*, A **24**, 390 (1968).
17) C. J. Humphreys and P. B. Hirsch: *Phil. Mag.*, **18**, 115 (1968).
18) G. Radi: *Acta Cryst.*, A **26**, 41 (1970).
19) J. M. Cowley: *Progress in Materials Science*, **13**, 269 (1967).
20) A. Howie and Z. S. Basinski: *Phil. Mag.*, **17**, 1039 (1968).
21) M. Wilkens: *Phys. Stat. Sol.*, **5**, 175 (1964); **6**, 939 (1964).
22) E. M. Yoffe: *Phil. Mag.*, **21**, 833 (1970).

Typical Problems in Electron Microscopy

M. J. GORINGE

Department of Metallurgy, University of Oxford - Oxford, England

C. R. HALL

Cavendish Laboratory, University of Cambridge - Cambridge, England

1. Introduction.

The following problems cover three main topics: i) basic diffraction theory (problems 1-11; see A. Howie: this volume), ii) contrast caused by defects (problems 12-20; see L. M. Brown: this volume) and iii) radiation damage (problems 21-26; see M. J. Makin: this volume). An outline solution is appended for each. For further references to basic material the following standard texts are suggested: stereographic projection and crystal symmetry, Phillips ([1]); reciprocal lattice and diffraction, James ([2]); electron diffraction and electron microscopy, Hirsch *et al.* ([3]).

2. Problems.

Problem 1 – Find the co-ordinates of the points in the reciprocal lattice if in the real cell atoms are at positions ρ_j (see Howie, p. 276) given by:

i) $(0, 0, 0), (0, \frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2}, 0), (\frac{1}{2}, 0, \frac{1}{2})$ in a cubic cell;

ii) $(0, 0, 0), (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ in a cubic cell;

iii) $(0, 0, 0), (2a/3, a/3, c/2)$ in a hexagonal cell.

Problem 2 – Index the patterns of Fig. 2.1 which are for: *a*) fcc, *b*) bcc, *c*) hcp, *d*) NaCl. What is the approximate beam direction in each case?

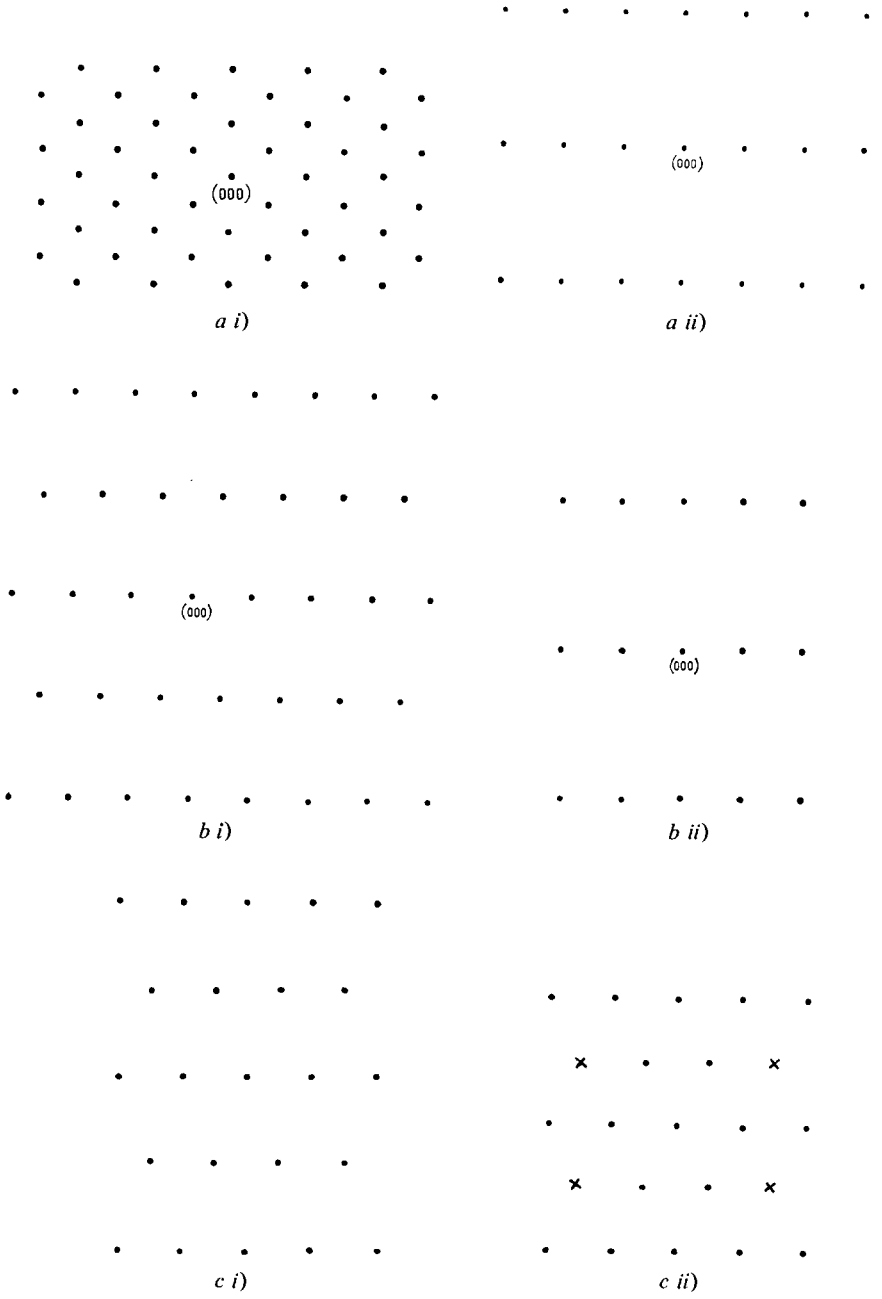
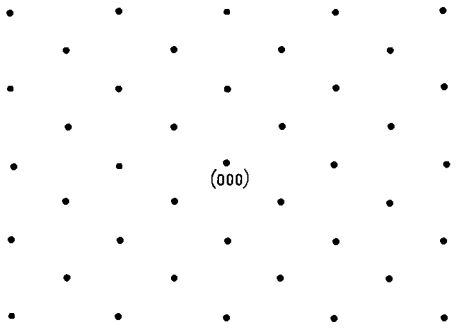
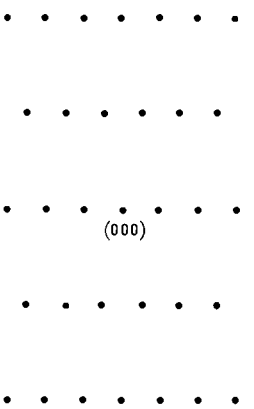


Fig. 2.1

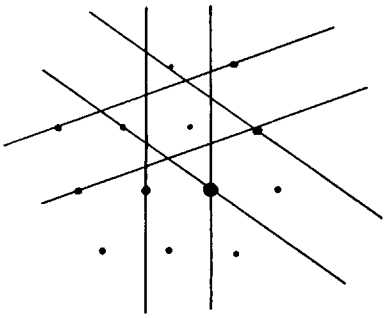


d i)



d ii)

Fig. 2.1



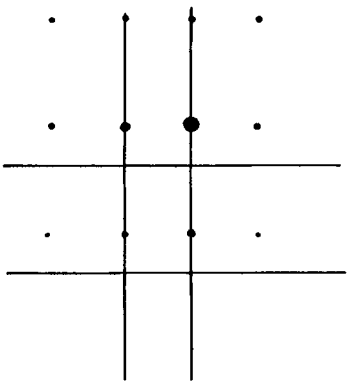


Fig. 3.1.

Problem 3 – The two diffraction patterns of Fig. 3.1 correspond to copper at 100 keV. Determine the approximate orientation for each and, assuming that in order to get from one to the other the crystal is tilted so that the Kikuchi lines, other than the pair passing through the spots, move away from the centre of the pattern, find the angle of tilt between them ($a = 3.61$ Å).

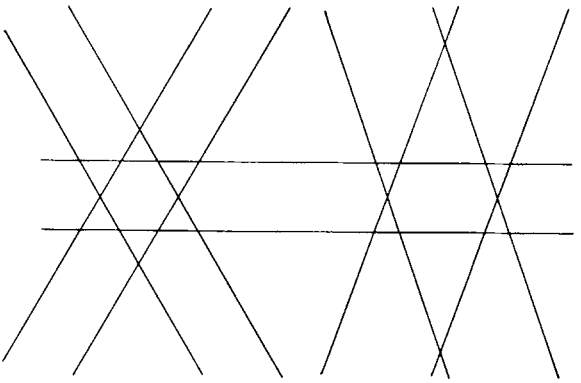


Fig. 4.1.

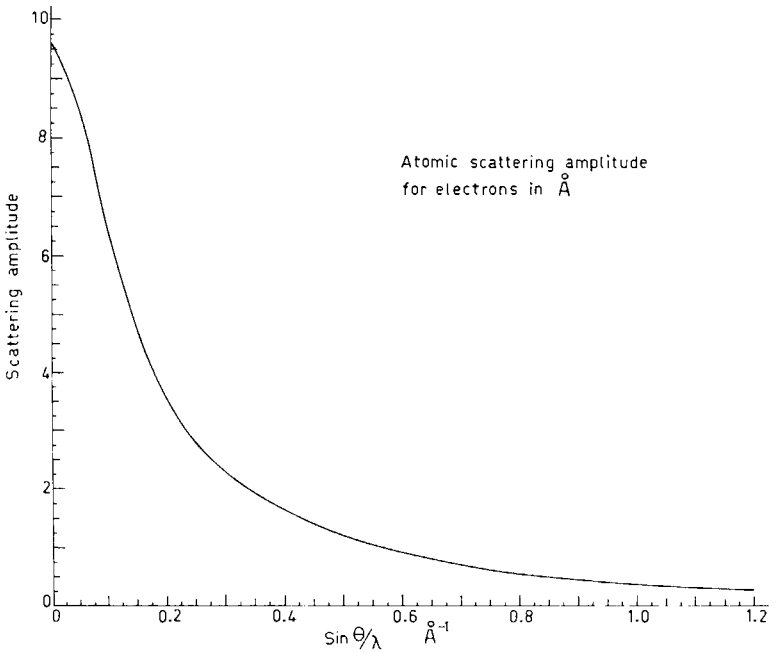


Fig. 5. – Atomic scattering factor (not relativistically corrected).

Problem 4 – Calculate the accelerating voltage of the microscope used to take the diffraction pattern of aluminium ($a = 4.08$ Å) shown in Fig. 4.1.

Problem 5 – Given the graph of nonrelativistically corrected scattering factors as a function of $\sin\theta/\lambda$ (Fig. 5), calculate the extinction distances for the following reflections: 111, 110, 220, 222, 333, assuming that the material is: *a*) fcc ($a = 4$ Å); *b*) bcc ($a = 3.1$ Å); *c*) diamond structure ($a = 5.2$ Å). [Atoms in the diamond unit cell are at (0, 0, 0); $(\frac{1}{2}, \frac{1}{2}, 0)$; $(0, \frac{1}{2}, \frac{1}{2})$; $(\frac{1}{2}, 0, \frac{1}{2})$; $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$; $(\frac{3}{4}, \frac{3}{4}, \frac{1}{4})$; $(\frac{1}{4}, \frac{3}{4}, \frac{3}{4})$; $(\frac{3}{4}, \frac{1}{4}, \frac{3}{4})$.]

Assume $E = 80$ keV, when $\lambda = 0.042$ Å and $m/m_0 = 1.16$.

Problem 6 – The bend contour in the photograph of Fig. 6 corresponds to a 111 reflection in Al at 100 kV, the extinction distance for this reflection being 560 Å. Estimate the thickness at a number of points along the contour.

Problem 7 – The potential $V(\boldsymbol{r})$ and the wave function $\psi_{\boldsymbol{k}}(\boldsymbol{r})$ describing an electron of wave vector $\boldsymbol{k}$ in a perfect centro-symmetric crystal may be written:

$$V(\boldsymbol{r}) = \sum_{\boldsymbol{g}} V_{\boldsymbol{g}} \exp\left[2\pi i \boldsymbol{g}\cdot\boldsymbol{r}\right],$$

$$\psi_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_{\boldsymbol{g}} C_{\boldsymbol{g}}(\boldsymbol{k}) \exp\left[2\pi i(\boldsymbol{k}+\boldsymbol{g})\cdot\boldsymbol{r}\right],$$

where the summations extend over the reciprocal lattice vectors $\boldsymbol{g}$. Derive the equations satisfied by the $C_{\boldsymbol{g}}(\boldsymbol{k})$ and hence find the forms of the four possible waves which are propagated in a cubic crystal when the (220), (200) and (020) planes are simultaneously at the reflecting position. Calculate the intensity distributions in each of the Bloch waves around the atomic positions.

Problem 8 – A certain (hypothetical) absorption process gives rise to a uniform probability of absorption in a cube of side αd centred on each atom, where d is the distance between the Bragg planes and $\alpha < 1$. What is the ratio of the absorption distance of the well-transmitted Bloch wave to that of the strongly absorbed wave at the exact reflecting orientation? What happens to the ratio as α tends to zero?

Problem 9 – A crystal of nickel is prepared in the form of a wedge with its upper surface parallel to (100) and its lower surface to (110). Electrons of energy 100 keV enter through the upper surface and travel in the (001) plane perpendicular to the edge of the wedge, falling on the (020) planes at the exact Bragg angle. If the extinction distance for this reflection is 250 Å calculate the angular splitting of the reflected beam due to refraction as it leaves through the lower surface.

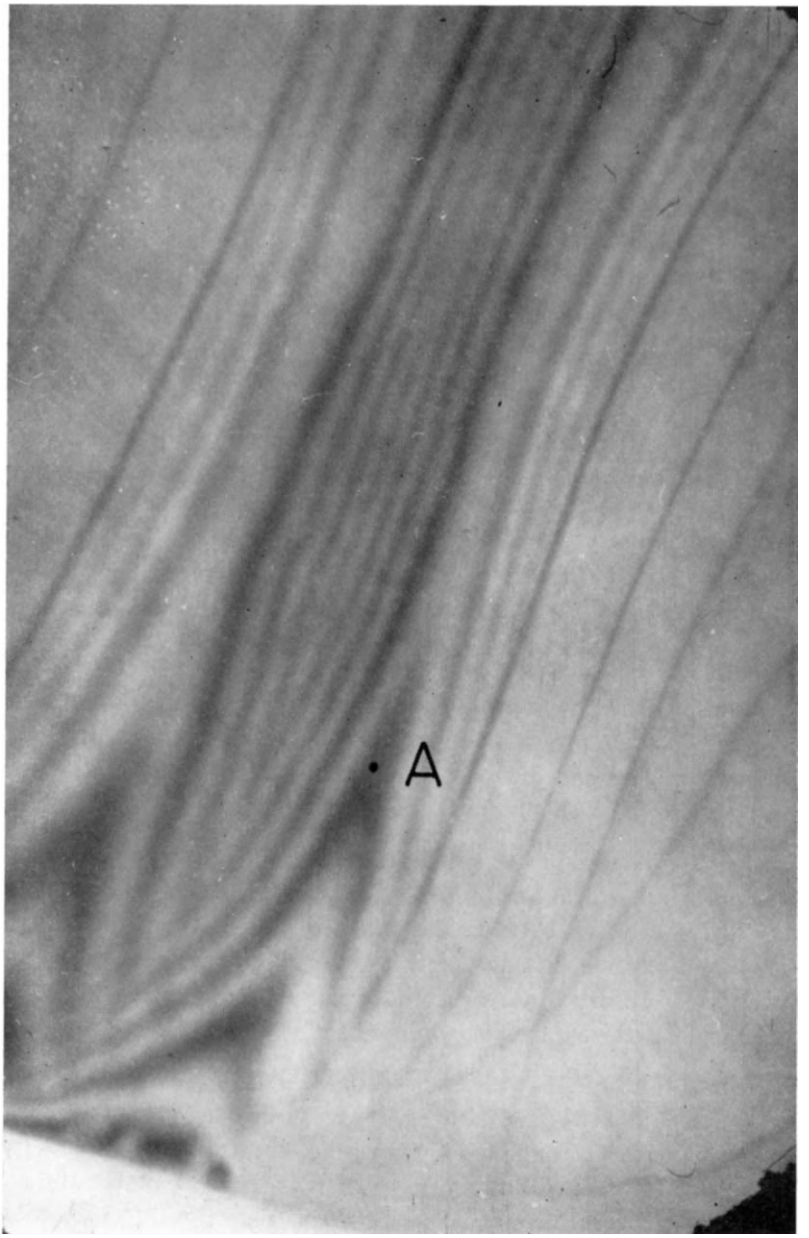


Fig. 6.

Problem 10 – Use the phase-grating approximation (see Howie, this volume) to obtain the intensity distribution produced at the reflecting position by two superposed crystals each of thickness t which have potentials in the $(x\text{-}y)$ plane given respectively by:

$$V_0 + V_1 \cos(2\pi g_1 x), \qquad V_0 + V_1 \cos(2\pi g_2 x).$$

Problem 11 – A crystal of thickness t has a phase-grating potential given by

$$V_0 + V_1 \cos(2\pi g_x x) + V_1 \cos(2\pi g_y y) + V_2 \cos(2\pi(g_x x + g_y y)).$$

Calculate the intensity on the phase-grating approximation of the diffrated beam having $\boldsymbol{g} = (g_x, g_y, 0)$.

Problem 12 – Draw an edge dislocation in a foil with $\boldsymbol{g}\cdot\boldsymbol{b} = 1$, showing the Bragg planes. On which side of the dislocation does the image lie for orientations of the matrix such that *a*) $s > 0$ and *b*) $s < 0$?

Problem 13 – Draw a coherent misfitting sphere in a foil showing the Bragg planes. Why is there a line of no contrast perpendicular to $\boldsymbol{g}$ through the centre of the defect? Show that there will always be such a line of no contrast for displacements which are symmetrical with respect to the Bragg plane through the centre of the defect.

Problem 14 – Show that the moiré fringe spacing which arises from the two spots $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ as in the diffraction pattern of Fig. 14 is

$$\frac{1}{|\boldsymbol{g}_1 - \boldsymbol{g}_2|}.$$

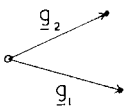


Fig. 14.

Two similar lattices rotated by a small angle $\delta\theta$ give rise to a rotation moiré pattern. If this is misinterpreted by assuming that the lattices have different spacings but are parallel (parallel moiré) what lattice parameter difference would be deduced?

Problem 15 – The micrograph of Fig. 15 was taken using an objective aperture of size and position indicated on the inset diffraction pattern (which

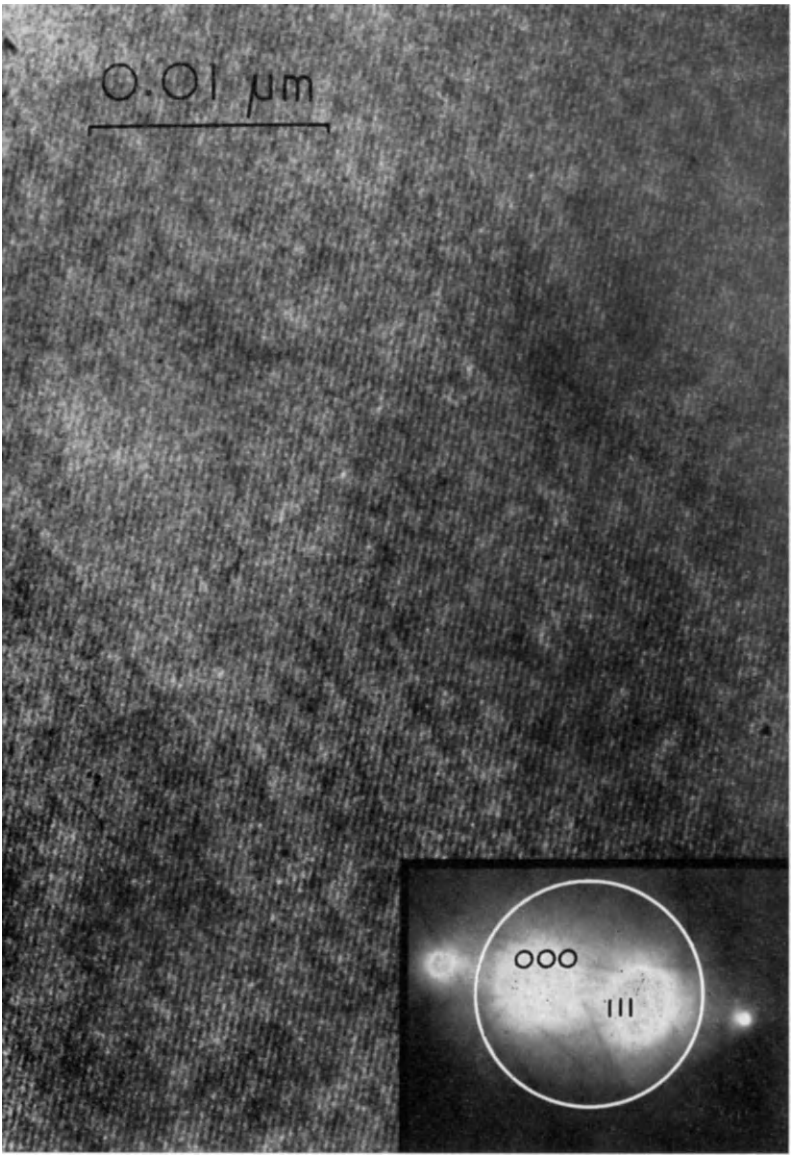


Fig. 15. – Micrograph taken using objective aperture of size and position as in inset diffraction pattern (correctly oriented).

is correctly oriented). Explain the nature of the closely spaced fringes on the micrograph and calculate the size of the unit cell (cubic) of the specimen material. [Micrograph courtesy of I. L. F. Ray.]

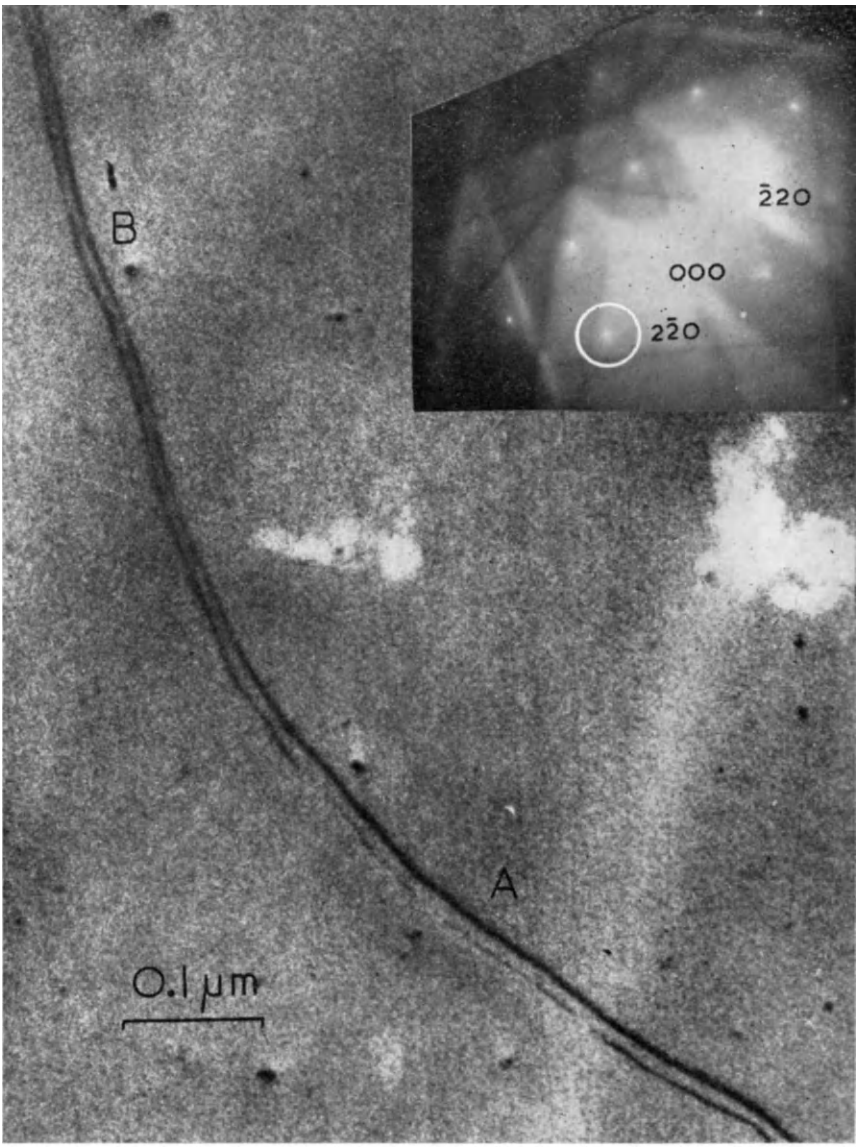


Fig. 16.1. – Electron micrograph (negative print) of Cu-10 at % Al taken using an objective aperture of size and position indicated in the inset diffraction pattern (correctly oriented).

Problem 16 – The micrograph of Fig. 16.1 of copper-10 at% aluminium (fcc) was taken using an objective aperture of size and position indicated on the inset diffraction pattern (which is correctly oriented). The micrograph (which is a negative print) is of a foil with normal very near [111] and the dislocation is dissociated according to the scheme

$$\boldsymbol{b} = \frac{a}{2}[1\bar{1}0] \rightarrow \boldsymbol{b}_1^P + \boldsymbol{b}_2^P = \frac{a}{6}[1\bar{2}1] + \frac{a}{6}[2\bar{1}\bar{1}] \tag{16.1}$$

and near A the dislocation line direction $\boldsymbol{u}$ is parallel to $[11\bar{2}]$.

a) Calculate $\boldsymbol{g}\cdot\boldsymbol{b}_1^P$, $\boldsymbol{g}\cdot\boldsymbol{b}_2^P$ and $\boldsymbol{g}\cdot\boldsymbol{R}$ for the two partial dislocations and the connecting stacking fault. Use the results to explain the visibility of both partials and the absence of stacking fault contrast.

b) Assuming that the dark lines accurately define the positions of the partials, estimate the stacking fault energy, γ, of the material from the formula

$$\gamma = \frac{\mu b^2}{24\pi\Delta}\cdot\frac{(2-\nu)}{(1-\nu)}\left[1 - \frac{2\nu\cos 2\alpha}{(2-\nu)}\right], \tag{16.2}$$

where μ is the shear modulus, ν the Poisson ratio, Δ the separation of the partials, $\boldsymbol{b}$ the total Burgers vector of the dislocation, and α the angle between $\boldsymbol{b}$ and $\boldsymbol{u}$.

c) Note that eq. (16.2) predicts values of Δ depending on $\boldsymbol{u}$. Confirm this variation by measurements near A and B.

d) Calculate the value of the deviation parameter $s_{2\bar{2}0}$ for the dark field beam used to form the micrograph.

e) Using the co-ordinate system of Fig. 16.2 for the total dislocation in edge orientation we have at point B a displacement $\boldsymbol{R} = \boldsymbol{R}_1^P + \boldsymbol{R}_2^P$ such that

$$\boldsymbol{g}\cdot\boldsymbol{R}_i^P = \frac{\boldsymbol{g}\cdot\boldsymbol{b}_i^P}{2\pi}\left(\theta_i + \frac{\sin 2\theta_i}{4(1-\nu)}\right), \qquad i = 1, 2. \tag{16.3}$$

The observed positions of the « weak-beam » peaks occur for values of x and z where

$$s_g + (\mathrm{d}/\mathrm{d}z)(\boldsymbol{g}\cdot\boldsymbol{R}) = 0 \tag{16.4}$$

and

$$(\mathrm{d}^2/\mathrm{d}z^2)(\boldsymbol{g}\cdot\boldsymbol{R}) = 0\,. \tag{16.5}$$

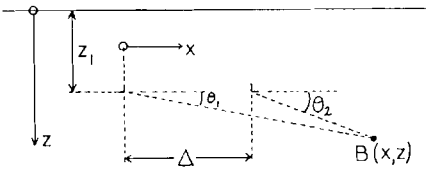


Fig. 16.2. – Co-ordinate system used in calculation.

Show that the observed separation Δ_{obs} is given by

$$\Delta_{\mathrm{obs}} = (\Delta^2 + 4/c^2)^{\frac{1}{2}}\,, \tag{16.6}$$

where

$$c = -s_g \Big/ \left[\frac{\boldsymbol{g}\cdot\boldsymbol{b}^P}{2\pi}\left(1 + \frac{1}{2(1-\nu)}\right)\right] \tag{16.7}$$

and

$$\boldsymbol{g}\cdot\boldsymbol{b}^P = \boldsymbol{g}\cdot\boldsymbol{b}_1^P = \boldsymbol{g}\cdot\boldsymbol{b}_2^P.$$

f) Using the results of *d*) and *e*) estimate the correction necessary to the value of γ calculated in *b*).

Assume $a = 3.66$ Å, $\mu = 4\cdot10^{11}$ dyne cm^{-2}, $\nu = \frac{1}{3}$ and electron wavelength $\lambda = 0.032$ Å. Note that *e*) entails considerable algebra and only the results are required for *f*).

[Micrograph courtesy D. J. H. Cockayne and I. L. F. Ray.]

Problem 17 – Calculate:

a) the maximum density of dislocations that can be observed under 2-beam bright field conditions in copper (assume foil thickness 2000 Å; 100 keV electrons);

b) the maximum stacking fault energy, γ, that can be measured by the node method (see Fig. 17) assuming

$$\frac{\gamma w}{\mu b^2} = \frac{1}{3}\,.$$

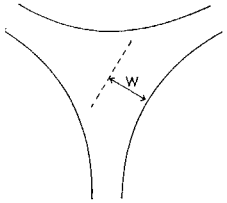


Fig. 17.

In copper and gold, no extended nodes are seen. What are the minimum values for γ_{Au} and γ_{Cu}? (Data as per Table.)

Material	Extinction distances (Å)			a (Å)	μ (dyne cm^{-2})
	111	200	220		
Cu	242	281	416	3.608	$4 \cdot 10^{11}$
Au	159	179	248	4.070	$2.8 \cdot 10^{11}$

(See *Problem 16* above for an alternative method of measuring γ.)

Problem 18 – In the light of your answers to *Problem 17* do you think that the resolution of the electron microscope limits present-day observations in solid-state physics?

Problem 19 – A specimen of thickness t contains a small inclusion of thickness Δz which scatters electrons differently from the matrix. The crystal is viewed under 2-beam dynamical conditions at the Bragg position ($s = 0$) and the effective extinction distances are ξ_{gm} for the matrix and ξ_{gi} for the inclusion, which may be assumed to have the same crystal structure as the matrix. Calculate the visibility of the inclusion as a function of *a*) its depth in the specimen and *b*) the specimen thickness. How would you expect these results to be modified if the specimen is not precisely at the Bragg position (*i.e.* $s \neq 0$)?

Problem 20–a) Show that under 2-beam conditions the change in transmitted intensity due to an imperfection is proportional to the real part of the change in the well-transmitted Bloch wave amplitude. (Assume that the crystal is thick enough to reduce the amplitude of the absorbed Bloch wave to zero.)

b) Given the equation governing the scattering of Bloch waves into the well-

transmitted waves (here denoted $\psi^{(1)}$)

$$\frac{\mathrm{d}\psi^{(1)}}{\mathrm{d}z} = 2\pi i\beta'_g\left[\sin^2\frac{\beta}{2}\psi^{(1)} - \frac{1}{2}\sin\beta\exp\left[2\pi i\Delta kz\right]\psi^{(2)}\right] \tag{20.1}$$

(see this volume, Howie, eq. (51); Goringe, eq. (30); or Hirsch *et al.*, [3] eq. (12.25)); by making the following substitutions to eliminate intraband scattering,

$$\psi^{(1)} = \Psi^{(1)}\exp\left[2\pi i\sin^2(\beta/2)\boldsymbol{g}\cdot\boldsymbol{R}\right] \tag{20.2}$$

and

$$\psi^{(2)} = \Psi^{(2)}\exp\left[2\pi i\cos^2(\beta/2)\boldsymbol{g}\cdot\boldsymbol{R}\right], \tag{20.3}$$

using the fact that $\beta'_g = (\mathrm{d}/\mathrm{d}z)(\boldsymbol{g}\cdot\boldsymbol{R})$ and integrating (assuming $\Psi^{(2)}$ constant over integration), show that the effect of an imperfection at depth z_0 is

$$\Delta\Psi^{(1)} = -\Psi^{(2)}\pi i\sin\beta\exp\left[2\pi i\Delta kz_0\right]\int_{z-z_0=-\infty}^{z-z_0=\infty}\beta'_g(z-z_0)\cdot$$
$$\cdot\exp\left[2\pi i[\Delta k(z-z_0)+\boldsymbol{g}\cdot\boldsymbol{R}\cos\beta]\right]\mathrm{d}(z-z_0)\,. \tag{20.4}$$

Hence show that if $\beta'_g(z-z_0)$ is odd then $\mathrm{Re}(\Delta\Psi^{(1)})$ is proportional to $\cos 2\pi\Delta kz_0$ and if $\beta'_g(z-z_0)$ is even then $\mathrm{Re}(\Delta\Psi^{(1)})$ is proportional to $\sin 2\pi\Delta kz_0$.

Deduce the depth-variation in the visibility of *c*) stacking faults, *d*) dislocations and *e*) centres of strain.

Problem 21 – Calculate the maximum knock-on energy along *a*) [111], *b*) [100] and *c*) [110] that can be transferred to atoms in a copper target bombarded along [100] by protons of energy 5 MeV. [At.wt Cu = 63.6.]

Problem 22 – In copper exposed to a flux of $1\cdot10^{13}$ cm^{-2} s^{-1} of neutrons of energy 10 keV calculate *a*) the concentration of primary knock-ons per year, *b*) their mean energy and *c*) the total point defect concentration produced, assuming the hard-sphere model and that no recombination takes place. [Copper is fcc, $a = 3.608$ Å, atomic weight 63.6, displacement energy $E_d = 19$ eV, total neutron cross-section $\sigma = 3\cdot10^{-24}$ cm^2.]

Problem 23 – Calculate the energies of a cluster of 1000 vacancies in copper when in the form of *a*) a spherical void, *b*) a Frank loop and *c*) a perfect loop on (111). [Assume surface energy $\gamma' = 1670$ erg cm^{-2}, stacking fault energy $\gamma = 85$ erg cm^{-2}, lattice constant $a = 3.608$ Å, shear modulus $\mu = 4 \cdot 10^{11}$ dyne cm^{-2} and Poisson's ratio $\nu = \frac{1}{3}$.]

Problem 24 – If a copper sample bombarded by α-particles contains 10^{15} cm^{-3} equilibrium gas bubbles 200 Å in diameter at 0°C, calculate *a*) the volume swelling and *b*) the volume of gas at NTP per cm^3 of sample. Calculate *c*) the gas bubble density and *d*) volume swelling if the gas is redistributed into bubbles 100 Å in diameter. [Assume helium gas is perfect; surface energy of copper = 1670 erg cm^{-2}.]

Problem 25 – Calculate the growth factor G if $2 \cdot 10^{-5}$% burn-up in α-uranium results in the formation of 10^{15} cm^{-3} interstitial loops 200 Å in diameter. [α-uranium is orthorhombic with $a = 2.85$ Å, $b = 5.87$ Å and the loops lie on (010).]

Problem 26 – If there are 10^{15} cm^{-3} Frank loops 50 Å in diameter present in an irradiated crystal calculate the critical shear stress assuming that the strength of each loop intersected exceeds μb^2 *i.e.* the dislocations bow between the loops rather than cutting through them. [Assume $\mu = 4 \cdot 10^{11}$ dyne cm^{-2}, $a = 3.608$ Å.]

3. Solutions.

Problem 1 – To find the co-ordinates of the points in the reciprocal lattice if in the real cell atoms are positioned at:

i) $(0, 0, 0), (0, \frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2}, 0), (\frac{1}{2}, 0, \frac{1}{2})$ in a cubic cell (fcc) ;

ii) $(0, 0, 0), (\frac{1}{2}, \frac{1}{2}, \frac{1}{2})$ in a cubic cell (bcc) ;

iii) $(0, 0, 0), (2a/3, a/3, c/2)$ in a hexagonal cell (hcp) ;

Solution. The unit vectors of the reciprocal lattice are given by

$$\boldsymbol{a}^* = (\boldsymbol{b} \times \boldsymbol{c})/V, \quad \boldsymbol{b}^* = (\boldsymbol{c} \times \boldsymbol{a})/V, \quad \boldsymbol{c}^* = (\boldsymbol{a} \times \boldsymbol{b})/V,$$

where V, the volume of the unit cell, is given by $\boldsymbol{a} \cdot (\boldsymbol{b} \times \boldsymbol{c})$.

Which points actually exist within this lattice depends upon the space group of the real cell, and can be found either by looking in the *Interna-*

tional Tables for X-ray Crystallography, vol. 1, or in the case of simple structures, by working out F (Howie, (3), p. 276), the scattering amplitude of the unit cell.

In i) the reciprocal cell is also cubic, and if the edge of the real cell is a, it has edge $a^* = 1/a$. Taking one allowed point to be at the origin it is found that the nearest points are at positions such as (a^*, a^*, a^*), $(2a^*, 0, 0)$, $(2a^*, 2a^*, 0)$ etc. Thus the allowed points lie on a body centred cubic lattice, giving the rule that hkl are all odd or all even in fcc patterns.

Similarly for ii) the edge of the cubic reciprocal cell is again given by $1/a$, but in this case the allowed points lie on a face centred lattice with co-ordinates such as $(a^*, a^*, 0)$, $(2a^*, 0, 0)$, $(0, 2a^*, 0)$, the rule in this case being that $h + k + l$ is even.

In iii) the crystal cell is hexagonal, so that the reciprocal cell is also hexagonal, but is rotated by 30° about the c axis relative to the real cell.

Application of the above formula gives:

$$a^* = b^* = 2/a\sqrt{3}\,, \qquad c^* = 1/c\,.$$

Using the four co-ordinate representation of points in this lattice (*i.e.* $(hkil)$ where $i = -h-k$) we find, using the *International Tables*, volume 1, that all values of h, k and l are allowed except those for which $(h-k) = 3n$ $(n = 0, 1, 2, \ldots)$, when l has to be an even number. Thus the points $(0, 0, 0, c^*)$, $(0, 0, 0, 3c^*)$, $(0, 0, 0, 5c^*)$, $(2a^*, -a^*, -a^*, c^*)$ etc. are absent. Note however that reflected beams corresponding to these points are usually present in electron diffraction patterns as a result of multiple reflection (see *Problem 2 c*).

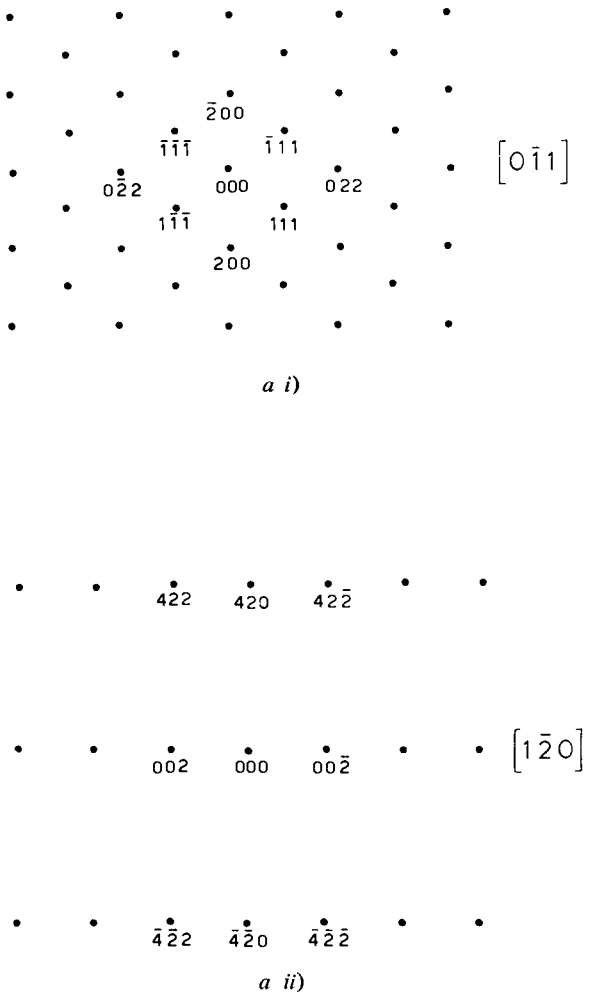
Problem 2 – To index the patterns in Figs 2.1*a*)-*d*) (which are for *a*) fcc, *b*) bcc, *c*) hcp, *d*) NaCl) and to find the approximate beam direction in each case.

Solution. By measuring the spot separation along principal directions the ratio of the corresponding d-spacings and hence, for the cubic materials, the ratio of the two values of $h^2 + k^2 + l^2$ can be found. This must be a ratio of two fairly small integers, and the two smallest which fit approximately can be found. That these give the correct values for hkl for the two spots can be checked from the angle between the spots and the spacings of other pairs of spots in the pattern. The spots can then be indexed in a consistent manner, (see Figs 2.2) and the normal (which is the approximate beam direction) found as the vector normal to any two directions in the pattern.

The hexagonal pattern is more difficult, and unless there is any obvious

symmetry recourse has to be made to trial and error. A list of the ratios of the spacings of the lower order planes and the angles between them is a useful aid. Usually in practice the approximate camera length is known.

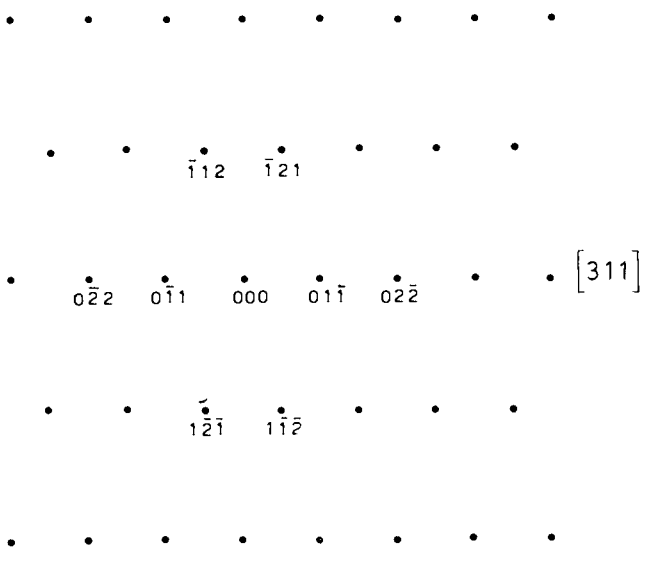
The normal to each pattern is denoted [*uvw*] in the Figures. The spots marked as crosses in Fig. 2.2*c*) ii) are reflections corresponding to points absent from the reciprocal lattice but which are present by multiple reflection.



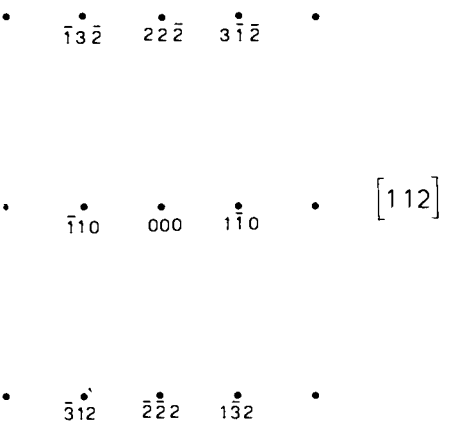
a i)

a ii)

Fig. 2.2


$\bar{1}\,1\,2$
$\bar{1}\,2\,1$
$0\,\bar{2}\,2$
$0\,\bar{1}\,1$
$0\,0\,0$
$0\,1\,\bar{1}$
$0\,2\,\bar{2}$
[311]
$1\,\bar{2}\,\bar{1}$
$1\,\bar{1}\,\bar{2}$

b i)


$\bar{1}\,3\,\bar{2}$
$2\,2\,\bar{2}$
$3\,\bar{1}\,\bar{2}$
$\bar{1}\,1\,0$
$0\,0\,0$
$1\,\bar{1}\,0$
[112]
$\bar{3}\,1\,2$
$\bar{2}\,\bar{2}\,2$
$1\,\bar{3}\,2$

b ii)

Fig. 2.2

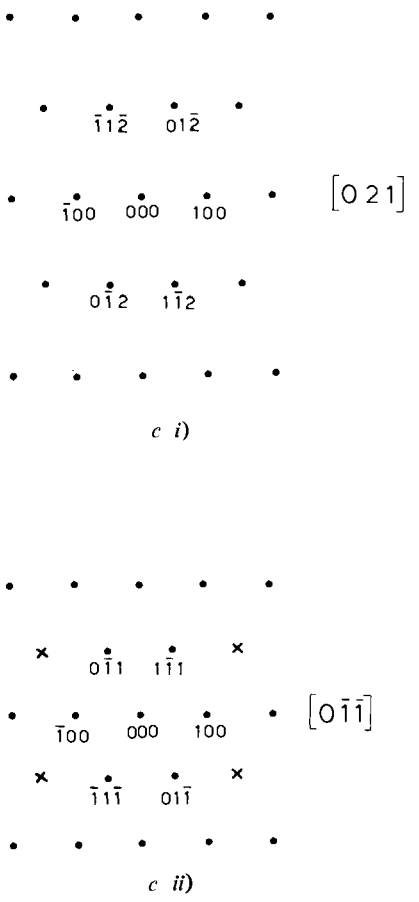


Fig. 2.2

Problem 3 – The two diffraction patterns in Fig. 3.1, Section **2**, correspond to copper at 100 keV. Determine the approximate orientation of each and find the angle of tilt between them (for copper $a = 3.61$ Å).

Solution. The two orientations correspond to the beam being approximately parallel to [110] and [211], and the Kikuchi lines have been so indexed in Fig. 3.2. The angle between these two directions is 30°. For the $20\bar{2}$ reflection twice the Bragg angle (the spacing of the Kikuchi lines) is 1.63°. Hence by measurement the orientation in the upper pattern is found to be

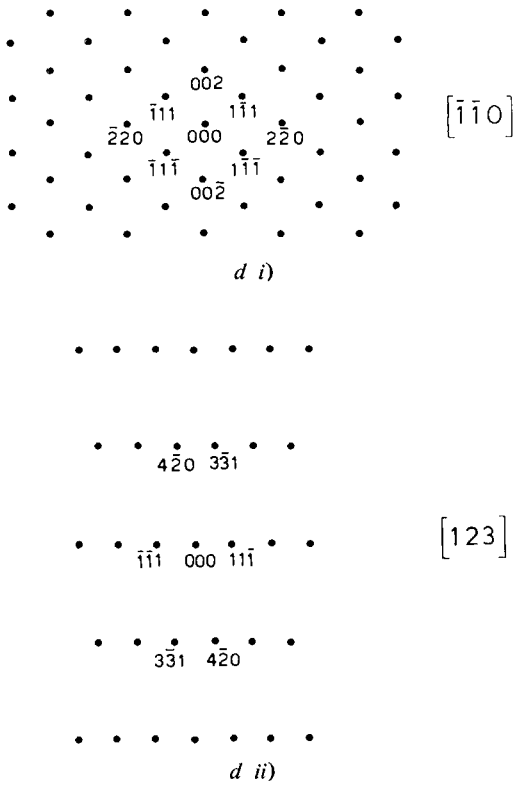


Fig. 2.2

0.68 of this angle away from the position at which [110] would bisect the direct and diffracted $(1\bar{1}1)$ beams; *i.e.* 1.11° away. Similarly in the other pattern the orientation is 1.47° away from the position in which [211] would lie in this direction. Hence the angle between the patterns is $30° - 1.11° - -1.47° = 27.42°$.

Problem 4 – Calculate the accelerating voltage of the microscope used to take the diffraction pattern of aluminium in Fig. 4.1 ($a = 4.08$ Å).

Solution. In Fig. 4.2 the two poles have been identified as [111] and [343], which are an angle of 8.0° apart. Hence the separation of the $\bar{3}\bar{3}1$ and $\bar{3}\bar{3}\bar{1}$ Kikuchi lines are measured to be 2.75°. The *d*-spacing for this reflection is $4.08/(19)^{\frac{1}{2}} = 0.94$ Å. Hence from the Bragg equation the wavelength of the electrons is 0.045 Å, and the energy is thus approximately 70 keV.

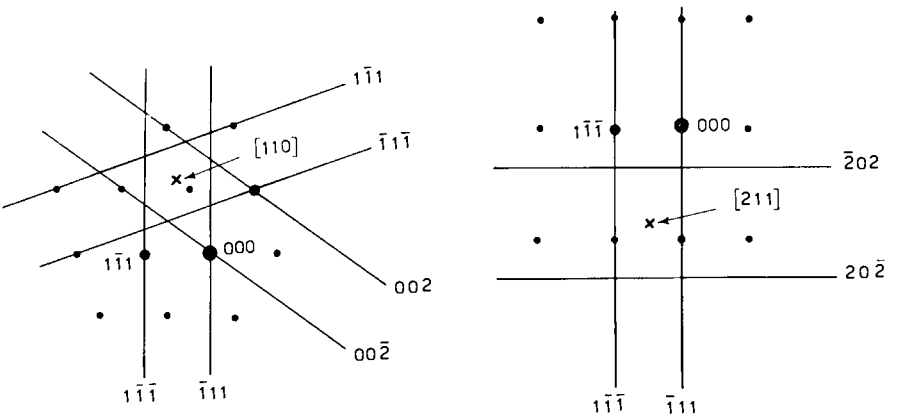


Fig. 3.2.

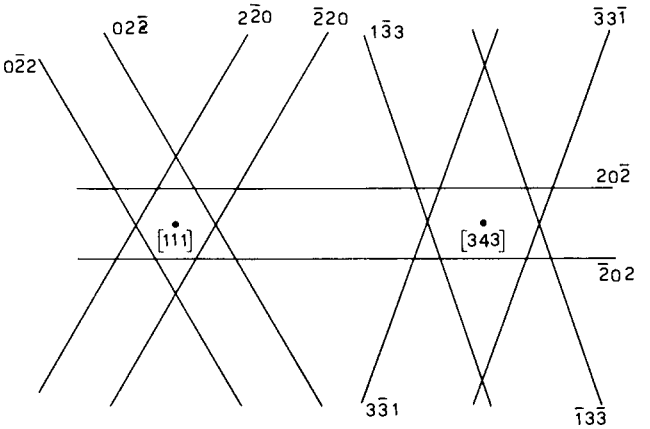


Fig. 4.2.

Problem 5 – Given the nonrelativistically corrected graph of scattering factor as a function of $\sin\theta/\lambda$ (Fig. 5, Section **2**) calculate the extinction distances for the following reflections: 111, 110, 220, 222, 333, assuming that the material is *a*) fcc ($a = 4$ Å), *b*) bcc ($a = 3.1$ Å), *c*) diamond structure ($a = 5.2$ Å). (Atoms in the diamond unit cell are at $(0, 0, 0)$; $(\frac{1}{2}, \frac{1}{2}, 0)$; $(0, \frac{1}{2}, \frac{1}{2})$; $(\frac{1}{2}, 0, \frac{1}{2})$; $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$; $(\frac{3}{4}, \frac{3}{4}, \frac{1}{4})$; $(\frac{1}{4}, \frac{3}{4}, \frac{3}{4})$; $(\frac{3}{4}, \frac{1}{4}, \frac{3}{4})$). The energy is 80 keV, λ=0.042 Å. m/m_0=1.16.

Solution. The extinction distance ξ_g is given by (see Howie, eqs (30) and (35)):

$$\xi_g = \frac{\pi V_c \exp[M_g]}{(m/m_0)\lambda F_g},$$

where

$$F_g = \sum_j f_j \exp[-2\pi i \boldsymbol{g}\cdot\boldsymbol{r}_j],$$

f_j is the scattering amplitude of the atom located at $\boldsymbol{r}_j$ in the unit cell of volume V_c. The effect of thermal vibrations is taken into account by the factor $\exp[M_g]$: this is neglected in the present calculations, but it should be noted that it can be important, especially in the case of higher order reflections. Substitution of the appropriate f_j into the above formulae gives:

	111	110	220	222	333
a) fcc	312 Å	—	544 Å	697 Å	1290 Å
b) bcc	—	320 Å	710 Å	1020 Å	—
c) diamond	383 Å	—	454 Å	—	1390 Å

Problem 6 – The bend contour in Fig. 6, Section **2**, corresponds to a 111 reflection in Al at 100 keV, the extinction distance for this reflection being 560 Å. Estimate the thickness at a number of points along the contour.

Solution. The simplest method of thickness estimation is to count the number of thickness fringes in from the edge of the specimen at the exact reflecting position: thus A is at a thickness of about $2.5\xi_g$.

An alternative method makes use of the spacing of the subsidiary fringes across a bend contour. On the two-beam dynamical theory light fringes occur in bright field at values of thickness t and deviation parameter s given by

$$t^2 = n^2/(s^2 + 1/\xi_g^2), \qquad \text{where } n = 1, 2, 3, \ldots$$

Dark fringes occur at positions found by replacing n by $(n - \frac{1}{2})$. The local value of s can be estimated from the fact that the total change in s across the double contour for one of the reflections is given by g^2/k, which has the value 0.0066 Å^{-1} for the 111 reflection in Al at 100 keV. Thus opposite the point A

on the micrograph the next dark fringe in (for which $n = 4$) is roughly 7/32 of the way across the contour and hence is at a value of s of about 0.0015. Therefore:

$$t^2 = 12.2/((0.0015)^2 + (0.0018)^2),$$

i.e. $t = 2.7\xi_g$.

If ξ_g is large enough and g is also large (*e.g.* a 311 reflection in aluminium) then $1/\xi_g$ can be neglected in comparison with s, the fringes become evenly spaced with a spacing of δs, and t is given approximately by

$$t = 1/\delta s.$$

Problem 7 – The potential $V(\boldsymbol{r})$ and the wave function $\psi_k(\boldsymbol{r})$ describing an electron of wave vector $\boldsymbol{k}$ in a perfect centro-symmetric crystal may be written:

$$V(\boldsymbol{r}) = \sum_g V_g \exp[2\pi i \boldsymbol{g}\cdot\boldsymbol{r}], \qquad \psi_k(\boldsymbol{r}) = \sum_g C_g(\boldsymbol{k}) \exp[2\pi i(\boldsymbol{k}+\boldsymbol{g})\cdot\boldsymbol{r}].$$

Where the summations extend over the reciprocal lattice vectors $\boldsymbol{g}$. Derive the equations satisfied by the $C_g(\boldsymbol{k})$ and hence find the forms of the four possible waves which are propagated in a cubic crystal when the 220, 200 and 020 planes are simultaneously at the reflecting position. Calculate the intensity distributions in each of the Bloch waves around the atomic positions.

Solution. We use the simplifying relationships

$$V_g = \frac{h^2}{2me} U_g, \qquad K^2 = \chi^2 + U_0.$$

Substitution of the trial solution into the wave equation yields another equation in which the sum of a set of complicated exponential terms equals zero: putting the coefficient of each exponential equal to zero gives a set of equations of the form:

$$(K^2 - (\boldsymbol{k}+\boldsymbol{g})^2) C_g(\boldsymbol{k}) + \sum_{h\neq 0} U_h C_{g-h}(\boldsymbol{k}) = 0.$$

In the present case for convenience we write the $C_g(\boldsymbol{k})$ as C_0, C_1, C_2 and C_3, and using the fact that the end of $-\boldsymbol{k}_0$ lies above the centre of the square defined by the four diffraction spots, these equations can be written in matrix

form as:

$$\begin{bmatrix} K^2-k_z^2-g^2/2 & U_{200} & U_{220} & U_{200} \\ U_{200} & K^2-k_z^2-g^2/2 & U_{200} & U_{220} \\ U_{220} & U_{200} & K^2-k_z^2-g^2/2 & U_{200} \\ U_{020} & U_{220} & U_{200} & K^2-k_z^2-g^2/2 \end{bmatrix} \begin{bmatrix} C_0 \\ C_1 \\ C_2 \\ C_3 \end{bmatrix} = 0,$$

where $g = g_{200}$.

The allowed values of k_z are found by putting the determinant of the matrix equal to zero. This leads to four values for $K^2-k_z^2-g^2/2$ which are: U_{220}, U_{220}, $U_{220} \pm 2U_{200}$. There are also four corresponding column vectors of C_n. Because of the symmetry of the problem all the C_n in any one vector will have the same magnitude, differing only in phase. The relationships between them is more easily found from the fact that a rotation of 90° must leave the wave functions unaltered. This rotation multiplies each component by a factor $\exp[in\pi/2]$, where n has to be an integer to ensure that the function returns to its original form under 360° of rotation. The possible solutions are thus:

$$n=0 \begin{bmatrix} a \\ a \\ a \\ a \end{bmatrix}, \quad n=1 \begin{bmatrix} b \\ ib \\ -b \\ -ib \end{bmatrix}, \quad n=-1 \begin{bmatrix} c \\ -ic \\ -c \\ ic \end{bmatrix}, \quad n=2 \begin{bmatrix} d \\ -d \\ d \\ -d \end{bmatrix}.$$

Matching these to a unit amplitude incident plane wave at the crystal surface gives $a = b = c = d = \frac{1}{4}$. Substituting these C_n into the original form of the solution gives the following x-y dependence of the wave amplitudes:

$$\begin{aligned} n=0: &\quad \cos(2\pi x/a)\cos(2\pi y/a), \\ n=1: &\quad \sin(2\pi x/a)\cos(2\pi y/a), \\ n=-1: &\quad \cos(2\pi x/a)\sin(2\pi y/a), \\ n=2: &\quad \sin(2\pi x/a)\sin(2\pi y/a). \end{aligned}$$

The squares of these amplitudes give the intensities in the four waves relative to the atomic positions. For $n = 0$ the intensity is a maximum along the

rows of atoms and as scattering occurs mainly in this region this wave is strongly absorbed. For $n = 2$ the intensity is a minimum at this point and this wave is thus well channelled. The waves for $n = \pm 1$ are not so strongly scattered as the wave with $n = 0$ but are less well channelled than the fourth wave.

Problem 8 – A certain (hypothetical) absorption process gives rise to a uniform probability of absorption in a cube of side αd centred on each atom, where d is the distance between the Bragg planes and $\alpha < 1$. What is the ratio of the absorption distance of the well-transmitted Bloch wave to that of the strongly absorbed wave at the exact reflecting orientation? What happens to the ratio as α tends to zero?

Solution. The ratio required is the ratio of the absorption probability of the strongly absorbed wave to that of the well-transmitted one. The strongly absorbed wave at the reflecting position has the form:

$$A \exp [2\pi i k_z z] \cos(\pi x/d) .$$

The probability of this wave being absorbed is given by the product of the wave intensity and the local absorption probability, integrated over some appropriate volume such as the unit cell. This becomes, since the absorption probability is zero outside the cube of side αd:

$$\text{probability} \propto \underset{\text{cube side } \alpha d}{\iiint} A^2 \cos^2(\pi x/d)\, \mathrm{d}x\, \mathrm{d}y\, \mathrm{d}z = A^2(\alpha d)^2 \left(\alpha d + \frac{d}{\pi} \sin(\alpha\pi)\right) .$$

For the other wave this quantity is $A^2(\alpha d)^2(\alpha d - (d/\pi)\sin(\alpha\pi))$.

The ratio is thus $(\alpha\pi + \sin(\alpha\pi))/(\alpha\pi - \sin(\alpha\pi))$. This tends to infinity as α tends to zero, corresponding to strong localisation of the scattering process about each atom and hence negligible absorption of the well-transmitted wave.

Problem 9 – A crystal of nickel is prepared in the form of a wedge with its upper surface parallel to (100) and its lower surface to (110). Electrons of energy 100 keV enter through the upper surface and travel in the (001) plane perpendicular to the edge of the wedge, falling on the (020) planes at the exact Bragg angle. If the extinction distance for this reflection is 250 Å calculate the angular splitting of the reflected beam due to refraction as it leaves through the lower surface.

Solution. We assume that the angles involved are sufficiently small that the curved dispersion surface can be approximated by a plane. The relevant section through the surface containing the points corresponding to the waves excited is thus as shown in Fig. 9:

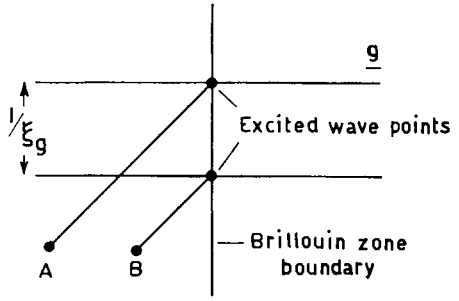


Fig. 9 – Section through dispersion surface.

Upon leaving the crystal through the lower surface waves with wave vectors which end upon A and B are generated. The separation A-B is $4 \cdot 10^{-3}$ Å^{-1}, and as $k = 27$ Å^{-1} at 100 keV the angular separation of the two reflected waves is $(4/27) \cdot 10^{-3}$ rad$= 1.46 \cdot 10^{-4}$ rad.

Problem 10 – Use a phase-grating approximation treatment to obtain the intensity distribution produced at the reflecting position by two superposed crystals each of thickness t which have potentials in the x-y plane given respectively by:

$$V_0 + V_1 \cos(2\pi g_1 x); \qquad V_0 + V_1 \cos(2\pi g_2 x).$$

Solution. The wave at the bottom of the crystal is given by:

$$\psi(x, y, z) = \exp[2\pi i \boldsymbol{\chi} \cdot \boldsymbol{r}] \exp\left[2\pi i \int_0^t \frac{m}{h^2 \chi} V(x, y, z)\, \mathrm{d}z\right].$$

Substituting the given potential and expanding the exponential (t small):

$$\psi(x, y, z) = \exp[2\pi i \boldsymbol{\chi} \cdot \boldsymbol{r} + 4\pi i m V_0 t / h^2 \chi] \left\{1 + \frac{4\pi i m V_1 t}{h^2 \chi} \cos(2\pi \bar{g} x) \cos(\pi \Delta g x)\right\},$$

where $\bar{g} = (g_1 + g_2)/2$, $\Delta g = g_1 - g_2$; expanding the first cos function as the sum of two exponentials and picking out the term in $\exp[2\pi i(\boldsymbol{\chi} + \boldsymbol{g})\cdot\boldsymbol{r}]$ gives for the amplitude of one of the diffracted waves:

$$\text{diffracted amplitude} = (2\pi i m V_1 t/h^2\chi)\cos(\pi\Delta g x)\,.$$

The intensity of the image formed using this beam thus oscillates with position across the specimen as $\cos^2(\pi\Delta g x)$, appearing as moiré fringes. The direct beam intensity can be found as the initial intensity less the intensity in the two diffracted beams.

Problem 11 – A crystal of thickness t has a phase-grating potential given by:

$$V_0 + V_1\cos(2\pi g_x x) + V_1\cos(2\pi g_y y) + V_2\cos(2\pi(g_x x + g_y y))\,.$$

Calculate the intensity of the diffracted beam having $\boldsymbol{g} = (g_x, g_y, 0)$.

Solution. At the bottom of the crystal we write the wave as

$$\psi(\boldsymbol{r}) = \exp[2\pi i\boldsymbol{\chi}\cdot\boldsymbol{r} + 2\pi i m V_0 t/h^2\chi]\cdot$$
$$\cdot\exp\left[\frac{2\pi i m V_0 t}{h^2\chi}\left(V_1\cos(2\pi g_x x) + V_1\cos(2\pi g_y y) + V_2\cos(2\pi g_x x + 2\pi g_y y)\right)\right].$$

The second exponential is expanded as before except that second order terms are now included in order to take account of the scattering to $(g_x, g_y, 0)$ via g_x and then g_y or *vice versa*. The main terms which can give intensity in the beam under consideration are:

$$\frac{2\pi i m t}{h^2\chi}\left[V_2\cos\left(2\pi(g_x x + g_y y)\right) + \frac{2\pi i m t}{h^2\chi}\left(V_1^2\cos(2\pi g_x x)\cos(2\pi g_y y)\right)\right].$$

Now $\cos(2\pi g_x x)\cos(2\pi g_y y)$ can be written as

$$\tfrac{1}{2}\left(\cos(2\pi(g_x x + g_y y)) + \cos(2\pi(g_x x - g_y y))\right).$$

Hence the coefficient of the $\cos(2\pi(g_x x + g_y y))$ term is:

$$\frac{2\pi i m t}{h^2\chi}\left(V_2 + \frac{\pi i m t}{h^2\chi}V_1^2\right)$$

and the intensity of the diffracted beam is therefore

$$\frac{\pi^2 m^2 t^2}{(h^2\chi)^2}\left(V_2^2 + \frac{\pi^2 m^2 t^2}{(h^2\chi)^2} V_1^4\right).$$

Problem 12 – Draw an edge dislocation in a foil with $\boldsymbol{g}\cdot\boldsymbol{b} = 1$, showing the Bragg planes. On which side of the dislocation does the image lie for orientations of the matrix such that *a*) $s > 0$ and *b*) $s < 0$?

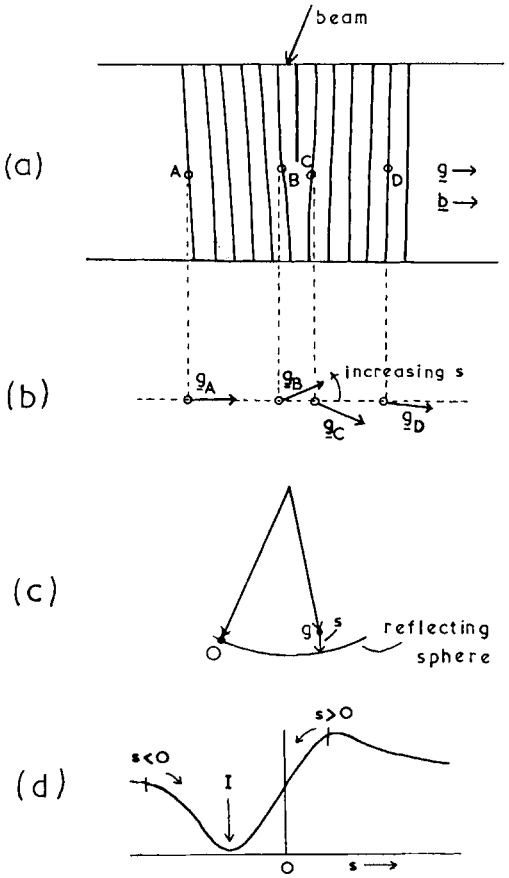


Fig. 12. – *a*) Schematic section of foil showing reflecting planes near an edge dislocation with $\boldsymbol{g}\cdot\boldsymbol{b} = 1$. *b*) Local orientation of reflection vector $\boldsymbol{g}$. *c*) Reflecting sphere. *d*) Typical bright field rocking curve.

Solution. The edge dislocation in a foil is drawn schematically in Fig. 12*a*), where the directions of the electron beam, $\boldsymbol{g}$ and $\boldsymbol{b}$ are indicated. The image position may be found by reference to the lower sections of the Figure. Figure 12*b*) shows how the reciprocal lattice varies locally around the dislocation, rotating in both directions (positions *B* and *C*) from the matrix orientation (*A* and *D*, slightly different due to the insertion of the extra half plane). The reflecting sphere (Fig. 12*c*)), however, is fixed in reciprocal space for the particular matrix orientation considered. Thus the effect of the local rotation is to increase or decrease the local value of the deviation parameter *s*. Inspection of a typical rocking curve (Fig. 12*d*)) shows that the dark line of the bright field dislocation image occurs for a local orientation corresponding to *I* (*s* negative). Thus for case *a*), where $s > 0$, the image *I* will be found to the right of the dislocation, at *C*, say, as the sense of the required rotation is negative. In case *b*), where $s < 0$, the image will be to the left, at *B*, say.

Problem 13 – Draw a coherent misfitting sphere in a foil showing the Bragg planes. Why is there a line of no contrast perpendicular to $\boldsymbol{g}$ through the centre of the defect? Show that there will always be such a line of no contrast for displacements which are symmetrical with respect to the Bragg plane through the centre of the defect.

Solution. The Bragg planes through and surrounding a coherent misfitting sphere are shown schematically in Fig. 13, where $\boldsymbol{g}$ is in the plane of the paper. A typical displacement $\boldsymbol{R}$ is shown. It may be seen that $\boldsymbol{R}$ is zero everywhere on the plane *AA*. As contrast is caused by the displacements $\boldsymbol{R}$ of the atoms from their ideal positions it follows that there must be a line of no contrast where the plane *AA* cuts the plane of projection of the image (perpendicular to the electron beam direction), *i.e.* a line perpendicular to $\boldsymbol{g}$. A similar argument follows for any symmetrical strain field; the plane *AA* is, by definition, displacement-free and thus must produce a line of no contrast.

Problem 14 – Show that the moiré fringe spacing which arises from the two spots $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ as in the diffraction pattern of Fig. 14 is $1/|\boldsymbol{g}_1 - \boldsymbol{g}_2|$. Two similar lattices rotated by a small angle $\delta\theta$ give rise to a rotation moiré pattern. If this is misinterpreted by assuming that the lattices have different spacings but are parallel (parallel moiré) what lattice parameter difference would be deduced?

Solution. If the amplitudes of the two beams contributing to the diffraction spots are φ_1 and φ_2 respectively, then the total dark field disturbance Ψ

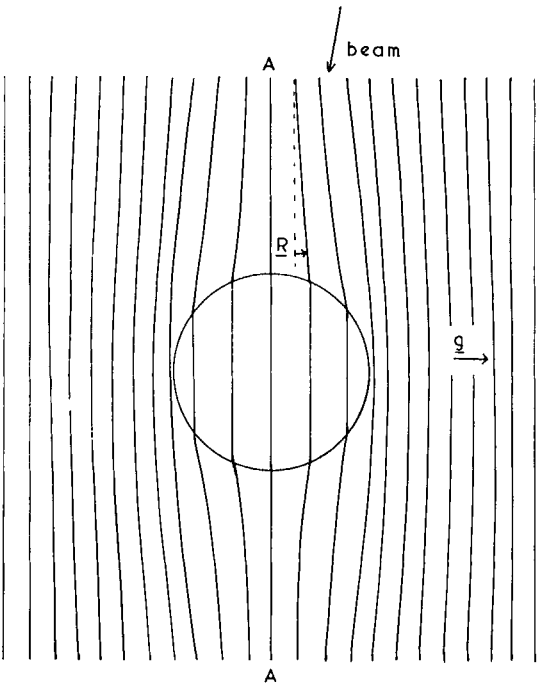


Fig. 13. – Schematic section of foil showing reflecting planes near a coherent misfitting sphere.

is given by

$$\Psi(\boldsymbol{r}) = \varphi_1 \exp\left[2\pi i(\boldsymbol{k} + \boldsymbol{g}_1)\cdot\boldsymbol{r}\right] + \varphi_2 \exp\left[2\pi i(\boldsymbol{k} + \boldsymbol{g}_2)\cdot\boldsymbol{r}\right] \tag{14.1}$$

and the intensity $I(\boldsymbol{r})$ by

$$I(\boldsymbol{r}) = |\Psi(\boldsymbol{r})|^2 =$$

$$= |\varphi_1 \exp\left[2\pi i(\boldsymbol{k} + \boldsymbol{g}_1)\cdot\boldsymbol{r}\right](1 + R\exp\left[2\pi i(\delta + (\boldsymbol{g}_2 - \boldsymbol{g}_1)\cdot\boldsymbol{r}\right])|^2 \tag{14.2}$$

$$= |\varphi_1|^2(1 + R^2 + 2R\cos 2\pi(\delta + (\boldsymbol{g}_2 - \boldsymbol{g}_1)\cdot\boldsymbol{r}))\,, \tag{14.3}$$

where $R \exp[i\delta] = \varphi_2/\varphi_1$.

Equation (14.3) is the equation of fringes of spacing $1/|\boldsymbol{g}_2 - \boldsymbol{g}_1|$ running perpendicular to $\delta\boldsymbol{g} = \boldsymbol{g}_2 - \boldsymbol{g}_1$. The bright field image, being complementary, exhibits similar fringes.

In the case of a rotation of two similar lattices by a small angle $\delta\theta$, $\delta g = g\,\delta\theta$. If this is misinterpreted as a parallel moiré with lattice parameter difference δa in a, then, since $g = 1/a$, it follows that $\delta a/a = \delta\theta$.

Problem 15 – The micrograph of Fig. 15 of Sect. **2** was taken using an objective aperture of size and position indicated on the inset diffraction pattern (which is correctly oriented). Explain the nature of the closely spaced fringes on the micrograph and calculate the size of the unit cell (cubic) of the specimen material. [Micrograph courtesy of I. L. F. Ray.]

Solution. The fringes are so-called lattice fringes produced under favourable conditions when two or more beams (000 and 111 here) combine to form the image. If the amplitudes (complex) of the two beams emerging from the bottom of the crystal are φ_0 and φ_g then as a function of position $\boldsymbol{r}$ in this surface the total disturbance Ψ is (in a similar way to the moiré fringes of *Problem 14* above)

$$\Psi(\boldsymbol{r}) = \varphi_0 \exp\left[2\pi i \boldsymbol{k}\cdot\boldsymbol{r}\right] + \varphi_g \exp\left[2\pi i(\boldsymbol{k}+\boldsymbol{g})\cdot\boldsymbol{r}\right]. \tag{15.1}$$

Hence the intensity distribution $I(r)$ is given by

$$I(r) = |\Psi|^2 = |\varphi_0|^2(1 + R^2 + 2R\cos(2\pi g r + \delta)), \tag{15.2}$$

where $R\exp[i\delta] = \varphi_g/\varphi_0$ and r is measured parallel to $\boldsymbol{g}$.

Equation (15.2) shows that the intensity distribution is cosine periodic with spacing $\Delta r = 1/g$ between maxima, the « lattice fringe » visibility being greatest when $R = 1$, *i.e.* $|\varphi_0| = |\varphi_g|$. The fringes in the micrograph of Fig. 15 have spacing $\Delta r = 0.01 \cdot 10^{-6}/32$ m leading to a value for the unit cell side $a = \Delta r\sqrt{3} = 0.01 \cdot 10^{-6} \times \sqrt{3}/32\,\text{m} = 5.41$ Å, which agrees very well with the tabulated value of 5.417 Å for silicon (the specimen material).

It should be noted that although the periodicity of the fringes accurately follows the internal periodicity of the lattice planes the positions of the fringes do not necessarily correspond to the positions of actual lattice planes, which is a most important reservation when the distortions of lattice planes near dislocations, for example, are being considered. The lattice images produced by such defects must be carefully computed (see, for example Fig. 13 of M. J. Goringe: « Computing Methods », this volume).

Problem 16 – The micrograph (Fig. 16.1 of Sect. **2**) of copper-10 at % aluminium (fcc) provided was taken using an objective aperture of size and

position indicated on the inset diffraction pattern (which is correctly oriented). The micrograph (which is a negative print) is of a foil with normal very near [111] and the dislocation is dissociated according to the scheme

$$\boldsymbol{b} = \frac{a}{2}[1\bar{1}0] \rightarrow \boldsymbol{b}_1^P + \boldsymbol{b}_2^P = \frac{a}{6}[1\bar{2}1] + \frac{a}{6}[2\bar{1}\bar{1}] \tag{16.1}$$

and near A the dislocation line direction $\boldsymbol{u}$ is parallel to $[11\bar{2}]$.

a) Calculate $\boldsymbol{g}\cdot\boldsymbol{b}_1^P$, $\boldsymbol{g}\cdot\boldsymbol{b}_2^P$ and $\boldsymbol{g}\cdot\boldsymbol{R}$ for the two partial dislocations and the connecting stacking fault. Use the results to explain the visibility of both partials and the absence of stacking fault contrast.

b) Assuming that the dark lines accurately define the positions of the partials estimate the stacking fault energy γ, of the material from the formula

$$\gamma = \frac{\mu b^2}{24\pi\Delta}\cdot\frac{(2-\nu)}{(1-\nu)}\left[1 - \frac{2\nu\cos 2\alpha}{(2-\nu)}\right], \tag{16.2}$$

where μ is the shear modulus, ν the Poisson ratio, Δ the separation of the partials, $\boldsymbol{b}$ the total Burgers vector of the dislocation, and α the angle between $\boldsymbol{b}$ and $\boldsymbol{u}$.

c) Note that eq. (16.2) predicts values of Δ depending on $\boldsymbol{u}$. Confirm this variation by measurements near A and B.

d) Calculate the value of the deviation parameter $s_{2\bar{2}0}$ for the dark field beam used to form the micrograph.

e) Using the co-ordinate system of Fig. 16.2, in Sect. **2** for the total dislocation in edge orientation we have at point B a displacement $\boldsymbol{R} = \boldsymbol{R}_1^P + \boldsymbol{R}_2^P$ such that

$$\boldsymbol{g}\cdot\boldsymbol{R}_i^P = \frac{\boldsymbol{g}\cdot\boldsymbol{b}_i^P}{2\pi}\left(\theta_i + \frac{\sin 2\theta_i}{4(1-\nu)}\right), \qquad i = 1, 2. \tag{16.3}$$

The observed positions of the « weak-beam » peaks occur for values of x and z where

$$s_g + (\mathrm{d}/\mathrm{d}z)(\boldsymbol{g}\cdot\boldsymbol{R}) = 0 \tag{16.4}$$

and

$$(\mathrm{d}^2/\mathrm{d}z^2)(\boldsymbol{g}\cdot\boldsymbol{R}) = 0. \tag{16.5}$$

Show that the observed separation Δ_{obs} is given by

$$\Delta_{obs} = (\Delta^2 + 4/c^2)^{\frac{1}{2}}, \tag{16.6}$$

where

$$c = -s_g \Big/ \left[\frac{\boldsymbol{g}\cdot\boldsymbol{b}^P}{2\pi}\left(1 + \frac{1}{2(1-\nu)}\right)\right] \tag{16.7}$$

and

$$\boldsymbol{g}\cdot\boldsymbol{b}^P = \boldsymbol{g}\cdot\boldsymbol{b}_1^P = \boldsymbol{g}\cdot\boldsymbol{b}_2^P .$$

f) Using the results of *d*) and *e*) estimate the correction necessary to the value of γ calculated in *b*).

Assume $a = 3.66$ Å, $\mu = 4\cdot10^{11}$ dyne cm^{-2}, $\nu = \frac{1}{3}$ and electron wavelength $\lambda = 0.032$ Å. Note that *e*) entails considerable algebra and only the results are required for *f*).

Solution.

a) $\boldsymbol{g} = (1/a)[2\bar{2}0]$, $\boldsymbol{b}_1^P = (a/6)[1\bar{2}1]$, $\boldsymbol{b}_2^P = (a/6)[2\bar{1}\bar{1}]$, $\boldsymbol{R} = (a/3)[111]$

therefore $\boldsymbol{g}\cdot\boldsymbol{b}_1^P = 1$, $\boldsymbol{g}\cdot\boldsymbol{b}_2^P = 1$, $\boldsymbol{g}\cdot\boldsymbol{R} = 0$.

