Aberration correction past and present

BY P. W. HAWKES*

CEMES-CNRS, B.P. 94347, 31055 Toulouse cedex, France

Electron lenses are extremely poor: if glass lenses were as bad, we should see as well with the naked eye as with a microscope! The demonstration by Otto Scherzer in 1936 that skilful lens design could never eliminate the spherical and chromatic aberrations of rotationally symmetric electron lenses was therefore most unwelcome and the other great electron optician of those years, Walter Glaser, never ceased striving to find a loophole in Scherzer’s proof. In the wartime and early post-war years, the first proposals for correcting \( C_s \) were made and in 1947, in a second milestone paper, Scherzer listed these and other ways of correcting lenses; soon after, Dennis Gabor invented holography for the same purpose. These approaches will be briefly summarized and the work that led to the successful implementation of quadupole–octopole and sextupole correctors in the 1990s will be analysed. In conclusion, the elegant role of image algebra in describing image formation and processing and, above all, in developing new methods will be mentioned.

Keywords: aberrations; electron lenses; aberration correction; history

1. The early years

The atmosphere immediately after the first electron microscopes were constructed was heady. Ernst Ruska, at first unaware of Louis de Broglie’s work, was upset to learn that electrons too have a wavelength but soon established that the electron wavelength in his microscope would be very much shorter than that of visible light and hence formed no obstacle to subatomic resolution. It was tacitly assumed that the necessary lenses could be designed; the paper by Otto Scherzer (1936) in which a positive-definite form of the spherical aberration integrand was published was therefore extremely unwelcome. Reactions were numerous and varied: optimum lens designs were sought but the boldest—and to our eyes, the most outrageous—attempt to circumvent Scherzer’s finding was made by Walter Glaser, the other great electron optician of the first decades of electron optics. He observed that among the many ways of writing the spherical aberration integral is the expression

\[
C_s = \int_{z_o}^{z_i} A h^4 \, dz, \tag{1.1}
\]

*hawkes@cemes.fr

One contribution of 14 to a Discussion Meeting Issue ‘New possibilities with aberration-corrected electron microscopy’.
in which

\[ A = \frac{1}{48} \left( 5\eta^2 B'^2 \frac{\hat{\varphi}}{\phi} - \eta^2 B'^2 \frac{\hat{\varphi}}{\phi} + 4\eta^4 B^4 \right) \]

and \( h(z) \) is a particular solution of the paraxial equation of motion of electrons in a rotationally symmetric magnetic field \( B(z); \eta \) is a constant and \( \hat{\varphi} \) is the relativistically correct accelerating voltage. He argued that by setting \( A = 0 \) and solving the resulting differential equation for \( B(z) \), he would obtain a field distribution for which \( C_s = 0 \). That this was a forlorn hope can be seen immediately from the positive definite form

\[ C_s = \frac{1}{16} \int_{z_0}^{z_1} \left( \frac{\eta^4 B^4}{\phi^2} h^4 + 2(hB' + h'B)^2 \frac{\eta^2 h^2}{\phi} + 2 \frac{\eta^2 B^2}{\phi} h^2 h' \right) dz, \quad (1.2) \]

which can vanish only if all the individual terms in the integrand vanish and hence if \( B(z) = 0 \) everywhere. Glaser (1940a, b) did indeed find a solution but it was soon shown that this field was incapable of producing a real image of a real object, one of the necessary conditions in the derivation of equation (1.2). A similar result was shown to be true of electrostatic lenses. Glaser continued to fulminate against Scherzer’s ‘theorem’ for the rest of his life: we find doubts expressed in the Grundlagen der Elektronenoptik (Glaser 1952) and in his long chapter in the Handbuch der Physik (Glaser 1956a), where a footnote promises new evidence of a loophole. This was presented the same year (Glaser 1956b). Glaser’s earlier solution to the equation \( A = 0 \) had yielded a U-shaped field distribution. In this last attempt to overthrow Scherzer’s result, Glaser adopted a model U-shaped field, obtained from his own bell-shaped model

\[ B(z) = \frac{B_0}{1 + (z/a)^2} \quad (1.3) \]

by setting \( a = i\alpha \) to give

\[ B(z) = \frac{B_0}{1 - (z/\alpha)^2}. \quad (1.4) \]

This model had already been studied by Sturrock (1951b) and, with it, Glaser found values of \( C_s \) smaller than the theoretical minimum value. But in a corrigendum, Glaser (1956c) acknowledges that he has made a mistake in the calculation and, after correcting this, \( C_s \) falls among the permitted values. At least one person remains incredulous today (Nomura 2008; Hawkes 2009).

It is usual to regard a paper by Scherzer that appeared in 1947 as the next major event in the aberration correction saga. In deriving his positive-definite \( C_s \), Scherzer had made a number of (entirely reasonable) assumptions; namely, that the lens is static, rotationally symmetric, forms a real image of a real object, has no conductors on the axis, no discontinuities in the electrostatic potential distribution and no space charge and acts as a lens, not a mirror. In the 1947 article, each of these requirements is relaxed in turn and a possible correction system is proposed. Important though this paper is, for almost all the correctors devised in the following decades were inspired by it, some major proposals were made in the wartime years, two of which are even mentioned in an early textbook (Zworykin et al. 1945). These involve the use of an electron mirror and of a high-frequency lens but no references are given. Less well known is the section on ‘Possibilities of future
development’ in a small book published by Dennis Gabor (1945). (This is an ‘amplified version of a lecture, delivered on the 4 March 1943, before the Cambridge University Physics Society’; the preface is dated 22 July 1944 but an endnote tells us that publication has been delayed to 1945.) Here, Gabor first analyses in detail the proposal by Rudolph Kompfner (1941) to use a chopped illuminating beam and high-frequency excitation of an ‘electrostatic’ lens and shows that it is incapable of providing any real improvement over a static lens ‘The maximum gain by Kompfner’s suggestion’, writes Gabor, ‘is about...0.5. Though this appears an appreciable improvement, it would be almost certainly outweighed by the technical difficulties arising from the complication of the scheme’.

