UNIVERSITY OF ASTON IN BIRMINGHAM

Department of Physics

MINISTRY OF TECHNOLOGY

Royal Aircraft Establishment

THE USE OF COMPLEX SPATIAL FREQUENCY FILTERS IN CORRELATION PROCESSES

by

Gordon Elliott Keyte

Thesis submitted for the Ph.D. degree

1970

Abstract

A method of solving pattern recognition problems is to correlate the unknown data with the known pattern. Where the data is available in two-dimensional transparency form it is convenient to use optical systems; it can be shown that a coherent optical system, using a complex filter corresponding to the pattern being sought, can perform the required correlation operation.

The method of setting up the coherent optical system and of constructing the complex filter are described. The limitations of the recording material used in making the filter are discussed and results are presented showing the effect of these limitations on the output of the system.

Problems concerned with the practical use of the optical correlator are discussed, including the effect of optical aberrations and difficulties in using photographic input data. A method of overcoming some of the recording limitations is also presented.

Finally, some applications are described in which complex spatial filters have been used. The merits and disadvantages in this approach to pattern recognition are discussed and some suggestions for future applications are given.

TABLE OF CONTENTS

Page

Page

CHAPTER 6 THE OPTICAL CORRELATOR

CHAPTER 7 APPLICATIONS AND CONCLUSION

Page

CHAPTER 7 - continued

Chapter I

Correlation processes and pattern recognition

1.1 Pattern Recognition

The task of this chapter is to discuss what is meant by correlation processes and to show how such processes may be used to search for a particular signal or pattern.

In general, we are concerned with the recognition of a known signal or pattern, which may or may not be present in a field of interest. There are a number of examples:-

Recognition of targets in aerial photographs (and other specialised photographs).

Recognition of radar signals.

We may also extend this idea to the search for a particular pattern from a set of patterns. The set may be displayed sequentially or in the form of an array, in which case the pattern must not only be recognised, but also located. Examples here are character recognition and fingerprint recognition.

In some cases it may be desirable to compare one pattern closely with another, e.g. where a pattern must conform to within a specified tolerance to a master pattern.

In most cases one finds that the field of interest can contain other information as well as the pattern being sought. One frequently finds that this information exists whether or not the pattern is present, and the problem of pattern recognition is to consider whether the pattern is present in a background of miscellaneous information which tends to obscure the pattern. This information is often of no interest and is therefore regarded as noise by analogy with communications practice. In the case of character recognition, where one has a sequence of patterns, then all patterns other than the one being sought can be regarded as noise.

As an illustration of this idea one can consider the use of camouflage in nature and in war. Here a well known pattern (such as an animal, or ship) is camouflaged or painted in such a way as to make its' outlines blend into the background. It is as if one had deliberately added 'noise' (or redundant information) to the pattern so that the task of pattern recognition is made more difficult.

In some applications the task of pattern recognition is complicated by fact that some tolerance has to be allowed on the shape or size of the pattern. For example, in character recognition, the letter 'A' may assume a variety of actual shapes, but must still be identified as being letter A.

It is possible to allow a small tolerance on the pattern recognition process, such that small variations of the pattern being sought will be accepted. However, the introduction of such a tolerance has the disadvantage that other patterns may be mistaken for that being sought, giving false detections.

Thus, in addition to the basic problem of pattern recognition in a noisy background, one has to consider whether there is ambiguity between the pattern being sought and any similar pattern.

The problem may be expressed in mathematical form. Since we are concerned here with two dimensional patterns, any pattern may be described by a function dependent on two independent variables, e.g. $s(x, y)$.

If the field of interest, or input data is denoted as $f(x, y)$ we have to decide whether

$$f(x, y) = s(x, y) + n(x, y)$$

or

$$f(x, y) = n(x, y)$$

where $s(x, y)$ is the pattern or signal being sought, and $n(x, y)$ represents redundant information. Any of the functions can be complex.

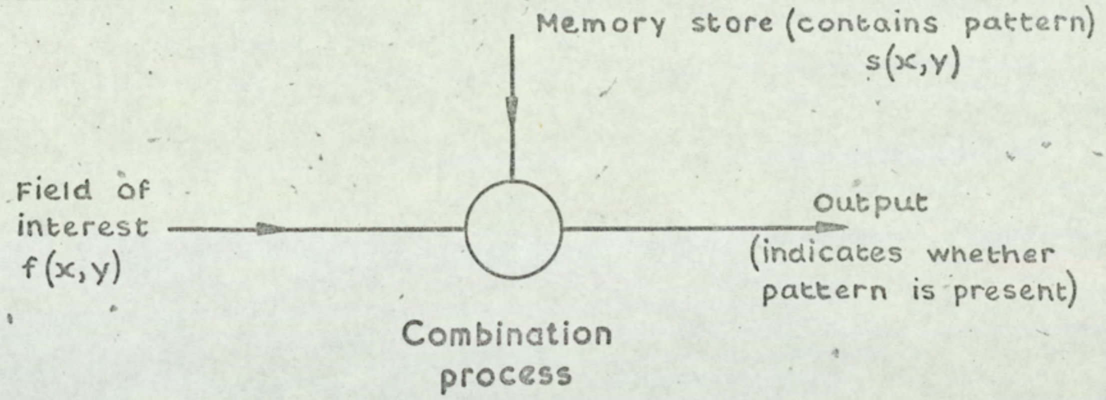
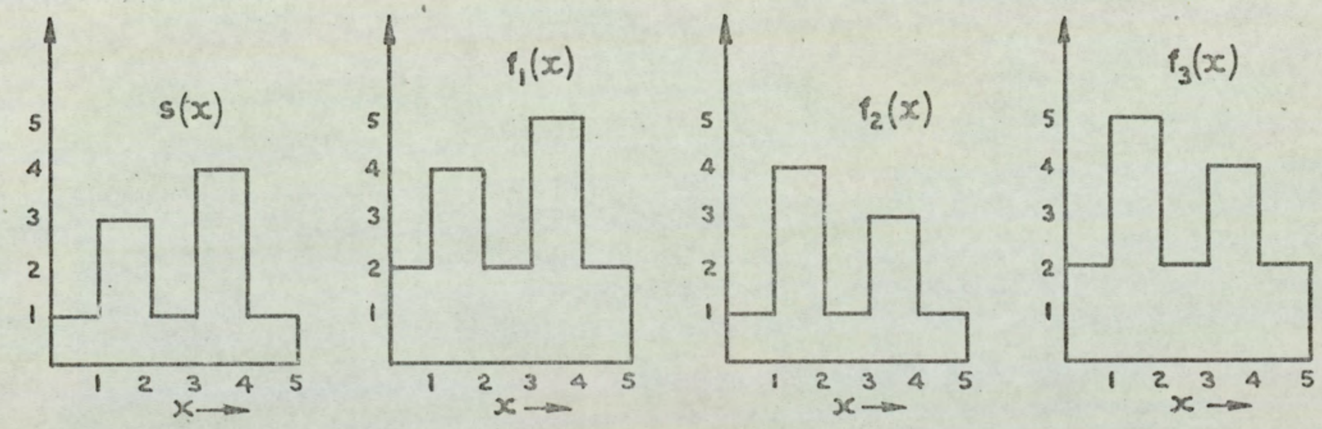


Fig. I. 1



$R[s(x).\ s(x)] = 5{\cdot}6$	$R[s(x).\ f_1(x)] = 7{\cdot}6$	$R[s(x).\ f_2(x)] = 5{\cdot}4$	$R[s(x).\ f_3(x)] = 7{\cdot}4$
$p(s(x).\ s(x)) = 1{\cdot}0$	$p(s(x).\ f_1(x)) = 1{\cdot}0$	$p(s(x).\ f_2(x)) = 0{\cdot}875$	$p(s(x).\ f_3(x)) = 0{\cdot}875$
autocorrelation	cross correlation	cross correlation	cross correlation

Fig. I. 2

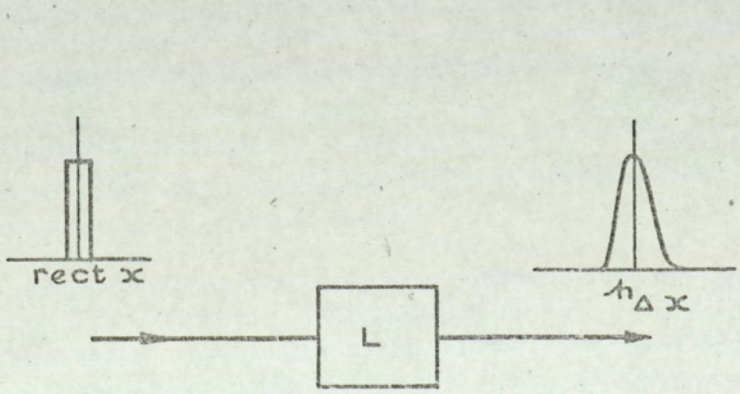


Fig. I.3

Since we shall be concerned with optical pattern recognition methods, it will be convenient here to show that both signal patterns and input data will often be in the form of transparencies, where the function is represented by a variation in density across the (x, y) plane. This information is transmitted through the system by light which carries the information in terms of amplitude and phase modulation. In the case of density variations in the transparency, there will be a corresponding variation in the intensity of the transmitted light. There will also be cases where transparent objects, having a variation in refractive index or thickness may be used. Here the transmitted light carries the information in terms of its phase modulation.

In general the signal or input function $f(x, y)$ will be complex and may be represented by an amplitude and phase component, i.e. $A(x, y) e^{-j\phi(x, y)}$.

1.2 Correlation - Statistical solution to pattern recognition

The nature of pattern recognition problems implies that some form of memory is required in which the pattern $s(x, y)$ to be sought may be stored. The operation of pattern recognition will then involve some process of combination of the memory with the input from the field of interest, as shown in Fig.1.1.

The output would be of a form which would indicate whether the input contained the signal or pattern being searched, ie, it need simply be a binary Yes:No output. In some cases additional information is required, for example, location of pattern in field of interest. For the present, consider simply the Yes:No output.

To simplify the problem consider a known one dimensional signal $s(x)$ and it is required to know whether such a signal is contained in an input $f(y)$. Although x and y have the same dimensions they do not necessarily at this stage refer to the same coordinates.

If the input $f(y)$ is in fact equal to $s(x)$ there will be an exact correspondence between all the ordinates; we say that the functions are correlated. On the other hand, if there is no connection between values of ordinates in $f(y)$ and those in $s(x)$ then there is no correlation. The pattern recognition problem is essentially one of finding the correlation between $s(x)$ and $f(y)$.

This may be done statistically by finding the correlation coefficient ρ given by[1] :-

$$\rho = \frac{E[\{s(x) - m_{s(x)}\}\{f(y) - m_{f(y)}\}]}{\sigma_{s(x)}\,\sigma_{f(y)}}$$

where $m_{s(x)}$ and $m_{f(y)}$ are the means of the two functions, $\sigma_{s(x)}$, $\sigma_{f(y)}$ are the standard deviations of the functions and $E[\ldots]$ evaluates the average of the product inside brackets.

