

Two Procedures for Phase Estimation from Visibility Magnitudes

R. J. Sault

School of Electrical Engineering,
University of Sydney, Sydney, N.S.W. 2006.

Abstract

Two procedures, which retrieve phase information from magnitude data, are examined. These procedures, one due to Bates, Fright and Garden and the other due to Fienup, are firstly reviewed, and then numerical results are presented to quantify the effectiveness of the algorithms. Some areas where Bates' algorithm could be further refined are suggested.

1. Fourier Phase Problem

In many imaging fields it is far easier to measure the Fourier transform (or visibility) of an object, rather than to measure the object's image directly. The visibility phase is usually far more difficult to measure than the visibility magnitude, and indeed is sometimes impossible to measure. In cases where phase information is absent (or extremely unreliable) we have the Fourier phase problem, namely estimating the visibility phase.

$$\phi(u, v) = \arg\{F(u, v)\},$$

from the visibility magnitude $|F(u, v)|$.

This paper first starts with a brief review of the problem, followed by considerations of existence and uniqueness of a solution. It then proceeds to a more detailed examination of two procedures, one due to Bates (1982), Fright and Bates (1982) and Garden and Bates (1982), and the other due to Fienup (1978, 1982). Computational experiences with the procedures on simulated data are discussed.

2. Statement of the Problem and Background

For notational simplicity only two-dimensional images will be considered. Extension to higher dimensional images is straightforward. We note, however, that the one-dimensional case (phase retrieval for signals) is very different, and contains many inherent ambiguities. Additionally the images will be assumed positive and compact (i.e. non-negative in a finite region and zero elsewhere). Again for notational simplicity, images will be assumed to be nonzero only in the region $M = [-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]$.

The Fourier phase problem can now be stated as

$$\begin{aligned} \text{Given} \quad & |F(u, v)| \\ \text{estimate} \quad & \phi(u, v) = \arg\{F(u, v)\} \\ \text{subject to} \quad & f(x, y) \geq 0, \quad (x, y) \in M \\ \text{and} \quad & = 0, \quad (x, y) \notin M, \\ \text{where} \end{aligned}$$

$$\begin{aligned} F(u, v) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \exp\{i 2\pi(xu + yv)\} dx dy \\ &= \mathcal{F}\{f(x, y)\} \\ &= |F(u, v)| \exp\{i \phi(u, v)\}. \end{aligned} \quad (1)$$

The problem could be equivalently expressed in estimating the phase from a power spectrum

$$P(u, v) = |F(u, v)|^2, \quad (2)$$

or an autocorrelation function

$$\begin{aligned} R(x, y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) f(x+p, y+q) dp dq \\ &= \mathcal{F}^{-1}\{P(u, v)\}, \end{aligned} \quad (3)$$

as these are simply related to the visibility magnitude.

Various forms of the phase problem appear in many scientific disciplines. In electron microscopy the image and its Fourier transform are both complex valued, and magnitude data are available in both domains. Gerchberg and Saxton (1972) have devised an iterative algorithm which will recover the phase. This algorithm will be discussed further in Section 7, as it is the basis of Fienup's procedures. In X-ray crystallography, the diffraction pattern is the Fourier transform of the atomic lattice. However, the symmetrical properties of the lattice heavily constrain the possible phases (Ramachandran and Srinivasan 1970), and consequently the phase problem is somewhat different from that considered here. In addition, the crystal lattice will be periodic which, as Bates (1982) has pointed out, produces a problem quite different from the compact image case.

In radio astronomy, Gull and Daniell (1978) have used the maximum entropy method (MEM) to solve the phase problem. As the magnitude data is undoubtedly noisy, MEM finds (by an iterative technique) the image which maximizes the entropy measure

$$H = \sum_{k,l} p(k, l) \log\{p(k, l)\},$$

where

$$p(k, l) = f(k, l) / \sum_{k,l} f(k, l).$$

The maximization is constrained so that the solution is consistent with the magnitude data. The MEM differs significantly from the algorithms by Bates, Fright and Garden

and by Fienup in that the latter two neglect noise, and essentially look for the unique image which is consistent with the data (exactly) and the constraints.

3. Existence and Uniqueness

Two immediate points of interest are existence and uniqueness of a solution to the phase problem. A necessary condition for existence of a compact positive image obviously is that the autocorrelation function must also be compact and positive. (Additionally, the autocorrelation of a real function will always be centrosymmetric or even.) An arbitrary compact positive centrosymmetric function, however, need not be a candidate for an autocorrelation function. Fienup *et al.* (1982) gave a simple counter-example. No sufficient condition to guarantee existence is known.

Image positivity requires that the power spectrum must satisfy properties such as (Papoulis 1965)

$$P(0,0) - P(u,v) \geq 0, \quad (4a)$$

$$\geq 4^{-n} \{P(0,0) - P(2^n u, 2^n v)\}. \quad (4b)$$

Physical existence of the image is often sufficient to indicate existence of a solution, but measurement imperfections and noise can make the data incompatible with a positive compact image.

Uniqueness is a far more interesting question. It is well known that without phase, certain 'trivial' characteristics of an image cannot be recovered (Bates 1982). In particular we cannot distinguish $F(u,v)$ from its conjugate mirror $F(-u,-v)$. This, however, only introduces an ambiguity of 180° rotation in the image, and does not affect the image's 'form' (Bates 1982). Even the crudest of phase estimates can be used to distinguish between these 'trivial' ambiguities.

To analyse less trivial ambiguities, let us consider dividing (sampling) an image into an array of picture elements, or pixels. If we sample at a rate of unity, we define the z -transform of an image (Bracewell 1978) as

$$F_z(z_1, z_2) = \sum_{k,l} f(k,l) z_1^{-k} z_2^{-l}. \quad (5)$$

As the image is compact, only a finite number of the pixels will be nonzero and $F_z(z_1, z_2)$ is a finite polynomial (probably with negative and positive powers). The z -transform of the image's autocorrelation function can be shown to be (Bracewell 1978)

$$P_z(z_1, z_2) = F_z(z_1, z_2) F_z(z_1^{-1}, z_2^{-1}). \quad (6)$$

This relates the polynomial $P_z(z_1, z_2)$ (whose coefficients can be derived from the visibility magnitude) to the product of two polynomials. As two (and higher) dimensional polynomials can almost never be factorized (reduced) (Bruck and Sodin 1979; Hayes and McClellan 1982; Fiddy *et al.* 1983), then it would be unlikely that $F_z(z_1, z_2)$ could be factorized. Consequently the factorization of $P_z(z_1, z_2)$ into two factors would usually be unique. Each of these factors could be taken as the image's z -transform. (This is the ambiguity between an image and its conjugate mirror image.) Although counter-examples to this uniqueness argument can easily be constructed, it is believed that these are pathological. The problem of uniqueness can be seen as a problem of guaranteeing the irreducibility of the image's z -transform. This approach has been taken by Fiddy *et al.* (1983).

It should be noted that the uniqueness argument fails for one-dimensional images, as polynomials in one dimension can always be factorized. This is a real ambiguity, and indeed one-dimensional problems will, in general, have many solutions. Fienup (1978) has noted that, in some instances, all solutions have similar forms and so the solution might be 'effectively' unique. Additional prior knowledge, such as symmetries or the constraints on the roots of the z -transform $F_z(z_1, z_2)$, can sometimes be used to guarantee unique solutions, even in one dimension.

The effects of noise on uniqueness in the two-dimensional problem is unclear. Bates (1982) conjectured that the noise present means that the image's z -transform is never reducible. Due to noise, however, there will be an infinite number of autocorrelation functions which are consistent with the measurements.

The computational experiences of Fienup (1978, 1982) tend to suggest that the solutions are usually unique, although Fiddy *et al.* (1983) have noted that Fienup's techniques work better when the solution is unique.

4. Canterbury Algorithm—Theory

Bates, Fright and Garden have proposed a method for solving the phase problem which they call the 'Canterbury algorithm' or 'crude phase estimation'. The algorithm consists of the three steps of firstly estimating a phase difference cosine, secondly estimating a phase difference, and finally estimating the phase.

Bates and Fright (1983) have used the Canterbury algorithm as one stage in a several-stage phase-retrieval procedure, however, only the Canterbury algorithm stage will be considered here. Other stages of the procedure were intended for use with low contrast or noisy images (these are not relevant to the tests performed here), or to tidy up the phase estimate using Fienup's algorithm (see Sections 7 and 8).