Kompfner’s paper was not, however, the first on the subject for, 2 years earlier, Neßlinger (1939) had discussed the chromatic aberration of a high-frequency lens. In their book, Brüche & Recknagel (1941) discuss high-frequency phenomena in some detail. Later, in a very short note in the Physikalische Blätter, Scherzer (1946) tells us that correction of a weak lens has been achieved by superposing a sinusoidal voltage onto the lens excitation; moreover, spherical and chromatic aberration can be removed simultaneously. No details of the geometry are given but it is probably similar to that illustrated in Neßlinger’s article; we are told that the period of the high-frequency field is of the same order of magnitude as the time spent by the electrons in the field—this implies a frequency of the order of hundreds of megahertz or even gigahertz, though the use of such a frequency seems unlikely at that date. Scherzer concludes that ‘Da die Korrekturmöglichkeiten nicht auf schwache Linsen beschränkt sind, besteht Hoffnung, daß bei Verwendung von Hochfrequenz-Linsen die theoretische Auflösungsgrenze des mit zeitunabhängigen Feldern arbeitenden Elektronen-Mikroskops erheblich überschritten werden kann’.

Gabor did not, however, confine himself to a negative estimate of any possible benefit from the use of high frequency. He went on to show that the introduction of space charge into the path of electrons could well act as a diverging lens; by combining such a lens with a standard lens, a doublet with a very short focal length could be obtained, and this in turn could be expected to have low spherical aberration (there is a hint of this in his note in Nature; Gabor 1942). Gabor explores this proposal at considerable length (11 pages of the book are devoted to it) but, in the end, he admits that ‘There is one more question which is naturally provoked by the suggestion of using an electron cloud as a lens: Will the cloud not act on the beam like frosted glass? This is a very difficult question to decide...It can be said only, that...from an unpublished theoretical investigation of the author it appears that the effect will not cause serious interference’. Space charge is the subject of a later paper (Gabor 1947); Gabor’s comments on this suggestion are unaltered in the second, revised edition of his book (Gabor 1948a). Soon after, Gabor gave the problem to Eric Ash as his PhD project; Ash’s conclusions are unequivocal: ‘Some of the electron optical properties of an isolated space charge cloud have been derived. The possibility of obtaining a combination of a space charge lens with a space charge free converging lens which is free from spherical or chromatic aberration has been investigated.

The discussion of achromatization leads to the conclusion that it is not possible to correct either the thin electric or the thin magnetic lens, if the resultant action of the combination is to be convergent.
The possibility of reducing the spherical aberration of an electron microscope objective is also rejected. The calculation shows that the random microscopic field due to the discrete nature of the space charge is greater than the small macroscopic field required for correcting the spherical aberration.

The system considered in both these cases was one in which the space charge fields do not overlap with the converging lens field; the space charge was considered to be uniform over its length. It is not believed that the conclusions would be changed if these restrictions were removed (Ash 1955; related work is reported by Ash & Gabor 1955).

None of this detracts from the importance of Scherzer’s paper of 1947, in which the various ways of correcting chromatic and spherical aberration by relaxing the conditions for which the earlier derivation of the positive-definite integrand in $C_s$ was valid are described. The paper begins with a study of cylindrical lenses, first for chromatic aberration correction and then, after incorporation of octopoles, for spherical correction. This is not the place to recapitulate Scherzer’s reflections in detail but it is important to note that his proposal is very detailed and carefully argued: it is much more than a preliminary speculative suggestion. This is followed by a section on ‘Charges in the vicinity of the axis’. First, Gabor’s idea of using space charge is summarized. Scherzer observes that the difficulty of holding the required space-charge distribution stable could be overcome by introducing a charged foil (in the context of spherical correction) and returns to the idea in a short paper a year later (Scherzer 1948), this time for chromatic correction. We note in passing that the use of such a device remained in his mind and he returned to it many years later (Scherzer 1980), by which time several very thorough studies of such ‘foil-lenses’ had been made. Gabor (1946) had suggested that spherical correction could be achieved by placing an electrode on the axis of an electrostatic lens and an experimental study was made by Dungey & Hull (1947); Scherzer devotes a paragraph to this. This section of Scherzer’s (1947) paper concludes with a short discussion of electron mirrors as correctors. The fact that their spherical aberration coefficient can be negative had been established in the 1930s (Hottenroth 1936, 1937; Recknagel 1936, also published as Recknagel 1937) but Scherzer points out the need for great care close to the plane of reflection, where the axial component of the electron velocity falls to zero.

The third section contains a detailed discussion of high-frequency lenses, which clearly attracted him—it must be admitted that the extreme simplicity of microwave circuitry compared with multipoles must have been very seductive. A possible lens design is illustrated and its properties analysed. In the concluding section, in which the various procedures are compared, Scherzer opines ‘rein gefühlsmässig’ that it is either the non-rotationally symmetric systems or the HF lenses that will be the first to reach a resolution of a few ångströms.

We have seen that Gabor, who had been involved in the development of electron optics since the late 1920s, gave considerable thought during the wartime years to aberration correction. It was in 1948 that he asked himself whether, in the absence of a satisfactory corrector of lens aberrations, it might be possible to correct the image—it is said that he hit on the solution while waiting for his turn on the tennis courts of the British Thompson–Houston Research Division in Rugby: holography (Gabor 1948b, 1949, 1951a). Despite some instructive but abortive experiments, holography had to await the invention of coherent sources of light and electrons, after which it became a subject in its own right. For a
very vivid account of the early work, see Johnston (2005). An early assessment by Allibone (1958) is also memorable. We return to this topic later but, before leaving it, it is amusing to note that holography, invented to cancel the effect of spherical aberration of electron lenses, is now being performed in aberration-corrected electron microscopes (e.g. Geiger et al. 2008).

2. Forty disappointing years

During the next four decades, all of the four major families of correction methods attracted considerable attention. Here, we can only provide a representative selection; some other projects and the corresponding references are to be found in earlier surveys (Hawkes 1980, 2007, 2008; Hawkes & Kasper 1989, ch. 41; Rose 2008a,b).