In statistical terms $E[\ldots.]$ evaluates the joint moments of the functions and is called the covariance of $s(x)$ and $f(y)$. It tells us how much alike $s(x)$ and $f(y)$ are. The deviations $\sigma_{s(x)}$, $\sigma_{f(y)}$ are used to normalise ρ so that it equals unity when the functions are exactly alike. The coefficient can be zero, indicating no correlation at all, and also -1, indicating negative correlation, i.e. between functions of equal shape but opposite sign.

Some modification to this expression can be made in the case of signal recognition (the following argument owes to ref. 1.) ρ can now be written as:-

$$\rho = \frac{E[s(x).f(y)] - m_{s(x)}.m_{f(y)}}{\sigma_{s(x)}\quad\sigma_{f(y)}}$$

where $E[s(x) \,.\, f(y)]$ is called the correlation function of $s(x)$ and $f(y)$.

If x, y are measured on the same system of coordinates, we will not be interested in their specific values, but rather in their difference. This difference can be regarded as a displacement of one pattern from another, and variation of the displacement will enable the patterns to be shifted over each other so that every possible position for correlation will be evaluated.

The function $f(y)$ can be replaced by $f(x-u)$ where u represents the shift.

$$\text{i.e.} \quad \rho_u = \frac{E[\, s(x) \cdot f(x-u)] - m_{f(x)} \cdot m_{s(x)}}{\sigma_{s(x)} \; \sigma_{f(x)}}$$

$$= \frac{R_u \; (s.f.) - m_{f(x)} \cdot m_{s(x)}}{\sigma_{s(x)} \; \sigma_{f(x)}}$$

where R_u (s.f.) is now the correlation function.

Here we have assumed that the function $f(x)$ is strictly stationary, i.e. that a shift of the origin has no effect on the probability distribution of $f(x)$.

The numerical value of R_u(s.f.) will vary as the shift 'u' is varied. The greater the value of R_u(s.f.), the greater the degree of correlation between the two functions. When the two functions are not equal, then the curve traced by R_u(s.f.) is called the cross correlation function; when the functions are identical the curve becomes the autocorrelation function.

In general, both curves will contain a maximum value of R_u(s.f.), although in some cases there may be several maxima, or peaks, in the curve. The maximum value of R_u(s.f.) for the autocorrelation indicates that the two functions overlap each other exactly and this can only be obtained when the functions are identical. One would not expect any of the cross correlation maxima to equal or exceed the value obtained at the autocorrelation maxima. There are circumstances when this can occur, however, as shown in Fig. 1.2.

Here the function $s(x)$ shown on the left of Fig. 1.2 is correlated first with itself giving the autocorrelation maxima and the autocorrelation coefficient $[\rho(s^2)]$. It is then correlated with each of the other functions shown, i.e. $f_1(x)$, $f_2(x)$ and $f_3(x)$.

In the case of $f_1(x)$, the shape of the function is the same as $s(x)$, but it's mean is greater. As a result, the cross correlation maxima is also higher, but calculation of the cross correlation coefficient, which includes the mean, gives the same result as for the autocorrelation.

For function $f_2(x)$, the shape is dissimilar to $s(x)$, although its mean is the same. Its correlation maxima and coefficient are both lower than that of the autocorrelation.

The cross correlation maximum of $f_3(x)$ is higher than that of the autocorrelation, even though the shape of the function is dissimilar to that of $s(x)$. Only when the means are included in the calculation, as for ρ, is the dissimilarity made evident, and a lower value than the autocorrelation is obtained.

If the denominator and mean terms are neglected, the correlation function is:

$$R_u \; [\, s(x)\; f(x-u) \,]$$

This may be written $\lim\limits_{A\to\infty} \frac{1}{2A} \int_{-A}^{A} s(x)\; f^*(x + u)\; dx$ under the condition that both functions are ergodic[1]. Here ergodic means that the statistical average of the function is also equal to the space average. The use of complex conjugate simply means that the resulting function will be real and not imaginary; it is necessary where complex functions are involved.

1.3 Linear Systems for correlation

We must now consider how the integral

$$\frac{1}{2A} \int_{-A}^{A} s(x)\; f^*\; (x + u)\; dx \quad \text{can be evaluated.}$$

Whatever process is used, there will be an input $f(x)$ and an output $g(x)$. If the input is made $f(x + b)$ then we require the output to be $g(x + b)$, i.e. the system used must be a fixed parameter system. This is clearly desirable since we shall operate over a large region, over which we require space invariance.

In addition, we require that if $f(x) = a_1 f_1(x) + a_2 f_2(x)$ then $g(x) = a_1 g_1(x) + a_2 g_2(x)$. This implies that a linear system is required.

The system must therefore be linear and space invariant.

Let the system be represented by an operator L, it is necessary to find out what form this operator takes.

Let the input to the system be a rect function rect x where

$$\text{rect}\,(x) = \frac{1}{\Delta x} \quad \text{for } |x| \leq \frac{\Delta x}{2}$$

$$= 0 \quad \text{otherwise}$$

and let the response be $h_{\Delta(x)}$

i.e. $L[\text{rect}\,(x)] = h_{\Delta(x)}$ as shown in fig 1.3.

Since the system is space invariant we can also put:-

$$L\ \text{rect}\,(x - n\Delta x)] = h_{\Delta(x - n\Delta x)}$$

Now any function $f(x)$ can be expressed as a summation of rectangular strips of width Δx.

i.e. $f(x) \approx \sum_{n=-N}^{N} f(n\,\Delta x)\ \text{rect}\,(x - n\,\Delta x)\,\Delta x$

Hence $g(x) \approx L[\sum_{-N}^{N} f(n\,\Delta x)\ \text{rect}\,(x - n\,\Delta x)\,\Delta x]$

$$\approx \sum_{-N}^{N} f(n\,\Delta x)\ h_{\Delta}\,(x - n\,\Delta x)\ \Delta x$$

Note that the justification for taking $f(n\,\Delta x)$ outside operator is that this function can be regarded as being constant over the interval Δx, whereas the rect function varies over this interval.

In the limit we have $g(x) = \int_{-\infty}^{\infty} f(x_o)\ h\ (x - x_o)\ dx_o$

where $n\,\Delta x = x_o$ = shift or displacement.

Thus we see that the output $g(x)$ may be expressed as a convolution of the input with the function $h(x)$

i.e. $g(x) = f(x) * h(x)$

Using fourier transform theory thus gives:

$$G(p) = F(p)\ .\ H(p)$$

where $G(p)$, $F(p)$ and $H(p)$ are all fourier transforms of $g(x)$, $f(x)$ and $h(x)$ respectively and measured in the frequency domain. Since we have been concerned with spatial dimensions (i.e. x, y) then the dimensions of p are those of spatial frequency.

There are two important conclusions here. One is that a linear system performs a convolution of the system with the input and the other is that the product of the transforms of input and system produce the output transform.

The system transform, because of its simplicity in use is frequently used in linear systems, and is called the transfer function. Its action is to weight different parts of the spatial frequency spectrum of the input, so that a modified spectrum is produced, which is inversely transformed to give the output.

Strictly speaking, it is immaterial whether in practice one performs the linear system operation by direct convolution or by multiplication of the transforms, since both will give the same result. However, in optics one finds that the fourier transform can be very easily formed, and it is generally more convenient to work with transforms. The function $H_{(p, q)}$ which is used to multiply the input function transform is called a filter, and in this particular context - a spatial frequency filter. Its impulse response when the input consists of a delta function, is simply $h(x, y)$.

Suppose we now have a linear system in which the filter transfer function is $H^*(p)$. [i.e. the complex conjugate of $H(p)$]

$$\text{The output } G(p) = F(p)\ H^*(p)$$

$$g(x) = \int f(u)\ h^*\ (x + u)\ du$$

where the impulse response of the filter $h^*\ (-u)$ has been used.

This result gives the correlation of $f(u)$ with $h^*(x)$ as required earlier, for pattern recognition. Provided that the linear system contains a filter of the form $h^*(-u)$ it can be used to perform correlation operations, and hence in pattern recognition.

1.4 Detection of signal in noise

The previous section has shown how a linear system with a filter $h^*(-u)$ can be used to perform correlation operations. Such a filter will give maximum response when correlated with its own signal - i.e. auto-correlation (assuming that means are equal). However, the cross correlation can easily have a value approximating to this maximum, as shown in Fig 1.2. The problem is to design a filter which will emphasize the signal at the expense of noise, i.e. to maximise the signal to noise ratio. Since in the output plane we shall be observing intensities or energies, the maximum possible ratio will be

$$\frac{\text{Total Signal Energy}}{\text{Mean Square Noise Energy}} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{|S(p)|^2}{Sn(p)}$$

where $\int_{-\infty}^{\infty} |S(p)|^2$ if the signal energy, and $\frac{1}{2\pi}\int_{-\infty}^{\infty} Sn(p)$ is the mean square noise energy.

This problem is one frequently encountered in electronic communication systems; the following derivation is from Brown[2].

Let us assume that the input is composed of both signal $s(x)$ and noise $n(x)$. Let the filter be $h(x)$.

Then the signal out of the system is $s(x) * h(x)$ and the noise out is $n(x) * h(x)$.

The fourier transform of the output signal is $S(p)\ H(p)$ and the signal out is therefore:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-jpx}\ H(p)\ S(p)\ dp$$

The value of x at which the peak of the signal correlation occurs is arbitary; we can put $x = 0$

$\therefore$ Signal out is $\frac{1}{2\pi}\int_{-\infty}^{\infty} H(p)\ S(p)\ dp$

The mean square noise out is $\frac{1}{2\pi}\int_{-\infty}^{\infty} Sn(p)\ |\ H(p)|^2\ dp$

[Note: one frequently knows only the power of the noise spectrum - i.e. spectral density, rather than the actual amplitude and phase. Hence it is usual to work in terms of the noise spectral density $S_n(p)$. The output noise spectral density is $S_g(p) = |H(p)|^2 S_n(p)$ and the average spectral density (or mean square noise) is $\frac{1}{2\pi}\int_{-\infty}^{\infty} S_g(p)\, dp = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_n(p)|H(p)|^2\, dp$.]

The signal to noise ratio is therefore $\dfrac{|(\frac{1}{2\pi})\int H(p)\,S(p)\,dp|^2}{(\frac{1}{2\pi})\int S_n(p)|H(p)|^2 dp}$.

and the problem is to find H(p) which maximises this ratio.

This is done using the Schwartz in equality.