Hence both partial dislocations are visible by kinematical theory and the stacking fault is invisible.

b) Measurements near A yield $\Delta \simeq 120$ Å and $\alpha = 90°$.

Hence $\gamma = 10.4$ erg cm^{-2} by substitution into eq. (16.2)

c) From eq. (16.2)

$$\Delta(\alpha) \propto 2 - \nu - 2\nu\cos 2\alpha .$$

Near B, $\alpha \simeq 45°$ and near A, $\alpha = 90°$

$$\frac{\Delta^B}{\Delta^A} = \frac{2 - \frac{1}{3}}{2 + \frac{1}{3}} = \frac{5}{7} .$$

Measured values of $\Delta^B_{obs}/\Delta^A_{obs}$ are in reasonable agreement.

d) $s_{2\bar{2}0} = -g^2/k = -g^2\lambda = -1.91\cdot10^{-2}$ Å^{-1} ($2\bar{2}0$ is outside the Ewald sphere therefore negative s_g).

e) For this particular case eq. (16.5) is satisfied for all x at $z = z_1$ by inspection, *i.e.* $\theta_1 = \theta_2 = 0$ and $\boldsymbol{g}\cdot\boldsymbol{b}_1^P = \boldsymbol{g}\cdot\boldsymbol{b}_2^P$ (solution *a*)). Differentiating eq. (16.3) with respect to z and substituting into eq. (16.4) with these simplifying factors yields a quadratic for x with solution

$$x = (2 + c\Delta \pm \sqrt{4 + c^2\Delta^2})/2c\,. \tag{16.8}$$

Hence $\Delta_{\text{obs}} = x_1 - x_2$, yielding eqs (16.6) and (16.7).

f) Note that as $s \to \infty$, $\Delta_{\text{obs}} \to \Delta$, $x_1 \to 0$ and $x_2 \to \Delta$. In this case $s = -1.91 \cdot 10^{-2}$ Å^{-1}, $\boldsymbol{g}\cdot\boldsymbol{b}^P = 1$, $\nu = \frac{1}{3}$ therefore $c = 6.85 \cdot 10^{-2}$. Thus $\Delta \simeq \simeq \Delta_{\text{obs}}(1 - 2/c^2\Delta_{\text{obs}}^2)$, *i.e.* $\Delta = \Delta_{\text{obs}}(1 - 3 \cdot 10^{-2})$. Hence the value of γ should be increased by 3%, which is well within the error of measurement.

Additional notes on computed results. This problem is an example of application of the high resolution « weak beam » technique, which has recently been discussed by Cockayne, Ray and Whelan ([4]) and Cockayne ([5]). The latter has carried out a number of calculations to check the accuracy of the image positions, the results being outlined below (figures courtesy D. J. H. Cockayne).

For weak-beam calculations it would be reasonable to apply the kinematical theory of image contrast (Hirsch, Howie and Whelan ([6])). Performing the kinematical integral $\varphi_g = \int \exp[2\pi i(s_g z + \boldsymbol{g}\cdot\boldsymbol{R})]\,dz$ over a range of columns near a dislocation yields curves of the form of Fig. 16.3, from which the

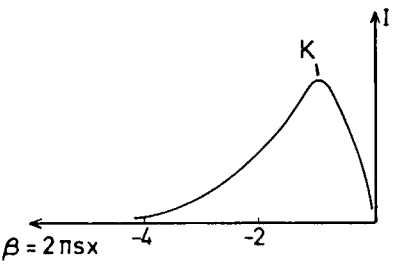


Fig. 16.3. – Typical kinematical image profile for a screw dislocation (Hirsch, Howie and Whelan, ([6])) with kinematical peak, *K*, indicated.

position *K* of the kinematical image may be found. Equations (16.4) and (16.5) for the position *W* of a weak-beam peak near a similar dislocation may be explained in terms of the columns of Fig. 16.4. The weak beam only acquires

intensity when $(s_g + (\mathrm{d}/\mathrm{d}z)(\boldsymbol{g}\cdot\boldsymbol{R})) = (s_g + \beta'_g) = D$ is small but by the same token may not lose any intensity gained when D is again large. For all columns to one side of the dislocation (right in Fig. 16.4) D is always large, giving a very low intensity in the weak beam. The same is true far to the other side.

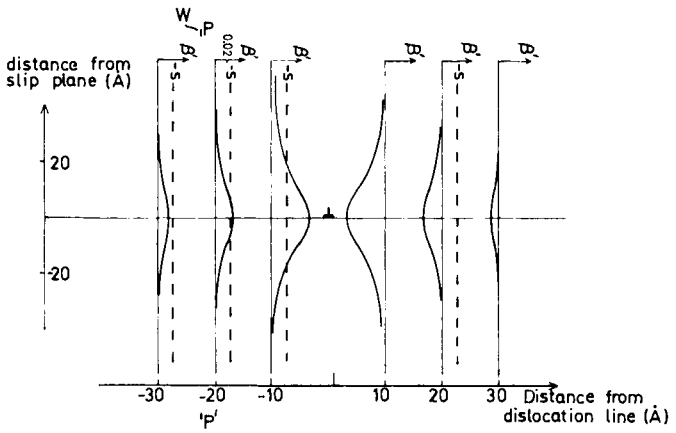


Fig. 16.4. – The value of β'_g down columns at various distances from a dislocation. Also indicated is the value of s_g ($-2.5\cdot10^{-2}$ Å^{-1}) for a typical weak-beam. The weak-beam peak, W, is expected to occur for the column PP'.

Near the dislocation, however, D is sometimes zero. The column for which D is small for the greatest depth is the one for which $s_g + \beta'_g = 0$ at a turning point of β'_g, *i.e.* when $(\mathrm{d}^2/\mathrm{d}z^2)(\boldsymbol{g}\cdot\boldsymbol{R}) = 0$ and this column (PP') gives the position W of the weak-beam peak. The exact theoretical position of the peak is found by integrating the n-beam dynamical equations for the orientation under consideration (see M. J. Goringe: « Computing Methods », this volume). The results of six-beam calculations are shown in Fig. 16.5 for an undissociated edge dislocation and in Fig. 16.6 for a dissociated edge dislocation in copper with partials separated by 50 Å. Inspection of Fig. 16.5 shows that the true peak lies between W and K, its exact position depending on dislocation depth and foil thickness. In Fig. 16.6 it can be seen that the separation of the two calculated peaks is within 7 Å of the separation given by eqs (16.4) and (16.5) (peak positions marked W). This last Figure also shows how the relative magnitudes of the two peaks varies with the depth of the dislocation in the foil, including, of course, the possibility of zero intensity in one « peak » (a possible cause of the varying visibility in the experimental micrograph of Fig. 16.1 of Sect. **2**).

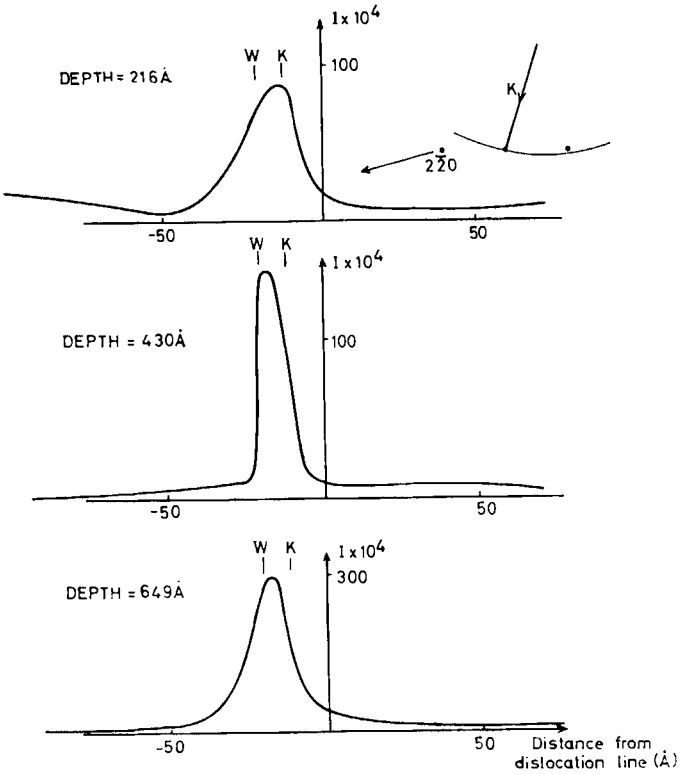


Fig. 16.5. – $g=2\bar{2}0$ weak-beam images of an undissociated edge dislocation with $b=(a/2)[1\bar{1}0]$ at various depths in a copper foil of thickness 860 Å. $s_{2\bar{2}0} = -2.47 \cdot 10^{-2}$ Å^{-1}, 100 keV electrons, six-beam calculation. *W* is the weak-beam position from eqs (16.4) and (16.5) and *K* the kinematical peak position.

The weak-beam technique, exemplified by this problem, obviously has a wide range of applications in the study of details of strain fields near defects (see Cockayne, Ray and Whelan ([4]) for some suggested topics). The accuracy achieved using the relatively simple criteria of eqs (16.4) and (16.5) is, of course, one of the most attractive features of the technique from the point of view of the experimentalist.

Problem 17 – Calculate:

a) the maximum density of dislocations that can be observed under 2-beam bright field conditions in copper (assume foil thickness 2000 Å; 100 keV electrons);

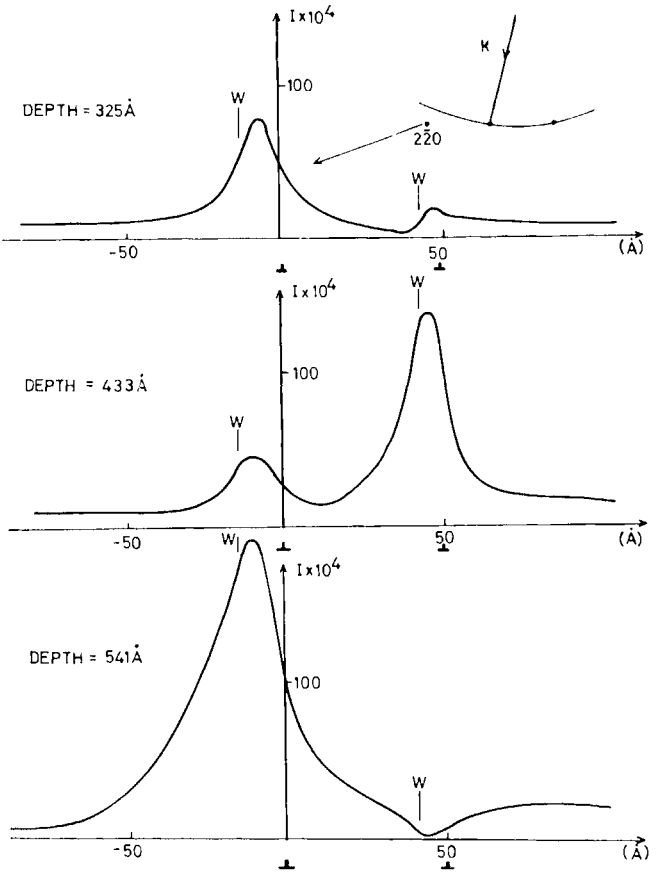


Fig. 16.6. – As Fig. 16.5 except the edge dislocation is dissociated according to the scheme in the text into two partials separated by 50 Å.

b) the maximum stacking fault energy, γ, that can be measured by the node method, *i.e.* by applying the formula $\gamma w/\mu b^2 = \frac{1}{3}$, where w is the width of the node as defined in Fig. 17 of Sect. **2**, μ the shear modulus and ***b***

Material	Extinction distances (Å)			a (Å)	μ (dync cm^{-2})
	111	200	220		
Cu	242	281	416	3.608	$4 \cdot 10^{11}$
Au	159	179	248	4.070	$2.8 \cdot 10^{11}$

the total Burgers vector of the dislocations involved. In copper and gold no extended nodes are seen. What are the minimum values for γ_{Au} and γ_{Cu}? (Data as per Table.)

Solution. Under the operating conditions defined, the width of the image of a dislocation image is $\sim \xi_g/3$. Thus in both *a*) and *b*) the best results will be obtained if it is possible to work with a reflexion $g = 111$.

a) An absolute upper limit to the dislocation density l ($cm \cdot cm^{-3}$) measurable in a foil of thickness t occurs when the dislocation images overlap, *i.e.* when fractional « image » area $= 1$. Thus $l_{max} \times t \times$ image width $= 1$, giving $l_{max} = 6 \cdot 10^{10}\ cm \cdot cm^{-3}$ by substitution of $t = 2000$ Å, image width $= \xi_{111}/3$. A more realistic upper limit might be about half this value.

b) Roughly the node width, w, must be greater than the dislocation image width for an extended node to be visible, *i.e.* $w > \xi_g/3$. Thus, if nodes are not visible, $\gamma_{min} = \mu b^2/\xi_g$. In order to make both partials of at least two arms of the node visible it is convenient to work with $g = 2\bar{2}0$ (see *Problem 16* above), when the stacking fault is invisible and, being on (111), is perpendicular to the beam and hence no geometrical correction factors are required. Substitution of values for the two materials yields $\gamma_{min}(Au) = 58.5\ erg\ cm^{-2}$ and $\gamma_{min}(Cu) = 61\ erg\ cm^{-2}$.

Problem 18 – In the light of your answers to *Problem 17* do you think that the resolution of the electron microscope limits present-day observations in solid-state physics?

Solution. Present-day electron microscopes have a resolution under ideal conditions in the region of a few Ångström units, which is well below the resolution limits set by image widths under the diffraction contrast conditions employed in studies in solid-state physics. Far more important for most studies are the deleterious effects produced by nonideal specimens, *e.g.* energy losses in thick specimens, which degrade the image through the chromatic aberration of the lenses. Assuming that the necessary theory could be developed to take inelastic scattering into account much more might be gained by the development of achromatic lenses of otherwise moderate resolution than by refinement of the lens for ideal operation. Of course, certain studies in the solid state (and many in biology) require the ultimate resolution *e.g.* lattices fringes (see *Problem 15* above) and their use in the study of dislocation cores, certain nucleation studies (see *e.g.* C.R.Hall: « Contrast calculations for small clusters of atoms », this volume), etc.

Problem 19 – A specimen of thickness t contains a small inclusion of thickness Δz which scatters electrons differently from the matrix. The crystal is

viewed under 2-beam dynamical conditions at the Bragg position ($s = 0$) and the effective extinction distances are ξ_{gm} for the matrix and ξ_{gi} for the inclusion, which may be assumed to have the same crystal structure as the matrix. Calculate the visibility of the inclusion as a function of *a*) its depth in the specimen and *b*) the specimen thickness. How would you expect these results to be modified if the specimen is not precisely at the Bragg position (*i.e.* $s \neq 0$)?

Solution. The situation considered is shown schematically in Fig. 19*a*). The amplitudes of the waves transmitted through column 1, which does not contain the inclusion are $\boldsymbol{\varphi}_1$, where

$$\boldsymbol{\varphi}_1 = \boldsymbol{P}\boldsymbol{\varphi}_0 , \tag{19.1}$$

while the corresponding transmission through column 2 is

$$\boldsymbol{\varphi}_2 = \boldsymbol{P}_3\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{\varphi}_0 . \tag{19.2}$$

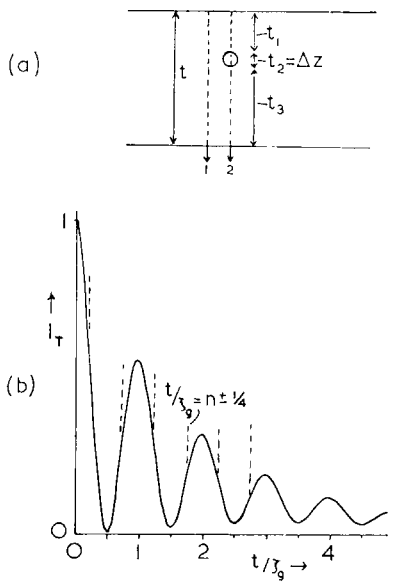


Fig. 19. – *a*) Section of foil thickness t containing an inclusion of thickness Δz at depth t_1. Columns 1 and 2 considered in the text are indicated. *b*) Typical bright field thickness fringe profile showing position of greatest visibility for « structure-factor » contrast.

In eqs (19.1) and (19.2) the scattering matrices $\boldsymbol{P}$ (see « Computing Methods », this volume) are of the form

$$\boldsymbol{P}_j = \begin{pmatrix} \cos\beta_j/2 & \sin\beta_j/2 \\ -\sin\beta_j/2 & \cos\beta_j/2 \end{pmatrix} \begin{pmatrix} \exp[2\pi i\gamma_j^{(2)} t_j] & 0 \\ 0 & \exp[2\pi i\gamma_j^{(1)} t_j] \end{pmatrix} \cdot \begin{pmatrix} \cos\beta_j/2 & -\sin\beta_j/2 \\ \sin\beta_j/2 & \cos\beta_j/2 \end{pmatrix}, \tag{19.3}$$

where $j = 1, 2, 3$, (or blank) and $\operatorname{ctg}\beta_j = s_j\xi_{gj} = w_j$.

For the principal case under consideration

$$\xi_{g3} = \xi_{g1} = \xi_g = \xi_{gm}, \quad \xi_{g2} = \xi_{gi} \quad \text{and} \quad s_j = 0 \text{ (all } j).$$

Under these conditions $\beta_j = \pi/2$ (all j) and the matrices involving trigonometric function simplify considerably. Multiplying out eq. (19.2)

$$\boldsymbol{\varphi}_2 = \begin{pmatrix} a & a \\ -a & a \end{pmatrix} \begin{pmatrix} \exp[2\pi i((t_1+t_3)\gamma_m^{(1)} + t_2\gamma_i^{(1)})] & 0 \\ 0 & \exp[2\pi i((t_1+t_3)\gamma_m^{(2)} + t_2\gamma_i^{(2)})] \end{pmatrix} \cdot \begin{pmatrix} a & -a \\ a & a \end{pmatrix} \boldsymbol{\varphi}_0, \tag{19.4}$$

where $a = \sin(\pi/4) = \cos(\pi/4)$. Equation (19.1) becomes

$$\boldsymbol{\varphi}_1 = \begin{pmatrix} a & a \\ -a & a \end{pmatrix} \begin{pmatrix} \exp[2\pi i t\gamma_m^{(1)}] & 0 \\ 0 & \exp[2\pi i t\gamma_m^{(2)}] \end{pmatrix} \begin{pmatrix} a & -a \\ a & a \end{pmatrix} \boldsymbol{\varphi}_0. \tag{19.5}$$

Equations (19.4) and (19.5) are of the same form, that of perfect crystal, but with different thicknesses. Writing t' as the thickness of perfect matrix crystal which is equivalent to the composite crystal it can be seen that

$$\Delta\gamma_m t' = \Delta\gamma_m(t_1 + t_3) + \Delta\gamma_i t_2, \tag{19.6}$$

where $\Delta\gamma = |\gamma^{(1)} - \gamma^{(2)}|$, or, since $\Delta\gamma = 1/\xi_g$

$$t' = t_1 + t_3 + t_2\frac{\xi_{gm}}{\xi_{gi}}, \quad \text{or} \quad t' = t + \Delta z\left(\frac{\xi_{gm}}{\xi_{gi}} - 1\right). \tag{19.7}$$

Inspection of the typical thickness fringe profiles of Fig. 19*b*) shows that small changes in the effective thickness will be most visible when $t/\xi_{gm} = n \pm \frac{1}{4}$, for small values of the integer n. The inclusion will then appear darker or lighter than background alternately on opposite sides of thickness fringes, the start of the sequence depending on the sign of $(\xi_{gm}/\xi_{gt} - 1)$. In the condition where $s = 0$ this « structure factor » contrast is completely independent of the depth of the inclusion in the foil.

When $s \neq 0$ the trigonometric matrices do not simplify as before since $\beta = \beta_1 = \beta_3 \neq \beta_2$. So, in addition to the structure factor effect discussed above, phase factors are introduced by the inclusion similar to those present at a stacking fault. There is thus a depth-dependent contribution to the contrast of a rather complicated form.

Problem 20 – a) Show that under 2-beam conditions the change in the transmitted intensity due to an imperfection is proportional to the real part of the change in the well-transmitted Bloch wave amplitude. (Assume that the crystal is thick enough to reduce the amplitude of the absorbed Bloch wave to zero).

b) Given the expression governing the scattering of Bloch waves into the well-transmitted waves (here denoted $\psi^{(1)}$),

$$\frac{\mathrm{d}\psi^{(1)}}{\mathrm{d}z} = 2\pi i \beta'_g \left[\sin^2 \frac{\beta}{2} \psi^{(1)} - \tfrac{1}{2} \sin\beta \exp\left[2\pi i \Delta k z\right] \psi^{(2)} \right] \tag{20.1}$$

(see this volume, Howie eq. (51), Goringe eq. (30); or [3], eq. (12.25)); making the following substitutions to eliminate intraband scattering,

$$\psi^{(1)} = \Psi^{(1)} \exp\left[2\pi i \sin^2(\beta/2) \boldsymbol{g} \cdot \boldsymbol{R}\right] \tag{20.2}$$

and

$$\psi^{(2)} = \Psi^{(2)} \exp\left[2\pi i \cos^2(\beta/2) \boldsymbol{g} \cdot \boldsymbol{R}\right], \tag{20.3}$$

using the fact that $\beta'_g = (\mathrm{d}/\mathrm{d}z)(\boldsymbol{g} \cdot \boldsymbol{R})$ and integrating (assuming $\Psi^{(2)}$ constant over integration), show that the effect of an imperfection at depth z_0 is

$$\Delta\Psi^{(1)} = -\Psi^{(2)} \pi i \sin\beta \exp\left[2\pi i \Delta k z_0\right] \int_{z - z_0 = -\infty}^{z - z_0 = \infty} \beta'_g(z - z_0) \cdot \exp\left[2\pi i \left[\Delta k (z - z_0) + \boldsymbol{g} \cdot \boldsymbol{R} \cos\beta\right]\right] \mathrm{d}(z - z_0). \tag{20.4}$$

Hence show that if $\beta_g'(z-z_0)$ is odd then $\mathrm{Re}(\Delta\Psi^{(1)})$ is proportional to $\cos 2\pi\Delta k z_0$ and if $\beta_g'(z-z_0)$ is even then $\mathrm{Re}(\Delta\Psi^{(1)})$ is proportional to $\sin 2\pi\Delta k z_0$.

Deduce the depth variation in the visibility of *c*) stacking faults, *d*) dislocations and *e*) centres of strain.

Solution.

$$a) \qquad \varphi_0(z) = C_0^{(1)}\psi^{1)} \exp\left[2\pi i k^{(1)} z\right] + C_0^{(2)}\psi^{(2)} \exp\left[2\pi i k^{(2)} z\right]. \tag{20.5}$$

Hence, assuming $\psi^{(2)}$, $\psi^{(2)*} \to 0$

$$|\varphi_0(z)|^2 \propto |\psi^{(1)}|^2,$$

i.e.

$$\Delta|\varphi_0(z)|^2 \propto \psi^{(1)}\Delta\psi^{(1)*} + \psi^{(1)*}\Delta\psi^{(1)}.$$

But in perfect crystals $\psi^{(1)} = \psi^{(1)*} = C_0^{(1)} = \cos\beta/2$, *i.e.* real and therefore $\Delta|\varphi_0(z)|^2 \propto \Delta\psi^{(1)*} + \Delta\psi^{(1)} = 2\mathrm{Re}(\Delta\psi^{(1)})$.

b) Differentiating eq. (20.2)

$$\frac{\mathrm{d}\psi^{(1)}}{\mathrm{d}z} = \frac{\mathrm{d}\Psi^{(1)}}{\mathrm{d}z} \exp\left[2\pi i \sin^2(\beta/2)\boldsymbol{g}\cdot\boldsymbol{R}\right] + 2\pi i\beta_g' \sin^2(\beta/2)\psi^{(1)} \tag{20.6}$$

and substituting from eqs (20.6) and (20.3) in eq. (20.1) yields

$$\frac{\mathrm{d}\Psi^{(1)}}{\mathrm{d}z} = -\pi i\beta_g' \sin\beta\,\Psi^{(2)} \exp\left[2\pi i(\Delta k z + \boldsymbol{g}\cdot\boldsymbol{R}\cos\beta)\right]. \tag{20.7}$$

Equation (20.4) follows by change of variable $z \to z - z_0$ and integration, assuming that $\Psi^{(2)}$, $\sin\beta$ and Δk are constants, and that the defect is so far from the foil surfaces that the limits of integration may be extended to infinity without error (*i.e.* the strain field of the defect must be localized).

Inspection of eq. (20.4) shows that, at the reflecting position ($\beta = \pi/2$), the integral is of the form

$$I = \int_{-\infty}^{\infty} \beta_g'(q) \exp\left[2\pi i\Delta k q\right] \mathrm{d}q = \int_{-\infty}^{\infty} \beta_g'(q) \left[\cos 2\pi q\Delta k + i \sin 2\pi q\Delta k\right] \mathrm{d}q,$$

where $q = z - z_0$.

If $\beta'_g(q)$ is odd then $I = i\,2\int_0^\infty \beta'_g(q)\sin 2\pi q\Delta k\,\mathrm{d}q = iP$, *i.e.* imaginary. Hence $\Delta\Psi^{(1)} = -\Psi^{(2)}\pi i \exp[2\pi i\Delta k z_0]\ iP$ and since, at the top of the crystal $\Psi^{(2)}(=\psi^{(2)})$ is real and P is real, $\mathrm{Re}(\Delta\Psi^{(1)}) \propto \cos 2\pi\Delta k z_0$. Similarly if $\beta'_g(q)$ is even then $I = 2\int_0^\infty \beta'_g(q)\cos 2\pi q\Delta k\,\mathrm{d}q = Q$ and $\mathrm{Re}(\Delta\Psi^{(1)}) \propto \sin 2\pi\Delta k z_0$.

As $\psi^{(1)} = \Psi^{(1)}\exp[i\delta]$, $\mathrm{Re}(\Delta\psi^{(1)})$ is also of the same form, and hence so is the visibility of the defect.

c) Stacking fault; β'_g even (in the sense that $\int\beta'_g\,\mathrm{d}z$ is nonzero).

d) Dislocations; (*e.g.* edge dislocations of *Problem 16* above) β'_g even.

e) Centres of strain; β'_g odd ($\boldsymbol{g}\cdot\boldsymbol{R}$ even, see *Problem 13* above).

Thus the visibilities vary with depth as $\sin 2\pi\Delta k z_0$ (*c*) and *d*)) and as $\cos 2\pi\Delta k z_0$ (*e*)) as shown schematically in Fig. 20.

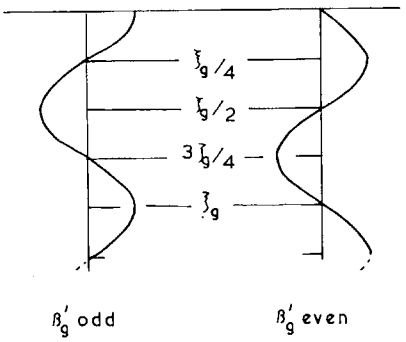


Fig. 20. – Depth variation of contrast as a function of symmetry of β'_g.

Note that this question is essentially a proof of the modified Bloch wave approach to contrast from defects developed by Wilkens ([7]).

Problem 21 - Calculate the maximum knock-on energy along *a*) [111], *b*) [100] and *c*) [110] that can be transferred to atoms in a copper target bombarded along [100] by protons of energy 5 MeV. [At. wt Cu = 63.6]

Solution. Equating momenta in 2 directions and energy (see Fig. 21) we have 3 equations:

$$\text{mtm}\ \rightarrow \qquad M_2 c\sin\theta = M_1 b\sin\phi\,, \tag{21.1}$$

$$\text{mtm}\ \downarrow \qquad M_1 a = M_2 c\cos\theta + M_1 b\cos\phi\,, \tag{21.2}$$

$$\text{energy} \qquad \tfrac{1}{2}M_1 a^2 = \tfrac{1}{2}M_2 c^2 + \tfrac{1}{2}M_1 b^2\,. \tag{21.3}$$

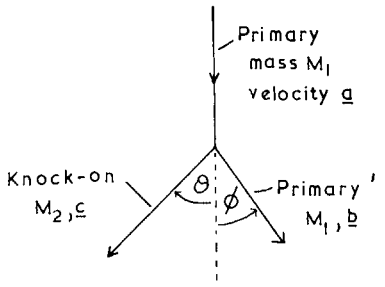


Fig. 21.

Eliminating ϕ and b from eqs (21.1)-(21.3) yields

$$c = \frac{2M_1 a \cos\theta}{M_1 + M_2} . \tag{21.4}$$

Hence knock-on energy

$$E[uvw] = E_i \Lambda \cos^2\theta , \tag{21.5}$$

where E_i = incident energy and $\Lambda = 4M_1M_2/(M_1 + M_2)^2$. The angles θ required in the three cases are the angles between [100] and [111], [100] and [110] respectively, *i.e.* $\arccos(1/\sqrt{3})$, $\arccos(1)$, $\arccos(1/\sqrt{2})$. Now $\Lambda = 4 \cdot 63.6/(64.6)^2 = 6.1 \cdot 10^{-2}$, and $E_i = 5 \cdot 10^6$ eV.

$\therefore$ *a)* $E[111] = E_i\Lambda/3 = 1.02 \cdot 10^5$ eV,

b) $E[100] = E_i\Lambda = 3.04 \cdot 10^5$ eV,

c) $E[110] = E_i\Lambda/2 = 2.03 \cdot 10^5$ eV.

Problem 22 – In copper exposed to a flux of $1 \cdot 10^{13}$ cm^{-2} s^{-1} of neutrons of energy 10 keV calculate *a)* the concentration of primary knock-ons per year, *b)* their mean energy and *c)* the total point defect concentration produced, assuming the hard-sphere model and that no recombination takes place. [Copper is fcc, $a = 3.608$ Å, atomic weight $= 63.6$, displacement energy $E_d = 19$ eV, total neutron cross-section $\sigma = 3 \cdot 10^{-24}$ cm^2].

Solution.

a) Taking cross-section $\sigma = 3 \cdot 10^{-24}$ cm^2,

neutron flux $\phi = 1 \cdot 10^{13}$ cm^{-2}s^{-1},

time $t = 60 \cdot 60 \cdot 24 \cdot 365$ s,

atom density $n = 4/(3.61 \cdot 10^{-8})^3$ cm^{-3},

we have for the concentration of primary knock-ons, C_p,

$$C_p = \sigma\phi tn = 8.05 \cdot 10^{19}\,\mathrm{cm}^{-3}.$$

b) On the hard-sphere model the differential cross-section $\sigma(\theta)\,\mathrm{d}\theta$ is $2\sigma \sin\theta \cos\theta\,\mathrm{d}\theta$, where σ = total cross-section (see Fig. 22).

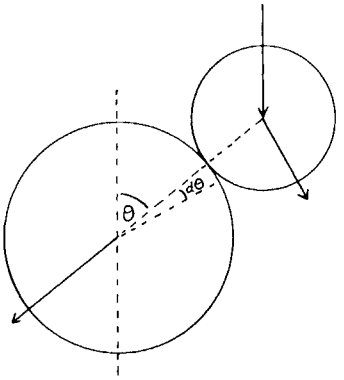


Fig. 22.

Hence $\bar{E} = \int_0^{\pi/2} E(\theta)\sigma(\theta)\,\mathrm{d}\theta / \int_0^{\pi/2} \sigma(\theta)\,\mathrm{d}\theta$ becomes $E_i/2$ by substitution of $E_i \cos^2\theta$ for $E(\theta)$ (see solution of *Problem 21* above), *i.e.* $\bar{E} = 10^4 \cdot 6.1 \cdot 10^{-2}/2 = 305$ eV.

c) The displaced atoms (with average energy $\bar{E}$) each cause further displacements by collison with other atoms. In this case the « incident » and « knock-on » have equal masses, *i.e.* $A = 1$ and so, on average, the energy is shared equally between the two. Hence, on average, the initial knock-on energy will be distributed equally over a number of atoms until the average energy falls below $2E_d$ (where E_d is the energy required to displace an atom), after which no additional knock-ons can be created. Hence the average number of knock-ons per primary event is $\bar{E}/2E_d$ for $\bar{E} > 2E_d$, 1 for $E_d < \bar{E} < 2E_d$ and 0 for $\bar{E} < E_d$. In this case $\bar{E} \gg E_d$ and so the concentration required is $c = \sigma\phi tn\,\bar{E}/2E_d = \sigma\phi tn\,E_i/4E_d = 6.45 \cdot 10^{21}\,\mathrm{cm}^{-3}$. Note that this is an unrealistically large defect concentration as the atomic density, n, of the undamaged copper is only $8.5 \cdot 10^{22}\,\mathrm{cm}^{-3}$, indicating that the assumptions of no recombination, etc., are incorrect for such a long irradiation.

Problem 23 – Calculate the energies of a cluster of 1000 vacancies in copper when in the form of *a*) a spherical void, *b*) a Frank loop and *c*) a perfect loop on (111). [Assume surface energy = 1670 erg cm^{-2}, stacking

fault energy = 85 erg cm^{-2}, lattice constant 3.608 Å, shear modulus = $4 \cdot 10^{11}$ dyne cm^{-2} and Poisson ratio = $\frac{1}{3}$.]

Solution.

a) If N is the number of vacancies, r_v the radius of the spherical void, a the lattice constant and γ' the surface energy then $\frac{4}{3}\pi r_v^3 = Na^3/4$ ($r_v = 14.1$ Å). The total energy of the system is the surface area of the sphere ($4\pi r_v^2$) multiplied by γ', *i.e.*

$$E_v = 4\pi r_v^2 \gamma' . \tag{23.1}$$

Thus $E_v = (9N^2\pi/4)^{\frac{1}{3}} a^2 \gamma' = 4.17 \cdot 10^{-10}$ erg (= 0.26 eV per vacancy).

b) If γ is the stacking fault energy and $\boldsymbol{b}$ the total Burgers vector of the loop ($a/2$ [110]), μ the shear modulus, ν Poissons ratio and r_F is the radius of the loop then $\pi r_F^2 = Na^2 3/4$ ($r_F = 42.4$ Å), (loop is on (111)) and

$$E_F = \frac{\mu b^2 r_F}{3(1-\nu)} \left[\ln \left(\frac{r_F}{b} \right) + \frac{5}{3} \right] + \pi r_F^2 \gamma = \tag{23.2}$$

$$= 2.47 \cdot 10^{-10} + 4.8 \cdot 10^{-11} = 2.95 \cdot 10^{-10} \text{ erg } (= 0.18 \text{ eV per vacancy}).$$

c) For the perfect loop on (111) $r_p = r_F$ (= 42.4 Å) and

$$E_p = \frac{\mu b^2 r_p}{2(1-\nu)} \left[\ln \left(\frac{r_p}{b} \right) + \frac{5}{3} \right] = \tag{23.3}$$

$$= 3.7 \cdot 10^{-10} \text{ erg} \quad (= 0.23 \text{ eV per vacancy}) .$$

The formulae for E_F and E_p are, of course, considerable approximations (Kuhlmann-Wilsdorf and Wilsdorf([8])). The number of vacancies chosen for the problem was a typical value for a cluster, for which it was found that the Frank loop was the most stable configuration. However, this is not true for large loops as may be seen from Fig. 23 where the various energies are plotted against number of vacancies.

Problem 24 – If a copper sample bombarded by α-particles contains 10^{15} cm^{-3} equilibrium gas bubbles 200 Å in diameter at 0 °C, calculate *a*) the volume swelling and *b*) the volume of gas at NTP per cm^3 of sample. Calculate *c*) the gas bubble density and *d*) volume swelling if the gas is redistributed into bubbles 100 Å in diameter. [Assume helium gas is perfect; surface energy of copper = 1670 erg cm^{-2}.]

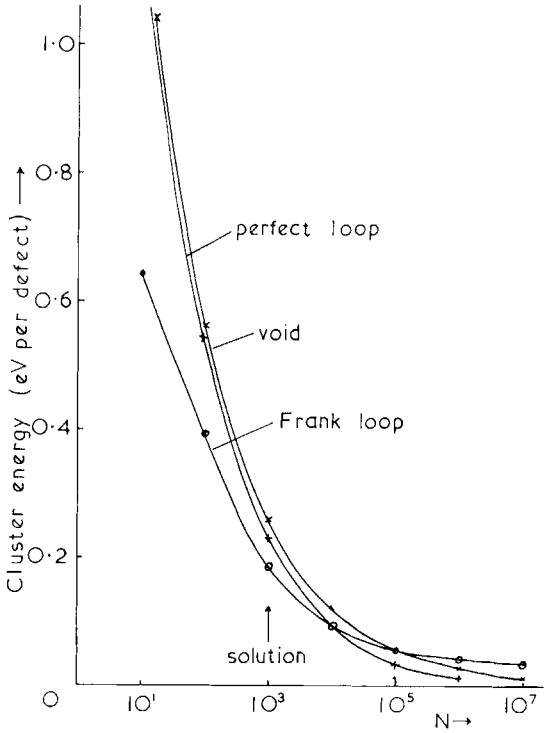


Fig. 23. – Cluster energies according to eqs (23.1), (23.2) and (23.3) as a function of number of vacancies, N, in the cluster.

Solution.

a) Volume swelling $= \Delta V/V = N4\pi r^3/3 = 4.2 \cdot 10^{-3}$.

b) Pressure p inside the gas bubble is related to the surface energy γ' and the bubble radius r by $p = 2\gamma'/r = 3.34 \cdot 10^9$ dyne cm^{-2}. Atmospheric pressure is $p_0 = 10^6$ dyne cm^{-2}. Hence volume of gas at *NTP* cm^{-3} is $(\Delta V/V)(p/p_0) = 14$ cm^3.

c) If $r_1 = 50$ Å then $p_1 = 2p$ and $(N_1 4\pi r_1^3/3) \cdot p_1 = (N 4\pi r^3/3) \cdot p$, *i.e.* $N_1 = Nr^3 p/r_1^3 p_1 = 4 \cdot 10^{15}$ cm^{-3}.

d) $\Delta V/V = N_1 4\pi r_1^3/3 = 2.1 \cdot 10^{-3}$.

Problem 25 – Calculate the growth factor G if $2 \cdot 10^{-5}$ % burn-up in α-uranium results in the formation of 10^{15} cm^{-3} interstitial loops 200 Å in diameter. [α-uranium is orthorhombic with $a = 2.85$ Å, $b = 5.87$ Å and the loops lie on (010).]

Solution. Interstitial loops lie on (010) and have Burgers vector $\frac{1}{2}[110]$. The increase in length Δl perpendicular to (010) is related to the growth factor for G and fractional burn-up B by $\Delta l/l = GB = Ab_{010}$, where A is loop area on (010) per unit area and b_{010} is effective Burgers vector of loop perpendicular to (010). Inspection of Fig. 25 shows that the effective Burgers vec-

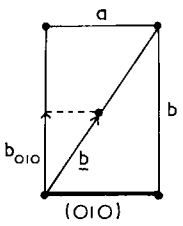


Fig. 25.

tor $\perp (010) = b/2 = 2.94$ Å. Loop area per cm$^3 = 10^{15} \cdot \pi \cdot 10^{-12}$ cm^2 cm^{-3}, *i.e.* there are $\pi \cdot 10^3$ extra planes cm^{-3}. Hence $\Delta l/l = \pi \cdot 10^3 \cdot 2.94 \cdot 10^{-8} = 9.23 \cdot 10^{-5}$ giving $G = 9.23 \cdot 10^{-5}/2 \cdot 10^{-7} = 462$.

Problem 26 – If there are 10^{15} cm^{-3} Frank loops 50 Å in diameter present in an irradiated crystal calculate the critical shear stress assuming that the strength of each loop intersected exceeds μb^2, *i.e.* the dislocations bow between the loops rather than cutting through them. [Assume $\mu = 4 \cdot 10^{11}$ dyne cm^{-2}, $a = 3.608$ Å.]

Solution. Just before the dislocation passes through the line of obstacles in its slip plane distance l apart it is semicircular in shape (see Fig. 26.1).

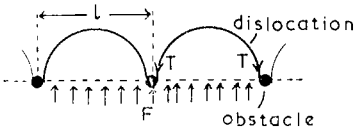


Fig. 26.1

If critical shear stress $= \tau_c$ then force acting on dislocation between two obstacles is $\tau_c bl$, which is just balanced by the dislocation line tension T $(= \frac{1}{2}\mu b^2)$ at each obstacle, *i.e.*

$$2T = \tau_c bl$$

(using the fact that $F = \tau b$, where F is the force per unit length on the dislocation in the direction of the Burgers vector $\boldsymbol{b}$).

The calculation of l proceeds as follows: the Frank loops are assumed to be randomly distributed on the four {111} planes and only those cutting through a number of slip planes, (111), need be considered, *i.e.* $\frac{3}{4}$ of the total. As {111} planes are inclined at 70° to each other the effective barrier density per unit area of slip plane becomes $\frac{3}{4} Nd \sin 70°$ (see Fig. 26.2), where N is the density of Frank loops of diameter d.

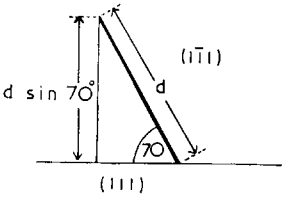


Fig. 26.2

Hence $l = (\frac{3}{4} Nd \sin 70)^{-\frac{1}{2}} = 5.35 \cdot 10^{-5}$ cm and

$$\tau_c = \frac{\frac{1}{2}\mu b^2}{bl} = 1.91 \cdot 10^8 \text{ dyne cm}^{-2},$$

using $b = a/2[110]$ and $T = \frac{1}{2}\mu b^2$.

REFERENCES

1) F. C. Phillips: *An Introduction to Crystallography*, 3rd ed., Longmans (1961).
2) R. W. James: *Optical Principles of the Diffraction of* X-*Rays*, Bell (1958).
3) P. B. Hirsch, A. Howie, R. B. Nicholson, D. W. Pashley and M. J. Whelan: *Electron Microscopy of Thin Crystals*, Butterworths (1965).
4) D. J. H. Cockayne, I. L. F. Ray and M. J. Whelan: *Phil. Mag.*, **20**, 1265 (1969).
5) D. J. H. Cockayne: *Ph. D. Thesis*. University of Oxford (1970).
6) P. B. Hirsch, A. Howie and M. J. Whelan: *Phil. Mag.*, A **252**, 499 (1960).
7) M. Wilkens: *Phys. Stat. Sol.*, **6**, 939 (1964).
8) D. Kuhlmann-Wilsdorf and H. G. F. Wilsdorf: *Journ. Appl. Phys.*, **31**, 516 (1960)

Transfer of Image Information in the Electron Microscope

F. A. LENZ

Institut für Angewandte Physik, Universität Tübingen - Tübingen, Germany

1. General theory.

The purpose of the electron microscope as that of any other optical or electron optical imaging device is to transmit information about properties of an object to an image. Therefore, we may consider it as an information channel and use some of the concepts and methods of information theory to describe the imaging properties of an electron microscope. In order to illustrate some of the basic concepts, let us start with the transfer of a signal which is a function of one variable only. An example is the transmission of an electrical signal along a telephone line. In this case, the signal may be a voltage or current, and the variable on which it depends is time. The input signal $S_0(t)$ which is entered into the transmission line on the input end gives rise to an output signal $S_1(t)$ at the output end of the line. If the transmission line is any good, the receiver at the output end should be able to conclude from the output signal $S_1(t)$ he is receiving on at least some of the information contained in the input signal $S_0(t)$. In the case of an imaging device the input and output signals depend on at least two variables x and y if an object surface is imaged to an image surface. x and y may stand for co-ordinates in these surfaces. If three-dimensional information on the object is to be transmitted, the input and output signals are functions of three variables. Using vector denotation, an input signal $S_0(\boldsymbol{r}_0)$ is fed into the transmission system at the object (input), and an output signal $S_1(\boldsymbol{r}_1)$ is received at the image (output). If the imaging device is any good, the receiver at the output end should be able to conclude from the output signal $S_1(\boldsymbol{r}_1)$ (the

image) on at least some of the information contained in the input signal $S_0(\boldsymbol{r}_0)$. If the transmission system is free of noise, the output signal S_1 will depend only on the input signal S_0 and on nothing else. Noise does not have to be audible: In the case of the electron microscope it means the source of any part of the output signal S_1 which is not due to the input signal S_0 but to such causes as mechanical vibrations of the microscope column, granularity of the photographic emulsion or fingerprints of a technical assistant on the micrograph.

Let us first neglect noise, not because there is not any but because it makes the theory simpler. Then there is a unique relation between input and output signal. In other words: Two or more different shots of the same object taken under exactly equal conditions should give two or more exactly identical micrographs. If there is some noise on the transmission line, and one has a reproducible input signal, one better records the output signal repeatedly in order to be able to distinguish which part of the output signal is real information and which part is due to noise.

We have seen that, neglecting noise, there is some unique relation between input and output signal. We call a system linear if this relation is linear. In other words, if the response of the system to one input signal S_0 is S_1 and the response to another input signal S_0' is S_1', then an input signal $\alpha S_0 + \beta S_0'$ would, in a linear system, produce an output signal $\alpha S_1 + \beta S_1'$ for arbitrary α and β. It is easier to treat linear than nonlinear systems. We shall therefore take care to define our input and output signals S_0 and S_1 so that they are related to each other by a linear relation at least to a good approximation. The transfer of electrical signals in electrical transmission lines can be made well enough linear. If, in an electron microscope, we define input and output signals as the amplitudes of an electron wave in the object and the image, they are also linearly related. This follows directly from the linearity of Schrödinger's or Dirac's wave equation. If we declare the mass thickness of the object as the input signal and the optical density of the developed photographic plate as the output signal, the linearity between input and output are no longer self-evident but at best a tolerable approximation. Most transfer theories are restricted to the case of linear transfer.

One function which can be used to describe the relation between input and output signals in a linear system is its impulsive response $G(t, t')$ or $G(\boldsymbol{r}_1, \boldsymbol{r}_0)$, respectively. It describes the response of the system to a short pulse $S_0(t) = \delta(t - t_0)$ in a one-dimensional transfer system or to an object consisting of one point only in an image transfer system, *i.e.* $S_0(\boldsymbol{r}_0) = \delta(\boldsymbol{r}_0 - \boldsymbol{r}_0')$. The delta function describing the short pulse or the object point, respectively,

is defined so that $\delta(t-t_0)$ equals zero for all times $t \neq t_0$ but is so large for $t = t_0$ that

$$\int \delta(t-t_0)\,\mathrm{d}t = 1 \tag{1.1}$$

if the interval of integration contains the time $t = t_0$. If the interval of integration does not contain $t = t_0$, the value of the integral equals zero. Correspondingly, the delta function in two-dimensional space is defined so that

$$\iint \delta(\boldsymbol{r}_0 - \boldsymbol{r}_0')\,\mathrm{d}\boldsymbol{r}_0 = \iint \delta(x_0 - x_0')\,\delta(y_0 - y_0')\,\mathrm{d}x\,\mathrm{d}y = 1 \tag{1.2}$$

if the two-dimensional interval of integration contains the point $\boldsymbol{r}_0'$ with the co-ordinates x_0', y_0'. Otherwise, the value of the integral equals zero. This definition of the delta function can be extended to more than two dimensions. The definition of the delta function implies that

$$\int_{-\infty}^{+\infty} A(t')\,\delta(t-t')\,\mathrm{d}t' = A(t)\,; \qquad \iint A(\boldsymbol{r}_0')\,\delta(\boldsymbol{r}_0 - \boldsymbol{r}_0')\,\mathrm{d}\boldsymbol{r}_0' = A(\boldsymbol{r}_0)\,. \tag{1.3}$$

In other words: Any arbitrary function $A(t)$ can be written as a linear superposition of delta functions $\delta(t'-t)$ with a weight function $A(t')$. Since we have assumed that $G(t, t')$ is the response of the linear system to the input signal $\delta(t-t')$, the output signal $S_1(t)$ of an arbitrary input signal

$$S_0(t) = \int_{-\infty}^{+\infty} S_0(t')\,\delta(t-t')\,\mathrm{d}t' \tag{1.4}$$

can be written as

$$S_1(t) = \int_{-\infty}^{+\infty} S_0(t')\,G(t, t')\,\mathrm{d}t'\,. \tag{1.5}$$

The transfer properties of a linear transfer system are therefore completely described by its impulsive response $G(t, t')$. Mathematicians and theoretical physicists refer to the impulsive response as « Green's function ». For

signals with more than one dimension, we have correspondingly

$$S_1(\boldsymbol{r}_1) = \iint S_0(\boldsymbol{r}_0)\, G(\boldsymbol{r}_1, \boldsymbol{r}_0)\, \mathrm{d}\boldsymbol{r}_0\,, \tag{1.6}$$

where $G(\boldsymbol{r}_1, \boldsymbol{r}_0')$ is the impulsive response of the linear imaging system to a delta function $\delta(\boldsymbol{r}_0 - \boldsymbol{r}_0')$. The integration (1.6) is extended over the object surface.

The transfer properties of a good electric transmission line should not depend on time. In other words: If the same message $S_0(t)$ is transmitted at two different times t_1 and t_2, say today and tomorrow, then the two input signals $S_0(t - t_1)$ and $S_0(t - t_2)$ should produce the same output signals $S_1(t - t_1)$ and $S_1(t - t_2)$, apart from a shift $t_2 - t_1$ in time. This independence of the transfer properties on time can be expressed by saying that the impulsive response is a function not of the two separate variables t and t' but only a function of one variable, *viz.* the difference $t - t'$:

$$G(t, t') = G(t - t')\,. \tag{1.7}$$

The response to a short pulse at time $t = t'$ will be the same as to a pulse at $t = t''$, only with a time delay of $t'' - t'$ between both. If the signal has more than one dimension such as in imaging devices, the corresponding property of the system would be that the image disk of an object point at position $\boldsymbol{r}_0 = \boldsymbol{r}_0'$ is the same as the image disk of an object point at $\boldsymbol{r}_0 = \boldsymbol{r}_0''$, only displaced to another position in the image. The shift in the image may be different from $\boldsymbol{r}_0'' - \boldsymbol{r}_0'$ because the image may be magnified with respect to the object. This desirable property of an imaging system that all object points at $\boldsymbol{r}_0 = \boldsymbol{r}_0'$ would produce an image disk of equal shape around the point $M\boldsymbol{r}_0'$ (M is the magnification) in the image plane or in the image space is called isoplanacy. It can be expressed by saying that the impulsive response is a function not of two separate vectors $\boldsymbol{r}_1$ and $\boldsymbol{r}_0$ but only of the difference $\boldsymbol{r}_1 - M\boldsymbol{r}_0$.

$$G(\boldsymbol{r}_1, \boldsymbol{r}_0) = G\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\right). \tag{1.8}$$

The condition of isoplanacy is not precisely satisfied in optical and electron optical imaging systems. If the system has aberrations depending on $\boldsymbol{r}_0$ such as distortion, third-order astigmatism, or coma, the isoplanacy condition (1.8) is violated, *i.e.* the image disk of an off-axis point looks different from that

of an axis point. Aberrations depending only on the initial *direction* of an electron trajectory such as spherical aberration, defocusing, axial astigmatism and axial coma do not affect isoplanacy. If the field of view is sufficiently small, the condition of isoplanacy can always be considered to be approximately satisfied.

In the isoplanatic approximation we can write eqs (1.5) and (1.6) as

$$S_1(t) = \int_{-\infty}^{+\infty} S_0(t') G(t-t') \mathrm{d}t' \tag{1.9}$$

and

$$S_1(r_1) = \iint S_0(r_0) G\left(\frac{r_1}{M} - r_0\right) \mathrm{d}r_0 . \tag{1.10}$$

Integrals of this type are called convolution integrals. To understand the physical meaning of the linear relation between the input signal S_0 and the output signal S_1 the following consideration may be useful.

The input signal which we have considered above as a linear superposition of delta functions, can, according to Fourier's theorem, also be considered as a linear superposition of sinusoidal functions:

$$S_0(t) = \int_{-\infty}^{+\infty} s_0(f) \exp[-2\pi i f t] \mathrm{d}f . \tag{1.11}$$

Because of the linearity of the transfer system, each Fourier component $s_0(f) \exp[-2\pi i f t]$ of the input signal corresponding to a frequency f can be transformed to the corresponding Fourier component of the output signal and then summed up (or rather integrated up). In other words: In the expression

$$S_1(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} s_0(f) \exp[-2\pi i f t'] \mathrm{d}f G(t-t') \mathrm{d}t' \tag{1.12}$$

we can *first* integrate over t' and *then* over f. The integration over t' is nothing else but a Fourier transform of G:

$$\int_{-\infty}^{+\infty} \exp[-2\pi i f t'] G(t-t') \mathrm{d}t' = \exp[-2\pi i f t] \int_{-\infty}^{+\infty} \exp[2\pi i f t'] G(t') \mathrm{d}t' . \tag{1.13}$$

Equation (1.12) and (1.13) can be interpreted as follows: The calculation of the output signal $S_1(t)$ from a given input signal $S_0(t)$ can be performed in the following steps: First, the Fourier transform $s_0(f)$ of $S_0(t)$ is formed. Then $s_0(f)$ is multiplied by the Fourier transform of the impulsive response G to obtain the Fourier transform of $S_1(t)$. The Fourier transform $T(f)$ of the impulsive response G is called the *transfer function* of the system:

$$T(f) = \int_{-\infty}^{+\infty} \exp[2\pi i f t]\, G(t)\, \mathrm{d}t\,. \tag{1.14}$$

According to eqs. (1.12) and (1.13), the product of s_0 with the transfer function $T(f)$ yields the output signal by another Fourier transform:

$$S_1(t) = \int_{-\infty}^{+\infty} s_0(f) T(f) \exp[-2\pi i f t]\, \mathrm{d}f\,. \tag{1.15}$$

Let us for a while assume that the input signal is a sine or cosine function:

$$S_0(t) = A \exp[-2\pi i f_0 t]\,. \tag{1.16}$$

A comparison with eq. (1.11) shows that this is equivalent with a Fourier transform $s_0(f)$ of S_0 (an « input spectrum »)

$$s_0(f) = A\delta(f - f_0)\,. \tag{1.17}$$

As we have seen, the output spectrum $s_1(f)$ is obtained by multiplying the input spectrum $s_0(f)$ by the transfer function

$$s_1(f) = AT(f)\,\delta(f - f_0)\,. \tag{1.18}$$

The output signal $S_1(t)$ is, according to eq. (1.15), the Fourier transform of the output spectrum $s_1(f)$:

$$S_1(t) = \int_{-\infty}^{+\infty} s_1(f) \exp[-2\pi i f t]\, \mathrm{d}f = AT(f_0) \exp[-2\pi i f_0 t]\,. \tag{1.19}$$

In other words: A sinusoidal input function with frequency f_0 and amplitude A is received at the output as a function of the same frequency but with an amplitude $AT(f_0)$. A transmission system for which $T(f_0)$ equals one for all frequencies f_0 would be an ideal system because the output signal would always be identical with the input signal. According to Fourier's theorem any arbitrary input function can be written as a linear superposition of sinusoidal functions with different frequencies f and amplitudes $s_0(f)$. Each of them is transmitted and forms a Fourier component $s_1(f) = T(f)s_0(f)$ at the output. They only have to be linearly superimposed to form the output signal $S_1(t)$.

If the signals are two-dimensional as in image transfer systems the same reasoning can be applied. We have, however, to use different variables $\boldsymbol{r}_0$ and $\boldsymbol{r}_1$ in the input and output signals, respectively, because the co-ordinates in the object and in the image plane do not have the same meaning.

Let us define the Fourier transforms of input and output signal and of the impulsive response G

$$S_0(\boldsymbol{r}_0) = \iint s_0(\boldsymbol{f}) \exp\left[-2\pi i \boldsymbol{f}\boldsymbol{r}_0\right] \mathrm{d}\boldsymbol{f}\,; \qquad s_0(\boldsymbol{f}) = \iint S_0(\boldsymbol{r}_0) \exp\left[2\pi i \boldsymbol{f}\boldsymbol{r}_0\right] \mathrm{d}\boldsymbol{r}_0\,; \tag{1.20}$$

$$S_1(\boldsymbol{r}_1) = \iint s_1(\boldsymbol{f}) \exp\left[-\frac{2\pi i}{M} \boldsymbol{f}\boldsymbol{r}_1\right] \mathrm{d}\boldsymbol{f}, \qquad s_1(\boldsymbol{f}) = \iint S_1(\boldsymbol{r}_1) \exp\left[\frac{2\pi i}{M} \boldsymbol{f}\boldsymbol{r}_1\right] \frac{\mathrm{d}\boldsymbol{r}_1}{M^2}, \tag{1.21}$$

$$T(\boldsymbol{f}) = \iint G(\boldsymbol{t}) \exp\left[2\pi i \boldsymbol{f}\boldsymbol{t}\right] \mathrm{d}\boldsymbol{t}\,. \tag{1.22}$$

In eq. (1.22), $\boldsymbol{t}$ stands as a substitution for

$$\boldsymbol{t} = \frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\,. \tag{1.23}$$

As in the case of one-dimensional signal functions it can again be shown that it follows from eq. (1.10) that

$$s_1(\boldsymbol{f}) = T(\boldsymbol{f}) s_0(\boldsymbol{f})\,. \tag{1.24}$$

The « space frequency » $\boldsymbol{f}$ is now a vector with two components f_x and f_y. The area elements $\mathrm{d}\boldsymbol{r}_0$ and $\mathrm{d}\boldsymbol{r}_1$ in eqs (1.20) and (1.21) stand for

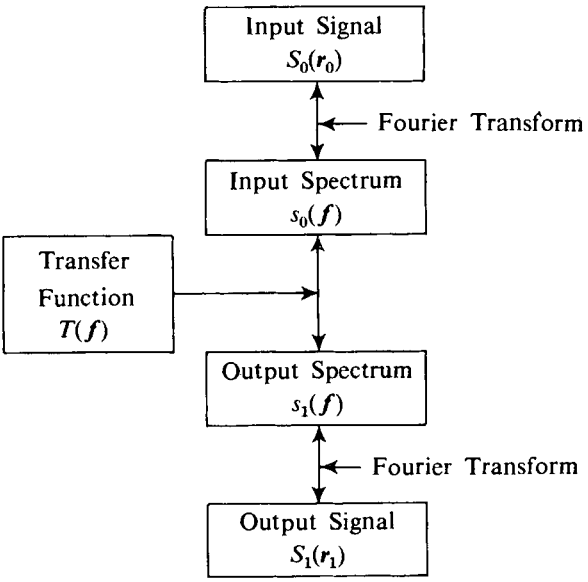
$$\mathrm{d}\boldsymbol{r}_0 = \mathrm{d}x_0 \mathrm{d}y_0\,, \qquad \mathrm{d}\boldsymbol{r}_1 = \mathrm{d}x_1 \mathrm{d}y_1\,, \tag{1.25}$$

such as the element $\mathrm{d}\boldsymbol{f}$ stands for

$$\mathrm{d}\boldsymbol{f} = \mathrm{d}f_x \, \mathrm{d}f_y \,. \tag{1.26}$$

The Fourier transforms (1.20) correspond to the expansion of the input (= object) signal in a series of sinusoidal components each of which is denoted by its space frequency $\boldsymbol{f}$ and its amplitude $s_0(\boldsymbol{f})$. The vector $\boldsymbol{f}$ with the components f_x and f_y denotes the direction of the sinusoidal component (plane wave) associated with each Fourier component. The vector $\boldsymbol{f}$ is perpendicular to the wave fronts of this plane wave, and its length $|\boldsymbol{f}|$ is the inverse of the repeat of the sinusoidal component.

Using the concept of the transfer function, the linear relation between the input and output signal can be described by the following diagram.



If the transfer function or the impulsive response of a system is known, the relation between S_0 and S_1 is uniquely defined, and one can conclude on S_1 if S_0 is known and *vice versa*. If, on the other hand, the relation between S_0 and S_1 were known empirically by taking a great number of micrographs of different objects with known properties, one would be able to determine the transfer function $T(\boldsymbol{f})$.

2. Amplitude transfer and contrast transfer function.

In the general theory we have derived relations between input and output signals in linear transmission systems but we have not specified what the physical nature of these signals is in electron microscopy. Since the image is transferred by means of electrons and since the propagation of these electrons in space can be described by a linear wave equation such as Schrödinger's, an obvious definition would be to identify the input signal with the wave amplitude in the object plane and the output signal with the wave amplitude in the image plane. The condition of linearity is exactly fulfilled in this case. The transfer function can be derived from a study of the propagation of the electron wave through the lenses and apertures of the electron optical imaging system.

The input signal depends on the conditions of illumination and on the interaction of the illuminating beam with the object. Let us first assume the illumination to be coherent in direction of the optical axis which we identify with the z axis of a Cartesian or cylindrical system of co-ordinates The wave amplitude of the incoming primary wave from the condenser, before it enters the object, would be a plane wave $\exp[2\pi ikz]$ with a wave number k depending on the acceleration voltage U

$$k = \frac{1}{\lambda} = \frac{1}{h}\sqrt{2em_0U\left(1+\frac{eU}{2m_0c^2}\right)} = \frac{1}{h}\sqrt{2em_0U^*}\,. \tag{2.1}$$

In eq. (2.1), h denotes Planck's constant. In high-resolution transmission microscopy the object can be considered as nonabsorbing. Practically all electrons entering the object from the condenser side leave it again on the image side because the probability for all interactions removing electrons from the beam, such as backscattering or bremsstrahlung production close to the short wavelength limit, is very small. The interaction between the primary electron beam and a thin object can therefore be understood as a local distortion of the electron wavefronts due to the local variations of the electrostatic potential within and between the atoms. Within an atom the potential is more positive than in the surrounding vacuum, and consequently the local wavelength is shorter than the vacuum wavelength. The resulting distortion of the wavefronts may be referred to as phase-shifting, diffraction or scattering, three different names for the same physical process.

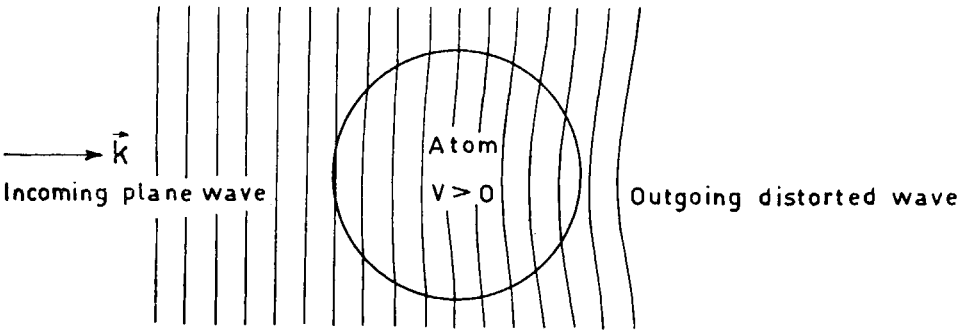


Fig. 1. – Interaction of atom and electron wave.

The amplitude of the distorted wave after passing through the object, which, without this interaction, would be a constant, is now

$$S_0(x_0, y_0) = \exp[i\eta(x_0, y_0)], \qquad S_0(\boldsymbol{r}_0) = \exp[i\eta(\boldsymbol{r}_0)] \tag{2.2}$$

a complex function of the co-ordinates x_0, y_0 in the object plane. S_0 is the input signal, and $\eta(\boldsymbol{r}_0)$ is the phase shift. The fact that the object is treated as nonabsorbing is expressed by the constance of object current density

$$j(\boldsymbol{r}_0) \sim |S_0(\boldsymbol{r}_0)|^2 \equiv 1 \tag{2.3}$$

immediately behind the object.

For weak phase objects, *i.e.* if $\eta(\boldsymbol{r}_0)$ is small compared with 2π for all $\boldsymbol{r}_0$, we have

$$\eta(\boldsymbol{r}_0) = 2\pi \int_0^t \left(\frac{1}{\lambda(x_0, y_0, z)} - \frac{1}{\lambda_0} \right) \mathrm{d}z =$$

$$= \frac{2\pi e m \lambda}{h^2} \int_0^t \varphi(x_0, y_0, z)\, \mathrm{d}z = \frac{2\pi e}{hv} \int_0^t \varphi(x_0, y_0, z)\, \mathrm{d}z \,. \tag{2.4}$$

The integration in eq. (2.4) is extended over the thickness t of the object. Equation (2.4) is relativistically correct if the relativistic expressions for m, λ and v are used.

All information about the object which the electron wave is carrying is contained in $S_0(\boldsymbol{r}_0)$ or $\eta(\boldsymbol{r}_0)$, respectively. In order to calculate the corresponding

output signal $S_1(\boldsymbol{r}_1)$, *i.e.* the wave amplitude in the image plane, we have to know the transfer function T or its Fourier transform, the impulsive response. The impulsive response G is the response of the imaging system to a point source in the object, *i.e.* the image wave amplitude in the diffraction disk which forms the image of an object point. The classical method of determining the wave amplitude in the diffraction disk is the application of Kirchhoff's integral formula. If the surface of integration in Kirchhoff's integral is the back focal plane of the objective lens, the evaluation of the integral is equivalent to the Fourier transform leading from the output spectrum s_1 to the output signal S_1. Kirchhoff's integration is extended only over the transparent part of the objective aperture. It has further to take into account the phase shift due to aberrations and defocusing. This phase shift is closely related to the wave aberration which is defined as the local distance between the real wave front and an ideal wave front, *i.e.* a sphere around the geometrical image point.

Each point in the back focal plane of the objective lens corresponds to one space frequency $\boldsymbol{f}$. If the object were a periodic structure whose object signal contained only one or a small number of space frequencies, then the wave function in the back focal plane would be zero except for steep local intensity maxima, one for each space frequency. In other words, the wave function in the back focal plane is the diffraction pattern of the object with a diffraction length equal to the focal length of the objective lens. Each space frequency $\boldsymbol{f}$ in the object (input signal) corresponds to one Bragg angle, *i.e.* one direction of a diffracted wave in object space. Each direction in object space corresponds to one point in the back focal plane. These two statements can be combined into one, saying that each space frequency $\boldsymbol{f}$ corresponds to one point $\boldsymbol{r}_B$ in the back focal plane

$$\boldsymbol{r}_B = l\lambda\boldsymbol{f}. \tag{2.5}$$

In eq. (2.5) l denotes the focal length of the objective lens (the letter f being reserved for space frequencies). Equation (2.5) should look familiar to people who have worked with electron diffraction of crystals where the position vector $\boldsymbol{r}$ of an intensity maximum in the diffraction diagram is equal to the product of the diffraction length l, the wavelength λ and the reciprocal lattice vector $\boldsymbol{f}$ denoting a space frequency (inverse of the spacing of lattice planes) in the periodic structure of the crystal. The effect of the aberrations is to shift the phase of the wave function in the back focal plane where the phase shift depends on $\boldsymbol{r}_B$ which, according to eq. (2.5) can be interpreted as a phase shift depending on space frequency.

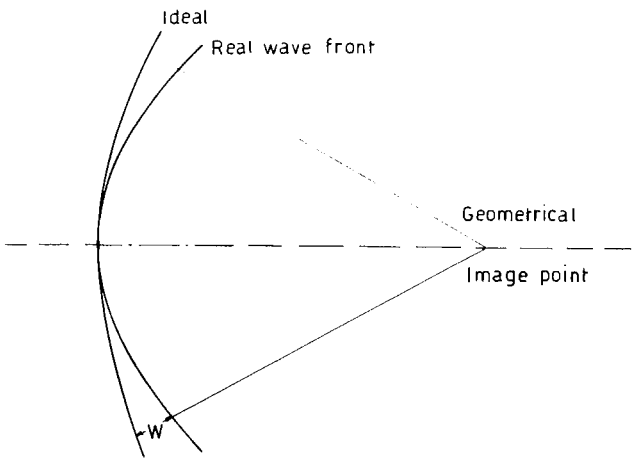


Fig. 2.

The first Fourier transformation transforming the input signal $S_0(\boldsymbol{r}_0)$ into the input spectrum $s_0(\boldsymbol{f})$ corresponds to the formation of the diffraction pattern of the object neglecting lens aberrations and apertures. These are taken into account by the transfer function

$$T(\boldsymbol{f}) = \frac{1}{M} \exp\left[-\frac{2\pi i}{\lambda} W(\boldsymbol{f})\right] B(\boldsymbol{f}), \tag{2.6}$$

which describes the phase shift $2\pi W/\lambda$, and the effect of an aperture by the aperture function $B(\boldsymbol{f})$. W is the wave aberration introduced by aberrations and defocusing. $B(\boldsymbol{f})$ is assumed to equal 1 in the transparent parts of the aperture, and to vanish for the opaque parts. If the lens suffers from spherical aberration, axial astigmatism and defocusing the wave aberration can be written as

$$\left.\begin{aligned} W(\boldsymbol{r}_B) &= \frac{C_s}{4l^4}(x_B^2 + y_B^2)^2 + \frac{\Delta z}{2l^2}(x_B^2 + y_B^2) - \frac{C_A}{2l^2}(x_B^2 - y_B^2), \\ W(\boldsymbol{f}) &= \frac{C_s}{4}\lambda^4 f^4 + \frac{\Delta z}{2}\lambda^2 f^2 - \frac{C_A}{2}(f_x^2 - f_y^2)\lambda^2. \end{aligned}\right\} \tag{2.7}$$

In eq. (2.7), C_s is the third-order spherical aberration coefficient. Its definition is the usual one, *i.e.* it implies that a geometrical electron trajectory leaving the axis point of the object plane under an angle α against the axis

intersects the image plane in a point at a distance $C_s|M|\alpha^3 + O(\alpha^5)$ from the axis. Δz stands for defocusing in object space. It is counted negative if the object is closer to the objective lens than the plane conjugated to the recording plane (screen or photographic plate). C_A is the coefficient of astigmatism. Its definition implies that the geometrical astigmatic lines, referred to object space have a distance of $2C_A$ from each other and a distance of C_A from the geometrical disk of least confusion. Figure 3 shows the dependence of wave aberration on spherical aberration and defocusing for zero astigmatism.

Knowing the wave aberration $W(f)$ and the aperture function $B(f)$ we can use the transfer function $T(f)$ to calculate the image wave amplitude $S_1(r_1)$ if we know the object wave amplitude $S_0(r_0)$, *i.e.* we can conclude from a given object on the corresponding image and *vice versa*. But unfortunately, wave amplitudes are not observable quantities. What we can observe in the image are such quantities as current density, contrast, optical density, etc., and they are not linearly related to any property of the object. It can, how-

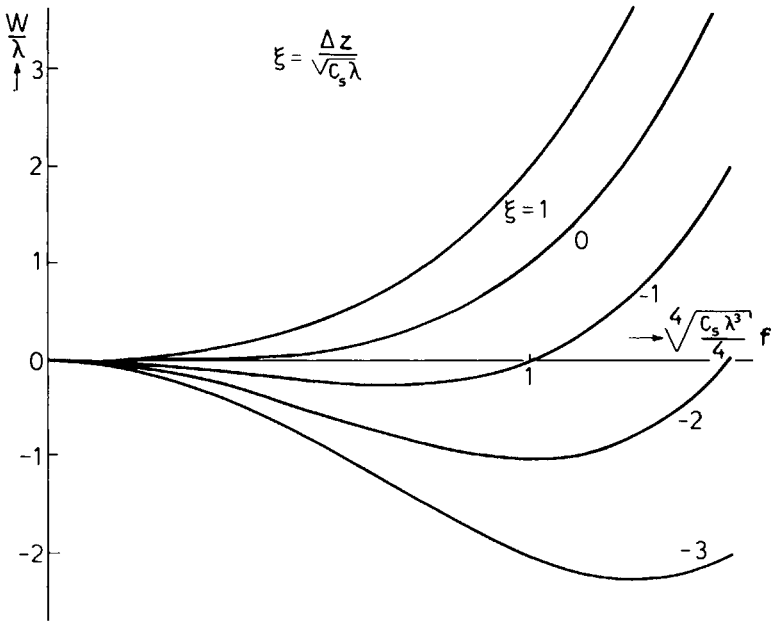


Fig. 3.

ever, be shown that the contrast in the image of a weak phase object is at least approximately a linear function of the phase shift η. To show this, let us treat the amplitude transfer of a weak phase object. For $\eta \ll 2\pi$, eq. (2.2)

can be written as

$$S_0(\boldsymbol{r}_0) = \exp[i\eta(\boldsymbol{r}_0)] = 1 + i\eta(\boldsymbol{r}_0) + O(\eta^2)\,. \tag{2.8}$$

The object spectrum follows by Fourier transformation, neglecting second and higher-order terms in η,

$$s_0(\boldsymbol{f}) = \int S_0(\boldsymbol{r}_0) \exp[2\pi i \boldsymbol{f}\boldsymbol{r}_0]\,\mathrm{d}\boldsymbol{r}_0 = \delta(\boldsymbol{f}) + i\int \eta(\boldsymbol{r}_0) \exp[2\pi i \boldsymbol{f}\boldsymbol{r}_0]\,\mathrm{d}\boldsymbol{r}_0\,. \tag{2.9}$$

$s_0(\boldsymbol{f})$ describes the angular distribution of the wave behind the object. The delta function stands for the undiffracted primary beam in axial direction. The second term on the right-hand side is the complex scattering amplitude of the object. If $\eta(\boldsymbol{r}_0)$ in eq. (2.9) is replaced by the expression in eq. (2.4) one obtains

$$s_0(\boldsymbol{f}) = \delta(\boldsymbol{f}) + \frac{ie}{\hbar v}\iiint \varphi(x_0, y_0, z) \exp[2\pi i \boldsymbol{f}\boldsymbol{r}_0]\,\mathrm{d}x_0\,\mathrm{d}y_0\,\mathrm{d}z\,. \tag{2.10}$$

We see that the second term on the right-hand side is a three-dimensional Fourier transform of the potential distribution within the scatterer. The integral on the right-hand side is known as the scattering amplitude of the scatterer. In the special case that the scatterer is an atom, it is called the atom form amplitude. Its absolute square is the differential scattering cross-section. Let us introduce an abbreviation $A(\boldsymbol{f})$ for this quantity:

$$A(\boldsymbol{f}) = \int \eta(\boldsymbol{r}_0) \exp[2\pi i \boldsymbol{f}\boldsymbol{r}_0]\,\mathrm{d}\boldsymbol{r}_0\,, \tag{2.11}$$

so that eq. (2.9) can be written as

$$s_0(\boldsymbol{f}) = \delta(\boldsymbol{f}) + iA(\boldsymbol{f})\,. \tag{2.12}$$

According to eq. (1.24), the image (output) spectrum $s_1(\boldsymbol{f})$ follows from the object (input) spectrum by multiplication with the amplitude transfer function

$$s_1(\boldsymbol{f}) = T(\boldsymbol{f}) s_0(\boldsymbol{f}) = T(0)\,\delta(\boldsymbol{f}) + iA(\boldsymbol{f})T(\boldsymbol{f})\,. \tag{2.13}$$

Performing the inverse Fourier transformation we obtain the output signal

(the image wave amplitude)

$$S_1(\boldsymbol{r}_1) = T(0) + i\int A(\boldsymbol{f})T(\boldsymbol{f}) \exp\left[-2\pi i \boldsymbol{f}\frac{\boldsymbol{r}_1}{M}\right] \mathrm{d}\boldsymbol{f}. \tag{2.14}$$

In bright field microscopy, $B(0) = 1$, and it follows from eq. (2.6) that $T(0) = 1/M$. The first term on the right-hand side of eq. (2.14) describes the bright background of the bright field image whose current density is M^{-2} times the primary current density in the object. The second term describes a small modulation of this background. It is small because we have assumed the phase shift η is small and because A is defined as the Fourier transform of this phase shift. In dark field microscopy, $B(0) = T(0) = 0$, and the background is dark. It is evident from eq. (2.14) that the contrast in a dark field image of a weak phase object exceeds that of the bright field image.

Let us define contrast C in the bright field image by

$$C(\boldsymbol{r}_1) = \frac{|S_1(\boldsymbol{r}_1)|^2 - 1/M^2}{1/M^2} = M^2|S_1(\boldsymbol{r}_1)|^2 - 1\,. \tag{2.15}$$

Replacing S_1 in eq. (2.15) from eq. (2.14) and neglecting second order terms we obtain

$$C(\boldsymbol{r}_1) = iM\int A(\boldsymbol{f})[(T(\boldsymbol{f}) - T^*(-\boldsymbol{f})] \exp\left[-2\pi i \boldsymbol{f}\frac{\boldsymbol{r}_1}{M}\right] \mathrm{d}\boldsymbol{f}. \tag{2.16}$$

If the aperture function $B(\boldsymbol{f}) = B(-\boldsymbol{f})$, *i.e.* if $B(\boldsymbol{f})$ has two-fold symmetry around the optical axis and if further $W(\boldsymbol{f}) = W(-\boldsymbol{f})$ then we have

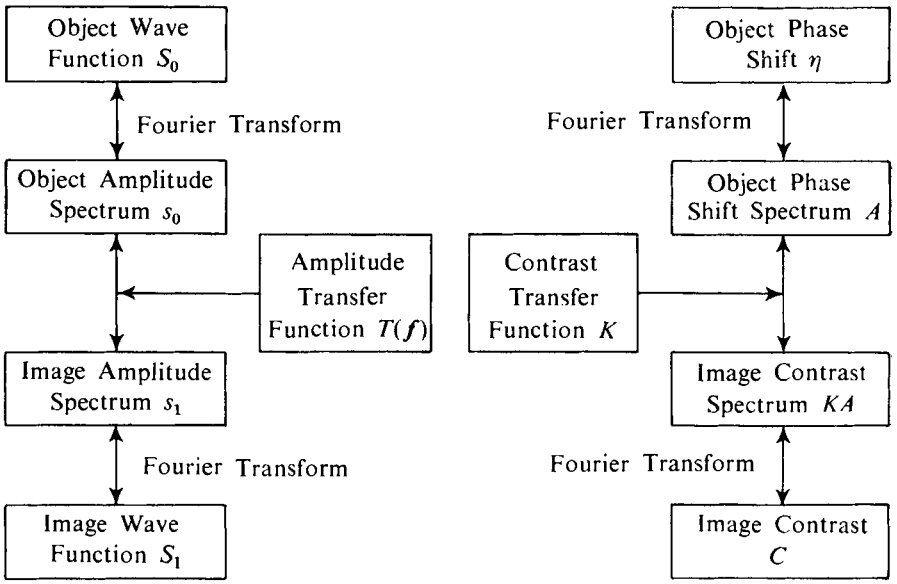
$$C(\boldsymbol{r}_1) = 2\int A(\boldsymbol{f})B(\boldsymbol{f}) \sin\left(\frac{2\pi}{\lambda} W(\boldsymbol{f})\right) \exp\left[-2\pi i \boldsymbol{f}\frac{\boldsymbol{r}_1}{M}\right] \mathrm{d}\boldsymbol{f}. \tag{2.17}$$

Equations (2.17) and (2.11) define a linear relation between the real contrast $C(\boldsymbol{r}_1)$ and the real phase shift $\eta(\boldsymbol{r}_0)$. If we define a contrast transfer function

$$K(\boldsymbol{f}) = 2B(\boldsymbol{f}) \sin\left(\frac{2\pi}{\lambda} W(\boldsymbol{f})\right), \tag{2.18}$$

this relation can be interpreted as follows: The input signal for contrast

transfer is now $\eta(\boldsymbol{r}_0)$. Its input spectrum $A(\boldsymbol{f})$ is multiplied by the contrast transfer function $K(\boldsymbol{f})$ to obtain the output spectrum. The inverse Fourier transform (2.17) then generates the output signal, *i.e.* the image contrast. This is explained in the following diagrams.