Experimental work on correctors consisting of cylindrical lenses and octopoles began in the late 1940s in Darmstadt, where Scherzer gave Robert Seeliger the task of constructing and testing such a corrector for his PhD project. Although mechanical difficulties proved a major obstacle, Seeliger still managed to show that the spherical aberration coefficient of such a device (two cylindrical lenses and three octopoles) could be negative and hence be used as a corrector, at least in principle. His work is recorded in a series of articles in the recently launched Optik (Seeliger 1948, 1949a,b, 1951, 1953; cf. Scherzer 1949). Seeliger’s device was subsequently transferred to Tübingen, where Gottfried Möllenstedt (1956) again demonstrated that the combination had the desired properties.

These attempts to correct spherical aberration caught the attention of Geoffrey Archard, who was at Aldermaston Court, the AEI research establishment. He published several proposals for correction inspired by Scherzer’s cylindrical lenses. Doubtless, the possibility of producing a corrected model of the AEI electron microscope was a motive for these studies. The opening terms of the potential in a cylindrical lens are those of a round lens and a quadrupole and Archard is most often remembered for pointing out that the cylindrical lenses of Scherzer’s proposal and Seeliger’s experiments could advantageously be replaced by quadrupoles. He also described combined elements producing quadrupole and octopole fields simultaneously and contributed significantly to Scherzer’s idea (in the 1947 paper) for chromatic correction, resuscitated many years later by Weißbäcker & Rose (2000, 2001, 2002) and by Maas et al. (2000, 2001, 2003) and Henstra & Krijn (2000). Archard’s scattered publications are listed in full here for convenience (Archard 1954a,b, 1955a–c, 1958, 1960, 1966; Archard et al. 1960). An even more ‘non-rotationally symmetric’ proposal for chromatic correction was made by Gabor at the international congress on electron microscopy held in Paris in 1950 (Gabor 1951b, 1953); here, the correcting element has a helical optic axis.

It was clear to all the early workers that these complicated sets of multipoles would be difficult to align and adjust and so, for his Cambridge PhD project, Jack Burfoot searched for the corrected electrostatic lens without rotational symmetry having as few electrodes as possible. (It should be remembered that it was not yet clear whether it would be better to build an autonomous system that acted as a corrected objective lens or to design a corrector to be used in conjunction with a good round objective.) He described a four-electrode lens, each of the electrodes producing multipole fields, but concluded that they could not be machined with

Phil. Trans. R. Soc. A (2009)
the necessary precision (Burfoot 1953). Archard (1958) later suggested that the Burfoot design was unnecessarily complicated and that the same result could be achieved with four simpler elements: a combined quadrupole–octopole, a combined round lens and octopole, a second combined quadrupole–octopole and a pure quadrupole.

AEI was not the only company interested in aberration correction. In 1955, Glaser spent some time in New York at the Farrand Optical Company, manufacturers of an electrostatic electron microscope. There, he worked extensively on the optics of quadrupoles and octopoles. Some results are reproduced in his long chapter in the Handbuch der Physik (Glaser 1956a); the full calculations, never published by Glaser, are however available in manuscript (a copy is in the collections of the Cambridge University Library and there are doubtless copies elsewhere). This work was continued by Hermann Dušek in Vienna, where he presented his dissertation, directed by Glaser, in 1958 (Dušek 1959), shortly before Glaser’s death in February 1960 (see Grümm & Schiske 1996).

The following decade witnessed numerous attempts to develop aberration correctors of different kinds. In Darmstadt, Meyer investigated very thoroughly the parasitic aberrations of lenses and quadrupole–octopole correctors in order to understand why Seeliger’s system had been unsuccessful. The many such aberrations arising from misalignments were identified and the idea of using multipole stigmators to correct them was put forward. Meyer also studied external sources of image blur, notably AC fields and mechanical vibrations, and designed anti-vibration mountings and shielding against ambient fields (Meyer 1956, 1958a,b, 1961a,b). In Cambridge, I reconsidered the forms of the aberration coefficients of quadrupoles and octopoles, using the methods introduced by Glaser and perfected by Peter Sturrock (1951a, 1955); formulae for all the aberrations of quadrupoles had been established by Alexander Melkich (1947a,b) but it proved possible to find much simpler expressions and, thanks to the eikonal approach, to demonstrate any relations between the coefficients (Hawkes 1965a,b). Hans Deltrap (1964a,b) demonstrated convincingly that a quadrupole–octopole system can correct spherical aberration but the corrector was tested only on an electron-optical bench, not on an electron microscope. David Hardy showed that combined electrostatic–magnetic quadrupoles can be achromatic (or have negative chromatic aberration), as predicted by Kel’mann & Yavor (1961); this finding was rediscovered by Albert Septier (1963) and generalized by Hawkes (1965c). Michael Thomson used the early Cambridge computer EDSAC II to calculate quadrupole properties and Robin Preston later attempted to find expressions for the potential in a mixed quadrupole–octopole lens (Thomson 1966, 1967, 1972; Hardy 1967; Preston 1968,1969). In the University of California, Philip Meads (1963) studied the aberrations of quadrupoles and included in his calculations the aberrations of gradient as well as those of position; he also listed numerous relations between the various coefficients and later examined the parasitic aberrations (Meads 1967). Another early study is that of Whitmer (1956).

Another major centre for quadrupole studies was the A.F. Ioffe Physico-technical Institute in Leningrad; the ‘Russian quadruplet’ was introduced there and a stream of papers on the optical properties of quadrupoles and quadrupole quadruplets appeared by Kel’mann, Yavor, Ovsyannikova, Fishkova, Baranova, Dymnikov and others. Their work is summarized in a number of
review articles and books (Yavor 1968; Baranova & Yavor 1986, 1989 (transl.); Hawkes 1970, 1980). This Russian work had a distinguished predecessor in Aleksei Strashkevich, whose first publications on electron optics appeared (in Ukrainian) in 1937. He published many articles on quadrupole theory and his results are collected in two books (Strashkevich 1959, 1966). There were many other studies of quadrupoles as correctors and of course as guiding elements (strong-focusing lenses) in particle accelerators. Many of these are listed and classified in Hawkes (1966a, 1970).