The numerator is modified by putting $\sqrt{S_n}(p) / \sqrt{S_n}(p)$ into the integral and assuming $S_n(p) > 0$.

$$\text{i.e. } \frac{S}{N} = \frac{|(\frac{1}{2\pi})\int \frac{S(p)}{\sqrt{S_n}(p)} H(p)\sqrt{S_n}(p)\, dp\,|^2}{(\frac{1}{2\pi})\int S_n(p)\,|H(p)|^2\, dp}$$

The Schwartz inequality states that;

$$|\int \alpha(x)\,\beta^*(x)\,dx\,|^2 \leq \int |\alpha(x)|^2\,dx \int |\beta(x)|^2\,dx$$

with equality when $\alpha(x) = (\text{const.}\ \beta(x))$

The numerator becomes $(\frac{1}{2\pi})^2 (\int |\frac{S}{S_n}|^2\ dp)\ (|H|^2 S_m\, dp)$

where $\alpha = S(p) / \sqrt{S_n}(p)$ and $\beta^* = H(p)\sqrt{S_n}(p)$

and $\frac{S}{N} = \frac{1}{2\pi}\int |\frac{S(p)}{S_n(p)}|^2\ dp$

when $\frac{S(p)}{\sqrt{S_n}(p)} = (H(p)\sqrt{S_n}(p))^*$

or $H(p) = \frac{S^*(p)}{S_n(p)}$

Thus the form of H(p) which gives maximum response is equal to the complex conjugate divided by the mean square noise. The numerator is to be expected, it was shown to be necessary for correlation with a known signal.

The action of the denominator is to weight those frequencies where the noise power is lowest; where the noise power is high, both signal and noise are attenuated.

In some cases the noise spectrum is uniform and $S_n(p)$ is constant - i.e. "white" noise.

In such cases similar argument shows that the signal/noise ratio is

$$\frac{1}{N}\left(\frac{1}{2\pi}\int |S(p)|^2 \, dp\right) \quad \text{when} \quad H(p) = k\, S^*(p) \quad (k = \text{constant})$$

Note that this form of H(p) gives the maximum possible signal/noise ratio since by Parsenals theorem

$$\frac{1}{2\pi}\int |S(p)|^2 \, dp = \int |f(x)|^2 \, dx = \text{total signal energy.}$$

The dimensions of the ratio are equal (energy units) but numerator is total power, whereas denominator is average power.

Since $H(p) = kS^*(p)$ the signal output becomes

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-jpx}\, kS^*(p)\, S(p)\, dp$$

$$= \frac{k}{2\pi}\int_{-\infty}^{\infty} s(u - x)\, s^*(-x)\, dx \quad \text{(by convolution theorem)}$$

Putting $-x = x$

$$= \frac{k}{2\pi}\int_{-\infty}^{\infty} s(x + u)\, s^*(x)\, dx \sim \text{the required autocorrelation function.}$$

1.5 Summary

The problem of pattern recognition is essentially one of measuring correlation. Although this may be done using a linear system, such a system performs a convolution, in which the input is multiplied by an inverted function. The convolution operation can be used in correlation processes provided that a correction is made for the inverted function; this is implemented by using the complex conjugate of the function concerned. The recording of the complex conjugate of the signal or pattern to be sought is called a matched filter; the convolution of the linear system enables the signal in such a filter to be correlated with the input.

The response to an input containing redundant information can be improved by making the filter of the form:-

$$H(p) = \frac{S^*(p)}{S_n(p)}$$

where $S_n(p)$ is the mean square noise or redundant information in the input.

This implies foreknowledge of input data, but it is reasonable to assume that the noise distribution will be fairly constant in any given application. Thus $S_n(p)$ would not change appreciably within a wide range of inputs.

In some cases, such as character recognition, the entire range of redundant patterns may be known in advance, and in theory, it ought to be possible for the matched filter to reject all characters other than the one to which it is matched.

The function $\frac{1}{S_n}(p)$ will include zero spatial frequency and therefore the d.c. terms of both signal and noise will be attenuated. This is equivalent to subtracting the means of the input and signal functions as required in the statistical correlation expression $[R_u(s(x)\ f(x)) - m_s m_f]$. It will be shown in later chapters how this subtraction occurs in practice.

The matched filter formula appears to be the optimum solution for recognising signals in a noisy background. It is not the only solution however, since there are other filters which can be used on correlators, but these do not give the optimum response.

In particular, it is possible for a multiplexed filter to be envisaged, which will correlate with a number of functions. Such a filter would be of use in character recognition and is discussed later.

We must now consider how optical systems can be used as a linear system and how a matched filter (or any complex spatial frequency filter) can be made.

<u>References</u>

1. "Random Signals and Noise" Darenport and Root (McGraw-Hill).

2. "Analysis of Linear Time Invariant Systems" W. M. Brown (McGraw-Hill).

Chapter 2

Coherent Optical Transforming Systems

2.1 Introduction

This chapter shows how optical systems may be used to carry out the convolution processes mentioned in Chapter 1.

It has already been stated that the convolution integral may be expressed in the spatial domain

$$\text{i.e.} \int f(x)\, h\,(u - x)\, dx$$

or in the spatial frequency domain

$$\text{i.e.} \quad F(p) \,.\, H(p).$$

The spatial domain operation could be realised in practice simply by moving the one function over the other. If transparencies of the functions were available, then a very simple correlator could be made as shown in Fig.2.1(a). The diffusing screen effectively scans one function over the other, and a complete convolution pattern is built up on the screen. A photograph of the result is shown in Fig.2.1(b). Strictly speaking, the two patterns should not be equal in size, since the "shadow" of 1st transparency will be bigger than the other transparency.

This method is unsatisfactory because no correction is made for the total amount of light transmitted by each transparency. The method evaluates the function $E\,[\,S(x)\,.\,f\,(x - u)] \;=\; R_u\,(s.f)$ but does not find $[m_{f(x)} \,.\, m_{s(x)}]$ (see Chapter 1). As a result, it is possible for cross correlation intensities to be obtained which are greater than the auto-correlation intensity.

There are several methods of overcoming this difficulty, in particular, Kowalkski[1] has described a coherent light correlator which uses a small opaque stop in the centre of the fourier transform plane to obstruct the average light coming from the transparencies. This means that the correlator evaluates $E\,[\,S(x)\,.\,f\,(x - u)] - (m_{f(x)}\, m_{s(x)})$ and gives unambiguous results.

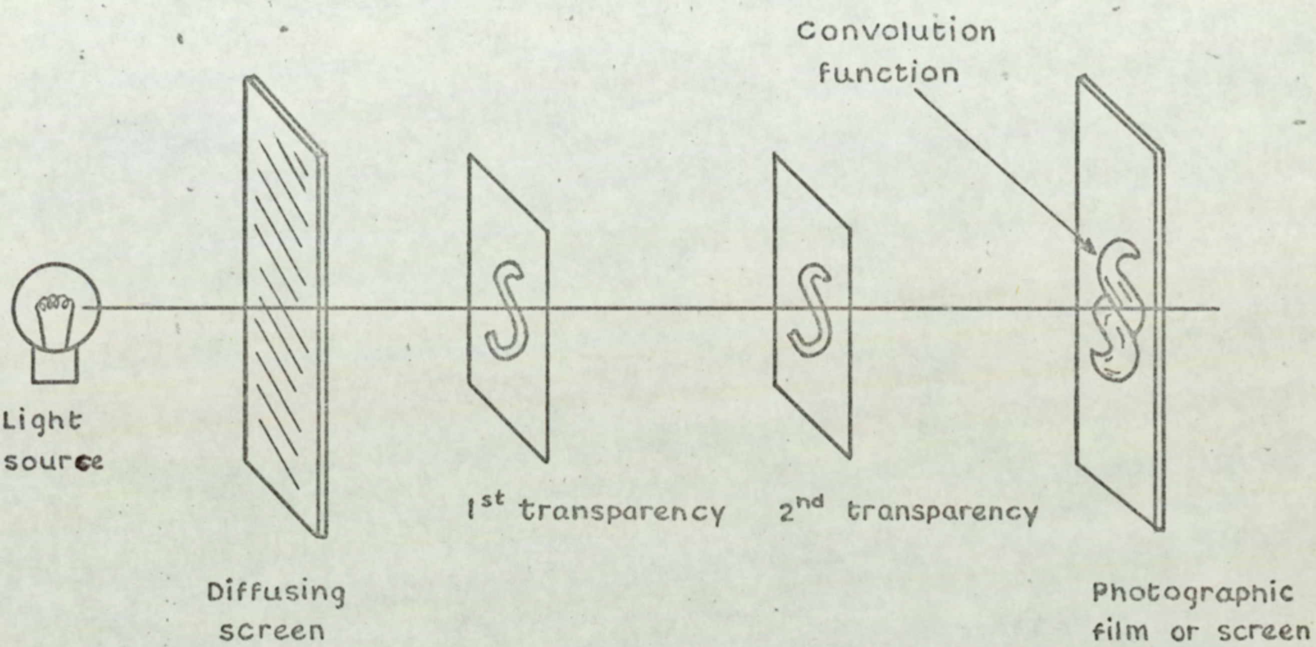


Fig. 2.1 a

Fig. 2.1 b Autocorrelation pattern of 'S'

In the following sections it will be shown how coherent optical systems can be used for both direct convolution (spatial domain processing) and spatial frequency domain processing. The advantages of the latter will be discussed.

2.2 Propagation of Light

In the theory given below, the propagation of light is described by assuming light to be a scalar quantity. Whilst this assumption is valid in most practical cases, it is not strictly accurate, and the assumption is unacceptable for very small apertures and for points very close to an aperture edge. For these last named cases one has to consider light as a vector quantity, where it is represented by electric and magnetic fields at right angles to each other. The propagation of light may then be described by the Maxwell equations.

For most of this work, the scalar theory of light will be acceptable, and it is convenient to discuss the propagation of light using this assumption. A satisfactory expression describing the transmission of light is then given by the Kirchoff integral[2];

i.e.
$$u_o(p) = -\frac{1}{4\pi}\int_{\Sigma}\left[u_o \frac{d}{dn}\left(\frac{1}{s}\cdot e^{jks}\right) - \frac{1}{s}\cdot e^{jks}\frac{du_o}{dn}\right] d\Sigma \quad \ldots . \quad 2.1.$$

where $u_o(p)$ describes the light amplitude at a point p, Σ is a surface containing the light wave, and s is the distance of the point p from element of surface $d\Sigma$. n is the normal to the surface Σ at $d\Sigma$. This integral assumes the light wave at Σ to have the form $u = u_o e^{-j\,wt}$ and describes the transmission of a monochromatic wave of maximum amplitude u_o.