While phase contrast can be understood and explained only using wave optical aspects, another type of contrast has been discussed since the early days of electron microscopy, the so-called scattering absorption or amplitude contrast. It can be explained without using wave-optical concepts by saying that the atoms in the object scatter a fraction of the incoming electron current by scattering angles large enough to be intercepted by the objective aperture. The characteristic features of this type of contrast can also be explained in terms of the amplitude transfer theory. Let us suppose that the phase shift $\eta(\boldsymbol{r}_0)$ is so large that it makes sense to continue the expansion (2.8) by an additional second-order term:

$$S_0(\boldsymbol{r}_0) = \exp\left[i\eta(\boldsymbol{r}_0)\right] = 1 + i\eta(\boldsymbol{r}_0) - \frac{\eta^2(\boldsymbol{r}_0)}{2} + O(\eta^3)\,. \tag{2.19}$$

Let us, for the sake of simplicity, consider a sinusoidal variation of phase

shift $\eta(\boldsymbol{r}_0)$ with the space frequency $f_x = f_0$; $f_y = 0$:

$$\eta(\boldsymbol{r}_0) = \eta_0 \cos(2\pi f_0 x_0) . \tag{2.20}$$

Then the object wave function is

$$S_0(\boldsymbol{r}_0) = 1 + \frac{i}{2}\eta_0 [\exp[2\pi i f_0 x_0] + \exp[-2\pi i f_0 x_0]] - \\ - \frac{1}{8}\eta_0^2 [\exp[4\pi i f_0 x_0] + 2 + \exp[-4\pi i f_0 x_0]] + O(\eta_0^3) . \tag{2.21}$$

The corresponding object spectrum is

$$s_0(\boldsymbol{f}) = \delta(\boldsymbol{f}) + \frac{i}{2}\eta_0 \delta(f_y)[\delta(f_x + f_0) + \delta(f_x - f_0)] - \\ - \frac{\eta_0^2}{8}\delta(f_y)[\delta(f_x + 2f_0) + 2\delta(f_x) + \delta(f_x - 2f_0)] + O(\eta_0^3) . \tag{2.22}$$

Let us now assume that the space frequency is so high, and the objective aperture is so narrow that $T(f_0, 0)$ and $T(2f_0, 0)$ both vanish. In this case we have an image spectrum

$$s_1(\boldsymbol{f}) = T(\boldsymbol{f}) s_0(\boldsymbol{f}) = \frac{1}{M}\delta(\boldsymbol{f})\left(1 - \frac{1}{4}\eta_0^2\right) \tag{2.23}$$

and

$$S_1(\boldsymbol{r}_1) = \frac{1}{M}\left(1 - \frac{1}{4}\eta_0^2\right) . \tag{2.24}$$

The effect is a uniform reduced background intensity, and the space frequency f_0 is not resolved. An example is a thin foil of some amorphous material imaged under conditions at which the atoms or other local variations of potential are not resolved. Then regions containing many such atoms or potential variations appear darker in the image than regions containing less scatterers. This type of « area » contrast is compared with phase contrast in the following table.

	Space frequencies f_0 and $2f_0$ intercepted by aperture $T(f_0) - T(2f_0) = 0$	$2f_0$ intercepted, f_0 not intercepted $T(f_0) \neq 0$; $T(2f_0) = 0$	f_0 and $2f_0$ not intercepted $T(f_0) \neq 0$; $T(2f_0) \neq 0$
η small $\sin\eta \ll 1$ $\eta^2 \ll \eta$	No contrast	Phase contrast linear in η, f_0 resolved	Phase contrast linear in η, f_0 resolved
η larger	Amplitude contrast proportional with η^2, f_0 not resolved	Nonlinear phase contrast containing higher harmonics. Loss in background intensity, f_0 resolved	

Example: Image of a phase edge.

Let us assume that an object consists of two half-planes each of which is homogeneous, but because of a difference in thickness or in mean inner potential they produce different phase shifts:

$$S_0(\boldsymbol{r}_0) = \begin{cases} \exp[-i\varphi/2], & \text{for } x_0 \leqslant 0, \\ \exp[i\varphi/2], & \text{for } x_0 > 0. \end{cases} \tag{2.25}$$

In order to simplify the problem let us assume that

$$T(\boldsymbol{f}) = \begin{cases} 1, & \text{for } |\boldsymbol{f}| \leqslant f_0, \\ 0, & \text{for } |\boldsymbol{f}| > f_0. \end{cases} \tag{2.26}$$

This corresponds to a circular objective aperture within which the wave aberration is negligible.

The object spectrum is

$$s_0(\boldsymbol{f}) = \int S_0(\boldsymbol{r}_0) \exp[2\pi i \boldsymbol{f}\boldsymbol{r}_0]\, \mathrm{d}\boldsymbol{r}_0 = \delta(f_y)\left[\delta(f_x)\cos\frac{\varphi}{2} - \frac{1}{\pi f_x}\sin\frac{\varphi}{2}\right]. \tag{2.27}$$

Multiplication with the contrast transfer function (2.26) and Fourier transfor-

mation yields the image wave function

$$S_1(\boldsymbol{r}_1) = \cos\left(\frac{\varphi}{2}\right) + \frac{2i}{\pi}\sin\left(\frac{\varphi}{2}\right)\mathrm{Si}\left(\frac{2\pi f_0 x_1}{M}\right), \tag{2.28}$$

where

$$\mathrm{Si}(u) = \int_0^u \frac{\sin v}{v}\,\mathrm{d}v \tag{2.29}$$

is the « sine integral ». For the image contrast we obtain

$$C(\boldsymbol{r}_1) = \sin^2\left(\frac{\varphi}{2}\right)\left[\frac{4}{\pi^2}\mathrm{Si}^2\left(\frac{2\pi f_0 x_1}{M}\right) - 1\right]. \tag{2.30}$$

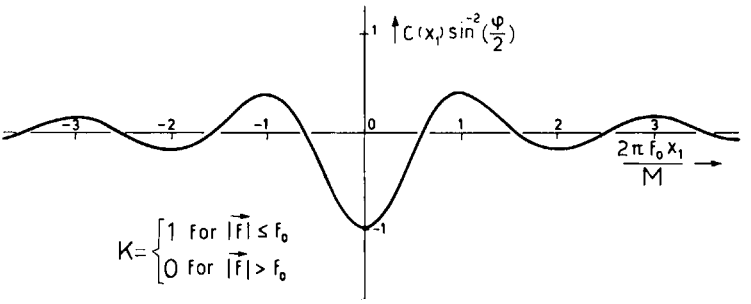


Fig. 4. – Contrast in the image of a phase edge.

3. Zonal plates and other interventions in the back focal plane of the objective.

Let us apply the contrast transfer theory to the case of a phase shifting point in the object, and let us ask the question how the aperture function $B(\boldsymbol{f})$ must be chosen if we want to achieve maximum bright field contrast in the image of this point. If the phase shifting interaction in the object is assumed to be localized in a point we have

$$\eta(\boldsymbol{r}_0) = \eta_0\,\delta(\boldsymbol{r}_0)\,. \tag{3.1}$$

The corresponding phase spectrum is

$$A(\boldsymbol{f}) = \int \eta(\boldsymbol{r}_0)\exp\left[2\pi i \boldsymbol{f}\boldsymbol{r}_0\right]\mathrm{d}\boldsymbol{r}_0 = \eta_0\,. \tag{3.2}$$

Multiplying by the transfer function we obtain the image contrast spectrum

$$A(f)K(f) = 2\eta_0 B(f) \sin\left(\frac{2\pi}{\lambda} W(f)\right). \tag{3.3}$$

By Fourier transform we obtain the image contrast

$$C(r_1) = 2\eta_0 \int B(f) \sin\left(\frac{2\pi}{\lambda} W(f)\right) \exp\left[-2\pi i f \frac{r_1}{M}\right] \mathrm{d}f. \tag{3.4}$$

The contrast $C(0)$ in the center of the image disk is

$$C(0) = 2\eta_0 \int B(f) \sin\left(\frac{2\pi}{\lambda} W(f)\right) \mathrm{d}f. \tag{3.5}$$

When the integration over the space frequencies f is performed, which is equivalent to an integration over the back focal plane where the objective aperture is arranged, there will be positive and negative contributions from different bands of space frequencies. Space frequencies for which the sine

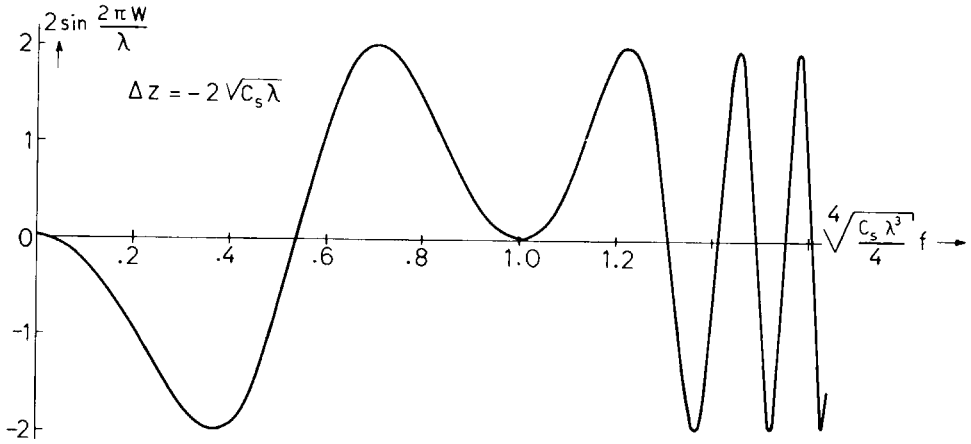


Fig. 5. – Contrast transfer function without aperture, $B(f) = 1$.

function in the integrand has a positive value, add to $C(0)$. For other space frequencies the sine function has a negative sign, and they will cancel at least part of the contrast. Hoppe's idea of using annular ring systems to

improve the electron microscopical image amounts to dimensioning the apertures so that either all negative contributions or all positive contributions to $C(0)$ are intercepted by the aperture stop. Figures 5 ÷ 7 show the con-

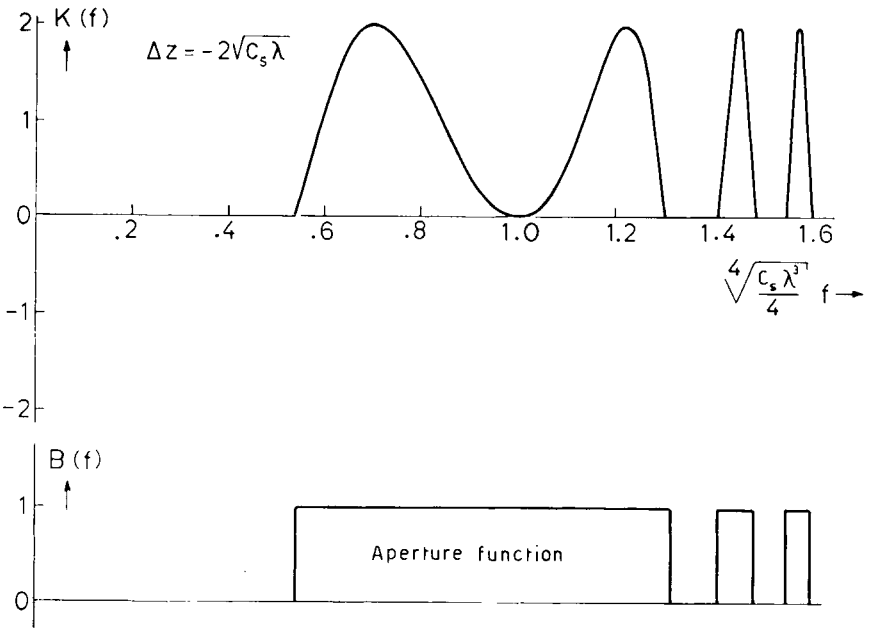


Fig. 6. – Contrast transfer function for maximum positive contrast, $B(0)=1$.

trast transfer functions K and the aperture functions B for the case that the wave aberration is given by eq. (2.7) with $C_A = 0$ and $\Delta z = -2\sqrt{C_s \lambda}$. Figure 5 shows the contrast transfer function K if no aperture is used ($B \equiv 1$), Fig. 6 for an aperture which leaves through all positive contributions to contrast, and Fig. 7 the same for negative contributions. The aperture system which helps to image an object point with maximum contrast is not necessarily ideal for all other types of objects. Apertures consisting of a system of concentric rings leave through some bands of space frequencies and intercept others. If an observer is interested in properties of an object which are mainly in some fixed space frequency region, then it would be unwise to intercept a frequency band in this region. For example, if an observer is interested in atomic distances of the order of 1 Å, his objective aperture should be transparent in the region of space frequencies around 1 Å^{-1} which corresponds to an aperture radius of $r_B = l\lambda$ Å^{-1} according to eq. (2.5).

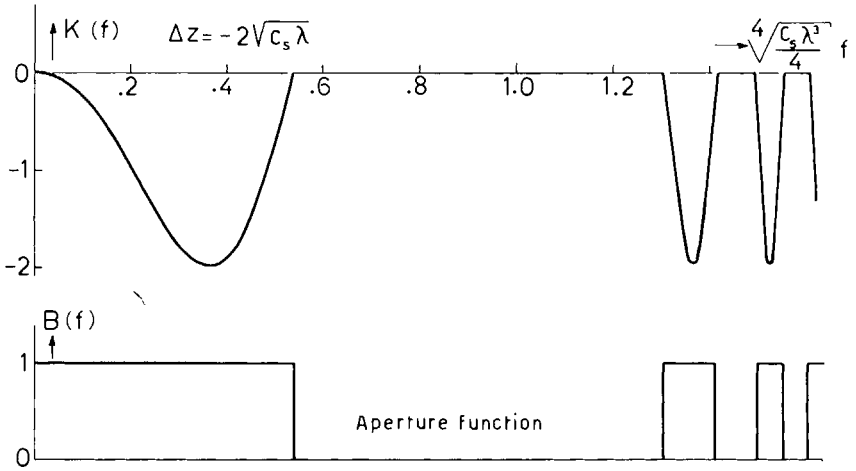


Fig. 7. – Contrast transfer function for maximum negative contrast, $B(0)=1$.

As we have seen, the Hoppe zonal plate is the aperture which optimizes the contrast of a point object in a bright field image. Let us now consider the effect of zonal plates on a dark field image. According to eq. (2.14) the wave amplitude S_1 in dark field is

$$S_1(\boldsymbol{r}_1) = i\int A(\boldsymbol{f})T(\boldsymbol{f}) \exp\left[-\frac{2\pi i}{M}\boldsymbol{f}\boldsymbol{r}_1\right] \mathrm{d}\boldsymbol{f}. \tag{3.6}$$

According to eq. (3.2) we have for a point object $A = \eta_0$. In the geometrical image $\boldsymbol{r}_1 = 0$ of this point we have

$$S_1(0) = i\eta_0 \int T(\boldsymbol{f})\,\mathrm{d}\boldsymbol{f} = \frac{i\eta_0}{M}\int \exp\left[-\frac{2\pi i}{\lambda} W(\boldsymbol{f})\right] B(\boldsymbol{f})\,\mathrm{d}\boldsymbol{f}. \tag{3.7}$$

As in eq. (3.5) we have again an integrand whose real and imaginary parts are changing their signs. If different space frequency intervals are not to cancel each other's contributions to the absolute value of $S_1(0)$, $B(\boldsymbol{f})$ must again be chosen so that only ring-shaped areas of the objective aperture are transparent for which

$$n + c < \frac{W}{\lambda} < n + c + \frac{1}{2}, \qquad n \text{ integer, } c \text{ arbitrary}. \tag{3.8}$$

For $c = 0$, this is the same condition as for maximum positive bright field contrast. For $c = \frac{1}{2}$ it coincides with the condition for maximum negative bright field contrast. Since in dark field microscopy the phase relation of the diffracted electrons with respect to the primary electrons does not matter, any other value of c in the condition for the ring radii would also be acceptable. The most important conclusion is, however, that a Hoppe zone plate designed for maximum contrast in the bright field image of a point object will also maximize the intensity in the center of the dark field image of the same point object.

Most other interventions in the back focal plane such as a filament across the center intercepting the primary beam or narrow circular apertures surrounded by a phase shifting ring may have the effect of increasing contrast in the image of an object but not necessarily in the space frequency region in which the observer is interested. In order to design an optimum aperture one must know the space frequency region of main interest. Then one can design an aperture which produces a maximum of the amplitude or contrast transfer function around this space frequency of main interest. Having done this, one may expect to find this space frequency in all image areas corresponding to object areas in which this space frequency occurs, even if it does so only as a second or higher harmonic of a lower space frequency.

4. The effects of illumination on image transfer.

We have so far restricted ourselves to coherent illumination in the direction of the optical axis. Even when dark field images were discussed, it was assumed that the primary beam had axial direction, and the aperture was symmetric with respect to the axis. In practical dark field microscopy, however, conditions are often different: The primary beam is inclined with respect to the axis so that it does not intersect the back focal plane of the

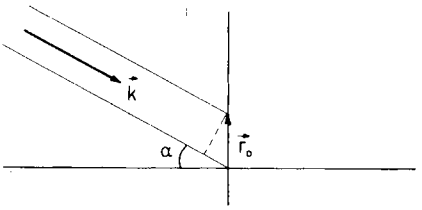


Fig. 8. – Oblique coherent illumination.

objective lens in its axis point but in another one. If the objective aperture is opaque in this point, the primary beam is intercepted, and a dark field image results. It is obvious that oblique illumination introduces a preferential direction in the image. In this case the transfer function is not a function of only the absolute value of the space frequency f but it also depends on its direction.

The oblique coherent illumination produces an additional phase shift $2\pi \boldsymbol{k}\boldsymbol{r}_0$ in excess of the one given in eq. (2.4) describing the interaction of primary beam and object. If we assume that the object is so thin that its thickness times the angle between $\boldsymbol{k}$ and the axis is smaller than the smallest details we want to observe we may treat the object as infinitely thin, and we have instead of eq. (2.4)

$$\eta(\boldsymbol{k}, \boldsymbol{r}_0) = 2\pi \boldsymbol{k}\boldsymbol{r}_0 + \eta(0, \boldsymbol{r}_0)\,. \tag{4.1}$$

In eq. (4.1) and in the following we may treat $\boldsymbol{k}$ as a vector with two components k_x and k_y only, because $\boldsymbol{r}_0$ lies in the object plane which we have assumed to be perpendicular to the axis. This is because $\eta(\boldsymbol{k}, \boldsymbol{r}_0)$ contains all information about the object which enters the transfer system, and because $\eta(\boldsymbol{k}, \boldsymbol{r}_0)$ does not depend on k_z. We can now apply the transfer theory to determine the image contrast, replacing $\eta(\boldsymbol{r}_0)$ by $\eta(\boldsymbol{k}, \boldsymbol{r}_0)$. Equation (2.8) reads now, in the case of oblique coherent illumination

$$S_0(\boldsymbol{k}, \boldsymbol{r}_0) = \exp\left[i\eta(\boldsymbol{k}, \boldsymbol{r}_0)\right] = (1 + i\eta(0, \boldsymbol{r}_0)) \exp\left[2\pi i \boldsymbol{k}\boldsymbol{r}_0\right]. \tag{4.2}$$

The object spectrum becomes

$$s_0(\boldsymbol{k}, \boldsymbol{f}) = \int S_0(\boldsymbol{k}, \boldsymbol{r}_0) \exp\left[2\pi i \boldsymbol{f}\boldsymbol{r}_0\right] \mathrm{d}\boldsymbol{r}_0 = \\ = \delta(\boldsymbol{k} + \boldsymbol{f}) + i \int \eta(0, \boldsymbol{r}_0) \exp\left[2\pi i(\boldsymbol{k} + \boldsymbol{f})\boldsymbol{r}_0\right] \mathrm{d}\boldsymbol{r}_0\,. \tag{4.3}$$

Using the abbreviation (2.11) we have now

$$s_0(\boldsymbol{k}, \boldsymbol{f}) = \delta(\boldsymbol{k} + \boldsymbol{f}) + iA(\boldsymbol{k} + \boldsymbol{f})\,. \tag{4.4}$$

The physical meaning of this equation is that each point in the diffraction pattern of the object in the back focal plane has been shifted from $\boldsymbol{r}_B = l\lambda \boldsymbol{k}$ to $\boldsymbol{r}_B = l\lambda(\boldsymbol{k} + \boldsymbol{f})$. The primary beam ($\boldsymbol{f} = 0$) no longer corresponds to the axis

point in the back focal plane but to $l\lambda\boldsymbol{k}$. The axis point ($\boldsymbol{r}_B = 0$) corresponds now to the space frequency $\boldsymbol{f} = -\boldsymbol{k}$. If we now multiply the object spectrum $s_0(\boldsymbol{k}, \boldsymbol{f})$ by the amplitude transfer function $T(\boldsymbol{f})$ we find the image spectrum

$$s_1(\boldsymbol{k}, \boldsymbol{f}) = T(-\boldsymbol{k})\,\delta(\boldsymbol{k} + \boldsymbol{f}) + iA(\boldsymbol{k} + \boldsymbol{f})T(\boldsymbol{f})\,. \tag{4.5}$$

Performing the inverse Fourier transform we obtain the image wave amplitude

$$S_1(\boldsymbol{k}, \boldsymbol{r}_1) = T(-\boldsymbol{k}) \exp\left[\frac{2\pi i}{M}\,\boldsymbol{k}\boldsymbol{r}_1\right] + i\int A(\boldsymbol{k} + \boldsymbol{f})T(\boldsymbol{f}) \exp\left[-2\pi i \boldsymbol{f}\frac{\boldsymbol{r}_1}{M}\right] \mathrm{d}\boldsymbol{f}. \tag{4.6}$$

If $B(-\boldsymbol{k}) = 0$, *i.e.* if the primary beam is intercepted by the aperture stop, we have dark field imaging with a preferential direction, and the image wave amplitude becomes

$$S_1(\boldsymbol{k}, \boldsymbol{r}_1) = i\int A(\boldsymbol{k} + \boldsymbol{f})T(\boldsymbol{f}) \exp\left[-2\pi i \boldsymbol{f}\,\frac{\boldsymbol{r}_1}{M}\right] \mathrm{d}\boldsymbol{f}. \tag{4.7}$$

If, on the other hand, $B(-\boldsymbol{k}) = 1$ we have a bright field image with a wave amplitude

$$S_1(\boldsymbol{k}, \boldsymbol{r}_1) = \frac{1}{M} \exp\left[-\frac{2\pi i}{\lambda}\,W(-\boldsymbol{k})\right] \exp\left[\frac{2\pi i}{M}\,\boldsymbol{k}\boldsymbol{r}_1\right] + \\ + i\int A(\boldsymbol{k} + \boldsymbol{f})T(\boldsymbol{f}) \exp\left[-2\pi i \boldsymbol{f}\frac{\boldsymbol{r}_1}{M}\right] \mathrm{d}\boldsymbol{f}. \tag{4.8}$$

It is not self-evident that coherent illumination always yields the best images. It can be shown that even for an arbitrary incoherent illumination a contrast transfer function can be defined as long as weak phase objects are imaged and the isoplanatic approximation holds. The illumination is called incoherent if the condenser aperture α is large so that the beam can no longer be called parallel. If the variation in wave vector $\boldsymbol{k}$ within the primary beam is so large that the phase differences $2\pi\boldsymbol{k}\boldsymbol{r}_0$ (compare eq. (4.1)) vary by an amount comparable to or larger than 2π, then the phase relations between two points in a distance $|\boldsymbol{r}_0|$ from each other are destroyed, and two such points are « incoherently illuminated ». If, on the other hand, the variation of $\boldsymbol{k}$ is so small that the phase differences $2\pi\boldsymbol{k}\boldsymbol{r}_0$ vary by less than $\pm\pi/2$, then two points at a distance $|\boldsymbol{r}_0|$ from each other are « coherently illuminated ».

Whether some illumination is coherent or not, depends not only on the condenser aperture but also on the size $|\boldsymbol{r}_0|$ of the object details one wants to observe.

According to eq. (4.2), the object wave function for an incoming electron with wave vector $\boldsymbol{k}$ can be written as

$$S_0(\boldsymbol{k}, \boldsymbol{r}_0) = S_0(0, \boldsymbol{r}_0) \exp [2\pi i \boldsymbol{k} \boldsymbol{r}_0] . \tag{4.9}$$

According to eq. (1.10), the corresponding image wave function can be written as

$$S_1(\boldsymbol{k}, \boldsymbol{r}_1) = \int S_0(0, \boldsymbol{r}_0) G\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\right) \exp [2\pi i \boldsymbol{k} \boldsymbol{r}_0] \, \mathrm{d}\boldsymbol{r}_0 . \tag{4.10}$$

The image current density is, apart from an irrelevant constant factor

$$|S_1(\boldsymbol{k}, \boldsymbol{r}_1)|^2 = \iint S_0(0, \boldsymbol{r}_0) S_0^{*} (0, \boldsymbol{r}_0') G\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\right) G^*\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0'\right) \cdot \\ \cdot \exp [2\pi i \boldsymbol{k} (\boldsymbol{r}_0 - \boldsymbol{r}_0')] \, \mathrm{d}\boldsymbol{r}_0 \mathrm{d}\boldsymbol{r}_0' . \tag{4.11}$$

For incoherent illumination, all the current densities corresponding to different $\boldsymbol{k}$ vectors occurring in the primary beam are superimposed upon each other incoherently. The image current density becomes

$$j(\boldsymbol{r}_1) = \int |S_1(\boldsymbol{k}, \boldsymbol{r}_1)|^2 F(\boldsymbol{k}) \, \mathrm{d}\boldsymbol{k} = \iiint S_0(0, \boldsymbol{r}_0) S_0^*(0, \boldsymbol{r}_0') \cdot \\ \cdot G\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\right) G^*\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0'\right) F(\boldsymbol{k}) \exp [2\pi i \boldsymbol{k} (\boldsymbol{r}_0 - \boldsymbol{r}_0')] \, \mathrm{d}\boldsymbol{r}_0 \, \mathrm{d}\boldsymbol{r}_0' \, \mathrm{d}\boldsymbol{k} . \tag{4.12}$$

$F(\boldsymbol{k})$ is a distribution function describing the angular distribution of the primary beam from the condenser. It is defined so that $F(\boldsymbol{k}) \, \mathrm{d}\boldsymbol{k} = F(k_x, k_y) \, \mathrm{d}k_x \, \mathrm{d}k_y$ is the probability that an incident electron has a direction such that the x and y components of its wave vector lie within the intervals $\{k_x, k_x + \mathrm{d}k_x\}$ and $\{k_y, k_y + \mathrm{d}k_y\}$. This distribution function is assumed to be normalized so that

$$\int F(\boldsymbol{k}) \, \mathrm{d}\boldsymbol{k} = 1 . \tag{4.13}$$

The integration over $\boldsymbol{k}$ in eq. (4.12) can be performed if we introduce the

Fourier transform of the distribution function $F(\boldsymbol{k})$:

$$\Phi(\boldsymbol{r}_0) = \int F(\boldsymbol{k}) \exp\left[2\pi i \boldsymbol{k}\boldsymbol{r}_0\right] \mathrm{d}\boldsymbol{k} \,. \tag{4.14}$$

Then we have

$$j(\boldsymbol{r}_1) = \iint S_0(0, \boldsymbol{r}_0) S_0^*(0, \boldsymbol{r}_0') G\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\right) G^*\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0'\right) \Phi(\boldsymbol{r}_0 - \boldsymbol{r}_0') \, \mathrm{d}\boldsymbol{r}_0 \, \mathrm{d}\boldsymbol{r}_0' \,. \tag{4.15}$$

Let us now assume that we have a weak phase object, *i.e.* that $S_0(0, \boldsymbol{r}_0)$ can be expressed by eq. (2.8)

$$S_0(0, \boldsymbol{r}_0) = 1 + i\eta(\boldsymbol{r}_0) \,. \tag{4.16}$$

Replacing S_0 from (4.16) in (4.15) and neglecting second-order terms in η we obtain for the image current density

$$\begin{aligned} j(\boldsymbol{r}_1) = j_B + i \iint \eta(\boldsymbol{r}_0) \, \Phi(\boldsymbol{r}_0 - \boldsymbol{r}_0') G\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\right) G^*\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0'\right) \mathrm{d}\boldsymbol{r}_0 \, \mathrm{d}\boldsymbol{r}_0' - \\ - i \iint \eta(\boldsymbol{r}_0') \, \Phi(\boldsymbol{r}_0 - \boldsymbol{r}_0') G\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\right) G^*\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0'\right) \mathrm{d}\boldsymbol{r}_0 \, \mathrm{d}\boldsymbol{r}_0' \,. \end{aligned} \tag{4.17}$$

In eq. (4.17), j_B is an abbreviation for the background current density

$$j_B = \iint \Phi(\boldsymbol{r}_0 - \boldsymbol{r}_0') G\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\right) G^*\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0'\right) \mathrm{d}\boldsymbol{r}_0 \, \mathrm{d}\boldsymbol{r}_0' \,. \tag{4.18}$$

If we again define contrast $C(\boldsymbol{r}_1)$ by

$$C(\boldsymbol{r}_1) = \frac{j(\boldsymbol{r}_1) - j_B}{j_B} \,, \tag{4.19}$$

we have

$$C(\boldsymbol{r}_1) = \int \eta(\boldsymbol{r}_0) \Gamma(\boldsymbol{r}_1, \boldsymbol{r}_0) \, \mathrm{d}\boldsymbol{r}_0 \,, \tag{4.20}$$

where the impulsive response Γ is given by

$$\begin{aligned} \Gamma(\boldsymbol{r}_1, \boldsymbol{r}_0) = \frac{i}{j_B} \int \Big[\Phi(\boldsymbol{r}_0 - \boldsymbol{r}_0') G\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\right) G^*\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0'\right) - \Phi(\boldsymbol{r}_0' - \boldsymbol{r}_0) \cdot \\ \cdot G\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0'\right) G^*\left(\frac{\boldsymbol{r}_1}{M} - \boldsymbol{r}_0\right) \Big] \mathrm{d}\boldsymbol{r}_0' \,. \end{aligned} \tag{4.21}$$

Substituting a new variable of integration

$$t = \frac{r_1}{M} - r_0', \tag{4.22}$$

it can be shown that the impulsive response Γ is a function not of r_1 and r_0 separately but only of the combination $r_1/M - r_0$. In other words: The isoplanacy condition is not destroyed by incoherent illumination. (4.20) can now be written as

$$C(r_1) = \int \eta(r_0)\, \Gamma\left(\frac{r_1}{M} - r_0\right) \mathrm{d}r_0 . \tag{4.23}$$

This is again a convolution integral so that we can define a transfer function for the Fourier transform of Γ. If the impulsive response $G(r_1/M - r_0)$ for coherent amplitude transfer is known, Γ can be calculated from eq. (4.21) for any arbitrary angular distribution $F(k)$ of the illuminating beam. The Fourier transform of Γ takes into account not only the electron optical properties of the imaging system behind the object but also the conditions of illumination. C, η and Γ are real functions. It should, however, be noted that, in the case of partial coherence, the linear terms in η may not be large compared to the second order terms.

REFERENCES

General Theory:

O. SCHERZER: *Journ. Appl. Phys.*, **20**, 20 (1949).
K.-J. HANSZEN, B. MORGENSTERN, *Zeits. angew. Phys.*, **19**, 215 (1965).
F. LENZ, *Laboratory Investigation*, **14**, 808 (1965); *Optik*, **22**, 270 (1965).

Zone Correction Plates:

W. HOPPE: *Naturwiss.*, **48**, 736 (1961).
W. HOPPE: *Optik*, **20**, 599 (1963).
G. MÖLLENSTEDT, R. SPEIDEL, W. HOPPE, R. LANGER, K. H. KATERBAU and F. THON: *Proc. 4th Eur. Conf. Electr. Micr. Rome 1968* (Rome, 1968), vol. **1**, p. 125.
F. LENZ, *Optik* **21**, 489 (1964).

Light Optical Theory of Phase Contrast:

F. ZERNIKE: *Phys. Zeits.*, **36**, 848 (1935).

Optical Space Frequency Analysis by Fraunhofer Diffraction of Electron Micrographs:

A. KLUG and J. E. BERGER: *Journ. Mol. Biol.*, **10**, 565 (1964).
F. THON: *Zeits. Naturfor.*, **21***a*, 476 (1966).

Image Reconstruction in Light Optical Microscopy:

A. MARECHAL and P. CROCE: *Compt. Rend. Acad. Sci. Paris*, **237**, 607 (1953).

Image Reconstruction in Electron Microscopy:

K.-J. HANSZEN: *Proc. 4th Eur. Conf. Electr. Micr., Rome 1968* (Rome, 1968), vol. **1**, p. 153,
P. SCHISKE: *Proc. 4th Eur. Conf. Electr. Micr., Rome 1968* (Rome, 1968), vol. **1**, p. 145.
D. J. DE ROSIER and A. KLUG: *Nature*, **217**, 130 (1968).
R. LANGER and W. HOPPE: *Optik*, **24**, 470 (1966-67); **25**, 413, 507 (1967).

Phase Contrast Electron Microscopy

F. THON

Siemens AG - Berlin, Germany

This material has been contributed from the standpoint of an electron microscopist, who experiments in the high resolution field. It deals basically with the information which can be gained from a high resolution phase contrast image.

A discussion of the mechanism for obtaining phase contrast with conventional objective lenses in Sect. **1** leads to the demand for better transfer-conditions.

In Sect. **2** we investigate, how contrast transfer is influenced by differently shaped apertures in the back focal plane of the objective lens. Also these interventions cannot establish optimum transfer conditions.

The prospects for realizing phase contrast transfer, which could be called ideal from the theoretical standpoint, will be discussed in Sect. **3**. Some principal experiments into this direction will be described.

Finally, in Sect. **4** we shall introduce some methods which allow one to improve the quality of electron microscopic images by subsequent light-optical reconstruction. The general aim will be: improvement of the information transfer conditions of the electron microscope in order to achieve interpretable images. Except in Sect. **1** we report recent investigations within the scope of the course.

1. Conventional phase contrast imaging.

1'1. Introduction.

By conventional phase contrast imaging we understand imaging by means of a high-performance electron microscope, using high magnifications, where

the objective aperture is sufficiently large and circular. Thus, our subject is phase contrast in the high resolution field and our aim is to outline how this phase contrast is dependent on the wave aberration of the objective lens.

There is no doubt that phase contrast is also present and useful in defocused images of medium or even low resolution. But in these cases things are much less complicated, since phase contrast then is just a means to enhance details, which are anyhow visible due to strong scattering absorption contrast. It is mainly in the range below 10 Å where phase contrast becomes the dominant contrast mechanism.

The imaging properties in the high resolution field with special emphasis on phase contrast have been thoroughly investigated in recent years theoretically and experimentally by several authors ([1-18]). A summary of the results will be given in this Section as far as they are of importance for present practical work. The chosen way of treatment seems to be the most appropriate one to describe conventional phase contrast imaging and also the special techniques, which will be discussed in the following Sections. A more complete theoretical treatment of the basic problems can be found in the contribution of Lenz in this book.

It should be mentioned that most of the basic problems and methods to be discussed in the following Sections are, in a modified way, also valid for scanning microscopy.

1·2. Theory.

We have to deal with the mechanism by which information about object properties is transferred to the electron microscopic image. Let us assume our imaging system to be linear and the illumination of the object to be coherent. If the object is a weak phase object, the image contrast is at least approximately a linear function of the phase shift introduced by the specimen. When dealing with linear systems, it is useful to decompose a complicated input, *i.e.* our object properties, into a number of more simple inputs, to calculate the response of the system to each of these « elementary » functions, and to superimpose the individual responses to find the total response. In other words: we can assume our object $g(x)$ (one-dimensional treatment) to be composed of a large number of sinusoidally varying transmission gratings, each with a different period length Λ or spatial frequency $1/\Lambda \equiv F$. The spatial frequencies of these gratings vary from zero up to some maximum value F_{max}.

To make use of the principle of Fourier decomposition, we can formally write

$$g(x) = \int_{-\infty}^{+\infty} G(F) \exp [-2\pi i F x] \mathrm{d}F , \tag{1}$$

so that

$$G(F_0) \exp [-2\pi i \cdot F_0 x] \tag{2}$$

describes a single grating of spatial frequency F_0 and amplitude $G(F_0)$.

In a next step we look at one of these elementary gratings with a period length or reciprocal spatial frequency Λ (Fig. 1). When a plane electron

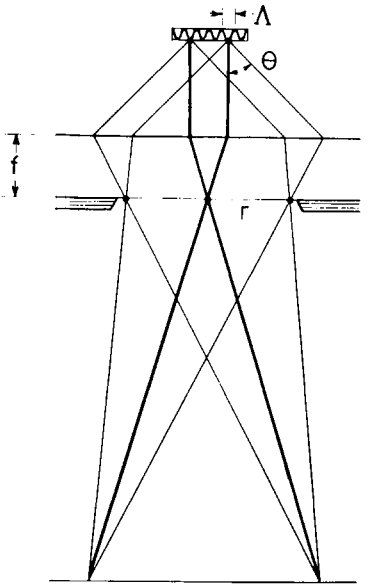


Fig. 1. – Imaging of a periodic specimen (schematically).

wave is incident on the object plane, this grating gives rise to a diffracted wave at an angle θ to the optical axis according to the fundamental grating equation, which can be written in a simplified form

$$\theta = \frac{\lambda}{\Lambda} \tag{3}$$

in the case of small angles used in electron microscopy. λ denotes the electron wavelength.

Introducing the focal length f of the objective lens and a radius r in the exit pupil of this lens, where all waves diffracted at the same angle are focused, we may write:

$$\Lambda = \frac{\lambda \cdot f}{r}. \tag{4}$$

This means that each point in the diffraction plane of the objective lens, this plane in a first approximation being identical with the back focal plane of the lens, corresponds to one specific diffraction angle θ and, consequently, to one certain reciprocal spatial frequency Λ or frequency F.

Thus, in the case of periodic objects, only very limited regions of the objective lens take part in the imaging process. Most of the advantages present in lattice plane imaging are due to this fact.

In the case of an amorphous object, we have a number of elementary gratings with different period length Λ_i and we have a lot of partial waves diffracted at different angles, thus the wave vectors hit the back focal plane of the lens at different points given by:

$$r_i = \frac{\lambda \cdot f}{\Lambda_i}. \tag{5}$$

This means, each spatial frequency contained in the object function is transformed to one specific point within the lens aperture, as it is for the one-dimensional case schematically drawn in Fig. 2. The superposition of all the diffracted waves in the image plane finally yields an image intensity distribution, taking account of amplitudes and phases. This description is valid in the case of coherent illumination. It is one key point for understanding the interventions in the back focal plane, which will be discussed later in Sect. **2** and **3**.

There is one complication in electron microscopy due to the fact that at least high resolution objects have to be considered as nonabsorbing. The interaction between the primary electron beam and the specimen leads to phase shiftings of the electron waves. Thus, the objects behave like phase objects in light microscopy, and the elementary gratings discussed above are phase-gratings.

As a consequence, a focused electron image would not show any contrast, if the objective lens possessed no aberrations and the aperture were suffi-

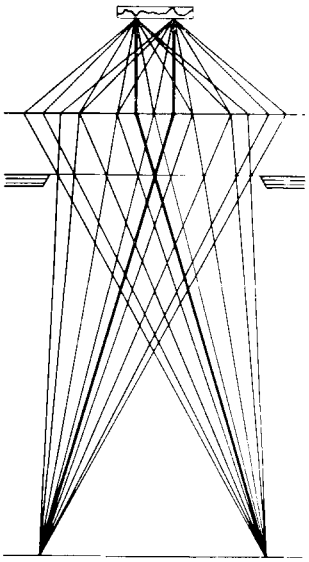


Fig. 2. – Imaging of an amorphous specimen (schematically).

ciently large. The reason is the same as in light optics, it is in a simplified form demostrated in Fig. 3: in the case of a weak phase object, we can assume a diffracted wave with a small amplitude and a phase difference $\pi/2$ compared

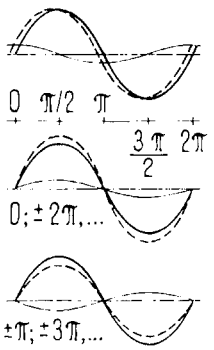


Fig. 3. – For explanation of phase contrast.

to the primary wave. The both waves mentioned are represented by the straight-line oscillations on top in Fig. 3. The interference of these two waves results in a wave with the same amplitude as the primary wave, only the

phase position of it is a little bit different (dotted line). Due to the fact that the amplitude is unaltered, there will be no variation of intensity and consequently no contrast in the image. To achieve contrast, we have to introduce an additional phase shift which brings the primary and the diffracted wave either into equal or into opposite phase position. Then, the interference of the primary wave and the diffracted wave will result in a wave with increased or decreased amplitude, as shown in the center and the lower part of Fig. 3. Consequently the image intensity will be altered and phase contrast arises.

These conditions were already recognized in 1947 by Boersch [19] and he suggested several methods to introduce the necessary phase shift. However, they are combined with extremely high experimental difficulties. We will return to this point in Sect. 3.

The common way to introduce phase shifts in order to get phase contrast is to make use of the wave aberration of the objective lens. Wave aberration is defined as the local distance between an ideal wave front and a real wave front, which is aspheric due to spherical aberration and defocusing. It can be expressed by the corresponding phase shift γ depending on the diffraction angle θ or the reciprocal spatial frequency Λ.

According to Scherzer [20] (modified for thick lenses):

$$\gamma = \frac{\pi}{2\lambda} \cdot (C_{\ddot{o}} \cdot \theta^4 - 2\Delta z \theta^2) , \tag{6}$$

where $C_{\ddot{o}}$ denotes the coefficient of spherical aberration and Δz the defocus value, *i.e.* the distance between the real object plane and the plane which is actually imaged. The function $\gamma(\theta)$ is plotted in Fig. 4 for a number of Δz values. It is clearly to be seen that the phase shift due to aberrations of the objective lens depends strongly on the co-ordinates in the back focal plane. Therefore, it is different for each spatial frequency.

Assuming that phase shifts

$$\gamma = (2n-1) \cdot \pi/2 , \qquad n = \pm 1, \pm 2, \ldots , \tag{7}$$

lead to maximum phase contrast, we find from (6) with (7)

$$\Lambda = + \lambda \left[\frac{\Delta z}{C_{\ddot{o}}} \pm \left[\left(\frac{\Delta z}{C_{\ddot{o}}} \right)^2 + \frac{(2n-1)\lambda}{C_{\ddot{o}}} \right]^{\frac{1}{2}} \right]^{-\frac{1}{2}} . \tag{8}$$

Equation (8) describes the dependence of phase contrast on spherical aberra-

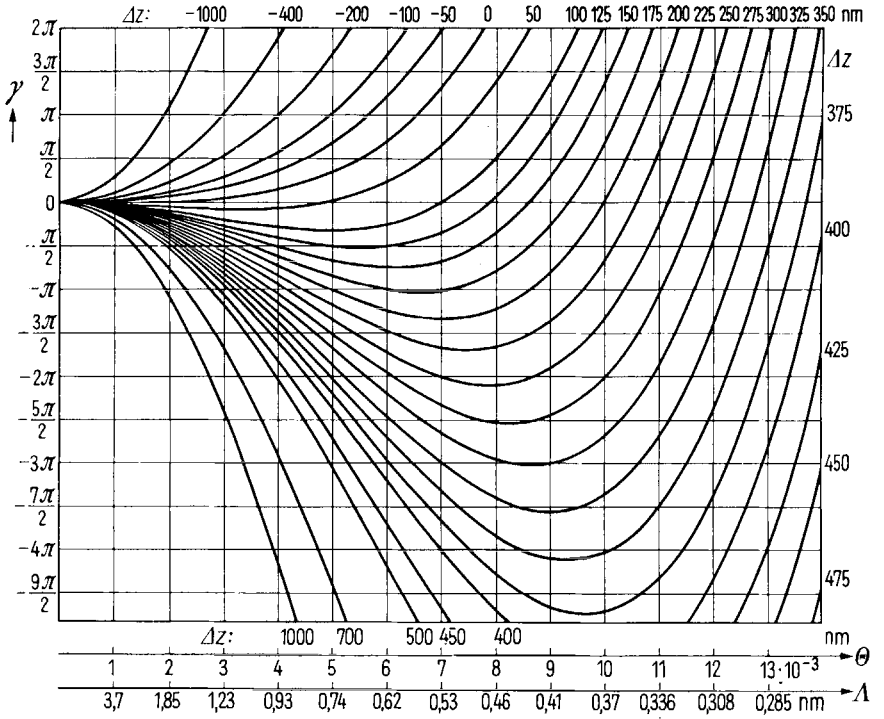


Fig. 4. – Phase shift γ due to defocus Δz and spherical aberration in dependence of diffraction angle θ and reciprocal spatial frequency Λ according to eq. (6) with $C_{\ddot{o}} = 4$ mm and $\lambda = 3.7 \cdot 10^{-9}$ mm.

tion and defocusing (6). With $C_{\ddot{o}} = 0$, which is impossible in practice, follows

$$\Lambda = + \left(\frac{2\lambda \Delta z}{1 - 2n} \right)^{\frac{1}{2}}. \tag{9}$$

We call a graph according to (8) a phase contrast transfer characteristic. Figure 5 shows one with $C_{\ddot{o}} = 4$ mm, $\lambda = 3.7 \cdot 10^{-9}$ mm and $n = +10 \ldots -16$. The dotted curves correspond to (9) with $n = 0$ and $n = +1$. The comparison shows that under the chosen conditions the influence of spherical aberration is negligible, when $\Lambda > 1.2$ nm.

From a phase contrast transfer characteristic, one can immediately read which reciprocal spatial frequencies Λ_t are transferred with maximum phase

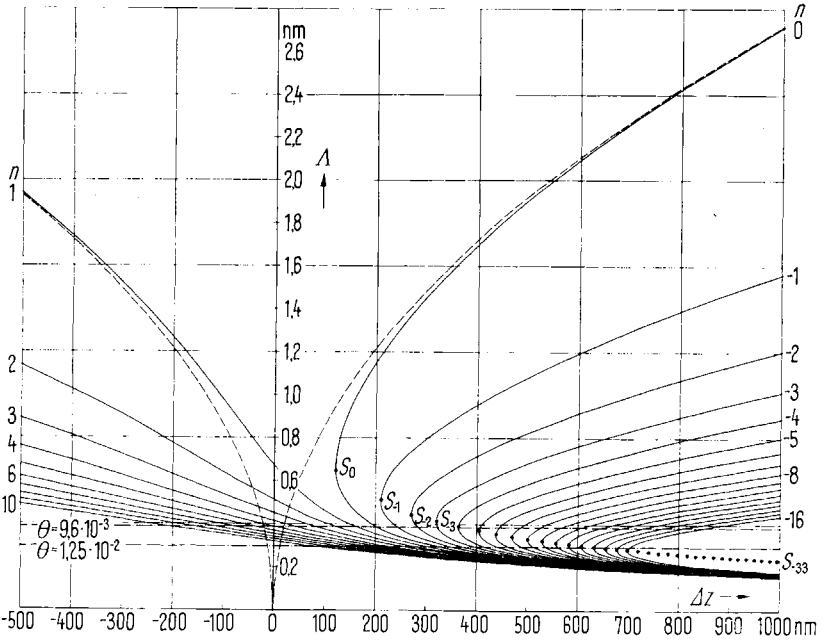


Fig. 5. – Defocusing dependence of the reciprocal spatial frequencies Λ according to eq. (8) (solid curves) and eq. (9), where $C_{\bar{o}} = 4$ mm and $\lambda = 3.7 \cdot 10^{-9}$ mm.

contrast at a given defocus value Δz. At each defocus value, including $\Delta z = 0$, always several specific spatial frequencies are transferred simultaneously. It follows from theory that adjacent frequency bands are transferred with opposite sign of contrast. A phase shift $-\pi/2$ gives positive phase contrast. Negative phase contrast arises by phase shifts of $+\pi/2$. Multiples of 2π are equivalent.

The contrast goes down to zero with phase shifts $\gamma = n\pi$. A curve system analogous to (8) can be calculated from (6). The curves lie between the curves of Fig. 5.

Good contrast for relatively large frequency bands is to be expected in the vicinity of the vertices S_n with $n \leqslant 0$. From (8) we get the co-ordinates

$$S_n \equiv \{+(1-2n)^{\frac{1}{4}} \cdot \lambda^{\frac{1}{2}} \cdot C_{\bar{o}}^{\frac{1}{2}}, \quad +(1-2n)^{-\frac{1}{4}} \cdot \lambda^{\frac{3}{2}} \cdot C_{\bar{o}}^{\frac{1}{2}}\}. \tag{10}$$

The dotted horizontal lines in Fig. 5 mark the influence of a limitation of the objective aperture for two certain values. For all Λ values smaller than

indicated by such a line, the diffraction orders are intercepted by the objective aperture. Consequently these frequencies are not transferred to the image.

For reasons of simplicity we have been discussing one-dimensional specimens only. The results are also valid in the two- or three-dimensional case. For example we can define Λ the period length of a two-dimensional spatial frequency with the components (u, v) and α the azimuthal angle. Then:

$$\Lambda = +(u^2 + v^2)^{-\frac{1}{2}}, \qquad \alpha = \operatorname{arctg} \frac{u}{v}. \tag{11}$$

If the lens field shows exact rotational symmetry, the defocusing is independent of α and is again given by (8).

In practice there is never an ideal rotational symmetry, so an axial astigmatism has to be taken into account. This can be done in the following way ([17]). It is sufficient to consider first order axial astigmatism. Then two planes with a maximum defocus difference can be found, these planes being perpendicular on each other, and both containing the optical axis. These planes are hatched in Fig. 6, their defocus values are Δz_1 and Δz_2, respectively.

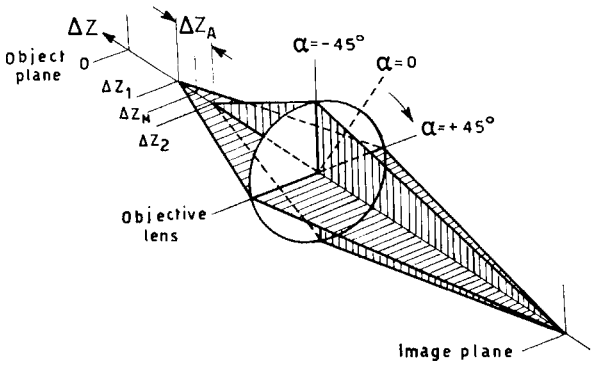


Fig. 6. – For derivation of eq. (16).

We call

$$\Delta z_A = (\Delta z_1 - \Delta z_2) \tag{12}$$

the astigmatic defocus difference.

The defocus value of an intermediate plane, marked by $\alpha = 0$ is

$$\Delta z_M = (\Delta z_1 + \Delta z_2)/2 \,. \tag{13}$$

The azimuthal dependency of defocus Δz can be written

$$\Delta z(\alpha) = \Delta z_M + (\Delta z_A \cdot \sin 2\alpha)/2 \,. \tag{14}$$

The phase shift now is

$$\gamma = \pi(C_{\ddot{o}} \theta^4 - 2 \cdot \Delta z_M \theta^2 - \Delta z_A \theta^2 \sin 2\alpha)/2\lambda \,. \tag{15}$$

Finally, we get

$$\Lambda = +\lambda \left[\frac{2\Delta z_M + \Delta z_A \cdot \sin 2\alpha}{2C_{\ddot{o}}} \pm \left[\left[\frac{2\Delta z_M + \Delta z_A \cdot \sin 2\alpha}{2C_{\ddot{o}}} \right]^2 + \frac{(2n-1)\lambda}{C_{\ddot{o}}} \right]^{\frac{1}{2}} \right]^{-\frac{1}{2}} . \tag{16}$$

This equation enables us to calculate the azimuthal dependency of the spatial frequencies with arbitrary values of Δz_M and Δz_A.

1·3. Experimental demonstrations.

For experimental demonstration of the theoretical statements it has been proved extremely useful to investigate high resolution images of thin carbon foils by means of a light optical diffractometer ([11]).

It turned out that carbon films have approximately a white frequency spectrum, that means all spatial frequencies, which are of interest in high resolution imaging, are contained in the specimen. Thus, one can check the transfer conditions of a lens by considering the frequency spectrum of a carbon foil image, which was taken under definite operating conditions. Additionally, these techniques give information about imaging parameters.

To evaluate the frequency spectrum of an image intensity distribution, *i.e.* the power spectrum, a light optical diffractometer is most convenient. It is not absolutely necessary to use exact parallel illumination for this purpose ([11]). A simple arrangement, as proposed in 1969 by Mulvey ([21]) and schematically shown in Fig. 7, also allows one to determine the most important parameters of the image, *e.g.* the defocus difference Δz and the astigmatic defocus difference Δz_A, with completely sufficient accuracy. As a light source

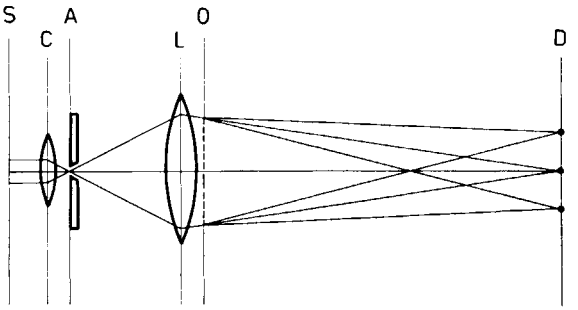


Fig. 7. – Scheme of a simple light optical diffractometer. S = source; C = condenser; A = aperture; L = lens; O = object; D = diffraction plane.

S one should use in any case a laser ((1÷10) mW) to have sufficient intensity. As a condenser C a light microscope objective lens may be used. The size of the effective source should be limited by an aperture A of about 10 μm in diameter. A camera lens L ($f \approx 100$ mm) works as a diffraction lens, focusing the rays diffracted by the object O, *i.e.* the optical density distribution on the electron plate, in a plane D. Thus, a two-dimensional Fourier transformation of the electron image can be performed. Registration of the intensity distribution in plane D yields knowledge of the power spectrum of the electron image. The electron plate may be placed immediately against the lens or at another convenient place between L and D. This enables one to vary the scale of the Fourier transform, which is extremely useful if different electron optical magnifications have been used in taking the electron micrographs.

Figure 8 gives an example for the evaluation of a phase contrast electron image. At the bottom left-hand corner, a section of the image structure to be examined is shown, and above the light optical diffraction pattern of the structure. This pattern clearly proves that actually only selective frequency bands are transferred, and not the whole frequency spectrum. The radii r_L in the diffraction plane correspond to reciprocal spatial frequencies A of the electron microscopic object according to the equation

$$A = \frac{\lambda_L \cdot f_L}{V \cdot r_L}, \tag{17}$$

where λ_L is the wavelength of the laser light, f_L is the focal length of the diffraction lens, and V the electron optical magnification.

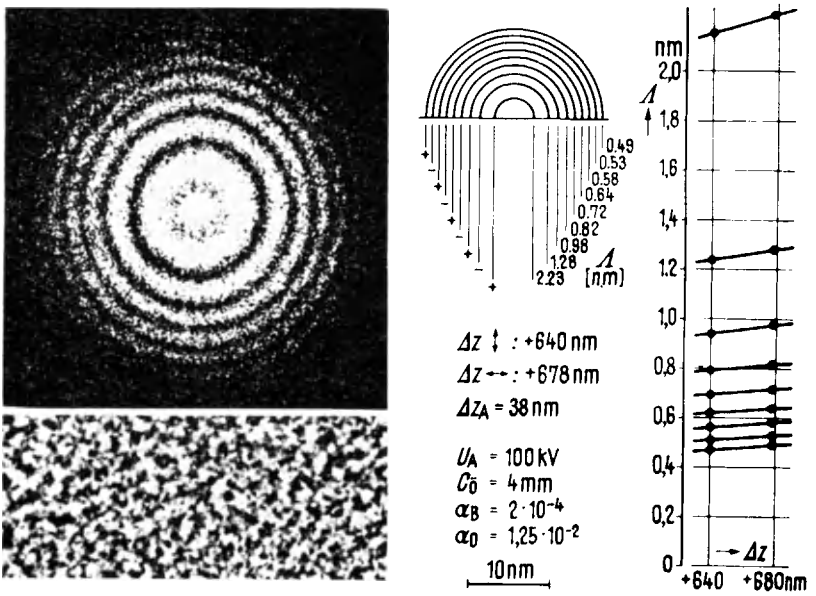


Fig. 8. – Example for the evaluation of phase contrast image structure.

Equation (17) is valid in the case of parallel illumination of the specimen, if the diffraction lens is arranged behind the specimen. In the case of the simpler arrangement described by Fig. 7, A becomes dependent on the position of the electron plate with respect to the diffraction plane D.

In both cases it is useful to calibrate the arrangement, using a grating with known period length.

The A values corresponding to the diffraction maxima in the horizontal direction of the pattern are marked at the scheme on the right hand side of Fig. 8. This combination of A values fits exactly at a defocus value $\Delta z = +678$ nm to the phase contrast transfer characteristic, a section of which is shown.

An evaluation in the vertical direction of the slightly elliptic pattern yields somewhat different A values. The corresponding defocus value is $\Delta z = +640$ nm, thus revealing an astigmatic defocus difference $\Delta z_A = 38$ nm. The accuracy of the method is much higher compared to others known before. And it is essential to know the exact values of defocus and axial astigmatism for any further interpretation of the image structure. Figure 9 demonstrates, how the image structure, and, correspondingly, the diffraction patterns vary with different defocus. Four special defocus values

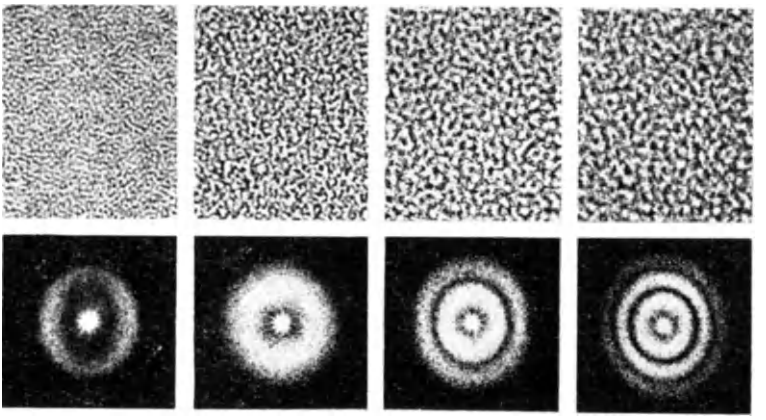


Fig. 9. – Focusing series on a carbon foil and the corresponding light optical diffraction patterns.

have been used for the 4 micrographs shown. The first (from left) corresponds approximately to Gaussian imaging, $\Delta z = 0$. The others are close to the vertices S_0, S_{-1} and S_{-2} in terms of Fig. 5.

$\Delta z = 0$ gives a very poor imaging as there has been only one small frequency band ($n = 1$) transferred to the image, as a consequence of the finite illuminating aperture. The optimum defocus according to Scherzer [20] provides in case of a lens with $C_{\ddot{o}} = 4$ mm, as used in the experiments, a frequency band from about 5 Å to 15 Å to be transferred with observable contrast. If the objective aperture has been chosen appropriately, then only one relatively wide frequency band contributes to the image. This seems to be the only reasonable way to use phase contrast for the investigation of unknown specimens, if a circular aperture is used and no special means for the evaluation of the micrographs are available. Clearly, this frequency band is not wide enough for all applications. And the resolution attainable is not very high.

The transfer limit can be shifted toward higher frequencies or smaller A values by using a lens with a smaller coefficient of spherical aberration. This will be combined with a further decrease in width of the frequency band.

If we want to expand the transmitted frequency range, we have to defocus the lens more. But then gaps in the frequency spectrum occur, as can be seen from micrographs 3 and 4 from the left in Fig. 9, and adjacent frequency bands are transmitted with opposite phase position. Without

any interventions in the back focal plane, as will be discussed in Sect. **2** and **3**, it is advisable to use the Scherzer optimum defocus.

The experimental values following from the evaluation of a whole focusing series are in excellent agreement with theory as demostrated by Fig. 10.

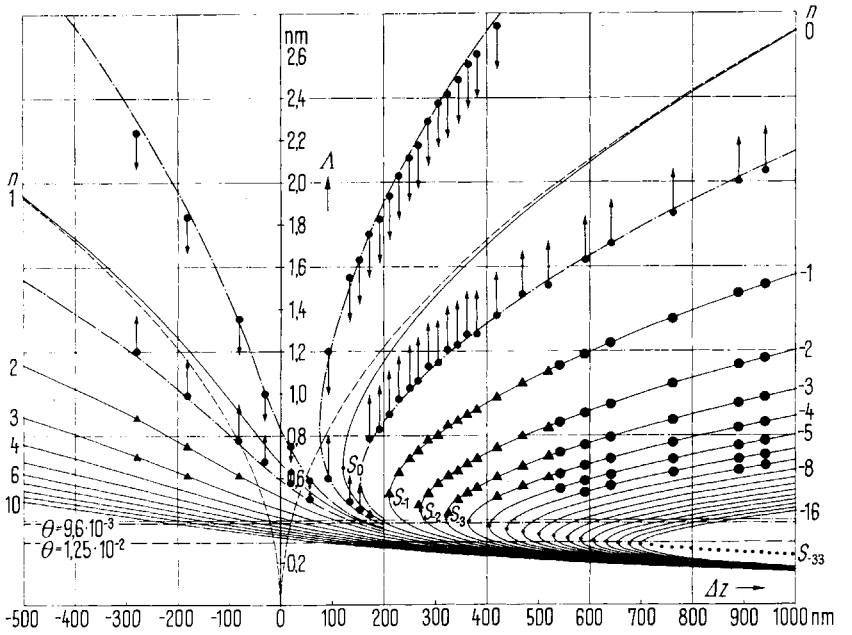


Fig. 10. – Defocusing dependence of the transmission of reciprocal spatial frequencies Λ, as measured from two focusing series, in comparison with the theoretical curves.

In cases of $n = 0$ and $n = 1$, not the diffraction maxima but the bandwidths were determined from the patterns and plotted together with curves (dot-dashed), calculated from

$$\Lambda = +\lambda\left[\frac{\Delta z}{C_\delta} \pm \left[\left(\frac{\Delta z}{C_\delta}\right)^2 + \frac{(2n-1\pm 0.6)\lambda}{C_\delta}\right]^{\frac{1}{2}}\right]^{-\frac{1}{2}}. \tag{18}$$

Equation (18) has been derived in the same way as eq. (8) but using

$$\gamma = (2n-1\pm 0.6)\pi/2\,, \tag{19}$$

instead of eq. (7). The term 0.6 has been established from experimental results ([18]).

According to theory, adjacent frequency bands are transferred with contrast of opposite sign. Consequently when going from positive to negative defocus there is a contrast reversal and *vice versa*. This is demonstrated in Fig. 11. This figure shows images of the same area of a carbon foil, taken

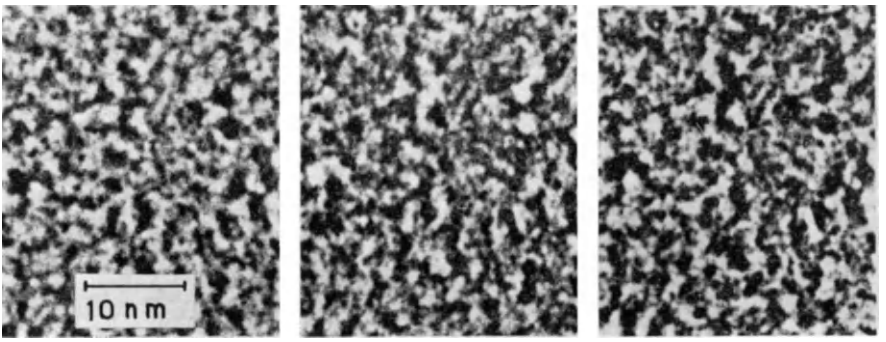


Fig. 11. – Image structures of a carbon foil at different defocus values Δz. There is a complete contrast reversal from the micrograph at the left to the micrograph in the middle. The micrograph on the right hand side is a photographic negative from the middle one.

at $\Delta z = -940$ nm (left) and $\Delta z = +960$ nm (centre), respectively. A comparison of the two images confirms the theoretical predictions. The micrograph on the right-hand side is a photographic negative of the centre one; this makes the comparison easier.

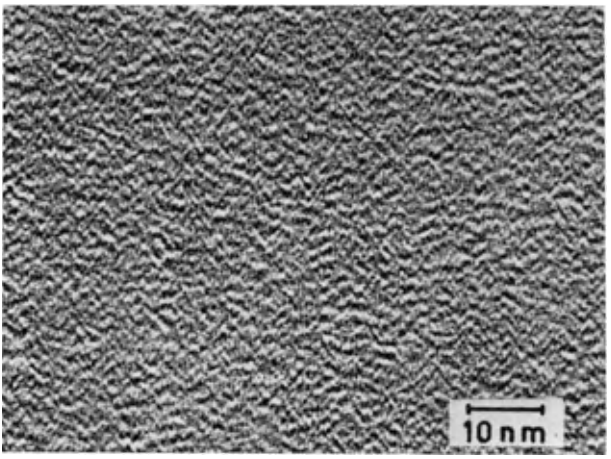


Fig. 12. – Image structure of a carbon foil taken with an axial astigmatism $\Delta z_A = 364$ nm.

Equation (16), which describes the influence of axial astigmatism on phase contrast transfer, has been checked experimentally in the following way ([17]).

Figure 12 is a section of an image structure taken with a relatively highly astigmatic objective lens. The astigmatic defocus difference amounts to $\Delta z_A = 364$ nm. This is large enough to be set in experiments with sufficient accuracy.

Figure 13 shows the corresponding light optical diffraction pattern. It looks quite different from the pattern shown in Fig. 8, although the astigmatism of the latter is only about a factor of ten less. This demonstrates the high sensitivity of the method.

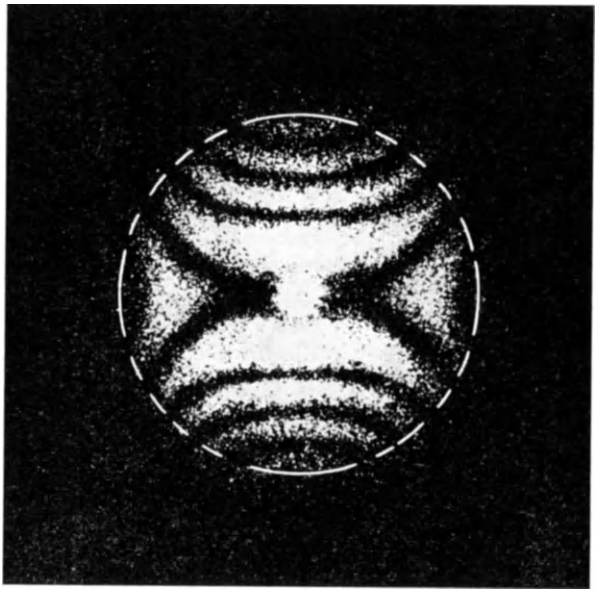


Fig. 13. – Light optical diffraction pattern of the image structure shown in Fig. 12.

For a calculation of a two-dimensional frequency spectrum according to eq. (16) we need besides of Δz_A the value Δz_M or one of the main defocus values. The only way we can get it with sufficient accuracy is to take it from the optical diffraction pattern. From Fig. 13 follows $\Delta z_1 = +358$ nm in the vertical direction. Using this value, we calculated from eq. (16) the azimuthal dependency of the reciprocal spatial frequencies $\Lambda \geqslant 5$ Å with

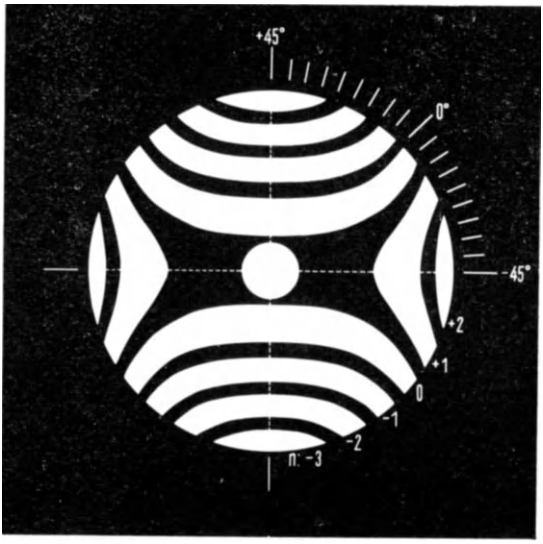


Fig. 14. – Calculated power spectrum according to eq. (16), where $\Delta z_A = 364$ nm and $\Delta z_1 = +358$ nm.

$\Delta z_A = 364$ nm. The calculated Fourier transform (Fig. 14) is in good agreement with the experimental one. Equation (16) obviously describes the influence of first order axial astigmatism correctly.

It is now necessary to consider the influence of the finite illuminating aperture since our theoretical considerations were based on the assumption that the illuminating aperture was exactly zero. In practice, however, its value is in the range of 10^{-3} to 10^{-4} rad. It is usual to consider an object detail of size d to be illuminated effectively coherent even in the case of an extended source, if

$$d \approx \frac{\lambda}{\pi \cdot \alpha_B}, \tag{20}$$

where α_B is the illuminating aperture. With $\lambda = 3.7 \cdot 10^{-9}$ mm, as used in experiments, we have $d \approx 10$ Å to 100 Å, when $\alpha_B = 10^{-3}$ to 10^{-4}. Therefore, with $\alpha_B \approx 5 \cdot 10^{-4}$ the illumination should be effectively coherent for all details of interest.

It should be pointed out, however, that the coherence condition (20) is not satisfactorily applicable in the case of defocused imaging using a lens with spherical aberration.

The diffracted intensity of each spatial frequency fills a finite angle of approximately $2\alpha_B$. If the phase shift according to the wave aberration of the objective lens is strongly varying within such a range, the phase contrast will be destroyed. This means, the value of α_B, which can be tolerated, depends on the defocus value of the objective lens.

In experiments, an increase of α_B from $2 \cdot 10^{-4}$ to $2 \cdot 10^{-3}$ had little effect in the vicinity of the first vertex S_0 of the transfer characteristic (Fig. 5), whereas phase contrast for high spatial frequencies was completely destroyed at defocus values about three times stronger, using the higher illuminating aperture.

Finally, Fig. 15 demonstrates the influence of a limited objective aperture. If the $+1$ and -1 diffraction order of an elementary grating are intercepted by an aperture, this grating cannot be transferred to the image.

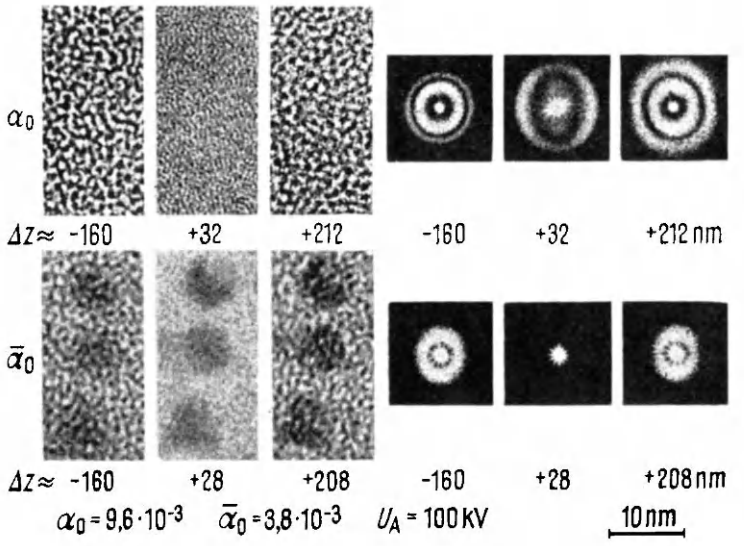


Fig. 15. – Image structures of a carbon foil (left) and corresponding diffraction patterns (right), taken with two different limitations of the objective aperture α_0 and $\bar{\alpha}_0$.

Accordingly, with the values used in experiment, transfer limits result at $\Lambda = 4$ Å (upper row) and 10 Å (lower row), respectively. In agreement with the diffraction formula for partially coherent illumination

$$\delta_{pp} = \frac{0.77\lambda}{\alpha_0}, \tag{21}$$

where α_0 is the objective aperture and δ_{pp} can be considered a point-to-point separation, no image details closer than about 7 to 8 Å appear in the micrographs shown in the lower row. The decrease in resolution can clearly be seen from the light optical diffraction patterns on the right-hand side of Fig. 15. A comparison with the transfer curves of Fig. 5 makes it understandable, why phase contrast is almost totally eliminated with a defocus value of $\Delta z = +28$ nm (lower row in Fig. 15). This reveals a possibility to separate phase contrast and scattering absorption contrast components by appropriately choosing the objective aperture and the defocus value.

Summarizing, one can say that the conventional method to produce phase contrast by defocusing is very unsatisfactorily, if circular apertures are used. With a defocus value close to the vertix S_0 of the transfer characteristic, which value corresponds to the Scherzer optimum defocus, only one relatively wide continuous frequency band is transferred if the objective aperture has been chosen appropriately. But the resolution, which can be obtained under those conditions, is restricted to values lower than desirable. With stronger defocus values, higher spatial frequencies will be transferred but, due to the occurrence of selective frequency bands and a reversal of contrast, the interpretation of the micrographs becomes complicated or even impossible.

Therefore techniques, which improve the phase contrast transfer properties, are urgently needed. Those methods will be subject of the following Sections.

REFERENCES (Section 1)

1) K.-J. HANSZEN, B. MORGENSTERN and K.-J. ROSENBRUCH: *Zeits. Angew. Phys.*, **16**, 477 (1964).
2) F. THON: *Proc. 3rd Eur. Reg. Conf. on Electron Microscopy, Prague 1964* (Prague, 1964), vol. **1**, p. 127.
3) L. ALBERT, R. SCHNEIDER and H. FISCHER: *Zeits. Naturfor.*, **19***a*, 1120 (1964).
4) R. D. HEIDENREICH and R. W. HAMMING: *Bell System Techn. Journ.*, **44**, 207 (1965).
5) K.-J. HANSZEN and B. MORGENSTERN: *Zeits. Angew. Phys.*, **19**, 215 (1965).
6) F. THON: *Zeits. Naturfor.*, **20***a*, 154 (1965).
7) F. LENZ: *Optik*, **22**, 270 (1965).
8) C. B. EISENHANDLER and B. M. SIEGEL: *Journ. Appl. Phys.*, **37**, 1613 (1966).
9) L. REIMER: *Zeits. Naturfor.*, **21***a*, 1489 (1966).
10) K.-J. HANSZEN: *Zeits. Angew. Phys.*, **20**, 427 (1966).
11) F. THON: *Zeits. Naturfor.*, **21***a*, 476 (1966).
12) F. THON: *Proc. 6th Int. Conf. on Electron Microscopy, Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 23.
13) F. THON: *Physikertagung* (1966), *Vorabdruck der Fachberichte*, Stuttgart, p. 101.

14) R. LANGER and W. HOPPE: *Optik*, **24**, 470 (1966/67).
15) F. THON: *Phys. Bl.*, **23**, 450 (1967).
16) F. THON: *Siemens Review*, **36**, 24 (1969) (3rd Special Issue).
17) F. THON: Paper at the *Meeting of the German Society of Electron Microscopy*, Marburg (1967).
18) F. THON: *Thesis*, Tübingen (1968).
19) H. BOERSCH: *Zeits. Naturfor.*, **2***a*, 615 (1947).
20) O. SCHERZER: *Journ. Appl. Phys.*, **20**, 20 (1949).
21) T. MULVEY: *Conference on Non-Conventional Electron Microscopy*, Oxford (1969). See also *Proc. 7th Int. Conf. on Electron Microscopy, Grenoble 1970* (Paris, 1970), vol. **2**, p. 65.

2. High resolution microscopy using special apertures.

2·1. Introduction.

In this second Section we are going to discuss how contrast transfer can be influenced and possibly improved by inserting special apertures into the back focal plane of the electron microscope. Common to all methods to be discussed is the fact that a certain amount of the electron beam after having passed the object is intercepted completely while the remaining part will pass uninfluenced.

Only when applying the first method (zone correction plates) is phase contrast in its direct meaning still employed. Image contrast is, as pointed out in Sections **1**, attained by interference of the primary beam and the $\pm$ 1st diffraction orders.

By inserting an asymmetric aperture into the back focal plane (semicircular aperture) we can show that the value of contrast no longer depends on the wave aberration of the objective lens. Thus the frequency dependence of contrast is eliminated, too. On the other hand further complications arise as will be shown. With the third method (dark field) the intensive zero-diffraction order is intercepted completely. As in high resolution electron microscopy dominantly coherent illumination is used, this method may lead to quite intricate imaging conditions.

2·2. Zone correction plates.

Since the contrast of adjacent spatial frequency bands changes its sign, as pointed out in Section **1**, the chance of interpreting high resolution images becomes quite small.

The function $K(\Lambda)$ describing contrast transfer oscillates between positive and negative values, according to the local variation of the wave aberration. An example for such a function, which can in principle be derived from the Fig. 4 or 5, is shown in Fig. 16.

The dotted line in the diagram stands for an ideal contrast transfer function. Such an ideal transfer function is supposed to be of constant value

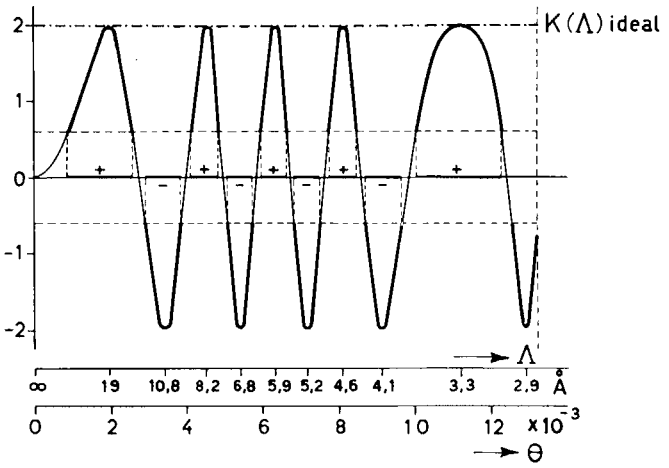


Fig. 16. – Phase contrast transfer function of an electron objective lens with $C_{\ddot{o}} = 4$ mm at a defocus value $\Delta z = +500$ nm (accelerating voltage 100 kV). The real CTF is oscillating between positive and negative values. The contrast has to exceed a value of about 0.6 to become perceptable in the electron micrograph.

from zero up to the highest relevant spatial frequencies. The real conditions when imaging with a circular aperture are quite far from this.