In the 1970s, two determined attempts were made to put into practice the aberration correction techniques that had been shown to be promising. In Darmstadt, Scherzer, Rose and a succession of PhD students made systematic studies of the obstacles in the way of correction and went a long way to overcoming them (Koops 1978). But Scherzer died in 1982 and the project was dropped. The theoretical support for the project was provided by Harald Rose, who has written an extremely detailed and readable history of the project (Rose 2008a; see also Rose 2009), and I refer to this first-hand account rather than recapitulate it here. The other project was launched by Albert Crewe in the University of Chicago. In the 1960s, Crewe had succeeded in making a stable field-emission source with which he was able to develop the first STEM; he also attempted to obtain funding for the construction of a very high-voltage electron microscope at the Argonne National Laboratory (of which he was director at the time). In 1960, the first images from the Toulouse million-volte electron microscope had been published and the instrument was soon being heavily used; other laboratories, notably the Cavendish Laboratory, were already following Dupouy’s example. A major American project was therefore a natural aspiration and generous funding was obtained for an international workshop, which was held in 1966 at Argonne; during this, scientists from Europe, the USA and Japan prepared designs for a high-voltage, high-resolution electron microscope, which was to include a quadrupole–octopole corrector (Gilroy et al. 1966; Hawkes 1966b). Funding for the project was not approved but Crewe did not forget the corrector design that formed part of the blueprint and the conviction of the group of electron opticians working on the corrector that there was no fundamental reason why it would not work (Crewe et al. 1968). For six years (1972–1978), Beck and Crewe attempted to correct the spherical aberration of the probe-forming lens of a STEM but with no success; parasitic aberrations resulting from inhomogeneity of the materials used were blamed for the failure but the absence of fast diagnostic and control technology was undoubtedly another reason (Thomson 1972; Beck & Crewe 1974, 1976; Beck 1977; Crewe 2009). Among the other shorter-lived projects, we should not overlook the work of Müssig (1971a,b) in Dresden and the many papers by Daniel Dhuicq and Albert Septier (e.g. Dhuicq & Septier 1968; see also Septier 1966).

At the end of the same decade, a new approach to aberration correction emerged that was to have immense repercussions: sextupole correction. I had shown in 1965 that sextupoles exhibit a third-order aberration of the same nature as the spherical aberration of round lenses but did not see any use for this at the time as sextupoles have second-order aberrations. Rose & Plies (1973) and Plies (1973, 1974; see also Plies 2002) likewise included sextupoles in their very thorough study of the optics of general systems but did not consider the possibility of exploiting this third-order aberration for the same reason. But in

Phil. Trans. R. Soc. A (2009)
Beck showed that a pair of sextupoles could be disposed in such a way that the second-order axial aberration is balanced out; for subsequent work in Chicago, see Crewe (1980a, b, 1982, 1984) and Crewe & Kopf (1980a, b), whose ideas were tested by Chen & Mu (1991). Although Beck’s first design was not entirely satisfactory as it was not free of other unwanted aberrations, it was soon followed by a new configuration, proposed by Rose (1981), that did not suffer from this handicap. This was the forerunner of the sextupole correctors produced by Haider and colleagues at the CEOS Company with which aberration-corrected transmission electron microscopes are equipped today.

Looking back, we perceive that correction based on non-rotationally symmetric elements is the only technique that has been brought to a successful conclusion, in the sense that commercial instruments are equipped with such devices. Before leaving these decades of exploration, however, let us examine the fate of the other approaches that once seemed promising.

We have seen that, despite Gabor’s pessimistic estimate, Scherzer originally favoured high-frequency lenses (as well as non-rotationally symmetric elements). An experimental and numerical study of such lenses was embarked on in Cambridge by Lawrence Oldfield (1971, 1973a, b, 1974, 1976), who showed that the paraxial properties agreed well with the theoretical predictions. However, he was not able to obtain any general results concerning the aberrations and the problem of energy spread, and hence of chromatic aberrations, remained open. Some further work on the subject was reported by Vaidya (1972, 1975a, b) and Pandey & Vaidya (1977) and the aberration coefficients of microwave cavity lenses were listed by Matsuda & Ura (1974a, b). For a recent revival of the use of high frequency, see Schönhense & Spiecker (2002) and Schönhense et al. (2006).

Space charge and foil lenses have attracted much more attention. The optics of such arrangements have been thoroughly investigated by Gianola (1950), Hauke (1958), Verster (1963), Rus (1965), Barth (1967), Dekkers & Le Poole (1968), Typke (1968/1969, 1972a, b, 1976), Hoch et al. (1976), Thomson (Thomson & Jacobsen 1971; Jacobsen et al. 1971; Thomson 1973) and Munro & Wittels (1977). The contents of these studies are summarized in Hawkes (1980) and not repeated here. The most far-reaching practical studies of the correcting capacity of a thin conducting foil placed in the path of an electron beam are those of Maruse, Ichihashi, Hibino and others in the University of Nagoya. Their progress and successes are charted in the following papers: Maruse et al. (1970a, b), Ichihashi & Maruse (1971, 1973), Hibino & Maruse (1976), Hibino et al. (1977, 1978, 1981, 1998), Hanai et al. (1982, 1984, 1994, 1995, 1998) and Matsuda et al. (2005). For another project, see van Gorkum (1983) and van Gorkum & Beirens (1986). More recently, van Aken has devised a very different kind of foil lens, in which the abnormally high transparency of thin foils to very low energy electrons is exploited (van Aken et al. 2002, 2004; van Aken 2005).

The suggestion that a lens free of spherical aberration could be designed by placing a conductor on the axis was revived by Wilska (1961) and examined further by Lenz & Wilska (1966/1967) and Kunath (1968, 1970, 1972, 1976) and independently by Marais (1970) and Communay & Marais (1971).