The integral in its form in 2.1 is very general; for a spherical wave, emitted by a single point source;

$$u_o = \frac{A}{r}\, e^{jk\,r}$$ where A is amplitude of wave and r distance from source;

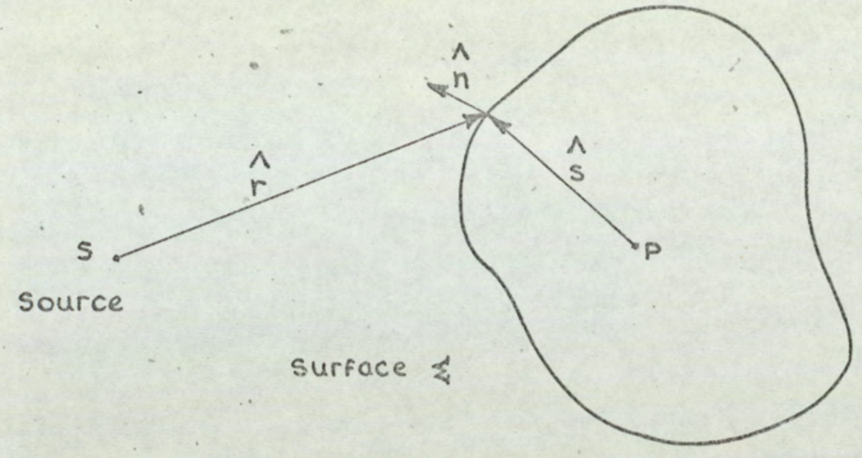


Fig. 2.2

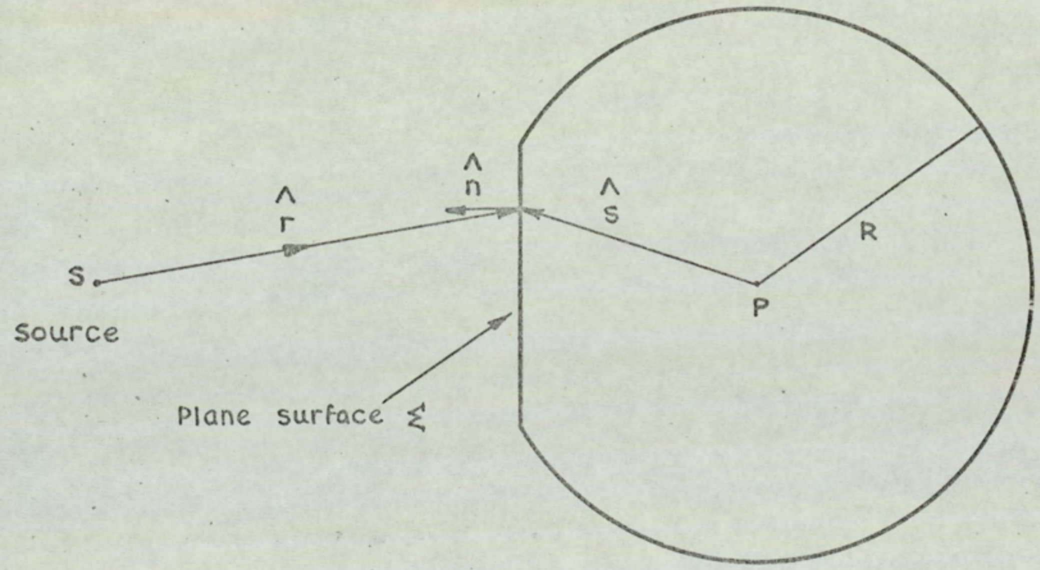


Fig. 2.3

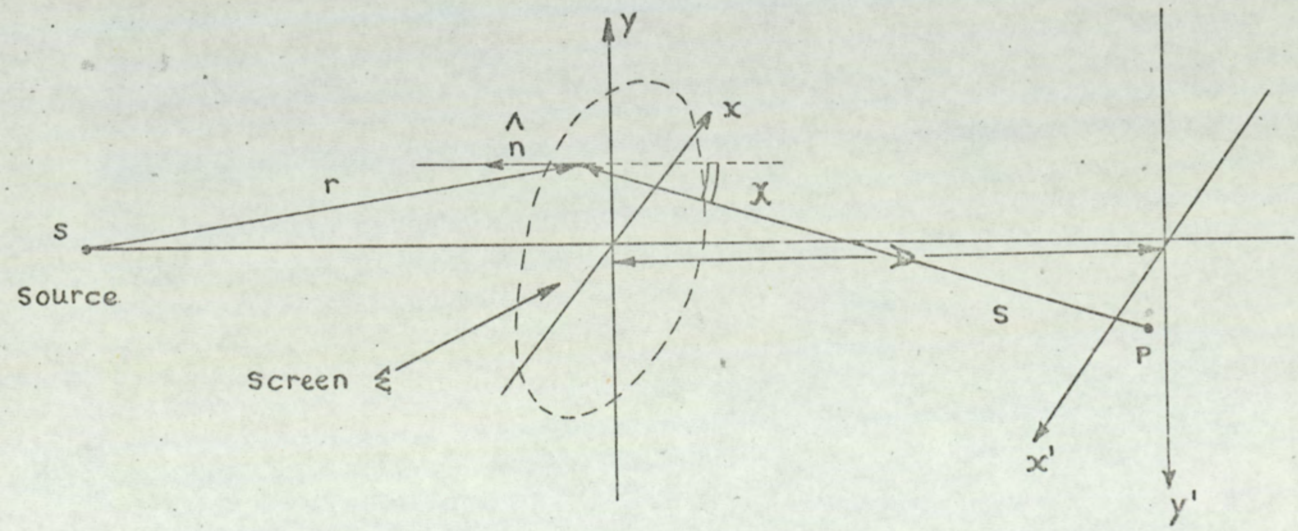


Fig. 2.4

$$\text{and } u_o(p) = \frac{-jkA}{4\pi} \int_{\Sigma} \frac{1}{r.s} \left[\text{Cos } (\hat{n}, \hat{s}) - \text{Cos } (\hat{n}, \hat{r}) \right] e^{jk(r + s)} \, d\Sigma \qquad \ldots\ldots \quad 2.2$$

where r and s are assumed large compared with the light wavelength λ and $\hat{n}$ is a unit vector normal to Σ (see Fig. 2.2). The integral now gives the light amplitude at P after transmission through surface Σ from source S.

The surface Σ is now considered to be represented by a plane disc capped by a sphere of radius R and centre P, as shown in Fig. 2.3.

It can be shown that contributions from the plane and spherical surfaces may be regarded as separate quantities, and that those from the spherical surface vanish as $R \to \infty$.

Thus if the light wave is assumed to pass through a plane screen Σ, infinite in extent, the resulting light amplitude at P is:-

$$u_o(p) = \frac{-jkA}{4\pi} \int_{\Sigma} \frac{1}{r.s} \left[\text{Cos } (\hat{n}. \hat{s}) - \text{Cos } (\hat{n}. \hat{r}) \right] e^{jk(r + s)} \, d\Sigma \qquad \ldots\ldots \quad 2.3$$

(N.B. The source, surface Σ and P are considered to be many wavelengths apart).

In the absence of screen Σ, the wave at P is $u(p, t) = u_o(p) \, e^{-iwt}$ and in presence of screen it is given by 2.3.

Further simplification of the integral in 2.3 must consider the relative positions of source, screen and point P, and their effect on the values of Cos $(\hat{n}. \hat{s})$, Cos $(\hat{n} \hat{r})$ and $e^{jk(r + s)}$. The evaluation for the case of Fraunhofer diffraction is described in section 2.3.

2.3 Application to Fraunhofer Diffraction

Fraunhofer Diffraction occurs when the distances from source to diffracting aperture, and aperture to observation point are large compared with the wavelength of light. The problem is to evaluate $u_o(p)$ when Σ is an open aperture of unspecified shape in an infinite screen. It is convenient to define a function f(x, y) which is defined all over Σ as being unity where there is an aperture and zero elsewhere. The equation describing the propagation of light must be multiplied by this function, i.e. :-

$$u_o(p) = \frac{-jA}{2\lambda} \iint_{\Sigma} \frac{e^{jk(r+s)}}{r.s.} \quad f(x, y) \; [\mathrm{Cos}\,(\hat{n}, \hat{r}) - \mathrm{Cos}\,(\hat{n}, \hat{s})]\, d\Sigma \quad \ldots\ldots \quad 2.4$$

In the Fraunhofer diffraction case, angle $(\hat{n}, \hat{r})$ is small and $\mathrm{Cos}\,(\hat{n}, \hat{r}) \approx 1$ (see fig. 2.4). Also, since $\mathrm{Cos}\,(\hat{n} . \hat{s})$ is in a negative sector, it is negative, i.e.: $[\mathrm{Cos}\,(\hat{n}. \hat{r}) - \mathrm{Cos}\,(\hat{n}. \hat{s})] = [1 + \mathrm{Cos}\,\chi]$.

The distance of source from Σ, r is slowly varying over Σ and can be regarded as constant. In addition, in most cases the distance s is nearly equal to D, and $\frac{1}{s} = \frac{1}{D}$, (for an $^{f}/_{4}$ system the difference is less than 3%.) Also, taking $\frac{Ae^{jkr}}{r}$ as a single constant, we have,

$$u_o(p) = \frac{-jA}{2\lambda D} \iint_{\Sigma} e^{jks} f(x, y) \; (1 + \mathrm{Cos}\,\chi) \; d\Sigma \quad \ldots\ldots\ldots \quad 2.5$$

The factor $(1 + \mathrm{Cos}\,\chi)$ is called the obliquity factor since χ measures the angle between the ray leaving Σ and the normal at Σ. Contributions at P are reduced with large values of χ. In Fraunhofer cases, χ is small, and $\mathrm{Cos}\,\chi \approx 1$, $\therefore$ $(1 + \mathrm{Cos}\,\chi) \approx 2$. If the coordinates of P are (x', y') then

$S = \sqrt{D^2 + (x' - x)^2 + (y' - y)^2}$ which may be simplified to

$S = D + \frac{1}{2D}\left([x' - x]^2 + (y' - y)^2\right)$ and

$$u_o(p) = \frac{K}{D} \iint_{\Sigma} f(x, y) \; e^{j\frac{\pi}{2\lambda D}} \quad (x' - x)^2 + (y' - y)^2 \quad dx\, dy \; \ldots . \; 2.6$$

where $d\Sigma = dx, dy$, and a constant $e^{j\frac{2\pi D}{\lambda}}$ has been absorbed into constant K.