To understand the Canterbury algorithm, we recall that the image is confined to the region $[-\frac{1}{2}, \frac{1}{2}] \times [-\frac{1}{2}, \frac{1}{2}]$. The sampling theorem (Bracewell 1978) can now be applied to the object's visibility. By sampling at the Nyquist rate (which is unity in this case), the visibility can be expressed using sinc function interpolation:

$$F(u, v) = \sum_{n, m = -\infty}^{\infty} F(m, n) \operatorname{sinc}(u - m) \operatorname{sinc}(v - n), \quad (7)$$

where $\operatorname{sinc}(x) = \{\sin(\pi x)\}/\pi x$.

If sampling is performed at higher rates (i.e. closer spacing in the visibility plane), then there will be redundancy in the samples. The Canterbury algorithm uses this known redundancy to estimate the phase.

The visibility midway between two sampling points on the u -axis is given by

$$F(k + \frac{1}{2}, l) = \sum_{m = -\infty}^{\infty} F(m, l) \operatorname{sinc}(k - m + \frac{1}{2}). \quad (8)$$

Truncating the infinite series to the two dominant terms, we have

$$F(k + \frac{1}{2}, l) \approx \{F(k, l) + F(k + 1, l)\} \operatorname{sinc}(\frac{1}{2}). \quad (9)$$

The truncation to two terms would introduce a large error. It is known that there are better interpolation kernels for two-point interpolation than the sinc function.

Representing the interpolation weight by W , we then have

$$F(k + \frac{1}{2}, l) \approx W \{F(k, l) + F(k + 1, l)\}, \quad (10a)$$

$$|F(k + \frac{1}{2}, l)| \exp\{i\phi(k + \frac{1}{2}, l)\} \approx W[|F(k, l)| \exp\{i\phi(k, l)\} + |F(k + 1, l)| \exp\{i\phi(k + 1, l)\}]. \quad (10b)$$

Taking the magnitudes of both sides, and manipulating the equation, we get

$$\cos\{\phi(k + 1, l) - \phi(k, l)\} = \frac{\{|F(k + \frac{1}{2}, l)|/w\}^2 - |F(k, l)|^2 - |F(k + 1, l)|^2}{2|F(k, l)||F(k + 1, l)|}, \quad (11)$$

which can be interpreted as an application of the cosine rule (of elementary trigonometry).

Thus by sampling at twice the Nyquist rate (and thus measuring terms like $|F(k + \frac{1}{2}, l)|$), an approximation to the cosine of the phase differences can be obtained. The technique could be easily generalized to non-cartesian systems (such as the radial data of radio astronomy), although the justification and practice might be more difficult.

Although the approximations might seem gross, the errors should decrease as the sampling rate is increased. In other words, for sufficiently close sampling, even linear interpolation should prove reasonably accurate.

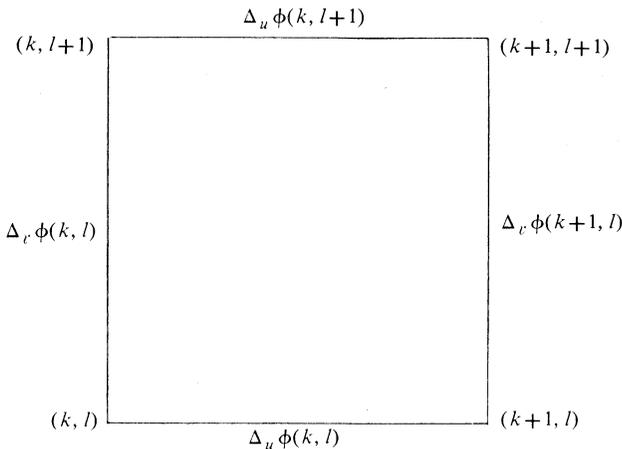


Fig. 1. Phase differences on a cartesian grid.

Fright and Bates (1982) suggested several methods for choosing the interpolation weight W , for example that W be kept constant within a region and chosen so as to minimize an error measure within that region. The region may vary in size from a few points to the entire u - v plane. Alternative arguments, based on differential calculus, would imply that linear interpolation (i.e. $W = 0.5$) would be the best choice.

We define the phase differences in the u and v directions as

$$\Delta_u \phi(k, l) = \phi(k + 1, l) - \phi(k, l), \quad \Delta_v \phi(k, l) = \phi(k, l + 1) - \phi(k, l),$$

respectively. The absolute values of these phase differences are estimated by taking an inverse cosine of the cosine estimates of equation (11) (note that the cosine function loses the sign and only the absolute value is retrievable). To retrieve the sign, we consider the four phase differences around a square (see Fig. 1).

The sum of phase differences around a closed path must be zero (modulo 2π). If the sign of one phase difference is known, then the signs of the other three can be chosen so as to minimize the deviation from zero (modulo 2π) of the sum of the phase differences around the path. There will generally be some discrepancy, which we call the phase mismatch (it will be used as an error measure later). We call this the 'one-sign method' for estimating the sign of differences.

The algorithm is a bootstrap one. If the sign of one phase difference is known, three others can be estimated. The procedure is started by arbitrarily choosing a sign for the first phase difference, a choice which is a manifestation of the ambiguity between the image and its conjugate mirror image. A natural phase difference to choose is one involving the origin.

We note that if the signs of two phase differences around a path (rather than just one) are known, then a similar procedure can be followed to estimate the signs of the remaining two differences. We call this the 'two-sign method' for estimating the sign of differences. The algorithm can be programmed so that the 'one-sign method' need be used only near the positive u and v axes. The 'two-sign method' can be used for estimating all the other signs.

The algorithm is completed by estimating the phases from phase differences. A least-squares method, similar to that analysed by Hunt (1979), could be used. Fright and Bates (1982) used a much simpler technique whereby phases and phase difference signs are determined simultaneously. They assumed that $\phi(k, l)$ and $\phi(k+1, l)$ and the signs of the three remaining differences (see Fig. 1) had been determined. Then ideally, this gave

$$\begin{aligned}\phi(k, l+1) &= \phi(k, l) + \Delta_v \phi(k, l) \\ &= \phi(k+1, l) + \Delta_v \phi(k+1, l) - \Delta_u \phi(k, l+1).\end{aligned}\quad (12)$$

Due to phase mismatch, these two phases will not generally be the same and so $\phi(k, l+1)$ is taken as the phase average of the two, and similarly for $\phi(k+1, l)$. To start the procedure, it is known that the phase at the origin is zero for positive images. A neighbouring point, say $(1, 0)$, is then given the phase $\pm |\Delta_u \phi(0, 0)|$ (the sign is again arbitrary). We call this the 'phase averaging method'.

5. Canterbury Algorithm—Computational Aspects

To investigate the properties of the Canterbury algorithm, a number of known images were Fourier transformed, and their phases discarded. All images were relatively simple, consisting of 1–5 gaussians, discs or delta functions. The images were nonzero in a rectangle varying in size from 17×17 to 50×50 pixels. The images were always zero-extended to allow various sampling rates in the visibility domain. The zero-extended images were of sizes varying from 64×64 to 512×512 pixels. No noise was modelled.

In our implementation of the Canterbury algorithm, the interpolation weight was made constant throughout the u - v plane; usually it was set to 0.5 (i.e. linear interpolation). 'Phase averaging' to share the phase mismatch was performed (the more sophisticated techniques of Hunt 1979 were not used). The 'one-sign method' was used to determine phase difference signs near the positive u and v axes, while the 'two-sign method' was used elsewhere.

The image array output by the Canterbury algorithm is one-quarter the size of the input magnitude array (if the input magnitude is $N \times N$, the output image is $\frac{1}{2}N \times \frac{1}{2}N$). As it is convenient for the input and output arrays to be the same size, two methods were chosen to extend the image to $N \times N$. Firstly the image could be zero-extended, using the assumption that the image is compact. The alternative is to perform two-point interpolations on the visibility, before the Fourier transform, using the same interpolation weight as used in the cosine rule. This is consistent with the assumption that two-point interpolation is accurate. Edge effects in the interpolation were handled so as to make the algorithm exact when reconstructing a simple delta function. The two methods were found to give almost identical results.

Two error measures were chosen for analysis. These were the average phase mismatch error ε_p (the absolute value of the phase mismatch averaged over the $u-v$ plane) and the visibility magnitude error ε_v , defined as

$$\varepsilon_v = \left(\sum_{k,l} \left| |F(k,l)| - |G(k,l)| \right|^2 \right)^{\frac{1}{2}}, \quad (13)$$

where $F(k,l)$ is the measured visibility magnitude and $G(k,l)$ is the estimate of the visibility by the Canterbury algorithm, and the sum is over all measured points in the $u-v$ plane. Various aspects of the results are now discussed.