More favourable transfer conditions than in the case of the circular aperture could be attained according to a proposal of Hoppe ([1]), if only equiphase waves were admitted to the imaging process while intercepting the waves of opposite phase direction. This means that only certain ringlike sections of the objective lens are to transmit waves while the others have to be opaque. A detailed theoretical treatment of the effectiveness of the method has been given by Hoppe and Langer ([2]).

In the diagram of Fig. 16 such a zonal intervention could be illustrated by cutting off the negative values of the function $K(\Lambda)$ thus giving rise to additional gaps in the spatial frequency spectrum. According to the calcula-

tions of Hoppe and Langer (2) for the case of imaging single atoms this should result in a positive effect on the shape of the image point intensity distribution.

There were two main difficulties which had to be overcome in trying to realize Hoppe's idea.

First of all, new techniques of preparing such zone correction apertures with their extremely minute dimensions had to be found. Second an extraordinary high experimental skill had to be applied while working with these correction plates in the microscope. The problems connected with the preparation of the plates have been solved by Möllenstedt and colleagues (3). The method of plate preparation by means of an electron optical demagnification system was published elsewhere (4). An example of a zone correction plate produced by these authors is shown in Fig. 19 in the top left hand corner. The diameter of this plate is about 60 μm, the axial thickness is less than 1 μm.

To understand where the experimental difficulties in applying these plates arise from, let us discuss Fig. 17. This shows the transfer characteristic of

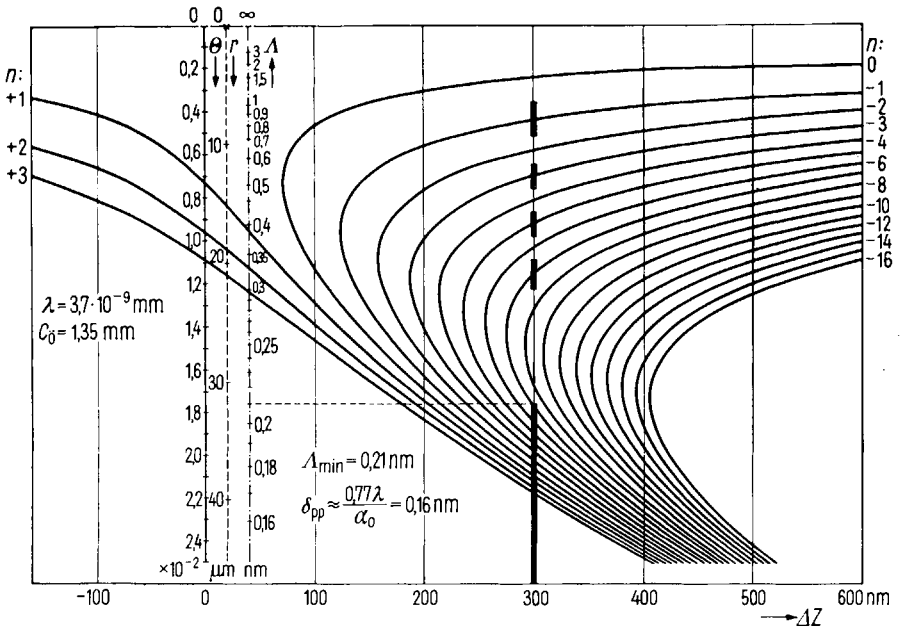


Fig. 17. – Phase contrast transfer characteristic of a special objective lens with $C_{\delta}=1.35$ mm. The intercepting effect of a zone correction plate is indicated by the broad dark lines at $\Delta z = +300$ nm.

a special objective lens for use with the Elmiskop 101. The coefficient of spherical aberration in this case amounts to 1.35 mm. For practical reasons the diffraction angle θ and the co-ordinate in the aperture plane are plotted linearly, instead of the reciprocal spatial frequency A. Therefore this representation looks a little bit different from that used in Fig. 5. At the defocus value $\Delta z = +300$ nm, those frequency bands which are intended to be eliminated by the zone plate of the employed type, are indicated by broad dark lines. The elimination of every second frequency band ensures that only frequency bands of identical sign of contrast are transferred to the image plane. The transfer limit for the reciprocal spatial frequencies determined by this plate is 2.1 Å. According to the equation for point to point resolution in the case of partially coherent illumination, as indicated in Fig. 17, a true resolution of 1.6 Å should be possible.

From Fig. 17 it is quite obvious that the quality of imaging with zone plates is very sensitive to deviations from the calculated defocus value, to small amounts of axial astigmatism and demands accurate alignment of the plate in respect to the optical axis.

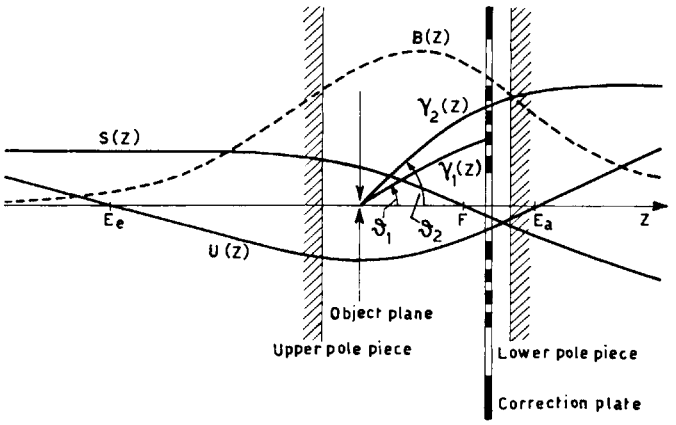


Fig. 18. – Field curve $B(z)$ and electron path $S(z)$ in the case of the special lens for use with the Elmiskop 101. The position of the zone correction plate is indicated.

In Fig. 18 the electron optical conditions in the case of the special objective lens mentioned above are schematically shown.

The field curve B is indicated by a dotted line. The electron path S, which enters parallel to the optical axis, determines the back focal plane at point F. The electron path U, which is parallel to the optical axis at the position of

the object, determines the entrance pupil E_e and the exit pupil E_a. The zone plate should be arranged in the diffraction plane of the objective lens at E_a. In the case of strong lenses, this plane is always located inside the lower pole piece bore. It is therefore inconvenient to position the zonal plate at this point. The plate however, can be located in the gap between F and E_a, providing the plate dimensions are correspondingly altered. If one follows the electron path γ_1 which is stopped by the plate, and γ_2, which passes through an opening in the plate, it can be clearly seen that the dimensions must be reduced. The axial position for which the zonal plate has definitely been calculated, must be kept to within a tolerance of not greater than 100 μm. The requirements governing the lateral aligment are far more severe. In order to obtain a correct elimination of the intended spatial frequencies, the plate axis must not deviate more than 0.5 μm from the optical axis. Furthermore, the astigmatic defocus difference of the objective lens should be less than 4 nm.

Currently known methods do not enable such an accurate compensation of the axial astigmatism to be made to within any degree of certainty, and it is purely by coincidence, when it so happens. The defocus value, for which the zonal plate is calculated, must likewise be set within a tolerance of about 4 nm. This could be met in the case of the arrangement employed, because it was possible to produce focusing series with defocus intervals of 3.9 nm. Further requirements are: The illuminating aperture must be less than $4 \cdot 10^{-4}$ rad; the relative fluctuations of the lens current and high voltage must be less than $2 \cdot 10^{-6}$. The plate must naturally be completely free from electrostatic charging during the whole investigation. The requirements on the whole are quite difficult to meet.

Nevertheless the effect of zone correction plates could already be demonstrated in experiments, using carbon foils as testing objects ([4,5]). Figure 19 shows a micrograph taken with a zone plate. The plate is shown in the top left hand corner, it was shadow photographed while in the diffraction plane of the objective lens. The image structure of the carbon foil looks more uniform than in the case of a circular aperture at the same defocus. The light optical diffraction pattern in the bottom right hand corner clearly shows intensity for A values as low as 3 Å. That means, in practice a point to point resolution of 2.3 Å is attained. Details with such distances and with distances even slightly smaller are actually contained in the image.

The effect of the zone plate on the transfer properties can be better recognized by comparing electron images and their corresponding light-optical diffraction patterns, the images taken with and without a zone plate at approx-

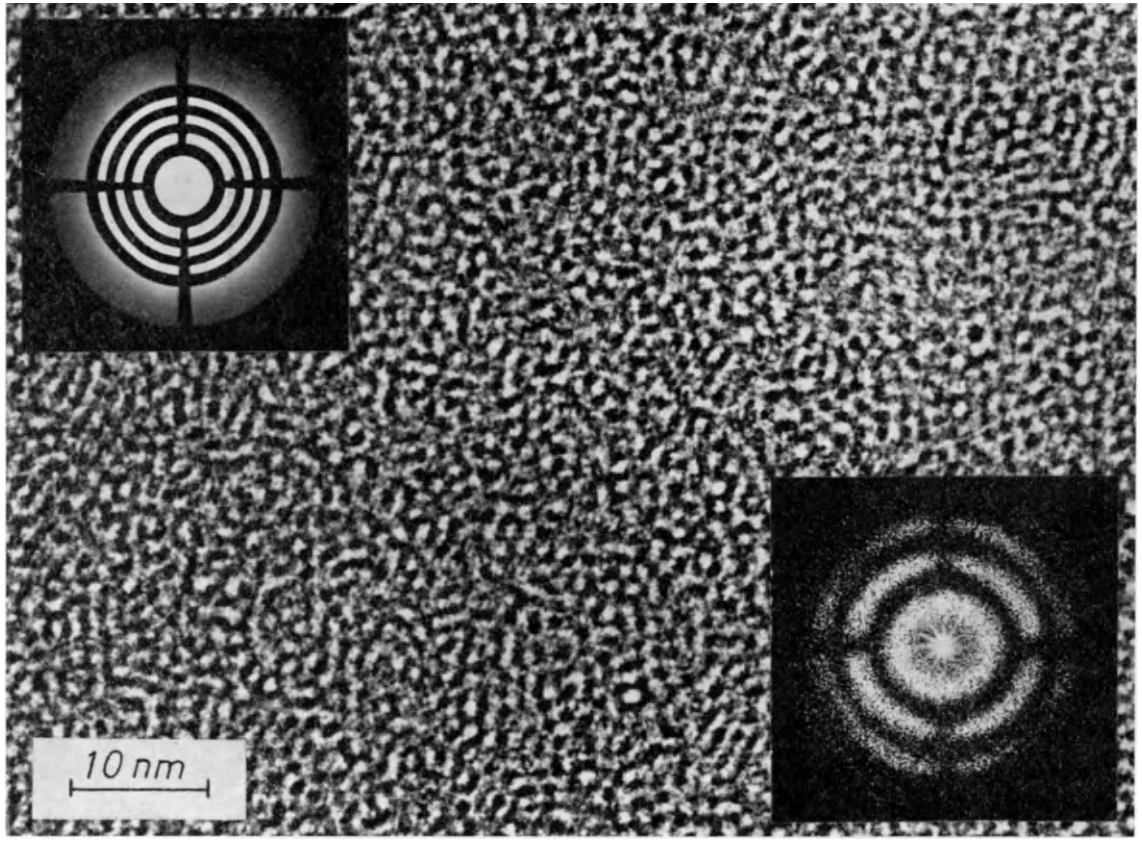


Fig. 19. – Image of a carbon foil, taken with the zone correction plate, a photograph of which is arranged in the top left-hand corner. Bottom right-hand corner: Light optical diffraction pattern of the image structures.

imately the same defocus value under identical operating conditions. The result of such an experiment is shown in Fig. 20, it was made using an

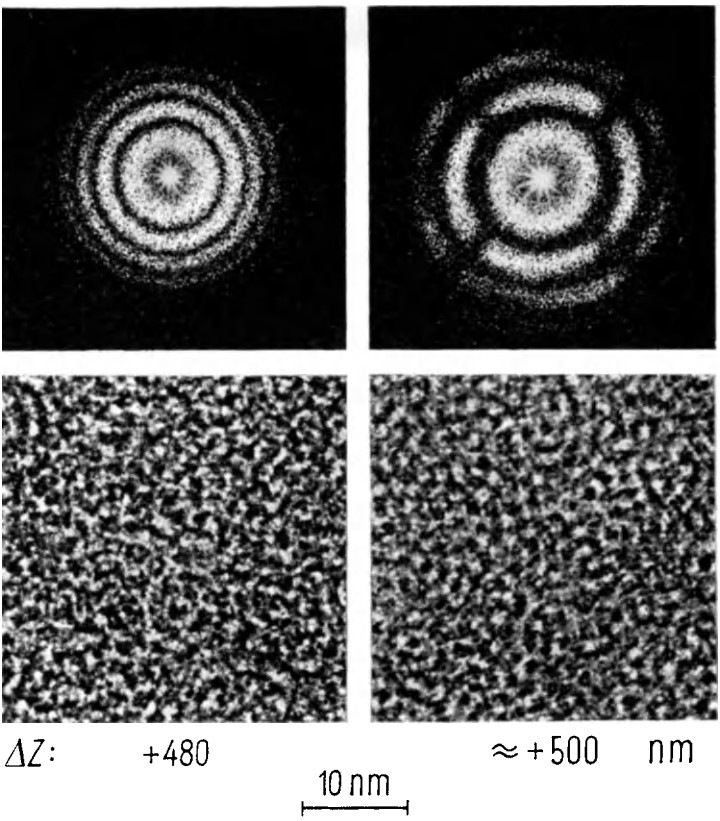


Fig. 20. – Comparison of carbon foil images and of the corresponding light optical diffraction patterns in case of circular objective aperture (left) and zone correction plate (right).

Elmiskop IA. The comparison of the diffraction patterns shows, that in the case of the zone plate (right) the transfer properties, especially for higher spatial frequencies, are remarkably improved. The 2nd, 4th and 6th frequency bands have been eliminated, but therefore the remaining bands are broader than the corresponding bands (1st, 3rd and 5th band) in the case of the circular aperture (left). Additionally, diffraction intensity is indicated

for another band, that does not occur in the image taken with a circular aperture.

In principle, zone plates of the type used would enable one to achieve a point to point resolution of 1.6 Å. Up to now it has not been possible to attain this value.

The cause for this could be due to difficulties in applying the zone plates but also other limitations, for instance anomalous energy distributions of the electron beam can have affected the imaging process. It is also possible that the objects used did not fulfill the necessary requirements. In any case the recent experiments confirm in principle the possibility of the application of zonal correction plates in high resolution electron microscopy.

The big experimental difficulties in applying zone plates initiated experiments for zonal filtering in light optical reconstruction. This technique is an alternative to zonal correction in bright-field phase contrast only. Within the microscope itself zonal correction plates for dark-field microscopy are also applicable. Still there has not been a convincing demonstration of the effectiveness of zone plates applied to high resolution imaging of real objects.

2·3. Semicircular apertures.

If either the + 1st or the — 1st diffraction order is eliminated from the imaging process, then the contrast of a spatial frequency is no longer dependent on the wave aberration of the objective lens. A theoretical treatment has been published by Hanszen and Morgenstern (6) for the case where this asymmetrical elimination of diffracted intensity is achieved by oblique illumination of the specimen, using a circular objective aperture. In principle, this technique allows the whole frequency spectrum to be transferred simultaneously. This would be classified as an ideal condition. Unfortunately, this asymmetrical intervention causes lateral phase shifts, which are again dependent on the wave aberration of the objective lens. It means that different displacements of the individual spatial frequencies occur in the image plane. This has no consequence, as long as only one single frequency is relevant to the imaging process, such as in the case of lattice imaging. In the case of amorphous objects, it leads to a decisive limitation.

According to calculations made by Hanszen (6) for a one-dimensional object, a better resolution should be obtainable even with amorphous objects, if balanced conditions in respect of illumination angle, wave aberration and size of aperture are met. In the actual two-dimensional case, however, the

azimuthal dependency of the aperture limitation produces a negative effect, as demonstrated recently ([7]).

The azimuthal dependency of the aperture limitation and of the transfer conditions is excluded, if the necessary elimination is attained by intercepting one side of the aperture using a special semicircular aperture ([7,8]). The illumination, of course, has to be axial. Figure 21 shows schematically (on the right-hand side) the shape and the arrangement of this type of aperture.

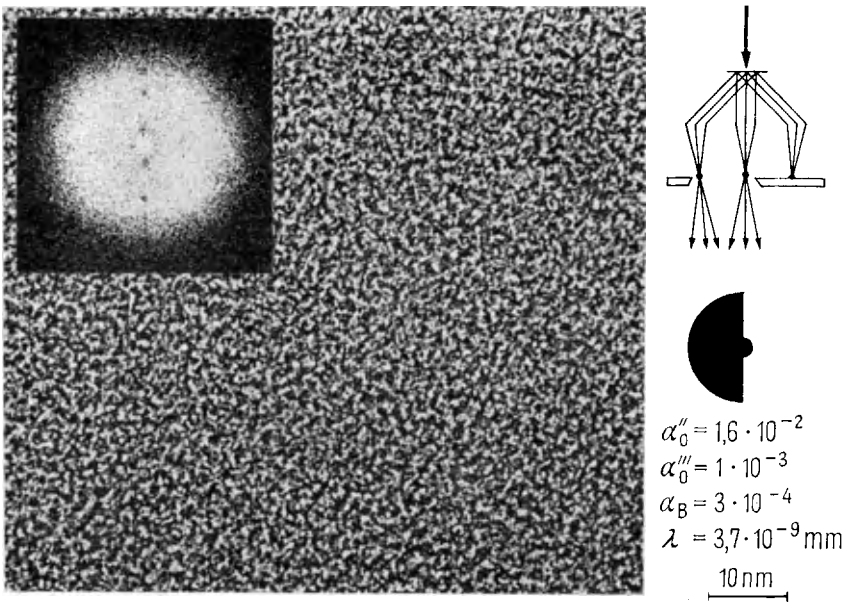


Fig. 21. – Carbon foil image, taken with a semicircular aperture. Shape of the aperture and data used in the experiment are shown on the extreme right.

The open area is not a correct semicircle, there is an additional small open area which allows the primary beam to pass through. This has been found necessary for experimental reasons. The carbon foil image shown in Fig. 21 has no preferential orientation, the structure details correspond to distances less than 3 Å in the object plane. The light optical diffraction pattern in the top left hand corner shows a rotational symmetry and a continuous distribution of diffraction intensity, as expected from the theoretical considerations. There are still three-beam interferences indicated in a very small ver-

tical region of the pattern. This is a result of the aperture not having been exactly centered.

It is surprising, how little the appearance of the image structure is altered with relatively large changes of the defocus value. This is demonstrated by Fig. 22. The upper row shows a focusing series taken with three-beam interference conditions, that means using a circular aperture, over a range of 305 nm, and the corresponding light optical diffraction patterns above. The total as well as the detailed image structure changes quite drastically. The series below were taken with a semicircular aperture. Here the defocus range is extended to 330 nm, and we notice very little change in the structure. For example identical details with distances of about 0.5 nm can be detected

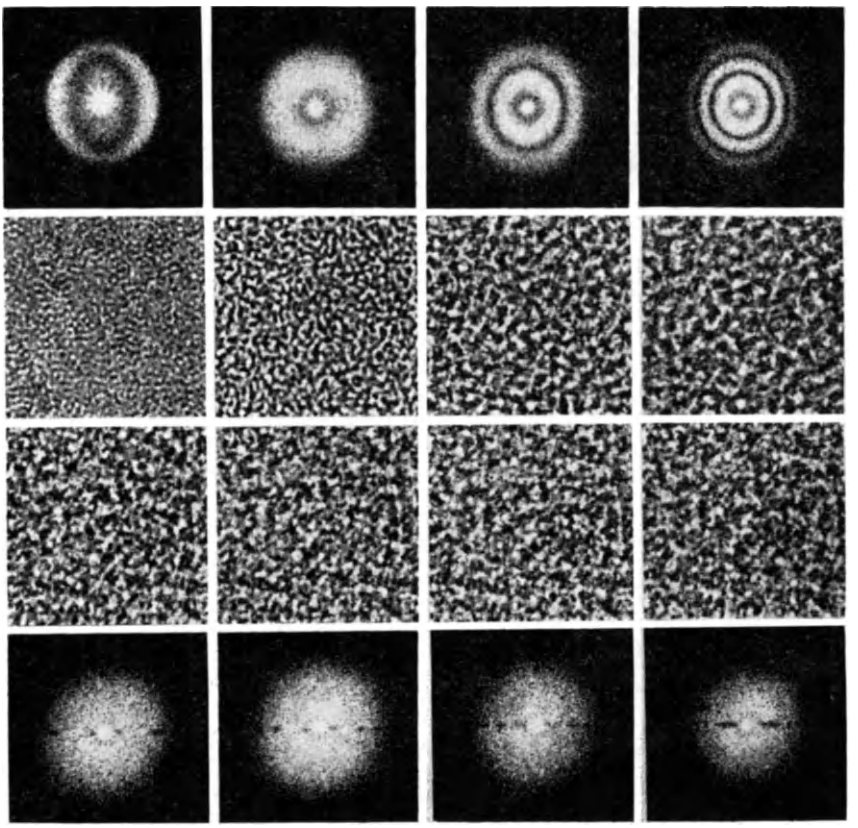


Fig. 22. – Focusing series on carbon foils and corresponding light optical diffraction patterns in case of circular aperture (top) and semicircular aperture (bottom). The defocusing range is about 320 nm in both cases.

over the whole defocus range. In the case of three-beam interference, as in the upper row, details of this size would not be identifiable due to a multiple contrast reversal.

Thus, the transfer properties seem to be quite ideal in the case of semicircular apertures, but we still have to take into consideration the lateral phase shifts which are included in this method. Further considerations on this subject will be made in the near future.

Very recently a new method for reconstruction of complex image functions has been suggested by Hoppe, Langer and the author ([9]). Two consecutively taken exposures using complementary arranged semicircular apertures enables one to reconstruct the Fourier coefficients of a complex image amplitude with equal weight for all Fourier components.

Also very recently Hanszen ([10,11]) pointed out that an image taken with a semicircular aperture can be considered a single-side band hologram. A micrograph of this type can be perfectly reconstructed using a matched light optical reconstruction system. From a theoretical point of view, this reconstruction method is almost ideal.

First experiments suffered from insufficient quality of the electron micrographs due to axial astigmatism which was introduced by the semicircular apertures. We are now going to provide means which should prevent the apertures from being contaminated.

2·4. Dark field methods.

H. Boersch ([12]) introduced dark field imaging into electron microscopy in 1936. This technique has been proved extremely useful in some types of investigations. During the last years Dupouy ([13]) has shown excellent dark field images in the field of medium resolution. He used dark field apertures with a central stop, thus maintaining axial illumination of the specimen.

In the high resolution field, however, there seem to be some restrictions on the applicability of this technique. The reason lies probably in the fact, that the high degree of coherence usually present when working under high resolution conditions, is disadvantageous in dark field electron microscopy.

In imaging carbon foils with dark field apertures of the type shown in Fig. 23 on the right hand side, we got image structures with details in the 3 Å region ([7]). The question arose, whether these structures are due to real dark field contrast analogous to scattering absorption contrast in bright field imaging, or whether they are results of an interference between the

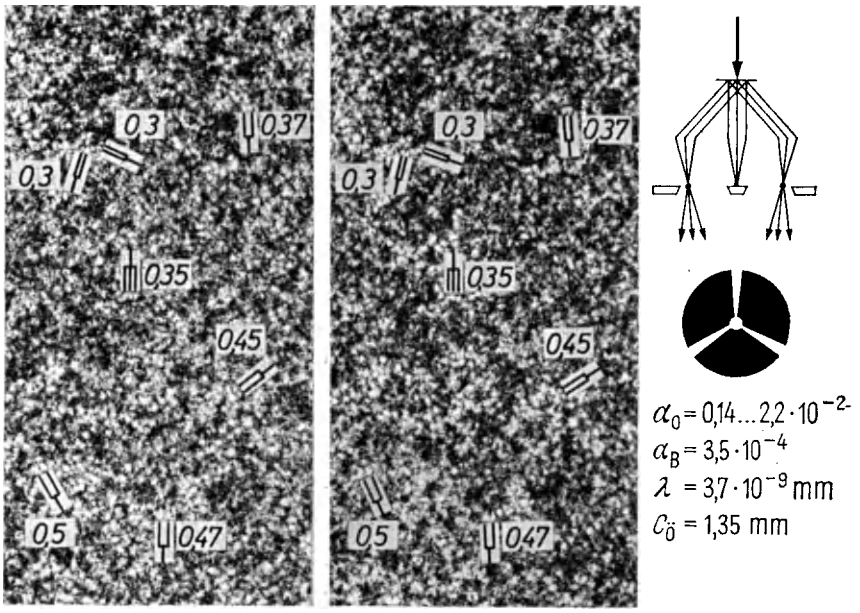


Fig. 23. – Dark field images of a carbon foil; resolution test.

+ 1st and the — 1st diffraction orders. In the latter case the image structure would show half-spacings, which would be most unfavourable for interpretation of the images.

Half-spacings, however, should not occur, if in addition to the zero beam one half of the aperture plane is also intercepted by a special aperture of the type shown in Fig. 24 on the extreme right. Figure 24 allows a comparison of the two experiments under discussion: In case of the semicircular dark field aperture the transfer limit is lower compared to the two-sided case. This seems to indicate that half-spacings really have been eliminated. Obviously we do have to take interferences between several diffracted waves into account in those cases, where the strong zero order does not any more control the contrast. As a result of one-sided interference we cannot exclude the possibility that even differential frequencies contribute to the image structure under the given coherent illuminating conditions.

One should therefore be very careful in interpreting high resolution dark field images, when they have been taken with a coherent illumination of the object. This is supported by results of Hanszen ([14]) who made light optical experiments analogous to ours.

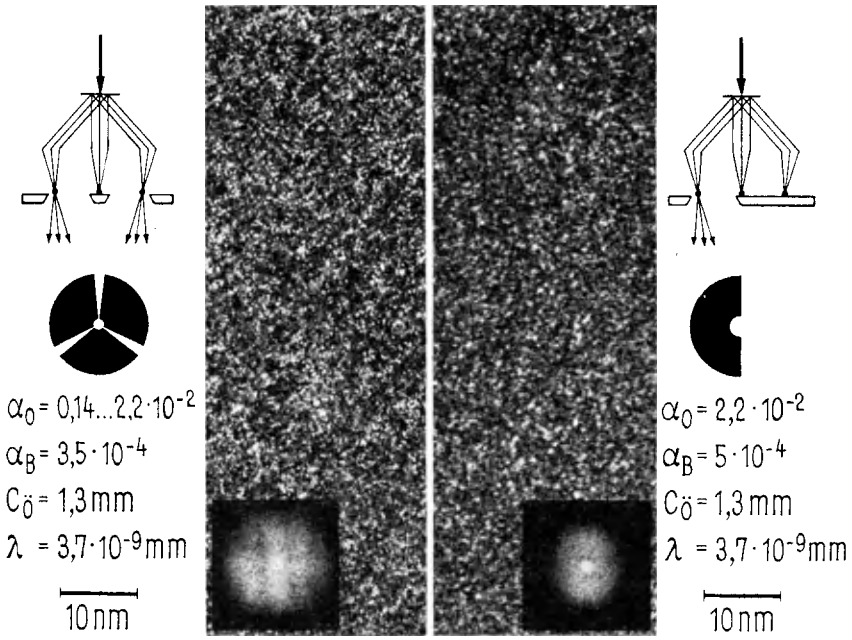


Fig. 24. – Two different types of dark field imaging.

Dark field microscopy in the high resolution field has therefore still to be investigated further. For this reason there is again enhanced interest in methods which make use of the phase contrast mechanism, but avoid the disadvantages of circular apertures or even of zone plates. A method, which theoretically fulfilles these requirements, will be discussed in the next Section.

REFERENCES (Section 2)

1) W. HOPPE: *Naturwiss.*, **48**, 736 (1961).
2) R. LANGER and W. HOPPE: *Optik*, **24**, 470 (1966/67).
3) K.-H. VON GROTE, G. MOELLENSTEDT and R. SPEIDEL: *Optik*, **22**, 252 (1965).
4) W. HOPPE, K.-H. KATERBAU, R. LANGER, G. MOELLENSTEDT, R. SPEIDEL and F. THON: *Siemens Review*, **36**, 24 (1969) (3rd Special Issue).
5) G. MOELLENSTEDT, R. SPEIDEL, W. HOPPE, R. LANGER, K.-H. KATERBAU and F. THON: *Proc. 4th Eur. Conf. on Electron Microscopy, Rome 1968* (Rome, 1968), vol. **1**, p. 125.
6) K.-J. HANSZEN and B. MORGENSTERN: *Zeits. Angew. Physik*, **19**, 215 (1965).

7) F. THON: *Proc. 4th Eur. Reg. Conf. on Electron Microscopy, Rome 1968* (Rome, 1968), vol. **1**, p. 127.
8) F. THON and D. WILLASCH: *Proc. 7th Int. Conf. on Electron Microscopy, Grenoble 1970* (Paris, 1970), vol. **1**, p. 3.
9) R. LANGER, W. HOPPE and F. THON: *Optik*, **30**, 538 (1970).
10) K.-J. HANSZEN: *Zeits. Naturfor.*, **24***a*, 1849 (1969).
11) K.-J. HANSZEN: *Proc. 7th Int. Conf. on Electron Microscopy, Grenoble 1970* (Paris, 1970), vol. **1**, p. 21.
12) H. BOERSCH: *Ann. Phys.*, **27**, 75 (1936).
13) G. DUPOUY: *Journ. Microscopie*, **5**, 655 (1966).
14) K.-J. HANSZEN: *Zeits. angew. Phys.*, **27**, 125 (1969).

3. Prospects of high resolution microscopy using phase plates.

3·1. General aspects.

From the previous Section we know that even by applying zone correction plates, semicircular apertures etc., the contrast transfer conditions of the objective lens can only be improved to a certain degree. Also these techniques do not provide that all spatial frequencies contained in the object are transferred to the image with maximum contrast simultaneously, or they cause lateral phase shifts. Thus, they do not really provide optimum conditions, *i.e.* a phase contrast transfer function which is constant in the frequency range from 0 up to the highest relevant frequencies.

The purpose of this Section is to discuss, whether better conditions could be achieved by not only using the phase shifting effect of the objective lens itself and intercepting parts of the electron waves, but by introducing additional phase shifts.

In light optics an ideal phase contrast imaging is possible by inserting a phase shifting plate of constant thickness into the imaging system according to the Zernike method on condition that the object is in focus and the objective lens spherically corrected.

In electron microscopy phase shifting can also be attained. According to a proposal of Boersch [1] electrostatic fields or material foils with an inner potential should be suited to produce phase contrast. Corresponding experiments were done by Kanaya *et al.* [2] and by Fert *et al.* [3]. We will discuss their investigations later.

A decisive complication in electron phase contrast microscopy, especially in the high resolution field, arises from the wave aberrations of the objective

lens which are caused by the unavoidable spherical aberration and by defocusing, if any. In consequence the profile of an ideal phase shifting foil to be inserted in the back focal plane of such an objective lens has to be quite complicated in order to achieve the desired correction for each point of the exit pupil, *i.e.* for all spatial frequencies of the object. In principle, however, a foil of suitable material with suitable variation of thickness should be able to establish the required conditions as Hanszen ([4]) pointed out in 1965.

If phase contrast electron microscopy is to be realized by phase shifting plates of this type, two questions are mainly to be discussed:

1) Is it possible to prepare such complicated plates?

2) How far will the scattering of electrons within such a plate increase the background intensity and diminish the image contrast of small object details?

3·2. Preparation of phase plates.

Let us start with a discussion of the first question: Fig. 25 shows at the top a sectional drawing of a carbon phase plate calculated for a special objective lens of the Siemens Elmiskop 101. The accelerating voltage is 100 kV, the nominal defocus value chosen is + 283 nm and the calculated resolu-

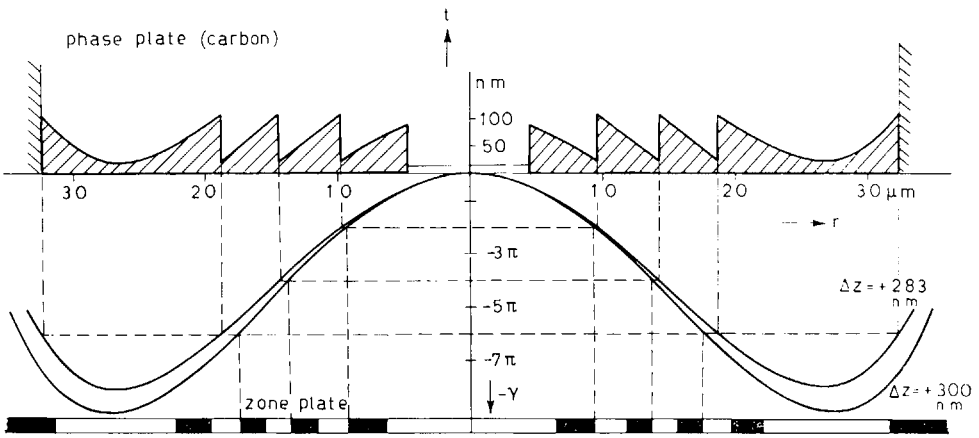


Fig. 25. – Profile of a phase plate thickness t in dependence of radius r in the back focal plane of the objective lens (top), phase shift γ by spherical aberration and defocusing for two different defocus values (centre) and scheme of a zone correction plate (bottom).

tion limit for point-to-point separation is 1.6 Å. Notice that in the vertical direction a 1000 times enlarged scale is used compared to the horizontal axis; thus the radial variation of thickness is exaggerated. In fact, the diameter of the plate is 60 μm, the central hole is about 7 μm in diameter, and the thickness varies between about 200 Å and 1200 Å.

Below the plate profile in Fig. 25 is plotted the phase difference γ between the diffracted waves and the undiffracted beam in dependence of the radius r in the back focal plane, this phase shift being introduced by spherical aberration of the objective lens and a defocus value of $\Delta z = +283$ nm (upper curve) and $\Delta z = +300$ nm (lower curve), respectively. In order to attain the desired additional phase shift for maximum phase contrast at each point of the exit pupil, the thickness of the plate has to vary in the manner shown above. If possible, the thickness should be kept small because of the inherent scattering. That is why drops of phase at values of 2π are used, thus building up the plate stepwise. A central hole has been chosen to avoid an interaction of the very intense undiffracted beam and the plate, which certainly would produce undesired effects caused by electrostatic charging, etc. Below the plotted phase shift γ a Hoppe-type zone correction plate for a calculated

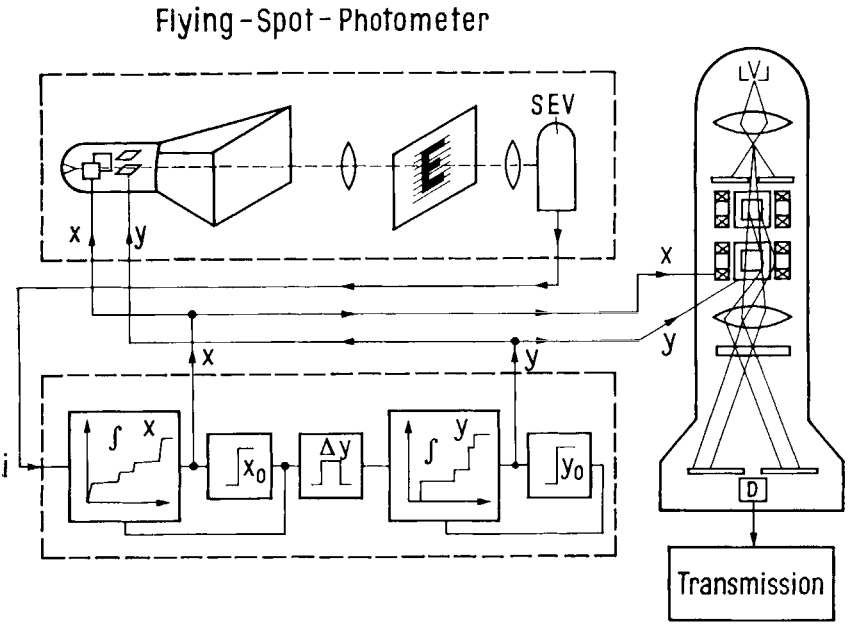


Fig. 26. – Principle of a speed-controlled electron optical microrecorder.

defocus value $\Delta z = +300$ nm is schematically drawn with the same lateral scale. The analogy is quite clear.

With a high degree of probability it will soon be possible to prepare such complicated plates with sufficient accuracy. Mueller ([5]) described how to use a Siemens Elmiskop 101 as a speed-controlled microrecorder. As a writing material he used the contamination layer which grows upon a specimen which is irradiated by electrons in a normal residual gas atmosphere. Figure 26 demonstrates the idea in principle: the transparency value of a pattern is transformed into an electrical signal by a flying-spot-photometer. Here capital *E* is used as an example. A steering-logic adjusts the sweep velocity of the scanning beam proportional to the optical density at any point of the pattern. Synchronized with this is an electromagnetic deflection system above the objective lens of the electron microscope. The deflection system controls the position and velocity of a very minute electron probe scanning on a carbon foil which is arranged in the back focal plane of the lens by means of a special specimen cartridge. As the probe has a diameter of less than 100 Å, contamination traces grow up very rapidly on top and bottom of the irradiated parts of the foil. The thickness of the contamination layer at each point of the foil depends on the local velocity of the scanning beam, on the diameter of the probe and on the partial pressure of the hydrocarbon molecules. The diameter of the probe and the distance of scanning lines

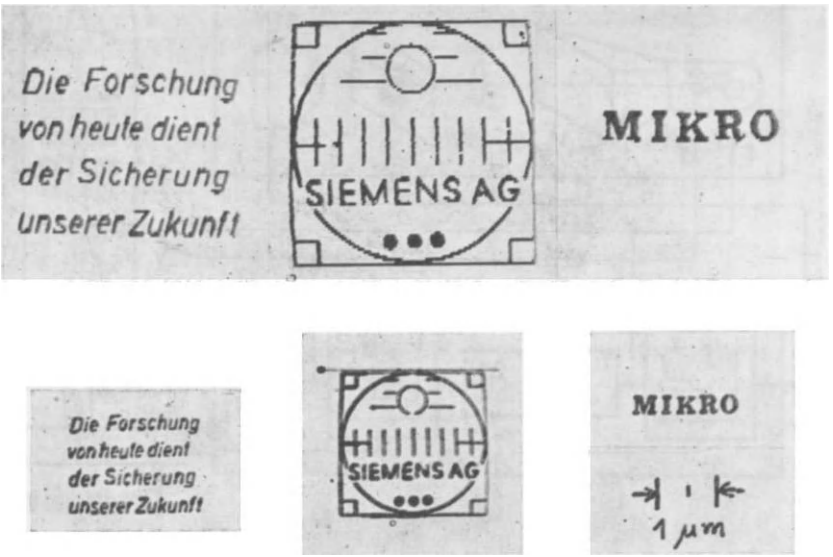


Fig. 27. – Some examples of microwriting.

can be controlled. A semiconductor detector monitors the thickness of the contamination layer continuously by measuring the electron density variation. In this way the growing of the contamination layer can be supervised.

Some applications of such an electron optical microrecorder are shown in Fig. 27. The recorded examples including the magnification scale were written on a carbon film, using a line distance of 500 Å (top) and 280 Å (bottom).

Fig. 28. – Half-tone picture produced by an electron optical microrecorder. The size of the micro-picture was $(5 \times 5)\ \mu m^2$.

To prepare phase plates, the microrecorder must be capable of producing continuously varying thickness distributions. This condition is met by the speed controlled type. Figure 28 shows as an example the first half-tone picture ever produced with an electron optical microrecorder. The size of the record is $5 \times 5\ \mu m^2$. The line distance was 300 Å, the half-width of the probe about 50 Å.

For the preparation of specimens with rotational symmetry it is convenient to control the writing process using polar co-ordinates. Such a steering-

logic is under construction. It is only a matter of time until phase plates as described will be available for first experiments.

It may be mentioned that the use of phase plates in the microscope demands special care to preserve the profile of the plates from contamination. The vacuum conditions have to be so good that an increase in thickness is negligible.

3·3. The self-scattering of phase plates.

The second question in discussion concerns the effect of electron scattering caused by a phase plate on the contrast of fine image details. There are actually no experimental results known with respect to this problem. Kanaya *et al.* ([2]) and Fert *et al.* ([3]) did some experiments concerning phase contrast microscopy about ten years ago. Kanaya inserted a perforated film into the back focal plane of the objective and let the zero beam pass a central hole of about 10 μm in diameter. Thus he prevented the image from being destroyed by electrostatic charging.

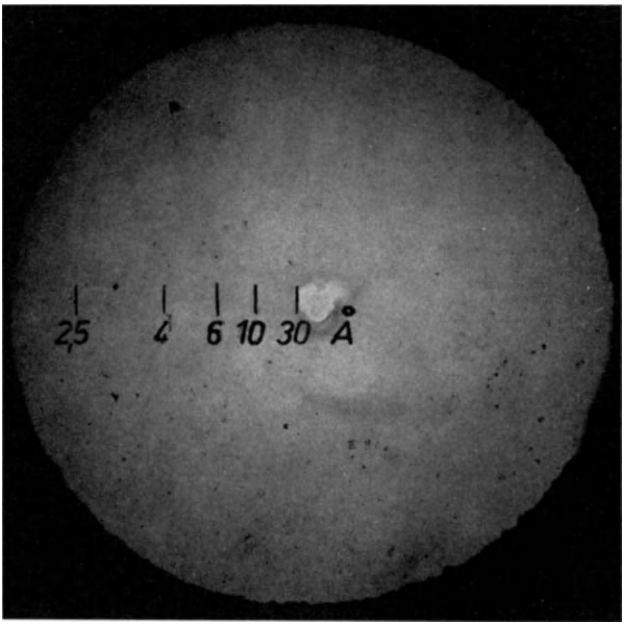


Fig. 29. – Electron micrograph of a phase plate of uniform thickness with a centre hole. The marks indicate the radii at which the diffracted orders corresponding to the noted reciprocal spatial frequencies pass the aperture plane.

With a central hole of 10 μm in diameter, however, only diffracted waves corresponding to object details smaller than about 10 Å could be affected by the phase shifting foil. But such a resolution could normally not be achieved in electron microscopy at that time. Therefore these experiments cannot give an answer in respect to the application of phase plates in the high resolution field.

For this reason some basic experiments were done by the author ([6]): a 60 μm hole in a metal foil aperture was covered with an evaporated carbon film. The thickness was determined by measuring the scattering properties, it was about 1400 Å. A central hole was burnt in by electron irradiation in presence of oxygen. An electron image of such a plate is shown in Fig. 29. The average diameter of the central hole is 3.2 μm in this case. This means, all ± 1st diffraction orders corresponding to reciprocal spatial frequencies smaller than 30 Å are shifted in phase by this plate. In Fig. 29 for some reciprocal spatial frequencies the distances from the centre are marked, where the wave vectors of the diffracted waves hit the back focal plane.

This plate was used in imaging thin carbon films under high resolution conditions. It is clear that these experiments with plates of even thickness

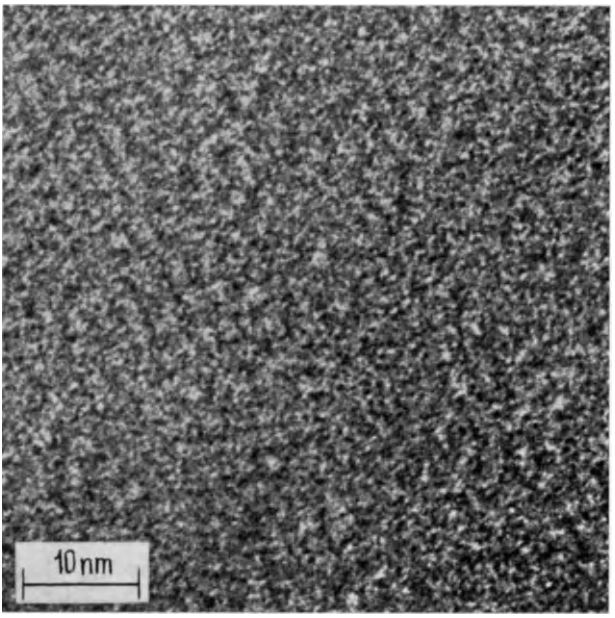


Fig. 30. – Image structures of a carbon film. The micrograph was taken using the phase plate of Fig. 29.

are only to answer the question whether the contrast of fine details is considerably decreased or even totally destroyed when diffracted amplitudes go through phase shifting material of such a thickness.

After some experimental difficulties, especially caused by charging, it was possible to attain image structures including the 3 Å region with good contrast. An example is shown in Fig. 30. Figure 31 shows the corresponding light optical diffraction pattern. The attenuation in diffraction intensity which is indicated for the innermost maximum is due to the fact that the phase shift close to the edge of the central hole was especially strong since a contamination layer had been grown there during the oxidation of the centre area. Figure 31 proves that the spatial frequencies of the object are transferred to the image

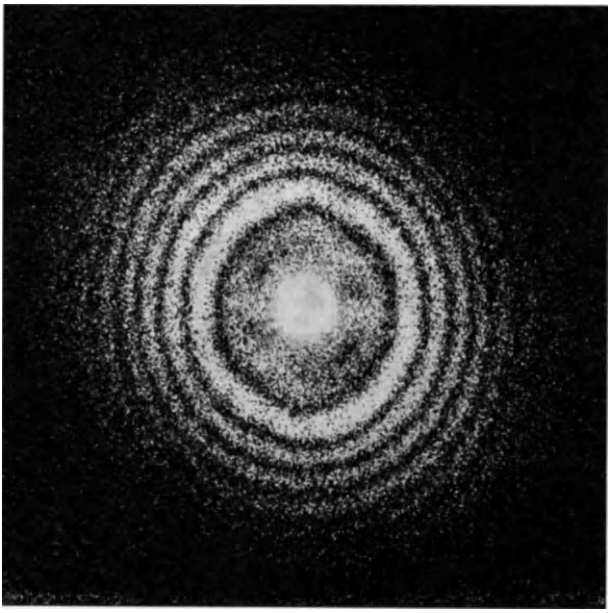


Fig. 31. – Light optical diffraction pattern of the image structure shown in Fig. 30.

in the usual manner. Only the combination of frequency bands is different from that which would normally occur at the chosen defocus value as a consequence of the additional phase shift introduced by the plate.

In addition, the contrast of 5 Å details in micrographs was measured. There was almost no difference in contrast of image structures taken with a phase plate of the type described or taken with an open circular aperture.

The measured intensity modulation was in both cases about 25% compared to the background intensity.

In the meantime, Badde and Reimer ([7]) calculated the influence of a scattering phase plate on the electron image. The authors determined the decrease of the coherent part of the electron beam intensity which is expected to be caused by a phase plate with a profile like that shown in Fig. 25. Their results can be summarized as follows: In the case of an objective lens with a coefficient of spherical aberration $C_{\ddot{o}} = 1.3$ mm, using 100 kV electrons, the contrast attainable for a Pt-atom (C-atom) should be 15.5% (3.1%) without any phase plate under optimum defocus conditions. Having an ideal phase plate without any inherent scattering, the contrast should amount to 48% (11%). And finally, using a real phase plate as described above, 32% contrast should be attainable in the case of a Pt-atom, and 7% in case of C-atom. This is about twice as much as under optimum conditions without using a phase plate. These results are very encouraging. A comparison between these calculations and experimental results will be possible in the near future ([8]).

3·4. Interferometry with the electron microscope.

Here a first result of another possible application method of phase shifting in the back focal plane of an electron objective lens will be discussed. This method also has imaging of carbon foils and subsequent light optical diffraction at the image structures as its basis.

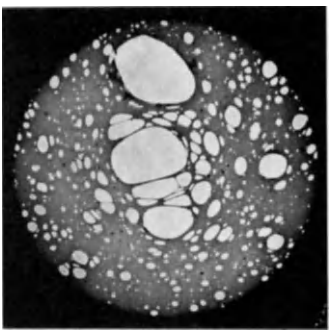


Fig. 32. – Electron image of a perforated foil, which was used for a basic experiment in interferometry.

According to a proposal of Hoppe ([9]) interferometry using the electron microscope should be possible in the following way. The material to be examined is mounted on an aperture in such a way that only a sector of the open area is covered. Thus, in imaging a carbon film, only in a certain azimuthal range an additional phase shift γ due to this material occurs. This gives rise to an azimuthal dependency in the frequency spectrum of the image structures which can be used to determine the additional phase shift and the inner potential U_i of the material.

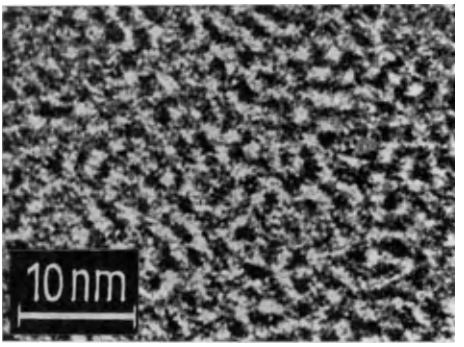


Fig. 33. – Image structures of a carbon foil. The micrograph was taken using the perforated film of Fig. 32 as objective aperture.

A basic experiment for testing the method was done using a perforated foil which is shown in Fig. 32. In a vertical direction a certain area is uncovered at least on one side. When inserting this foil into the back focal plane the image structures shown in Fig. 33 result. The light optical diffraction pattern of the image structures (Fig. 34) reveals a corresponding alteration of the frequency spectrum in the vertical direction. From this the inner potential U_i of the foil can be calculated, if the thickness t is known:

$$U_i = \frac{\gamma \cdot \lambda \cdot U_0}{\pi \cdot t},$$

where λ is the electron wavelength and U_0 the accelerating voltage. If the inner potential is known then the thickness t of the foil can be determined. A rough estimation shows that these quantities may be determined with a tolerance not larger than $\pm 5\%$. This is comparable in its accuracy with other, much more complicated methods.

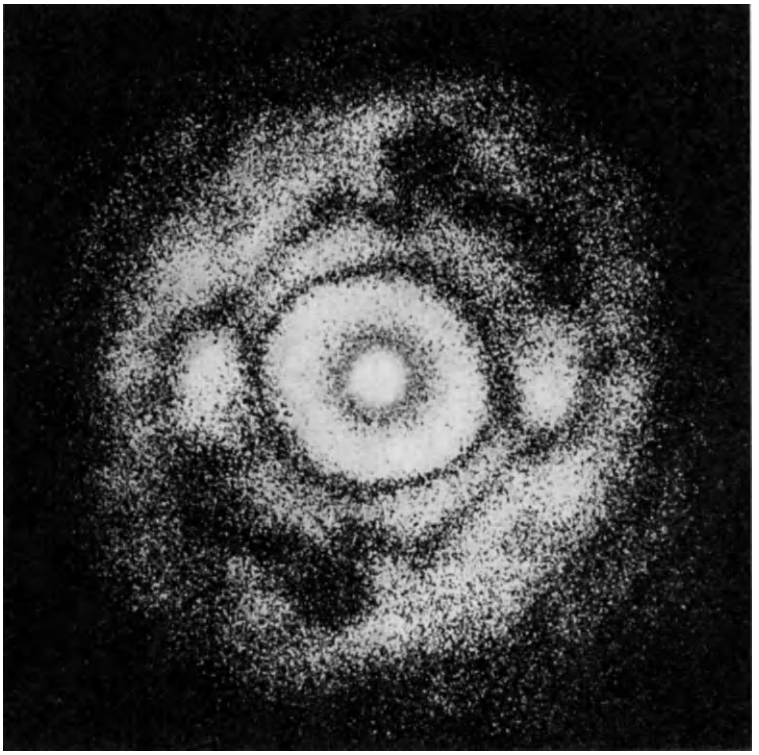


Fig. 34. – Light optical diffraction pattern of the image structures shown in Fig. 33.

REFERENCES (Section 3)

1) H. BOERSCH: *Zeits. Naturfor.*, **2***a*, 615 (1947).
2) K. KANAYA and H. KAWAKATSU: *Proc. 4th Int. Conf. on Electron Microscopy, Berlin 1958* (Springer, Berlin, 1960), vol. **2**, p. 308.
3) J. FAGET, M. FAGOT and C. FERT: *Proc. 2nd Eur. Reg. Conf. on Electron Microscopy, Delft 1960* (Delft, 1961), vol. **1**, p. 18.
4) K.-J. HANSZEN and B. MORGENSTERN: *Zeits. Angew. Physik*, **19**, 215 (1965).
5) K.-H. MUELLER: *Dissertation*, Tübingen (1970); and *Proc. 7th Int. Conf. on Electron Microscopy, Grenoble 1970* (Paris, 1970), vol. **1**, p. 183.
6) F. THON: Paper at the *Meeting of the German Society of Electron Microscopy*, Wien (Sept. 1969).
7) H. G. BADDE and L. REIMER: *Zeits. Naturfor.*, **25***a*, 760 (1970).
8) F. THON and D. WILLASCH: *Proc. 7th Int. Conf. on Electron Microscopy, Grenoble 1970* (Paris, 1970), vol. **1**, p. 31.
9) W. HOPPE: private communication.

4. Spatial filtering in optical reconstruction of high resolution phase contrast images.

4·1. Introduction.

The main aim of the investigations described in this Section is the improvement of phase contrast information in high resolution electron images by application of coherent spatial filtering techniques in optical reconstruction.

Similar techniques were first proposed and used by Maréchal and Croce [1] in light optics to remove undesired defects in photographic images due to incorrect imaging.

Optical filtering of electron micrographs has been very successfully applied to the investigation of biological particles with spatially extended periodicity by Klug and Berger [2] and by Klug and De Rosier [3]. In this case an improvement in interpreting image details of stained specimens has been achieved by eliminating phase contrast and admitting scattering absorption to the reconstructed image only. In addition to this the image is reconstructed by using only the relatively strong diffraction orders caused by periodic image details. So it is even possible to extract the image of one side of a particle from a two-sided image. However the staining does not allow a resolution better than about 20 Å. In order to get a higher resolution it seems to be necessary to avoid staining and to make use of phase contrast information.

Further investigations with optical reconstruction devices have been published by Hanszen [4]. He compared the transfer function of a defocused optical imaging system with spherical aberration with the imaging properties of an electron objective lens. The correspondence between the two systems may also be used to get a better quality of electron micrographs by reconstruction.

Here recent experiments of Thon and Siegel [5] are to be discussed. The aim of these investigations is to improve the information obtained from high resolution phase contrast electron micrographs. The basic procedure consists of zonal filtering with a Hoppe-type zone correction plate [6] and not applied within the electron microscope as realized by Möllenstedt *et al.* [7] but in a subsequent optical reconstruction step.

4·2. Theoretical considerations.

We assume the isoplanacy condition to be satisfied and the electron microscopic specimen to be a weak phase object. Thus the intensity distribution $d(x, y)$ in the image plane of a defocused objective lens with spherical aberration may be written

$$d(x, y) = F(\varphi \cdot K) + 1 , \tag{22}$$

where F stands for Fourier transformation, φ is the scattering amplitude of the object and K is the contrast transfer function

$$K = 2B_M \cdot \sin W . \tag{23}$$

Here the function B_M accounts for apertures and filters in the back focal plane of the electron objective lens. W is the aberration function for the case of defocusing and spherical aberration.

The amplitude distribution $D(u, v)$ in the back focal plane of the aberration-free diffraction lens in a light optical reconstruction apparatus can be written

$$D(u, v) = F(d) = 2\varphi \cdot B_M \cdot B_L \cdot \sin W + \delta , \tag{24}$$

where δ is Dirac's delta-function and B_L takes account of a filter in the diffraction plane. As B_M and B_L can be exchanged in this equation, the application of apertures and filters within the electron microscope and within the light optical reconstruction apparatus is equivalent under the conditions noted above.

4·3. Arrangements.

The above mentioned method of zonal correction is intended to allow the passage of only equiphase waves and to obstruct the passage of those waves of opposite phase position ([6]). The use of zone correction plates in the microscope has proved to be extremely difficult due to the minute dimensions of the plates and the most critical demands regarding lateral alignment, compensation of axial astigmatism and accuracy of focusing. According to the theory the same correction effect should be achieved by taking the elec-

tron micrograph using a circular aperture and subsequent zonal filtering of the image structure using an optical diffractometer.

If the reconstruction arrangement is properly dimensioned and no serious imaging defects occur, zonal filtering should be much easier to handle this way than in the electron microscope. But, in fact, several difficulties can arise, since even at highest magnifications the interesting details in high resolution electron micrographs are comparable in size with the granulation of the photographic plates and the contrast of the finest details is quite weak. Special care has to be employed in selecting the lenses for the reconstruction device as even small aberrations can falsify the image structure completely.

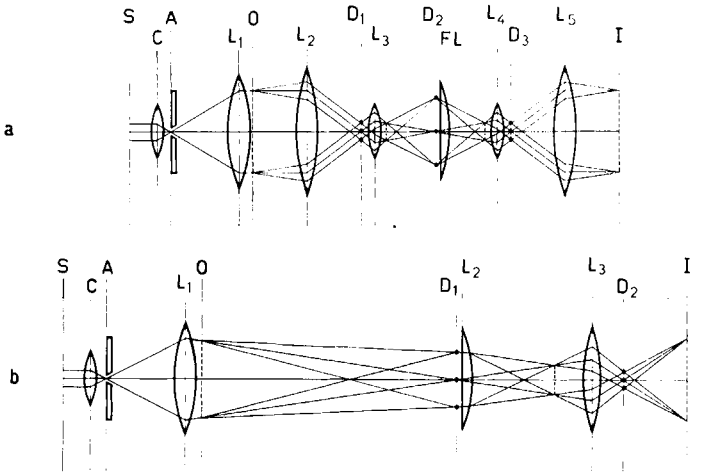


Fig. 35. – Two arrangements for optical filtering, schematically.

Figure 35 *a* and *b* shows schematically two possible arrangements for filtering techniques. The upper one (*a*) allows a very short length of the apparatus: the coherent light coming from the effective source A is collimated by lens L_1. Thus a parallel beam illuminates the object O, *i.e.* the photographic plate with the electron micrograph. Its primary Fraunhofer diffraction pattern in the back focal plane D_1 of lens L_2 is enlarged by lens L_3. The plane D_2 of the enlarged diffraction pattern is the plane of symmetry of the whole optical arrangement. Here the filters are inserted. The amplitude distribution behind the filter is transformed by the field lens FL and the lenses L_4 and L_5 into the image distribution I. This arrangement fulfills all requirements for the application of exact Fourier transformation to the imaging process.

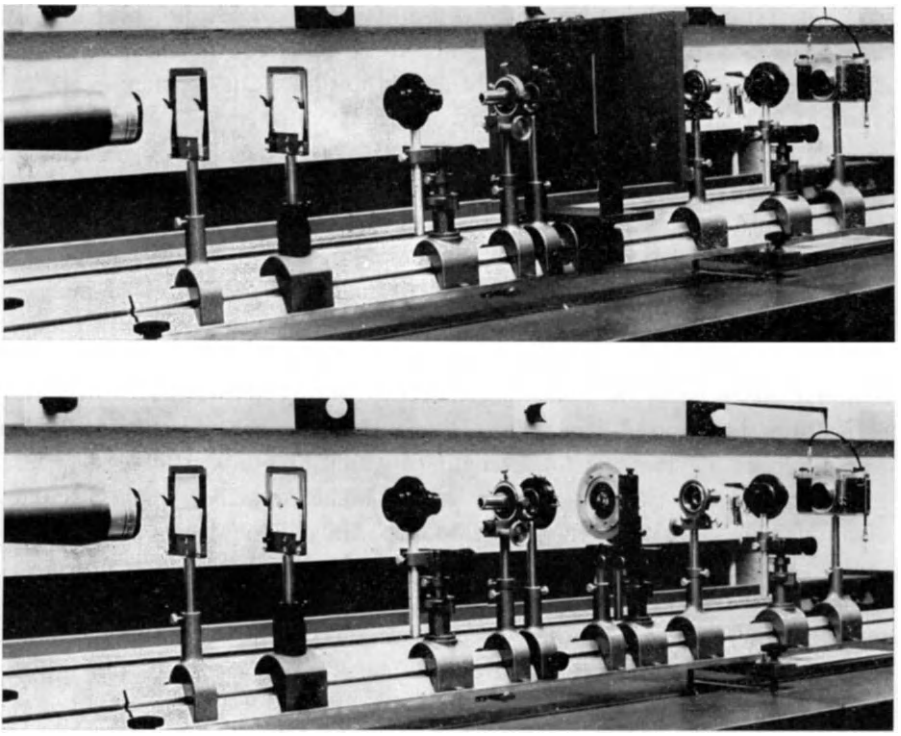


Fig. 36. – Device for optical filtering. Top: registration of the Fourier-transform. Bottom: reconstruction set-up.

Figure 36 shows an arrangement according to the principle described above, which was partly used for the experiments. The upper photograph shows how the diffraction pattern is registered, the lower one shows the actual reconstruction apparatus. As a light source a He-Ne laser of 15 mW power was used in this case. Other experiments were performed using 5 or 8 mW lasers. In Fig. 36 only an attachment to the laser is visible on the left-hand side. This attachment combines an expanding lens, a spatial filter and a collimating lens. The next carrier holds a neutral density filter for attenuation of the very intense beam during the adjustment of the set-up. Another carrier takes the diffracting object, *i.e.* the electron micrograph. It is arranged exactly in the front focal plane of the diffraction lens (camera lens $f = 135$ mm). A microscope objective ($10\times$, $f = 18$ mm) enlarges the Fourier transform and projects it on the registration plane of a Polaroid camera body, in front of

which is a shutter. It has been proved useful to take a Polaroid sheet film for registering the diffraction pattern, since it is quite convenient also in our case to cut the filter masks directly from the film. The diffraction pattern is about 30 mm in diameter, and its outermost diffraction maximum corresponds to 3 Å details at a magnification of 250 000 in the electron microscope.

The lower photograph in Fig. 36 demonstrates the reconstruction set-up. The Polaroid back was replaced by a mask-holder with a rotatable cross-table and a field lens. After this the lenses are symmetrically arranged to the filtering plane as mentioned above. Finally the reconstructed image is photographed by a 35 mm camera body. The lengt h of the diffractometer from the object plane to the registration plane of the first Fourier transform is about 50 cm, thus the total length from the object to the reconstruction plane does not exceed 1 m. This is remarkably short compared to other reconstruction devices the total length of which amounts to 4 or 5 meters.

In general one can say there are much higher demands in the accuracy of adjustment with this reconstruction apparatus than with a mere diffractometer. Lens aberrations, especially those of the microscope objectives, have a very critical influence on the quality of the reconstructed images. It is absolutely necessary to test the performance of the device very carefully with a suitably chosen test object before starting real reconstruction experiments.

A more simple arrangement, which was used for another part of our experiments, is schematically shown in Fig. 35*b*. Here a weakly converging illumination of the object was chosen, so that only three instead of six lenses were needed. This arrangement also meets the requirements for Fraunhofer diffraction but it has the disadvantage of a larger total length of about 2.5 m.

4·4. Experiments.

First the arrangement was tested without any filtering. Figure 37 demonstrates the quality of the reconstructed images of four micrographs of a carbon foil taken at different defocus values of the electron microscope objective lens. On the left hand side sections of the original electron images are shown. All sections show the same area of the specimen. The centre row gives the corresponding light optical diffraction patterns of these image structures. On the right hand side the reconstructions of the original micrographs are shown. Each detail of the original plates is accurately reconstructed which proves the reliability of the arrangement. The images even have a better

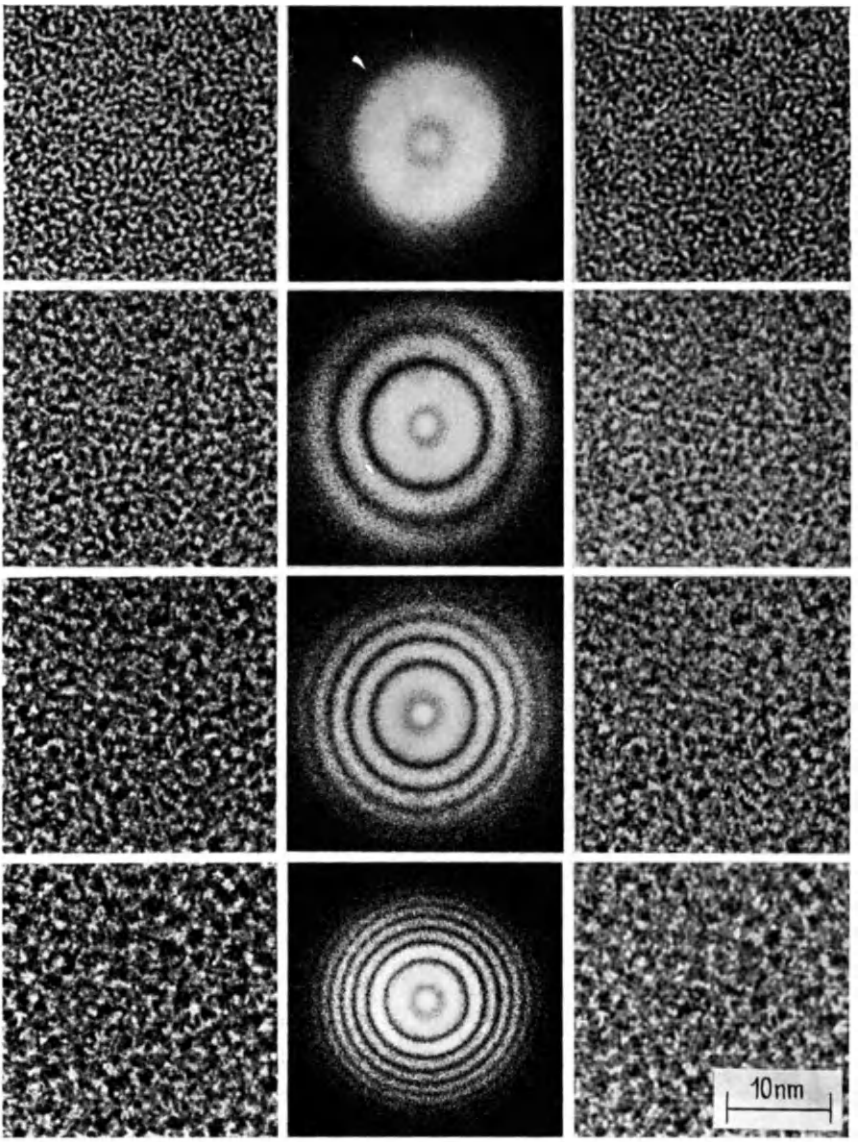


Fig. 37. – Originals, Fourier-transforms and reconstructions (right) of four different defocused micrographs of identical sections of a carbon foil.

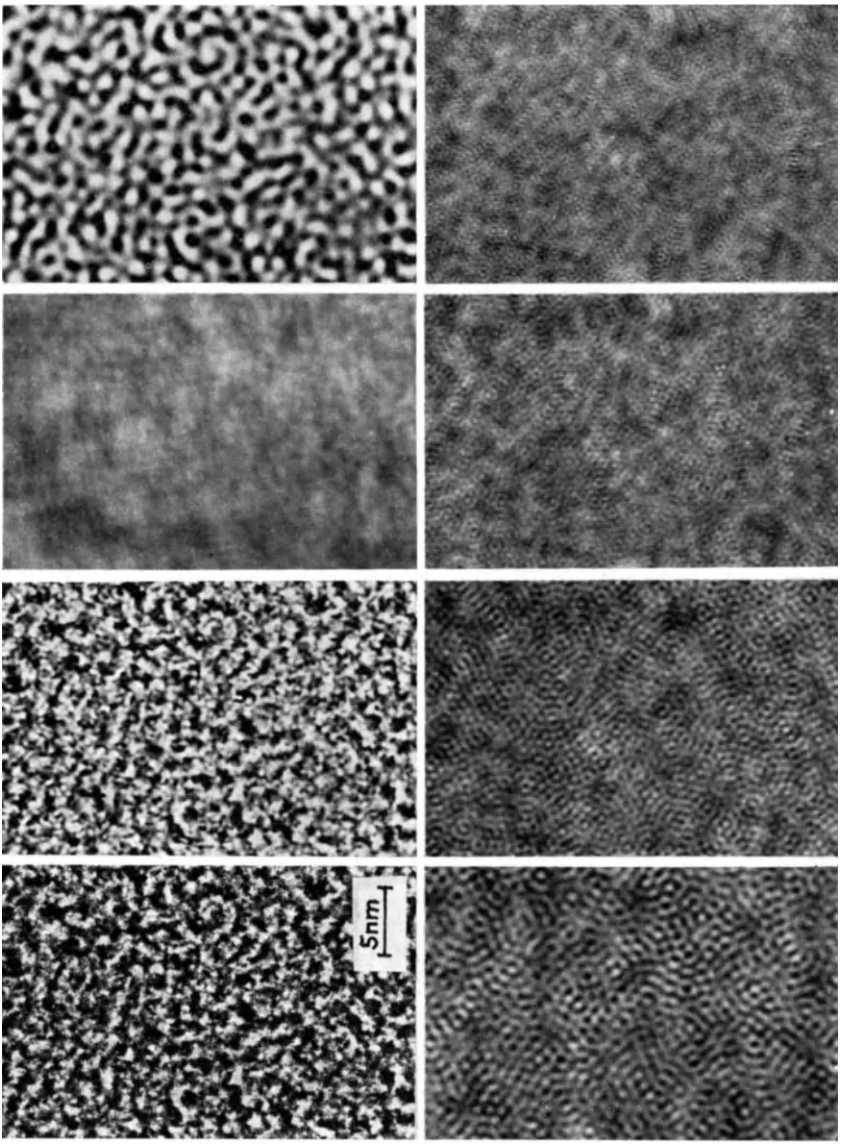


Fig. 38. – Decomposition of an image structure (third micrograph from the top in Fig. 37).

quality than the original micrographs since the noise of the photographic emulsion has been eliminated due to the limited size of the apertures used.

Figure 38 shows, as another experiment, the decomposition of the image structure of the third micrograph from top in Fig. 37. Starting with the original micrograph the upper row continues with the reconstructed image admitting all spatial frequency bands contained in the electron image. The next image is a result of the zero order alone with slight disturbances as a consequence of experimental difficulties. On the right hand side only the zero and first frequency bands contribute to the reconstruction. In the lower row the second, third, fourth and fifth frequency bands, consecutively each combined with the zero order, are contained in the reconstructed image. It is clearly seen that the image structure is determined by the selected frequency band, and the details are uniform in size. This proves that linear modulation transfer is guaranteed.

Figure 39 illustrates a practical attempt at a reconstruction. *a*) and *b*) are micrographs of the same area of a carbon foil taken with widely different defocus values. *c*) and *d*) show the corresponding diffraction patterns. The two reconstructions, *e*) and *f*), show identical image structures since the same

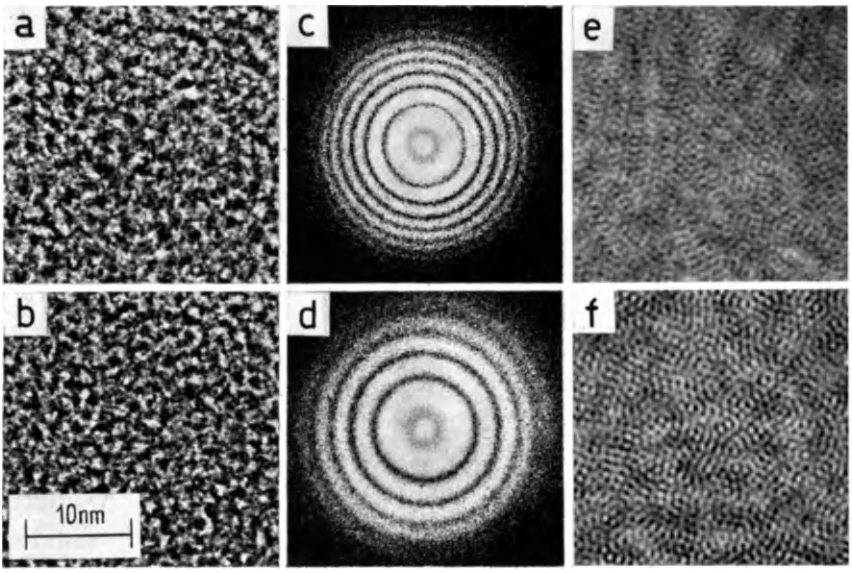


Fig. 39. – Reconstruction using the same frequency range of two different defocused micrographs of a carbon foil. Left: originals. Centre: Fourier-transforms. Right: reconstructions.

range of frequencies were used in the two reconstructions. The range used besides the zero order corresponded to passing the frequencies of the third-frequency band in *c*), or part of the second-frequency band in *d*). This procedure is a first step towards the total reconstruction of an image from a number of different defocused micrographs.

4·5. **Zonal filtering.**

The experiments described above were to demonstrate the capability of the reconstruction arrangement in principle. The next step implies zonal filtering. But first some comments on the experimental skill that has to be applied for zonal filtering within the electron microscope on one hand and in the reconstruction apparatus on the other hand. Figure 40 shows a com-

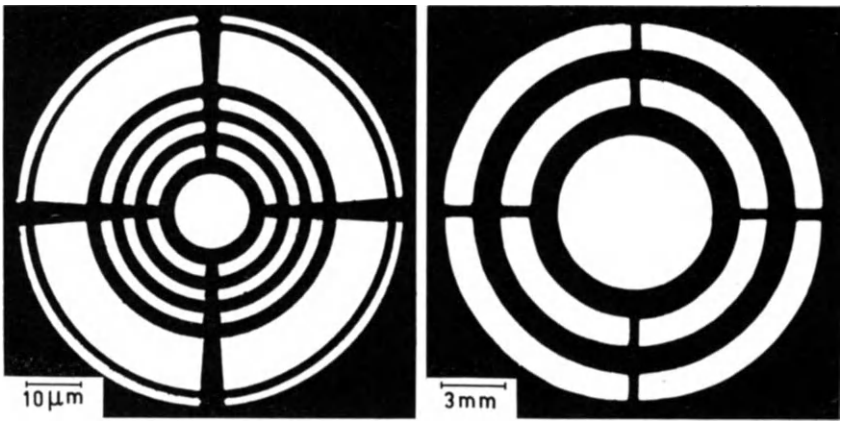


Fig. 40. – Geometrical dimensions of a zone correction plate for use within the electron microscope (left) and one used in optical reconstruction (right). The latter one is 500× larger than the one used in the microscope.

parison between a zone correction plate for use within the electron microscope (left) and one used in optical reconstruction especially in respect to their geometrical dimensions. Unlike the most tricky preparation of the minute plates to be used in the electron microscope ([8]) the light optical plate can easily be cut from the photograph of the diffraction pattern in a few minutes, since it is about 500× larger than the miniature one. The micro-

scopic zone plate has to be calculated for a special defocus value; this defocus value has to be met within a tolerance of 40 Å. In addition, the axial astigmatism of the objective lens has to be compensated almost completely. Otherwise the zone plate would be effective only in a certain azimuthal direction due to its rotational symmetry. The macroscopic plate has not to be calculated, the mask can be prepared from the diffraction pattern of an arbitrary micrograph of any given defocus value. Within the electron microscope the zone plate must be aligned within an axial tolerance of 100 μm and a lateral tolerance of 0.5 μm to the optical axis. Using the reconstruction arrangement the mask can be adjusted within a few seconds by observing the diffraction pattern. Of course also other kinds of filters, for instance phase shifting masks could be inserted into the system.

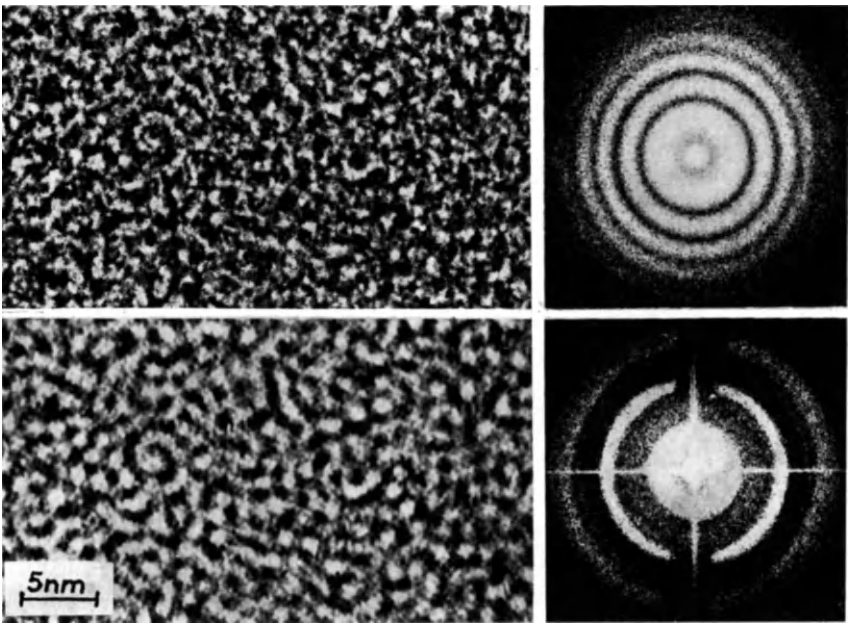


Fig. 41. – Top: electron micrograph of a carbon foil taken with a circular aperture and its Fourier-transform. Bottom: reconstruction of the micrograph with zonal filtering and corresponding Fourier-transform.

The effectiveness of zonal filtering will be shown by some examples. Figure 41 shows in the upper row an original micrograph of a carbon foil, taken with a circular aperture, and its Fourier-transform. The bottom row

shows the reconstruction of this micrograph with zonal filtering having been applied. From the corresponding diffraction pattern of the filtered image structures, finally, it can easily be seen that certain frequency bands have been obstructed and that the zone correction mask was properly adjusted.

As an example for a micrograph taken with a zone corrected electron objective lens Fig. 19 can be taken. It demonstrates that the image structure is in the same typical way different from that taken with a circular aperture. The optical diffraction pattern is quite similar to the corresponding pattern in Fig. 41, too.

A comparison of micrographs showing the same object field, taken with a zone corrected objective lens on one hand and taken with a circular aperture with subsequent light optical zonal filtering on the other hand, would be desirable for further confirmation of the results. Because of the high experimental requirements to be met when taking the electron micrographs under high resolution conditions (zone plate and circular aperture have to be changed without disturbing the compensation of astigmatism to any degree) this step has not yet been realized.