When the limits of the integral are changed to infinity this integral becomes a convolution integral and may be written in the form

$$u_o(p) = u_o(x', y') = K\, f(x, y) * d\psi\,(x, y, d) \quad \ldots\ldots \quad 2.7$$

where $d = \frac{1}{D}$, $\psi\,(x, y;\; d) = e^{j\frac{\pi}{\lambda D}[x^2 + y^2]}$

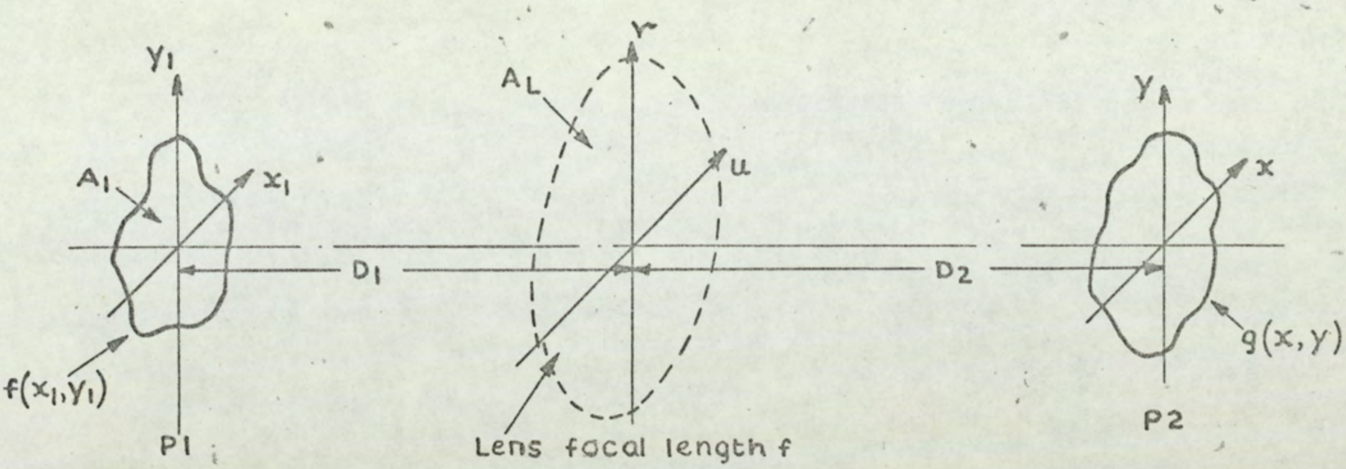


Fig.2. 5

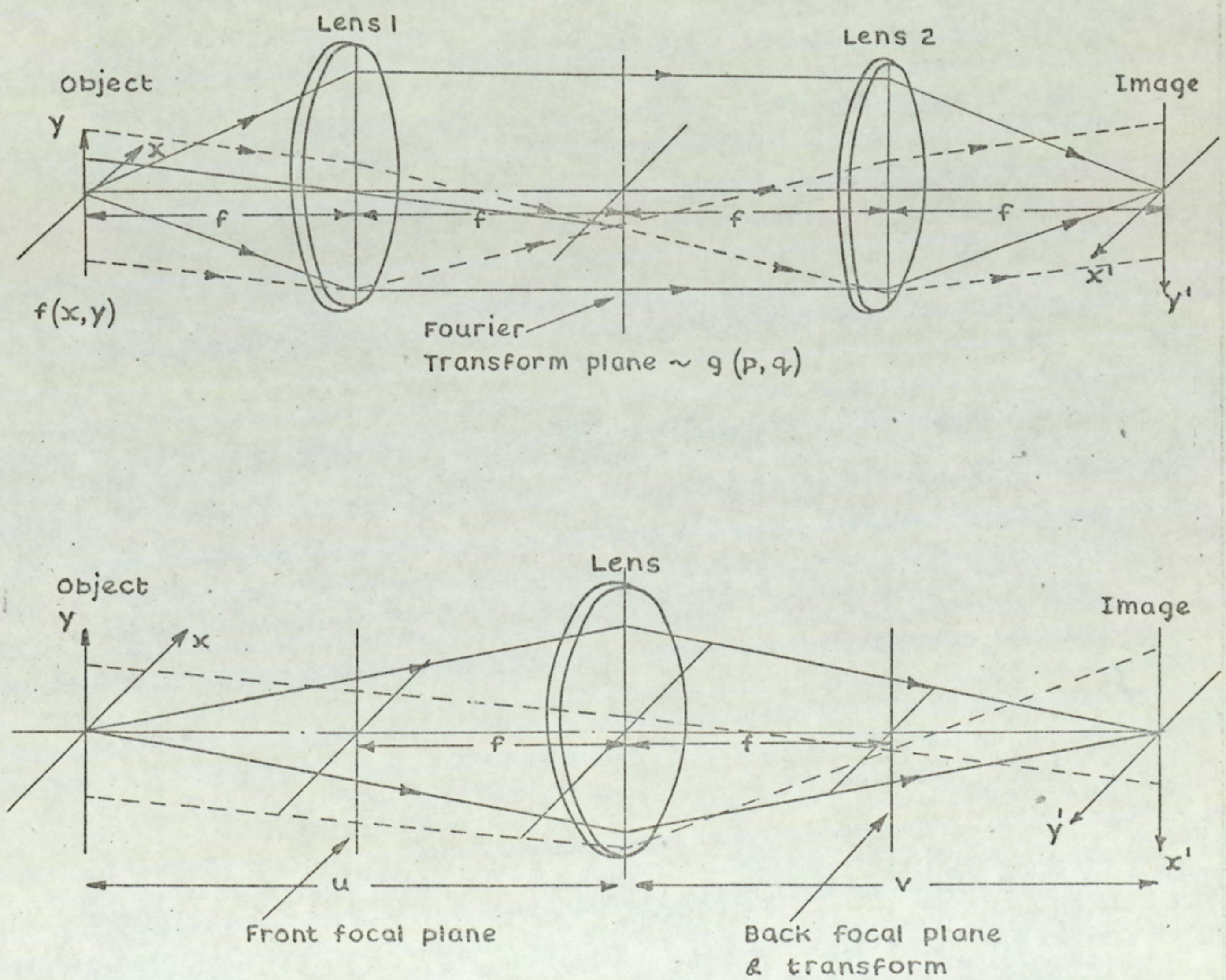


Fig.2.6 Image formation in coherent light

and $*$ denotes convolution. (The convention is to replace the variables $(x' - x)$, $(y' - y)$ by x and y in the form of 2.7).

The light amplitude at P may therefore be expressed as a convolution of the aperture function $f(x, y)$ with the function ψ describing the space through which the light passes. The form of 2.7 has been derived because it is suited for describing the behaviour of coherent optical systems, where the direct application of the Kirchoff integral would be too unwieldy. Note that 2.7 applies only to monochromatic light, emitted by a single point source.

2.4 Application of diffraction formula to lens systems

The propagation of light through a lens is now considered, as shown in Fig 2.5. For this purpose, it is convenient to think of a lens as affecting only the phase of the light which it transmits. The lens is assumed to be infinite in extent, so that its' aperture can be ignored. Since the action of a lens is to retard the wavefront, the phase change it introduces is negative. This change varies over the lens aperture, and may be described by the expression:- $(e^{-j^k/_{2f}(u^2 + v^2)})$ where f is the focal length of the lens (assumed to be positive and spherical) and u, v measure coordinates in the lens plane.

Thus a plane wave which is transmitted by the lens, will emerge as a curved, almost spherical wave having the phase $e^{-j^k/_{2f}(u^2 + v^2)}$. The phase change is peculiar to the lens itself, and all wave fronts, irrespective of their shape will suffer the same amount of phase change as described by the exponential term.

It is convenient to denote the phase change introduced by the lens as $\bar{\psi}(u, v, f)$ where f denotes its focal length.

Let $r_1(u, v)$ be the complex amplitude just before lens and $r_2(u, v)$ be the complex amplitude just after the lens. The light distribution in plane P_2 is given by $g(x, y)$ where

$$g(x, y) = K_1 r_2(u, v) * d_2 \psi(u, v; d_2) \quad \ldots\ldots\ldots \quad 2.8$$

which follows from the discussion in section 2.7. Since the lens introduces the phase change $\bar{\psi}(u, v; f)$ we can put,

$$r_2(u, v) = r_1(u, v) \cdot \bar{\psi}(u, v; f).$$

$$\therefore g(x, y) = K_2 \left[r_1(u, v) \cdot \bar{\psi}(u, v, f) \right] * d_2 \psi(u, v, d_2) \quad \ldots\ldots \quad 2.9$$

The complex amplitude $r_1(u, v)$ may also be written;

$$r_1(u, v) = Kf(x_1, y_1) * d_1 \psi(x_1, y_1; d_1)$$

$$\therefore g(x, y) = K_3 \left\{ \left[f(x_1, y_1) * \psi(x_1 y_1; d_1) \right] \bar{\psi}(u, v, f) \right\} * \psi(u, v, d_2) \quad \ldots\ldots \quad 2.10$$

Putting $\rho = \frac{(x-u)^2 + (y-v)^2}{D_2} - \frac{(u^2+v^2)}{f} + \frac{(u-x_1)^2 + (v-y_1)^2}{D_1}$

$$g(x, y) = K_4 \iint_{A_1} f(x_1, y_1)\, dx_1\, dy_1 \iint_{A_L} e^{j\frac{\pi}{\lambda}\rho}\, dudv \quad \ldots\ldots\ldots\ldots \quad 2.11$$

The form of $g(x, y)$ in the back focal plane of the lens may be found by putting $D_2 = f$ and letting $D_1 = f/m$ where m is some constant.

$$\therefore g(x, y) = K_4\, e^{-j\pi/\lambda f\,(1-m/m)\,(x^2+y^2)} \iint_{A_1} dx_1\, dy_1\, f(x_1\, y_1)\, e^{-j2\pi/\lambda f(x x_1 + y y_1)}$$
$$\iint_{-\infty}^{\infty} e^{j\pi/\lambda f\,[(u - x_1 - x)^2 + (v - y_1 - y)^2}\, dudv \quad \ldots\ldots\ldots \quad 2.12$$

where the limits on the second integral are taken to infinity by assuming the lens to be very large. (This implies that the lens accepts all the radiation from plane P_1.

The form of the second integral is $\iint_{-\infty}^{\infty} e^{j\frac{\pi}{2}(\xi^2 + \eta^2)}\, d\zeta\, d\eta$ 2.13

where $\xi^2 = \frac{2}{\lambda f}(u - x_1 - x)^2$

and $\eta^2 = \frac{2}{\lambda f}(v - y_1 - y)^2$.

Equation 2.13 is a Fresnel integral which becomes constant when evaluated over infinite limits. It may therefore be included in K_4. The limits of A_1 may also be taken to infinity since we are concerned with diffraction from the finite aperture described by $f(x_1, y_1)$

$$\therefore \; g(x, y) = K_5 \, e^{-j\frac{\pi}{\lambda f}\left(\frac{1-m}{m}\right)(x^2 + y^2)} \iint_{-\infty}^{\infty} f(x_1, y_1) \, e^{-j\frac{2\pi}{\lambda f}(x x_1 + y y_1)} \, dx_1 \, dy_1 \qquad \ldots\ldots\; 2.14.$$

Putting $p = \frac{x}{\lambda f}$ and $q = \frac{y}{\lambda f}$ gives

$$g(p, q) = K_5 \, e^{-j\pi\lambda f\left(\frac{1-m}{m}\right)(p^2 + q^2)} \iint_{-\infty}^{\infty} f(x_1, y_1) \, e^{-j2\pi(px_1 + qy_1)} \, dx_1 , dy_1 \qquad \ldots\ldots\; 2.15.$$

Equation 2.15 describes the light amplitude in the back focal plane of lens (P_2) after transmission from P_1 where it was described by $f(x_1, y_1)$, and through lens of focal length f. The diffraction effects of the light leaving P_1 and its transmission through space with phase change at lens give the form of $g(p, q)$. Diffraction by the lens itself is not considered since it is regarded as infinite in extent.