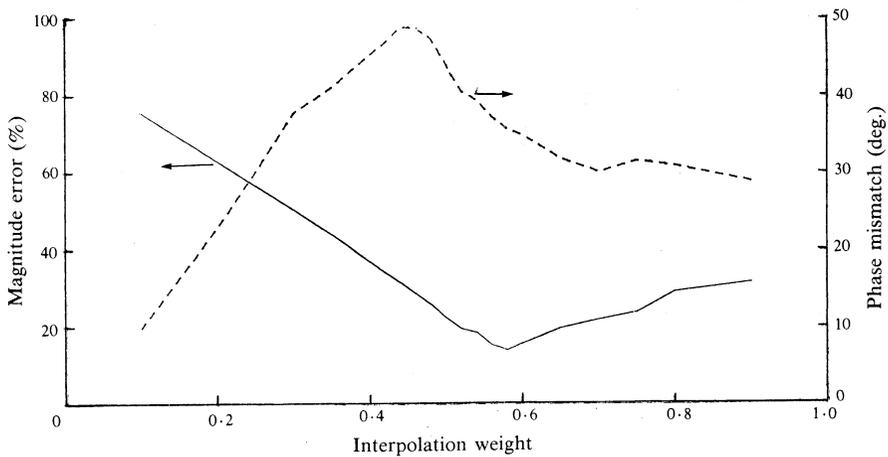


Fig. 2. Variation of magnitude error (solid curve) and average phase mismatch (dashed curve) with interpolation weight for an image shown in Fig. 3. The magnitude error has been normalized by the object's energy.

(a) Effect of Varying the Interpolation Weight

Fig. 2 gives a plot of the magnitude error and the phase mismatch error against interpolation weight, for one particular image, shown in Fig. 3. Fig. 4 gives four reconstructions of this image for various interpolation weights. The phase mismatch error is surprisingly large. Fright and Bates (1982) have suggested that a good choice for W is that which minimizes the magnitude error ε_v . For this case it is $W = 0.58$, while a subjective examination of the reconstructions would indicate that $W = 0.52$ is the 'best' weight for this image. Even for the 'best' interpolation weight, the reconstruction is only approximate.

Fig. 3. A typical image.

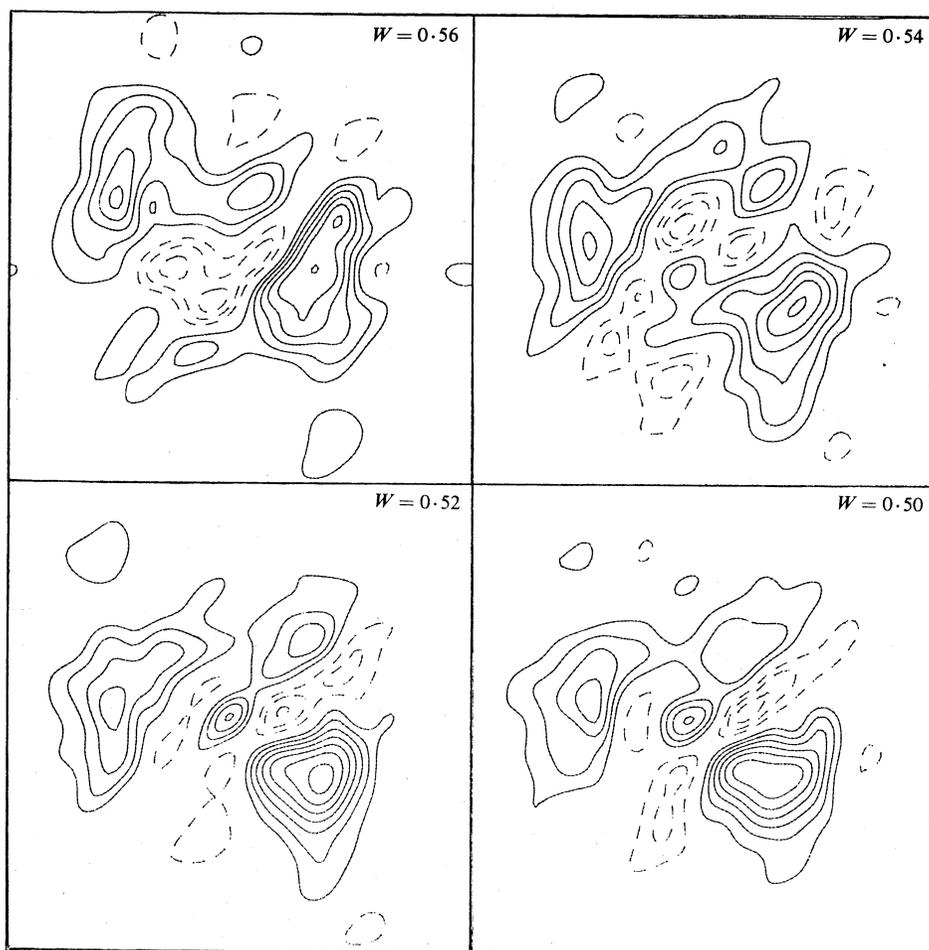
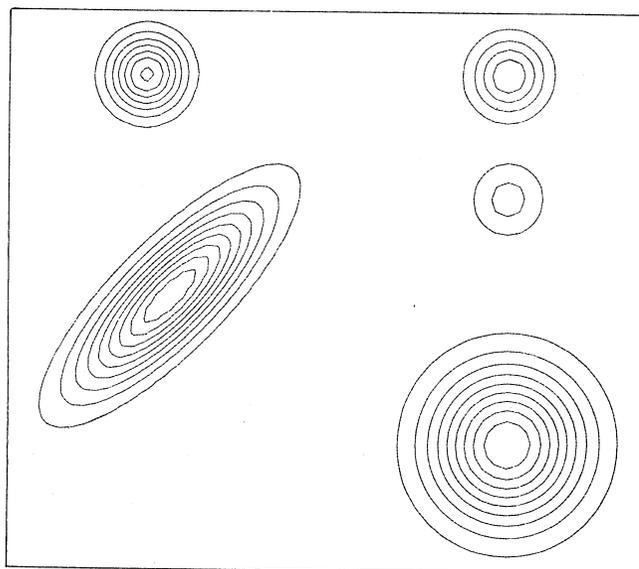


Fig. 4. Four Canterbury algorithm reconstructions of Fig. 3, using various interpolation weights W .

(b) Variation of Image Complexity

A sequence of four similar images were made, each consisting of a central gaussian, surrounded by a number of smaller gaussians of half the size (the number of smaller gaussians was varied from zero to three). The smaller gaussians were arranged so that peaks in the autocorrelation function would not interfere. The results of the sequence, shown in Figs 5 and 6, are very good for the simple images, but there is an increasing failure of the algorithm as the image complexity increases.

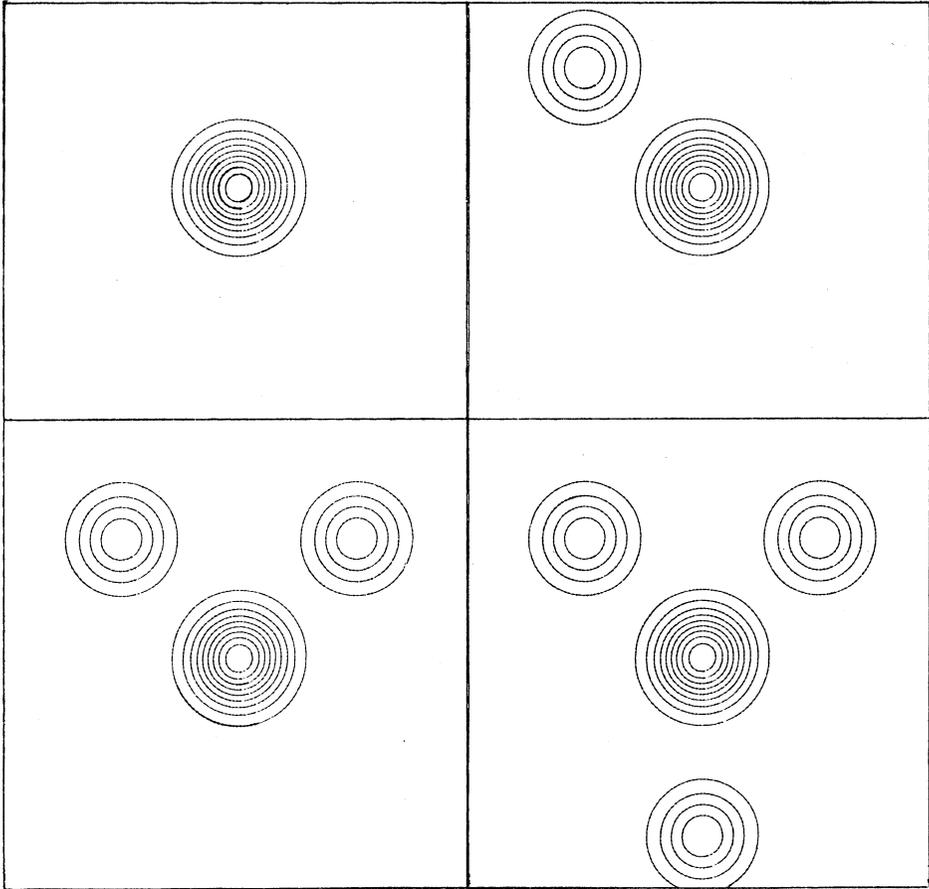


Fig. 5. Four original images, in order of increasing complexity.