Figure 42 finally shows a first application of zonal filtering to a biological object. A defocused image of a T_4-phage tail with superimposed phase structures (left) was reconstructed after zonal filtering (right). By comparing these two pictures there does not seem to be a great improvement by the zonal filtering applied. But we cannot expect to see much effect on the relatively

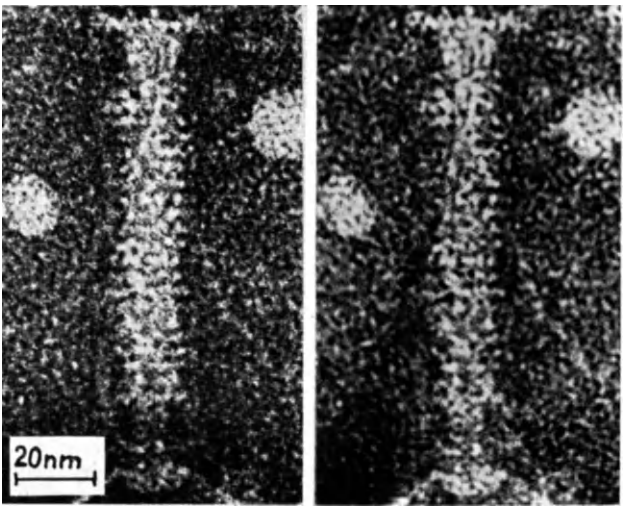


Fig. 42. – Micrograph of a T_4-phage-tail (left) and its reconstruction with zonal filtering.

coarse structures of this stained specimen. There is, however, a remarkable change in the behaviour of very fine details visible. Our experiments are still in progress, and at the moment we are looking for specimens which are more adequate to check the usefulness of this type of zonal filtering.

REFERENCES (Section 4)

1) A. MARÉCHAL and P. CROCE: *Compt. Rend. Acad. Sci.*, **237**, 607 (1953).
2) A. KLUG and J. T. BERGER: *Journ. Mol. Biol.*, **10**, 570 (1964).
3) A. KLUG and D. J. DE ROSIER: *Nature*, **212**, 29 (1966).
4) K.-J. HANSZEN: *Proc. 4th Eur. Reg. Conf. on Electron Microscopy, Rome 1968* (Rome, 1968), vol. **1**, p. 153.
5) F. THON and B. M. SIEGEL: *Proc. 7th Int. Conf. on Electron Microscopy, Grenoble 1970* (Paris, 1970), vol. **1**, p. 13. See also: *Berichte der Bunsengesellschaft für Physikalische Chemie*, **74**, 1116 (1970).
6) W. HOPPE: *Optik*, **20**, 599 (1963).
7) G. MOELLENSTEDT, R. SPEIDEL, W. HOPPE, R. LANGER, K.-H. KATERBAU and F. THON: *Proc. 4th Eur. Reg. Conf. on Electron Microscopy, Rome 1968* (Rome, 1968), vol. **1**, p. 125. See also: *Siemens Review*, **36**, 24 (1969) (3rd Special Issue).
8) K. H. v. GROTE, G. MOELLENSTEDT and R. SPEIDEL: *Optik*, **22**, 252 (1965).

Contrast Phenomena in Electron Images of Amorphous and Macromolecular Objects

A. C. VAN DORSTEN

Laboratory of Electron Microscopy, University of Amsterdam - Amsterdam, Holland

1.

1·1. Introduction.

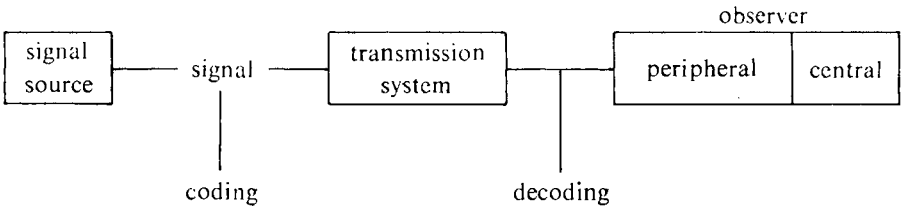
In the context of this lecture the word amorphous is used to describe electron microscope objects of noncrystalline nature, the structural parameters of which are in general nonperiodic or only statistically determined. This class of objects includes irregular objects as well as ordered structures, such as macromolecular particles or fibrous material.

1·2. Information theoretical aspects.

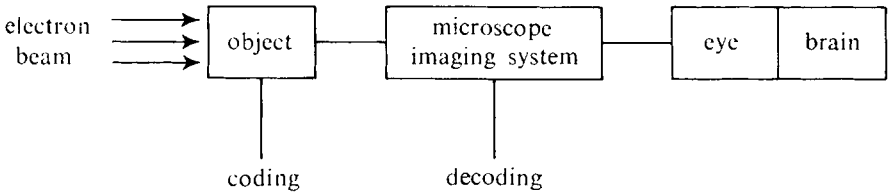
Although obtaining high resolution in electron micrographs is of considerable theoretical and practical importance, there are fields where contrast in extended areas, exceeding the theoretical resolution limit by a factor of 10 or even 20, appears to be decisive for the contribution the electron microscope can be expected to make for extending knowledge about structure. Macromolecular objects are a typical example.

Before entering into a discussion of the more specific aspects of the subject, a survey will be given of some very general aspects of microscopical observation, which is, in fact, a communication process. Communication, in this instance, can be defined as the collecting and handling of physically or otherwise detectable signals by an individual.

The different parts of a communication process form a communication chain. A typical chain is the observation chain, which can be represented as follows:



More specific for the electron microscope:



In the transmission microscope the object codes information onto the beam. The observer has some control over the coding process by adjusting the beam, *e.g.* by changing the degree of coherence or the accelerating voltage. The imaging system transmits and decodes the signals, converting it into a visual signal the perception of which can be related to the object. Part of the signal, however, cannot be related to the object and is regarded as noise, by analogy with the transmission of acoustical signals through electrical transmission systems first investigated basically. The noise can enter the chain at various points at which energy is supplied to the system, or where time-dependent irregular interaction occurs.

The signal is a carrier of information and is supplying itself the energy for its detection (« live » information, Brillouin ([1])). In electron microscopy, however, it is customary to store the information in micrographs as « dead » information, a scalar function of the two place co-ordinates defining the *structural information*, and the intensity (or density) for each point, defining the *metric information*. The output signal can thus be identified with the recorded image. Because of the finite resolving power of the microscope there is a quantization of the image co-ordinates, and thus of the structural information. This simply accounts for the well-known fact that the system acts as a low-pass filter for the spatial frequencies to be transmitted.

The image can thus be split up into a finite number of image elements, each corresponding to the size of the smallest object detail to be observed multiplied by the magnification.

The next consideration is to further specify the metric information by assigning a variable measure to each element in the form of a scale of distinguishable intensity steps. This proper scale is determined by a property of the observer or the detector he uses, or both. Because there is always an amount of uncertainty in every observation or measurement of a physical quantity, the measure can be only one within certain limits of probability.

The maximum number of gradation steps that can be distinguished by a human observer in a single observation is physiologically determined, and practically never exceeds ten. The actual number of recognizable levels of gradation, a property of the recorded picture, appears to be an important factor in the appreciation of the picture. For the normal case of a half-tone picture printed on white paper, it appears that the human observer has a typical preference for the presence in the image of about 5 perceivable brightness steps, irrespective of the size of detail or the nature of the picture. Images having a larger or a smaller number of steps are not considered good pictures, and are rejected as being too soft or too contrasty. They are apparently more difficult to use for the extraction of information than the preferred ones (Van Dorsten ([2])). The value 5 agrees very well with the measured channel capacity for the perception of brightness of the human observer (Eriksen and Hake ([3])).

The number of possible perceivable images is finite, but very large. If the picture consists of m elements, each having p possible states of brightness, the total number of possible pictures equals p^m. According to information theory the logarithm to the base of 2 of this number represents the information content of each of the possible pictures, $m \ln p$ bits. We can thus conclude that the optimal information content of a half-tone picture consisting of m elements will be close to $2.3\,m$ bits.

The foregoing outlines the digital concept of images or pictures, as sets of numbers defining the co-ordinates of each elementary area corresponding to the resolution, as well as the local intensity or optical density. Optical information regarding an object can be stored and handled in digital form, and subsequently reconverted into its analog form, the conventional picture. Figure 1 illustrates the connection between the digital and the analog representation of an image. The limited p-range restricts the retrieval of information by way of vision.

The photographically recorded image, however, suffers from similar limi-

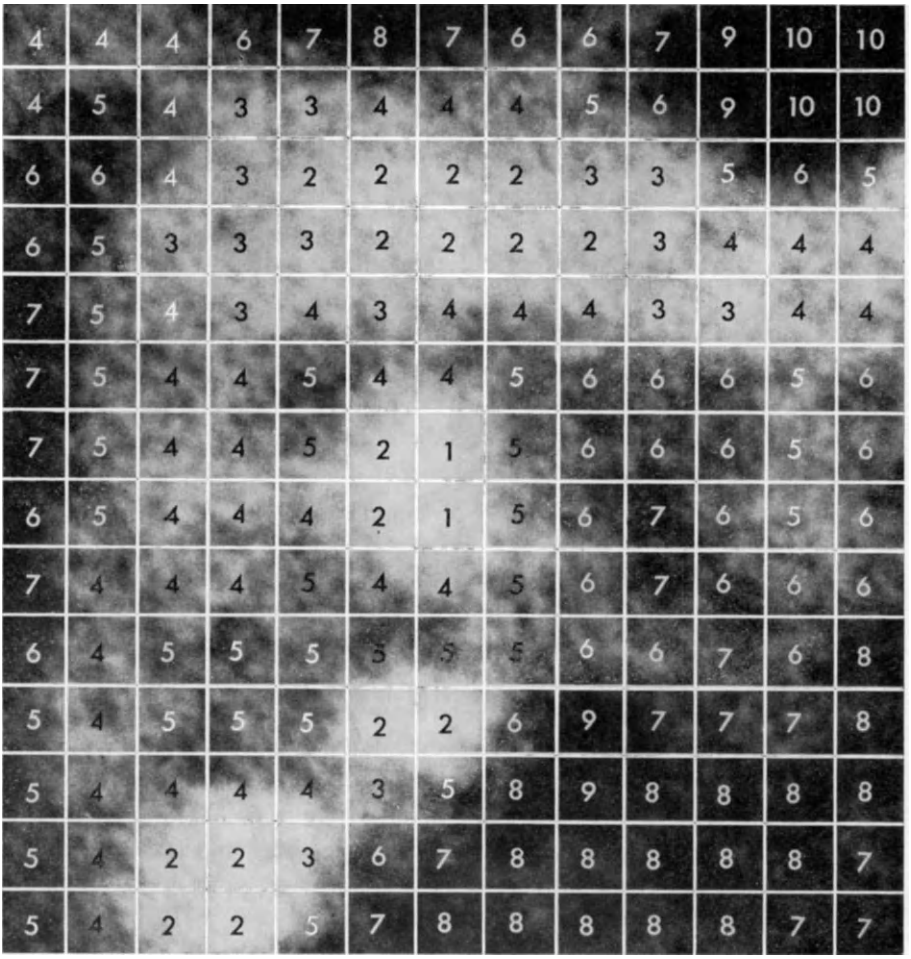


Fig. 1. – Digital and analog aspect of an image. Any image can be approximated by a quantized scalar function of two independent quantized place co-ordinates.

tations, as the recording of more than 5 gradation steps becomes progressively more difficult. This similarity explains the general success and adequacy of the photographic process. The limitations of vision and recording just mentioned could be overcome by the use of a linear detector for measuring electron densities in the image, combined with digital recording. This would offer the possibility of collecting and storing considerably more information than with the conventional analog version of photographic recording.

1·3. Speed of picture reading.

If the viewing of electron micrographs with the purpose of extracting information by an expert is assumed to be a process not very different from normal reading, an interesting point arises. It has been established that reading at normal speed is associated with the processing of about 25 bits of information per second. This enables an estimate to be made of the time required for reading an image field of given magnification and known resolution. The conclusion is that in many cases in the nonverbal communication of showing pictures not sufficient time is given to get the message through. Electron micrographs containing much information have to be studied during a correspondingly long time.

1·4. Statistical effects.

The recognition of structure patterns, particularly near visual thresholds, such as acuity of vision or minimum contrast discrimination, can be assumed to be greatly influenced by memorized knowledge or anticipation. In general there is in microscopy a typical preponderance of small detail for interpretation, and in this connection it seems appropriate to point out a fundamental difference between optical images, such as photographs or the retinal image, and the electron optical image. The number of electrons per image element during the observation time is several orders of magnitude smaller then the number of photons, and the statistical fluctuations are accordingly larger. As a result, for a given electron-optical magnification and a given sensitivity of the recording device, the number of possible gradation steps that can be recorded decreases with the size of image elements chosen. This holds for any arbitrary recording device suitable to translate an electron density distribution into a half-tone picture, when there are limits of response.

The integrated and recorded electron distribution in the image will have to contain the wanted number of distinguishable gradation steps as a minimum condition, since the subsequent photographic recording or any following information-converting device can only reduce, never increase the p-value.

The combined effect of electron fluctuations and grain distribution fluctuations on the photographically recorded image is to be considered as noise, but as the average number of developable grains produced per electron is of the order of 50 (Frieser and Klein [4]; Digby, Firth and Hercock [5]), this noise can be considered as being solely determined by the electrons.

The electrons reaching each image element in a given interval of time can be considered as a random variable obeying a Poisson distribution, since the number of possibile events is large and the probability of any individual event occurring in the time interval and area considered is small. Under these conditions the mean $\bar{n}$ and the variance σ^2 are equal to the expected number of events, in this case the expected average number of electrons, n_e per image element, during exposure or observation time. The varying value of n_e as a function of the place co-ordinates constitutes the information content of the image.

For the recording of detail it is essential that the numbers of electrons falling on two neighbouring image elements, n_1 and n_2, differ in a significant way, that is to say with a difference considerably larger than the expected statistical fluctuation. The latter is the difference of two random variables, the respective expected number of electrons falling on the two image elements considered, n_1 and n_2. Assuming that there is *no correlation*, and applying the rule that the variance of the difference (or the sum) of two independent random variables is equal to the sum of their variances, we conclude that the variance of the difference $\sigma_{1,2}^2 = n_2 + n_1$, and the standard deviation $\sigma_{1,2} = \sqrt{n_2 + n_1}$. In order to fulfil the condition that the difference be significant, we use the concept of signal-to-noise ratio. Considering $n_2 - n_1$ as the signal and $\sqrt{n_2 + n_1}$ as the associated noise, and choosing the signal-to-noise ratio equal to q, we find for a single step the condition

$$(n_2 - n_1)/\sqrt{n_2 + n_1} = q\,.$$

A simple calculation gives the required average number of electrons per object element, n_e, for p-gradation steps as:

$$n_e = \frac{p(p+1)}{2} q^2\,.$$

Figure 2 shows the curves for four values of signal-to-noise ratio q. The value 1 gives a hardly recognizable image; $q = 6$ gives already a picture of good quality in which noise is noticeable but not disturbing. The $p = 5$, $q = 6$ picture requires 540 electrons per image element during the observation time, and is a good standard. The curves of Fig. 2 can be used to find the minimum number of electrons per image element, n_e, necessary to eliminate the effect of statistical electron fluctuations for a wanted number of gradation steps to be recorded, as a minimum condition for picture quality

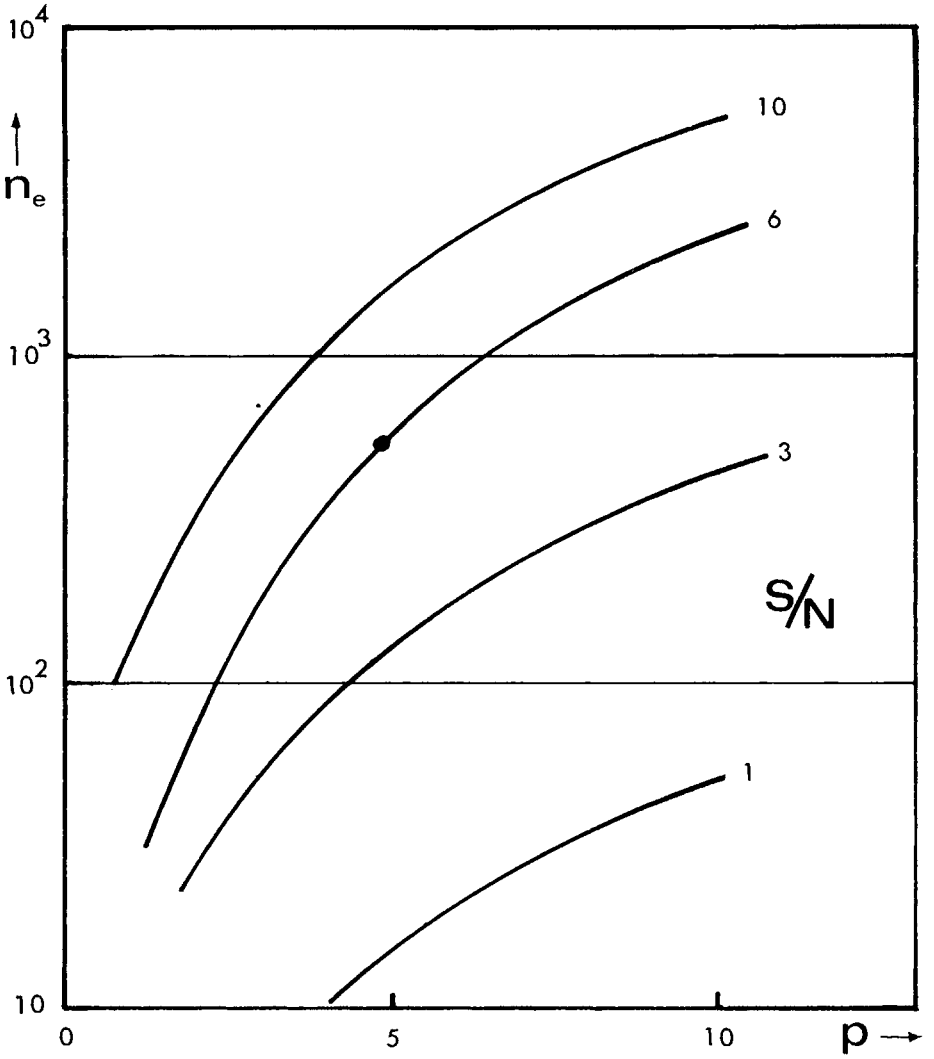


Fig. 2. – Number of electrons per object element (or image element in the least dense area) n_e, required during the time of observation for producing p statistically distinguishable density levels in the image, for different values of the signal-to-noise ratio S/N. The dot marks a good standard picture, with $S/N = q = 6$, and $n_e = 540$.

down to the smallest resolvable details, irrespective of the recording device used.

In case of recording on a photographic emulsion the required minimum magnification for fulfilling the fluctuation condition follows from the sensi-

tivity of the emulsion to electrons. Too low a magnification would lead to over-exposure, resulting in a considerable loss of information, or alternatively, with correct exposure, would give an undue amount of noise in the smallest detail. The fluctuation condition holds for all recording devices, including television pick-up tubes and image intensifiers. In the latter case statistical effects have a much higher probability of entering into the interpretation of the image, because under the conditions justifying the use of such a device the input intensity is low and the integration time short, so that the number of electrons per element during observation becomes relatively small. The fluctuation condition is also applicable in the case of a high-resolution transmission scanning electron microscope.

The assumption made that there is no correlation between neighbouring elements is not tenable in the case of typical phase structure, when there is a strong negative correlation. Usually a two-tone picture is then wanted, and the value $p = 2$ can be taken for the fluctuation condition.

1·5. Extended-area contrast.

In the further discussion of contrast phenomena we will next consider the contrast in extended areas substantially larger than the resolving power of the electron-optical system. This case is of importance for the study of, for instance, the quaternary structure or conformation of macromolecules, and for the identification of their subunits. This type of contrast is often referred to as *amplitude contrast*, or *absorption scattering contrast*, or, especially in electron metallography, as *diffraction contrast*. In Lenz's treatment of transfer of image information in the electron microscope, in this volume, it appears as a second-order term in eq. (2.19). The contrast mechanism involved is the removal out of the imaging beam of electrons scattered over relatively large angles by the aperture, causing a fictitious amplitude modulation in the object function, the intensity distribution just behind the object. Historically this form of contrast has been the first observed and discussed (von Borries [6], Leisegang [7]). Calculation of the actual value, depending on the object, the aperture and the accelerating voltage, requires detailed knowledge of the process of electron scattering in the object. Significant theoretical investigations made by Lenz [8] and by Burge and Smith [9] have provided the necessary data within the limits of plausible assumptions and approximations. The most striking result is the simultaneous occurrence of *elastic scattering* over relatively large angles by the field of the nucleus of the

scattering atom, and small-angle *inelastic scattering* by the interatomic electrons. At current accelerating voltages and with aperture angles of the order of 0.01 rad as used, all electrons scattered outside this aperture by an object of normal thickness consisting of light elements have suffered 1 elastic scattering. The electrons scattered inside the aperture, thus taking part in the image formation, have been scattered either elastically or inelastically. The larger part of the inelastic group is contained within an angle of, say, $3 \cdot 10^{-4}$ rad. Experimentally it appears that the intensity of the transmitted beam as observed in the image follows the same law as is found in optical absorption, especially in noncrystalline objects.

By placing a homogeneous thin object into the object plane the intensity drops to I_1, following the relation $I_1 = I_0 \exp[-\varrho z/\varrho z_0]$, z being the thickness of the object in the beam direction and ϱ the specific density, ϱz the mass thickness of the object area concerned. It is to be noted that there appears to be hardly any influence of the atomic number; therefore in principle the mass distribution in the object can be determined from the recorded electron distribution in the image when the instrument parameters are known, but no element discrimination is possible. The thickness for which the exponent equals -1 corresponds to a phenomenological mean free path z_0 for an electron to be scattered outside the aperture. This mean free path is not very different from that for overall elastic scattering.

Contrast is by some authors being defined as $\ln I_0/I_1$, but we will use as a definition $K_{01} = (I_0 - I_1)/(I_0 + I_1)$, the « coefficient of visibility » introduced by Michelson in optics, and widely used in the treatment of optical contrast transfer.

For a thickness z_0, $K = 0.46 \approx 0.5$, as I_1/I_0 then equals $1/e$. It is a characteristic thickness discriminating between opacity and translucence, and was given the name « clearing thickness » (Aufhellungsdicke) by von Borries ([6]). The dependence of this amplitude or scattering absorption contrast on accelerating voltage, aperture, and mass thickness enables, as mentioned, to determine the third dimension of the object in terms of density per unit area for extended areas, a procedure often called quantitative electron microscopy (Zeitler and Bahr ([10,11])).

Figure 3 shows the voltage dependence of K for given ϱz and aperture α.

This curve illustrates all relevant features of extended-area contrast in a semi-quantitative way; it holds in good approximation for a carbon film of $1.50 \cdot 10^{-5}$ g/cm^2 and a rather small aperture of the order of 1 mrad.

The K values for voltages higher than 100 kV are estimates, as no accurate experimental or theoretical data were available. The voltage scale is relativ-

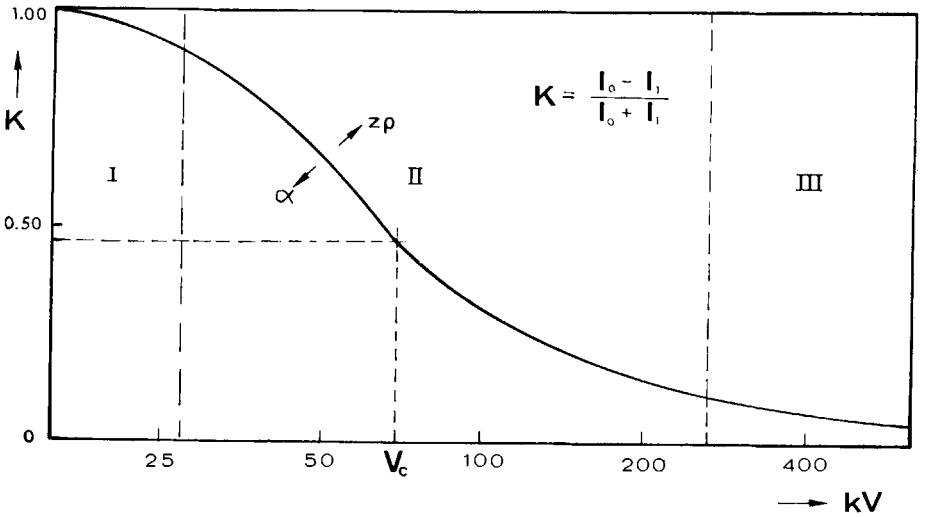


Fig. 3. – Extended area contrast as a function of accelerating voltage. For detailed explanation see text. According to a tentative empirical scaling rule the contrast K is a function of $(V \cdot z^{-\frac{1}{2}} \cdot \alpha^{\frac{1}{2}})$ (van Dorsten ([12])). I, II and III, regions of respectively low-voltage, conventional and high-voltage microscopy. In the latter region the contrast is practically pure phase contrast in areas small enough to be illuminated coherently.

istic. The K value 0.46 corresponds to the « clearing voltage » V_c, for which the mass thickness equals the clearing thickness.

Increasing the mass thickness ϱz or the aperture α shifts the curve as indicated by the arrows. The regions I, II and III are typical for low-voltage, conventional and high-voltage microscopy. At very low voltages even the thinnest objects appear black on the screen and internal structure is not or hardly observed; phase contrast is practically absent, as too few electrons are scattered inside the aperture. The contrast is pure amplitude contrast. At very high voltages thin objects appear transparent; there is hardly any amplitude contrast, but the relatively large fraction of elastically scattered electrons passing through the aperture can produce a considerable degree of phase contrast. Conventional electron microscopy has established itself in the intermediate range of voltages, where both forms of contrast are available. In general practice there is the typical preference for the slightly underfocussed image, with a balance between the two forms of contrast and an additional enhancement of contours by Fresnel diffraction.

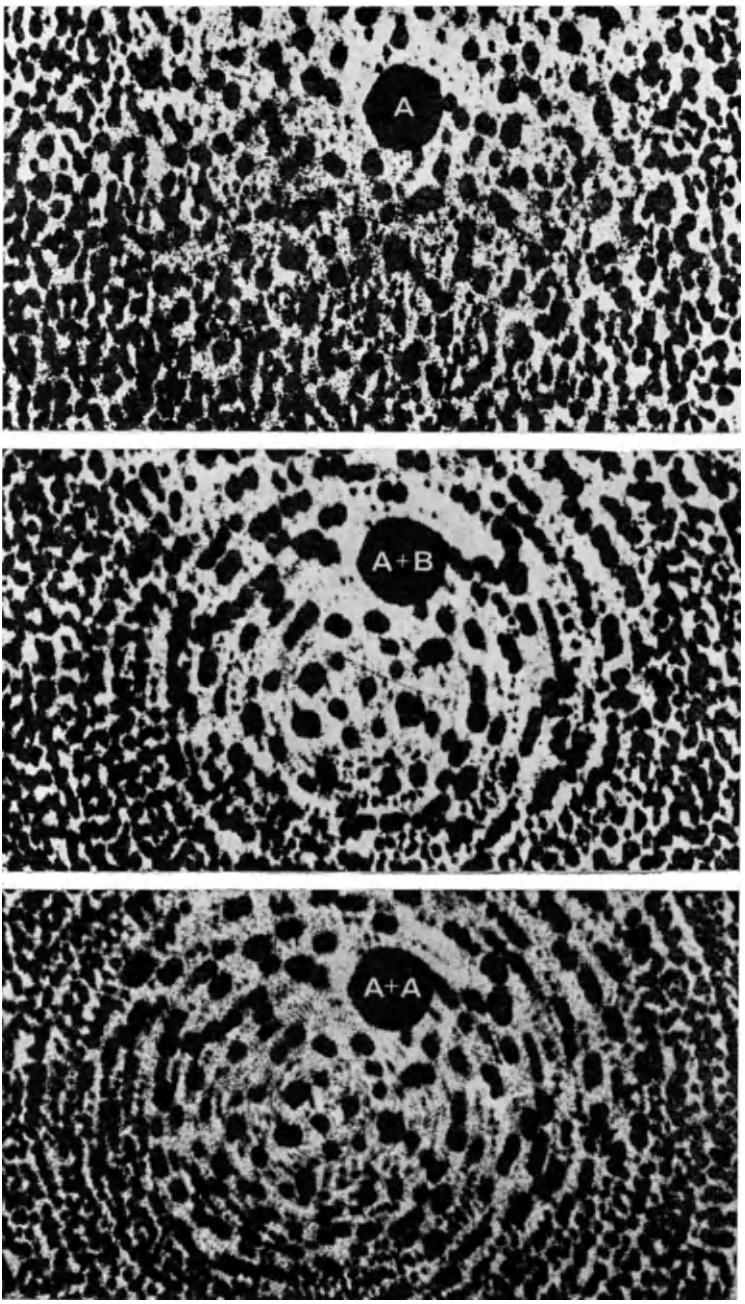
The extended-area contrast discussed in the foregoing remains of considerable interest in the electron microscopy of macromolecular objects, for which detail substantially larger than the resolving power of the microscope

can already present serious contrast problems. The contrast wanted in this case is pure amplitude contrast, with the recorded electron density in the image as a measure for the local mass thickness in the corresponding object area. Phase contrast of the rather coarse structure has to be avoided by proper focussing, but the Fresnel diffraction at the larger detail cannot be disregarded, as it improves the visibility of contours in slightly defocussed images. As pointed out, the optimal visual impression requires the information content of the image to be considerably lower than the amount of information that can be recorded by means of a suitable linear detector. As a consequence, digital recording and adapted methods of data processing do seem an attractive alternative to merely viewing pictures. On the other hand, the remark can be made that some possibilities of the sense of vision deserve more attention in electron microscopy.

One of them is stereoscopic observation: two pure amplitude contrast pictures of the same area taken at a suitable relative angle can provide accurate information about the third dimension (Cole ([13]), Helmcke ([14]), Nankivell ([15])). One requirement for good stereoscopy is sharpness, and it seems possible to « de-blurr » normal pictures by means of image-processing methods (Kovàsznay and Joseph ([16])), and to use the resulting modified pictures for stereo evaluation.

A second facility of the sense of vision, less well known, is the perception of correlation in « near fit » superimposition of pictures of random grain distributions. It can be used to estimate or determine the amount of noise in such objects, and is illustrated in Fig. 4. The method consists in making two independent pictures of the same area, and superimposing the two pictures in the form of diapositive transparencies in such a way that there is a small rotational error in the relative position. The resulting picture suggests the presence of ring-like arrangements of object spots, having a common centre at the common point of the two separate pictures. This apparent centre may lie inside or outside the picture. In case there is a relatively large amount of noise in the form of large numbers of spurious specks, for instance photographic granularity, the real object granularity reveals itself by what could be called the correlation pattern. If two identical diapositives, made from one single picture, are superimposed, there is an additional correlation for the noise; comparison of the two cases shows the amount of noise (*).

(*) The method was found by accident in 1956 by the author and his co-worker at that time, H. F. Premsela, at Philips Research Laboratories (Eindhoven, Holland), and communicated at the Stockholm Electron Microscopy Conference in the same year, but was not included in the proceedings.


A
A+B
A+A

Visualizing random grain structure is a marginal case of extended-area contrast. The method just described gives a largely qualitative and statistical answer to a statistical question. The typical pattern is only seen between the approximate limits of two values of an apparent radius R. If the average particle size is d, the average distance between particles D (supposed $D > d$), and the small angle of rotation φ, the range of R is roughly defined by: $d/\varphi < R < D/\varphi$. The study of individual particles belongs to the case of small-area contrast to be treated in the next Section.

2.

2·1. **Small-area contrast.**

In Sect. **1** the information collecting process in the electron microscope was considered to be mass thickness sampling in object areas significant for the coarser structure of the specimen. If the sampling area is gradually reduced to a size approaching the electron-optical resolution limit, the atomic nature of matter becomes progressively more influential in the electron density distribution in the image. The usual procedure in the theoretical treatment of electron images of thin objects is to make use of an essentially optical model, either a pure phase object or a combined phase and amplitude object.

◀ Fig. 4. – Correlation method for identifying random grain patterns in noisy pictures and visualizing the amount of noise. Two independent pictures resulting from a repeated observation of the same object area are superimposed with a deliberate rotation error of *e.g.* 3°. This causes a typical visual impression of seeing ring-like arrangements of spots centred around a point inside or outside the picture, the common point of the two pictures. The occurrence of such a correlation pattern is a proof of at least partial identity of the two observations. If two identical pictures of a single observation are superimposed, the noise appears as a correlation pattern also. Comparison of the two cases is an indication for the amount and nature of the noise. The object is: clusters of tungsten atoms deposited on a SiO film in a gas discharge. A: one single observation; $A+B$: superimposition of two independent observations; $A+A$: superimposition of two identical single observations. Comparison of $A+B$ and $A+A$ shows that the recording noise, mainly photographic grain, has a much finer structure than the granularity of the object. Magnification: 136 000×.

In the object function, that is the wave function immediately behind the object, the amplitude modulation is largely a fictitious one, because the electrons missing in the image are not absorbed in the object but by the physical aperture in the objective lens. The optical model does not account for elastic and inelastic scattering processes affecting the larger part of the transmitted electrons. The present treatment is based on general wave theory and originates from Bremmer ([17,18]). The special adaptation to the electron-optical case has been described in a paper by Bremmer and van Dorsten ([19]).

The problem has been simplified by not taking into account the image formation itself. Instead, the structure of the wave field immediately behind the object, forming the input effect of the subsequent electron-optical imaging system, is derived. The region concerned is the Fresnel region in which the resulting wave amplitude can be considered to result from the interference between the geometrical-optical primary wave and the diffracted or scattered wave generated in the object; this concept is closely connected with the representation by Rubinowicz of Kirchhoff's diffraction theory of image formation (Rubinowicz ([20])). Furthermore the main difference with existing phase contrast theories consists in taking into account the inelastically scattered electrons, as well as the elastically scattered ones. The problem lies in the borderland of optics, diffraction and scattering, which more or less specifically apply to coarse structure, periodic structure and statistically determined structure, respectively. The phase structure of the wave after passage through the object can be expected to show resolvable periodicities of the order of size of interatomic distances as well as coarser ones, as first pointed out by von Borries and Lenz ([21]). An extensive experimental and theoretical investigation of the imaging of such structures has been made by Thon ([22]) and is reported by him in his volume.

For practical electron microscopy the simultaneous occurrence of amplitude contrast and phase contrast presents serious problems in interpretation. In the following approach the structure of the wave field in and closely behind the object will be described. The complex wave amplitude in this region resulting from interaction between the electron beam and the atoms constituting the object defines a shadowlike object function, which will be called the real object function. The idealized image of this real object function would show already most of the characteristics of the actually observed image. The latter, of course, could be worked out in principle by one of the known methods, of which the contrast transfer theory developed by Hanszen ([23]) seems the most logical. Perhaps the most general theoretical treatment is the one given by Lenz in this volume.

2'2. Structure of the wave field immediately behind the object.

To derive the real object function the stationary phase method is applicable. The pertaining Schrödinger-function results as the sum of the primary wave constituting the incident electron beam and the single spherical waves generated in the scattering centres.

For the simplest case of perpendicular incidence of the beam on the object the supposedly coherent primary beam can be represented as:

$$\psi_{pr} = AN_0 \exp [ik_0 z] ;$$

the optical axis is the z-axis. The factor AN_0 indicates the proportionality with the electron density N_0, and k_0 stands for the de Broglie wave number $2\pi/\lambda_0$ of the incident electrons (Fig. 5). The scattered wave generated by any single scattering centre Q has the form

$$\psi_{pr}(Q) B \frac{\exp [ik_a QP]}{QP} f(\theta_s) .$$

QP is the distance to Q, B a scattering factor, and $f(\theta_s)$ the dependence on the scattering angles θ_s between QP and the incident beam. The wave number k_a equals k_0 for elastic scattering; for inelastic scattering $k_a < k_0$. The general case with k_a complex includes absorption, but this does not occur in fully amorphous objects. The summation of all scattered waves can be approxi-

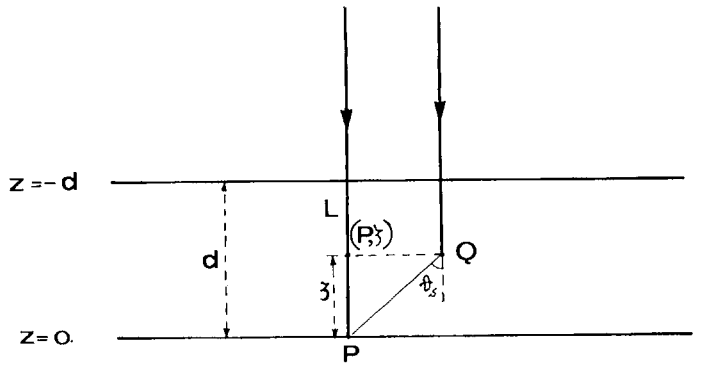


Fig. 5. – Transilluminated object with in the exit plane the point P, for which the resulting wave amplitude is to be derived; Q arbitrary scattering centre; L line, in the direction of incidence, of stationary phase of the scattered waves for all points Q.

mated for small scattering angles. For objects not exceeding a thickness l^2/λ_0, with l a characteristic length to be observed in the object, the main contribution can be calculated directly using the saddle-point approximation, which is applicable because there is a stationary phase: the phase of the scattered wave as a function of the position of Q appears to have a minimum value for all points Q on the straight line L parallel to the incident beam.

The first approximation in which the summation over the discrete scattering points is taken over an assumed continuous density $n_s(Q)$, yields the expression for the real object function observed in the plane $z = 0$, for the case of only one type of scattering, either elastic or inelastic:

$$\psi(P) = \psi_{\mathrm{pr}}\left[1 - \varepsilon(x_P, y_P) + B\frac{2\pi i}{k_a} f(0) \cdot \int_{-z}^{0} \mathrm{d}\zeta\, n_s(P, \zeta)\{1 - \varepsilon(P, \zeta)\} \exp\left[i(k_0 - k_a)\zeta\right]\right]. \tag{1}$$

Here the place function $\varepsilon(Q)$ has been introduced, which accounts for any observable attenuation of ψ_{pr} on arrival in Q. For a plane object with boundaries the planes $z = -d$ and $z = 0$, this factor will obviously be approximately proportional to $\int_{-d}^{z_Q} n_s(x, y, z)$.

Phenomenologically, as observed in the final image, the losses have to do with electrons absorbed by the aperture or those playing no role because of a considerable degree of chromatic aberration. The term $(1 - \varepsilon)$ between the brackets in (1) accounts for the primary wave partially attenuated by scattering, whereas the integral term stands for the total contribution of all scattered waves. The integral is to be taken all along the line L over the thickness d of the object.

2·3. Special case of pure phase modulation.

This case occurs when the expression between brackets in (1) is of the form $1 + id\varphi$, which in our approximation equals $\exp[id\varphi]$; it means prevalent elastic scattering in relatively small angles with $k_a = k_0$, that is in thin objects observed with fast electrons. For this case, typical for high-voltage microscopy, when ε is very small and may be neglected, one gets:

$$\psi(P) = \psi_{\mathrm{pr}}\left\{1 + B\frac{2\pi i}{k_0} f(0) \int_{-z}^{0} \mathrm{d}\zeta\, n_s(P, \zeta)\right\}, \tag{2}$$

B being real. In the saddle-point approximation used, the factor i results from the total contribution of all scattering centres, assumed to be continuously distributed and lying in a plane perpendicular to L defined by $\zeta = \text{const}$, and being illuminated sufficiently coherently. In practical cases, with discrete scattering centres, the mutual interferences as well as those with the primary wave are confined to a small area around the point of intersection of L with the plane concerned. The radius of this region is comparable to some sort of Fresnel fringe width of size $\sqrt{h\bar{\lambda}_0}$, h being the distance between the point observed, P, and the plane.

It is to be noted here that the degree of coherence of the incident wave plays an important role. Because the dominating phase structures appear only as image structure as a result of axial or zonal defocussing (spherical aberration) of the objective lens, the interpretability or object-similarity of the image requires an optimal degree of coherence, giving still a good display of first-order interferences, but causing the higher orders to be largely cancelled by phase-shifted superposition. This corresponds to an illuminating angle at which in the defocussed image of an edge only one Fresnel fringe distinctly appears.

2·4. The occurrence of amplitude modulation in the real object function.

We next consider the case of inelastic interaction, in which $k_a < k_0$ and B is, in general, complex. This causes the term proportional to B in (2) to become also complex, and this means that any phase modulation will neccssarily be accompanied by an amplitude modulation. Neglecting, as before, the loss factor ε leads to the following expression in the place of (1), in which the splitting into a real and an imaginary part is clearly represented:

$$\psi(P) = \psi_{\mathrm{pr}}\left[1 - \frac{2\pi|Bf(0)|}{k_a}\int_{-d}^{0} \mathrm{d}\zeta\, n_s(P,\zeta)\sin\{(k_0 - k_a)\zeta + \arg B\} + \right.$$

$$\left. + \frac{i2\pi|Bf(0)|}{k_a}\int_{-d}^{0} \mathrm{d}\zeta\, n_s(P,\zeta)\cos\{(k_0 - k_a)\zeta + \arg B\}\right]. \qquad (3)$$

The real part of the expression in brackets, that is the amplitude modulation, will dominate because in the intensity $|\psi|^2$ this part contains a first-order contribution in the scattering parameter B, whereas the imaginary part con-

tains only a second-order contribution in B. Apparently this means that we have a typical example of an amplitude modulation as a result of the interferences between inelastically scattered electrons. The pertaining Fourier integrals (with finite integration intervals, determined by the thickness of the object) become particularly influential in case the object structure in the direction of the incident beam shows periodicities of the order of $2\pi/(k_0 - k_a)$. This periodic distance can be approximated by a length $\lambda_0^2/\Delta\lambda_0$, with $\Delta\lambda_0$ representing the increase of the de Broglie wavelength associated with energy loss by inelastic scattering. A further condition for the possibility of observation of this effect is that the chromatic aberration shall not be too large, in order that the slowed-down electrons are not much defocussed.

This type of interference that can also be characterized by $k_0 - k_a = k_l$, $k_l = 2\pi/l$ being the wave number of the periodicity concerned, is of a type well known in general wave optics. It occurs, for instance, in the diffraction of light by ultrasonic waves. This type of optical diffraction was predicted by Brillouin as early as 1921, and observed experimentally several years later (Brillouin ([24,25])). Interference by more or less coherent interaction effects of electrons having suffered inelastic energy losses in electron microscope images of thin metal foils have been observed by Watanabe ([26]).

In general, observations on energy loss indicate distinctly discrete values of k_a in a number of cases. The conditions for interference require well-defined k_l values, which will occur in amorphous substances with a statistical distribution of some kind, in case the thickness of the object is at least of the order of $\lambda^2/\Delta\lambda_0$. In carbon, with characteristic losses of about 20 eV per inelastic interaction, the required minimum thickness is of the order of 400 Å for single scattering at conventional voltages, and accordingly smaller for multiple scattering. In crystals which generally show effects of strong coherent scattering, this type of interference can be expected to show phenomena of this kind as a result of a sort of resonance. Experiments combining high resolution with energy filtering could probably further clarify this point. In amorphous objects the occurrence of granulation, also in the parafocal region, might probably be explained along these lines.

2·5. **Appearance of phase structure.**

Going back to elastic scattering processes we note that according to (2) only integrals of the type $\int d\zeta\, n_s(P, \zeta)$ play a role. These integrals depend on the inner potential but not on its distribution in the beam direction.

The integrals are analogous to those in optical cases where the path of integration is to be taken along a geometrical-optical ray, when $n_s(P, \zeta)$ represents the local value of the refractive index. Such integrals become important in cases where the path through the medium does not exceed the mentioned Fresnel length l^2/λ_0. In practical electron microscopy the resolution limit d_0 will define the minimum characteristic length that can be observed. The condition for the object thickness, $d < d_0^2/\lambda_0$, is to be fulfilled in order to justify application of the theory outlined here.

Structural features of the object, in particular in the beam direction, cannot be expected to reveal themselves in the case of pure phase modulation unless there is some amplitude modulation at the same time. As n_s appears only as a summation in the beam direction, the phase modulation, even after conversion into amplitude modulation by deliberately achieved or resulting amplitude modulation, when phase contrast appears in the image, cannot supply direct structural information in single images. There is therefore no possibility for stereo observation of phase structure.

The interesting question may arise whether phase structure corresponding to interatomic distances can be expected to show up in cases when summation over, for instance, 50 atomic distances takes place, or whether a steady mean value is reached. The answer to this question has been given by von Laue ([27]) in order to explain X-ray pictures of multiple layers of uniform granular matter. Increasing the number of layers does not at all smear out the relative differences in intensity; the overall periodicity is conserved, even when individual grains cannot be identified in the shadow image, and stereo observation appears to be impossible. This phenomenon seems important for the interpretation of micrographs of macromolecules. Phase structure, coming to the fore already on very slight defocussing, may easily lead to false interpretations, and may greatly distort the amplitude part of the image. It is therefore important to be able to discriminate between the two forms of image structure. The best criterion for a phase structure is the dependence of the spatial frequencies in the image on focussing, and the disappearance of this structure altogether in the near-focus (parafocal) region (van Dorsten and Premsela ([28])) under favourable operating conditions, especially the absence of axial astigmatism.

2·6. Objects exceeding in thickness the Fresnel length l^2/λ_0.

A theory for this case has been developed by Bremmer ([18]), following a method that can be used in all cases involving the well-known Sommerfeld

integral for the spherical wave function

$$\psi = \frac{\exp[ik_a QP]}{QP}.$$

In the treatment concerned use is made of the symbolic expression for the factor $n_s(P,\zeta)\{1-\varepsilon(P,\zeta)\}$ in (1), namely:

$$\exp\left[\frac{i\zeta}{2k_a}\Delta_2\right]\cdot n_s(P,\zeta)\{1-\varepsilon(P,\zeta)\}, \tag{4}$$

where $\Delta_2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ stands for the transverse two-dimensional Laplace operator relevant to the point P. Such a symbolic expression passes into an explicit expression, *e.g.* when the function $h(x_P, y_P, \zeta)$, to which it is to be applied, is represented by its Fourier integral:

$$h(x_P, y_P, \zeta) = \int_{-\infty}^{+\infty} dk_x \int_{-\infty}^{+\infty} dk_y\, g(k_x, k_y, \zeta) \exp[i(k_x x_P + k_y y_P)].$$

Applying the operator then simply means the substitution $\Delta_2 = -k_x^2 - k_y^2$. The expansion of the exponential function in (4) leads to a corresponding expansion of the real object function, with the first term in:

$$\exp\left[\frac{i\zeta}{2k_a}\Delta_2\right] = 1 + \frac{i\zeta}{2k_a}\Delta_2 + \ldots, \tag{5}$$

yielding once more the fore-mentioned saddle-point approximation. For a characteristic length l, Δ_2 means, as an order of magnitude, a multiplication by l^{-2}. The simplest case occurs when the exponential in (5) may be neglected altogether, which involves the condition

$$\frac{|\zeta|}{k_a}\cdot\frac{1}{l^2} \ll 1.$$

Considering the fact that $k_a \approx k_0$ and $|\zeta| < d$, the condition becomes $d \ll k_0 l^2$, which is equivalent to the condition previously found that the thickness of the object should not exceed the Fresnel length l^2/λ.

The expression (1), corrected following (4), includes the diffraction phenomena, because the exponential in (4) becomes especially prevalent for

structure in the direction perpendicular to the direction of incidence, that is in the plane of the object. Given an object having periodic structure in the x-direction, the corresponding representation of the density of the scattering centres,

$$n_s(x) = A \sum_{-\infty}^{+\infty} \delta(x - nl) = \frac{A}{l} \sum_{-\infty}^{+\infty} \exp\left[2ni\left(\frac{n}{l}\right)x\right],$$

shows that not only single lengths l, but also fractions l/n play a role.

The foregoing condition requires that for all values of n, $d^2 \ll (l/n)^2/\lambda_0$, and therefore cannot be fulfilled rigorously. The occurrence of the associated diffraction phenomena is fully accounted for by the exponential in (4). The presence of higher harmonics of the spatial frequence $1/l$ is inherent to the distribution of the scattering centres as determined by the atomic nature of matter, that is with narrow, sharp maxima resembling delta functions.

2·7. **Defocusing effects.**

The general theory of Fresnel fringes leads to the expectation that at a distance z behind an object plane, object structures with periodicities of the order of magnitude of $\sqrt{\lambda_0 z}$ present themselves prevalently. Attention to this phenomenon was first drawn by von Borries and Lenz ([21]) in connection with the well-known extra focal granular appearance of thin foils mentioned earlier in this paper. A quantitative treatment of this effect is possible using an earlier described method by Bremmer ([17]).

Suppose the real object function, the wave function in the plane $z = 0$, is given in phase and amplitude by a complex function u_0 as:

$$u_0(x, y) \exp[-ikvt], \qquad \text{with } k = 2\pi/\lambda_0.$$

The corresponding wave function in the region $z > 0$ can be represented, again using symbolic expressions, as

$$u(x, y, z) = \exp\left[ikz\sqrt{1 + \Delta_2/k^2}\right] u_0(x, y) \exp[-ikvt], \tag{6}$$

in which $\Delta_2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is, as before, the two-dimensional Laplace operator. The negative sign in the exponential accounts for the fact that the wave function in the region $z > 0$ results exclusively from waves travelling

in the positive z-direction. The general expression (6) can be simplified if only structures $l_x = 2\pi/k_x$ and $l_y = 2\pi/k_y$ are considered, which are large compared to the wavelength. This renders Δ_2/k^2 a small quantity, because

$$\left|\frac{\Delta_2}{k^2}\right| = \left|\frac{k_x^2 + k_y^2}{k^2}\right| = \frac{\lambda_0^2}{l_x^2 + l_y^2} \ll 1 .$$

The expression (6) can now be approximated by the first two terms of the binomial expansion of the roots:

$$u(x, y, z) = \exp\left[ik(z - vt)\right] \exp\left[\frac{iz}{2k}\Delta_2\right] u_0(x, y) . \tag{7}$$

This simple expression explains, for instance, at once the diffraction fringes near an edge, as it yields the corresponding Fresnel integral. In order to investigate the effect of special spatial frequencies $1/l$ in the object, we put

$$u_0(x, y) = 1 - \varepsilon u_1(x/l, y/l) = 1 - \varepsilon u_1(\xi, \eta) ,$$

in which ε, for unit amplitude of the incident wave, determines the order of magnitude of the change in wave amplitude and phase caused by the object. The function u_1, as well as its derivatives to the dimensionless variables ξ and η, are now of the order of magnitude of unity. Expression (7) transforms into:

$$u(x, y, z) = \exp\left[ik(z - vt)\right]\left\{1 - \varepsilon \exp\left[\frac{iz}{2kl^2}\Delta_2'\right] u_1(\xi, \eta)\right\} , \tag{8}$$

with the new transverse Laplace operator $\Delta_2' = \partial^2/\partial\xi^2 + \partial^2/\partial\eta^2$. In the Fresnel region, for small z values when $0 < |z| < l^2/\lambda^0$, the wave function distinctly shows local object structure, whereas in the Fraunhofer region, for larger values of z, waves originating from the whole illuminated area of the object interfere. In the former region the parameter $z/2kl^2 = (1/4\pi)\,\lambda_0 z/l^2$ is still small, and the expression (8) can be replaced, using the exponential expansion, by:

$$u(x, y, z) = \exp\left[ik(z - vt)\right]\left(1 - \varepsilon u_1 - \frac{i\varepsilon z}{2kl^2}\Delta_2' u_1 + \frac{\varepsilon z^2}{8k^2 l^2}\Delta_2'^2 u_1 \ldots\right) ,$$

with the once iterated transverse Laplace operator $\Delta_2' = (\partial^2/\partial\xi^2 + \partial^2/\partial\eta^2)^2$. Taking into account the terms up to the second order in z/kl^2 and to the first

order in ε, we find the following expression for the corresponding intensity:

$$|u(x, y, z)|^2 = 1 - 2\varepsilon \operatorname{Re} u_1 + \frac{\varepsilon z}{kl^2} \Delta_2'(\operatorname{Im} u_1) + \frac{\varepsilon z^2}{4k^2 l^4} \Delta_2'^2(\operatorname{Re} u_1) . \qquad (9)$$

This expression describes the experimentally observable effects of defocusing very clearly. The phase structure determined by $\operatorname{Im} u_1$ exists only for $z \neq 0$ and thus becomes invisible at exact focus. The intensity changes sign with z, in other words there is reversal of contrast by going through focus. The amplitude part, determined by $\operatorname{Re} u_1$, does not change in magnitude or sign in the neighbourhood of $z = 0$. The positive sign of the third term can be understood, as $\Delta_2'(\operatorname{Im} u_1) > 0$ indicates a convex curvature of the equiphase planes, which means a focusing effect with increasing intensity for $z > 0$. Furthermore (9) also brings out the influence of the presence of characteristic lengths in the object. For a given structure function $u_1(\xi, \eta)$, the intensity can reach maxima for specific spatial frequencies. The following is a simple example of the case of periodic phase structure of a sinusoidal grating:

$$u_1(x/l, y/l) = i \cos\left(\frac{2\pi x}{l}\right) = i \cos(2\pi\xi) .$$

The original expression (8) for this case becomes

$$u(x, y, z) = \exp[ik(z - vt)] \left\{1 - i\varepsilon \exp\left[-\frac{2i\pi^2 z}{kl^2}\right] \cos(2\pi\xi)\right\}$$

and the intensity, approximated to first order in ε,

$$|I|^2 = 1 - 2\varepsilon \sin\left(\frac{2\pi^2 z}{kl^2}\right) \cos(2\pi\xi) ,$$

the optimum occurs if

$$\sin\left(\frac{2\pi^2 z}{kl^2}\right) = \pm 1 , \quad \text{or} \quad l^2 = \frac{2\pi z}{k(n + \frac{1}{2})} = \frac{\lambda_0 z}{(n + \frac{1}{2})} , \qquad n \text{ integer} .$$

To conclude, we can say that defocussing to an amount z will make periodic structure of the order of size of $\sqrt{2\lambda_0 z}$ well observable, but in addition also higher spatial frequencies. The reproduction of such structure in the image depends entirely on the contrast transfer process in the imaging part of the

electron microscope. For example, the Abbe or Rayleigh condition for the image-side aperture has to be fulfilled; the objective aperture acts as a low-pass filter for the spatial frequencies to be transmitted and therefore determines the z values for the parafocal region, in which phase structure is not or hardly observable.

3.

3·1. Contrast enhancing procedures by means of specimen treatment and image conversion and image processing.

In this Section we will consider the processes influencing contrast and abstract from the object itself as much as possible. What we have vaguely defined in the Introduction of Sect. **1** as the amorphous object may consist of material of arbitrary density. In all cases where the specific density is sufficiently high, there are no contrast problems for the electron microscopical recognition of size, shape and possible structure down to dimensions of the order of 5 to 10 times the theoretical resolution limit of the microscope. Contrast problems arise in the case of objects consisting of the lighter elements, such as natural and synthetic macromolecular specimens and organic polymers in particulate or filamentous form. The density may vary from 2.6, for compact globular particles such as plant viruses or certain enzymes, to 1, for polystyrene and even slightly smaller values for loosely built proteins consisting of a considerable number of globular subunits forming the quaternary structure of a larger particle. It is to be noted here that the support film mostly used, the carbon film, has a density of about 2.

It has been mentioned in Sect. **1** that for the type of object considered and the size of detail to be observed the indicated method of retrieval of structural information is through amplitude contrast. Phase contrast, extensively dealt with in this volume by other authors, in this case has to be avoided or at least reduced to the smallest possible degree by very accurate focussing in the parafocal region, that is with a focussing error not exceeding 100 to 200 Å. It may be remarked here that according to simple diffraction theory the natural depth of focus, the region in which no physical changes in the image are observable, is of the order of 50 Å in conventional electron microscopy; for this reason, statements sometimes found in legends of micrographs such as $\Delta f = -10$ Å, or even $\Delta f = 0.0$ Å, have no sense, as the reference point

has an uncertainty of the order of 50 Å. The available focussing steps in a high-resolution instrument, however, should permit an accuracy of 50 Å.

3·2. **Staining.**

Increasing contrast by specimen treatment has been known since the early days of electron microscopy. The biological material studied then did require fixation for the preservation of structure under the conditions of observation and the most widely used osmium tetroxide (Palade ([29])), a highly reactive compound, is retained at reaction sites, which appear with high contrast in the image on account of the high electron scattering power of the heavy metal.

Osmium tetroxide was found to react with fatty acids, phospholipids and related organic compounds: the primary sites of reaction appear to be the C=C double bonds, the uptake being about one atom of osmium per double bond (Stoeckenius and Mahr ([30])). Other biologically significant compounds such as nucleic acids require different staining agents, for instance phosphotungstic acid or uranyl acetate. Strong effects can be forcibly obtained: in certain cases the mass of isolated particles could be increased by a factor of 4, as reported by Hall ([31]), who in the same investigation demonstrated the usefulness of embedding particles in dense material, salts of heavy metals already in use for staining. After the work of Brenner and Horne ([32]) this method became a standard technique known as *negative staining*, as the counterpart of the *positive staining* first described. The object and some of its structure reveals itself by the absence of dense material and appears in reverse contrast. Properly applied the method permits visualizing particles including their possible surface corrugations and in some cases porous structure, tubular formations and further structural features. For a reason to be mentioned later the method is, however, not a high-resolution method in the strict sense. Positive staining, and to a lesser degree also negative staining, can affect structure by chemical interaction. The staining process may in some cases reveal structure caused or strongly influenced by its very application.

3·3. **Anomalous contrast.**

A discussion of negative staining in connection with contrast would seem incomplete without mentioning an interesting contrast anomaly described bu Müller and Meyerhoff ([33]). They found that particles of various nature embedded in phosphotungstate showed a remarkable loss of absolute con-

trast, compared with identical particles in the same specimen, lying in an area of the support film free of tungstate. Although there have been some controversies on the reality of the effect, that was given the name anomalous contrast by the authors, its observation was found to be correct beyond any doubt in a number of cases. Visually observed contrast phenomena can be frequently associated with properties of the sense of vision, as for instance the Mach effect, the subjective but remarkably strong increase of contrast at contrast boundaries. The anomalous contrast, however, is also found in extended areas remote from any boundary. It appears with particulate objects of diverse nature, small inorganic crystals as well as polystyrene latex globules or virus particles. A tentative explanation satisfying the non-physicist authors in first instance, based on the assumption of the building-up of positive electric charges under the influence of secondary electron emission from the phosphotungstate during exposure to the electron beam, creating hypotetical « microlenses », was refuted by Lippert ([34]). Although the effect of electric fields cannot be ruled out completely in many cases, the first explanation given does seem very improbable. An alternative explanation suggested here is to consider the phenomenon of anomalous contrast as caused by the dual nature of the electron scattering in the type of object concerned, making it a typical two-component medium. One component is the heavy atoms causing practically exclusively elastic scattering over relatively large angles, the other the light elements, for which the low-angle inelastic scattering events outnumber the elastic ones by a considerable factor. As follows from scattering theory, confirmed by experiment, the ratio between inelastic and elastic scattering interactions is largely independent on the electron energy but strongly dependent on the atomic number. For $Z = 25$ the ratio is unity, and the following simple expression holds in good approximation: $n_{in}/n_{el} = = 25/Z$. A quantitative treatment in any special case would require more explicite data on the scattering processes involved than are presently available from theory and experimental work, and in addition an exact physical description of the object: dimensions in the beam direction, mass thickness distribution and atomic composition, data that are usually either unknown or only known as rough estimates. However, even without detailed analysis it will be clear that in the dense parts of the specimen, where both components are present, there will be on the average at least one elastic interaction with a heavy atom and several inelastic interactions with the lighter component for each incident electron. As a result the average transmitted electron will have suffered an angular deviation of the order of 10^{-2} rad and an energy loss equal to a multiple of the average energy loss per inelastic interaction,

that is n_{in} times 20 eV, if the light component is taken to be carbon. If it is further assumed that the mass thickness of the light component is of the order of the clearing thickness, we know from theory that n_{in} is about 4, and thus the energy loss about 80 eV. Due to the elastic scattering the fraction of these electrons passing through the objective aperture will fill the whole aperture uniformly. If the microscope is focussed for electrons having the energy of the incident beam this group of electrons is rather strongly defocussed to circles of confusion of a size of several hundreds of Ångstroem units. If for example, the accelerating voltage is 80 kV, the focal length of the objective 1.8 mm and the image-side semi-aperture angle 1.5×10^{-2} rad, the diameter of the circles of confusion for electrons having lost 80 eV is about 400 Å, if the coefficient of the chromatic aberration is taken to be 0.7 times the focal length.

Because in reality n_{in} shows a statistical distribution, there is a whole range of sizes of the chromatic circles of confusion, causing a practically uniform background of slightly decelerated electrons. If the area of the compound scatterer has a sharp boundary, there will be a focus-dependent gradual fade-out from dark to light in the image on the screen, with a noticeable maximum of intensity inside the light area close to the boundary. This edge effect can be easily observed in many micrographs of negatively stained areas appearing semi-transparent. If the stain appears dark, as a result of excessive mass thickness or a small objective aperture, or both, this edge effect disappears. Small areas enclosed by the stain, either holes or embedded particles of relatively low density, show up with increased intensity on account of the overshooting of the decelerated electrons, and this means the possibility of anomalous contrast in such areas.

According to the explanation just given, conditions favourable for the occurrence of anomalous contrast are: semi-transparent areas consisting of a combination of heavy and light elements, a sufficiently large objective aperture (or no aperture at all), and focussing for the electrons having suffered no energy loss. There is also a noticeable influence of the accelerating voltage. Increasing the voltage reduces the effect as the result of the decrease of the elastic scattering cross-section and the decrease of the mean scattering angles, both of which reduce the relative number of inelastically scattered electrons falling inside the objective aperture. The effect of anomalous contrast disappears largely when the image is strongly underfocussed and the decelerated electrons are thus brought closer to focus. In fact, when anomalous contrast appears, the image is difficult to focus, a property found more or less with

practically all negatively stained specimens. The resulting pictures exhibit nearly always a typical partial unsharpness.

3·4. Shadow casting.

Evaporating heavy metal at a small angle onto the support film carrying the specimen, introduced by Williams and Wyckoff ([35]), has become an indispensable method of increasing contrast and simultaneously revealing local differences in the physical thickness of the object. Observed as a negative print the effect produced resembles the effect of oblique illumination of a corrugated surface. The type of picture produced is too well known to be discussed into detail here, but it can be said that the method is not a fully satisfactory high-resolution method because the one-sided deposit of metal, in order to be effective, has to have a minimum thickness exceeding the resolving power of the electron microscope by a considerable factor. The visible effect, therefore, is more suggestive of the result of a snow storm than of the immaterial touch of light, detail being covered, profiles smoothed. A firm belief in the simplicity of the principle has led the electron microscopists to choosing the heaviest metals for shadow casting, even metals requiring very high temperatures for evaporation. The formation of small aggregates during evaporation or during exposure to the electron beam in the microscope can be the cause of disturbing structure in the metal coating. Apparently, nucleation followed by the growth of separate small crystals takes place. In some cases the use of alloys or the simultaneous evaporation of two different metals reduces this unwanted effect as a result of continuous changes in the ratio of the two components in the condensing material, which renders the formation of possible crystal phases more difficult. A combination of gold and palladium, for instance, is much better in this respect than gold, which is notorius for aggregating very easily in the beam. Good results are found with a combination of tantalum and tungsten; platinum simultaneously evaporated with carbon is the method of choice of many electron microscopists. Apart from these empirical facts, one might expect the occurrence of similar phenomena as observed with negative staining and described in the foregoing, when objects consisting of light elements such as carbon are treated with heavy elements for shadowing. Effects of this kind have been receiving little attention but do indeed exist. There are good reasons to believe that, in fact, they determine the ultimate limits of resolution of the shadow casting methods. As an example we will report here on some observations in connection with an investigation on shape

and possible structure of small particles of special biological interest, ribosomal subunits, that can be isolated from certain bacteria. These particles which appear as globules measuring about 300 Å have been studied by Nanninga ([36]) in their isolated form, using the best standard of shadow casting technique. The particles isolated in pure state by ultracentrifugation were freeze-dried upon a carbon foil prepared by evaporation on a freshly made cleavage plane of mica, and supported by the network of a holey formvar film. The shadowing was done by a simultaneous evaporation of carbon and platinum in high vacuum under a fixed angle of approximately 50 degrees. A careful examination of the resulting micrographs reveals features requiring further explanation, see Fig. 6. The most striking observation when studying the micrographs is, that whereas the ribosomal particle itself shows not very well-defined boundaries and a distinctly unsharp outline at the side of maximum thickness of the shadowing material, the outline of the shadow on the support film is remarkably sharp. From the shape of the sharp shadows of several particles in the field of observation, having random orientations in respect of the direction of shadowing, it appeared to be possible to conclude to an icosahedral shape of the particle. The various shadows appear bounded by straight line sections forming parts of polygons and were shown to agree very well with this shape, as was proved by means of a cardboard model in a parallel beam of light.

A question arising from this observation is: how can a particle showing itself with a not well-defined outline produce a sharp shadow? The possible answer is that it is, as in the case of negatively stained macromolecular objects, again the two-component nature of this particular object that is causing a high proportion of inelastically scattered electrons from regions containing maximum amounts of both the light and the heavy component. Thus the most densely shadowed parts, which are the most elevated and therefore thickest parts of the compound specimen, cannot be focussed sharply as a result of chromatic aberration. On the other hand, the areas in which the shadow on the support film can be seen are the thinnest and least dense part of the specimen. The total amount of the light component measured in the beam direction in these regions is too small to cause multiple inelastic scattering, so that the corresponding image area is hardly affected by chromatic aberration.

A second feature, frequently observable at careful examination, is the presence of « ghosts » in the image, areas slightly brighter than the background in the negative print and surrounding the particles. These areas show approximately the same shape as the particles, but larger by a factor

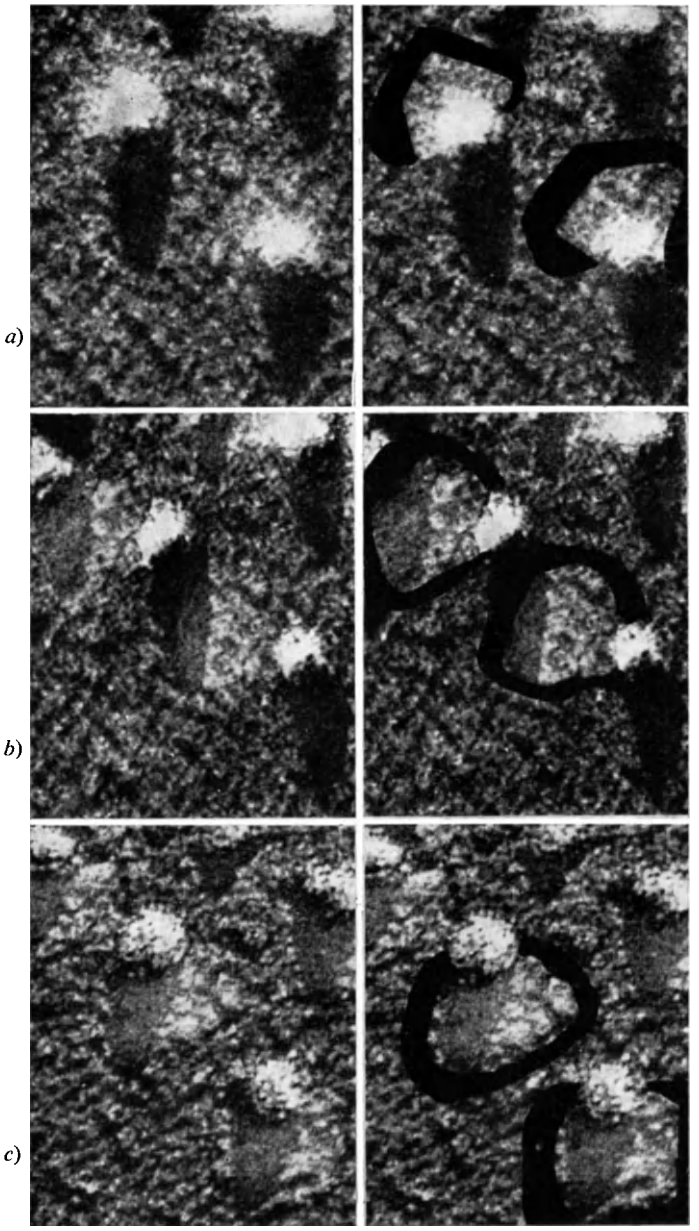
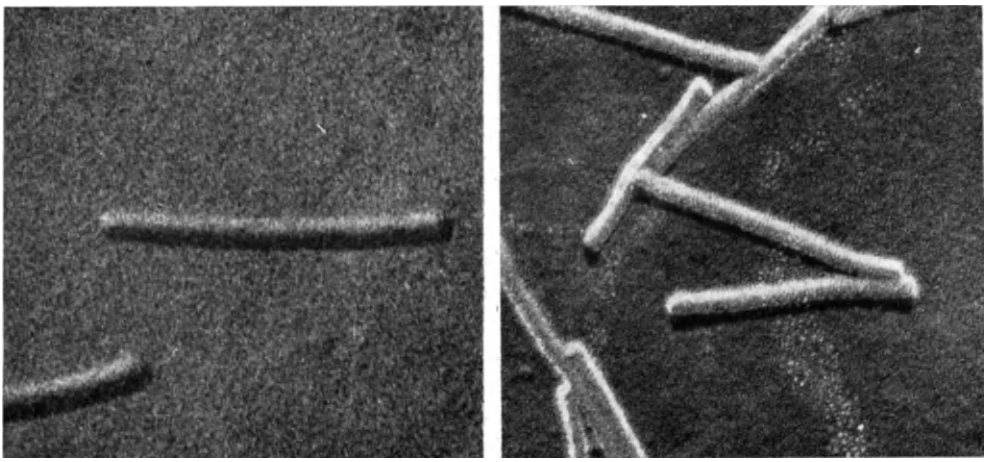


Fig. 6. – Freeze-dried 50 S ribosomal subunits of Bacillus Subtilis, 290 000×. Shadow: platinum/carbon at about 50°, (Nanninga ([3])). The shadow appears sharper than the particle itself. Note « ghosts », areas slightly brighter than the background, the position of which changes when the specimen is tilted or rotated with respect to the beam: *a*) specimen plane approximately perpendicular to the beam; *b*) tilted 30°; *c*) tilted and rotated. The ghost areas are outlined in black on the duplicate pictures on the right. The effects are attributable to multiple inclastic scattering and single elastic scattering at the light and heavy components of the specimen respectively. Full explanation in text.

of the order of 2. The effect apparently occurs as the result of a superimposition of a defocussed image of multiply inelastically scattered electrons and a focussed image of elastically scattered electrons. The sharpness of the boundary can be seen as the result of the fact that the energy loss curve for a sequence of, say, 4 or 5 small-angle scattering processes will be considerably more peaked than the curve for a single inelastic scattering event, and the energy width of such groups of electrons must be very small. The effects just described may seem of a very subtle nature; they do, however, determine the lower limits of observation in practically all high-resolution work with shadow casting as the basic technique. A suggestion presenting itself as a logical consequence would be the use of lighter elements for shadow casting. It is known that the helical structure of the protein coat of tobacco mosaic virus can be hardly observed using the best of heavy-metal shadow casting. We could, however, observe that chromium shadowing did reveal this structure, even under the presence of considerable granularity. The difference with tungsten shadowing is striking; the latter shows a very small granulation but hardly any surface structure on the virus rods, Fig. 7. The problem of optimal shadow casting, however, does remain of a complex nature, as the metal used may adhere to the object at preferred sites, some sort of decoration effect taking place.



a) *b)*

Fig. 7. – Tobacco mosaic virus rods, about 200000×. Influence of material used for shadow casting: *a)* tungsten; *b)* chromium. Chromium gives a coarser granularity than tungsten, but is more effective than the latter in bringing out the fine structure of the helical protein coat.

3·5. Atomic injection.

A relatively recent new idea worth mentioning is applying the heavy staining component by atomic injection, as proposed by Manley ([37]). First experiments were done with cesium. A beam of cesium ions having an energy of between 50 and 800 eV was directed onto an unstained biological specimen inside a vacuum system; a density of between 1 and 3 atoms per square Ångstroem was obtained in a time varying from 2.5 to 7 minutes at a beam current of $15 \cdot 10^{-6}$ A/cm^2. As the author points out, the ions are probably neutralized to neutral atoms close to or at the surface by recombination with an electron, so the expression *atomic injection* seems well chosen. The final place of acceptance of the implanted atoms will depend on atomic and molecular properties of the specimen, as well as on the energy of injection, which latter can be varied at choice. The method evidently prevents surface migration and aggregation of the staining agent. Staining effects could be demonstrated, but a full evaluation of the method has not yet been made and any indication of its limits of usefulness would be conjectural.

3·6. Choice of electron optical parameters.

For a given specimen the contrast, as mentioned in Sect. **1**, depends on two electron optical parameters: the objective aperture and the accelerating voltage. Reducing the voltage drastically appears to be far more effective for increasing contrast than choosing a smaller aperture. The latter method suffers from the disadvantage that diffraction unsharpness is introduced when the aperture is substantially reduced beyond the optimum value, apart from the practical problem to keep a very small aperture sufficiently free from contamination. The possibilities of low-voltage microscopy, with voltages between 6 and 15 kV were investigated by Wilska ([38]) and by Van Dorsten and Premsela ([39]). High-contrast pictures of extremely thin specimens consisting of light elements, mainly carbon, were obtained at the cost of some loss of resolution, evidently caused by the presence of a large proportion of inelastically scattered electrons in the image. Real pioneer work to improve the low-voltage electron microscope was done by Wilska ([40]), who showed that the use of an electrostatic projector lens of special design, adjusted to act as a filter lens, could greatly improve picture quality by elimination of the inelastically scattered electrons. Trying to attain high resolution and energy filtering at the same time at moderate and low voltages does seem a highly interesting goal for future development.

All methods for increasing contrast in the typical low-density thin amorphous object discussed thus far suffer from limitations in connection with resolution or contrast, or both. They also do influence the beam-specimen interactions, sometimes making the interpretation of images uncertain. The subject of contrast improvement, however, is by no means exhausted if one looks into the last link in the communication chain, the part between the final electron image and the ultimate visual perception. With a given specimen and a given electron optical system, the electron density distribution in the final image is completely determined and the remaining possibilities of influencing contrast for the observer are lying entirely in the detector used and its adaptation to the human visual system. The use of some devices and methods in this final link of the chain will be briefly described and discussed.

3·7. Image conversion.

Converting the image on the fluorescent screen of an electron microscope into a video signal by means of a simple television camera and reproducing it on a television picture tube could be regarded as a technical curiosity offering the possibility of presenting electron microscope images simultaneously to larger groups of observers, until Haine and Einstein ([41]) pointed out that specially adapted pick-up devices could in principle improve the transfer of information from object to an observer, in particular when the luminance of the image was too low for good direct visual observation. The authors, making use of the effect of electron bombardment induced conductivity, showed that a target of amorphous selenium exposed to the image forming beam inside the microscope column, and scanned with a low-voltage beam as in a vidicon pick-up tube, could be used. In later years more practical solutions were developed, avoiding the use of delicate devices inside the poor vacuum of the projection chamber of the microscope, and enabling operation of the electron microscope at extremely low electron densities in the image.

One of these devices consists of an external plumbicon television pick-up tube optically coupled to a small transmission fluorescent screen inside the microscope column by means of a fiber optics window (van Dorsten, Broerse and Premsela ([42])). This system constitutes at present probably one of the best compromises between performance on the one hand, and complexity and cost of equipment on the other. A more sophisticated system is based on the use of the so-called SEC (secondary emission conductivity) tube; a

description and a full discussion of the observation methods involving a video chain is given by Kübler in this volume.

The new element in our discussion of contrast phenomena in the class of objects concerned is that, using a video technique, control of contrast becomes available. The electronic part of the television chain allows adjusting the number of brightness levels from black to white transmitted through the chain, and also, within limits, the total intensity range. The important aspects of image intensification with such systems will be left out of the discussion here.

Reducing the number of levels within a given range means increasing contrast at the cost of a decrease of the metric information mentioned in Sect. **1**, and a possible loss of structural information, depending on the properties of the object. It will be clear that contrast manipulation allows emphasizing certain features by suppressing other features. The process is basically irreversible, as lost information cannot be regained.

The user of the microscope, however, will not always be thinking in terms of information according to the definitions of information theory. He will be mostly concerned with the *meaning* of his observations in a particular field of application. His interest is in the *semantic information* as distinct from the *signal information* considered thus far. Making use of prior knowledge about the object, he may want a yes-or-no answer to a question regarding the presence of a certain feature, being prepared to accept the loss of information irrelevant for this restricted purpose. It will be obvious that such a destructive analysis of results has to be done in the last part of the communication chain, that is after the image formation. As mentioned in the discussion of statistical effects in Sect. **1**, the preponderance of small detail for interpretation is typical for microscopy, and manipulating of contrast exclusively in the smallest detail is a physical process approaching the mathematical process of differentiating, in other words the use of derivatives. The use of such mathematical operations was given the name image processing by Kovásznay and Joseph ([16]) and is also referred to as optical processing. Introduction of this principle must be regarded as a new departure in information retrieval from optical images.

If we regard an image simply as a scalar function of two independent variables, the place co-ordinates, this function $F(x, y)$ can be thought to be modified into a new function, $f(x, y)$, akin to the original one in the following way:

$$f(x, y) = (1 - c^2 \Delta_2) F(x, y) ,$$

with Δ_2 the two-dimensional Laplace operator and c^2 a constant to be chosen

as a measure for the required effect. Performing this operation means adding a certain amount of negative second derivative to the picture for each of its elements and results in *contour enhancement.* If there are good reasons to assume that the picture obtained from an imaging system has been degraded by impaired resolving power or loss of contrast, or both, the operation of contour enhancement can improve the image by partly restoring the original object function and a de-blurring effect can be obtained.

A further possible useful operation involves the first derivative. In this case the condition that the operation shall be independent of rotation requires only even functions to occur in the operator, the simplest of which is the square, yielding the operator $c(\nabla)^2$, with

$$(\nabla)^2 = \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2,$$

and c a constant. The resultant function $f(x, y) = c(\nabla)^2$ consists of the square of the absolute magnitude of the gradient vector for each point. This function is always either zero or positive, depending on whether the point considered is in an area of constant intensity or not.