Consider the case when $m = 1$, i.e. P_1 in front focal plane of lens. The exponential term outside the integral vanishes, and

$$g(p, q) = K_5 \iint_{-\infty}^{\infty} f(x_1, y_1) e^{-2\pi(px_1 + qy_1)} \, dx_1 \, dy_1 \qquad \ldots\ldots\; 2.16.$$

The variables (x_1, y_1) and (p, q) are directly related and equation 2.16 describes an exact Fourier transform relationship between $g(p, q)$ and $f(x_1, y_1)$. Thus for $m = 1$, $g(p, q)$ is the Fourier transform of the aperture $f(x_1, y_1)$ through which the wave has passed.

In the case of $m \neq 1$ the exponential term remains and

$$g(p, q) = K_5 \, e^{j\pi\lambda f n(p^2 + q^2)} \iint_{-\infty}^{\infty} f(x_1, y_1) e^{-j2(px_1 + qy_1)} \, dx_1 \, dy_1$$

where $n = \left(\frac{1 - m}{m}\right)$ and fourier transform is exact to a quadratic phase factor.

The analysis shows that lenses have an inherent integrating property, and can form the fourier transforms of any input function. The function $g(p, q)$ in the back focal plane is called the Fraunhofer diffraction pattern of $f(x_1, y_1)$. This pattern normally forms at large distances from the aperture $(f(x_1, y_1))$ but the use of a lens enables this pattern to occur at much smaller distances. The fourier transform relationship is not peculiar to the lens itself; it is the form taken by the diffraction pattern at large distances from the aperture.

The effect of the quadratic phase factor would only be observed if the phase of $g(p, q)$ were studied. Thus the intensity distribution $|g(p, q)|^2$ is the same irrespective of the value of $(\frac{1 - m}{m})$.

2.5 Properties of lens transforming systems

The diffraction pattern formed in the back focal plane of a lens obeys all the properties of Fourier transforms. There is an inverse relationship between the dimensions of $f(x_1, y_1)$ and $g(p, q)$, if x_1 is large then p is small and vice versa.

The dimensions of p and q are measured in spatial frequency; this logically follows since $p = {}^{x}/_{\lambda f}$ whose dimensions are ${}^{1}/_{\text{period}}$ = frequency. The fourier transform represents the spatial frequency spectrum of the function $f(x_1, y_1)$; by analogy with the theorem of Fourier analysis. Because of the inverse relationship, it follows that low spatial frequencies (i.e. wide apertures) are near the centre of the transform and high spatial frequencies (narrow apertures) appear well away from the centre. The undiffracted light is focussed at the centre of the transform and which thus represents zero spatial frequency, or "d.c." light. The centre of the transform is on the axis joining point source and centre of lens.

The fourier transform, or spatial frequency spectrum of any function available in transparency form can be formed simply by illuminating the function with collimated light from a monochromatic point source and observing the pattern in the focal plane of a lens.

The addition of a further lens, behind the fourier transform will perform another fourier transform operation, which in turn will reproduce the original object function. If the transparency is not placed in the front focal plane of the transforming lens, then its fourier transform will be formed with the addition of a quadratic phase factor. However, the image of the transparency will be formed without the need for an additional lens; i.e. the inverse transformation is accomplished automatically. This is shown in Fig.2.6.

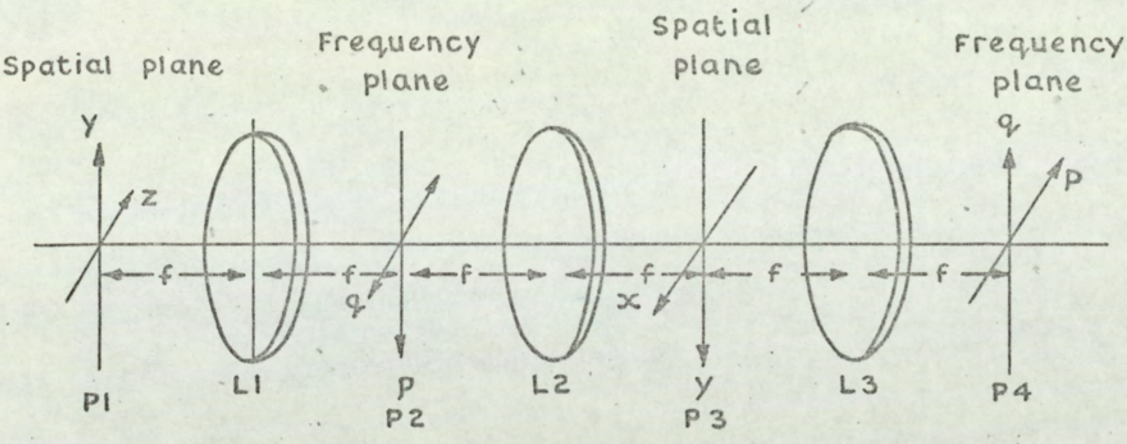


Fig 2.7 Cascade of lenses

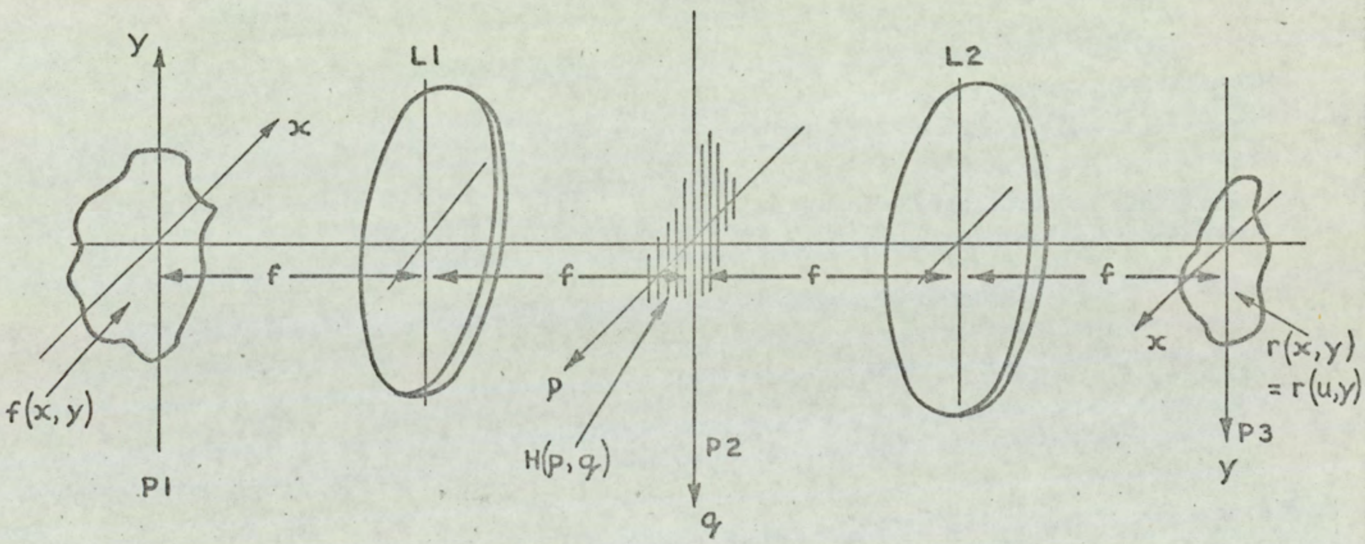


Fig. 2.8

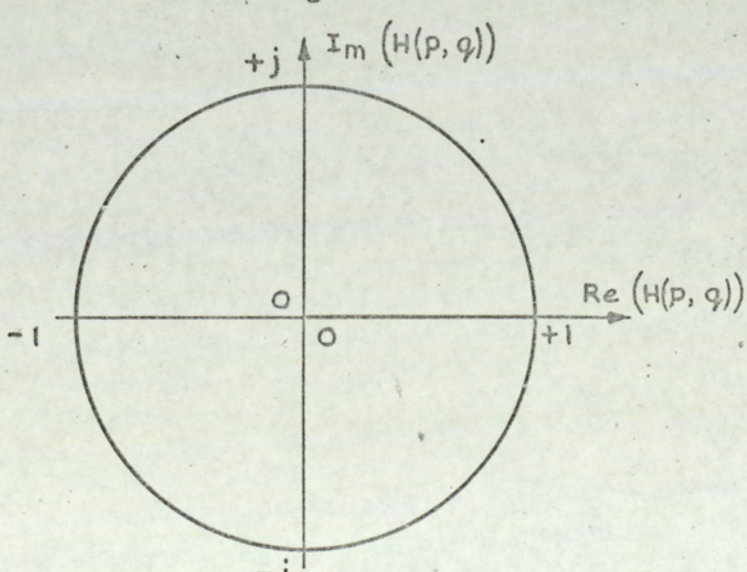


Fig. 2.9 Unit circle for spatial filters

The latter property of lenses was discovered by Abbe, and has lead to the Abbe theory of microscopic vision. This theory has very close connections with coherent optical systems, particularly since the effect of modifying the complex amplitude in the focal plane of the imaging lens has been found to have a pronounced effect on the final image.

It will be more convenient to consider the case of objects in the front focal plane of the transform lens, as this approximates to the practical optical processing systems. Two lenses can be used to form a sequence of transforming operations. In the case shown in Fig. 2.6, the second lens L2 introduces the phase factor (or kernel) exp $-j \frac{2\pi}{\lambda f} (px + qy)$ and the image plane (or back focal plane of L2) contains;

$$f'(x, y) = \iint_{-\infty}^{\infty} g(p, q)\, e^{-j \frac{2\pi}{\lambda f} (px + qy)}\, dpdq$$

The inverse transformation of $g(p, q)$ is

$$f(x, y) = \iint_{-\infty}^{\infty} g(p, q)\, e^{j \frac{2\pi}{\lambda f} (px + qy)}\, dpdq$$

i.e. it is equal to $f'(x, y)$ with reversal of sign of (x, y)

The output is therefore $f(-x, -y)$. This is the original transparency but with coordinates reversed in sign, i.e. an inverted image.

A whole sequence of lenses can be arranged to perform any number of transformations. The kernel at each transformation will be $e^{-j\frac{2\pi}{\lambda f} (px + qy)}$, but the convention is to arrange that each transformation is followed by an inverse transformation, and that the signs of alternate planes are reversed, as in Fig. 2.7.

Each spatial domain is followed by a spatial frequency domain. Distances between planes are equal to the focal lengths of the intervening lenses.

An important parameter of optical transforming systems, (particularly for information processing) is the amount of data or information which can be transmitted through the system. This amount is limited by the apertures of the lenses in the system, the rim of the lens acting as an aperture stop. A numerical method of assessing an optical system by the amount of information transmitted is the space-bandwidth product, or SBP, in analogy with electronic systems (Time-bandwidth product).