One other possible correction method is, perhaps surprisingly, proving useful in certain contexts. We have mentioned that the spherical aberration coefficient of electron mirrors can be negative and that a combination of round lens and (round) mirror can hence, in principle, be free of spherical aberration. Several
ways of exploiting this have been explored over the years (Zworykin et al. 1945; Kasper 1968/1969 a,b; Rempfer & Mauck 1985, 1986, 1992; Shao & Wu 1989, 1990a,b; Rempfer 1990a,b; Crewe et al. 1995, 2000; Preikszas & Rose 1997; Rempfer et al. 1997; Tsai 2000), of which the most successful have been those of Gertrude Rempfer and colleagues in the University of Portland (Oregon) and of the partners in the SMART project and the PEEM project (Fink et al. 1997; Feng et al. 2005; Schmid et al. 2005; see Könenkamp et al. 2008). Simple though the idea may seem, it is fraught with practical difficulties, stemming essentially from the fact that the incident and image-forming beams occupy the same space and must somehow be separated without degrading the optics. In one proposal (Crewe 1992a,b; Crewe et al. 2000; Tsai 2000), it was even suggested that the beam should traverse the specimen twice, once in the forward direction as in any other transmission electron microscope and again on its return journey after reflection at the mirror. Two configurations have survived, in one of which (Könenkamp et al. 2008) the incident and reflected beams are parallel; a deflection system translates the incident beam towards the specimen and a second deflection system translates the reflected beam away from it. Considerable care is needed to ensure that these deflections do not introduce aberrations of their own. This is proving useful for low-energy electron microscopy (LEEM). The other forms part of a more complicated instrument, the Spectromicroscope for all Relevant Techniques (SMART); here the incident and reflected beams pass through a 90° magnetic deflection field as in the early Castaing–Henry analyser. For a full description, see Hartel et al. (2002).

Why, despite all these efforts and investment, did none of these aberration correction projects succeed at the time? Many reasons have been advanced and, in the case of the quadrupole–octopole systems, they can all be summarized in the word adjustment. Whatever the configuration, a large number of electrodes and/or polepieces have to be machined and positioned with the highest accuracy and their excitations have to be highly stable and correctly chosen. Added to this is the fact that a quadrupole–octopole system is particularly ill-adapted to correct a conventional transmission electron microscope, where correction is required over the whole region imaged. We thus have the situation in which a highly perfected objective lens, designed to provide the best possible resolution, is coupled to a series of quadrupoles which have comparatively huge aberrations; the octopoles must then cancel both the large quadrupole aberrations and the small objective aberration. No calculation is needed to show that this is a highly unstable arrangement, in which the slightest imperfection will be disastrous. Conversely, such a system should be much more suitable for a probe-forming system, a SEM or a STEM, in which correction is required only on or in the close neighbourhood of the optic axis. Sextupoles do not suffer so badly from this problem as they have no linear ‘paraxial’ effect. They have second-order effects which have to be eliminated by operating them in pairs but the azimuthal difference is far less severe than with quadrupoles, which have a linear diverging effect in one principal section and a linear converging effect in the other.

However, no multipole system will be perfect and provision must always be made for supplementary weak multipole correctors, which we may think of as generalized stigmators, and these aggravate the complexity of devices that are already complicated. The fundamental reason for the failure of the earlier correctors is that there was no way of operating them effectively. In order to
excite the main elements (quadrupoles and octopoles) as well as the numerous
generalized stigmators correctly, fast diagnostic tools are indispensable, with
direct feedback and routines to ensure that the process converges. The hardware
and software to perform these tasks did not become available until the 1990s,
the decade in which success was finally achieved. Auto-tuning was an essential
tool and has since become much more sophisticated in order to correct many
high-order parasitic aberrations as well as the spherical aberration of fifth order.
Looking ahead, we see that four approaches have been followed: that of Zach and
Haider for the SEM (e.g. Zach 2000); an image-based procedure for the STEM
(Krivanek et al. 1997b); a diffractogram-based method for STEM (Dellby et al.
2001; Krivanek et al. 2008a–c); and a diffractogram-based technique for TEM
(Haider et al. 2008b,c).

3. Success at last

Since it is clearly difficult to correct a TEM objective by means of quadrupoles
and octopoles, Joachim Zach and Max Haider revived such correctors for the
low-voltage SEM. Here, correction is needed, not to improve the quality of a
lens already operating at its highest performance level, but to render a relatively
mediocre system better. Zach (1989) described the design and in due course the
corrector performed satisfactorily (Zach & Haider 1994, 1995a,b; Zach 2006).
Meanwhile, Rose had shown that the objections to correction by means of
sextupoles could be overcome and he, together with Haider and Urban, applied
successfully to the Volkswagen Foundation for funding to build and test such
a corrector. The device was built in Heidelberg in the European Molecular
Biology Laboratory (EMBL) and the first results were reported by the middle
of the decade (Haider et al. 1995). As always, misalignment and constructional
inaccuracies caused problems and an iterative program was found that could
solve them (Uhlemann & Haider 1998). The corrector was then installed in a high-
performance transmission electron microscope and, in 1998, Haider published the
first images obtained with a corrected commercial TEM (Haider et al. 1998a–c).

During the same years, Ondrej Krivanek and Nicklas Dellby, working in the
Cavendish Laboratory, Cambridge, set out to correct the spherical aberration of
a Vacuum Generators STEM by means of a quadrupole–octopole system of the
now-familiar type. Unlike all their predecessors, however, they were able to call
on fast computer control to adjust the system and, in particular, the additional
correction coils whose role was to cancel the parasitic aberrations. In 1997, at
the Institute of Physics EMAG meeting in Cambridge, Krivanek showed that
correction had at last been achieved in a commercial instrument (Krivanek et al.
1997a). The meeting was also a celebration of the centenary of J.J. Thomson’s
papers in The Electrician and Philosophical Magazine confirming the particle
nature of the electron, a happy coincidence.

Behind the dates given in the foregoing paragraphs is an interesting
story. The evidence that correction had been achieved is of two kinds: first,
demonstrations that the corrector is behaving satisfactorily based on Zemlin
tableaux or Ronchigrams and, second, corrected images. For the quadrupole
correctors, evidence of the first kind is to be found in the 1997 papers by
Krivanek et al. (1997a,b) and, in much more detail, in Krivanek et al. (1999)
and Dellby et al. (2001), and, for the sextupole device, in the paper by Uhlemann and Haider, published in 1998 but submitted on 5 May 1997. Turning now to the question of corrected images, I am indebted to Urban for the following information about the German project. The first corrected images obtained with the sextupole corrector were recorded by Bernd Kabius on a Philips CM200 microscope in Heidelberg on 24 June 1997. The following day, these images were shown to Karel van der Mast and Ben Bormans of Philips Electron Optics in Eindhoven; the images were sufficiently impressive to overcome van der Mast’s earlier reticence and the decision to produce a corrected commercial instrument was taken that same day. An account of this work was submitted to Nature soon after but was not accepted for publication; Science likewise rejected it. After further negotiations, however, Nature did accept the paper, which appeared in the issue of 23 April 1998 (Haider et al. 1998a).