The high values indicate high-intensity gradients, that is to say contours. The resulting image becomes a genuine line pattern if the function $f(x, y)$ is restricted to high threshold values by clipping. The operation thus performed is called *contour outlining.*

The principle of image processing just described in brief seems of great interest for electron microscopy, and especially for the observation of amplitude contrast in macromolecular objects. It is a real asset of a video system that it provides the image information in analog form as a time signal at any time during observation, without requiring a rapid access memory device, by virtue of the scanning system. Operations as mentioned in the foregoing can be performed entirely by electronic methods with specially designed circuitry.

Because for the macromolecular object, in order to match the resolution of the video system, the magnification has to be high, the information content of the image is relatively low, and the requirements on bandwidth therefore moderate. This can be directly related to the cost of the equipment.

A practical resolution limit of 5 Å can be attained using a simple video system equipped with a plumbicon camera, without any image processing, for high-contrast amplitude objects or phase objects.

The pictures for final observation can be photographed from a monitor tube and a good deal of image processing could be done on-line, with adapted

circuitry. Limitations imposed by the photographic process and the visual system of the observer (accurate focussing and watching and control of operating conditions) can, in principle, be reduced or removed.

A more ambitious scheme would consist in using digital recording of a considerably larger number of brightness levels than can be recorded by means of the photographic process or an analog signal, in combination with a strong nonlinear response, for instance a cube law, in order to obtain a significant increase of contrast in small detail.

Contour enhancement and contour outlining, as well as other operations as the case may be, can then be done by computer processing of the digital signal. For a comprehensive survey, see Mendelsohn *et al.* ([43]).

Image processing as a new technique of data processing and information retrieval has received the greatest possible stimulus from space research. A very considerable expenditure for development in this field has already resulted in basic and technical knowledge worth to be studied by workers in related fields. The method recently used for processing the observation data from the planet Mars, obtained by means of the Mariner spacecrafts, as described by Leighton and co-workers ([44]), are certainly worthy of the attention of electron microscopists.

In conclusion it can be said that, of the contrast enhancing procedures discussed for low-contrast objects in electron microscopy, low-voltage microscopy with energy filtering, and optical processing offer a better outlook than staining and shadow casting methods. The motto for progress could be: more contrast with less staining.

REFERENCES

The references listed below concern primarily original approaches and concepts new at their time, not latest developments or achievements, with exception of the last two references which are indicative of a state of the art.

1) L. BRILLOUIN: *Journ. Appl. Phys.*, **25**, 595 (1954).
2) A. C. VAN DORSTEN: *Proc. 2nd Eur. Reg. Conf. on Electron Microscopy, Delft 1960* (Delft, 1960), vol. **1**, p. 64.
3) C. W. ERIKSEN and H. W. HAKE: *Journ. Exp. Psychology*, **49**, 323 (1955).
4) H. FRIESER and E. KLEIN: *Zeits. angew. Phys.*, **10**, 337 (1958).
5) N. DIGBY, K. FIRTH and R. J. HERCOCK: *Journ. Phot. Sc.*, **1**, 194 (1953).
6) B. VON BORRIES: *Zeits. Naturfor.*, **9a**, 51 (1949).
7) S. LEISEGANG: *Zeits. Phys.*, **132**, 183 (1952).
8) F. LENZ: *Zeits. Naturfor.*, **9a**, 185 (1954).
9) R. E. BURGE and G. H. SMITH: *Proc. Phys. Soc.*, **79**, 673 (1962).

10) E. ZEITLER and G. F. BAHR: *Journ. Appl. Phys.*, **30**, 940 (1959).
11) E. ZEITLER and G. F. BAHR: *Laboratory Investigation*, **14**, 946 (1965).
12) A. C. VAN DORSTEN: *Laboratory Investigation*, **14**, 819 (1965).
13) A. COLE: *Proc. 27th Annual Meeting EMSA* (1969), p. 400.
14) J. G. HELMCKE: *Laboratory Investigation*, **14**, 933 (1965); *Optik*, **11**, 201 (1954); **12**, 253 (1955).
15) J. F. NANKIVELL: *Optik*, **20**, 171 (1963).
16) L. S. G. KOVÀSZNAY and H. M. JOSEPH: *Proc. I.R.E.*, **43**, 560 (1955).
17) H. BREMMER: *Physica*, **28**, 469 (1952).
18) H. BREMMER: *Symposium on Quasi-Optics*, Polytechnic Institute of Brooklyn (1964), p. 415.
19) H. BREMMER and A. C. VAN DORSTEN: *Zeits. angew. Phys.*, **27**, 219 (1969).
20) A. RUBINOWICZ: *Die Beugungswelle in der Kirchhoffschen Theorie der Beugung*, Warschau, Polnischer Verlag der Wissensch. (1957, 1966).
21) B. VON BORRIES and F. LENZ: *Proc. Ist. Eur. Reg. Conf. on Electron Microscopy, Stockholm 1956* (Amquist and Wiksell, 1957), p. 60,
22) F. THON: *Zeits. Naturfor.*, **20***a*, 154 (1965); **21***a*, 476 (1966).
23) K. J. HANSZEN: *Naturwiss.*, **54**, 125 (1967).
24) L. BRILLOUIN: *Ann. de Phys.* **17**, 103 (1921).
25) L. BRILLOUIN: *La diffraction de la lumière par des ultrasons*, Hermann (1933).
26) M. WATANABE: *Journ. Phys. Soc. Japan*, **17**, 569 (1962).
27) M. VON LAUE: *Naturwiss.*, **30**, 205 (1942).
28) A. C. VAN DORSTEN and H. F. PREMSELA: *Proc. of the Conf. on Electron Microscopy, Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 21.
29) E. PALADE: *Journ. Exp. Med.*, **95**, 285 (1952).
30) W. STOECKENIUS and S. C. MAHR: *Laboratory Investigation*, **14**, 1196 (1965).
31) C. E. HALL: *Journ. Biophys. Biochem. Cytol.*, **1**, 1 (1955).
32) S. BRENNER and R. W. HORNE: *Biochim. et Biophys. Acta*, **34**, 103 (1959).
33) G. MÜLLER and K. H. MEYERHOFF: *Nature*, **201**, 590 (1964).
34) W. LIPPERT: *Naturwiss.*, **51**, 408 (1964).
35) R. C. WILLIAMS and R. W. G. WYCKOFF: *Journ. Appl. Phys.*, **15**, 712 (1944).
36) N. NANNINGA: *Proc. Natl. Acad. Sci. U.S.*, **61**, 614 (1968).
37) J. H. MANLEY: *Science*, **154**, 424 (1966); **158**, 1585 (1967).
38) A. P. WILSKA: *Proc. 2nd Eur. Reg. Conf. on Electron Microscopy, Delft 1960* (Delft, 1960), vol. **1**, p. 165.
39) A. C. VAN DORSTEN and H. F. PREMSELA: *Proc. 2nd Eur. Reg. Conf. on Electron Microscopy, Delft 1960* (Delft, 1960), vol. **1**, p. 105.
40) A. P. WILSKA: *Laboratory Investigation*, **14**, 825 (1965).
41) M. E. HAINE and P. H. EINSTEIN: *Proc. 2nd Eur. Reg. Conf. on Electron Microscopy, Delft 1960* (Delf, 1960), vol. **1**, p. 96.
42) A. C. VAN DORSTEN, P. H. BROERSE and H. F. PREMSELA: *Proc. 6th Int. Conf. on Electron Microscopy, Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1**, p. 275.
43) M. L. MENDELSOHN, B. H. MAYALL, J. M. S. PREWITT, R. C. BOSTROM and W. G. HOLCOMB: *Advances in Optical and Electron Microscopy*, R. Barer and V. E. Cosslett eds, Academic Press (1968), vol. **2**, p. 77.
44) R. B. LEIGHTON, N. H. HOROWITZ, B. C. MURRAY, R. P. SHARP, A. G. HERRIMAN, A. T. YOUNG, B. A. SMITH, M. E. DAVIES and C. B. LEOVY: *Science*, **165**, 684 (1969); *Science*, **165**, 787 (1969).

Contrast Calculations for Small Clusters of Atoms

C. R. HALL

Cavendish Laboratory, University of Cambridge - Cambridge, England

1. Introduction.

The diffraction theory of electron microscope image formation can be used to predict the contrast due to any object, provided that the scattering factors of the atoms and their arrangement within the object is known. A number of papers have already appeared reporting calculations of image intensities and in some cases comparing them with experimental observations. The calculations of Eisenhandler and Siegel [1] however neglected the complex nature of the scattering factor, while in the case of the calculations of Heidenreich [2] a real scattering factor was again used and in addition there was some doubt about the exact structure of the graphite sample used for the experiments. Zeitler [3] on the other hand has shown that the imaginary part of the scattering factor is important in giving rise to contrast from heavier atoms near exact focus. Reimer [4] has recently calculated the image contrast due to single atoms and small clusters of atoms (up to 19 atoms) of platinum, using a complex scattering factor. The present calculations (most of them carried out as part of a project with Professor R. L. Hines while he was on leave in Cambridge 1968-69) are for clusters of gold atoms, which can easily be prepared for experimental observation by evaporation in vacuum. The experiments carried out by Professor Hines provided some support for the calculated image contrast (Hall and Hines [5]), and more recent experiments, providing better confirmation and which use the calculated intensities to obtain information about cluster thickness, are to be reported elsewhere

((Hines and Hall ([6])). The results described here are in general agreement with those of Reimer and should be typical of the contrast of materials in this part of the periodic table.

2. Outline of the method of calculation.

The atomic scattering factor for gold as a function of angle α was evaluated for 80 keV electrons by the standard phase shift method. This gives a scattering factor f with an amplitude $f(\alpha)$ and a phase $\eta(\alpha)$. The wave $\exp[2\pi i\boldsymbol{k}\cdot\boldsymbol{r}]$ scattered into the objective aperture at (α, φ) from an incident plane wave $\exp[2\pi i\boldsymbol{k}_0\cdot\boldsymbol{r}]$ by an atom at $\boldsymbol{r}_j$ in the object has an amplitude A given by:

$$A = f(\alpha)\exp[i\eta(\alpha)]\exp\left[2\pi i(\boldsymbol{k}_0-\boldsymbol{k})\cdot\boldsymbol{r}_j + \frac{i\pi\alpha^2}{\lambda}\Delta L - \frac{i\pi c_s\alpha^4}{2\lambda} + \right.$$
$$\left. + \frac{i\pi c_a\alpha^2}{2\lambda}\cos 2\varphi\right]. \qquad (1)$$

ΔL is the amount of defocus, and c_s and c_a are the spherical and astigmatic aberration coefficients respectively. For an object consisting of a single atom the amplitude $S(\varrho)$ at ϱ in the image due to these scattered waves is found by adding the contribution at ϱ from all parts of the aperture. Combining this with the amplitude of the unscattered wave gives a total image amplitude $1+S(\varrho)$: the image intensity is therefore given by $|1+S(\varrho)|^2$.

In principle the evaluation of $S(\varrho)$ involves a two-dimensional integration over the aperture for each point in the image. However if c_a is put equal to zero (no astigmatism) then for an aperture of semiangle α_0 and a single atom on the axis we have:

$$S(\varrho) = \frac{2\pi i}{\lambda}\int_0^{\alpha_0} f(\alpha)\,J_0\left(\frac{2\pi\alpha\varrho}{\lambda}\right)\exp\left[i\eta(\alpha) + \frac{i\pi\alpha^2}{\lambda}\Delta L - \frac{i\pi c_s\alpha^4}{2\lambda}\right]\alpha\,\mathrm{d}\alpha. \qquad (2)$$

This single integral can be evaluated on a computer much more rapidly than the previous double integral, and the resultant $S(\varrho)$ is circularly symmetrical. The total amplitude at $\boldsymbol{R}$ in the image plane due to atoms at $\boldsymbol{r}_j$ in the object

plane is now simply given by:

$$A = 1 + \sum_j S(|\boldsymbol{R} - \boldsymbol{r}_j|) \,. \tag{3}$$

(For convenience positions in the image are represented in object co-ordinates.) The image intensity is given by the square of the modulus of A as before.

Thus to compute the image contrast expected using a microscope with a particular c_s and aperture size α_0, $S(\varrho)$ is first found by eq. (2) as a function of ϱ for a range of values of ΔL, using the given values of α_0 and c_s. These values of $S(\varrho)$ are stored as two sets of tables within the computer (one for the real and one for the imaginary part). By interpolation eq. (3) can subsequently be evaluated for any given set of $\boldsymbol{r}_j$ and $\boldsymbol{R}$ for each of the defocus values for which $S(\varrho)$ is stored. Even for large numbers of atoms the evaluation of eq. (3) is rapid so that a large number of atomic configurations can be investigated once the initial calculations of $S(\varrho)$ have been carried out.

It is also possible to investigate, with the use of very little extra computer time, the effect of changing the phase of the scattered radiation by a constant amount relative to the unscattered radiation. This corresponds to the effect of a simple phase plate with a small hole in the middle. All that is necessary is to change the phase of $\sum_j S(|\boldsymbol{R} - \boldsymbol{r}_i|)$ before finding $|A|^2$. The treatment of a phase plate more complicated than this requires the addition of an extra phase term inside the integral in eq. (2), so that $S(\varrho)$ has to be evaluated for each phase plate geometry and strength. This results in a loss in generality and a considerable increase in computation time. For these calculations a value of 1.3 mm was used for c_s except where stated otherwise.

3. Results of calculations.

3·1. Single atom images.

By carrying out calculations for a range of values of α_0, the aperture size which gives the best calculated contrast for a single atom was found to be about 0.01 radians semiangle. This is approximately the size given by the criterion that the extra phase shift introduced at the edge of the aperture should not exceed $3\pi/2$.

Reducing the aperture below this value reduces the image contrast and

increases the image size, as would be expected. On the other hand increasing the aperture has very little effect upon the size of the image or its contrast. This is because the additional Fresnel zones included by the larger aperture have a relatively small area due to the α^4 variation in phase caused by the spherical aberration. Thus these zones contribute relatively little to the amplitude but tend to cancel each other out giving rise to minor oscillations in the image contrast and size as the aperture increases above the optimum.

The main results of the single atom contrast calculations are summarised in Fig. 1 *a*), where it can be seen that the best normal (dark) contrast, which

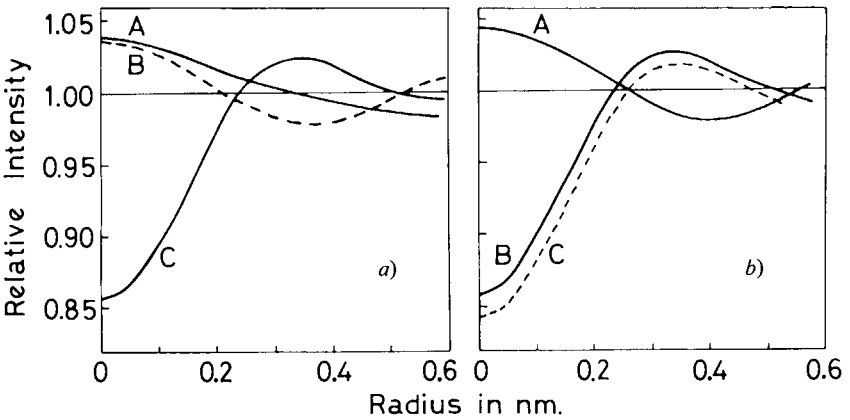


Fig. 1. – Radial intensity curves for a single atom using the optimum aperture. *a*) *A*, $\Delta L =$ $= -60$ nm; *B*, $\Delta L = 0$; *C*, $\Delta L = 80$ nm. *b*) *A*, $\Delta L = 0$, *f* real; *B*, $\Delta L = 80$ nm, *f* real; *C*, $\Delta L = 0$, with the phase of the scattered wave increased by $\pi/2$. (Courtesy of *Phil. Mag.*)

occurs at 80 nm defocus, corresponds to an intensity at the image centre 15% below background, and that the image diameter is about 0.4 nm. At exact focus there is some contrast, but the centre of the image is light, being some 4% above the background, and the image diameter is again about 0.4 nm. The maximum bright contrast occurs at a defocus of around -60 nm, when the contrast is still about 4%, but the image diameter is now about 0.6 nm. The effect upon the single atom image of using a real, rather than complex scattering factor is shown in curves *A* and *B* in Fig. 1 *b*), which corresponds to curves *B* and *C* respectively in Fig. 1 *a*). It is evident that the imaginary part of the scattering factor has little effect either upon the optimum contrast or upon the contrast at exact focus. Curve *C* in Fig. 1 *b*) is the optimum contrast found when the phase of the scattered wave is shifted by various amounts relative to the unscattered wave: it is obtained

at exact focus with a phase shift of $\pi/2$. It is seen that the contrast thus obtained is very little greater than that obtained under normal conditions at 80 nm of defocus. The reason for this can be understood from Fig. 2, which shows the total phase shift of the scattered wave across the aperture at 80 nm defocus. Because the defocus and the spherical aberration act in opposite senses the relative shift in phase of the scattered radiation across much of the aperture is of the order of $\pi/2$, the optimum value. Defocus and spherical aberration thus combine to give rise to fairly effective phase contrast. The reason for the small effect of the imaginary part of the scattering is that, as can be seen (it gives the phase shift at zero angle in Fig. 2),

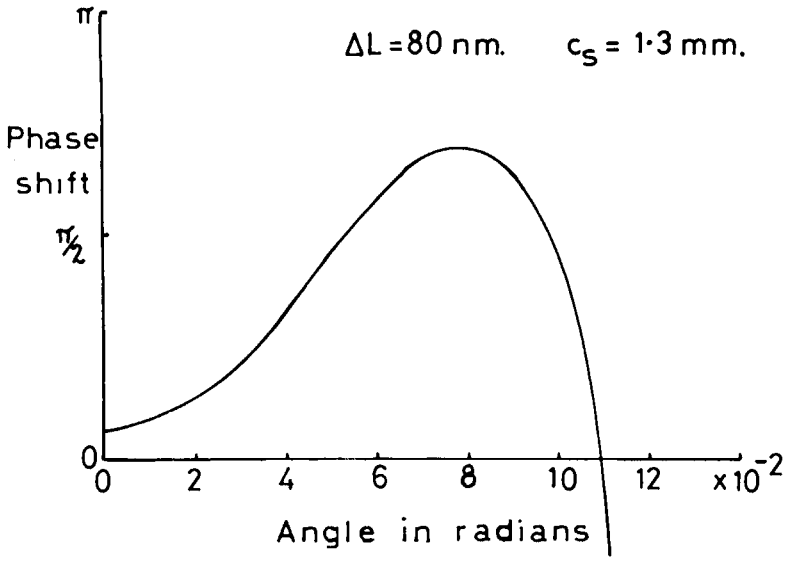


Fig. 2. – Phase across objective aperture of the wave scattered by a gold atom for 80 keV electrons, c_s = 1.3 mm and 80 nm defocus.

it contributes relatively little to the total change in phase. Thus the contrast at exact focus in Fig. 1 *a*) and *b*) is due mainly to spherical aberration rather than a complex scattering factor.

3·2. Pairs of atoms.

These calculations were carried out for pairs of gold atoms at a range of separations: a number of values of c_s were used and for each the corresponding optimum aperture size for single atom image contrast was used.

The resulting images are no longer circularly symmetric and a two-dimensional representation of the contrast is needed. This is most conveniently achieved by evaluating the contrast at a number of points on a grid by the method outlined previously and displaying the results on a cathode ray tube. By adjusting the controls of the tube the effective contrast of the image can be varied: in general the contrast of the images shown here is greater than would be expected from an image with the calculated contrast using normal photographic recording. In Fig. 3 simulated images for four different values of c_s are shown, together with the atomic spacing used in each case. For each c_s there are three images corresponding to the different amounts of defocus noted beneath each. These particular results were selected as showing the atom spacing at which the existence of two distinct images can still just be recognised. This spacing was found to be approximately proportional to $c_s^{-\frac{1}{4}}$, a result which applies also to self-luminous objects. This increase in effective resolution at smaller values of c_s is of course achieved with an in-

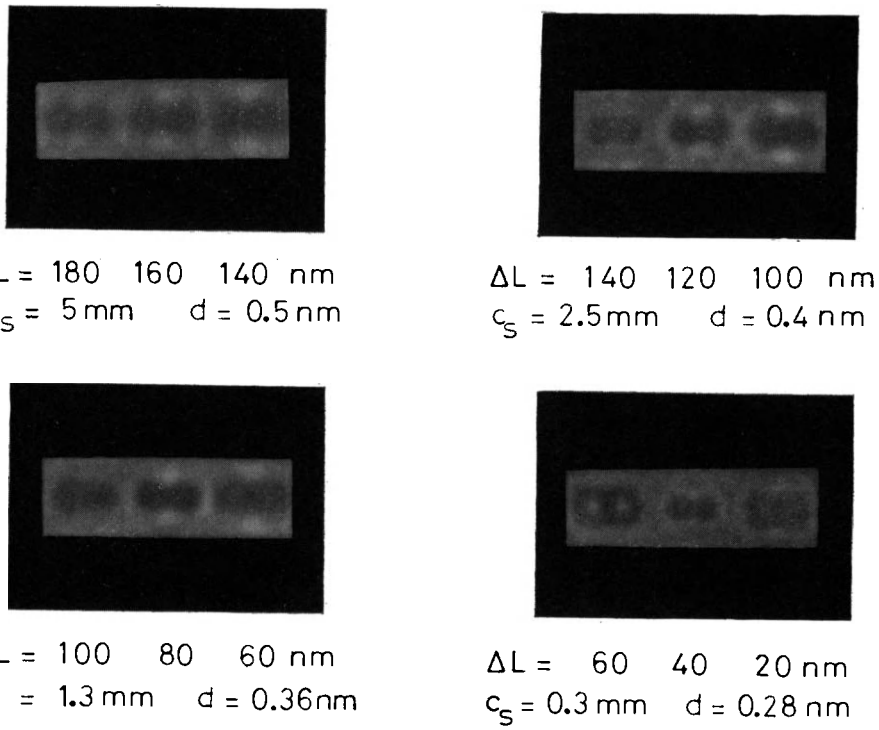


Fig. 3. – Cathode ray tube display of contrast due to two gold atoms at different separations d, defocus ΔL and c_s

creased aperture size, with the result that the contrast is more sensitive to defocus. Thus while in the upper pictures a change in focus of 20 nm on either side of the optimum has no great effect upon the contrast, in the images for $c_s = 0.3$ mm the contrast changes from dark to light in this distance.

Thus a microscope with this c_s and this potential resolution would have to be operated with some care in order to use the best degree of defocus, and it would have to be sufficiently stable to prevent the contrast being lost as a result of fluctuations in focal position.

3·3. Larger clusters of atoms.

Images of larger clusters of atoms will not in general be circularly symmetrical, and in principle it is necessary to calculate the contrast over a grid of points as before. However, this becomes fairly expensive in terms of computer time for the larger clusters, and it has proved possible in practice to get a good understanding of the contrast of the more symmetric islands and to study its variation with the different parameters by plotting the contrast along a single radial line. In each case the atoms in the single layer clusters have been taken to be close-packed with the arrangement and spacing (0.29 nm) found in the 111 planes in the solid. Subsequent layers have been added so as to give the face centred cubic structure.

In Fig. 4 *a*) the contrast along a line passing through one of the atoms in a three-atom cluster is shown. The curves shown are for the best light and dark contrast which is obtained by defocus, and the best dark contrast given by defocus and phase shift. As in the case of the single atom there is little contrast at exact focus, and the maximum contrast of each type arises at roughly the same defocus distances found for the single atom. Also as for the single atom the phase plate makes little difference, indicating that a considerable amount of phase contrast is again produced by the combined effects of spherical aberration and defocus. The best contrast corresponds to an intensity at the centre of the image which is 20% below background, which should be readily visible.

Calculations have also been made for a flat hexagonal island of seven atoms, and the remaining curves in Fig. 4 are for this configuration along a line passing between two outer atoms: this contrast is very similar in magnitude and behaviour to that along a line passing through an outer atom. The contrast at exact focus is again weak, the best bright contrast, 10% above background, occurring at −100 nm defocus. The image which would probably be interpreted as the best dark contrast is given by 80 nm of defocus

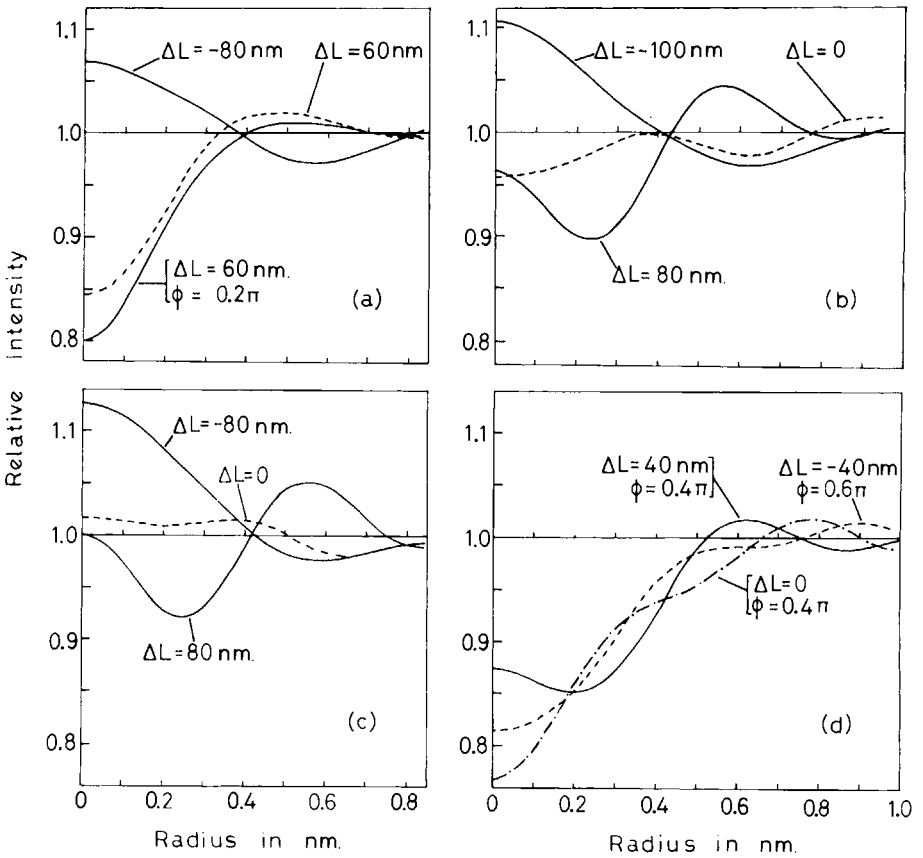


Fig. 4. – *a*) Radial intensity for 3 atoms; *b*) for 7 atoms; *c*) 7 atoms and real f; *d*) 7 atoms with phase shift. (Courtesy of *Phil. Mag.*)

and has a minimum intensity 10% below the background: in this case the image is roughly the same size as the object.

For this cluster the imaginary part of the scattering factor and a shift in phase of the scattered radiation are both more important than for a single atom. Figure 4 *c*) shows the effect of using a wholly real scattering factor. The contrast at exact focus is reduced, and the best light and dark contrast are both modified slightly as well. A phase plate is found to increase substantially the best dark contrast: curves for three combinations of phase shifts and defocus which give contrast near the optimum are shown in Fig. 4 *d*). This presumably arises because the scattering from the larger clusters is con-

centrated at smaller angles within the aperture, where defocus and spherical aberration have less effect, and where a phase plate can thus be important. An island of this size thus ought to be made more visible by a simple phase plate. The effect upon this contrast of reducing c_s and correspondingly increasing the aperture size is to increase the contrast but to cause it to oscillate more rapidly with defocus (Reimer ([4])), as with the pairs of atoms in Fig. 3. The results for larger clusters show a similar trend. Figures 5 and 6

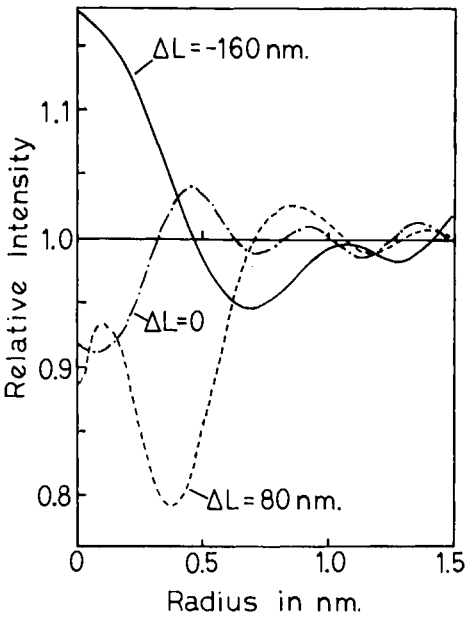


Fig. 5. – Radial contrast from a cluster of 13 atoms with 6 atoms on top. (Courtesy of *Phil. Mag.*)

show the intensity along a radial line for a cluster of 19 atoms (13 + 6) and of 64 atoms (37 + 27) respectively. The image at exact focus of the 19 atom island is smaller than the object, while the optimum contrast of about 20% occurs at 80 nm of defocus, when the object and image sizes are similar. Reversed contrast again occurs on the other side of focus, being at best 15% above background at — 160 nm defocus. For the largest cluster, radius about 1 nm, the contrast oscillates as the focus is changed, the form of the image frequently having little resemblance to the object. Dark and light centre contrast is found on both sides of focus, the centre being surrounded by rings of alternating but weaker contrast. The cluster

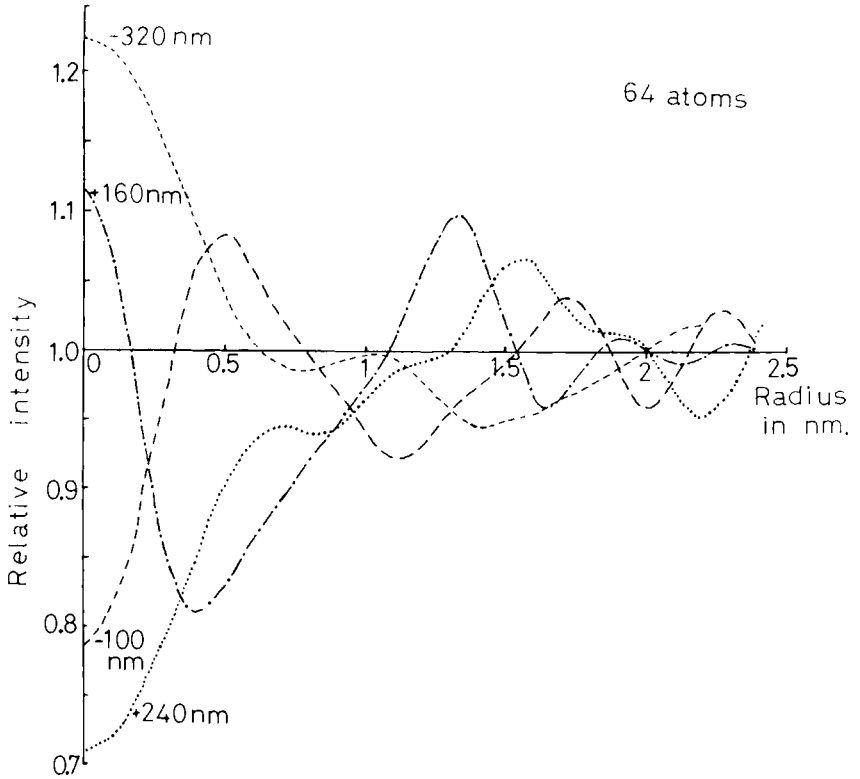


Fig. 6. – Radial contrast from a cluster of 37 atoms with 27 atoms on top.

in these latter calculations is roughly similar in size to the gold islands in the micrographs in Fig. 7 (*e.g.* island *A*). It can be seen that the « bullseye » contrast of some of the islands at different amounts of defocus is similar in form to that plotted in Fig. 6. It is not possible to attempt a detailed comparison as the calculations have not been taken as far as the nominal amount of defocus used to take the micrographs, the thickness of the clusters is not known and the aperture was somewhat larger than the optimum.

A general impression of the way in which the contrast of small islands will vary with defocus is given by Fig. 8, which shows the intensity at the centre of the 3, 7 and 19 atom clusters over the focal range ± 300 nm. These curves are slightly misleading since the maximum intensity change at the very centre of the image may not correspond to maximum overall contrast. Thus from Fig. 8 it might be inferred that 140 nm of defocus would give maximum

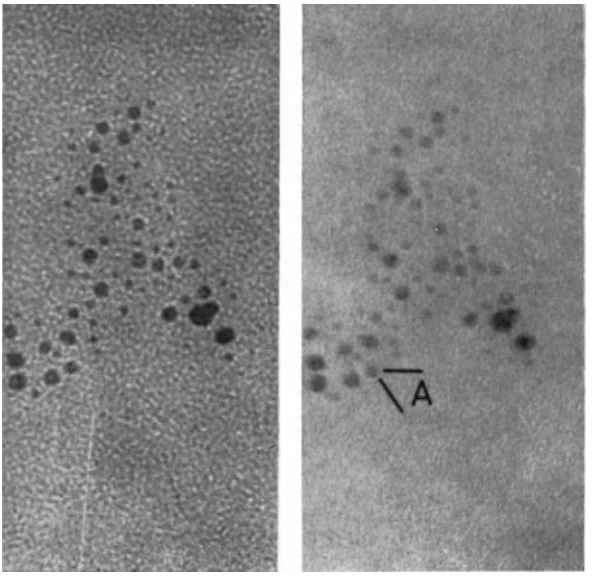
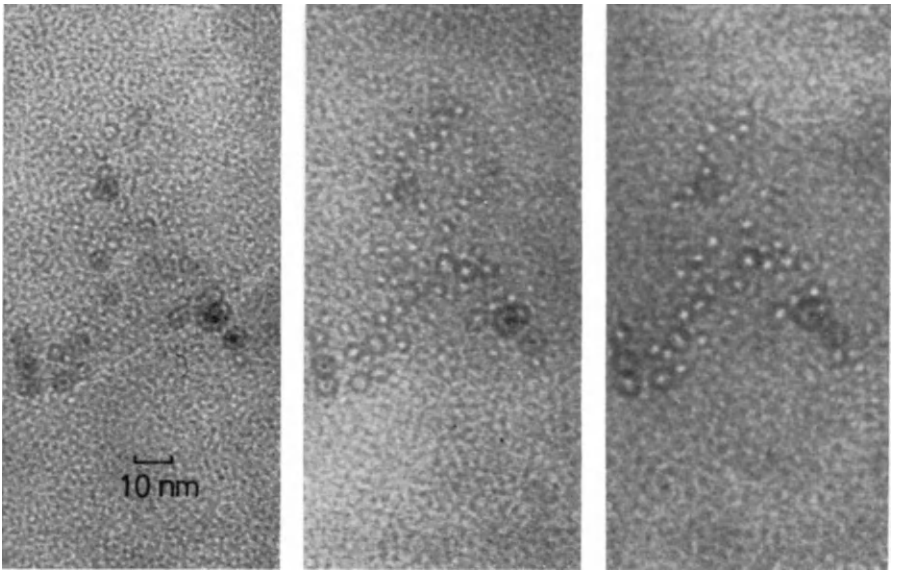


Fig. 7. – Micrographs of a group of gold islands at a number of different amounts of defocus. (Courtesy of *Phil. Mag.*)



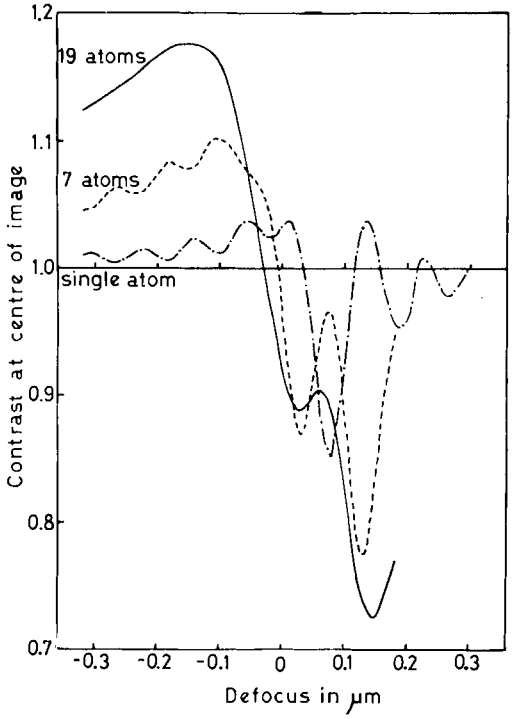


Fig. 8. – Contrast of the centre of a cluster as a function of defocus. (Courtesy of *Phil. Mag.*)

island visibility for the 19 atom cluster, when in fact the image is smaller than the object and would also be less visible than the image at 80 nm of defocus. However the defocus for best contrast is fairly close to the value obtained from Fig. 8.

3·4. Effect of astigmatism.

The effect of astigmatism is to introduce an extra phase change in the scattered wave of the form $(i\pi/2\lambda)\, c_a \alpha^2 \cos 2\varphi$ (see eq. (1)). As a consequence the amplitude in the objective aperture is no longer circularly symmetrical, and strictly a two-dimensional rather than a one-dimensional numerical integration over the aperture is required for each value of c_a and each defocus. However a good impression of the effect of astigmatism upon

image contrast can be obtained without the use of a lot of computer time by using an approximate formula given by Born and Wolf ([7]). This formula can be obtained more straightforwardly by expanding the factor

$$\exp\left[\frac{i\pi}{2\lambda} c_a \alpha^2 \cos 2\varphi\right]$$

in eq. (1) to become

$$1 + \frac{i\pi}{2\lambda} c_a \alpha^2 \cos 2\varphi \,.$$

The first term simply leads to the amplitude scattered in the absence of astigmatism. The second gives rise to an astigmatism amplitude which can be integrated analytically over the angular co-ordinate in the objective aperture to leave a single radial integration for the computer as before. The extra contribution to $S(\varrho)$ in eq. (2) is then:

$$\frac{2\pi^2 c_a}{\lambda^2} \cos 2\psi \int_0^{\alpha_0} f(\alpha) J_2\left(\frac{2\pi\alpha\varrho}{\lambda}\right) \exp\left[i\eta(\alpha) + \frac{i\pi\alpha^2}{\lambda}\Delta L - \frac{i\pi c_s \alpha^4}{2\lambda}\right] \alpha^3 \,\mathrm{d}\alpha \,. \qquad (4)$$

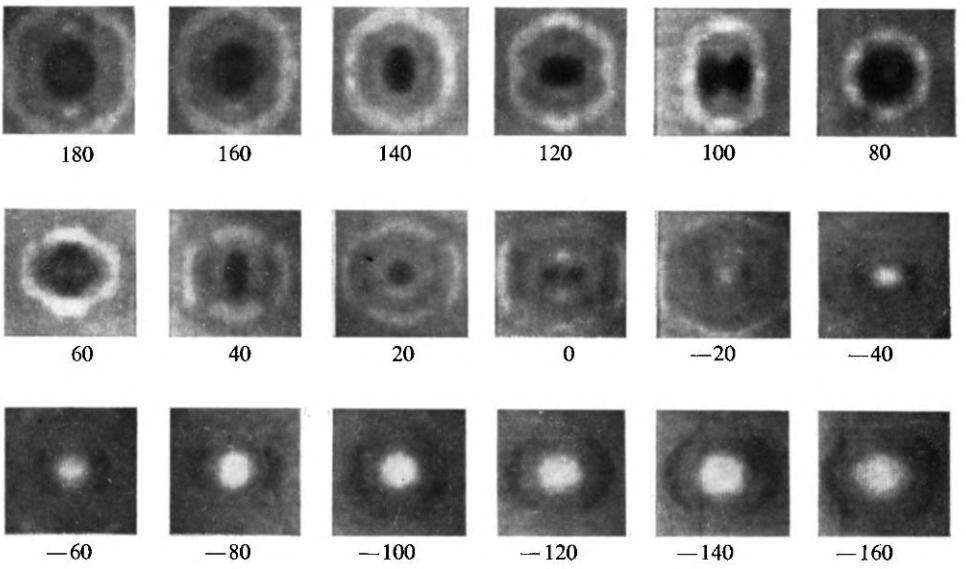


Fig. 9. – Calculated contrast of a seven atom cluster at the amounts of defocus indicated in nm including the effects of spherical aberration and astigmatism.

This amplitude contains a term $\cos 2\psi$, where ψ is the angle between ϱ, the radial direction in the single atom image, and the direction from which φ is measured. Consequently the single atom image is no longer circularly symmetrical. It can also be seen that this extra amplitude is proportional to c_a, so that it can be evaluated with $c_a = 1$ and stored as before, the interpolated number being multiplied by the particular value of c_a of interest when it is combined with the original amplitude to give the total amplitude. The higher terms neglected in the expansion give amplitude contributions with higher symmetry and therefore tend to give contrast which is more nearly circularly symmetrical. The image of a cluster of atoms is found by adding the amplitudes in the image plane as before.

As an illustration of the effects which might be expected under extreme conditions, the intensity of a hexagonal cluster of 7 atoms with $c_a = 42$ nm, displayed in the same way as Fig. 3, is shown in Fig. 9. On a small scale the astigmatism has an appreciable effect on the form of the image, so that for example at exact focus the cluster looks like a pair of small dots at a separation of 0.6 nm.

4. Conclusions.

a) The contrast from a small object varies considerably with defocus: in order to interpret fine detail on micrographs correctly it is necessary to take a through-focus series of exposures to obtain an adequate amount of information. Single micrographs taken at an unknown defocus can be misleading.

b) For very small objects it seems that a simple phase plate will not give much in the way of extra visibility: for coarser structures however the technique could still be useful.

c) The calculations suggest that it should be possible, using existing microscopes, to see single atoms as well as small clusters: 10% of contrast should be visible. The problems which need to be overcome seem to be:

i) Substrate preparation: the contrast from the graphite substrate visible in Fig. 7 obscures the smaller clusters. A single crystal substrate, as opposed to an amorphous film, should not give any contrast if the Bragg reflected beams are excluded by the aperture. The contrast in Fig. 7 is prob-

ably due to surface contamination introduced during the substrate preparation prior to the island deposition.

ii) Mechanical stability needs to be extremely good if very fine detail is not to be blurred out. In addition the electrical stability has to be higher than that of many microscopes at the moment: the stability required depends upon c_s and upon the aperture size, but for $c_s = 1.3$ mm and the optimum aperture of 0.01 rad and thus a maximum desirable ripple in ΔL of say 10 nm, with an objective focal length of 2 mm, a stability of better than about 3 in 10^6 is needed on both the lens currents and the accelerating voltage.

iii) Finally, electron noise in the beam which reaches the recording plate gives rise to spurious contrast on a small scale, the smaller the scale the greater the contrast. (For a review see Valentine ([8])). At the maximum magnification given by most microscopes (10^5 or a little more) this contrast becomes comparable with that expected ($\sim 10\,\%$) from atoms and small clusters on a scale roughly equal to the image size expected from a group of a few atoms. This will need to be overcome either by increasing the magnification or by using less sensitive plates which require more electrons per unit area to produce an image.

Acknowledgements.

I am grateful to Professor R. L. Hines for many stimulating discussions, to the Director of the Mathematical Laboratory for computing time and to M. S. Spring for running data with his imaging program.

REFERENCES

1) C. B. Eisenhandler and B. M. Siegel: *Journ. Appl. Phys.*, **37**, 1613 (1966).
2) R. D. Heidenreich: *Journ. Electron Microscopy*, **16**, 23 (1967).
3) E. Zeitler: *Proc. 6th Int. Conf. of Electron Microscopy, Kyoto 1966* (Maruzen, Tokyo, 1966), vol. **1** p. 43.
4) L. Reimer: *Zeits. Naturfor.*, **24***a*, 377 (1969).
5) C. R. Hall and R. L. Hines: *Phil. Mag.*, **21**, 1175 (1970).
6) R. L. Hines and C. R. Hall: *Proc. 7th Int. Conf. on Elecron Microscopy, Grenoble 1970* (Paris, 1970) vol. **1**, p. 33.
7) M. Born and E. Wolf: *Principles of Optics*, Pergamon Press (1965), p. 474.
8) R. C. Valentine: *Advances in Optical and Electron Microscopy*, Academic Press (1966), vol. **1**, p. 180.

Some Aspects of Lorentz Microscopy

R. H. WADE

Centre d'Etudes Nucleaires - Grenoble, France

Magnetic-field distributions are usually imaged in transmission electron microscopy by off-focus techniques. In the first of these three lectures an attempt is made to situate Lorentz microscopy in the general background of phase microscopy. The second lecture deals with the interpretation of the off-focus images in trying to define the conditions under which a geometrical approach gives a good approximation to wave optics. The experimental attempts to solve the domain wall problem are reviewed in the third lecture in which we indicate why and where these attempts fail.

The field of Lorentz microscopy still awaits its full justification which will come when domain wall, ripple and stripe domain structures have been fully solved. This will be achieved only after careful consideration of the optimum working condition. We hope that this series of lectures taken with those of Wohlleben may contribute to form the foundations of this future success, or if the worst comes to the worst to define clearly where the failure lies.

1. Introduction to Lorentz microscopy.

1·1. Image contrast in phase microscopy.

The action of a transmitting object on an incident plane wave can be described by a transmission function $f(x, y)$ where

$$f(x, y) = V(x, y)/V_0(x, y) .$$

V_0 and V represent respectively the wave functions in a reference (x, y) plane in the absence and presence of the object. For our purposes the incident electron illumination is the plane wave $\exp[ikz]$ directed along the z-axis incident on a thin object situated in the $(x_c, y_c, 0)$ plane. The disturbance on the exit surface of the object is then given by V. The quantities $|f|$ and $\arg(f)$ are the amplitude and phase of the object transmission function $f(x_c, y_c)$.

In the case of a one-dimensional periodic object $f(x_c) = f(x_c + \varepsilon)$ we can rearrange the Fresnel integral to give the disturbance $\psi(x, z)$ in a plane at distance z below the object in the form

$$\psi(x, z) = \exp\left[\frac{i\pi}{4}\right] \sum A_n \exp\left[i2\pi x \frac{n}{\varepsilon}\right] \cdot \exp\left[-i2\pi\lambda z \frac{n^2}{2\varepsilon^2}\right], \tag{1}$$

where the Fourier series expansion of the object function is:

$$f(x_c) = \sum A_n \exp[i2\pi x_c n/\varepsilon] .$$

and

$$A_n = \frac{1}{\varepsilon} \int_{-\varepsilon/2}^{+\varepsilon/2} \mathrm{d}x_c f(x_c) \exp[-i2\pi x_c n/\varepsilon] .$$

All of classical microscopy of a periodic object is contained in eq. (1). We find a reproduction of the object function $f(x)$ in any defocussing plane satisfying the condition $z = 2m\varepsilon^2/\lambda$ where m is an integer and $m = 0$ corresponds to the in-focus image. For a pure phase object $f(x_c) = \exp[i\varphi(x_c)]$, no contrast is visible in these planes. In light microscopy various techniques have been developed to render such objects visible. These methods utilize the fact that the diffraction image formed at the focal plane of an imaging lens corresponds to the object spectrum A_n. Any manipulation in the focal plane, F of Fig. 1, altering the amplitudes A_n will produce a corresponding change in the image.

As an example we consider a weak unidimensional symmetrical phase object of periodicity ε.

$$f_1(x_c) = \exp[i\varphi(x_c)] = \sum_{-\infty}^{+\infty} c_n \exp[i2\pi n x_c/\varepsilon] \simeq 1 + 2i \sum_{1}^{\infty} b_n \cos 2\pi n x_c/\varepsilon . \tag{2}$$

The image intensity $|f_1|^2$ is unity. In the Schlieren method of introducing an image contrast all spectra on one side of the zero order are blocked off

by the screen D_1. The image function becomes

$$f_2(x) = 1 + i\sum_{1}^{\infty} b_n[\cos(2\pi n x_c/\varepsilon) + i\sin(2\pi n x_c/\varepsilon)]\,,$$

$$|f_2|^2 \simeq 1 - 2\sum_{1}^{\infty} b_n \sin(2\pi n x_c/\varepsilon)\,.$$

The effect of the screen D_1 in the back focal plane F has been to introduce an intensity distribution in the image plane I proportional to the derivative of the object phase function $\varphi(x_c)$.

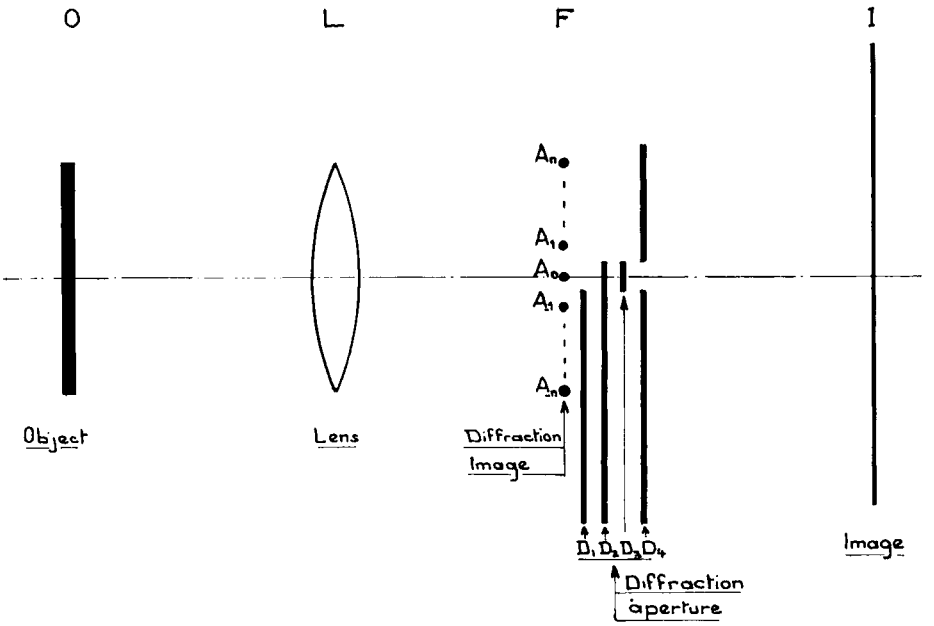


Fig. 1. – A periodic phase object O is imaged at I by the lens L. The diffraction image formed at the focal plane F shows discrete maxima with amplitudes A_n corresponding to the appropriate Fourier components of the object spectrum. The diffraction apertures D_1, D_2, D_3, D_4, can be used to introduce contrast into the final image.

Other dispositions of the screen are possible. Those in use in classical optics are shown in the Fig. 1. The screen position D_2 intercepts the zero-order maximum and one side of the diffraction image (oblique dark field); the screen D_3 intercepts only the central maximum (dark field). The bright-

field method commonly used in electron microscopy (D_4) allows only the central maximum to pass into the image. No resolution is possible, only local variations of the zero-order diffraction maximum are detected.

The phase contrast method first used (Zernike (1942)) is of considerable interest in that it produces an intensity fluctuation in the image directly proportional to the phase structure $\varphi(x_c)$ of the object. This method makes use of eq. (2) in which the phase of the constant term, or rather the corresponding zero order term in the object spectrum, is modified by an advance of $\pi/2$.

$$|f_3|^2 \simeq 1 + 4 \sum_{1}^{\infty} b_n \cos(2\pi n x_c/\varepsilon) .$$

The contrast is increased if the quarter wave plate used to effect the phase change is also partially absorbing.

The diffraction spectrum of an isolated or a nonperiodic object, consists of a continuous amplitude distribution rather than of the discrete maxima associated with a periodic object. Figure 2 shows schematically the diffraction amplitudes given by these different objects. A screen AB can be placed between the discrete maxima of the periodic object Fig. 2*a*) so as to exclude a certain number of terms A_n from the image function. In the cases *b*) and *c*) the edge B of the screen intercepts a non-zero intensity in the diffraction image. The image function will contain a contribution due to diffraction from the aperture edge.

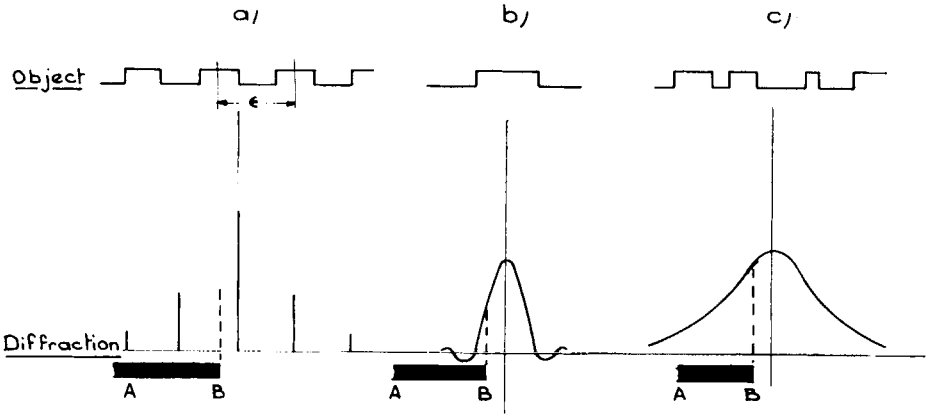


Fig. 2. – Showing *a*) a periodic object, *b*) an isolated object and *c*) a nonperiodic object. Below each is shown the corresponding diffraction image. In *a*) the aperture AB placed between the discrete maxima does not itself contribute to the final image. In the other cases the edge B is illuminated and this will modify the image intensity distribution.

1·2. Abbe theory of image formation.

It is useful at this point to present briefly the Abbe theory of image formation; this theory considers that a lens forms an image of an object by a two stage process. The formation of the diffraction image at the back focal plane F of the lens L (Fig. 1) is considered to be stage one. The diffraction amplitude $f(s)$ is related to the object function $f(x_c)$ by:

$$f(s) = \int_{-\infty}^{+\infty} f(x_c) \exp [i2\pi s x_c] \, \mathrm{d}x_c \,, \tag{3}$$

where we have considered the lens to have an infinite aperture. The final image $f(x')$ is associated to the diffraction image by the relation

$$f(x') = \int_{-\infty}^{+\infty} f(s) \exp [-i2\pi s x'] \, \mathrm{d}s \,, \tag{4}$$

which forms the second stage of the imaging process. The relations (3) and (4) can be demonstrated directly by successive applications of the Kirchhoff-Fresnel diffraction integral to the initial object function $f(x)$ using a function of the form $\exp [-i\pi u^2/\lambda f]$ to represent the action of the convex imaging lens. The expressions (3) and (4) above associate the image function $f(x')$ to the object function $f(x)$ by a double Fourier transformation. Since such a transformation is known to reproduce the initial function $f(x)$ we can write $f(x') = f(xM)$ where M is the magnification of the system.

The relations (3) and (4) show that a lens has the double property of displaying a harmonic analysis of an object in its focal plane F and of recombining this spectrum to form an image in the plane I. The action of an aperture in the focal plane F has its mathematical equivalent in an appropriate limitation of the range of the integral (4), or, in the case of a periodic object, of the summation (1).

In classical light optics phase objects are usually rendered visible by using the kind of manipulation which we have described in Sect. 1·1. The best amongst these techniques is the phase contrast method since it reproduces in intensity the phase structure of the object. As we shall presently see these techniques are seldom applicable to the case of a magnetic object imaged by electron microscopy.

1·3. Magnetic object as a phase object for electrons.

The role of a magnetic object in electron optics can be introduced via the concept of refractive index. We consider an electron wave interacting with a static electromagnetic field defined by the vector potential $\boldsymbol{A}$ and the scalar potential V; the magnetic and electrical fields $\boldsymbol{B}$ and $\boldsymbol{E}$ are defined by the relations $\boldsymbol{B} = \operatorname{curl} \boldsymbol{A}$ and $\boldsymbol{E} = \operatorname{grad} V$.

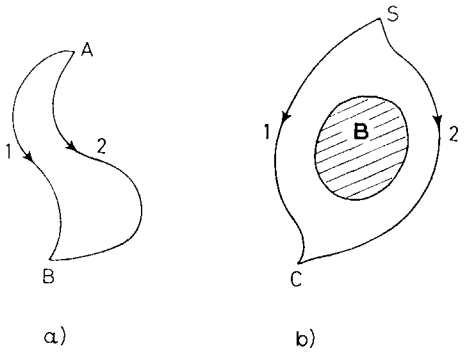


Fig. 3. – *a*) Illustrating the application of Fermat's principle to the passage of a ray between A and B. Amongst the possible paths are those marked 1 and 2. *b*) The two rays emitted by a source at S follow the paths 1 and 2 coming together to interfere at C. The enclosed magnetic field $\boldsymbol{B}$ influences the interference by the phase difference which it produces between the two rays.

We make use of the analogy between Fermat's principle of geometrical optics and the classical mechanics Maupertuis' principle. Fermat's principle states that the actual trajectory taken by a light ray between a point A and a point B, Fig. 3*a*), is such that the line integral $\int_B^A n \cdot \mathrm{d}r$ taken along the possible trajectories between A and B is minimum. Maupertuis' principle asserts that the path taken by a particle between the points A and B is such to minimise the action integral $\int_B^A \boldsymbol{p} \cdot \mathrm{d}\boldsymbol{r}$ taken along all possible paths; the momentum vector $\boldsymbol{p}$ in the presence of the vector potential $\boldsymbol{A}$ is not co-linear with the velocity $\boldsymbol{v}$

$$\boldsymbol{p} = m\boldsymbol{v} - e \cdot \boldsymbol{A}$$

(e is the electron charge). In what follows we limit our interest to the influence of the vector potential A. Consequently we put $V =$ constant. Since $v = \sqrt{2e(E_0 + V)/m}$, where E_0 is the accelerating potential, we can write the action integral in the form

$$\int_B^A \boldsymbol{p}\cdot \mathrm{d}\boldsymbol{r} = \int_B^A \mathrm{d}r\cdot\left[\sqrt{2me(E_0 + V)} - e\boldsymbol{A}\cdot\boldsymbol{u}\right],$$

where $\boldsymbol{u} = \boldsymbol{v}/v$.

The phase of a wave at B taken with respect to the phase at the point A, is given according to Fermat's principle by the integral $2\pi/\lambda\int_B^A n\cdot \mathrm{d}\boldsymbol{r}$ taken along the ray connecting A and B. The formal analogy between the two afore-mentioned principles allows a de Broglie wave to be associated with the movement of a particle between the points A and B which will be separated by the phase delay $2\pi/h\int_B^A \boldsymbol{p}\cdot \mathrm{d}\boldsymbol{r}$, where h is Plank's constant. One can consider that the medium between A and B possesses an anisotropic refractive index

$$n = \sqrt{1 + \frac{V}{E_0}} - (\lambda e/h)\cdot\boldsymbol{A}\cdot\boldsymbol{u}.$$

We now consider the situation schematised in Fig. 3*b*) in which two coherent electron beams issuing from a source S, come together to interfere at C after following the different trajectories 1 and 2 in field free space. The interference at C is determined by the phase difference $\Delta\varphi$ between the two beams which expressed in terms of the refractive index is

$$\Delta\varphi = \frac{2\pi}{\lambda}\left[\int_1 n\cdot \mathrm{d}r - \int_2 n\cdot \mathrm{d}r\right] = \frac{2\pi}{\lambda}\cdot\Delta r - 2\pi\cdot\frac{e}{h}\int \boldsymbol{B}\cdot \mathrm{d}\boldsymbol{S}.$$

Δr is the path length difference between the trajectories 1 and 2 whilst the second term is proportional to the flux Φ enclosed between the two trajectories. It is this term which is responsible for the action of a magnetic field region as a phase object for an electron beam.

In the case of a plane wave incident on a magnetic film of thickness a situated in the $(x_c, y_c, 0)$ plane, the phase difference φ_s between two points

$x_c^{(1)}$ and $x_c^{(2)}$ (y_c constant) is given by:

$$\varphi_s = \frac{e}{\hbar} \cdot a \int_{x_c^{(1)}}^{x_c^{(2)}} B_y(x_c)\, \mathrm{d}x_c \,. \tag{5}$$

The enclosed flux rule expressed by (5) enables us to construct the phase representation of a domain wall and of a periodic object. The flux density distributions of these two objects are shown in the upper part of Fig. 4. The

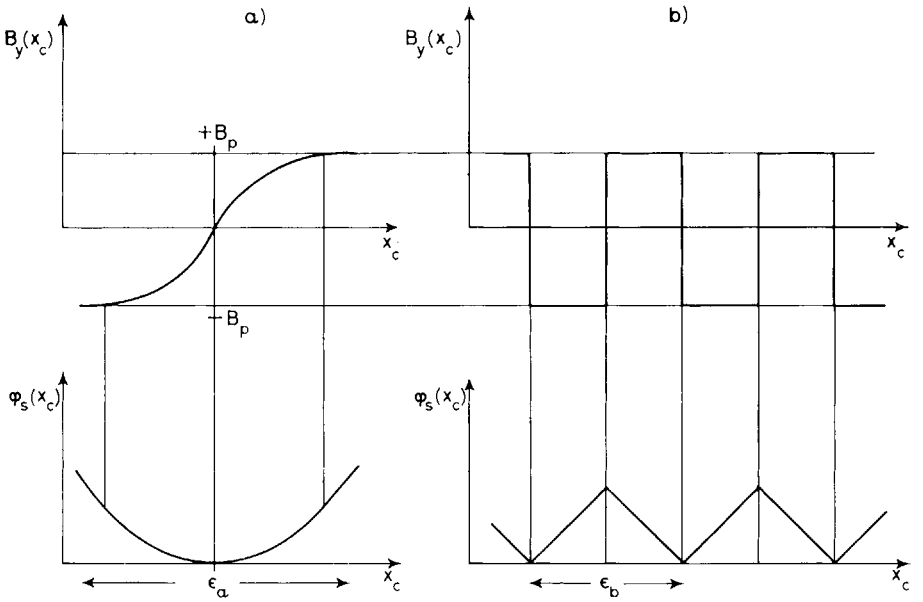


Fig. 4. – The one dimensional field distributions across a domain wall and a periodic object are represented in *a*) and *b*) respectively. Below are shown the corresponding phase representations constructed according to the enclosed flux rule (5).

phase has in both cases the symmetry property $\varphi_s(x_c) = \varphi_s(-x_c)$. The Fourier transform relation (3) can be used to show that for the phase object $\exp[i2\pi\varphi_s(x_c)]$ the real and imaginary parts of the diffracted amplitudes are also symmetric. They have in fact the form shown in Fig. 5.

We remark that no spatial separation between the real and imaginary diffraction amplitudes occurs in the example shown in the figure which cor-

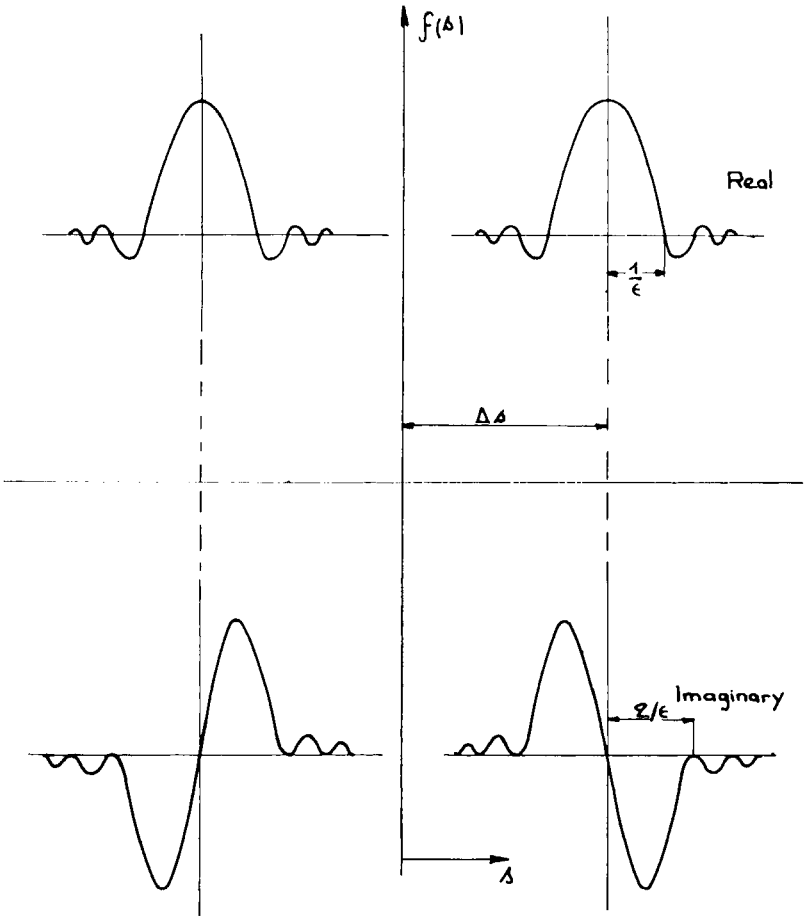


Fig. 5. – Diffraction amplitude envelopes of phase objects of the types shown in Fig. 4 which possess the symmetry property $\varphi_s(x_c) = \varphi_s(-x_c)$.

responds to the case of a strong phase object. Zernike phase contrast cannot be applied to such an object. It remains possible by the use of the diffraction aperture to make a Schlieren type observation. However as Wohlleben shows (1967) the image intensity distribution is extremely sensitive to the aperture position and for a nonperiodic object the intensity is also perturbed by diffraction from the edge of the aperture.

When $\Delta s \ll \varepsilon$ (see Fig. 5) the zero-order maximum of a periodic object is much enhanced with respect to the other diffracted maxima. The object

function can be written in the form $f_1(x_c)$ of eq. (2). This then is the case of a weak phase object. The diffracted amplitude of a nonperiodic object has a continuous lateral spread. An illumination source of finite width causes the direct beam ($s = 0$) to overlap the diffracted beams ($s \neq 0$) so that even a very narrow phase plate will not act separately on the direct and diffracted amplitudes. The phase plate itself gives a diffracted intensity in the final image plane. In addition there are severe practical difficulties associated with the fine scale of the magnetic scattering. In brief it seems difficult to apply this method to the types of magnetic object with which we are mainly concerned here.

In what follows we limit our attention almost exclusively to the out of focus method of introducing image contrast. This method is scarcely used in classical optical microscopy since in general no simple relation exists between the image intensity distribution and the object phase function. A mere casual glance at the images of periodic objects published by Fert and his collaborators (1961) offers a convincing proof of this fact. The method remains however for lack of something better the method most exploited in the so-called Lorentz microscopy which might be regarded by an unkind classical microscopist as poor man's phase microscopy.

2. The validity of geometrical optics.

2·1. The relationship between wave and geometrical optics.

The Kirchhoff-Fresnel diffraction integral, which represents the wave optical diffraction amplitude at a distance $z = d$ from an object, can be written in the form

$$\psi(x_d) = \frac{1}{\sqrt{\lambda d}} \int_{-\infty}^{+\infty} \mathrm{d}x_c \exp\left[i 2\pi \varphi_c(x_d, x_c)\right], \tag{6}$$

where in the case of Fresnel diffraction by a magnetic object the phase φ_c is given by

$$\varphi_c(x_d, x_c) = \frac{(x_c - x_d)^2}{2\lambda d} + \varphi_s(x_c) \tag{6a}$$

whilst for Fraunhofer diffraction

$$\varphi_c(x_d, x_c) = \frac{x_d \cdot x_c}{f\lambda} + \varphi_s(x_c)\,, \tag{6b}$$

f being the focal length of the imaging lens. In the above $\varphi_s(x_c)$ is given by eq. (5). Alternatively $\varphi_s(x_c)$ can be written $\varphi_s(x_c) = \int_{x_c^{(0)}}^{x_c^{(1)}} \mathrm{d}x_c \alpha(x_c)/\lambda$ where $\alpha(x_c)$ is the local deflection suffered by the electron beam. In the case of objects of the type shown in Fig. 4 we can write

$$\varphi_s(x_c) = \frac{\alpha_p}{\lambda} \int_{x_c^{(0)}}^{x_c^{(1)}} \mathrm{d}x_c \cdot f(x_c)\,. \tag{7}$$

α_p is the maximum deflection suffered by the electron beam and $f(x_c) = \alpha(x_c)/\alpha_p$ can be regarded as the normalised local deflection. In the case of the domain wall, Fig. 4*a*), we take the origin of co-ordinates at the wall centre.

The principle contribution to the integral (6) comes from the region or regions in the neighbourhood of the stationary phase points x_i for which $\partial\varphi_c/\partial x_c = 0$. The Taylor expansion of φ_c around x_i up to the second order is given by:

$$\varphi_c(x_d, x_c) = \varphi_c(x_d, x_i) + \Delta x_c \varphi_c'(x_d, x_i) + \frac{(\Delta x_c)^2}{2} \varphi''(x_d, x_i) + \ldots\,, \tag{8}$$

where $\Delta x_c = x_c - x_i$. Since the second term on the r.h.s. of eq. (8) is zero the stationary phase approximation to the integral (6) consists of substituting therein

$$\varphi_c(x_d, x_i) + \frac{(\Delta x_c)^2}{2} \varphi_c''(x_d, x_i)$$

for $\varphi_c(x_d, x_c)$. This yields

$$\psi_{\mathrm{S.P.}}(x_d) = \frac{2}{\sqrt{\lambda d}} \sum_{x_i} \exp\left[i2\pi\varphi_c(x_d, x_i)\right] \int_0^\infty \mathrm{d}x_c \exp\left[i\pi\,(\Delta x_c)^2 \varphi_c''(x_d, x_i)\right]\,.$$

The integral in the above expression is put in the form of a standard Fresnel integral $\int \mathrm{d}q \cdot \exp\left[i\pi q^2/2\right]$ by writing $q = \Delta x_c \sqrt{2\varphi_c''}$. This leads to the

form

$$\psi_{\text{S.P.}}(x_d) = \frac{1}{\sqrt{\lambda d}} \exp\left[\pm \frac{i\pi}{4}\right] \sum_{x_i} \exp\left[i2\pi\varphi_c(x_d, x_i)\right] / \sqrt{\varphi_c''(x_d, x_i)}\,, \tag{9}$$

where the positive and negative signs hold respectively for $\varphi_c'' > 0$ and $\varphi_c'' < 0$.

The diffraction amplitude in the case of Fresnel diffraction is obtained by substituting the expression (6*a*) for φ_c into (9) above. This yields

$$\psi_{\text{S.P.}}(x_d) = \sum_{x_i} \exp\left[\frac{i\pi}{4}\right] \cdot \frac{\exp\left[i2\pi\varphi_c(x_d, x_i)\right]}{(1 + d\cdot\alpha'(x_i))^{\frac{1}{2}}}\,, \tag{10}$$

with x_i defined by the relation

$$x_d = x_i + d\cdot\alpha(x_i)\,. \tag{11}$$

The demonstration of the relation between the stationary phase approximation and geometrical optics is readily made since the relation above resulting from the condition $\varphi'(x_d, x_i) = 0$ defines the classical electron trajectories. An electron passing the point x_i in the specimen is deflected through the angle $\alpha(x_i)$ and intersects that point P in the observation plane with coordinates (x_d, d) which satisfies the relation (11) above.

As shown in the Fig. 6 a beam element $\mathrm{d}x_i$ at the object plane gives rise to an element in the observation plane of width $\mathrm{d}x_d = \mathrm{d}x_i\ (1 + d\cdot\alpha'(x_i))$. In the case of an incident beam of intensity I_p the conservation of total current leads to the following expression for the intensity I at x_d:

$$\mathscr{I}(x_d) = \frac{I(x_d)}{I_p} = \frac{\mathrm{d}x_i}{\mathrm{d}x_d} = |1 + d\cdot\alpha'(x_i)|^{-1}\,.$$

In general several geometrical trajectories, ($x_1 P$, $x_2 P$, $x_i P$) in Fig. 6, may intersect at the point P. In this case the geometrical optics intensity is found by summing the contributions from each trajectory:

$$\mathscr{I}(x_d) = \sum_{x_i} |1 + d\cdot\alpha'(x_i)|^{-1}\,. \tag{12}$$

Comparison of the eqs (10) and (12) shows that the geometrical optics approximation gives an intensity distribution identical to that of the stationary phase method if a single term x_i contributes to the intensity $\mathscr{I}(x_d)$. That is

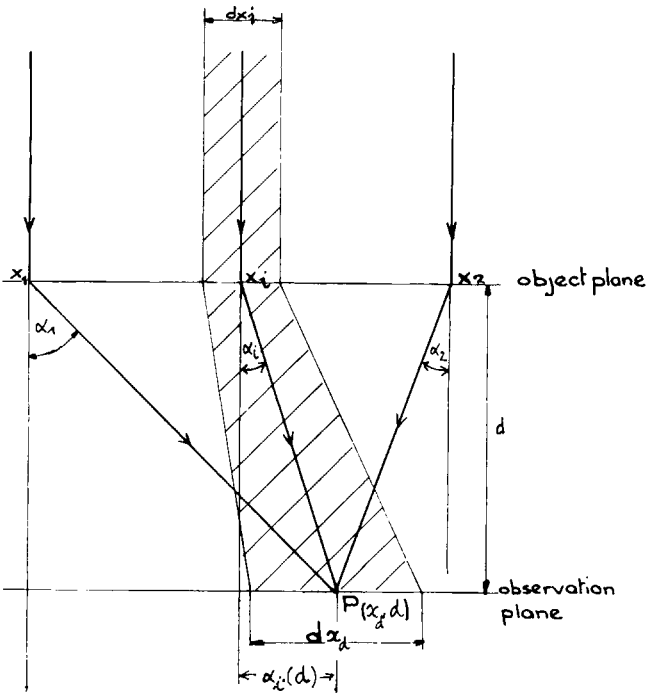


Fig. 6. – The geometrical image intensity at P is obtained by considering that the total current is conserved in the element around the ray x_iP. The element of width dx_i at the object is transformed by the difference in local deflections within the object to the width dx_d at P. Other trajectories x_2P, x_1P may contribute to the intensity at P.

only the one trajectory coming from the point x_i passes through the point x_d.

If several trajectories cross at P their mutual interference is taken account of by the stationary phase approximation, but not by the geometrical intensity expression which merely adds intensities.

2·2. Reduced parameters in the wave and geometrical optics equations.

The local electron beam deflection $\alpha(x_c)$ is given by

$$\alpha(x_c) = \frac{e}{h} \cdot \lambda a B_y(x_c) ,$$

where $B_y(x_c)$ is the y component of the magnetic field in the specimen. In the case of a domain wall aligned along the y axis $B_y(x_c) = B_p \cdot f(x_c/\omega_1)$ where

ω_1 is the domain wall half-width and $2\omega_1 = \omega$. The wave and geometrical optics expressions (6) and (12) contain the parameters: λ the electron wave length, $z = d$ the off-focus distance, a the film thickness, B_y the film magnetisation and the domain wall width ω. This embarassing number of parameters can be reduced in the following way. We write $\alpha(x_c) = \alpha_p \cdot f(x_c/\omega_1)$ as in eq. (7) and we note that $f(x_c/\omega_1) \to 1$ for $x > \omega_1$ whilst the wall is centred at $x = 0$ where $f(0) = 0$. Writing $X = x_c/\omega_1$ and $\mathcal{N} = x_d/\omega_1$ the diffraction integral (6) can be written in the forms:

Fresnel diffraction

$$\psi(\mathcal{N}) = \beta \int_{-\infty}^{+\infty} \mathrm{d}X \cdot \exp\left[i2\pi\left(\frac{\beta^2(X-\mathcal{N})^2}{2} + \beta_0^2 F(X)\right)\right]. \tag{13}$$

Fraunhofer diffraction

$$\psi(\mathcal{N}) = \beta \int_{-\infty}^{+\infty} \mathrm{d}X \cdot \exp\left[i2\pi(-\beta^2 X\mathcal{N} + \beta_0^2 F(X))\right]. \tag{14}$$

The geometrical intensity expression (12) can be written

$$\mathcal{I}(\mathcal{N}) = |1 + (\beta_0^2/\beta^2) \cdot f'(X)|^{-1}. \tag{15}$$

These expressions now contain only the two reduced parameters $\beta^2 = \omega^2/\lambda z$ and $\beta_0^2 = \omega\alpha_p/\lambda$, which together with the wall phase function $F(X)$ give a complete description of the problem. The reduced parameters are therefore extremely important in all problems of magnetic imaging. All of the problems which we discuss are treated in parametrised form. The parameter β_0^2 gives the strength in fluxon units of a magnetic object which can be described by the function $f(X)$, *i.e.* $\beta_0^2 = \frac{1}{2} \cdot \Delta\Phi/(h/e)$ where the magnetic inhomogeneity $\Delta\Phi = \Delta x \cdot \Delta B$ gives, in the case of a domain wall of width ω separating two domains of internal flux density B_p directed along $\pm y_c$ in a film of thickness a, $\Delta\Phi = \omega a B_p$.

2·3. The application of Wohlleben's criterion.

Wohlleben (1966) has applied the Heisenberg relation $\Delta x_c^{(1)} \Delta p_1 > h$ between the uncertainties in the electron position in the field region and the acquired lateral momentum p_1 to show that the detection of a magnetic flux

inhomogeneities $\Delta\Phi$ in the range $\Delta\Phi \leqslant h/e$ is possible only by abandoning the geometrical approximation in favour of a wave optical treatment. The quantity h/e is the flux quantum in the Gaussian system where h is Plank's constant, and e is the electron charge. The impact position $x_L^{(1)}$ of the electron beam deflected through α_1 by the object is given in the Fraunhofer diffraction plane by $x_L^{(1)} = L \cdot p_1/p_0$; L is a constant (in practice it is the effective camera length) and p_0, the component of electron momentum perpendicular to the object plane, is considered constant for small deflections $\alpha_1 = p_1/p_0$, see Fig. 7.

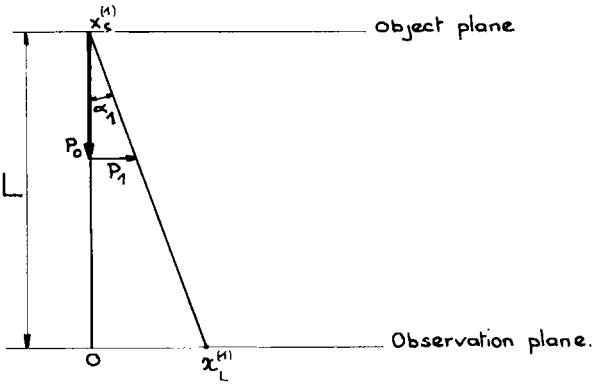


Fig. 7. – A measurement of the deflection α_1 carried out by detecting the impact position $x_L^{(1)}$ of the electron in an observation plane a distance L from the object is equivalent to measuring the lateral momentum p_1.

A measurement of $x_L^{(1)}$ is equivalent to a measurement of p_1. The object and image functions are related by a Fourier transformation which implies a Heisenberg like relation between the uncertainties of $x_c^{(1)}$ and $x_L^{(1)}$. In off-focus images the position $x_d^{(1)}$ in the observation plane is given by the relation (11) written in the form $x_d^{(1)} = x_c^{(1)} + d \cdot \alpha_1$. The object and image functions are related by a different transform, that described by the Fresnel integral.