Another class of images on which the algorithm was used consisted of a disc, with features on top of the disc. The algorithm produced very poor reconstructions for these images which looked, at times, like random noise (even when the magnitude and phase mismatch errors were quite low). Clearly these images are quite complex.

Finally the algorithm was tried on images which Fiddy *et al.* (1983) have shown to possess unique solutions. The uniqueness did not cause the algorithm to produce better results.

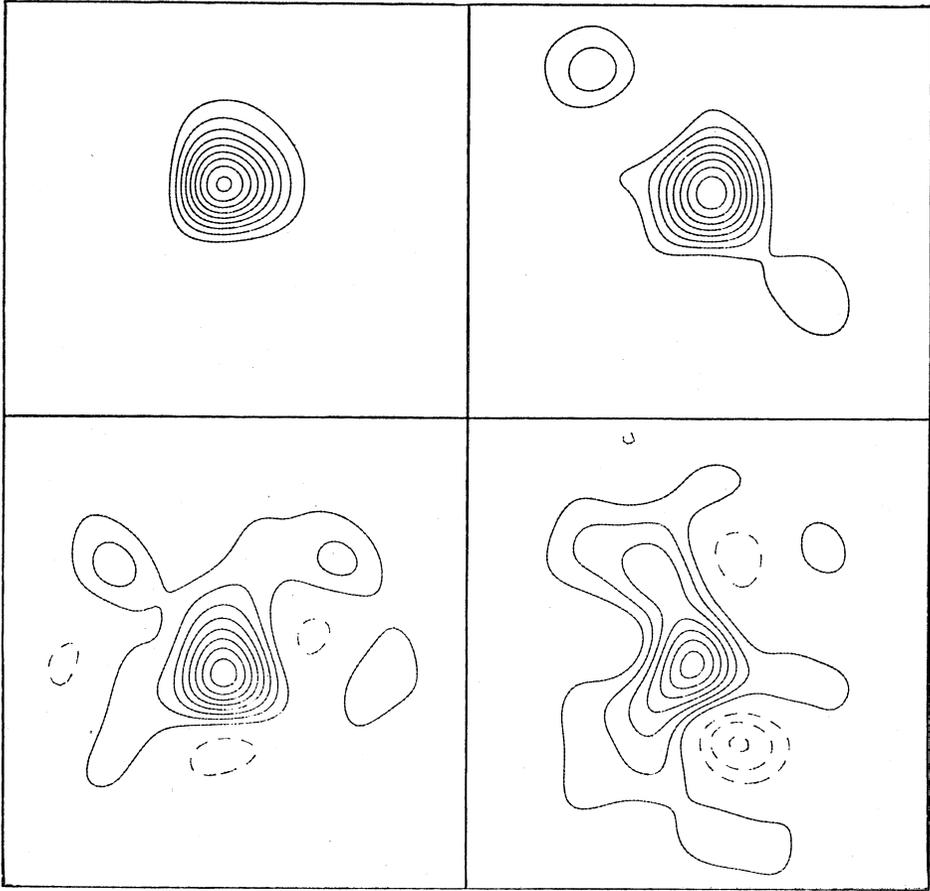


Fig. 6. Canterbury algorithm reconstructions of the corresponding images in Fig. 5. Note that the algorithm fails to reconstruct the more complex images.

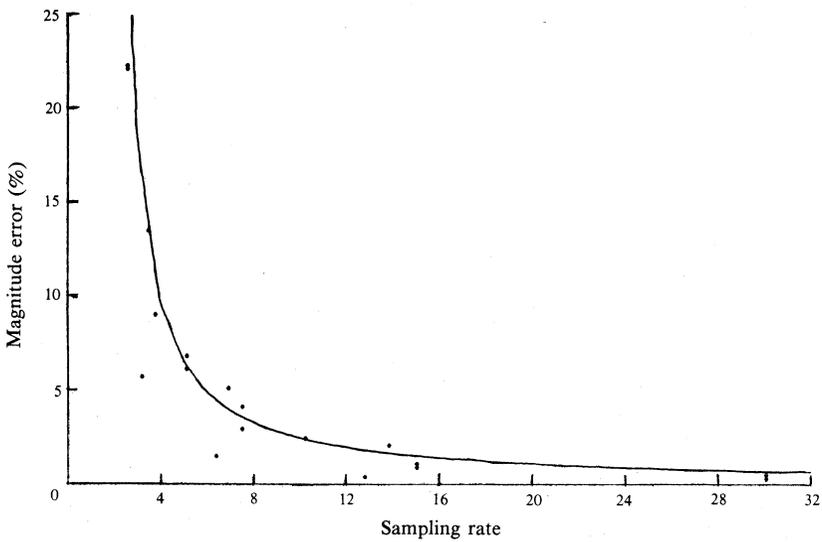


Fig. 7. Variation of magnitude error with sampling rate. The curve follows a best-fit line through the points.

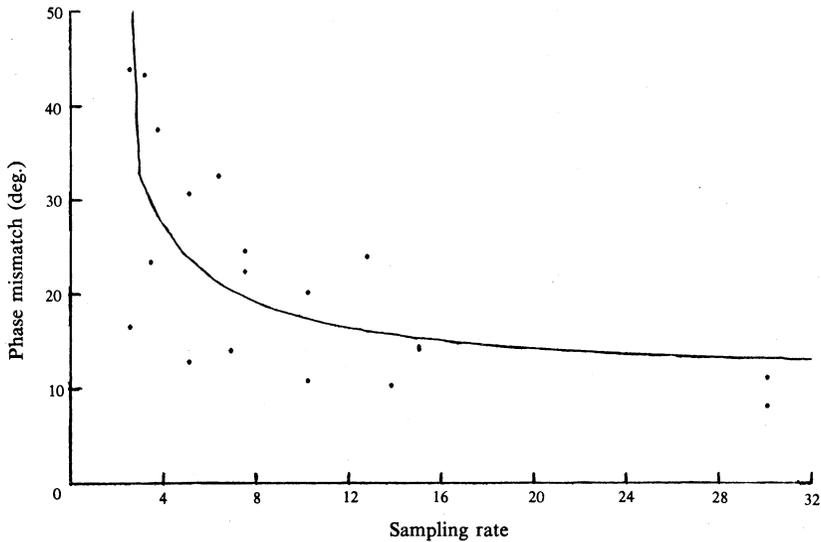


Fig. 8. Variation of average phase mismatch error with sampling rate. The curve follows a best-fit line through the points.

(c) Effect of Increasing the Sampling Rate

Figs 7 and 8 show plots of the magnitude error and phase mismatch error respectively for various sampling rates, for several images. It is encouraging that different images display more or less the same error curves. It is also encouraging that these error measures drop off as the sampling rate increases. This reduction in error might suggest an improvement in the estimation process. However, the reconstructed images did not reflect this improvement. Fig. 9 gives the original (see Fig. 3) and three reconstructions (at different sampling rates) of one image. The reconstructions are not particularly good. Examination of these suggests that the image with the lowest sampling rate is the best reconstruction. Indeed, it was found that the best reconstructions are never those carried out at the highest sampling rates.

(d) Phase Difference Cosine Estimation

The estimates of the phase difference cosine, produced by equation (11), will not, in general, be confined to the range $[-1, 1]$ (which of course it should be!). These bad cosine estimates then have to be clipped to this range. Typically the number outside the range is about 30–50% for an interpolation weight of 0.5, with the number increasing towards 100% for small weights, or decreasing towards 0% for large weights. Fig. 10 gives a typical plot of the bad cosines against the interpolation weight. Fig. 11 gives histograms of the estimated cosines, for the same image, for various sampling rates.