In optical systems the SBP is AR^2 where A is the maximum area of the input transparency, and R is the spatial frequency bandwidth in the fourier transform plane. Since the lens is usually symmetrical, this bandwidth is equal for both p and q directions, hence R^2. The value of R is limited by the finite diameter of the lens; if the lens diameter is reduced then R decreases, producing a smaller SBP. The value of SBP for optical systems can be very large, for example, a 10 mm square format containing detail of frequency 100 lines/mm would imply a SBP of 10^6 if this frequency were transmitted. This can easily be attained by an optical system of aperture $^f/_{16}$, which is a relatively small aperture for optics. The SBP of 10^6 compares very well with electronic systems where, in the case of a 1 msec signal of bandwidth 10^6 Hz, the time bandwidth product is 10^3.

The value of SBP for any given optical system depends on the exact configuration of lenses, input and filter planes. For a given size of input (A) the space bandwidth product will be increased by bringing input plane and transform lens close together, with a maximum when lens and input are adjacent. At this maximum, all the information contained in the input has been transmitted by the lens; in a cascaded optical system it will be necessary to consider the transmission of the following lens.

In some cases one finds that the space bandwidth product is not limited by the optical system, and the amount of information transmitted depends on the area of input and its resolution limit.

Although the SBP may be increased by placing input and transform lens together, the resulting fourier transform is not exact, and contains the phase factor $\exp^{-j\pi\lambda f\left(\frac{1-m}{m}\right)(p^2+q^2)}$. This spherical factor does not alter the intensity distribution in the transform and provided that phases are not important, does not give trouble. Suitable design of the optical system can provide correction for the spherical factor where phases are of importance.

The space invariant property of the optical system holds only when perfect lenses are considered, and when the cone of rays from every object point is transmitted by the optical system.

For an object in the front focal plane, a lateral shift in position produces a phase change at the fourier transform plane which varies linearly with spatial frequency. This means that phase variations in the transform itself will be preserved, the linear phase change can be used to locate the object in the input plane, as will be shown later.

When the object is in a plane close to the lens, as mentioned earlier, there will still be a linear phase change, even though the transform now contains a spherical phase factor. (This assumes the lens to be perfect). Provided that the lens aperture is sufficiently large to accept all information from the object and provided that the lens aberrations can be considered sufficiently small, then the optical system may be said to be space imvariant.

2.6 Spatial Filtering

The transforming property of coherent optical systems may be used for a variety of purposes. For example, analysis of spatial frequency spectra, and spatial frequency filtering. It is this latter example with which we are concerned.

The concept of spatial filtering depends on the fact that in a coherent optical system the interaction between the light wave and the filter is a multiplication of their complex amplitudes.

For instance, if the filter amplitude transmittance is described by $H(p, q)$ and the light wave is $F(p, q)$ then the light wave leaving the filter is

$$R(p, q) = F(p, q) . H(p, q)$$

In the case where $H(p, q)$ represents the fourier spectrum of some signal $h(x, y)$, and $F(p, q)$ is the spectrum of input data $f(x, y)$, the situation is shown in Fig. 2.8.

The resultant light wave leaving the filter is $R(p, q)$. This wave is transformed as it crosses to the next lens L2, its transformation appearing in the back focal plane of L2, at P3. This is given by,

$$r(x, y) = \frac{1}{4\pi^2} \iint_{-\infty}^{\infty} F(p, q)\, H(p, q)\, e^{-j(px + qy)}\, dpdq.$$

Since $F(p, q)$ and $H(p, q)$ are themselves transforms, the convolution theorem gives:-

$$r(u, v) = \iint_{-\infty}^{\infty} f(u - x, v - y)\, h(x, y)\, dxdy.$$

Here the units of u, v are identical to those of x, y. The variables u, v denote the shifting operation implied by the convolution.

A filter of the form $H(p, q)$ can thus be used to perform convolution operations on input data. Since it acts in the spatial frequency plane, it is called a spatial frequency filter. The convolution integral shows that every part of the input is filtered by the filter, the result being displayed in an output plane P3, whose dimensions are those of the input. The function $h(x, y)$ corresponds to the form of the filter in the spatial plane and is called the impulse response of the filter, (section 1.3).

The convolution operation could be performed more directly, by manipulating a transparency of the form $h(x, y)$ in the input plane P1. Since complex amplitudes may be multiplied this gives:

$$f(x - u, y - v)\, h(x, y).$$

This product will be transformed by L1 producing the transform

$$r(u, v) = F\left[f(x - u, y - v)\, h(x, y) \right] \text{ in the plane P2.}$$

At the point $p = q = 0$, this becomes:-

$$r(u, v) = \iint f(x - u, y - v)\, h(x, y)\, dxdy.$$

This is the form of a cross correlation integral. By putting $x - u \rightarrow u - x$ and $y - v \rightarrow v - y$ we have

$$r(u, v) = \iint f(u - x, v - y)\, h(x, y)\, dxdy$$

as obtained before. The results of the convolution for the values (u, v) are displayed only at the point p = q = 0, with values for other (u, v) being displayed sequentially.

The advantage of the previous method of filter synthesis (i.e. in the spatial frequency plane) is that the convolution operation is performed automatically, at the expense of an additional lens. Since the result of convolution over a whole input plane is instantaneous it will be necessary to consider only this form of filtering.

The general form of the spatial frequency filter is

$H(p, q) = |H(p, q)|\, e^{j\phi(p, q)}$

where $|H(p, q)| \leq 1$ since the filter is always passive in the case of transparencies and similar media.

A convenient way of describing different types of filters has been suggested by Cutrona[3] (et al) and Vander Lugt[4].

Since the maximum value of $|H(p, q)|$ is unity, all values of H(p, q) can be represented by points within a unit circle whose centre is the origin of the real and imaginary axes, and whose radius is unity, as shown in Fig. 2.9.

All forms of spatial filters can be represented within the unit circle, including the following special cases, which are taken from Ref.4.

2.6.1 Binary Filters

These filters can have only values zero and one on the real axis of the unit circle, with $e^{j\phi(p, q)} = 1$ (i.e. $j\phi(p, q) = 0$). This class of filter is not only the simplest form of filter but also the easiest to construct; it may be simply an opaque disc of appropriate diameter (high frequency pass filter) or an opaque screen containing an aperture of appropriate diameter, (low frequency pass filter). The action of the filter is simply that of a stop; restricting the number of spatial frequencies which are transmitted to the output plane.

The effect on the image depends on which frequencies are obstructed; in the case of low pass filters there may be removal of high frequency noise or, for instance, removal of dots in newsprint illustrations, leaving a 'pure' half tone image. In the case of high pass filters there will be removal of the d.c. term and low frequencies, which has the effect of enhancing the sharp edges in the image. This kind of filter can also be used to increase contrast and its use corresponds to differentiation of the input function.

Binary filters do not have to be symmetrical about the d.c. term; a wedge shaped filter can be used to obstruct all spatial frequencies at one particular orientation about the axis, i.e. between θ^{o} and ϕ^{o} and between $\pi + \theta^{o}$ and $\pi + \phi^{o}$. This can be of use in the processing of signals containing strong periodic signals which are to be suppressed, as in seismic data, for example.

The observation of contours in pure phase objects by the Schlieren method is another application of spatial filtering with a binary filter. Here all spatial frequencies to one side of the central d.c. term are obstructed, and contributions to the image come from one side of the transform only, enabling phase effects to be seen as intensity variations in the image.

2.6.2 Amplitude Spatial Filters

In these filters the function $|H(p, q)|$ can take any value from 0 to 1 along the real axis of the unit circle, with $j\phi$ again zero. They can be made very easily using photographic techniques.

A common application of this type of filter is in apodisation, where a gradual change in amplitude transmission is required. Similar masks or filters are often used in other positions than in the spatial frequency plane, but apodised spatial frequency filters are often used to reduce the undesirable "ringing" effect which is produced by diffraction at the sharp edge of a binary filter.

Vanderlugt has shown how amplitude filters may be used for least-mean-square error processing systems.

Suppose a signal $(s(x, y))$ is to be extracted from a background of noise $(n(x, y))$, then the filter which allows this to be done with minimum error is:

$$H(p, q) = \frac{S_s (p, q)}{S_s(p, q) + S_n(p, q)}$$

where $S_s(p, q)$ and $S_n(p, q)$ are the spectrum densities (or intensities) of signal and noise.

Amplitude filters may also be used to compensate for the transfer function $H_s(p, q)$ of an optical system. Here the filter is made of the form $^1/_{H_s(p, q)}$, and in theory would recover the original signal transmitted by the system. In practice $\frac{1}{H_s(p, q)}$ tends to large values for high frequencies and since the maximum value of $\frac{1}{H_s(p, q)}$ is limited to unity, for a passive filter, the result is not very effective.

2.6.3 Phase Spatial Filters

These filters are completely transparent, i.e. $|H(p, q)| = 0$ and can vary in thickness, i.e. have values between $+j$ and $-j$ on the unit circle. They have no effect on the modulus of the complex amplitude, and interact with the phase component.

The filters are made by introducing relief into films or by figuring glass plate. The manufacture of a continuously varying phase filter is difficult, and it is usual for the filter to advance or retard the phase of a function by $^\lambda/_2$. This can be done by depositing a thin film of thickness $^\lambda/_2$ on glass plate.

An application in which a special form of phase filters have been used has been described by Cutrona (et al) where the filter is synthesised by use of a series of lenses. The combination included conical and cylindrical/spherical lenses and used for the processing of radar data.

2.6.4 Amplitude and Phase filters

These filters are composed of the amplitude filters and phase filters described earlier, the composite filter having, in theory any value within or on the circumference of the unit circle. A wide variety of tasks can then be accomplished, including Zernicke's phase contrast method of viewing pure phase objects. In practice values are confined to the real or imaginary axes since continually varying values of $j\phi$ are difficult to construct.

Consider a pure phase object, this may be represented by:

$f(x, y) = e^{j\phi(x, y)}$, and if ϕ is small this becomes

$$f(x, y) \approx 1 + j\phi(x. y)$$

This fourier transform of this function is;

$$F(p, q) = \iint_{-\infty}^{\infty} [1 + j\phi(x, y)] e^{-j(px + qy)} \, dxdy$$

$$= A^2 \operatorname{Sinc}\left(p \frac{A}{2}\right) \operatorname{Sinc}\left(q\frac{A}{2}\right) + j\,\Phi(p, q)$$

where A is the length of the aperture edge.

$F(p, q)$ is displayed in the frequency plane of the system (say, plane P2, in Fig. 2.5).

The first term of $F(p, q)$ contains energy which mainly concentrated on the optical axis within a zone $\pi/_A$ wide. Since this zone contains little information about the phase variations of the object, the basis of Zernicke's method is to advance or retard the phase of the central zone by $\pi/_2$ so that the image plane contains the interference pattern formed by the central zone and the off-axis phase term $\Phi(p, q)$.

The filter must therefore correspond to H(p, q) where

$$H(p, q) = C. \exp(j\alpha) \; ; \quad |p| \leq \frac{2\pi}{A}, \; |q| \leq \frac{2\pi}{A}$$

$$= 1 \quad ; \text{ elsewhere.}$$

where C is an attenuating factor controlling the contrast of the interference pattern, and $|\alpha| \leq \pi/2$.