2·4. **The generalised criterion.**

Attempts to find a criterion equally applicable to Fresnel as to Fraunhofer diffraction have been based on the stationary phase approximation to the Fresnel-Kirchhoff diffraction integral. We have seen in Subsect. **2**.1 that the

stationary phase approximation is equivalent to the geometrical optics approach under the restriction that only a single trajectory passes through each point in the image plane. A discussion of the validity of the stationary phase approximation then implicitly considers the validity of geometrical optics.

Cohen (1967) argues that for a given observation point only the object region $x_i - \Delta x/2 < x_i < x_i + \Delta x/2$ contributes appreciably to the intensity $\psi(x_d)$ given by eq. (6). Structure within the interval Δx will not be resolved. In order to resolve two points in the sample separated by the distance Δx the waves originating at each of these points must be at least $\pi/4$ out of phase so that destructive interference occurs at the point x_d in the image plane. This leads to an inequality which we can write in parametrised form:

$$(\Delta X)^2 |\beta_0^2 f'(X_i) + \beta^2| \geqslant 1 . \tag{16}$$

Since the radius r of the first Fresnel zone is given in the stationary phase approximation by $r^2 = 1/\varphi''(x_i)$ we can give a more direct interpretation of the derivation of the inequality (16). It is supposed that $r(X_i)$ necessarily represents an uncertainty in position of the point X_i in the specimen since all of the zone r around X_i contributes to the intensity at $\mathcal{M}$. The attainable resolution ΔX_i has a lower limit given by $(\Delta X_i/r)^2 \geqslant 1$. Substitution for r yields (16).

Guigay and Wade (1968) aim to determine the conditions necessary to ensure that the wave and geometrical optics image intensities are the same. The second-order Taylor expansion of the phase is valid only in a limited region $\Delta X1 = |X - X_i|$ around the stationary phase point X_i. The remaining terms in the expansion assume a greater importance as $\Delta X1$ increases. We impose as condition on the extent of $\Delta X1$ that the second order term must be greater than that of the third order. This yields for $\Delta X1$

$$\Delta X1 < 3|\varphi''(X_i)/\varphi'''(X_i)| .$$

Furthermore the diffraction integral can be limited to a region $\Delta X2$ which represents a sufficient number of Fresnel zones to give a good approximation to the complete range of integration. For the stationary phase method to be valid it is necessary that $\Delta X1 > \Delta X2$. This requirement leads directly to the inequality

$$\beta_0^2 |f''(X)| < |\beta^2 + \beta_0^2 f'(X)|^{\frac{3}{2}} . \tag{17}$$

Neither of the criteria (16) and (17) are entirely satisfactory: that of Cohen

because it does not really treat the validity of the stationary phase method which is only used to calculate r; that of Guigay and Wade because if φ''' should be zero as it is at the centre of a domain wall it is necessary to consider the higher order terms in the expansion.

2·5. Another formulation of the generalised criterion (*).

The total phase of the Fresnel integral expressed in terms of the reduced parameters and normalised co-ordinates (13) is:

$$\varphi_c(X,\mathcal{N}) = 2\pi\left[\frac{\beta^2}{2}(X-\mathcal{N})^2 + \beta_0^2 F(X)\right].$$

We limit our discussion to the domain wall problem. For a given point $\mathcal{N}$ in the observation plane the first Fresnel zone around the object point X has a radius $r^2 = 1/\varphi_c''(X)$ according to the stationary phase approximation. The complete phase expression φ_c gives for the difference in phase between the points $(X+r,\mathcal{N})$ and $(X,\mathcal{N})$:

$$\Delta\varphi = \varphi_c(X+r,\mathcal{N}) - \varphi_c(X,\mathcal{N}).$$

Since the stationary phase method gives $\Delta\varphi_{\text{S.P.}} = \pi$ we require that

$$\Delta\varphi - \Delta\varphi_{\text{S.P.}} \ll \pi.$$

This inequality will be best satisfied for $\Delta\varphi = \pi$, *i.e.* for:

$$L(X) = \beta^2 r^2 \pm 2\beta_0^2 r f(X) \mp 2\beta_0^2 \langle F(X+r) - F(X)\rangle = 1. \tag{18}$$

The upper signs hold for the diverging case, the lower for the converging case.

2·5.1. *Application to the centre of the wall image.* – At the wall centre $f(0) = 0$ and $F(0) = 0$ so that we can write (18) in the form of the condition

$$\beta^2 r^2 + 2\beta_0^2 F(r) \to 1. \tag{19}$$

The first term on the l.h.s. is the geometrical intensity. All wall models have

(*) Derived from unpublished work carried out in collaboration with J. P. Guigay.

the same behaviour in that:

$$F(X) \to X^2/2\,, \qquad \text{for } X < 1\,,$$

$$F(X) \to X - \text{const}\,, \qquad \text{for } X \gg 1\,.$$

In Fig. 8 we plot $F(r)/r^2$ against $1/r^2$ for the function $F(r) = \ln \cosh(r)$. This corresponds to the wall model $\tanh(X)$. The plot shows that $F(r)$ rapidly approaches the limiting value of $r^2/2$ except for $1/r^2$ very small. Other

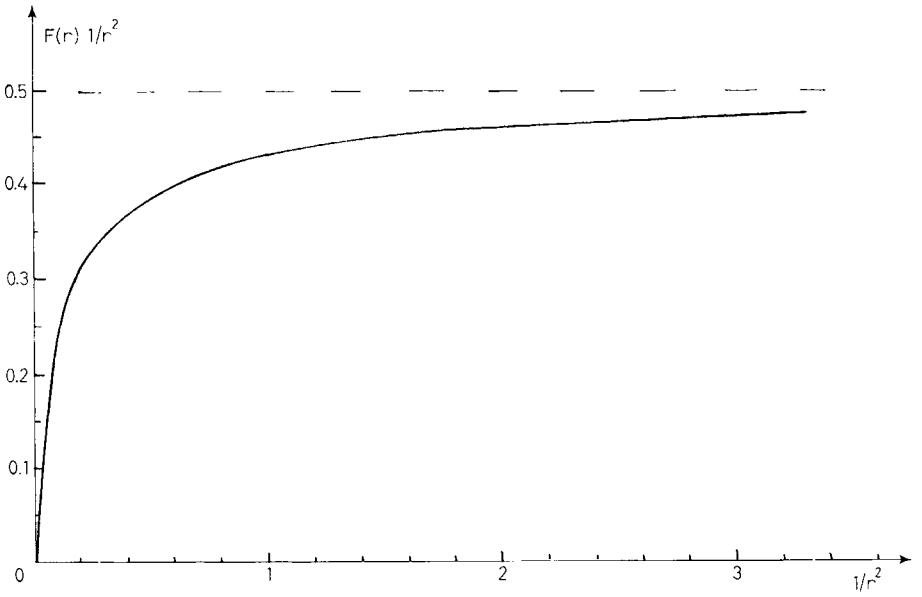


Fig. 8. – A plot of $F(r)/r^2$ against $1/r^2$ for the wall model $f(X) = \text{tgh}(X)$ for which $F(r) = \ln \cosh(r)$. Other wall models give similar curves.

wall models give essentially the same curve. Writing $F(r) = r^2/2$ for $1/r^2 > 1$ leads to the inequality

$$\beta^2 \pm \beta_0^2 > 1\,. \tag{20}$$

The satisfaction of the inequality (20) ensures that (19) is fulfilled. The inequality reduces to Wohlleben's criterion when $\beta^2 \ll 1$ which corresponds to the Fraunhofer diffraction condition. When $\lambda \to 0$ both β^2 and β_0^2 become very large and the inequality is satisfied. The negative sign holds for the converging wall image; $\beta^2 = \beta_0^2$ corresponds to the summit of the caustic where geometrical optics predicts an infinite intensity. The inequality (20) cannot be fulfilled at such a point giving a formal demonstration of the physically obvious fact that geometrical optics breaks down in such a region.

Alternatively we note that for a given value of $1/r^2$ we can write $F(r) = r^2/2 - k(r) \cdot r^2/2$. We can thereby express the condition (19) in the form

$$k(r) \cdot c(0) \to 0 ,$$

where $c(0)$ is the contrast at the wall centre. This condition is satisfied either for small contrasts or for small k which imposes a condition on $(\beta^2 \pm \beta_0^2)$ similar to (20) above.

2·5.2. *Application to the complete wall image.* – It is not possible to express the condition (18) in a simple form for positions away from the wall

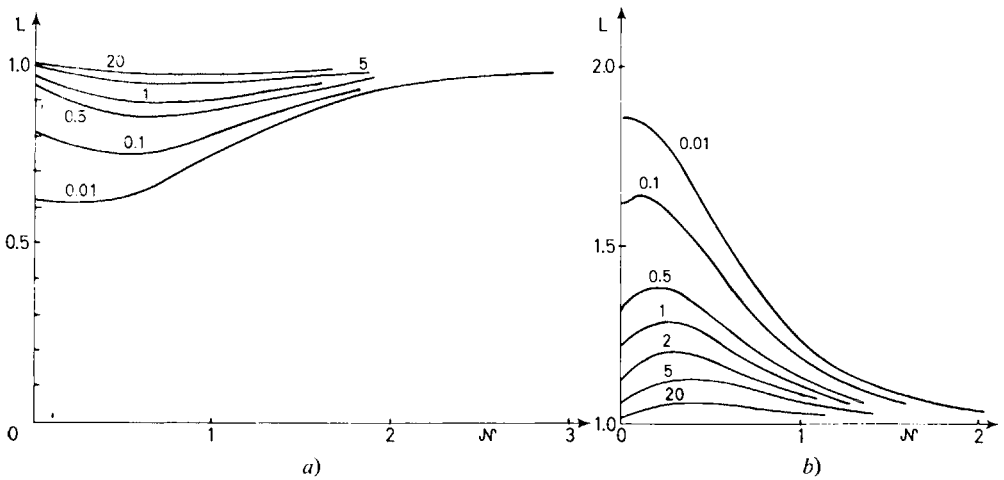


Fig. 9. – The phase difference $L(r)$ defined by eq. (18) plotted as a function of $\mathcal{N}$ the position in the image plane. *a*) A diverging wall with $I(0) = 0.5$; this imposes $\beta_0^2 = \beta^2$. The parameter β_0^2 is varied from 0.01 to 20 to cover the range from a weak to a strong phase object. *b*) A converging wall with $I(0) = 2$.

centre. We present therefore a numerical evaluation of L for the wall model $f(X) = \tanh(X)$. In the case of a diverging image we put $\beta_0^2 = \beta^2$ which imposes a central intensity of 0.5. In Fig. 9*a*) we plot L against $\mathcal{N}$ for various values of the parameters β_0^2, β^2. Comparison of the corresponding wave and geometrical image profiles indicates that for $0.8 \leqslant L \leqslant 1.2$ the two treatments can be considered in agreement. Similar curves are shown for the converging case in Fig. 9*b*) with $\beta^2 = 2\beta_0^2$.

Since the least favourable value of L in a given wall image is not far removed from the central value the condition (20) which we have applied to the image centre seems to offer a good estimation of whether geometrical optics may be valid throughout the wall image. In view of the simple form of the inequality (20) it is convenient to adopt it as a criterion.

2·6. Application to periodic objects.

The phase φ_c of the Fresnel integral

$$\varphi_c = 2\pi\left[\frac{\beta^2}{2}(X - N)^2 + \beta_0^2 F(X)\right]$$

has the additional property that $F(X) = F(X+1)$. A necessary condition for the application of the stationary phase approximation is that the width r of the first Fresnel zone be much less than the periodicity of the object which gives the condition

$$|\beta^2 + \beta_0^2 f'(\mathcal{N})| \gg 1 . \tag{21}$$

The periodic function $f'(\mathcal{N})$ has a mean value of zero so that for the condition (21) to be everywhere satisfied it is necessary that $|\beta^2| \gg 1$ or in terms of the defocussing distance $z \ll d_1$ where $d_1 = \varepsilon^2/\lambda$. The caustic surfaces below the object will have a saw-tooth form with the distance of the cusp given by $d_0 = 1/\alpha'_{\max}$. If $d_1 > d_0$ geometrical optics must be limited to the region $z \ll d_0$ whilst for $d_1 < d_0$ the condition $z < d_1$ must be satisfied. We may remark that $z = d_1$ corresponds to the first position of uniform image intensity as the object is defocussed by the imaging lens. Since the z periodicity of the image intensity is a wave optical effect predicted by eq. (1) it is evident that geometrical optics is forcibly limited to the region $z < d_1$.

3. Experimental investigations of magnetic structure.

3·1. The domain wall.

Most experimental attempts to measure domain wall widths have used the Fresnel image. For the present we will treat only this type of image. Independently of geometrical or wave considerations we can conceive two types of observation: measurements of position and measurements of intensity.

3·1.1. *Position measurements.* – Most early attempts involved looking for some geometrical property of the image which may be related to the domain wall width. Fuchs (1962) used the geometrical optics relation $d_0 = 1/\alpha'(0) =$

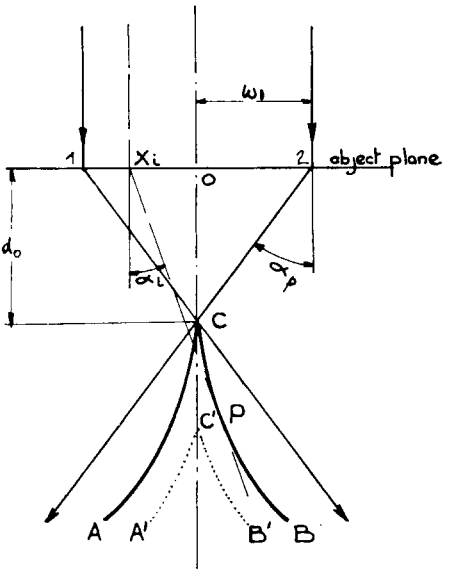


Fig. 10. – The caustic surface ACB associated with a domain wall has its cusp C a distance $d_0 = 1/\alpha'(0) = \omega_1/\alpha_p$ from the object plane. A ray from x_i deflected through the angle α_i is tangent to the caustic at P. The central maximum predicted by wave optics lies at C' below C. The surface of maximum intensity given by wave optics is distinct from the geometrical caustic. The extreme rays $1C$ and $2C$ represent the perfect lens case for which all the rays between $1C$ and $2C$ are focussed at C. The caustic surface becomes a point.

$= \omega_1/\alpha_p$ which determines the position of the caustic cusp C, Fig. 10. The deflection is measured in a separate diffraction experiment and d_0 must also be measured. In addition for defocussing distances $z > d_0$ the caustic surface divides into two branches the position of which as a function of z allows the wall structure to be determined through the relation $f'(X) = \beta^2/\beta_0^2$. Fuchs carries out this operation graphically; the experimental plot of $\mathcal{I}_{max}$ against z gives the caustic surface the tangent of which at the point P cuts the object plane at X_i. The angle α_i so determined plotted against X_i gives the wall structure.

We know intuitively and the inequalities (18), (20) demonstrate formally that geometrical optics breaks down in the region of the caustic surface where it predicts infinite intensities. The wave optical maxima lie below the caustic surface; Guigay shows that the wave optical cusp may be much below the object. It would be of interest to carry out experiments of this type but using a wave optical interpretation since experimentally it is interesting to work in the region of the caustic cusp because of the high intensity available in this region.

A simpler method was used by Wade (1966) in which abstraction being made of the internal wall structure it is noticed that measurements of the widths W_c and W_d of convergent and divergent wall images obtained at the same defocussing distance yield the wall width from the relation $W_d - W_c = 2\omega$, see Fig. 11. Plots of W_d and W_c against z yield straight lines of different

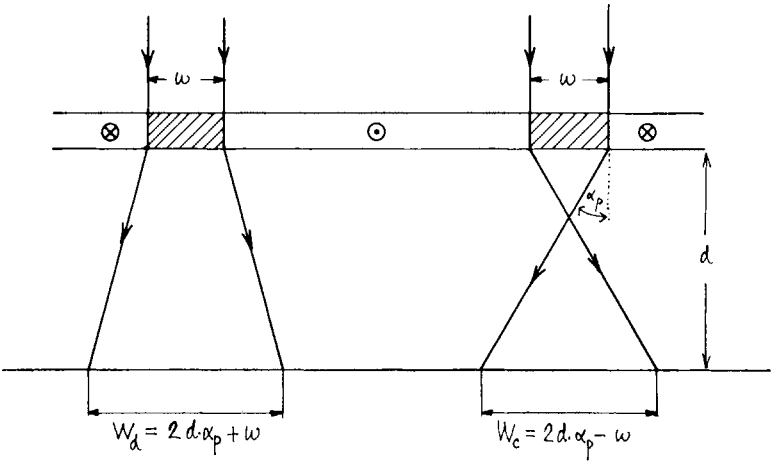


Fig. 11. – The simplified geometrical relation between the wall width ω and the projected image widths W_d and W_c.

slope which can be supposed to be due to the effect of the finite separation of the source and object and also due to the real wall structure. Wohlleben (1967) believes that the difference in slopes may be due to the wave optical dependence of the image profile edges on defocussing distance. This belief is supported by more recent work of Reimer and Kappert (1969) who have made a large number of numerical calculations of wall image profiles taking into account the finite illumination aperture.

Their calculations for iron films of thickness 200 Å and 500 Å, with wall widths 500 Å and 800 Å, correspond respectively to values of the parameter β_0^2 of 1.25 and 0.8. They find that the wave and geometrical optics image profiles are practically indistinguishable in the diverging case especially when the smoothing effect of the illumination aperture α_B is taken into account. In the converging wall image the interference fringes of separation δ are no longer resolvable for $\alpha_B > \delta/d$. Geometric and wave optical profiles are rather similar for large α_B. They find that extrapolation to $z = 0$ of the divergent profile halfwidth, which is rather insensitive both to α_B and the detailed wall model, gives a good estimate of ω subject to a correction factor dependent on the wall model.

3·1.2. *Intensity measurements.* – A wave optical analysis of the interference phenomena found in converging domain wall images was first made by Boersch *et al.* (1960, 1961, 1962) who found the fringe separation to be given by $\delta = \lambda(d+g)/2\alpha_p \cdot g$, where g is the source-object distance. This formula is in agreement with that of the Fresnel biprism but certain details of the fringe positions do not agree with the biprism formulation. The condition on the illumination for the fringes to be visible is $\alpha_B < \delta/d$. Since the average current intensity in the observation plane is given by $j = R\pi(\alpha_B/M)^2$, (R is the source brightness and M the image magnification), α_B can only be reduced, at the expense of the image intensity. As an example we show in Fig. 12 an image of interference fringes in permalloy film obtained with an exposure time of around five minutes despite the use of a point filament source for which R is greater than for conventional hairpin filament. Having chosen a model to represent the wall, the width of the wall ω is adjusted so that the intensity profile calculated using eq. (13) best fits the experimental microdensitometer trace.

Hothersall (1969) by numerical analysis shows the profiles to be very sensitive to the illumination aperture α_B. In addition he found that the calculated intensity profiles can be fitted to the experimental traces for different wall models by suitable adjustment of the width ω. He gives an empirical

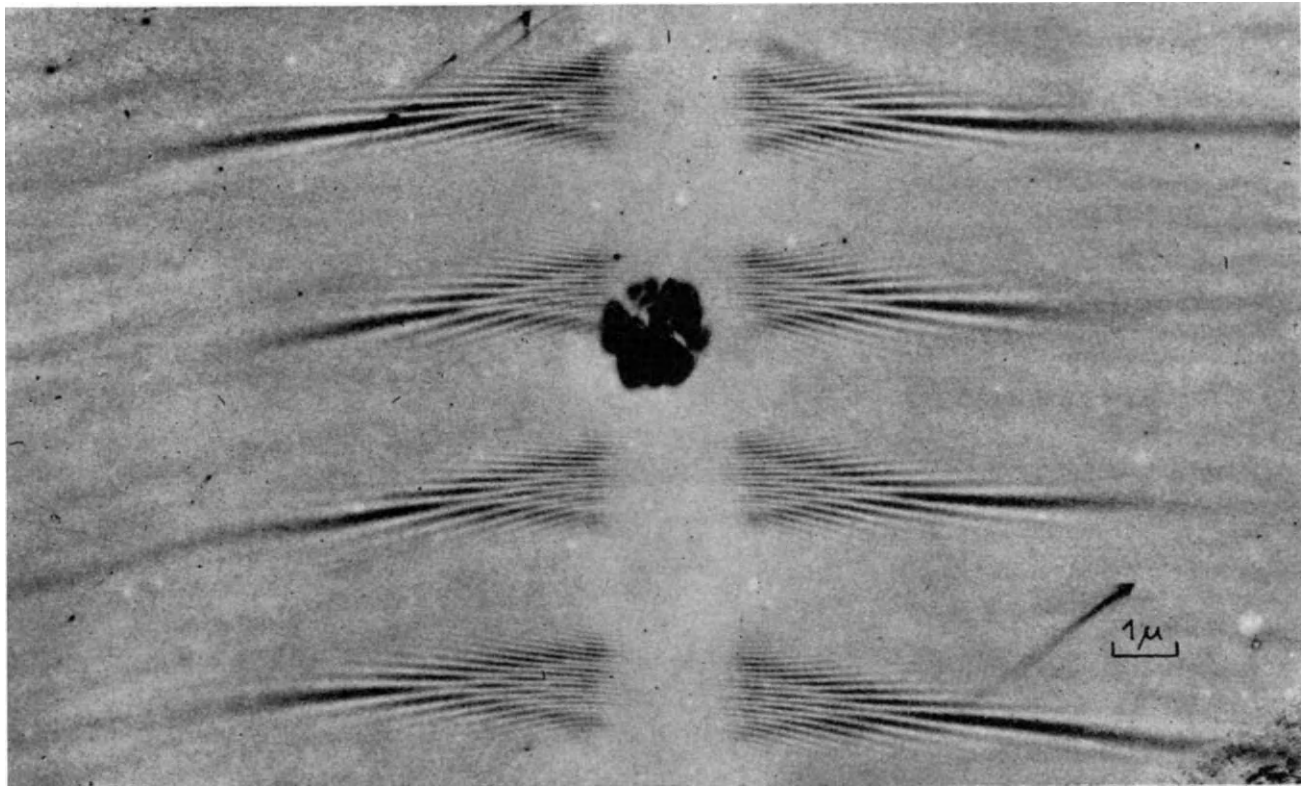


Fig. 12. – Interference fringes at a cross-tie wall image in a permalloy film about 500 Å thick. A pointed filament was used as source of electrons. The defocussing distance of the image is about 3 cm. The parameter β_0^2 decreases as the apexes of the cross-wall images are approached.

relation between the widths found for the different wall models. The experimental difficulties are not diminished by the background of incoherent electrons inelastically scattered in the specimen. The absolute measurement of intensity is rendered impossible. Hothersall side-steps this difficulty by measuring the intensity of the interference maxima and minima with respect to the mean background intensity.

A lesson to be drawn from Hothersall's results is that it is not very satisfactory to impose a precise wall model. It is the function F shown in Fig. 4 with $\varphi_s = \beta_0^2 F$ which describes the wall as a phase object. F has the following behaviour:

$$F(\mathcal{N}) = \mathcal{N}^2/2\,, \qquad \mathcal{N} \ll 1$$

$$F(\mathcal{N}) = \mathcal{N} - k\,, \qquad \mathcal{N} \gg 1\,,$$

where k is a constant, larger than 0.5. Guigay considers that the essential behaviour of the function F is contained in the two parameters ω, the wall width, and k which specifies the behaviour for large $\mathcal{N}$. The values of k corresponding to common wall models are summarized below:

$f(\mathcal{N})$	$\mathcal{N}$	tgh$(\mathcal{N})$	sin$(\mathcal{N})$
k	0.5	ln 2 = 0.69	$\pi/2 - 1 = 0.57$

Guigay (1970) combines the two functions F_1 and F_2 defined below with suitable weighting factors so as to vary the wall phase function $F = aF_1 + bF_2$, where $a + b = 1$, in a continuous manner.

$$F_1(\mathcal{N}) = \left\{ \begin{array}{ll} \mathcal{N}\,, & \mathcal{N} \leqslant 1 \\ 1\,, & \mathcal{N} \geqslant 1 \end{array} \right\}, \qquad k = \tfrac{1}{2}\,,$$

$$F_2(\mathcal{N}) = \mathcal{N}/\sqrt{1+\mathcal{N}^2}\,, \qquad k = 1\,,$$

He shows for example that the convergent wall interference fringe profiles shown in Fig. 13 are sensitive to k even for constant wall width. The physical reason for this is that the interference profile is sensitive to the phase of the rays coming from the domains on either side of the wall and that their phases depend on k, see Fig. 14. We may note that separate measurements of ω and k may allow F to be specified in a relatively precise manner.

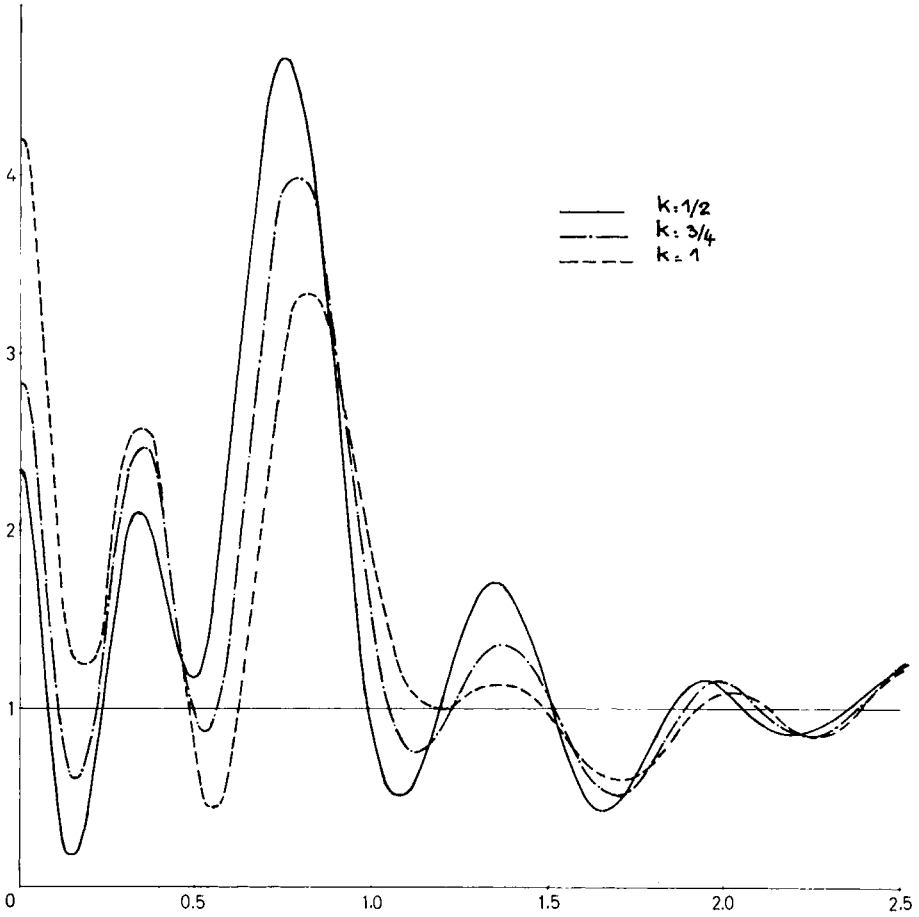


Fig. 13. – Intensity profiles for a converging wall with $\beta_0^2 = 0.5$ at a defocusing distance of $4d_0$. The three curves correspond to different values of k. ——— $k = \frac{1}{2}$, –·–·– $k = 3/4$, – – – $k = 1$. (Due to J. P. Guigay.)

A measurement of ω alone from the converging wall interference profile is without a great deal of significance. However as we shall presently see it is possible to measure ω independently of k using the divergent wall image.

Other investigations have been made using diverging wall images in the hope, based essentially on physical intuition, that a geometrical approach is valid. A formal justification must be sought in the generalised criterion (18), (20).

Warrington (1964) noticed that the value of $\mathscr{I}(0)$ is insensitive to the illumination aperture. He suggests that a measurement of $\mathscr{I}(0)$ could be useful to obtain ω. Suzuki *et al.* (1968) have used both this method and a profile fitting. Wall widths obtained by the two methods disagree. Guigay and Wade (1968) show that although a sufficiently small defocalisation is often necessary to enable the entire image to be treated geometrically the

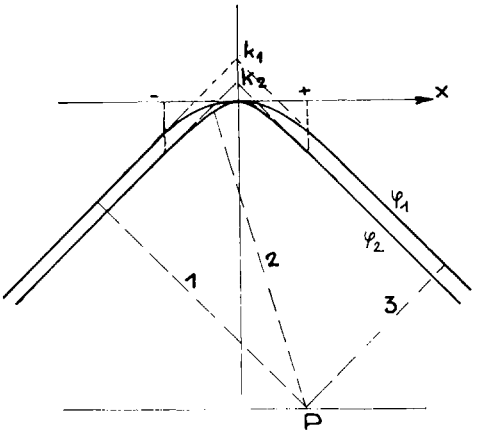


Fig. 14. – The phase profiles φ_1 and φ_2 correspond to different wall models which have the same width ω. The extrapolation to $X = 0$ of the linear portions of the curves give the values k_1 and k_2 respectively. The rays 1, 2, 3 which interfere at P have relative phases which differ for the two wall models. The interference profiles will then be different as shown in Fig. 13.

value of $\mathscr{I}(0)$ calculated geometrically is often valid for any defocussing distance. They also proposed the use of this ratio, for reasons complementary to those of Warrington, in the measurement of domain wall widths. A confirmation of this property is offered by the numerical calculations of Guigay and Wade and of Reimer and Kappert. The latter workers have encountered experimental difficulties due to inelastic background electrons which falsify the ratio.

3˙1.3. *Inversion procedures.* – Several attempts have been made to calculate the angular deflection within the domain wall region directly from an image intensity profile. The first method is that used by Fuchs in which $\alpha(X)$ is found graphically from a trace of the caustic surface, see Subsect. **3**˙1.1.

Another method exploited independently by Petrov *et al.* (1968) and by Cohen and Harte (1969) constitutes an inversion of the geometrical intensity profile for the case of a diverging wall. The geometrical intensity expression (15) is obtained from the relation $\mathrm{d}x_c I_p = \mathrm{d}x_d I(x_d)$ expressing the conservation of total current, see the Fig. 6. In the case of a domain wall image with normalised background intensity $\mathscr{I}(\mathscr{N})$ we have:

$$\mathrm{d}X = \mathscr{I}(\mathscr{N}) \cdot \mathrm{d}\mathscr{N}\,, \tag{22}$$

where X and $\mathscr{N}$ are related by $\mathscr{N}_i = X_i + d \cdot \alpha(X_i)$. Integrating both sides of (22) yields:

$$X_i = \int_A^{\mathscr{N}_i} \mathscr{I}(\mathscr{N}) \cdot \mathrm{d}\mathscr{N}\,. \tag{23}$$

The point A is in the uniform intensity region away from the wall image. The eq. (23) yields a curve of the form shown in Fig. 15*a*) where the straight

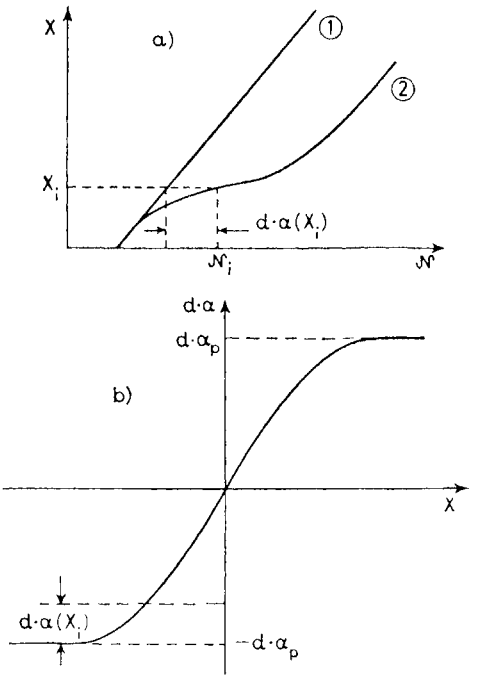


Fig. 15. – Showing the relationship between X_i and $\mathscr{N}_i$ obtained from eq. (23). Curve 1) holds for a uniform intensity region whilst curve 2) scans across a wall image. *b*) The profile of the local deflection $d\cdot\alpha(X_i)$ obtained from the curves 1) and 2) of *a*).

line (1) holds in the uniform intensity region and the curve (2) is produced as the wall image profile is scanned by $\mathcal{N}_i$. The curves (1) and (2) yield the curve of Fig. 15*b*) showing, $d \cdot \alpha(X)$, plotted against X which constitutes the solution to the problem.

The presentation above is essentially that of Petrov. The treatment of Cohen and Harte differs in that the solution is found numerically.

3·2. The ripple problem.

Although little quantitative work exists on this subject it seems generally accepted that the electron microscope will act like a filter in that a given defocussing will enhance a certain spatial periodicity present in the object. It is a well known fact that for a given object periodicity the spherical aberration of the objective lens can be compensated by a defocalisation in such a way that the well known lattice resolution test does not constitute a proof of the quality of the objective lens.

Equation (1) contains the defocussing term $[i2\pi z n^2 \lambda/2\varepsilon^2]$ the phase of which can be expressed in terms of the diffraction angle θ in the form $2z\theta^2/2\lambda$. The phase shift introduced by spherical aberration is $2\pi C_s \theta^4/4\lambda$. In the presence of spherical aberration eq. (1) will contain the term $\exp[i\varrho]$ where

$$\varrho = \frac{\pi}{2\lambda}(C_s \theta^4 - 2z\theta^2) .$$

We notice immediately that the in-focus image ($z = 0$) will be perturbed by the spherical aberration term. A defocussing of $z_s = C_s\theta^2/2$ allows the object function to be recuperated. In the case of an atomic lattice of period 4 Å, $z_s = 500$ Å for $\lambda = 4 \cdot 10^{-10}$ cm and $C_s \approx 0.1$ cm.

In general we will find a repetition of the object function for $\varrho = 2\pi m$ where m is an integer. For a phase object the corresponding image will be without contrast. Maximum contrast is found for $\varrho = (2m-1)\pi/2$. This yields

$$z_s = C_s\theta^2 - (2m-1)\lambda/\theta^2 . \tag{24}$$

We see from (24) that the effect of spherical aberration is to displace the origin of the classical z periodicity by the distance $C_s\theta^2$. For magnetic deflections and for object periodicities in the range 10^{-5} cm we have $\theta \sim 10^{-5} \div 10^{-4}$ rad. Taking the least favourable value $\theta = 10^{-4}$ rad yields

$C_s\theta^2 \simeq 2\cdot10^{-7}$ cm (we put $C_s \simeq 20$ cm) and $\lambda/\theta^2 \simeq 4\cdot10^{-2}$ cm. We can therefore ignore the effect of spherical aberration on the defocussed image which will be adequately represented by eq. (1).

In order to test the filtering action of the defocussed image we suppose a phase function of equally weighted sinusoidal terms. We consider the range $A\sum_1^N \sin 2\pi x_c/n\varepsilon$, where A is supposed small so that we are dealing with a weak phase object. The intensity in the Fresnel image can be expressed as

$$I(x, z) = 1 + 2A\sum_1^N \sin\frac{\pi\lambda z}{n^2\varepsilon^2}\cdot\sin\frac{2\pi x}{n\varepsilon}.$$

The defocussing planes of maximum contrast for the periodicity $n\cdot\varepsilon$ are given by $z = (2m-1)n^2\varepsilon^2/2\lambda$. We have calculated numerically the image intensity for the range $N = 10$, for $\varepsilon = 10^{-5}$ cm, *i.e.* the sine terms have a periodicity range from $(10^{-5}\div10^{-4})$ cm. We choose values of z to give maximum contrast.

The results show the intensity to well reproduce the smallest periodicities but that the larger periodicities are almost completely masked. We conclude that it may be rather difficult by direct observations to detect the entire periodicity range in an object containing a dispersion of spatial frequencies. The theory of image transfer combined with optical diffraction should be fruitful for treating this sort of problem.

REFERENCES

General articles and books:

C. FERT: *Traité de Microscopie Electronique*, Ed. C. MAGNAN, Hermann (1961), p. 333.

P. J. GRUNDY and R. S. TEBBLE: *Adv. in Phys.*, **17**, 153 (1968).

P. B. HIRSCH, A. HOWIE, R. B. NICHOLSON, D. W. PASHLEY and M. J. WHELAN: *Electron microscopy of thin crystals*, Butterworths (1965).

R. H. WADE: *Journ. de Phys.*, Colloque C2, Supplement aux n° 2-3, **29**, 95 (1968).

F. ZERNIKE: *Physica*, **9**, 686, 794 (1942).

Aharanov and Bohm Effect:

Y. AHARANOV and D. BOHM: *Phys. Rev.*, **115**, 485 (1959); **123**, 1511 (1961).

W. EHRENBERG and R. E. SIDAY: *Proc. Phys. Soc.*, **62**, 8 (1949).

Geometrical optics:

H. W. FULLER and M. E. HALE: *Journ. Appl. Phys.*, **31**, 238 (1960).

Wave optics:

H. BOERSCH, H. HAMISCH, D. WOHLLEBEN and K. GROHMANN: *Zeits. für Phys.*, **159**, 397 (1960); **164**, 55 (1961); **167**, 72 (1962).

M. S. COHEN: *Journ. Appl. Phys.*, **38**, 4966 (1967).

D. WOHLLEBEN: *Journ. Appl. Phys.*, **38**, 3341 (1967).

The validity of geometrical optics:

M. S. COHEN: *Journ. Appl. Phys.*, **38**, 4966 (1967).

M. S. COHEN and K. J. HARTE: *Journ. Appl. Phys.*, **40**, 3597 (1969).

J. P. GUIGAY and R. H. WADE: *Phys. Stat. Sol.*, **29**, 799 (1968).

R. H. WADE and J. P. GUIGAY: *Ecole d'Eté de Microscopie Electronique*, Perros Guirec (1969).

D. WOHLLEBEN: *Phys. Lett.*, **22**, 564 (1966); *Journ. Appl. Phys.*, **38**, 3341 (1967).

Observation of periodic objects:

O. BOSTANJOGLO and W. VIEWEGER: *Phys. Stat. Sol.*, **32**, 311 (1969).

R. P. FERRIER and I. B. PUCHALSKA: *Phys. Stat. Sol.*, **28**, 335 (1968).

R. H. WADE: *Phys. Stat. Sol.*, **19**, 847 (1967).

Wall width measurements:

M. S. COHEN and K. J. HARTE: *Journ. Appl. Phys.*, **40**, 3597 (1969).

E. FUCHS: *Z. Angew. Phys.*, **14**, 203 (1962).

J. P. GUIGAY: *Proc. 7th Int. Conf. on Electron Microscopy, Grenoble 1970* (Paris, 1970), vol. **2**, p. 605.

D. C. HOTHERSALL: *Phil. Mag.*, **20**, 89 (1969).

V. E. PETROV, N. N. SEDOV and G. V. SPIVAK: *IZV Akad. Nauk, USSR, Phys.*, **132**, 1185 (1968).

L. REIMER and H. KAPPERT: *Z. Angew. Phys.*, **26**, 58 (1969); **27**, 165 (1969).

T. SUZUKI, C. H. WILTS and C. E. PATTON: *Journ. Appl. Phys.*, **39**, 1983 (1968).

R. H. WADE: *Journ. Appl. Phys.*, **37**, 336 (1966).

D. H. WARRINGTON: *Phil. Mag.*, **9**, 261 (1964).

Magnetic Phase Contrast

D. WOHLLEBEN (*)

University of California - San Diego, Cal., U.S.A.

1. Introduction.

Lorentz microscopy is the general designation for a variety of techniques by which microscopic static magnetic field distributions can be studied through the deflection of electrons by the Lorentz force. A few papers concerned with magnetic fields in vacuo appeared more than 20 years ago (Ardenne 1943, Marton 1948) but most of the work in the past decade was motivated by the need of the computer industry for information on the ferromagnetic structure of thin films. It was shown in 1959 by Hale, Fuller and Rubinstein ([1]) and by Boersch and Raith ([2]) that ferromagnetic domains can be observed with high resolution in the conventional transmission electron microscope. This technique led immediately to the discovery of the ripple structure within domains and has since been very helpful in *qualitative* studies of wall configurations, ripples and stripe domains. However, a rather extensive effort to measure *quantitatively* the field distribution in domain walls and ripples met with disappointing results. Later, the technique failed even qualitatively to detect the Abrikosov flux line lattice in hard superconductors and the long period magnetization oscillations in antiferromagnetic chromium and rare earth metals, structures which should be easily resolved spatially in the conventional transmission electron microscope.

The main difficulty in high resolution Lorentz microscopy is the weakness of the interaction of the electrons with the magnetic field. This interaction

(*) Supported by the Air Force Office of Scientific Research under Grant No. AF-AFOSR-631-67.

was initially described by the classical Lorentz force. However, when sufficiently weak interactions are considered, quantum effects play an important role. Then the Lorentz force must be replaced by some quantum-mechanical equivalent, which incorporates field integrals rather than the fields themselves. The proper field integral for Lorentz microscopy is the magnetic flux. The flux causes a well-defined co-ordinate dependent phase shift in the incident electron beam, which is measurable and which describes the effect of a static magnetic field distribution on the electrons *completely*. Thus Lorentz microscopy is nothing but phase contrast microscopy with electrons for a special class of objects. Some well-known techniques to produce contrast from phase objects in electron optics are defocusing, the knife edge (Foucault) tecnique, dark field, Zernike phase contrast, low angle diffraction and interference microscopy. They all can be applied to study magnetic objects. In fact, the simple objects of Lorentz microscopy are probably the best available in electron microscopy for a quantitative study of the relationship between phase contrast and phase object:

Firstly, presumably the electron scattering on magnetic fields is completely elastic, contrary to electrostatic scattering on condensed matter, where the elastic and inelastic contributions are hard to separate experimentally.

Secondly, the derivation of the magnetic phase shift caused by conventional Lorentz objects is very simple. It does not suffer from the numerous complications of the calculation of the electrostatic phase shift which are caused by the three-dimensional periodicity of the lattice and by badly known atomic wave functions.

Thirdly, whereas the study of electrostatic phase shifts on an atomic scale is complicated in practice by the severe distortions introduced by the instrument at the high spatial frequencies involved, the spatial frequencies associated with conventional Lorentz objects are several orders of magnitude smaller, so that the imaging process can be considered as ideal.

Thus it seems that a good understanding of Lorentz contrast should be very helpful on the way to quantitative information retrieval from phase objects in general.

For a given object area the average magnetic phase shift can be either large or small compared to π (strong or weak object). In the following we introduce the quantum signal to noise ratio, *i.e.* the ratio of the number of transmitted electrons which are measurably affected by the magnetic object to those which are not. This ratio is roughly equal to the square of the involved magnetic flux measured in units of the quantum of flux h/e (h is Planck's quantum of action, e the electronic charge). If this ratio is larger

than one (strong object), the laws of classical geometric optics are more or less applicable. However, for sufficiently small flux (weak object), the signal to noise ratio can become smaller than one. Then diffraction effects dominate the contrast. They cause some complications in the *mathematical* analysis of the relation between contrast and object, which, however, are not insuperable. A more serious *experimental* difficulty associated with weak objects is the near cut-off of the electron scattering probability at the quantum of flux. This causes the need for a drastic increase of the illumination time over the classical value in order to record any information at all, *independent of the mode of detection.* There is an equivalent cut-off of the electrostatic scattering probability, since atoms and not too thick films are weak objects in the above sense. Thus the difficulty is a very general one.

At the present time it seems that experimental progress towards higher resolution in Lorentz microscopy will be difficult and can only come after some deeper theoretical understanding. Therefore basic concepts are emphasized in these lectures. No attempt is made to review past work. Practical examples are drawn in only when needed as background for theoretical discussion. Unfortunately, the discussion is not even theoretically complete, since there was no space to treat partial coherence. However, the treatment of the coherent case is sufficiently fruitful to produce some important guidelines for future experimental work. In particular, it is hoped that the concepts of scattering probability and signal to noise ratio, which are new to phase contrast microscopy, will obviate to experimenters the necessity to pay much attention than in the past to increasing the source brightness, to using detectors with better linearity than photographic material, and to recording the information in the regions of maximum contrast of the image space, even if that means facing the worst diffraction effects.

In Sect. **2** the Schrödinger wave function is derived in the presence of the magnetic object and is incorporated in the Kirchhoff integral. In Sect. **3** the geometric approximation to Lorentz contrast is derived, and the fluxon criterion is established to determine the limit of validity of that approximation. Section **4** presents a discussion of the most obvious diffraction effects in Lorentz microscopy. The quantum of flux appears directly in the image of domain walls in the defocused and the Foucault mode. Section **5** gives a derivation of the particle scattering probability and the signal to noise ratio in the image space. In Sect. **6** it is shown that the maximum contrast due to a given flux inhomogeneity in the defocused mode is nearly equal to the square root of the quantum signal to noise ratio for both strong and weak objects. In Sect. **7** the number phase uncertainty relation is applied to determine the best pos-

sible accuracy of the measurement of a phase shift as function of the illumination time. The domain wall is treated as an example. In Sect. **8** two ways to separate experimentally electric and magnetic contrast are proposed.

2. Electron wave function in the presence of a thin magnetic object.

The following is a derivation of the connection between the nonrelativistic Schrödinger wave function of the electron in the presence of a thin magnetic field distribution and the wave function without this field. Two approximations are made, which reflect realistically the experimental situation: 1) The Kirchhoff boundary conditions are assumed to hold (*i.e.* $\psi = 0$ and $\nabla\psi = 0$ on an integrating surface except over a small part of it). 2) The electrons propagate nearly parallel to the optical axis, before and after the magnetic object.

The derivation avoids the *ad hoc* use of the Aharonov-Bohm effect ([3,4]) which suffers from conceptual difficulties, since the choice of the geometric trajectories in the path integrals of the vector potential remains unjustified. Instead, well-understood concepts of wave propagation are employed, based on the Kirchhoff formalism and its stationary phase approximation.

Let the single component stationary Schrödinger wave function of the electrons be

$$\psi(\boldsymbol{r}) = \chi(\boldsymbol{r}) \exp[i\varphi(\boldsymbol{r})] \qquad (\chi, \varphi \text{ real}). \tag{1}$$

If there are no electric or magnetic fields, ψ satisfies

$$\mathscr{H}\psi = E\psi \qquad \left(\mathscr{H} = -\frac{\hbar^2\nabla^2}{2m}, \qquad E = \frac{\hbar^2 k^2}{2m}\right). \tag{2}$$

Equation (2) is identical with the Helmholtz equation

$$(\nabla^2 + k^2)\psi = 0 \qquad (k = 2\pi/\lambda = p_0/\hbar). \tag{3}$$

It is well known that if ψ satisfies eq. (3), and if ψ and $\nabla\psi$ are given on a closed surface S, ψ and $\nabla\psi$ can be calculated everywhere inside S by the Kirchhoff integral

$$\psi(\boldsymbol{r}_p) = (4\pi)^{-1}\int_S \mathrm{d}\boldsymbol{r}_s \left[r_{sp}^{-1} \exp[-ikr_{sp}]\, \nabla\psi(\boldsymbol{r}_s) - \psi(\boldsymbol{r}_s)\, \nabla(r_{sp}^{-1} \exp[-ikr_{sp}])\right] \tag{4}$$

$(r_{sp} \equiv |\boldsymbol{r}_s - \boldsymbol{r}_p|)$.

In practice, in an electron beam with small angular spread, $\psi(\boldsymbol{r})$ is calculated from the values of $\chi(\boldsymbol{r}_s)$ and $\varphi(\boldsymbol{r}_s)$ on a finite planar cross-section perpendicular to the beam in another parallel cross-section further down the electron path. If one chooses the z-axis of the co-ordinate system parallel to the mean momentum in the beam, one has $\partial\varphi(\boldsymbol{r})/\partial z \approx k$, since k_x, $k_y \ll k_z \leqslant k$. If one also assumes realistically $|\nabla\chi| \ll k$, *i.e.* that the amplitude of the wave function changes unnoticeably over the electron wavelength, eq. (4) reduces to

$$\psi(\boldsymbol{r}_j) = (2\pi)^{-1}\int_J \mathrm{d}\boldsymbol{r}_i\, ik\chi(\boldsymbol{r}_i) r_{ij}^{-1} \exp\left[i(kr_{ij} + \varphi(\boldsymbol{r}_i))\right] . \tag{5}$$

Here, $\boldsymbol{r}_j = (x, y, z_j)$ is a co-ordinate vector with its tip in the plane J, and $a_{ij} = z_i - z_j$. If a magnetic field $\boldsymbol{B}$ is switched on *somewhere* in space, it will in general give rise to a finite and co-ordinate dependent vector potential *everywhere* in space, even in regions where $\boldsymbol{B} = 0$. Then the Hamiltonian eq. (2) will go over to

$$\mathscr{H}' = (2m)^{-1}(-i\hbar\nabla + e\boldsymbol{A})^2 . \tag{6}$$

In this case ψ', the eigenfunction of $\mathscr{H}'$, no longer satisfies the simple Helmholtz eq. (3). This makes the Kirchhoff integral eq. (4) in general inapplicable for a calculation of ψ' from one plane to the next. Fortunately, however, if the fields are sufficiently weak ψ' can still be calculated with eq. (4) even in the field region, if this equation is used in conjunction with a series of appropriate gauge transformations of the vector potential.

The vector potential has the following properties:

In the field region:

$$\boldsymbol{B}(\boldsymbol{r}) = \nabla \times \boldsymbol{A}(\boldsymbol{r}) \qquad (\boldsymbol{B} \neq 0) . \tag{7}$$

Outside of the field region:

$$\boldsymbol{A}(\boldsymbol{r}) = \nabla\Lambda(\boldsymbol{r}) \qquad (\boldsymbol{B} = 0,\ \Lambda \text{ scalar}) . \tag{8}$$

It is well known that if eq. (8) holds, then

$$\psi'(\boldsymbol{r}) = \psi_0(\boldsymbol{r}) \exp\left[ie\Lambda/\hbar\right] . \tag{9}$$

$\psi_0(\boldsymbol{r})$ is the wave function *without* field. Therefore, in those regions of space where $\boldsymbol{B} = 0$, the wave function $\psi'(\boldsymbol{r})$ can be simply calculated by first setting $\boldsymbol{A} = 0$ and finding $\psi_0(\boldsymbol{r})$ via the Kirchhoff integral, and then writing down $\psi'(\boldsymbol{r})$ according to eq. (9). In the field region itself, a similar procedure can be adopted if the fields are so small that the vector potential changes only

slowly over one wavelength, or if

$$\lambda(e/\hbar)\partial A_i/\partial_j \ll k \qquad (i, j = x, y, z)\,. \tag{10}$$

This condition (which is related with the condition for applicability of the WKB approximation) is easily fulfilled in Lorentz microscopy. The field region can then be subdivided by many equidistant parallel planes such that $\boldsymbol{A}$ can be considered constant between two neighboring planes. Since $\boldsymbol{A} = \text{const}$ implies that eq. (8) is valid again, the wave function $\psi'(\boldsymbol{r})$ can be calculated successively and unambiguously across the field region in the same fashion as outside the field region with eqs (5) and (9).

Consider Fig. 1. The weak static magnetic field distribution is contained in a thin sheet between two parallel planes B and C. The distance $a_{bc} \equiv a$ is in general larger than the thickness of any film which might carry the magnetic field sources, because it must contain the external stray fields as well. The regions above and below the sheet are field free. A beam of fast, monoenergetic coherent electrons penetrates the illuminating plane A, the field sheet and the observation plane D. The wave function $\psi_0(\boldsymbol{r}_b)$ in the plane B is calculated from $\psi_0(\boldsymbol{r}_a)$ and $\nabla\psi_0(\boldsymbol{r}_a)$ with eq. (5) while holding $\boldsymbol{A} = 0$ and is then transformed to its form ψ' in the presence of the field by eq. (9). Next consider the propagation of the wave through the field sheet. Whereas there are no restrictions on the distances a_{ab} and a_{cd}, a is subjected to

$$a \ll k/k_x^2\,, \qquad a \ll k/k_y^2\,. \tag{11}$$

Here k_x and k_y are the largest lateral momentum components in the image space in the presence of the field. This condition assures that there is no contrast due to the magnetic field at the plane C. It also makes it possible to evaluate eq. (5) explicitly from B to C with the stationary phase approximation.

First assume $\boldsymbol{B} = 0$. If during the integration of eq. (5) the distance r_{bc} has changed such that $k\,\Delta r_{bc} \gg 1$ while still $\Delta r_{bc}/r_{bc} \ll 1$ the integrand will have undergone many oscillations with no noticeable change of amplitude. Clearly, the only lasting contributions to the integral come when the exponential does not oscillate during a variation of the integration variable, *i.e.* when

$$\partial\varphi_c/\partial x_b = 0\,, \qquad \partial\varphi_c/\partial y_b = 0 \qquad (\varphi_c \equiv k r_{bc} + \varphi(\boldsymbol{r}_b))\,. \tag{12}$$

This is the condition of stationary phase. It determines the center co-ordinate $\boldsymbol{r}_b^0$ of a small area of the plane B from which $\psi_0(\boldsymbol{r}_c)$ is built up. The extent of this area is determined roughly by the beginning of oscillations of the

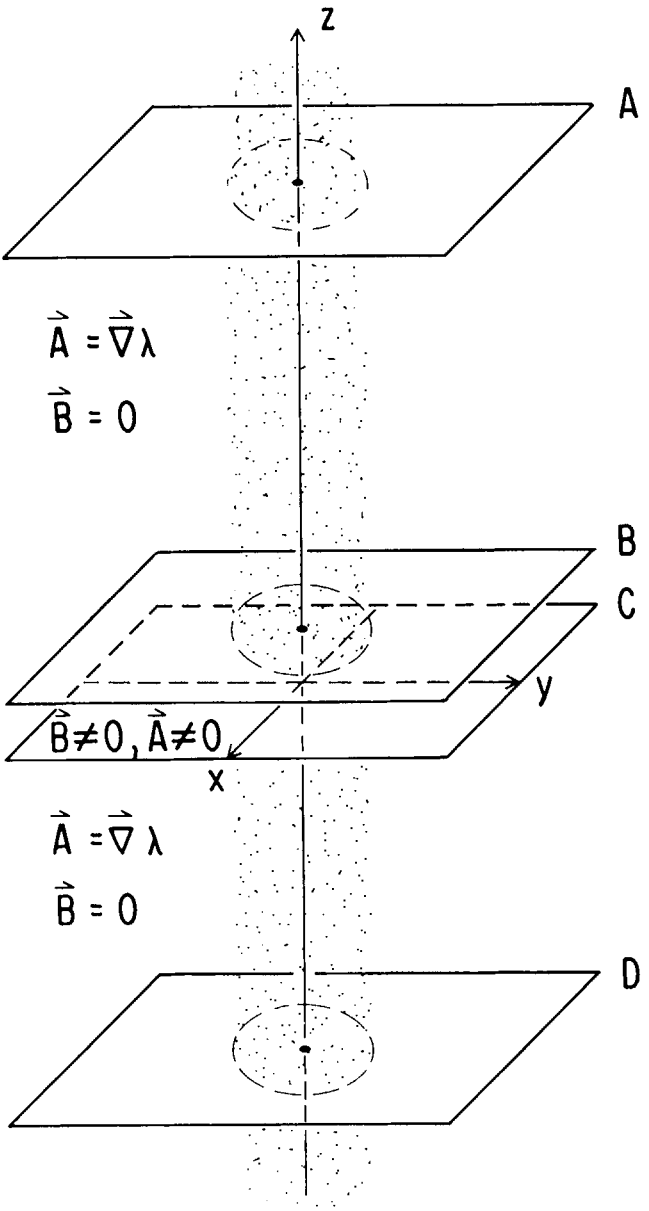


Fig. 1. – Reference planes in a beam of fast electrons with small divergence, scattering off a weak object.

exponential, *i.e.* by

$$\Delta\varphi_c = \varphi_c(\boldsymbol{r}_b^0 + \Delta\boldsymbol{r}) - \varphi_c(\boldsymbol{r}_b^0) \approx \pi/2 . \tag{13}$$

If $a^{-1} \ll k$ and $k_x, k_y \ll k$, an expansion of $\varphi_c(\boldsymbol{r}_b)$ about $\boldsymbol{r}_b^0$ to second order is sufficient far beyond $\Delta\varphi_c = \pi/2$. Thus the contributing area has the shape of an ellipse with two orthogonal axes Δx_b^0 and Δy_b^0 determined by

$$\frac{(\Delta x_b^0)^2}{2}\frac{\partial^2\varphi_c}{\partial x_b^2}(\boldsymbol{r}_b^0) \approx \pi/2 , \qquad \frac{(\Delta y_b^0)^2}{2}\frac{\partial^2\varphi_c}{\partial y_b^2}(\boldsymbol{r}_b^0) \approx \pi/2 . \tag{14}$$

The distance r_{bc} can be expanded

$$r_{bc} = (a^2 + (x_b - x_c)^2 + (y_b - y_c)^2)^{\frac{1}{2}} \approx a + (2a)^{-1}[(x_b - x_c)^2 + (y_b - y_c)^2] . \tag{15}$$

Inserting eq. (15) in eq. (14) yields

$$(\Delta x_b^0)^2 = \pi(k/a + \partial^2\varphi/\partial x_b^2)^{-1} , \qquad (\Delta y_b^0)^2 = \pi(k/a + \partial^2\varphi/\partial y_b^2)^{-1} . \tag{16}$$

Now, since $\partial\varphi/\partial x \approx k_x$, $\partial\varphi/\partial y \approx k_y$, if k_x, $k_y \ll k$, it is easy to show that the second derivatives of φ must be of order k_x^2, k_y^2. Therefore, the condition (11) on a implies that the second derivatives can be neglected in eq. (16). The limit of integration of eq. (5) then simply becomes a circle with radius $(\pi a/k)^{\frac{1}{2}} = (a\lambda/2)^{\frac{1}{2}}$, and eq. (5) yields for $\psi_0(\boldsymbol{r}_c)$, the wave function at $\boldsymbol{r}_c$ *without* field

$$\psi_0(x, y, 0) = i\chi_0(x, y, a)\exp[i(ka + \varphi_0(x, y, a))] . \tag{17}$$

Here it is assumed that χ and φ change so slowly with x and y that $\chi_0(x_b^0, y_b^0, a) \approx \chi_0(x_c, y_c, a)$ etc.

It must be emphasized that the wave function $\psi_0(\boldsymbol{r}_c)$ is exclusively determined by amplitude and phase of the wave function in a small circle with area $\pi a\lambda/2$ centered at $\boldsymbol{r}_b^0$. *It is independent of* $\psi(\boldsymbol{r})$ *anywhere else in the sheet.* Moreover, the wave functions in $\boldsymbol{r}_b^0$ and $\boldsymbol{r}_c$ are identical in the absence of a field, apart from the geometric phase shift ka. Finally, a calculation of $\psi_0(\boldsymbol{r}_c')$ in the neighborhood of the point $\boldsymbol{r}_c$ can depend on $\psi_0(\boldsymbol{r}_b)$ in the circle around $\boldsymbol{r}_b^0$ only, if $|\boldsymbol{r}_c' - \boldsymbol{r}_c| \leqslant (a\lambda/2)^{\frac{1}{2}}$, *i.e.* $\psi_0(\boldsymbol{r}_b)$ and $\psi_0(\boldsymbol{r}_c)$ depend on each other only within a cylinder with radius $(a\lambda/2)^{\frac{1}{2}}$ and axis $(\boldsymbol{r}_b^0 - \boldsymbol{r}_c)$.

Treating now the case of finite $\boldsymbol{B}$ in the sheet, it is assumed that eq. (10) holds. The sheet is subdivided by $N \gg 1$ parallel planes at distance a/N. Assume $\boldsymbol{A}(\boldsymbol{r}) = \boldsymbol{A}_n = \text{const}$ within a cylinder of radius $(a\lambda)^{\frac{1}{2}}$ between neighboring planes (Fig. 2). The problem of calculating $\psi'(\boldsymbol{r}_n)$ from $\psi'(\boldsymbol{r}_{n-1})$ is the

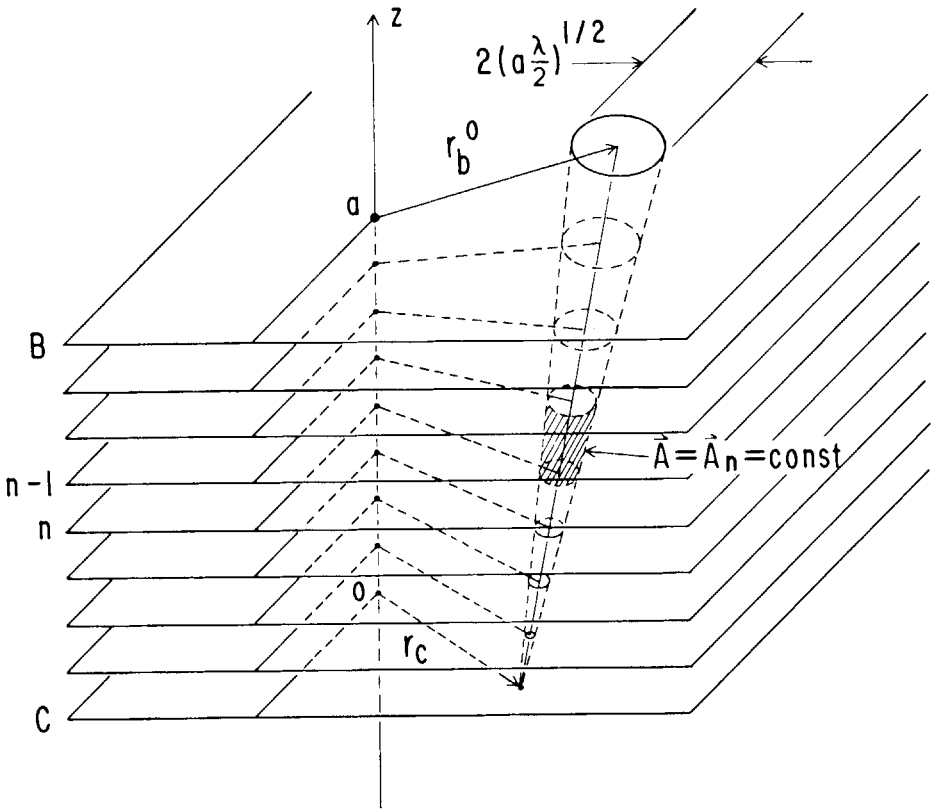


Fig. 2. – Development of the electron wave function in the magnetic field region.

same as calculating $\psi_0(\boldsymbol{r}_c)$ from $\psi_0(\boldsymbol{r}_b)$, except that the result which corresponds to eq. (17) must now be gauge transformed according to eq. (9). With this procedure a magnetic phase shift $\Delta\varphi$ is picked up at every plane.

$$\Delta\varphi_{n,n-1} = \frac{e}{\hbar}\frac{a}{N}(\nabla A)_n \cdot \frac{(\boldsymbol{r}_n - \boldsymbol{r}_{n-1}^0)}{|\boldsymbol{r}_n - \boldsymbol{r}_{n-1}^0|} \approx \frac{e}{\hbar}\frac{a}{N}\left(\frac{\partial A}{\partial z}\right)_n . \tag{18}$$

The total magnetic phase shift from $\boldsymbol{r}_b^0$ to $\boldsymbol{r}_c$ is a sum of N such terms which, for sufficiently large N goes to the integral

$$\Delta\varphi(x, y) = \frac{e}{\hbar}\int_a^0 \mathrm{d}z\, A_z(x, y, z) . \tag{19}$$

This path integral determines the phase unambiguously because A affects ψ' only locally according to eq. (6) so that ψ' and therefore also A *outside* the cylinders around the line $(r_b^0 - r_c)$ do not have any effect on the build up of $\psi'(r_c)$. Thus the wave function $\psi'(r_c)$ in the presence of the field follows from $\psi'(r_b^0)$ according to

$$\psi'(r_c) = i\psi'(r_b^0) \exp\left[i\left\{ka + \frac{e}{\hbar}\int_a^0 dz\, A_z(r_b^0 - r_c)\right\}\right], \tag{20}$$

or

$$\psi'(x, y, 0) = i\psi'(x, y, a) \exp\left[i\left\{ka + \frac{e}{\hbar}\int_a^0 dz\, A_z(x, y, z)\right\}\right].$$

Using eq. (9)

$$\psi_0(x, y, 0) \exp\left[i\frac{e}{\hbar}\Lambda(x, y, 0)\right] = i\psi_0(x. y, a) \exp\left[i\left\{\frac{e}{\hbar}\Lambda(x, y, a) + \right.\right.$$
$$\left.\left. + ka + \frac{e}{\hbar}\int_a^0 dz\, A_z(x, y, z)\right\}\right]. \tag{21}$$

The path integrals Λ must be referred to a common value, for which $\Lambda(0, 0, 0)$ is chosen here:

$$\left.\begin{aligned} \Lambda(x, y, a) &= \Lambda(0, 0, 0) + \frac{e}{\hbar}\int_0^a dz\, A_z(0, 0, z) + \frac{e}{\hbar}\int_{0,0}^{x,y} d\boldsymbol{r}_b' \boldsymbol{A}(x', y', a)\,, \\ \Lambda(x, y, 0) &= \Lambda(0, 0, 0) + \frac{e}{\hbar}\int_{0,0}^{x,y} d\boldsymbol{r}_c' \boldsymbol{A}(x', y', 0)\,. \end{aligned}\right\} \tag{22}$$

For the reasons just discussed, the difference between $\Lambda(0, 0, 0)$ and $\Lambda(0, 0, a)$ on both sides of the sheet near the origin is again unambiguous. Before inserting these expressions into eq. (21), a final gauge transformation is performed by multiplying both sides of eq. (21) with $\exp\left[-(ie\Lambda(x, y, 0)/\hbar)\right]$. This corresponds to adding a function $-\boldsymbol{A}(\boldsymbol{r}) = -\nabla\Lambda(\boldsymbol{r})$ which is chosen such that *in the space between planes C and D* it cancels completely that vector potential $\boldsymbol{A}(\boldsymbol{r})$ which arose from B in the sheet. This gauge transformation permits again the use of the Kirchhoff formalism in the image space between C and D. It also changes the vector potential in the sheet and above, but cannot cancel it in all spaces simultaneously. We call this gauge the Kirchhoff gauge. The wave function in r_c in the presence of the magnetic field is

obtained from eqs (21) and (22). Written in the Kirchhoff gauge (denoted by double primes) it is

$$\psi''(\boldsymbol{r}_c) = \psi'(x, y, 0) = i\psi_0(x, y, a) \exp\left[i\left\{\frac{e}{\hbar}\oint_L \boldsymbol{A}(\boldsymbol{r})\,\mathrm{d}\boldsymbol{r} + ka\right\}\right]. \tag{23}$$

The contour integral in eq. (23) runs around a rectangle with corners (0, 0, 0), (0, 0, a), (x, y, a) and (x, y, 0). It is clearly gauge invariant. Stokes law connects this contour integral with the enclosed magnetic flux Φ

$$\oint_L \mathrm{d}\boldsymbol{r}\cdot\boldsymbol{A}(\boldsymbol{r}) = \int_F \mathrm{d}\boldsymbol{f}\cdot\boldsymbol{B}(\boldsymbol{r}) = \Phi(F). \tag{24}$$

F is an area bordered by the contour L. In the present case, F is completely specified by the choice of $\boldsymbol{r}_c = (x, y, 0)$. Finally, in view of eq. (17) one obtains

$$\psi''(\boldsymbol{r}_c) = \psi_0(\boldsymbol{r}_c) \exp\left[i(e/\hbar)\,\Phi(\boldsymbol{r}_c)\right]. \tag{25}$$

In the Kirchhoff gauge, the Schrödinger wave function behind a weak magnetic object is simply the wave function without magnetic object with a phase shift proportional to the magnetic flux in the object.

The wave function $\psi''(\boldsymbol{r}_d)$ in the plane D, which is reproduced by the instrument in the photographic plate, follows from $\psi''(\boldsymbol{r}_c)$ (eq. (25)) via the Kirchhoff integral eq. (5):

$$\psi''(\boldsymbol{r}_d, p_0, \Phi, h) = (p_0/h)\int_C \mathrm{d}\boldsymbol{r}_c \psi_0(\boldsymbol{r}_c, p_0) r_{cd}^{-1} \exp\left[(i/\hbar)[p_0 r_{cd} + e\Phi(\boldsymbol{r}_c)]\right]. \tag{26}$$

In this equation, ψ'' is written to depend explicitly on all experimental variables, *i.e.* on $\boldsymbol{r}_d$, $p_0 = (2mE)^{\frac{1}{2}}$ and Φ and also on h.

The measured quantity is the time integrated probability current density $\boldsymbol{j}(\boldsymbol{r}_d)$, which is given by

$$\int_{t_0}^{t_1} \mathrm{d}t\,\boldsymbol{j}(\boldsymbol{r}_d) = \varrho(\boldsymbol{r}_d)\,\boldsymbol{v}(\boldsymbol{r}_d)(t_1 - t_0). \tag{27}$$

Since $v \approx \hbar k/m$ does not depend noticeably on $\boldsymbol{r}_d$ if $k \gg k_x$, k_y, the information on the flux distribution is extracted solely from the probability density

$$\varrho(\boldsymbol{r}_d, \Phi, p_0, h) = |\psi''(\boldsymbol{r}_d, \Phi, p_0, h)|^2. \tag{28}$$

Equations (26) and (28) constitute the basis for all calculations of Lorentz contrast. According to eq. (26), $\varrho(\boldsymbol{r}_d)$ depends in general on the magnetic flux distribution *everywhere* in the sheet. Of course different regions of the sheet can contribute with very different weight, depending on the distance a_{cd} between the planes C and D (« defocusing distance »), on the flux distribution $\Phi(\boldsymbol{r}_c)$, on the initial energy $E = p_0^2/2m$ and on the magnitude of the quantum of action.

3. Wave optical *vs.* geometric Lorentz contrast.

We shall now study the effect of varying h in order to find the correspondence limit of eq. (26). In order to simplify the discussion, a one-dimensional magnetic object is chosen as example, which is illuminated by a plane wave. The phase of $\psi''(\boldsymbol{r}_c)$ can be written as

$$\varphi_c(\boldsymbol{r}_c) = \varphi_0(\boldsymbol{r}_c) + (e/\hbar)\Phi(\boldsymbol{r}_c)\,. \tag{29}$$

With a plane incoming wave propagating parallel to the z-axis, and in the absence of electrostatic scattering, $\varphi_0(\boldsymbol{r}_c) = \text{const}$. Let the magnetic field in the sheet have only a y component, B_y, which depends only on x. Then

$$\Phi(x) = a\int_0^x B_y(x')\,\mathrm{d}x'\,. \tag{30}$$

Consider the stationary points x_c^0 in the integral eq. (26). They follow with eqs (29) and (30) from

$$\hbar\partial\varphi_c/\partial\xi_c = (x_c^0 - x_d)(p_0/d) + aeB_y(x_c^0) = 0\,. \tag{31}$$

Here a_{cd}, the defocusing distance, was renamed $a_{cd} \equiv d$ and kr_{cd} was expanded as in eq. (15), which is permissible if $dk \gg 1$. Equation (31) always has at least one, but can have several solutions. The number of solutions and their values depend on $B_y(x)$, d and p_0. However, note that they do not depend on h! Therefore the stationary points must somehow retain their significance in the correspondence limit. On the other hand, for each solu-

tion x_c^0, the area of contribution of $\psi''(\boldsymbol{r}_c)$ to $\psi''(\boldsymbol{r}_d)$ is determined by an expression analogous to eq. (13)

$$\left.\begin{aligned} &\Delta\varphi_c = \varphi_c(x_c^0 + \Delta x) - \varphi_0(x_c^0) \approx \pi/2\,, \\ &(\Delta x \equiv x - x_c^0; \quad \varphi_c \equiv \hbar^{-1}(p_0 r_{cd} + e\Phi))\,. \end{aligned}\right\} \tag{32}$$

$\Delta\varphi_c$, and therefore Δx *does* depend on h. In order to find the behaviour of the wave function in the classical limit, let $h \to 0$. Clearly, whatever the choice of p_0 and d, and whatever $\Phi(x)$, the limits Δx of the contributing areas centered at the stationary points will shrink. The wave function in $\boldsymbol{r}_d$ will depend less and less on the wave function *in the neighborhood* of the $\boldsymbol{r}_c^0$, until finally, in the limit $h = 0$, it will be determined by $\psi(\boldsymbol{r}_c^0)$, the wave function at the stationary points *only*. While the contributing area shrinks, the number of mutually independent areas in the integration plane C increases at the same rate. Then the probability current can be considered to run in very many separated fine straight cylinders from plane C to D. These cylinders are identified with the geometric trajectories. The appropriate transformation of the probability density from C to D is then to map an area element of C, $\mathrm{d}f_c \gg \lambda y$ (which contains many independent areas λd) along the trajectories defined by eq. (31) into the corresponding area elcment $\mathrm{d}f_d$ in D, such that the probability current element $v\mathrm{d}P = v\varrho\mathrm{d}f$ is conserved:

$$\varrho(\boldsymbol{r}_c^0)\,\mathrm{d}f_c^0 = \varrho(\boldsymbol{r}_d)\,\mathrm{d}f_d\,. \tag{33}$$

In the one-dimensional example

$$\varrho(x_c^0)\,\mathrm{d}x_c^0 = \varrho(x_d)\,\mathrm{d}x_d\,, \tag{33a}$$

with eq. (31)

$$\frac{\varrho(x_d)}{\varrho(x_c^0)} = \left(\frac{\mathrm{d}x_d}{\mathrm{d}x_c^0}\right)^{-1} = \left(1 + \frac{dae}{p_0}\,\frac{\partial B(x_c^0)}{\partial x_c^0}\right)^{-1}. \tag{34}$$

Equation (34) is the equivalent of eq. (28) in the correspondence limit; in other words, it is the formula for geometric optical contrast in Lorentz microscopy.

Equation (34) can of course be derived more directly from classical concepts. For the one-dimensional example, the classical Lorentz force in the sheet is

$$F_x(x, z) = (p_0 e/m) B_y(x)\,. \tag{35}$$

The momentum change which an electron experiences while passing the sheet is with $v = p_0/m$

$$p_x(x) = \int_{t_a}^{t_e} \mathrm{d}t F_x(x, z) = \int_{a}^{0} (\mathrm{d}z/v) F_x(x, z) = eaB_y(x) = p_0 \alpha(x) . \tag{36}$$

$\alpha(x)$ is the deflection angle. An electron which penetrates the sheet parallel to the z axis at x_c^0 will therefore penetrate the plane D at

$$x_d = x_c^0 + \alpha(x) d = x_c^0 + (ead/p_0) B_y(x) . \tag{37}$$

This equation is identical with eq. (31), from which in turn eq. (34) was derived with the assumption of well-defined geometric trajectories.

The quantum of action is finite. This raises the question whether the somewhat cumbersome expression eq. (26) or its simple approximation eq. (34) is applicable in practice. It turns out that a case must be made for both. Initially the importance of diffraction effects in Lorentz microscopy was overlooked by most workers in the field. Common experience misled to the belief that quantum effects in contrast of aperiodic objects are to be expected only when the diffraction at the aperture of the objective lens begins to limit the spatial resolution. However this is only one possible effect and is tied to direct observation of the object: With « in focus » operation the distance between the observation plane and the field region is of the order of the sheet thickness, so that eq. (11) holds also for the « defocusing distance ». Then the diffraction effects associated with the development of the wave function from the scattering object to the plane of the observation can be neglected against those which are associated with aberrations in the objective lens. For example, the radius of the contributing area would be $\Delta x_c \approx 0.5$ Å for $d = 5$ Å and $\lambda = 0.05$ Å. This distance, which is also equal to the distance of the first two Fresnel fringes in D for an absorbing half-plane in C, is smaller than the width of fringes which appear in the image due to Abbe's resolution criterion, if an objective lens of say 5 Å resolution is used with its optimal aperture. However, with increasing defocusing distance d the diffraction fringes associated with the development of the wave function from the scattering object to the observation plane increase in width like the contributing area, *i.e.* like $(d\lambda)^{\frac{1}{2}}$, whereas those associated with the diffraction on the objective aperture remain the same, namely $f\gamma$ (f is the focal length, γ the aperture angle). Eventually, $(d\lambda)^{\frac{1}{2}}$ will become larger than $f\gamma$.

To be sure, it was well known among specialists that the width of diffraction fringes increases with the square root of the defocusing distance. Already in 1943 Boersch ([5]) had followed experimentally the development of Fresnel fringes in the image space behind the opaque half plane and in 1952 Glaser ([6]) published more general theoretical drawings of the development of the Schrödinger wave function through the image space. These aspects of wave propagation have limited practical importance for in focus operation of the electron microscope and were therefore less known among material scientists. Lorentz microscopy was the first practically interesting case of electron phase microscopy at large defocusing distance. More recently, the quantitative experiments with phase contrast of amorphous objects by Thon ([7]) have drawn attention to the valuable information in out of focus images of electrostatic objects as well.

There is of course a large body of knowledge about phase contrast in light optics, much of which is concerned with the exact analogon of phase contrast in electron microscopy, although on a different scale. An important review article on light phase contrast with diffraction effects was written by Wolter ([8]) in 1956.