We note that the number of bad cosine estimates greater than 1 is far larger than those less than -1 . This is understandable as close sampling results in phase differences close to zero, and consequently cosines close to 1. We note also that while the percentage of bad cosine estimates was not found to decrease appreciably with an increase in the sampling rate, the histograms do improve (i.e. the cosine estimates cluster more in the range $[-1, 1]$).

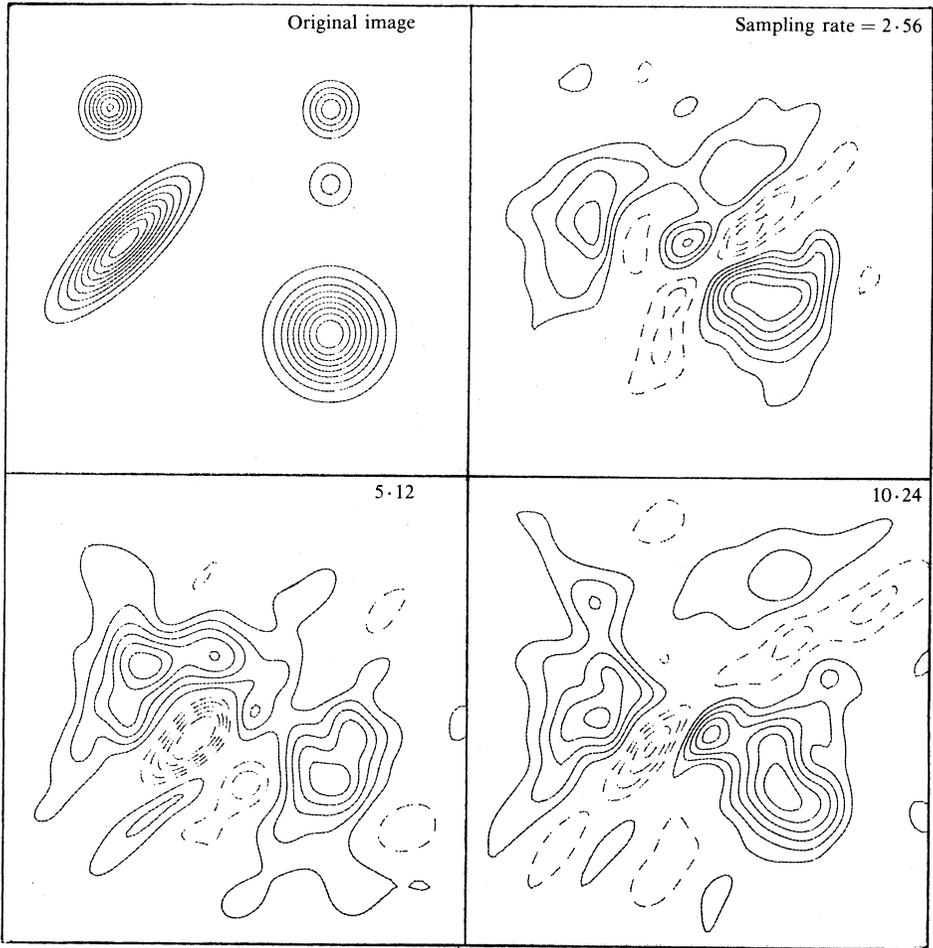


Fig. 9. Original image and three Canterbury algorithm reconstructions of Fig. 3 at different sampling rates.

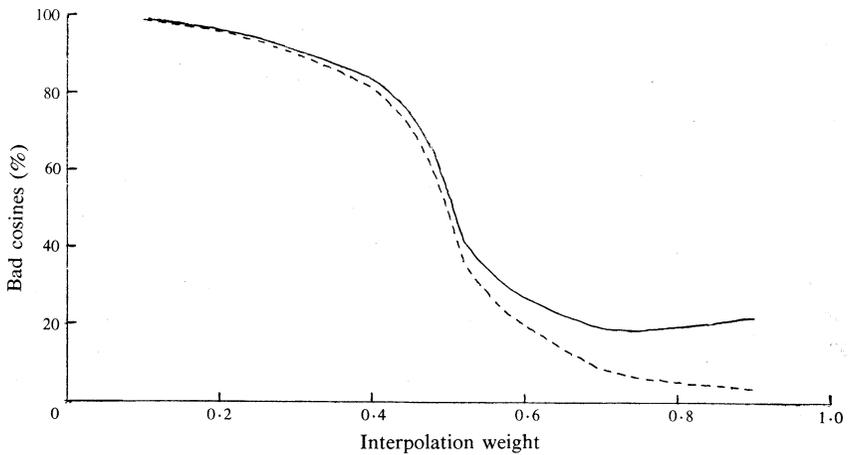


Fig. 10. Variation of cosine estimates outside $[-1, 1]$ with interpolation weight. The number of bad cosines is normalized by the total number of cosines estimated. The dashed curve represents cosines exceeding 1, while the solid curve represents the total outside the range $[-1, 1]$.

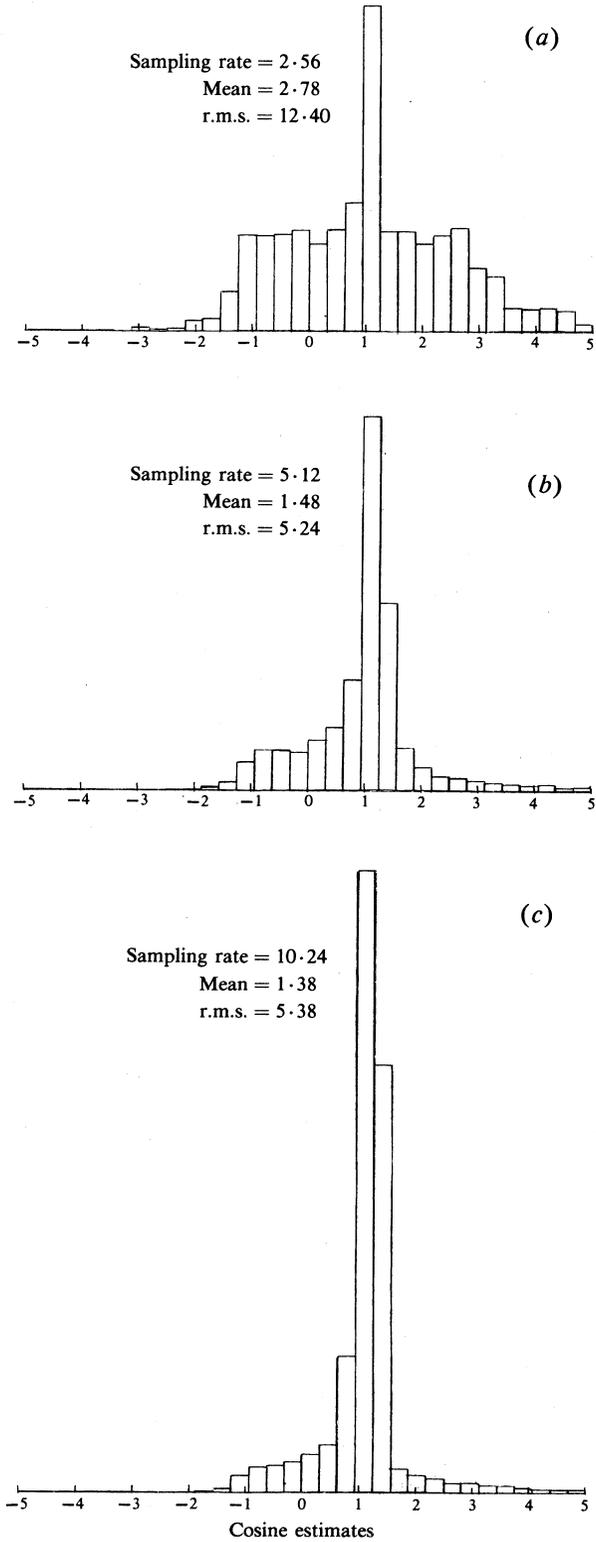


Fig. 11. Histograms of cosine estimates at different sampling rates. The estimates cluster more towards the limits $[-1, 1]$ as the sampling rate is increased.