For the quadrupole project, Krivanek has likewise sent me many details. He and L. M. (Mick) Brown applied to the Paul Instrument Fund of the Royal Society for funding for such a corrector project in 1994; the results obtained with the first version of the corrector were presented at ‘Microscopy and Microanalysis’, the annual congress of the Microscopy Society of America held in Cleveland in August 1997 (Krivanek et al. 1997b). The abstract tells us that an early version of the aberration diagnostic software was working on the date of submission (March 1997) and by the time of the meeting preliminary results showing that the corrector worked in principle were available. It was after that talk that Batson decided to order a corrector for the IBM STEM. Similar findings were presented at the Cambridge EMAG meeting (Krivanek et al. 1997a), after which Krivanek and Dellby moved to Seattle and began work on an improved corrector. A pair of pictures, before and after correction, showing that correction had indeed been achieved with the Cambridge instrument (for which the spherical aberration coefficient of the objective lens was 3.5 mm at 100 kV) were published; HAADF pictures were also obtained with a resolution of 2.3 Å, the corresponding figure for the uncorrected instrument being 2.8 Å. These were recorded in the summer and autumn of 1997 (Krivanek et al. 1998).

At Nion in Seattle, the Mark 2 corrector was working by the spring of 2000 in a more recent VG STEM and a better site. This corrector was installed on the IBM instrument on 9 June 2000; a HAADF image with 1.4 Å resolution had been obtained before shipment (in an uncorrected instrument the limit was about 2 Å). This is reproduced in Dellby et al. (2001), which also shows an image of silicon dumb-bells, recorded soon after installation. It was this instrument that furnished the first sub-Ångström resolution pictures shown in Batson et al. (2002).

Just as earlier studies on multipole correctors had shown the importance of parasitic aberrations, so these triumphant announcements of spherical aberration correction served as unwelcome reminders that chromatic aberration is also an axial aberration and an enemy no less formidable than spherical aberration. Unless it too could be corrected (or rendered inoffensive), the progress that had been achieved remained limited. Two approaches were possible. First, the energy spread of the incident beam could be kept so small that the chromatic aberration would have little effect. Alternatively, the chromatic aberration could be corrected as well as the spherical aberration. The commercial electron microscope manufacturers have largely adopted the first solution, because it could
be implemented rapidly but mainly because the second type of solution did not yet exist! Several types of monochromator were developed and are today incorporated in commercial instruments. The current in the beam is inevitably diminished since the role of the monochromator is to eliminate electrons with energies outside the acceptance window. As these devices do not, strictly speaking, effect any correction, we shall not discuss them here—they represent a truce, not a victory over the enemy. A list of representative references is to be found in Hawkes (2008) but the subject is in rapid evolution.

Not many ways of cancelling the chromatic aberration coefficient $C_r$ are known and all require quadrupoles. The simple rule that sextupole correctors are best for TEMs and quadrupoles for STEMs is therefore breaking down. One ingenious all-electrostatic suggestion was already present in Scherzer’s (1947) paper and requires us to find a quadrupole potential function, $p_2(z)$ that satisfies

$$p_2(z) = \phi'' - \frac{\phi'^2}{8\phi},$$

where $\phi(z)$ is the electrostatic potential on the optic axis ($z$) and primes denote differentiation with respect to $z$. Configurations for which this condition is closely satisfied are known (Henstra & Krijn 2000; Maas et al. 2000, 2001, 2003; Weißbäcker & Rose 2000, 2001, 2002). The resulting correctors can be made capable of correcting spherical aberration as well; we refer to the papers cited for many more details.

The other possible way of correcting chromatic aberration involves the use of combined electrostatic and magnetic quadrupoles. We have already mentioned that Kel’mán and Vavor pointed out as early as 1961 that the chromatic aberration coefficients of such combined quadrupoles can be of either sign and pass through zero. They can be included in a configuration designed to correct spherical aberration, which will then become capable of correcting both types of aberration. Work on such correctors is in active progress (Dahmen et al. 2008; Haider et al. 2008a,b,d; Hartel et al. 2008; Krivanek et al. 2008a,b,c).

In the foregoing account, one important contribution is missing; certainly theory played an essential part but so did computation of the properties of lenses and correctors. The two major contributors to this branch of electron optics are Bohumila Lencová (SPOC\textsuperscript{1}) at the Institute of Scientific Instruments (Czech Academy of Science) in Brno and Eric Munro, John Rouse and colleagues at MEBS\textsuperscript{2} in London. Each has produced software that has been gradually improved over the years to the point where all the elements found in modern microscopes can be calculated rapidly and with very high accuracy. The latest program suite from SPOC, EOD (Electron Optical Design), is capable of calculating the properties of round electrostatic and magnetic lenses, including unconventional geometries, deflectors, aberration correctors, guns and analysers; an ingenious test of the accuracy of the results forms part of the suite. For details, see Lencová & Zlámal (2005, 2007, 2008) and the SPOC website. The MEBS software is likewise capable of calculating the properties of a wide range of electron optical devices, surveyed with numerous illustrations in Liu et al. (2002) and Munro et al. (2006) and more recently

\textsuperscript{1}Software for Particle Optics Calculations, www.lencova.com

\textsuperscript{2}Munro’s Electron Beam Software, www.mebs.co.uk
in Liu et al. (2008); cf. their distant ancestor (Munro 1973). The MEBS group calculate higher order aberrations by means of differential algebra (and by means of aberration integrals when these are available) whereas Lencová prefers exact trajectory calculation followed by matching to the known polynomial expansions (Oral & Lencová in press).