The onset of quantum effects in the contrast of practical Lorentz objects can be estimated without regard to a particular experimental detection mode with the Heisenberg uncertainty relation. Consider Fig. 3. Two trajectories emerge from the plane C at x_1 and x_2 after having penetrated the field distribution sketched below. Since the fields in x_1 and x_2 differ, there is a corresponding difference in lateral momentum

$$\Delta p_x = ea\Delta B\,. \tag{38}$$

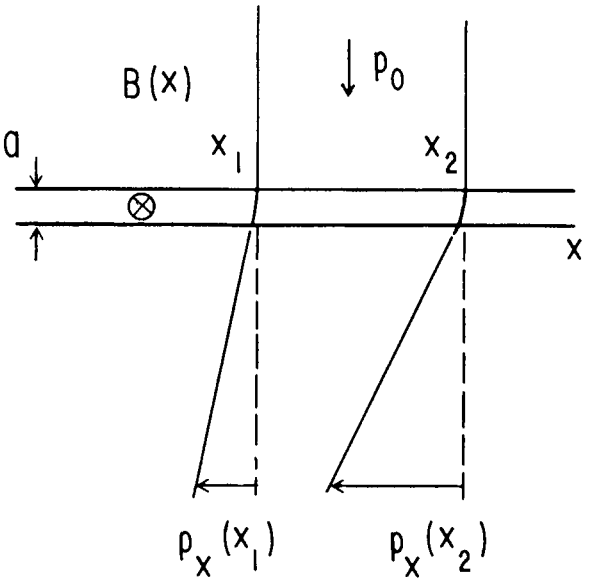
According to Heisenberg, the accuracy of a simultaneous measurement of the canonically conjugate variables x and p_x is

$$\Delta p_x \Delta x \geqslant h\,. \tag{39}$$

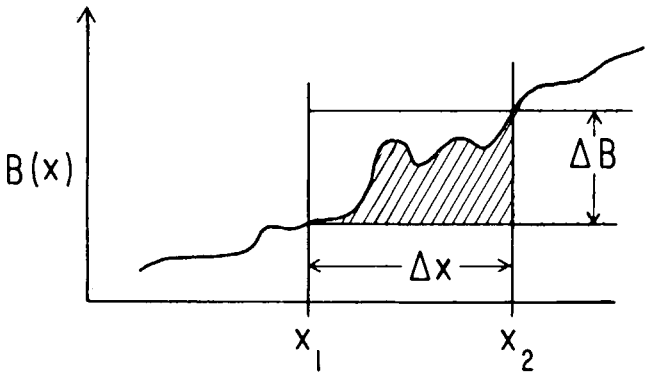
Inserting eq. (38) into (39) yields

$$\Delta\Phi \geqslant h/2e \qquad (\Delta\Phi = \Delta x\, a\, \Delta B/2) \tag{40}$$

if $B(x)$ is approximately represented by the first term in the Taylor expansion in the interval Δx. $\Delta\Phi$ is here the flux associated with the *difference* of B between x_1 and x_2. It will be redefined in Sect. **5**. The quantity h/e is a



$$\Delta p_x \, \Delta_x = e \, a \, \Delta B \, \Delta x \geq h$$



$$\Delta\phi \approx \Delta B \cdot a \cdot \Delta x / 2 \geq h / 2e$$

Fig. 3. – Application of the Heisenberg uncertainty relation to electron scattering from weak magnetic objects.

universal constant with the dimension of magnetic flux, $h/e = 4.7 \cdot 10^{-7}$ G c^2. We call $\Delta\Phi$, the flux associated with the *difference* of the fields between x_1 and x_2, the « inhomogeneity of flux ».

Suppose $\Delta\Phi = h/2e$. Then eq. (40) says that if the lateral momentum p_x is measured with accuracy $\Delta p_x \ll p_x(x_1) - p_x(x_2)$, it is impossible to measure also, on the *same* electron, where in the interval Δx the electron came from. Inversely, if the co-ordinate of penetration of an electron x is measured with accuracy $\Delta x \ll x_2 - x_1$, the uncertainty Δp_x of the momentum of this electron is larger than the difference of deflection at the ends of the interval. However it will be shown later, that measurements of x and p_x can be made with any desired accuracy, if the information is extracted from successive measurements on *many* electrons.

Equation (40) was initially ([9]) interpreted to mean that for $\Delta\Phi < h/2e$ the geometric approximation eq. (34) was inapplicable for contrast calculations. This rule is in general verified, in particular in all high-contrast modes, and the quantum of flux $h/2e$ often appears directly in the image, as will be shown next. There are, however, cases where the geometric approximation is never valid, even if $\Delta\Phi \gg h/2e$ (in and on the caustic mantle), and also cases where it is valid even if $\Delta\Phi \ll h/2e$ (outside the caustic mantle). The latter case will be discussed by Wade during these lectures ([10-12]).

Note that according to eq. (40) the onset of diffraction effects is independent of the energy of the incoming electron. This is a consequence of the peculiar velocity dependence of the Lorentz force, which makes the magnetic momentum transfer independent of the velocity (eq. (36)). In contrast to this, an application of the Heisenberg uncertainty relation to *electrostatic* deflections results in a resolution criterion where the transition from wave theory to geometric theory *does* depend on the incoming energy.

4. Practical manifestations of diffraction effects in Lorentz microscopy.

4·1. Domain walls in the defocused mode.

Very soon after introduction of the technique the possibility to measure the magnetization distribution in ferromagnetic domain walls and ripples attracted much attention. These structures extend over several 1000 Å and are therefore well within the spatial resolution of the transmission microscope.

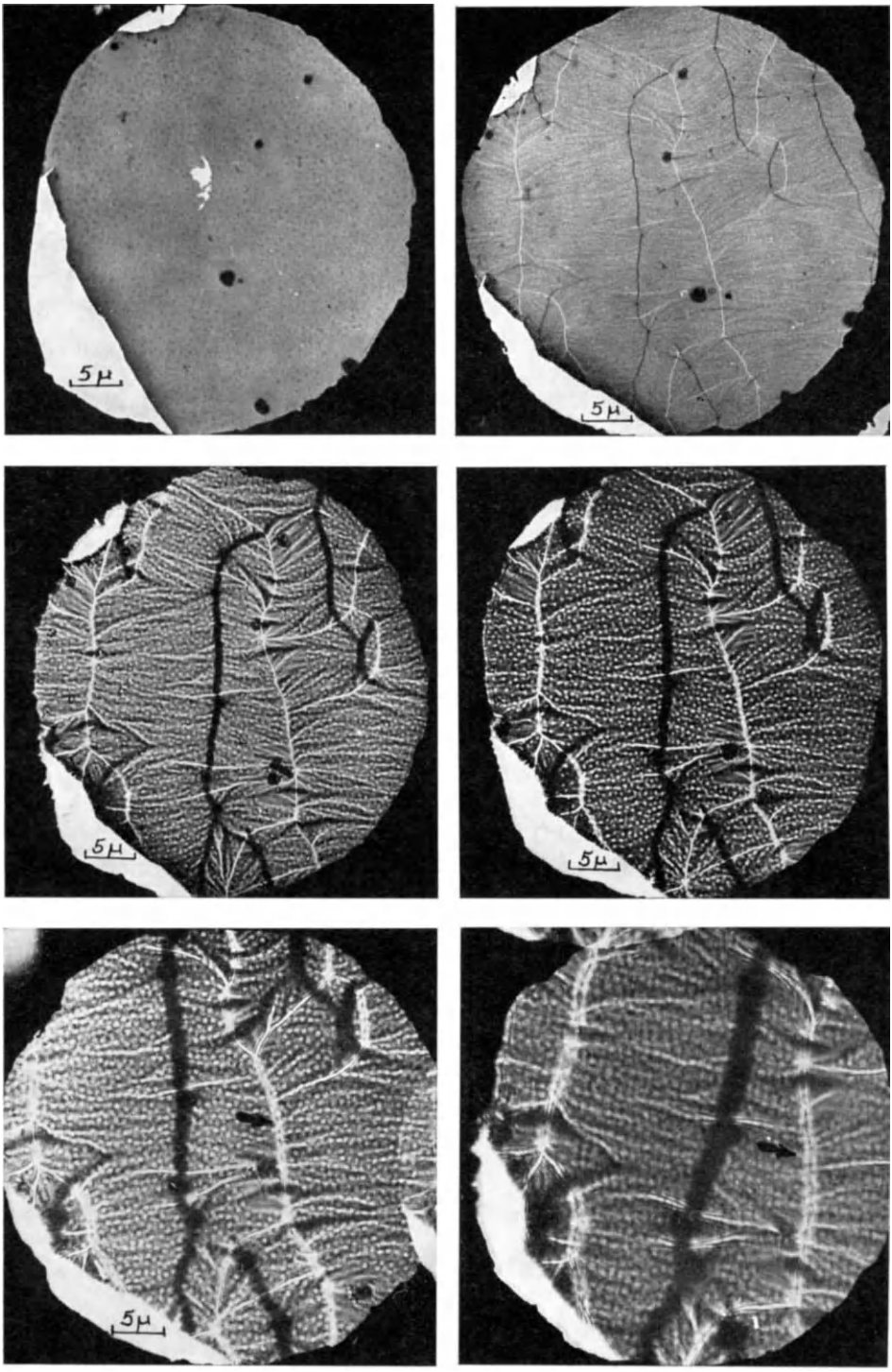


Fig. 4. – Contrast in the image plane behind a permalloy film as function of the defocusing distance.

Figure 4 shows a series of images taken from the same region of a permalloy film at increasing defocusing distance with a near point source in the illuminating plane A. No magnetic contrast is found in focus ($d = 0$). At small defocusing distance d, the domain walls begin to show up as fine dark or bright lines which broaden nearly linearly with increasing d. At still larger d the bright lines begin to split into several parallel bright fringes. The width of these fringes remains constant with respect to other distances between details in the film, while their number increases linearly with d. These fringes are the most obvious quantum effects in Lorentz microscopy. In first approximation they can be regarded as biprism interference fringes, and each such fringe can be loosely interpreted as the projection of one fluxon $h/2e$ from the magnetic sheet into the image plane.

Consider Fig. 5. Assume a ferromagnetic film in the sheet of thickness a. A wall of zero width runs at the center of the Figure (at $x = 0$) along a line perpendicular to the paper, separating two domains with homogeneous field B_p parallel and antiparallel to the wall. The geometric trajectories emerging from the near point source S at distance g above the films are deflected by an angle $+\alpha_p$ or $-\alpha_p$, depending on which side of the wall they penetrate the film. Then two wave trains overlap in the triangle below the film which seem to come from two virtual sources S' and S'' at mutual distance $2\alpha_p g$ in the illumination plane and intersect each other with the angle $2\xi = 2\alpha_p g/(g + d)$. The width of the fringes which result from the interference of these wave trains is (with eq. (36))

$$\delta_d = \lambda/2\xi = \lambda(g + d)/(g2\alpha_p) = h(g + d)/(g2eaB_p)\,. \tag{41}$$

One can refer the fringe width back into the plane C

$$\delta_c = \delta_d \cdot g/(g + d) = h/(2eaB_p)\,. \tag{42}$$

Clearly, the magnetic flux covered by the projected fringe width is

$$\Delta\Phi = B_p\,\Delta A = B_p a\delta_c = h/2e\,. \tag{43}$$

Independent of magnification and the electron energy, there is one fringe per fluxon $h/2e$ in the image of the film. Thus, the geometric approximation clearly breaks down when information on a scale finer than the fluxon is to be extracted from these regions of high contrast behind convergent walls. This is one example which confirms the original interpretation of eq. (40).

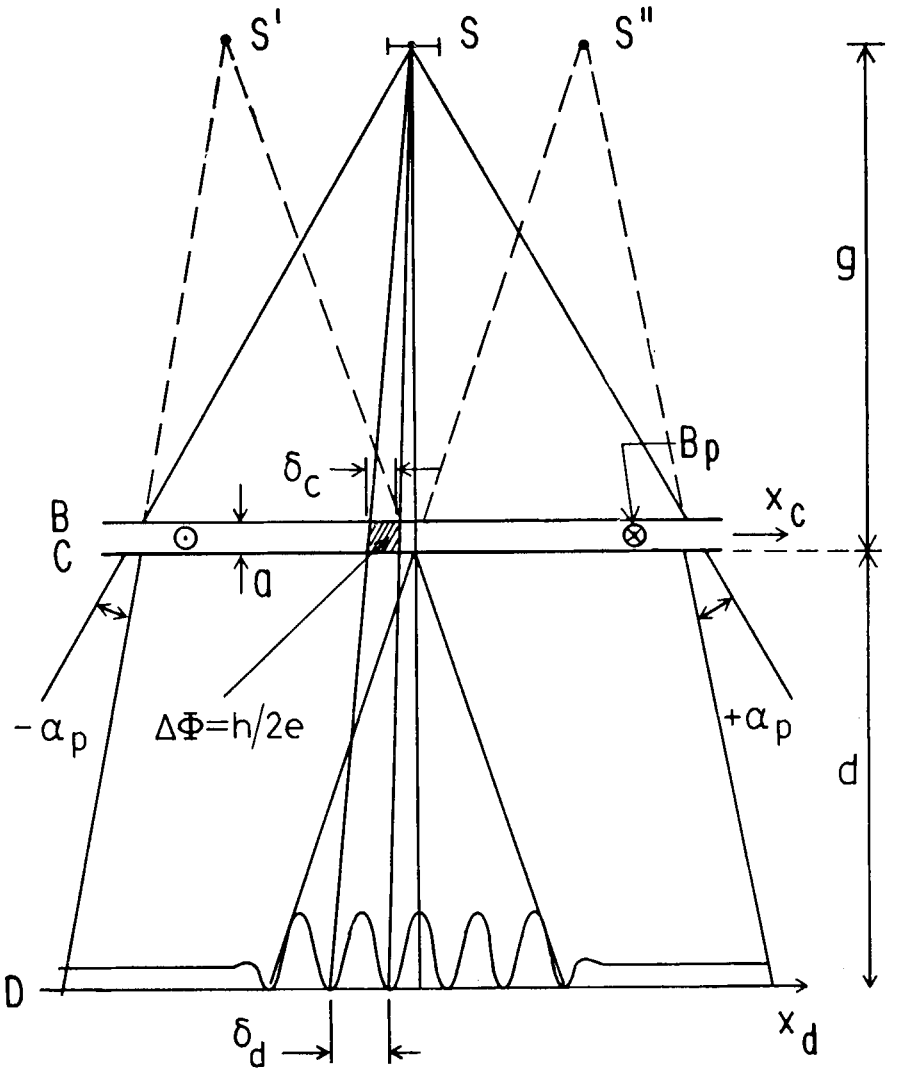


Fig. 5. – The appearance of the quantum of flux $h/2e$ in the defocused image of a convergent domain wall.

It is instructive to follow the contrast in the region behind the convergent wall as a function of the wall width and of the magnitude of the quantum of flux in order to clarify the relationship between the fundamental wave mechanical contrast (eqs (26) and (28)) and its geometric approximation (eq. (34)).

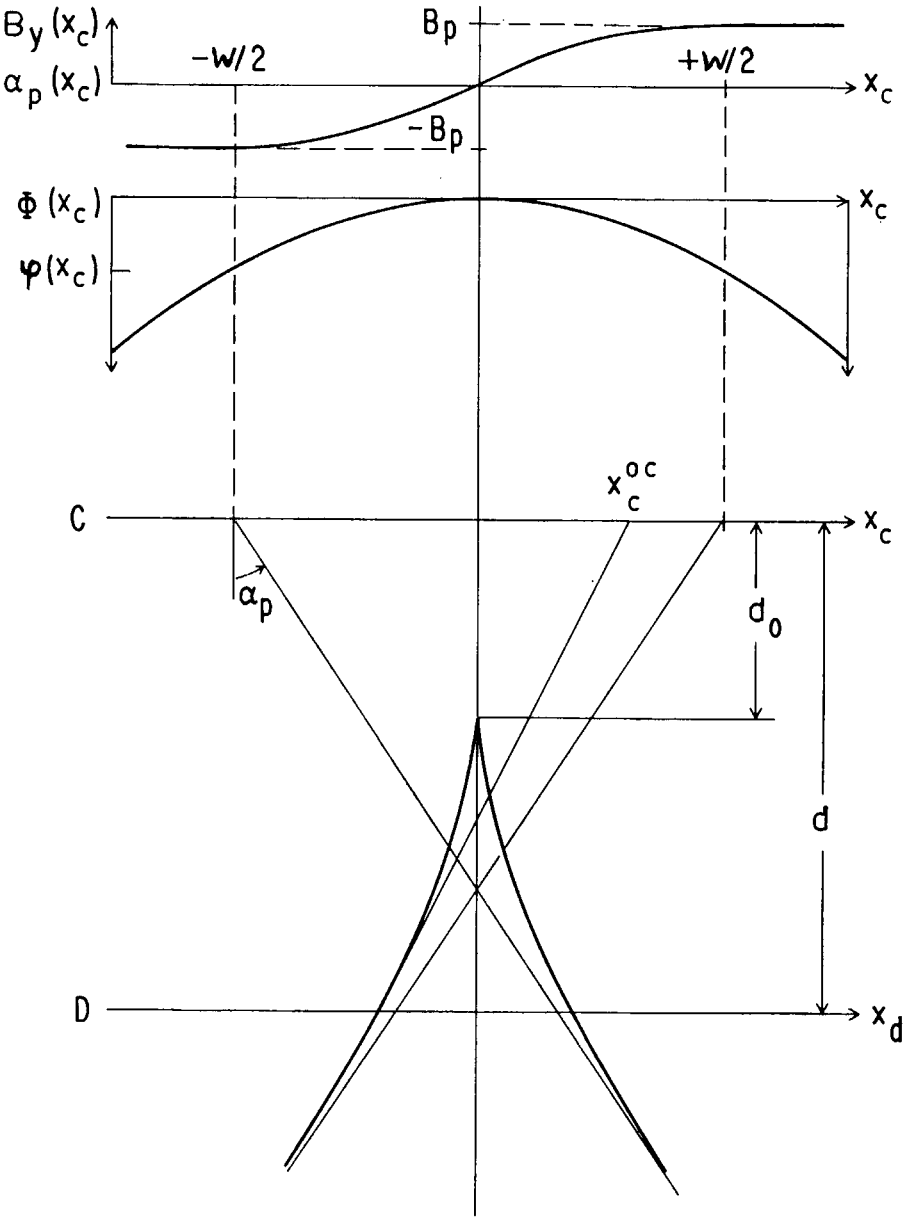


Fig. 6. – The caustic mantle behind a convergent domain wall.

Assume that the transition from $+B_p$ to $-B_p$ in the above domain wall is not a step function at $x = 0$, but occurs gradually over a characteristic distance $\Delta x = w$, the wall width (Fig. 6). Then the intensity in the image is also a continuously varying function of x_d and can, in the geometric approximation, be calculated from eq. (34). At certain points of the image plane it can happen that

$$\partial B(x_c^0)/\partial x_c^0 = -p_0/dae\,. \tag{44}$$

Then the geometric approximation predicts an infinity at x_d, the point defined by eq. (37). This point is called a caustic point.

Looking back at eq. (31), which defined the stationary points, it is clear that eq. (44) implies $\partial^2\varphi_c/(\partial x_c^0)^2 = 0$, *i.e.* that the defining equation of the geometric trajectories does not change with a small change of x_c. It follows that the caustic point is illuminated not by one, but a series of densely spaced stationary points x_c^0. Since eqs (31) and (44) are independent of h, the caustic points remain well defined and therefore ought to retain some physical significance if h is finite. The significance of caustic points to an evaluation of the Kirchhoff integral is that the limit of integration can clearly no longer be found from eq. (14), since eq. (44) is also equivalent to $\partial^2\varphi_c/\partial x_d^2 = 0$. To find the integration limit near caustic points one has to go back to eq. (13) and expand $\Delta\varphi_c$ to third order, or in general, to the next highest order derivative $\partial^n\varphi_c/\partial x_d^n$ which does not vanish in x_c^0. Clearly the limits of integration will be unusually large when φ_c changes so slowly around x_c^0, and the amplitude $\psi''(x_d)$ will build up to unusually large values in caustic points. On the other hand, it will never build up to infinity, as predicted by the geometric theory, since the contributing integration area must remain finite. Therefore, in caustic points the geometric approximation always breaks down badly.

The two-dimensional locus of all caustic points in the image space is called the caustic mantle. It is sketched schematically for the domain wall in Fig. 6. There is a minimum distance d_0 between object and all caustic points. The caustic mantle stretches to infinity asymptotically parallel to the first geometric trajectory which is deflected by the full angle $\pm\alpha_p$. The first nonvanishing derivative of φ_c is $\partial^4\varphi_c/\partial x_d^4$ in the tip of the caustic mantle, whereas everywhere else it is $\partial^3\varphi_c/\partial x_d^3$. It would seem therefore, that *the tip of the caustic mantle is the point of highest intensity in the image space of a domain wall.*

Figure 7 shows schematically the intensity distribution (eq. (28)) in an image plane which intersects the caustic mantle at a distance $d > d_0$ behind

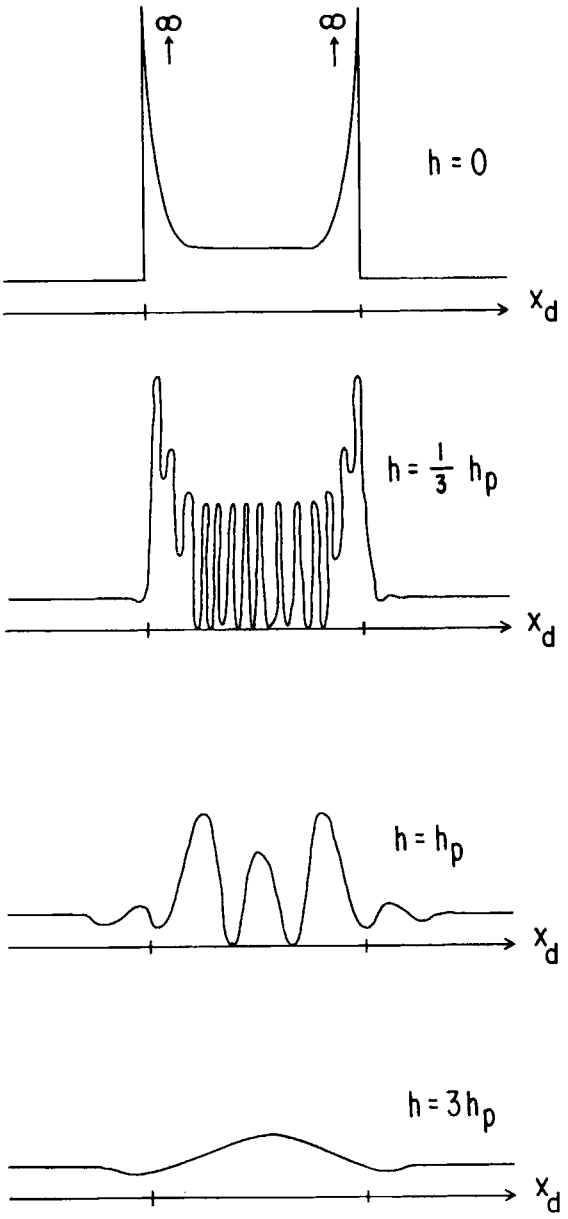


Fig. 7. – Demonstration of the correspondence principle in the strongly defocused image of a convergent domain wall.

a wall of finite width. The quantum of action grows from 0 to a large value from one image to the next. All other quantities $(B(x), p_0, g, d)$ are kept constant. Outside the caustic mantle, which stays of course fixed through all variations of h, every point of the image plane is illuminated from the neighborhood of a single stationary point in the plane C, while inside three separate stationary points contribute to the intensity. Two of the latter are outside of the wall in regions of constant $B(x_c^0) = B_p$ and contribute nearly equal amounts everywhere. The third is in the wall and contributes mostly near the caustic points, very little at the center of the image.

The geometric approximation is good nearly everywhere outside the caustic mantle, but clearly if h is finite, *on a sufficiently fine scale of observation* it is invalid nearly everywhere on and inside the caustic mantle (except in those points where the geometric intensity curve intersects the « real » curves). The fringe width near the center of the patterns (in the region of what is essentially two beam interference) is simply as before half the de Broglie wavelength associated with the maximum lateral momentum in the image space, *i.e.* $\Delta x_{fr} = (h/p_p)\cdot(g + d)/g = h(g + d)/g2eaB_p$.

Now, at sufficiently small h, the geometric curve is a rather good approximation if both the geometric and the diffraction pattern are blurred by one fringe width. This is actually often the practical situation since the source of illumination is necessarily finite, not a point source, and radiates incoherently from different points. Thus, in practice, if it is sufficient to extract information on a scale larger than biprism fringe width from the pattern, the geometric approximation is good enough in these high contrast regions of the image space. The first trouble occurs when the fringe width becomes comparable to the « half width » of the caustic peaks, *i.e.* comparable to the region where the stationary point in the wall contributes significantly. Then just one bit of information can be extracted geometrically from the pattern about the corresponding field region in the wall. Finally, there comes a point when the entire diffraction pattern can no longer be approximated by a geometric pattern even upon blurring of both. This occurs roughly when no more zeros of the intensity exist in the latter. Note that the integral of the difference between the intensity with and without object, $\int dx |\varrho(x, \Phi) - \varrho(x,0)|$ (which contains the information) remains nearly constant down to that point, but then drops to zero very fast. Note also that already before this point is reached, the maximum intensity of the diffraction patterns near the caustic peaks drops very rapidly with increasing h.

4·2. Domain walls in Foucault mode.

The quantum of flux can also appear directly in the image of the Foucault mode. Consider Fig. 8. The wave function in the image is calculated in the standard fashion for direct, in focus observation by a double Fourier transformation, first from the object plane C to the focal plane E (where the co-ordinate γ is proportional to the lateral momentum p_x in the beam), and then from E to the intermediate image plane F, (where the co-ordinate is $u = Vx_c$, with the magnification $V = 1$ in Fig. 8). Contrast in the Foucault mode is achieved by stopping with the objective aperture all electrons whose lateral momentum p_x is larger than a certain value. This is simulated mathematically by changing the integration limits of the second Fourier transformation from $-\infty$ and $+\infty$ to $-\infty$ and $\mu' \equiv \mu p_0/fh$, where μ is the co-ordinate of the aperture edge in the focal plane E and f the focal length of the objective

$$\psi''(u) = \int_{-\infty}^{\infty} dx_c \int_{-\infty}^{\mu'} d\gamma' \, \psi''(x_c) \exp\left[i\gamma'(u - x_c)\right] \qquad (\gamma' \equiv k\gamma). \tag{46}$$

In the neighborhood of the geometric image of a zero width domain wall, bordering a large domain, the wave function in F is

$$\psi''(u) = [3\pi/2 + \mathrm{Si}(\alpha' - \mu')u + i\,\mathrm{Ci}(\alpha' - \mu')u] \exp[i\alpha' u] + \\ + [-\pi/2 + \mathrm{Si}(\alpha' + \mu')u - i\,\mathrm{Ci}(\alpha' + \mu')u] \exp[-i\alpha' u].$$

Here $\alpha' = \alpha_p k = aB_p e/\hbar$ and $\mu' = \mu k/f = \mu p_0/f\hbar$. The corresponding probability density in F is shown in Fig. 8 for $\mu = 0$ and $\mu = \pm 0.4\alpha_p f$. For $\mu = 0$ the Si and Ci functions depend only on the object property α_p and show minima and maxima at distances $u_n = n\pi/\alpha' = nh/(2eaB_p)$ from the geometric edge u_w. Around u_w the width of the transition of ϱ from nearly zero to nearly ϱ_0 is determined essentially by the Si functions which have a first maximum at $u = u_w + \Delta u_1$. Thus, since $\Delta\Phi = \Delta u_1 \, aB_p = h/2e$, the width of the transition from dark to light covers a flux quantum $h/2e$ in the film as in the defocused mode. This situation is of course also independent of magnification and electron energy. It will be complicated in a finite wall, and ψ'' can then no longer be found analytically. However, it seems obvious that the intensity in the image will go from nearly zero at u_1 to nearly ϱ_0 at u_2 (*i.e.* the pattern will show nearly infinite contrast) only when the phase shift

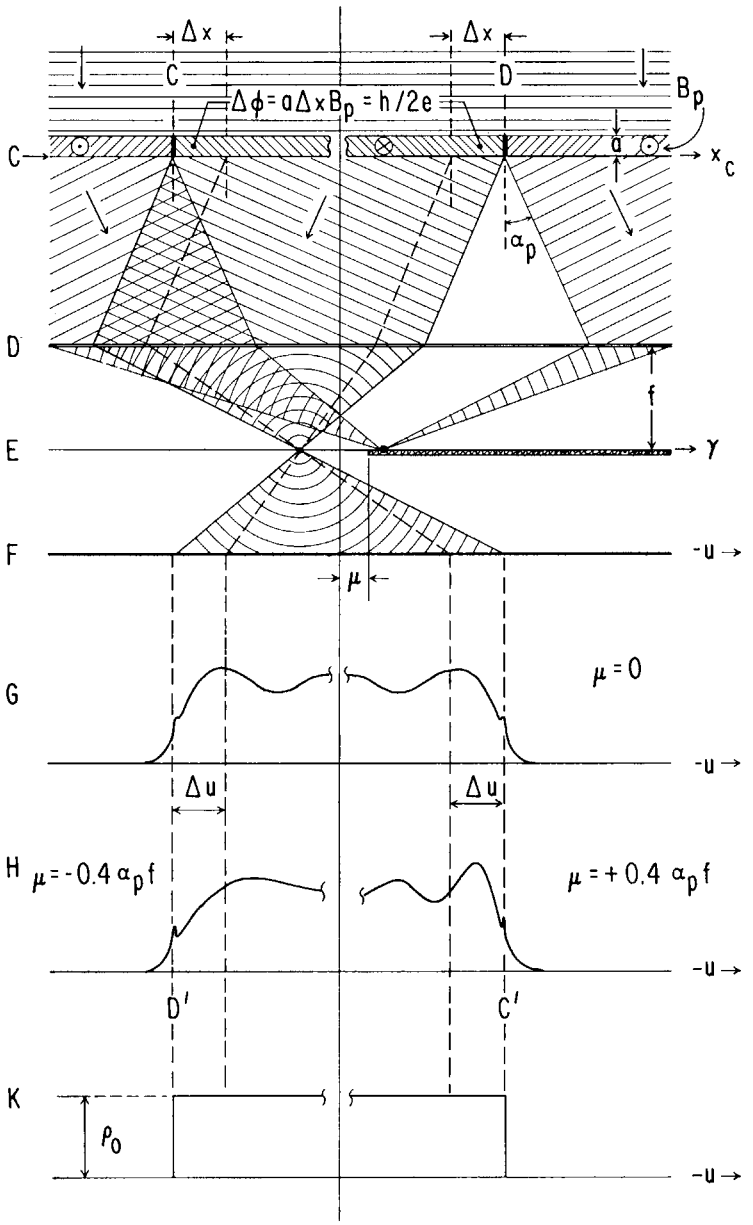


Fig. 8. – Diffraction on domain wall in the Foucault mode.

due to the object changes at least by π between x_1 and x_2. Otherwise the intensity difference will be smaller than ϱ_0. Again, with geometric concepts the wall cannot be located better than somewhere within a flux quantum $h/2e$.

Fraunhofer diffraction on periodic magnetic objects ([12,13]) is a third case where the quantum of flux manifests itself directly in the image. This is discussed by Wade in this volume.

5. Phase shift, scattering probability and signal to noise ratio.

It was shown in the last Section that when the magnetic flux inhomogeneity under study decreases to the order of a quantum of flux, the geometric approximation to Lorentz contrast (eq. (34)) breaks down in high contrast areas and the wave theory must be used. There is a one to one correspondence between the wave mechanical probability density (eq. (28)) and the magnetic flux distribution in the object plane. Therefore the information on the flux distribution can be in principle obtained with any desired accuracy by matching the calculated to the measured probability density, using the flux distribution in eq. (26) as a parameter. This calculation, although more complicated than an evaluation of the geometric formula, eq. (34), can of course be done, if necessary on a computer. However there is another more serious practical difficulty: the measured probability distribution $\varrho(\boldsymbol{r}_d)$ must be examined in ever decreasing steps $\Delta\varrho$, if the flux under study decreases below the quantum of flux. The probability density is plotted experimentally with accuracy $\Delta N = N^{\frac{1}{2}}$, where N is the number of recorded electrons. Since $\Delta\varrho/\varrho = \Delta N/N = N^{-\frac{1}{2}}$, high relative accuracy requires recording of a large number of particles, *i.e.* long illumination time.

For instance, if the magnetization step in the zero width wall is to be located in the Foucault mode within an interval of one tenth of the width of a flux quantum ($0.1\,\Delta x$ in Fig. 8), at least 100 electrons must be recorded in the image of that interval to distinguish it from the neighboring intervals where the probability (average probability density) is higher or lower by $\Delta\varrho \approx 0.1\varrho_0$. Classically, one or two electrons would have been sufficient in the interval adjoining the edge, since it is known that *no* electrons can arrive at the other side of the edge in the zero probability region. In other words, while classically it is only necessary to make a yes-no decision in an interval given by the desired measuring accuracy, in the real pattern a difference of probability from one interval to the next must be detected which is only a

fraction of the classical probability. This difficulty arises wherever probability differences must be measured *within a diffraction fringe.* It is therefore intimately connected with the transition from geometric to wave optical contrast, which in turn is tied to the strength of the object under study, measured in flux quantum units.

It will be shown now that the reason for this difficulty is independent of the mode of detection. It is simply a consequence of a cut-off of the electron scattering probability at the quantum of flux. The following is a derivation of the scattering probability from the phase shift which the incoming wave suffers in an elastically scattering object ([14]). It applies not only to magnetic phase shifts, but to electrostatic phase shifts and to light phase objects as well.

A complete measurement of the object properties in the sheet of Fig. 1 implies the measurement of the difference of the states in the presence and absence of the object, $\psi''(\boldsymbol{r})$ and $\psi_0(\boldsymbol{r})$, over all co-ordinates of the image space. A connection between these two states is given by eq. (25). Let the state without object have a general co-ordinate dependence of amplitude $\chi_0(\boldsymbol{r}_c)$ and phase $\varphi_0(\boldsymbol{r}_c)$ which is only restricted by the condition imposed in the beginning, namely that all particles travel nearly parallel to the z axis, *i.e.* $\partial\varphi_0/\partial x$, $\partial\varphi_0/\partial y \ll k$. The state with object can be split into two parts

$$\psi''(\boldsymbol{r}) = \beta\psi_0(\boldsymbol{r}) + \psi_s(\boldsymbol{r})\,. \tag{48}$$

β is a complex constant with modulus smaller or equal to one. $\beta\psi_0(\boldsymbol{r})$ is called the unscattered state. It contains no information about the object and is chosen as one of the components of $\psi''(\boldsymbol{r})$ for that reason only. The scattered state $\psi_s(r)$ is constructed from $\beta\psi_0(\boldsymbol{r})$ and $\psi''(\boldsymbol{r})$ by the condition that the probability P'' for the particles to be in the state $\psi''(\boldsymbol{r})$ is the sum of the probabilities P_0'' and P_s for them to be in the unscattered and the scattered state in the presence of the object. The probability for particles to be in a given state is the sum (integral) of the square of the amplitudes of that state at all values of the *independent* variables. The amplitude in the co-ordinate representation is $\chi(\boldsymbol{r})$, where $\boldsymbol{r}$ designates the set of three independent space co-ordinates. Thus

$$P'' = P_0'' + P_s\,, \qquad \int\limits_V \mathrm{d}\boldsymbol{r}\chi''^2(\boldsymbol{r}) = |\beta|^2 \int\limits_V \mathrm{d}\boldsymbol{r}\chi_0^2(\boldsymbol{r}) + \int\limits_V \mathrm{d}\boldsymbol{r}\chi_s^2(\boldsymbol{r})\,. \tag{49}$$

Equation (49) implies that ψ_s is orthogonal to $\beta\psi_0$, since inserting eqs (1)

and (48) into (49) forces

$$\int_V \mathrm{d}\boldsymbol{r}(\psi_s^* \beta \psi_0 + \psi_s \beta^* \psi_0^*) = 0 . \tag{50}$$

The constant β can be determined by inserting eq. (48) into (50)

$$2|\beta|^2 \int_V \mathrm{d}\boldsymbol{r}\chi_0^2 = \beta \int_V \mathrm{d}\boldsymbol{r}\psi_0 \psi''^* + \beta^* \int_V \mathrm{d}\boldsymbol{r}\psi_0^* \psi'' . \tag{51}$$

This equation is solved by (*)

$$\beta = \int_V \mathrm{d}\boldsymbol{r}\psi_0^* \psi'' \Big/ \int_V \mathrm{d}\boldsymbol{r}\chi_0^2 . \tag{52}$$

The scattering probability Σ is defined as the fraction of the particles in V which are in the scattered state in the presence of the object

$$\left.\begin{aligned} \Sigma &= N_s/N'' = P_s/P'' = 1 - |\beta|^2 \int_V \mathrm{d}\boldsymbol{r}\chi_0^2 \Big/ \int_V \mathrm{d}\boldsymbol{r}\chi''^2 , \\ \Sigma &= 1 - \Big| \int_V \mathrm{d}\boldsymbol{r}\psi_0^* \psi'' \Big|^2 \Big/ \int_V \mathrm{d}\boldsymbol{r}\chi_0^2 \cdot \int_V \mathrm{d}\boldsymbol{r}\chi''^2 . \end{aligned}\right\} \tag{53}$$

The volume in question is chosen to be the image space, *i.e.* a half space bordered by the plane C. Actually, in the present case the volume integrals, which are written with the assumption that the states depend on all three components of $\boldsymbol{r}$ independently everywhere in the image space, reduce to surface integrals ([14]), because all states satisfy the Helmholtz eq. (3) in the image space (in the Kirchhoff gauge). This equation fixes the magnitude of the momentum of the electrons. Therefore all states can only depend on *two independent momentum co-ordinates.* From the definition of the momentum operator $p_{\mathrm{op}} \equiv -i\hbar\nabla$ it follows immediately that the state in its co-ordinate representation $\psi(\boldsymbol{r})$ can also depend only on *two independent space co-ordinates.*

(*) Equation (51) produces also a set of spurious, unphysical solutions which arise because this equation is quadratic in β. These solutions do not occur when the projection operator technique is used.

This is reflected in the well known fact that if ψ satisfies eq. (3) and if the values of amplitude and phase of ψ are known as functions of the co-ordinates on a closed surface, they are known everywhere in the volume inside, in particular on another surface. Thus, in the beam of small divergence the Kirchhoff integral can be regarded as a mere unitary transformation of the state in its representation on one plane to its representation on another. Since the probabilities are independent of such unitary transformations, it is sufficient to evaluate the integrals of eq. (53) over any cross-section of the beam in V, in particular over the plane C. It is emphasized at this point that eq. (53) can handle both amplitude and phase objects or a mixture thereof. If one restricts oneself to a pure phase object one has in the plane C $\chi_0(\boldsymbol{r}_c) = \chi''(\boldsymbol{r}_c)$ so that the normalization integrals in eq. (53) are equal.

$$P_0 = \int_C \mathrm{d}\boldsymbol{r}_c \chi_0^2 = P'' = \int_C \mathrm{d}\boldsymbol{r}_c \chi''^2 = 1 . \tag{54}$$

Using eq. (1) the scattering probability becomes

$$\Sigma_c = 1 - \left| \int_C \mathrm{d}\boldsymbol{r}_c \chi_0^2(\boldsymbol{r}) \exp\left[i\varphi_s(\boldsymbol{r}_c)\right] \right|^2 \tag{55}$$

$$\left(\varphi_s(\boldsymbol{r}_c) \equiv \varphi''(\boldsymbol{r}_c) - \varphi_0(\boldsymbol{r}_c)\right).$$

Here φ_s is the scattering phase shift. This simple form of Σ_c is only valid if $|\nabla\varphi| \ll k$. The scattering probability depends on an average of the operator $\exp[i\varphi_s]$ over the plane C, weighed with the probability to find a particle at $\boldsymbol{r}_c$. In the plane C, the probabilities $\chi_0^2(\boldsymbol{r}_c)$ are independent of each other if $|\Delta\boldsymbol{r}_c| > \lambda$. It is therefore permissible to subdivide the full cross-section of the beam in C into a set of small areas A of any desired shape and apply eq. (55) to each of them separately, *i.e.* to calculate a number Σ_A for any subarea A of the object, as long as $A > \lambda^2$, and as long as χ_0^2 is renormalized for A instead of C (eq. (54)). In particular, an area with a radius equal to the resolution limit of the transmission microscope is large compared to λ^2, and a number Σ_A can be assigned to each such area of the in focus image of the object. The object area can of course (and often must in practice) be divided into much larger subareas. The number Σ_A^{-1} indicates the number of particles which must be passed through A before a difference between the wave function with and without object can be detected in A *in principle*. In other words, Σ_A *determines an absolute lower limit of the*

necessary illumination time. Equation (55), rewritten for the area A, is

$$\Sigma_A = 1 - \left| \int_A \mathrm{d}r_c \chi_0^2(r_c) \exp\left[i\varphi_s(r_c)\right] \right|^2 = 1 - \{\langle \cos\varphi_s \rangle_{0A}^2 + \langle \sin\varphi_s \rangle_{0A}^2\} - \\ = 1 - \nu \qquad \left(\int_A \mathrm{d}r_c \chi_0^2(r_c) = 1 \right). \tag{56}$$

Here $\langle O \rangle_{0A}$ designates an average of the operator O taken in the area A and weighed with the probability to find electrons there in the state without object.

For small phase shift, *i.e.* $\varphi_s \ll \pi$, ν can be expanded to yield $\nu \to 1 - -(\Delta\varphi_s)_{0A}^2$. For large phase shift, *i.e.* $(\Delta\varphi_s)_{0A} \approx 2\pi n$ (where $n \gg 1$ is an integer), the averages of the trigonometric functions are of order n^{-1} (disregarding certain singularities, where both averages can be zero simultaneously). Thus

$$\nu = \langle \cos\varphi_s \rangle_{0A}^2 + \langle \sin\varphi_s \rangle_{0A}^2 \begin{cases} \leqslant (\Delta\varphi_s)_{0A}^{-2} \ll 1 \\ = 1 - (\Delta\varphi_s)_{0A}^2 \approx 1 \end{cases} \quad \text{if } (\Delta\varphi_s)_{0A} \begin{cases} \gg \\ \ll \end{cases} \pi . \tag{57}$$

Therefore the qualitative behavior of the scattering probability as function of the phase shift is indicated by

$$\Sigma_A \begin{cases} \approx 1 \\ = (\Delta\varphi_s)_{0A}^2 \ll 1 \end{cases} \quad \text{if } (\Delta\varphi_s)_{0A} \begin{cases} \gg \\ \ll \end{cases} \pi . \tag{58}$$

For large phase shift, it suffices to send *one* particle through A to ascertain the presence of the object experimentally, *i.e.* the area scatters classically. However, for small phase shift *many* particles must be sent through A before the presence of the object can be ascertained. *For $(\Delta\varphi_s)_{0A} \approx \pi$ the scattering probability changes from classical to nonclassical,* from one to less than one.

The scattering probability is the ratio of the probability for electrons being in the scattered state to that for being in the state with object. Another useful number is the ratio of the probability to be in the scattered state to that for being in the unscattered state in the presence of the object:

$$\frac{P_s}{P_0''} \equiv \varkappa = \frac{\Sigma}{1-\Sigma} . \tag{59}$$

Since the particles in the state $\beta\psi_0$ do not contain any information about the object by definition, *the number $\varkappa$ has the meaning of a signal to noise ratio for the electrons in the beam.*

From the behaviour of Σ sketched above it follows that

$$\varkappa_A \begin{Bmatrix} \gg \\ \approx \\ \ll \end{Bmatrix} 1\,, \quad \text{if } (\Delta\varphi_s)_{0A} \begin{Bmatrix} \gg \\ \approx \\ \ll \end{Bmatrix} \pi\,. \tag{60}$$

Actually, quite generally $\varkappa \approx (\Delta\varphi_s)^2_{0A}$ for the whole range of $(\Delta\varphi_s)_{0A}$. Clearly, for $(\Delta\varphi_s)_{0A} < \pi$ some spatial analogon to phase sensitive detection in electronic engineering is needed for detection since the signal to noise ratio is smaller than one.

With the magnetic phase shift of eq. (25), the magnetic scattering probability is

$$\Sigma_A \begin{Bmatrix} = 1 \\ = \dfrac{e}{\hbar}(\Delta\Phi)^2_{0A} \ll 1 \end{Bmatrix}, \qquad \text{if } (\Delta\Phi)_{0A} \begin{Bmatrix} \gg \\ \ll \end{Bmatrix} h/2e\,. \tag{61}$$

The magnetic signal to noise ratio is

$$\varkappa_A \approx \left[\frac{e}{\hbar}(\Delta\Phi)_{0A}\right]^2 \begin{Bmatrix} \gg \\ \approx \\ \ll \end{Bmatrix} 1\,, \qquad \text{if } (\Delta\Phi)_{0A} \begin{Bmatrix} \gg \\ \approx \\ \ll \end{Bmatrix} h/2e\,. \tag{62}$$

It is a remarkable feature of magnetic objects that Σ and $\varkappa$ do not depend on the energy of the electrons.

In eqs (61) and (62) the quantity $(\Delta\Phi)_{0A}$ is the mean square fluctuation of the magnetic flux, weighed with the square of the amplitude of the wave function without magnetic object. Thus, if there is an elastically or inelastically scattering electrostatic object, its effect is written into $\varphi_0(\boldsymbol{r}_c)$ and $\chi_0(\boldsymbol{r}_c)$, *i.e.* eqs (61) and (62) are valid for magnetic scattering independent of other scattering. Now, usually the magnetic object in Lorentz microscopy is selected in those areas of the film where the electrostatic phase shift and the inelastic scattering do not depend much on $\boldsymbol{r}_c$. Moreover, the amplitude of the incoming wave does not change much over the areas of interest either. Then $\chi_0(\boldsymbol{r}_c) = \text{const}$ and $(\Delta\Phi)_{0A}$ reduces to

$$(\Delta\Phi)_A = \left[A^{-1}\int_A \mathrm{d}\boldsymbol{r}_c \Phi^2(\boldsymbol{r}_c) - \left(A^{-1}\int_A \mathrm{d}\boldsymbol{r}_c \Phi(\boldsymbol{r}_c)\right)^2\right]^{\frac{1}{2}}. \tag{63}$$

Equations (63) and (24) together give a rigorous and very useful definition of a magnetic flux inhomogeneity, which replaces the simple minded definition of $\Delta\Phi$ in eq. (40).

5·1. Example.

For an example of scattering probability and signal to noise ratio consider Fig. 9 and 10. A monochromatic but not necessarily parallel beam of fast electrons penetrates a slit of width q and length l and then a homogeneous magnetic field B which is parallel to the slit and is contained in the sheet of thickness a. The flux inhomogeneity is, from eq. (63)

$$(\Delta\Phi)_q = aqB(12)^{-\frac{1}{2}} = (12)^{-\frac{1}{2}}\,\Phi_I\,. \tag{64}$$

Φ_I is the illuminated flux. The scattering probability is

$$\left.\begin{aligned} \Sigma_q &= 1-\varepsilon^{-2}\sin^2\varepsilon\,, \\ \varepsilon &= Baqe/2\hbar = 3^{\frac{1}{2}}(\Delta\Phi)_q e/\hbar = \frac{1}{2}\frac{e}{\hbar}\Phi_I\,. \end{aligned}\right\} \tag{65}$$

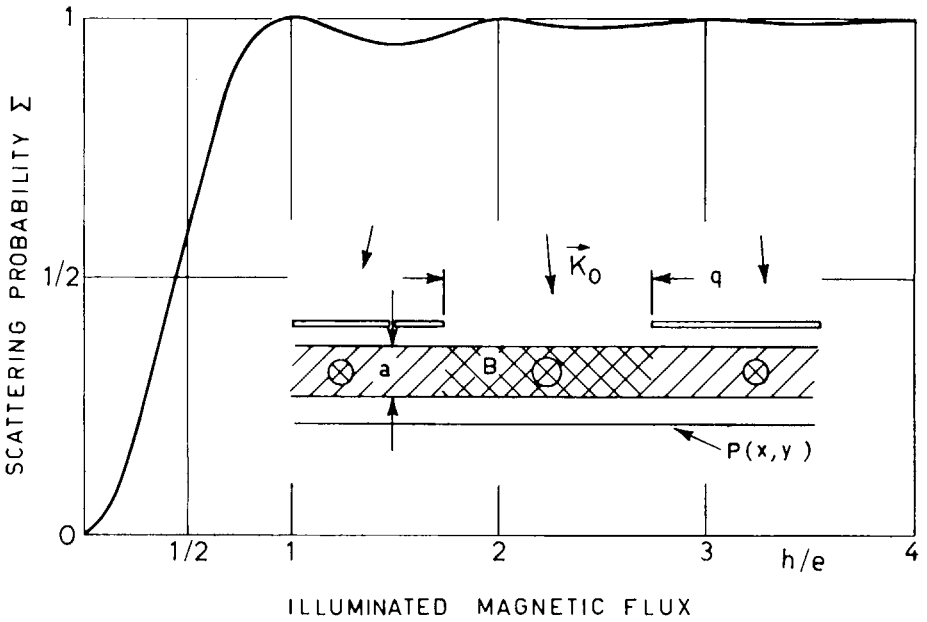


Fig. 9. – Magnetic scattering probability on a strip of homogeneous magnetic field. (Courtesy of *Journ. Appl. Phys.*)

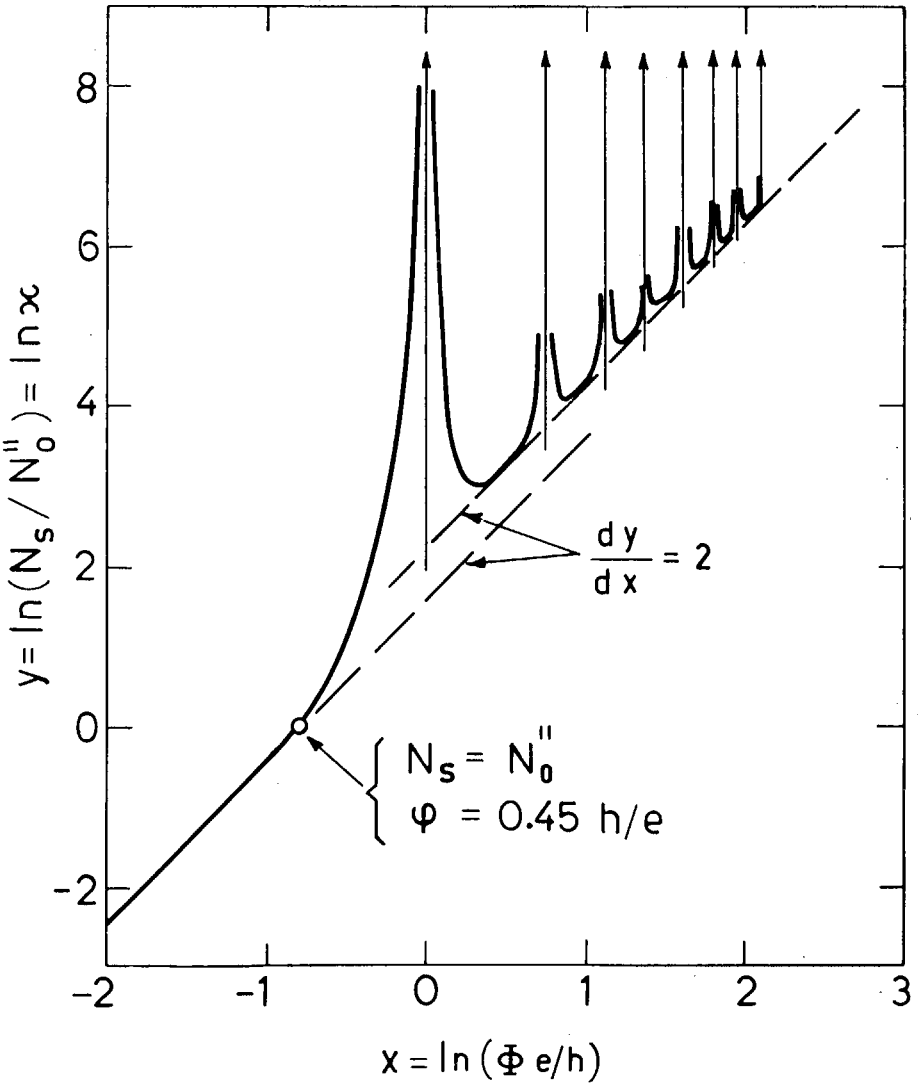


Fig. 10. – Magnetic signal to noise ratio behind a strip of homogeneous magnetific field. (Courtesy of *Journ. Appl. Phys.*)

In Fig. 9 the scattering probability is plotted as a function of the illuminated flux Φ_I. Note the cut-off near the flux quantum. Nearly every electron is scattered, if $\Phi_I \gg h/2e$, as expected classically. However, *below* $\Phi_I \sim h/2e$,

the majority of the electrons is not noticeably affected by the field which they have penetrated.

The signal to noise ratio (Fig. 10) is roughly equal to the square of the flux measured in units of the flux quantum throughout the flux range, but becomes infinite (classical) at certain points. It is equal to one for $\Phi_I \approx h/2e$.

6. Signal to noise ratio and maximum contrast.

A thorough discussion of experimental arrangements which approach optimal resolution is impossible without a lengthy treatment of partial coherence, which is outside of the scope of these lectures. However, within the present treatment a connection of the maximum possible contrast and the signal to noise ratio $\varkappa$ is pointed out which illustrates the central practical importance of the latter for both strong and weak objects.

6·1. Strong inhomogeneity.

Consider the following simple magnetic-field distribution in a domain wall:

$$B(x) = \begin{cases} B_p \operatorname{sign} x\,, & |x| > w/2\,, \\ B_p 2x/w\,, & |x| < w/2\,. \end{cases} \tag{66}$$

The caustic tip for this distribution is at $x = 0$ and $z = d_0 = p_0 w/2eaB_p$ (from eq. (44)). For these co-ordinates, *all* derivatives of φ_c higher than the second vanish if $|x| < w/2$. Therefore the caustic mantle degenerates into one point. This magnetic field distribution can be regarded as an ideal cylindrical lens for electrons with focal length d_0. Obviously the probability density in the focal point is the highest anywhere in the image space.

Consider the ratio of the probability densities in the focal point $(0, 0, d_0)$ with finite wall width w and with $w = 0$ (*). Assume that the wall is a strong inhomogeneity, $(\Delta\Phi)_w \gg h/2e$. With or without object, the amplitude in the focal point is proportional to the limit of integration, of eq. (26) if $\chi_0(x_c) = \text{const}$. Without object ($w = 0$), the amplitude is built up from two

(*) The latter describes the « state without object », although these is a field distribution $B(x) = B_p \operatorname{sign} x$ for $-\infty < x < \infty$. This point is clarified in Sect. 7.

stationary points at $x_c^0 = \pm \alpha_p d_0$, with a limit of integration $\Delta x = (\lambda d_0/2)^{\frac{1}{2}}$. With object, the integration goes clearly over the whole wall width, plus the outer limit of integration of the two stationary points of the case without object. We define the contrast by $K = (\varrho(w) - \varrho(0))/\varrho(0)$. It can be estimated by the ratio of the squares of the amplitudes with and without object.

$$K \approx \varrho(w)/\varrho(0) = w^2/2\lambda d_0 = \pi a w B_p(e/h) = (\Delta\Phi_s)_w \cdot 8e/h \qquad ((\Delta\Phi_s)_w \gg \pi) . \tag{67}$$

Here $(\Delta\Phi_s)_w$ was calculated from eqs (79), (63) and (24). From eq. (62) it is clear that in this point of maximum contrast

$$[\varrho(w)/\varrho(0)]_{\max} \approx \varkappa^{\frac{1}{2}} . \tag{68}$$

The maximum possible contrast in the defocused mode is within a factor of order one equal to the square root of the signal to noise ratio, or equal to the flux inhomogeneity measured in units of the flux quantum.

With the image plane D at the defocusing distance d, there will be a fringe with the peak intensity given by eq. (67) in the focal point. This fringe will have a width of order $\Delta\xi \approx h/2eaB_p$. It is clearly sufficient to send just a few electrons through the wall in order to *detect* the wall, since these electrons will be collected with very high probability within the fringe width around the focal point, if $(\Delta\Phi_s)_w \gg h/2e$. However, if one wants to *measure* $(\Delta\Phi_s)_w$ with a signal to noise ratio of one, one must know both $\varrho(w)$ and $\varrho(0)$ in the fringe width, *i.e.* one must count a sufficient number of particles in the fringe centred at the focal point. The requirement is

$$\Delta\varrho(w) \approx \varrho(0) . \tag{69}$$

Since $\Delta\varrho(w)/\varrho(w) = N_{wt}^{-\frac{1}{2}}$, in order to satisfy eq. (69), one has to record at least N_{wt} particles in the focal fringe, given by

$$N_{wt} \geq (\varrho(w)/\varrho(0))^2 \approx \varkappa . \tag{70}$$

The signal-to-noise ratio gives the minimum number of particles necessary to measure the magnitude of the flux inhomogeneity to the accuracy of one flux quantum in the maximum contrast point of the image space.

The above wall distribution is somewhat unrealistic. Several other distributions have been discussed, most frequent the hyperbolic tangent distribution (Fig. 11). Any distribution other than that of eq. (66) has a full caustic mantle instead of a single focal point. The contrast in the tip of the caustic

mantle is then still the maximum possible in the image space, for reasons discussed in Sect. **4**, but for the same wall width it will be lower than that of the ideal cylindrical lens. The situation is analogous to that of a lens with strong aberrations.

6·2. Weak inhomogeneity.

The domain wall does not lend itself to an easy discussion of the maximum contrast in the image space if it represents a weak inhomogeneity. For that case a periodic structure is chosen instead, which approximates the ripple or the Abrikosov flux line structure

$$B(x) = B_0 \sin(2\pi x/\tau) . \tag{71}$$

Here τ is the wavelength of the structure. The flux inhomogeneity is, from eqs (63) and (64)

$$(\Delta\Phi)_\tau = B_0 a\tau/2^{\frac{3}{2}}\pi . \tag{72}$$

With a point source at distance g above the film and $\chi_0(\boldsymbol{r}_c) = \text{const}$, if $(\Delta\Phi)_\tau \ll h/e$ one obtains from eqs (26) and (28):

$$\varrho(\Phi)/\varrho(0) = 1 + 2\tau a B_0(e/h) \sin(2\pi^2 gd/(g+d)\,\tau^2 k) \sin(2\pi x_d g/(g+d)\,\tau) . \tag{73}$$

Thus the maximum contrast in the image space is

$$K_{\max} = ((\varrho(\Phi) - \varrho(0))/\varrho(0))_{\max} = a\tau B_0(2e/h) = (\Delta\Phi)_\tau 2^{\frac{3}{2}}(e/\hbar) \approx 2^{\frac{3}{2}}\varkappa^{\frac{1}{2}} \ll 1 . \tag{74}$$

Again, the maximum contrast is nearly equal to the square root of the magnetic signal to noise ratio. Since it is also much smaller than one, $\varrho(\Phi) \approx \varrho(0)$ and $\Delta\varrho(\Phi) \approx \Delta\varrho(0)$. For a measurement of this maximum contrast, *i.e.* for $\Delta\varrho(\Phi) \leqslant 2\,(\varrho(\Phi) - \varrho(0))$, one must have

$$N_{t\tau}^{-\frac{1}{2}} = \frac{\Delta\varrho(\Phi)}{\varrho(\Phi)} \leqslant \frac{(\varrho(\Phi) - \varrho(0))_{\max}}{\varrho(0)} = 2^{\frac{3}{2}}\varkappa^{\frac{1}{2}} , \quad \text{or} \quad N_{t\tau} \geqslant (8\varkappa)^{-1} . \tag{75}$$

For weak inhomogeneities the *inverse* magnetic signal to noise ratio gives the minimum number of particles N_{t_τ} necessary to *measure* the maximum contrast in the same sense as in the case of the strong object. But in the case of

the weak object $N_{t\tau}$ is also the minimum number of particles necessary to *detect* the inhomogeneity anywhere in the image space.

The contrast of any inhomogeneity is nearly classical if the defocusing distance is small compared to d_m, where d_m is the distance at which the maximum contrast appears for the first time ([10-12]). For this reason small defocusing distances are often preferred experimentally ([11]). However, such arrangements are very unfavorable from the point of view of the illumination time, because their uncertainty product is much larger than $\frac{1}{4}$. For the present example of a weak periodic inhomogeneity, if $d_s = 0.1\,d_m$, the contrast is about 0.2 times the maximum possible. Following the same reasoning which led to eq. (75) one finds that $(N_{t\tau})_s \approx 25(N_{t\tau})_m$ where s and m designate the minimum number of particles which must be counted with defocusing distance d_s and d_m respectively in order to *detect* the inhomogeneity. Clearly, in view of the necessarily finite illumination time t, it is not permissible to work at small defocusing distance if one wants to attain the maximum possible information from the available number of particles. Similarly, from the point of view of illumination time the measurement of wall parameters from divergent wall images ([11,17]) is much inferior to that from convergent images.

7. Number phase uncertainty relation in phase-contrast microscopy

A magnetic field distribution has been measured with a certain accuracy if the correspondence between the measured probability density and the one calculated from a model of the magnetic field distribution is unambiguous within a difference equal to this accuracy. The problem is to determine the difference. Assume that the computer calculation can be done with any desired accuracy. Then the limit of obtainable information will be on the experimental side. Clearly, under otherwise ideal experimental conditions the limit of accuracy is given by the illumination time, since the probability can only be measured with finite accuracy. It will now be shown that with a slight redefinition of the scattering phase shift φ_s, the formalism developed in Sect. **5** can be used to determine the best possible accuracy of the measurement from the illumination time.

In deriving the scattering probability it was stated that the object manifests itself through a difference of two wave functions. One of the two, $\psi_0(r)$, was taken to be the wave function without field in the sheet. This wave function contains *a priori* information about the geometry of the beam, the energy,

the initial momentum distribution, etc., in other words, it is a *model* wave function which is in practice simulated by the computer and is rarely measured directly. It was later pointed out that ψ_0 can also contain all the effects of elastic and inelastic electrostatic scattering. Going still further, one can convince oneself easily that nothing prohibits $\psi_0(\boldsymbol{r})$ to be regarded as the wave function with *one* magnetic field distribution (containing also all effects of the source and electrostatic scattering) and $\psi''(\boldsymbol{r})$ to be a wave function with *another* magnetic field distribution (containing the *same* effects of source and electrostatic scattering as before). *Then φ_s is the difference between the magnetic phase shifts in the two models* and Σ and $\varkappa$ can be used to determine the distinguishability of the two models as a function of the accuracy with which the probability is plotted out experimentally.

It is intuitively plausible that the two wave functions are experimentally distinguishable if the number of « scattered » particles is larger than one. Now, $\chi^2 = \tilde{n}$ can be regarded as the number operator in second quantization. The integrals in eqs (49) and (54) are then expectation values of $\tilde{n}$, *i.e.* average numbers in the respective states. Therefore to have at least one scattered particle in the record one must have

$$\langle \tilde{n}_s \rangle_A = \int_A \mathrm{d}\boldsymbol{r}_c \chi_s^2 > 1 \,. \tag{76}$$

Equation (76) forces a renormalization of the integrals in eq. (54), since they are all connected through eq. (49). Σ and $\varkappa$ are independent of renormalization. From eqs (53) and (59) one has therefore simply

$$\Sigma \int_A \mathrm{d}\boldsymbol{r}_c \chi''^2 = \Sigma \langle \tilde{n}'' \rangle_A > 1 \,, \tag{77}$$

$$\varkappa \int_A \mathrm{d}\boldsymbol{r}_c \chi_0''^2 = \varkappa \langle \tilde{n}_0'' \rangle_A > 1 \,. \tag{78}$$

Inserting eq. (56) in (59) and using the result in eq. (78) gives

$$\langle \tilde{n}_0'' \rangle_A \frac{(\Delta \cos \varphi_s)_{0A}^2 + (\Delta \sin \varphi_s)_{0A}^2}{\langle \cos \varphi_s \rangle_{0A}^2 + \langle \sin \varphi_s \rangle_{0A}^2} \geqslant 1 \,. \tag{79}$$

This equation shows a remarkable similarity with the general uncertainty

relation between amplitude and phase of a particle wave function ([15])

$$(\Delta\chi)^2 \frac{(\Delta\cos\varphi)^2 + (\Delta\sin\varphi)^2}{\langle\cos\varphi\rangle^2 + \langle\sin\varphi\rangle^2} \geqslant \frac{1}{4}. \tag{80}$$

The averages in both eqs (79) and (80) are over the full set of independent variables, including the time. In time the amplitude in the plane C oscillates rapidly with frequency $\omega = E_0/\hbar = \hbar k^2/2m$. Thus *averaged over the illumination time* t

$$(\Delta\chi)^2_{At} = \langle\chi^2\rangle_{At} - \langle\chi\rangle^2_{At} = \langle\chi^2\rangle_{At} = \langle\tilde{n}\rangle_{At} = N_{At}. \tag{81}$$

The other averages in eq. (79) are time independent. If one accepts for the minimum uncertainty product $\frac{1}{4}$, which is the result of Schwartz' inequality, rather than one, which came in through the intuitive choice in eq. (76), eq. (79) becomes *identical* with the amplitude-phase uncertainty relation, and also becomes a number phase uncertainty relation.

Equation (79) then says that one must count a sufficient number of unscattered particles, *i.e.* plot experimentally the «unscattered» assumed state sufficiently accurately, before it is in principle possible to notice a change of the real state with respect to the assumed state.

In high-resolution phase-contrast microscopy one is mostly interested in *small* phase shifts, then

$$P_0'' \approx P'' = P_0 = \int_t \int_A \mathrm{d}t\, \mathrm{d}\boldsymbol{r}_c \chi_0^2(\boldsymbol{r}_c, t) = N_{0At} \quad (\varphi_s \ll \pi). \tag{82}$$

This in eq. (77) gives

$$N_{0At}(\Delta\varphi_s)^2_{0A} \geqslant \tfrac{1}{4}, \quad \text{or} \quad N_{0At} \geqslant (4\varkappa)^{-1}. \tag{83}$$

This general result should be compared with the special case of a weak ripple, as discussed in Sect. **6** (eq. (75)).

Inserting the magnetic phase shift for Lorentz microscopy and taking the square root of eq. (83)

$$2e(N_{0At})^{\frac{1}{2}}(\Delta\Phi)_{0A} \geqslant \hbar. \tag{84}$$

When a number N_{0At} of particles has passed through the area A during the illumination time t, $(\Delta\varphi_s)_{0A}$ in eq. (83) gives a lower limit to the set of all possible functional deviations $\delta\varphi_s(\boldsymbol{r}_c)$ of the mathematically assumed

phase shift distribution $\varphi_s(r_c)$ from the real one which can be detected (and corrected for) *in principle*. Actually, the mean square fluctuations in eqs (80) and (83) depend on the representation $\psi''(r_d)$, *i.e.* on the defocusing distance, or, more generally, on the surface chosen in the image space. Therefore the detection of the phase inhomogeneity $(\Delta\varphi_s)_{0.1}$ with the minimum number given by eq. (79) is only possible on a very special surface. This point was discussed briefly in Sect. **6**.

7·1. **Optimal resolution of a domain wall.**

We now calculate the optimal resolution of the magnetization distribution in a domain wall (Fig. 6) from the illumination time. It is assumed that an experimental setup can be found where the uncertainty product in eq. (83) is at its minimum value $\frac{1}{4}$. The magnetic field $B(x)$ varies slowly over the resolution limit of the instrument. The wall of width w and length l covers an area $A_w = lw$ in the plane C. This area is now divided into $n \gg 1$ strips of varying width $\Delta\xi_n$ parallel to the wall axis. The width of each strip is so small that the flux within a strip is smaller than $h/2e$ and that its variation is within a strip sufficiently accurately described by

$$\Phi(x) - \Phi(x_n) = (x - x_n)\frac{\partial \Phi}{\partial x}(x_n) + \frac{1}{2}(x - x_n)^2 \frac{\partial^2 \Phi}{\partial x^2}(x_n) =$$

$$= (x - x_n) aB(x_n) + \frac{a}{2}(x - x_n)^2 \frac{\partial B}{\partial x}(x_n) . \qquad (85)$$

Here x_n is a reference co-ordinate in the n-th strip. Assume $\chi_0(x_c) = \text{const}$ across the wall. During the illumination time t each strip receives $N_{0A_n} t$ electrons:

$$N_{0.1_n t} = l\Delta x_n (\alpha_B)^2 Rt . \qquad (86)$$

Here α_B is the illumination aperture and R the brightness of the source. The interest is in a *change* of $B(x)$ from one strip to the next. Therefore the width Δx_n is chosen such that a change ΔB can just be detected, with a signal to noise ratio of one. In other words, the problem is to detect the difference between a model where the second term on the right hand side of eq. (85) exists and one where it is zero. The second term represents a flux inhomo-

geneity $\Delta\Phi_s$ which can be calculated from eq. (63):

$$(\Delta\Phi_s)^2 = (45)^{-1} a^2 \Delta x_n^4 \left[\frac{\partial B}{\partial x}(x_n)\right]^2. \tag{87}$$

Equation (84) determines the width Δx_n for which one can get one bit of information on the change of the field in the n-th interval, if the equality sign is chosen. Then

$$(\Delta x_n)^5 = 45(\hbar/2e)^2 \left(l(\alpha_B)^2 R t a^2 \left[\frac{\partial B}{\partial x}(x_n)\right]^2\right)^{-1}. \tag{88}$$

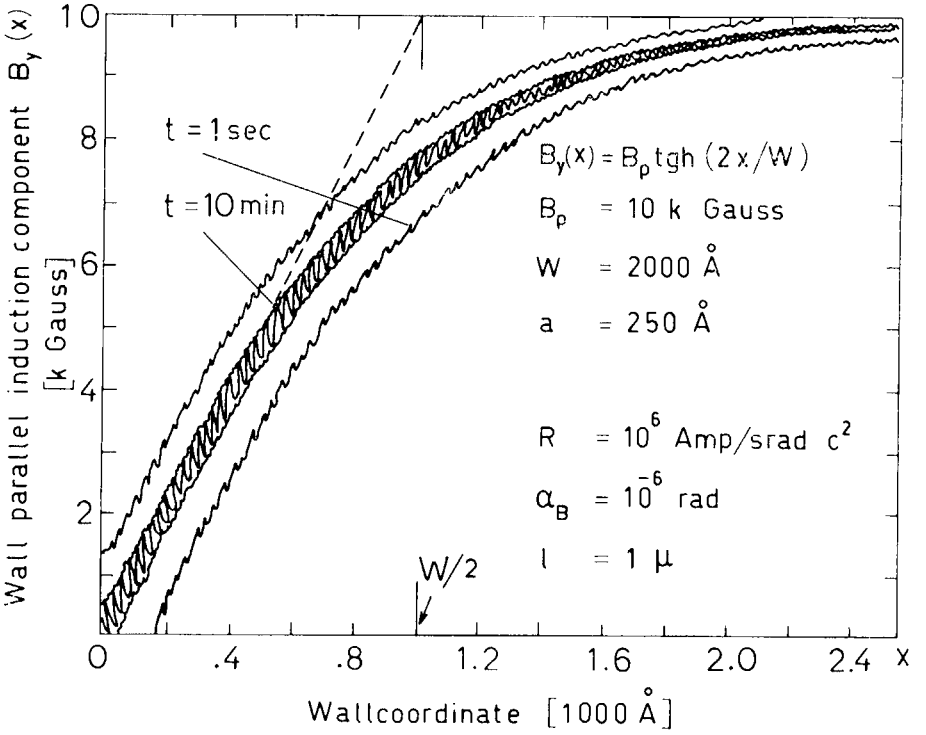


Fig. 11. Optimal resolution of a domain wall for two illumination times.

Clearly Δx_n is an extremely slow function of the illumination time, or of the brightness of the source. Figure 11 shows the optimal resolution of a domain

wall with a hyperbolic tangent distribution. The values of the wall parameters and the source are those commonly encountered in practical high-resolution measurements.

8. Separation of magnetic and electric contrast.

8·1. Separation at high energy.

In the previous Section it was shown that the maximum contrast which is possible in the image space is about equal to the flux inhomogeneity measured in units of the flux quantum. Generally speaking, the maximum contrast, in the defocused mode, is about equal to the mean square fluctuation of the phase shift.

This is of course also true for electric phase shifts. However, whereas the magnetic phase shift does not depend on the energy of the incoming beam (not even in the relativistic regime) it *does* in the electric case: the electric phase shift decreases with increasing energy, as will be shown next. Therefore it is possible to suppress appreciably the electric with respect to the magnetic contrast by working with high energy microscopes.

The calculation of the wave function in the plane C (Fig. 1) in the presence of an electrostatic object in the sheet between the planes B and C can be done along the lines of Sect. **2**. Since the energy of the incoming beam is assumed to be large compared to the binding energy of an electron in the object, the deflection is small, so that k_x, $k_y \ll k_z \approx k$, just as in the magnetic case. Then again within the sheet (the sheet thickness a is restricted by eq. (11)), the wave function in $\boldsymbol{r}_c$ is only dependent on a small area of the plane B around the stationary point $\boldsymbol{r}_b^0$. Therefore the electric potential can influence the wave function in $\boldsymbol{r}_c$ only on the line $(\boldsymbol{r}_b^0 - \boldsymbol{r}_c)$. Outside of the sheet the electric potential is constant, and one can set $V = 0$ on both sides of the sheet, if the object is grounded. The object is assumed to scatter elastically only. The Schrödinger equation inside the sheet is

$$H\psi = \left(-\frac{\hbar^2 \nabla^2}{2m} + eV\right)\psi = E\psi\,. \tag{89}$$

One has also

$$\hbar^2 k_0^2/2m = E_0 \gg eV\,. \tag{90}$$

Therefore one may write the wave vector in the presence of the field (observing eq. (90)) as

$$\left.\begin{aligned} &\hbar^2 k'^2/2m = E_0 + eV = (\hbar^2 k_0^2/2m)\cdot(1 + (2mVe/\hbar^2 k_0^2)), \\ &k' \approx k_0 + mVe/\hbar^2 k_0 = k_0 + Ve/\hbar v_0 . \end{aligned}\right\} \tag{91}$$

Here v_0 is the initial velocity of the electrons. The phase shift can be calculated in a similar fashion as in the magnetic case by subdividing the sheet by many planes and calculating the wave function from one plane to the next. The final result is

$$\psi_e(\boldsymbol{r}_c) = \psi_0(\boldsymbol{r}_c)\cdot\exp\left[i(ea\overline{V}(\boldsymbol{r}_c)/\hbar^2 v_0)\right] . \tag{92}$$

Here $\psi_e(\boldsymbol{r}_c)$ is the wave function in the presence of the electric field and $\overline{V}(\boldsymbol{r}_c) = \int_a^0 dz V(x, y, z)$. Therefore, the nonrelativistic electric phase shift decreases with $E_0^{-\frac{1}{2}}$, which means that *the electric signal to noise ratio decreases inversely proportional to the kinetic energy* (or inversely proportional to the square of the velocity). In the case of relativistic energies, eq. (92) remains valid. At very high energies $v_0 \to c$. Then the electric phase shift depends no longer on the accelerating voltage either.

At the same time the inelastic scattering, which was found to be a great obstacle to quantitative evaluation of high resolution Lorentz images ([4 11.16.17]) decreases with at least the same power of E_0. Therefore in high voltage electron microscopes electric and magnetic contrast should be much better separable than in conventional electron microscopes.

8·2. Separation by the parity operation.

The parity of the electric field is even, that of the magnetic field odd. Therefore, upon reversal of the direction of the initial momentum with respect to the field distribution, the magnetic interaction changes sign, whereas the electric does not. In other words, the magnetic phase shift changes sign, whereas the electric phase shift does not, if one turns the film in the sheet by 180° around an axis perpendicular to the optical axis. This procedure was once used ([2]) to prove the magnetic nature of the ripple contrast. The

difference of the probability densities of two records taken from reserved films under otherwise identical conditions gives directly twice the magnetic phase shift, if the flux inhomogeneity is weak. This method should therefore be very useful for quantitative study of ripple and the Abrikosov structure in the presence of strong elastic and inelastic electric scattering.

REFERENCES

Review Articles

P. B. HIRSCH, A. HOWIE, R. B. NICHOLSON, D. W. PASHLEY and M. J. WHELAN: *Electron Microscopy of Thin Crystals*, Butterworths, (1965), p. 388.
D. WOHLLEBEN: *Journ. Appl. Phys.*, **38**, 3341 (1967).
M. S. COHEN: *Journ. Appl. Phys.*, **38**, 4966 (1967).
R. H. WADE: *Journ. de Phys.*, Colloque C2, Suppl ment 2-3, **29**, 95 (1968).
P. J. GRUNDY and R. S. TEBBLE: *Adv. in Phys.*, **17**, 153 (1968).

Text References

1) H. W. FULLER, M. E. HALE and H. RUBINSTEIN: *Journ. Appl. Phys.*, **30**, 789 (1959); H. W. FULLER and M. E. HALE: *Journ. Appl. Phys.*, **31**, 238 (1960); H. W. FULLER and M. E. HALE: *Journ. Appl. Phys.*, **31**, 1699 (1960).
2) H. BOERSCH and H. RAITH: *Naturwiss.*, **20**, 574 (1959); H. BOERSCH, H. RAITH and D. WOHLLEBEN: *Zeits. f. Phys.*, **159**, 388 (1960).
3) W. FRANZ: *Physik. Berichte*, **21**, 686 (1940); W. EHRENBERG and R. E. SIDAY: *Proc. Phys. Soc.*, **62**, 8 (1949); Y. AHARONOV and D. BOHM: *Phys. Rev.*, **115**, 485 (1959).
4) H. BOERSCH, H. HAMISCH, D. WOHLLEBEN and K. GROHMANN: *Zeits. f. Phys.*, **164**, 55 (1961); H. BOERSCH, H. HAMISCH, K. GROHMANN and D. WOHLLEBEN: *Zeits. f. Phys.*, **167**, 72 (1962).
5) H. BOERSCH: *Physikal. Zeitsch.*, **44**, 202 (1943).
6) W. GLASER: *Handb. d. Phys.*, **33**, 332 (1952).
7) F. THON: *Zeits. f. Naturfor.*, **21***a*, 467 (1966).
8) H. WOLTER: *Handb. d. Phys.*, **24**, 555 (1956).
9) D. WOHLLEBEN: *Phys. Lett.*, **22**, 564 (1966).
10) J. P. GUIGAY and R. H. WADE: *Phys. Stat. Sol.*, **29**, 799 (1968).
11) M. S. COHEN and K. HARTE: *Journ. Appl. Phys.*, **40**, 3597 (1969).
12) O. BOSTANJOGLO and W. VIEWEGER: *Phys. Stat. Sol.*, **32**, 311 (1969).
13) R. H. WADE: *Phys. Stat. Sol.*, **19**, 847 (1967); M. J. GORINGE and J. P. JAKUBOVICS: *Phil. Mag.*, **15**, 393 (1967).
14) D. WOHLLEBEN: *Journ. Appl. Phys.*, **41**, 2551 (1970).
15) P. CARRUTHERS and M. M. NIETO: *Phys. Rev. Lett.*, **41**, 387 (1965).
16) D. HOTHERSALL: *Phil. Mag.*, **20**, 89 (1969).
17) L. REIMER and H. KAPPERT: *Zeits. angew. Phys.*, **27**, 165 (1969).

"*Ettore Majorana*" International Centre for Scientific Culture
A Series of Selected Publications directed by A. Zichichi

- *Symmetries in Elementary Particle Physics*,
  A. Zichichi ed. (1965).
- *Recent Developments in Particle Symmetries*,
  A. Zichichi ed. (1966).
- *Strong and Weak Interactions*,
  A. Zichichi ed. (1967).
- *Hadrons and their Interactions*,
  A. Zichichi ed. (1968).
- *Theory and Phenomenology in Particle Physics*,
  A. Zichichi ed., Part A and B (1969).
- *Subnuclear Phenomena*,
  A. Zichichi ed., Part A and B (1970).
- *Electron Microscopy in Material Science*,
  U. Valdrè ed. (1971).
- *Elementary Processes at High Energy*,
  A. Zichichi ed., Part A and B (1971).