This problem can be partially attributed to near zeros in the visibility magnitude, which cause a near zero divide in the cosine equation. It was found that there was a greater concentration of bad cosine estimates at higher visibility frequencies, where the visibility is usually close to zero. Indeed most cosine estimates outside the range $[-5, 5]$ belonged to high-frequency visibilities.

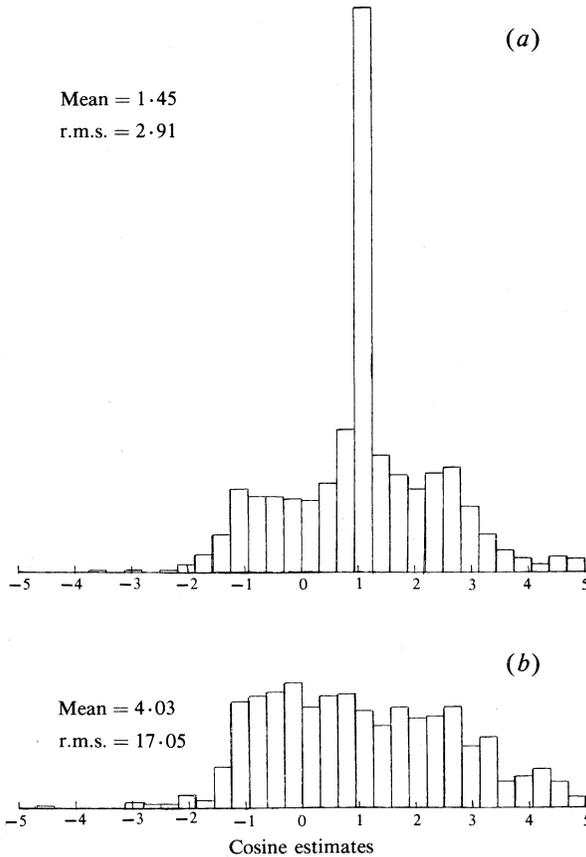


Fig. 12. Histograms of cosine estimates, for different parts of the $u-v$ plane, for the same reconstruction as in Fig. 11a: (a) low frequencies; (b) high frequencies. The low-frequency histogram is significantly better.

To examine the effect further, the $u-v$ plane was divided into two equal regions, one containing low-visibility frequencies, and the other containing high frequencies. Fig. 12 shows the histograms of the cosine estimates in these regions. The histogram of the cosine estimates at low frequencies (Fig. 12a) is significantly better. After a delta function was added to the image, the histograms improved significantly, but nevertheless there was no improvement in the quality of the reconstruction.

Fig. 13 gives density plots of (a) the estimated and (b) the actual phase difference cosines. It can be seen that the gross structure is correct, but that the details are not correct. Also there are large regions, at high frequencies, where the estimate is completely wrong.

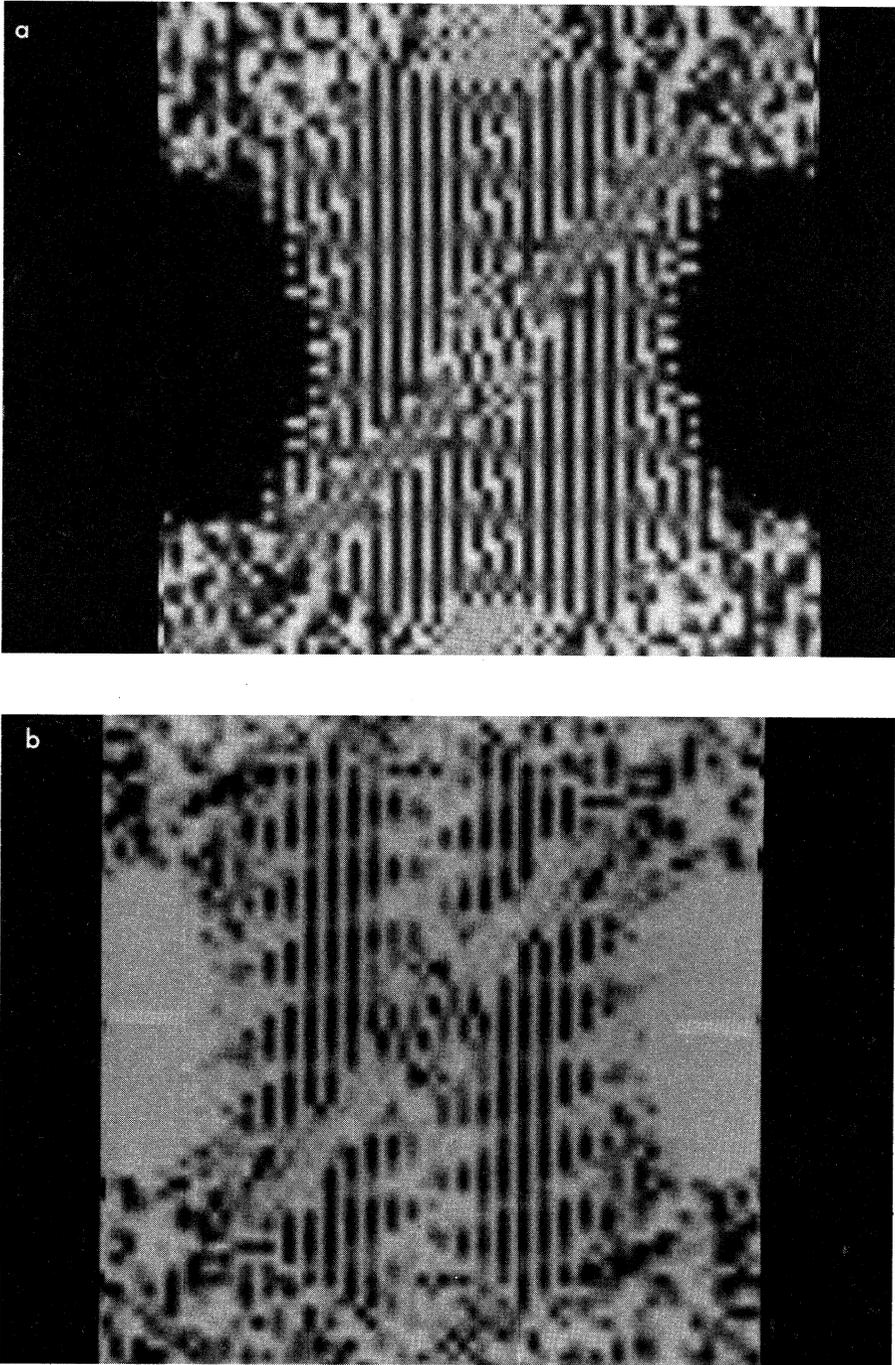


Fig. 13. Density plots of the phase difference cosines with the origin of the $u-v$ plane in the centre: (a) Canterbury algorithm estimate; (b) original image. The stips etc. at low frequencies have been reasonably estimated, but most high-frequency estimates are poor.

(e) *Phase Difference Estimation*

In the estimation of the phase difference from the cosine function, it should be noted that the inverse cosine function has infinite slope at -1 and 1 . This results in phase differences near 0 and π being very sensitive to errors in the cosine estimate.

One problem with the technique of estimating signs is that it often leads to ambiguities, and near ambiguities (we recall the assumption that the sign of one phase difference, around a square path, is known and the others are chosen so as to minimize the phase mismatch). If the known phase difference is zero (or near zero), there are two possible solutions, which are the opposite (negative) of each other, with the same phase mismatch error. There is no clear method for choosing the signs in this case.

As a specific example of this problem, we consider images, which Bates (1982) calls 'quasi-localized', for which every second visibility point is a two-point interpolation of its neighbours. The Canterbury algorithm is expected to be (nearly) exact for these images. For some quasi-localized images, this was indeed found to be the case, but for others (even those without strict ambiguities) the algorithm produced reconstructions which were appreciably in error. The latter images were found to have some phase differences of the order of 0.1° . The computer rounding error (23 bit mantissa used) and the sensitivity of the inverse cosine function combined to result in a wrong sign being chosen.

Finally, one would hope that the average phase mismatch error tends to zero as the sampling rate is increased. This was found to be the case, as can be seen by Fig. 8, which gives a plot of the average phase mismatch for several images for different rates of sampling. In addition it was found that the phase mismatch for high-frequency visibilities was much larger, which is consistent with the pooriness of the cosine estimates at high frequencies.

(f) *Estimation of Phase*

If an incorrect sign is chosen for a phase difference, then clearly this error will propagate outwards into the $u-v$ plane. This results in the phase estimate becoming inaccurate a number of sampling points away from the starting point. This effect appears to mitigate any improvement that one might gain by increasing the sampling rate. This is to say that, although the cosine estimates are better at higher sampling rates and that the phase mismatch decreases, the manner of estimating the phase tends to cause an avalanche of errors, so that higher sampling rates do not give better results.