With these program packages, new configurations can be explored rapidly and with great accuracy; so fast are the calculations that attempts to find optimum arrangements by programmed searches have been largely supplanted by trial-and-error in near-real time.

One other feature of these and current developments deserves to be stressed: the role of symmetry (Hawkes 2005). Certainly this has never been neglected. It was while considering the relation between system symmetry and permitted aberrations that I found that sextupoles have an axial aberration analogous to the spherical aberration of round lenses. It is the symmetry properties of antisymmetric quadrupole systems such as the Russian quadruplet that give them their valuable optical properties; notably, equality of focal length in the two principal sections. Symmetry considerations became of paramount importance when the aberrations of such devices as the $\Omega$-filter, the mandoline filter and their many siblings came to be investigated for the introduction of symmetry and antisymmetry cancels many of the aberrations (Rose & Plies 1973; Plies 1978; Rose & Krahl 1995; Rose 2003). The symmetry properties of such devices as Rose’s Superaplanator and Ultracorrector (Rose 2005) are the key to their success.

4. Correcting images, not lenses

We have seen that barely a year separates Scherzer’s proposals for lens correction and Gabor’s invention of holography. In one case, an improved image is generated by improving the optics of the instrument. In the other, the shortcomings of the optical elements are accepted and their deleterious effects on the image are removed in a further reconstruction step. It was not until the late 1960s and the 1970s that electron holography came into being thanks to the development of more coherent electron sources and the introduction of the electron biprism by Düker (1955) and Möllenstedt & Düker (1955, 1956) (see Möllenstedt 1991). In Europe, the publications of Wahl (Möllenstedt & Wahl 1968; Wahl 1974), which culminated in his Habilitationsschrift (Wahl 1975), form the first major study of off-line holography, while, in Japan, an early in-line hologram and reconstruction were published by Tonomura & Watanabe (1968).

The situation is, however, rendered more confused by the theoretical work on high-resolution image formation and processing of the same period. The notions of spatial frequency and transfer functions in microscope optics had been introduced into light optics by Pierre-Michel Duffieux more than 20 years before (Duffieux 1946), but became widely known only with the publication of Born & Wolf’s Principles of Optics (1959), which contained the first clear presentation in English of these ideas. In the 1960s, Karl-Joszef Hanszen and colleagues in Braunschweig introduced this approach into electron microscope image formation theory and were struck by the fact that the interference between the unscattered beam and the beam scattered in the specimen, which is at the origin of the image contrast,
is the same as that between the reference beam and the information-carrying beam in (in-line) holography. Much of the literature of image formation of the 1970s is therefore written in the language of holography and numerous attempts were made to reconstruct an improved electron image by inverse filtering. We cannot chart that work here, interesting though it is, but it certainly merits a historical study and assessment. The various strands—in-line holography, inverse filtering, Wiener filtering, optical reconstruction and the beginnings of digital image processing, the connection with the phase problem and the large literature that followed the appearance of the Gerchberg–Saxton (1972, 1973) algorithm—are all interwoven. This has not yet attracted the attention of historians of science, apart from a fascinating paper on George Stroke’s role in the story by Johnston (2004). The surveys by Hanszen give an insider’s idea of the subject (Hanszen 1971, 1973, 1982).

Just as in light optics, it is off-axial holography that has proved the most fruitful. Although there has been (and still is) some interest in the use of crystals to split a quasi-coherent electron beam into two parts, most of the work has been performed with the electron biprism, a thread placed in the path of the beam that creates two virtual sources, just as Fresnel’s biprism does for light.

The first in-line (Fraunhofer) hologram and its light-optical reconstruction were published by Tonomura & Watanabe (1968), and they are reproduced in Hawkes & Kasper (1994, fig. 54.1); see also Tonomura et al. (1968). Similar attempts were made in the University of Reims (Gallion et al. 1975; Troyon et al. 1976; Bonnet et al. 1978; Bonhomme & Beorchia 1980) and in Chicago by Saxon (1972a,b) and Munch (1975). Off-axis holograms were first obtained by Möllenstedt & Wahl (1968) and reconstructed using laser light. In the following year, Tonomura (1969) formed an off-axial hologram using not a biprism but a crystal to split the incident beam into two. In 1970, Tomita produced holograms from point sources (Tomita et al. 1970, 1972) and Wahl’s thorough study of the technique appeared soon after (Wahl 1974, 1975).

Our subject is, however, aberration correction and it was not for some years that Gabor’s original ambition was realized optically (Tonomura et al. 1979) and digitally (Franke et al. 1987; Lichte 1990, 1991, 1993; Fu et al. 1991). An attempt to recreate the conditions envisaged by Gabor was made by Juliet Rogers (1978a,b, 1980), who had an extraordinarily bad glass lens manufactured in order to model the optics of an electron microscope. Today, electron holography is widely used, mostly to study phase effects and magnetic field distributions in particular. A recent survey by Lichte & Lehmann (2008) describes many of the current areas of application.

For a historical account of the subject by one of the pioneers, see Tonomura (1999) as well as his beautifully written and informal book (Tonomura 1998), and for many references to the early literature, see Hawkes & Kasper (1994, ch. 63).

Digital image processing soon progressed far beyond holographic reconstruction. Three-dimensional reconstruction and the phase problem were the main preoccupations of the early years (the late 1960s and the 1970s) though more basic image improvement was implemented in the scanning electron microscope even earlier, since the sequential nature of image formation in that instrument lent itself to analogue processing (see Hawkes (1978) for a bibliography). Once these subjects had reached maturity, spherical aberration correction was incorporated.
into the routines and other approaches to exit-wave reconstruction appeared (e.g. Kirkland et al. 2008; Petersen et al. 2008); the incorporation of correction into electron tomography is proving difficult.