An interesting class of images are those for which the phase is always 0 or π (called centrosymmetric images because of their symmetry). As there is no problem in choosing the signs of the phase differences for these images, they should be expected to give better results. Fright and Bates (1982) have shown how to include this prior knowledge into the Canterbury algorithm, and it was found that the reconstructions were much improved. In addition the reconstructions were not found to degrade significantly as the sampling rate increased (as it did for most images). Fig. 14 gives an example of a reconstruction of a centrosymmetric image at various sampling rates.

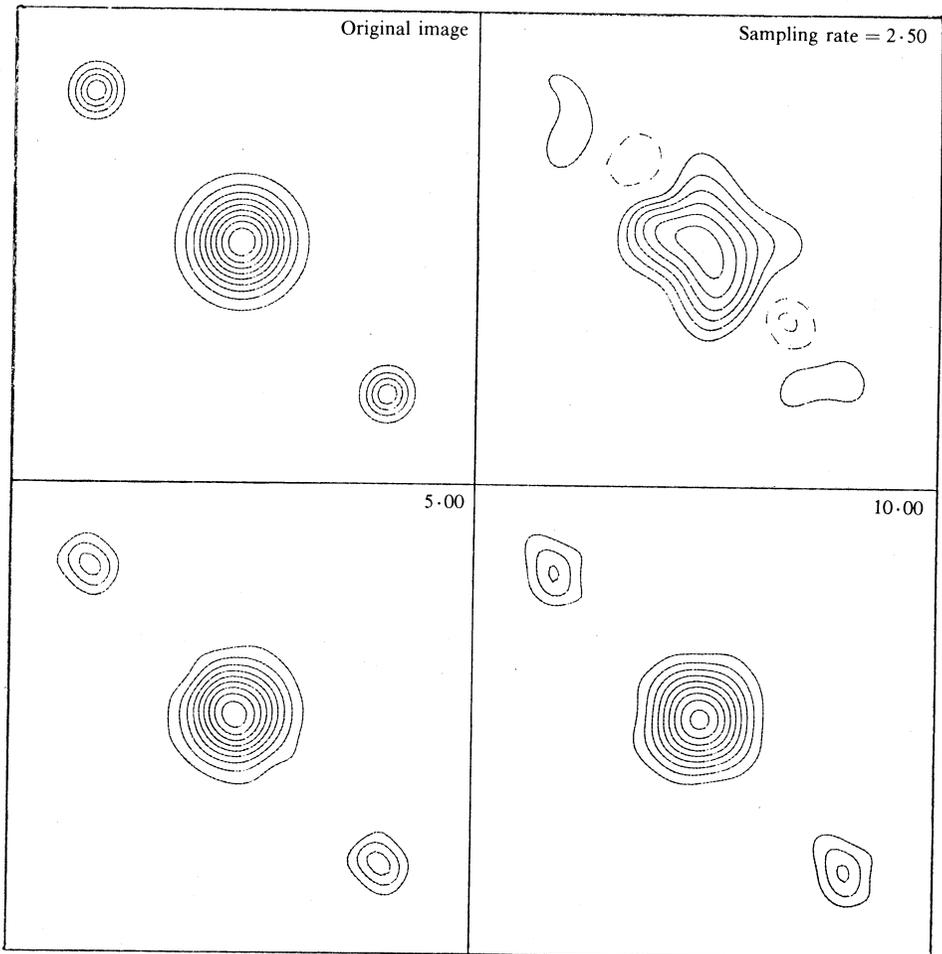


Fig. 14. Original centrosymmetric image and three Canterbury algorithm reconstructions using different sampling rates. The reconstructions do not degrade significantly as the sampling rate increases.

6. Canterbury Algorithm—Conclusions

In summary, the Canterbury algorithm fails to reconstruct anything but the simplest images. The algorithm also fails to estimate the phase differences accurately at high spatial frequencies. The major shortcoming of the algorithm is in the estimation of the sign of the phase difference and then combining the phase differences to estimate the phase. If this step could be performed more accurately, the algorithm should produce significantly better results, particularly as the sampling rate is increased.

7. Fienup's Iterative Procedures—Theory

An entirely different approach to the phase problem has been taken by Fienup (1978, 1982). His approach is to repeatedly Fourier transform an approximation to the object between the image and visibility domains. In each domain the approximation is modified to make it (or make it tend to be) more consistent with

the known image constraints (positivity and compactness) and measured data. There are several variants to his algorithm, but the two most successful are the so-called error-reduction algorithm and the hybrid input/output algorithm. If $g_i(m, n)$ and $G_i(k, l)$ are the approximations of the object's image and visibility at the i th iteration, then the algorithms can be expressed as follows. For the error-reduction algorithm we have

$$g'(m, n) = \mathcal{F}^{-1}\{G'(k, l)\}, \quad (14a)$$

$$\begin{aligned} g_i(m, n) &= g'(m, n), & (m, n) \in \gamma \\ &= 0, & (m, n) \notin \gamma, \end{aligned} \quad (14b)$$

$$G_{i+1}(k, l) = \mathcal{F}\{g_i(m, n)\}, \quad (14c)$$

$$G'(k, l) = G_i(k, l) |F(k, l)| / |G_{i+1}(k, l)|. \quad (14d)$$

For the hybrid input/output algorithm we have

$$g'(m, n) = \mathcal{F}\{G'(k, l)\}, \quad (15a)$$

$$\begin{aligned} g_i(m, n) &= g'(m, n), & (m, n) \in \gamma \\ &= g_{i-1}(m, n) - \beta g'(m, n), & (m, n) \notin \gamma, \end{aligned} \quad (15b)$$

$$G_{i+1}(k, l) = \mathcal{F}\{g_i(m, n)\}, \quad (15c)$$

$$G'(k, l) = G_{i+1}(k, l) |F(k, l)| / |G_{i+1}(k, l)|. \quad (15d)$$

Here γ is the set of points where the image constraints are violated (i.e. the approximation is negative or nonzero outside the known image extent) and β is a step-size parameter, typically from 0.1 to 1 (an image-dependent parameter).

The error-reduction algorithm is closely related to the algorithm used by Gerchberg and Saxton (1972) for the phase retrieval problem in electron microscopy, for which both the image and visibility are complex valued, and magnitude data are available in both domains.

We define the visibility magnitude error ε_v as for the Canterbury algorithm (equation 13), and the image constraint error (i.e. the amount that the image constraints are violated) as

$$\varepsilon_0 = \left(\sum_{(m,n) \in \gamma} \{g(m, n)\}^2 \right)^{\frac{1}{2}}. \quad (16)$$

It can be shown that for the error-reduction algorithm (if infinite precision arithmetic is used), the magnitude and image constraint errors form the decreasing sequence

$$\varepsilon_{v,i+1} \leq \varepsilon_{0,i+1} \leq \varepsilon_{v,i} \leq \varepsilon_{0,i}. \quad (17)$$

Although similar, in some respects, to the non-expansive mapping reconstruction algorithms (Schafer *et al.* 1981; Tom *et al.* 1981), the error-reduction mappings are generally expansive. The proof that the error decreases does not even assume, or prove, that a solution exists. It is possible that the error-reduction algorithm will stagnate, with the error remaining constant. Indeed, if the initial image estimate is centrosymmetric, or if no solution exists, then the error-reduction algorithm must stagnate.

In practice the error-reduction algorithm tends to be intolerably slow after the first few iterations. In addition it was found that computer rounding errors can cause the error to increase slightly, thus preventing convergence. This slowness caused Fienup to attempt other techniques, such as the hybrid input/output algorithm.

Fienup has found that the best strategy for reconstruction is to alternate between 5–10 iterations of the error-reduction technique and 10–30 iterations of the hybrid input/output technique. This is repeated until convergence (if it converges!). The image is tidied up, when it has converged, with a few extra iterations of the error-reduction algorithm.

Fienup *et al.* (1982) have determined a reliable method for estimating the size (extent) of the image from its autocorrelation function (the size of the image is simply half the size of the autocorrelation function). Fienup (1982) recommended using a random image, which satisfies the positivity/compactness constraints, as the initial estimate.

8. Fienup's Iterative Procedures—Computational Aspects

Following the methods of Fienup closely, we performed phase reconstructions on the same images as those used in investigating the Canterbury algorithm. The results and rate of convergence using a random initial image were similar to those reported by Fienup, and will not be repeated here. Although the procedures did not always converge, the resulting reconstructions were always found to be better than those of the Canterbury algorithm. The procedures were found to perform poorly when the visibility had a very large dynamic range.

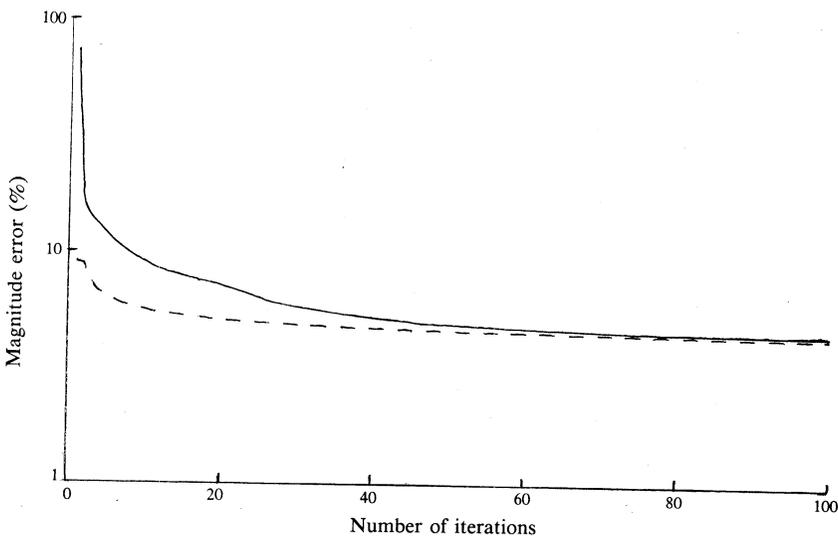


Fig. 15. Magnitude error (log scale) for various iterations of the error-reduction algorithm. The initial image estimate was either the Canterbury algorithm reconstruction (dashed curve) or a random image (solid curve). Both approach an error of about 4%.

A variant of Fienup's procedures, suggested by Bates and Fright (1983), is to start the iterations with the Canterbury algorithm reconstruction of the image, rather than a random image. However, this was not found to improve convergence or

the quality of the reconstruction significantly. In some cases starting with the Canterbury image was marginally better, in other cases it was marginally worse. A typical example of this is Fig. 15, which shows the magnitude errors using the two methods of starting Fienup's procedures. It can be seen that, although initially the magnitude error is much smaller when starting with the Canterbury reconstruction (dashed curve), the errors (when using the random initial image) reduce quickly to about the same level.

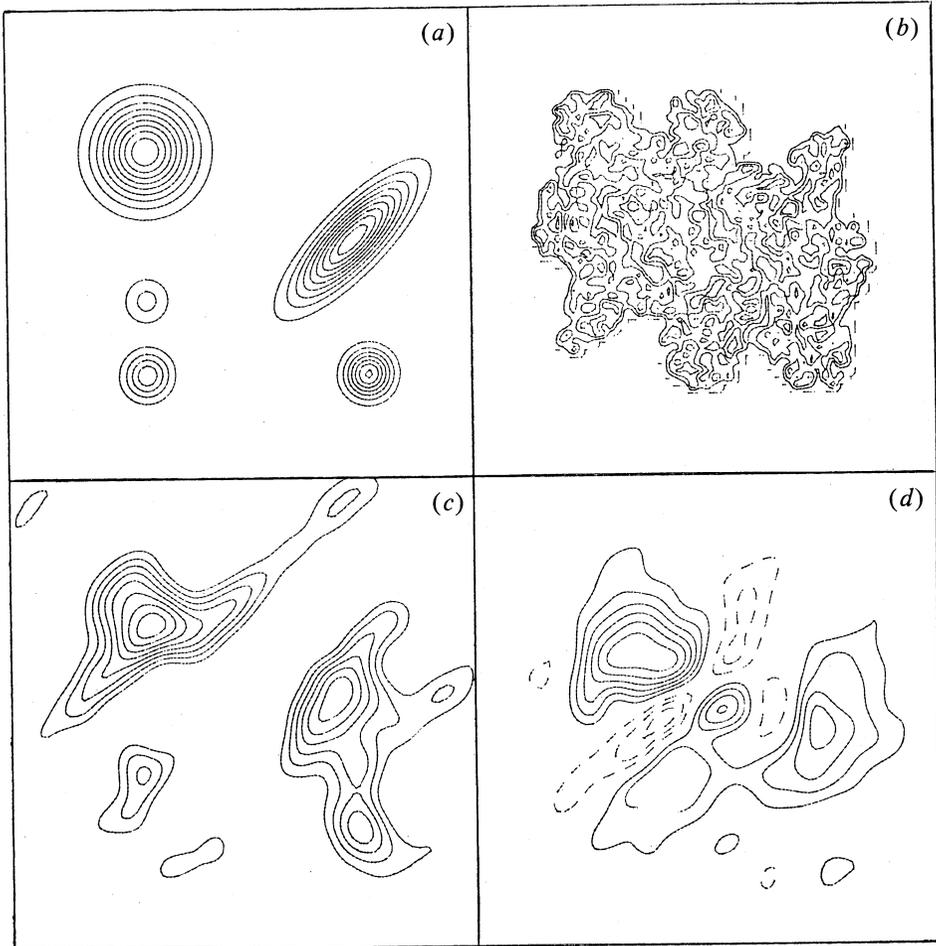


Fig. 16. Original image (a) and three reconstructions: (b) the random initial image of Fienup's procedures; (c) the random image after 10 iterations; (d) the Canterbury algorithm. The image of Fig. 16c produces a better reconstruction than the Canterbury algorithm.

Another example is shown in Fig. 16 of an original image and three reconstructions. The random image after 10 iterations (Fig. 16c) is a better reconstruction than that for the Canterbury algorithm (Fig. 16d).

9. Conclusions

Two phase restoration techniques, the Canterbury algorithm and Fienup's procedures, have been investigated. It was found that the Canterbury algorithm

produces poor results for all but the simplest of images. The major inaccuracy in the algorithm was found to be the estimation of the phase from the phase difference cosines. There is scope for future investigation to determine whether this step in the algorithm can be refined.

Results with Fienup's procedures indicate that in general there seems to be no advantage in using the Canterbury algorithm to produce an initial estimate of the image. A random initial image produces equally good results.

References

- Bates, R. H. T. (1982). *Optik (Stuttgart)* **61**, 247.
- Bates, R. H. T., and Fright, W. R. (1983). *J. Opt. Soc. Am.* **73**, 358.
- Bracewell, R. N. (1978). 'The Fourier Transform and Its Application' (McGraw-Hill: New York).
- Bruck, Y. M., and Sodin, L. G. (1979). *Opt. Commun.* **30**, 304.
- Fiddy, M. A., Brames, B. J., and Dainty, J. C. (1983). Enforcing irreducibility for phase retrieval in two dimensions. *Opt. Lett.* **8**, 96.
- Fienup, J. R. (1978). *Opt. Lett.* **3**, 27.
- Fienup, J. R. (1982). *Appl. Opt.* **21**, 2758.
- Fienup, J. R., Crimmins, T. R., and Holsztynski, W. (1982). *J. Opt. Soc. Am.* **72**, 610.
- Fright, W. R., and Bates, R. H. T. (1982). *Optik (Stuttgart)* **62**, 219.
- Garden, K. L., and Bates, R. H. T. (1982). *Optik (Stuttgart)* **62**, 131.
- Gerchberg, R. W., and Saxton, W. O. (1972). *Optik (Stuttgart)* **35**, 237.
- Gull, S. F., and Daniell, G. J. (1978). *Nature* **272**, 686.
- Hayes, M. H., and McClellan, J. H. (1982). *Proc. IEEE* **70**, 197.
- Hunt, B. R. (1979). *J. Opt. Soc. Am.* **69**, 393.
- Papoulis, A. (1965). 'Probability, Random Variables and Stochastic Processes' (McGraw-Hill: New York).
- Ramachandran, E. N., and Srinivasan, R. (1970). 'Fourier Methods in Crystallography' (Wiley: New York).
- Schafer, R. W., Mersereau, R. M., and Richards, M. A. (1981). *Proc. IEEE* **69**, 432.
- Tom, V. M., Quateri, T. F., Hayes, M. H., and McClellan, J. H. (1981). *IEEE Trans. Acoust. Speech Signal Process.* **29**, 1052.

Manuscript received 1 July 1983, accepted 26 January 1984