With the arrival of aberration-corrected electron microscopes, exit-wave reconstruction has entered a new period, which is not part of my subject here. Before concluding, I wish to draw attention to a body of work related to image processing that is still not widely known. In the early days of image processing, which was developed in many areas with little or no communication between them, numerous procedures and algorithms were inevitably duplicated though the vocabulary and notation used were often so different that their similarity was almost unrecognizable. In an earlier introduction to the subject, I wrote ‘Despite the stalwart efforts of the authors of many textbooks on image processing to impose a pattern on their subject, the reader is all too likely to perceive it as a magpie collection of methods, gadgets, knacks and contrivances’ and it was in the hope of imposing some order on this disparate material that an image algebra was developed. In this elegant mathematical toolbox, we have operands, of which the most important is the image, and operators, such as add, multiply, max and min. Although it is usual to think of the image as a matrix of binary or grey-level values, and this is of course included in the algebra, the definition of an image is much more general. An $F$-valued image $a$ on the coordinate set $X$ is the graph of a function

$$a : X \rightarrow F,$$

$$a = \{ (x, a(x)) | x \in X \}$$

$a(x) \in F$. Thus the pair $(x, a(x))$ characterizes a pixel of the image $a$, the value at the pixel identified by its coordinates $x$ being $a(x)$. The ‘value’ may be a grey-level or a binary value for a black-and-white image; it may also be something much more complicated: a set of values for a colour image, for example, or the energy-loss spectrum at the point in question. Another very important type of ‘image’ is the template, for which the value at each pixel is another image. Such images are encountered, for example, in the STEM, in which a far-field diffraction pattern is formed at the detector for each object-element; the user is not aware of this since the detector(s) adds all the pixel values captured (Hawkes 1995). In STEM ptychography (Rodenburg 2008), this view of the STEM image is at the heart of the method. Another way of picturing the template is as a discrete form of a function of two sets of two (or more) variables, $f(x, y; p, q)$ say. One pair of variables $(x, y)$ characterizes the position of the pixel in question in the main image while the other pair $(p, q)$ characterizes the position of each pixel in the subsidiary image located at $(x, y)$. The Fourier transform (and other such transforms) is represented by a template–image operation and templates play a central role in the branch of image processing known as mathematical morphology. Here, the basic operations of erosion and dilation of an image by a structuring element are represented by another template–image operation, in which the structuring element is characterized by the template. Image algebra has revealed many very interesting parallels between the (linear) convolutional techniques that are widely used for image enhancement and the (nonlinear) morphological operations. As a simple example, we reproduce the results of a traditional convolution of an image $a$ with
a template \( t \) (as in the Sobel filter) and of a morphological convolution of an image \( a \) with a structuring-element \( t \) (as in a dilation). For the first, we obtain a new image \( b \),

\[
b = \{ (y, b(y)) | y \in Y, \quad b(y) = \sum a(x) t_y(x) \}
\]

and for the second

\[
b = \{ (y, b(y)) | y \in Y, \quad b(y) = V a(x) + t_y(x) \}
\]

in which \( V \) signifies max. The resemblance is striking and is no coincidence. (We have followed the usual convention of writing \( t_y(x) \) instead of the cumbersome \( t(y)(x) \).) For further information and a full presentation, we refer to the seminal papers by Ritter et al. (1990), Ritter (1991) and Davidson (1992). Dougherty’s (1989a, b) account of a slightly different form of the algebra is likewise easy to follow. More recently, Ritter & Wilson (2001) have produced a large collection of image processing algorithms expressed in the terminology of image algebra.

Why do we mention this in the context of aberration correction? Image correction, which began with Gabor’s holography, has developed into a rich and varied field of study and many of the operations involved can be written compactly in algebraic form. Image algebra has so far been used only to represent and often to generalize image processing procedures. It would be a natural extension to use it to represent the image-forming process in microscopes, thus forming a continuous chain from the source to the image and thence to any image processing sequences that might be envisaged. No serious attempt to do this has yet been made, however.

5. Concluding remarks

It is now some 10 years since Krivanek et al. (1997a, b) and Haider et al. (1998a–c) announced that quadrupole–octopole and sextupole correctors, respectively, had been successful in correcting the spherical aberration of commercial high-resolution electron microscopes. Aberration-corrected TEMs and STEMs are now in regular service in numerous laboratories and their modes of operation, the interpretation of the images they produce and the enhancement or even the creation of contrast (a task that used to be entrusted to the spherical aberration) by means of real or virtual phase plates are active areas of research. With the correction of spherical aberration, the battle against chromatic aberration is in full swing. In other areas, notably for electron beam lithography but also for high-resolution electron microscopy, it is necessary to eliminate the other third-order geometrical aberrations and Rose has shown how this can be achieved with systems of considerable complexity, rendered manageable by intricate symmetries (Rose 2008a–c).

A few years ago, electron opticians were regarded as a near-extinct species. Today, the subject is in a healthy state of effervescence and we can confidently forecast many more exciting years for electron optics.
References


Britche, E. & Recknagel, A. 1941 *Elektronengeräte, Prinzipien und Systematik*. Berlin, Germany: Springer.


Duffieux, P.-M. 1946 L’Intégrale de Fourier et ses Applications à l’Optique (Besançon (printedin Rennes), privately printed).


Phil. Trans. R. Soc. A (2009)
Review. History of aberration correction


Phil. Trans. R. Soc. A (2009)
Hawkes, P. W. 1965a The geometrical aberrations of general electron optical systems. I. The conditions imposed by symmetry. II. The primary (third order) aberrations of orthogonal


Kunath, W. 1968 Zonal correction of spherical aberration by magnetic fields of rotational symmetry.


Phil. Trans. R. Soc. A (2009)


Review. History of aberration correction


Plies, E. 1974 Korrektur der Öffnungsfehler elektronenoptischer Systeme mit krummer Achse und durchgehend astigmatismusfreien Gaußschen Bahnen. II. *Optik* 40, 141–160.


Remper, G. F. 1990b Comment on ‘A study on hyperbolic mirrors as correctors’. *Optik* 86, 73–74.


*Phil. Trans. R. Soc. A* (2009)


Rus, P. J. 1965 *Correctie van sferische aberratie met behulp van een gaaslens*. Afd gehersverslag, Technische Hogeschool Delft.


Thomson, M. G. R. 1972 The primary aberrations of a quadrupole corrector system. Optik 34, 528–534.


Zach, J. & Haider, M. 1995b Correction of spherical and chromatic aberration in a low voltage SEM. *Optik* 93, 112–118.