The output is then r(x, y), which is the transform of the product

$H(p, q) \; F(p, q)$

i.e. $r(x, y) = C \exp(j\alpha) + \exp(j\pi/2)\phi(x, y)$

The intensity in the image plane is $|r(x, y)|^2$ or

$$C^2 + \phi^2(x, y) + 2C\phi(x, y) \operatorname{Cos}(\alpha - \pi/2)$$

Since ϕ is small,

$|r(x, y)|^2 = C^2 + 2\phi(x, y) \operatorname{Cos}(\alpha - \pi/2)$, and the image intensity now contains the phase information present in the original object. The contrast of the image is controlled by the attenuating factor C, it is also controlled by α, setting $\alpha = +\pi/2$ (+ve phase contrast) gives bright field contrast, setting $\alpha = -\pi/2$ (-ve phase contrast) gives dark field contrast.

Another use of amplitude and phase filters is for the correction of aberrations. Two basic methods are suggested by Tsijuichi. The coherent method uses a filter of the form

$$H(p) = \frac{2J_1(4\beta p)}{4\beta p}$$

where the wave aberration is assumed to be that for defocussing i.e. $kw(p) = 2\beta p^2$.

Differentiation of functions may be performed using amplitude and phase filters.

The phase variations can also be produced by using polarisation effects, and the filter has values which are strictly on the real axis of the unit circle. In this method two Vectograph films are used to record the positive and negative values of H(p, q) and the combination of films enables the amplitude components to be resolved into two oppositely directed components, at π radians out of phase.

2.6.5 Complex Spatial Filters

With this class of filter, all values on or within the circumference of the unit circle can be assumed. The range of tasks which can be accomplished is wide, probably the most important being the manufacture of matched filters, for which it is necessary to record the function $f^*(x, y)$, involving the recording of a continuously varying phase function.

There are two methods by which complex spatial filters can be made; one is by Holography, the other is by computer generation. Both forms of construction utilise the same basic principle to record the continuously varying phase function. By analogy with communications techniques, the function to be recorded is stored on a spatial carrier frequency, which consists physically of an array of fine lines drawn on a transparent substrate. The amplitude part of the filter function is recorded as a variation in contrast of the lines, whilst the phase is stored as a variation in position of the lines.

The holographically made filter is prepared on photographic film or plate of very high resolution, by making a direct recording of the interference pattern formed between the transform of the required signal and a reference wave, as described in Chapter 3.

The computer generated filter may also be prepared on photographic film, but the form of the fourier transform is computed by the computer. To avoid having to print a continuously varying grey scale, corresponding to the amplitude variations in the transform, the transform is printed as an array of small dots in much the same way as in halftones are made in the printing industry.

By altering the position of the regularly spaced dots in the transform, changes in phase can be simulated on the computer generated hologram. This type of hologram is particularly useful when the signal or pattern to be sought does not exist physically.

Both forms of filter can be used to record complex functions, and hence the task of pattern recognition by optical correlation becomes possible. The construction of holographic complex filters, their properties and applications will be discussed in the following chapters.

References to Chapter 2

1. D. C. Kowalksi, Bendix Technical Journal, Summer 1968
2. J. M. Stone, Radiation and Optics, McGraw-Hill
3. Cutrona et al, IRE Trans. Inf. Th. June 1960, 6 386
4. Vander Lugt, Optica Acta 15, 1, 1968, 1-33

Chapter 3

Construction of complex filters for correlation processes

3.1 Introduction

This chapter will discuss the method of constructing a complex filter using holographic techniques. It will also show how a matched filter can be made, and how such a filter can be used in an optical correlator for pattern recognition.

The first chapter has shown that pattern recognition is essentially a problem of correlating input data with some kind of memory. The mathematical operation is

$$\left[\int_{-\infty}^{\infty} f(x)\ s^*(x+u)\right] - m_s m_f$$

where s(x) is the signal or pattern in the memory, and f(x) is the input data.

The second chapter has shown that optical systems may be used to perform a convolution operation in either the space or spatial frequency plane. If the spatial frequency plane contains a function of the form $H^*(p,q)$ then the resultant integral becomes a cross-correlation integral, being identical in form to that required for pattern recognition.

It is considered reasonable at this stage to point out the distinction between convolution and correlation. In the former operation the mathematical description is

$$r(u) = \int f(x)\ s\ (u-x)\ dx$$

where u represents a shift of the function s(-x) with respect to f(x). Since the function s(-x) contains the negative ordinate it is effectively reversed, compared with s(x). The function may be regarded as being folded back on itself, hence the German "gefältung".

The convolution operation may also be expressed as a multiplication of the fourier transforms

$$r(u) = \int F(p).\ S\ (p)\ e^{j2\pi px}\ dp.$$

In the case when $S(p) = S^*(p)$

then the transform becomes

$$r(u) = \int f(x)\, s^*(-[u-x])\, dx \qquad \text{Using } S^*(p) \rightarrow s^*(-x)$$
$$= \int f(x)\, s^*(x-u)\, dx$$

which is a correlation integral. Here u is again a shift, but s(x) is not reversed or folded as in the convolution integral. Systems which perform convolution integrals can thus be used to evaluate correlation integrals when the frequency plano contains a complex conjugate.

The way in which a complex filter may be constructed using holography is now considered. The method owes much to the work of Vander Lugt[1,2] who first applied the techniques of holography to the preparation of spatial filters. In particular, Vander Lugt showed that a complete matched filter could be made, giving optimum signal recognition in the presence of noise.

For the present, consider the case where the noise spectrum is uniform and hence may be represented by a constant. It is necessary then to consider only the complex signal spectrum $s^*(p, q)$.

3.2 Holographic method of constructing complex filters

3.2.1 Historical

The method of recording complex functions using holography derives from the work of Gabor,[3,4] who was concerned with the problem of improving the limit of resolution of electron microscopes. As an approach to this problem Gabor proposed a two step method, which involved first, photographic recording information from an object, and second re-illumination of the record to produce a magnified image. The recording was made complete in both amplitude and phase information by allowing light from the object to interfere with a background wave, producing a pattern of interference fringes in which this information was retained. Re-illumination of the record or hologram, by a background wave caused images of the object to appear either side the recording. Disadvantage of this method was that the background and images were in line, and the former obscured the latter.

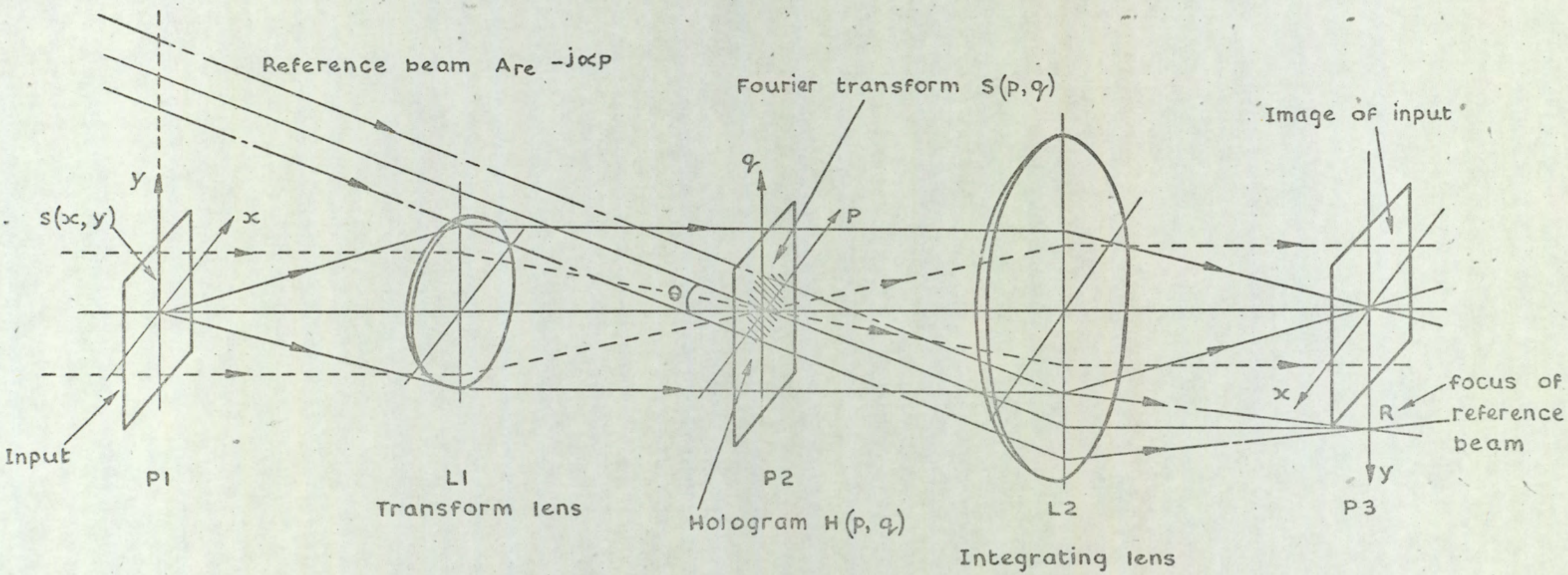


Fig. 3.1 Hologram synthesis in coherent optical system

Progress on this technique was slow until the laser was invented, providing an intense beam of coherent light. Using the laser, Leith and Upatnicks[5,6,7] were able to show that "off-axis" hologram could be made by interference between an oblique background or reference wave and the object beam. The images were then separated from each other and from the reference wave. Such a technique demands the use of light sources having a high coherence length, with lasers this can be of the order of 30-60 cms or higher. Development in the field of holography has been extremely rapid in the last five or six years, and there are many papers describing the mechanism of the process[8,9,10,11,12]. The following analysis gives a conventional description of the process.

This is still imperfectly understood, particularly when the hologram is considered in three dimensions, ie as a "volume" hologram.

3.2.2 Holographic recording

In order to record the complex function S(p, q) where S(p, q) is the transform of s(x, y), it is necessary to introduce a simple coherent reference wave, as shown in Fig 3.1.

The signal function s(x, y) is assumed to be available in transparency form, and its fourier transform S(p, q) is formed in the back focal plane of lens L1. The function is recorded on a photographic plate together with the oblique reference wave $A_r\ e^{-j\alpha p}$ where α is related to the angle of incidence θ by $2\pi \text{Sin}\ \theta = \alpha\lambda$.

The photographic plate is a square law detector and in the absence of the reference wave records only $|S(p, q)|^2$ thus losing all phase information in the complex function. Since the reference wave is present it records

$$|A_r\ e^{-j\alpha p} + S(p, q)|^2 \quad \ldots\ldots\ldots\ldots \quad 3.1$$

This may be expanded, giving

$$|A_r|^2 + |S(p, q)|^2 + A_r\ S^*(p, q)\ e^{-j\alpha p} + A_r\ S(p, q)\ e^{+j\alpha p} \quad \ldots\ldots\ldots \quad 3.2$$

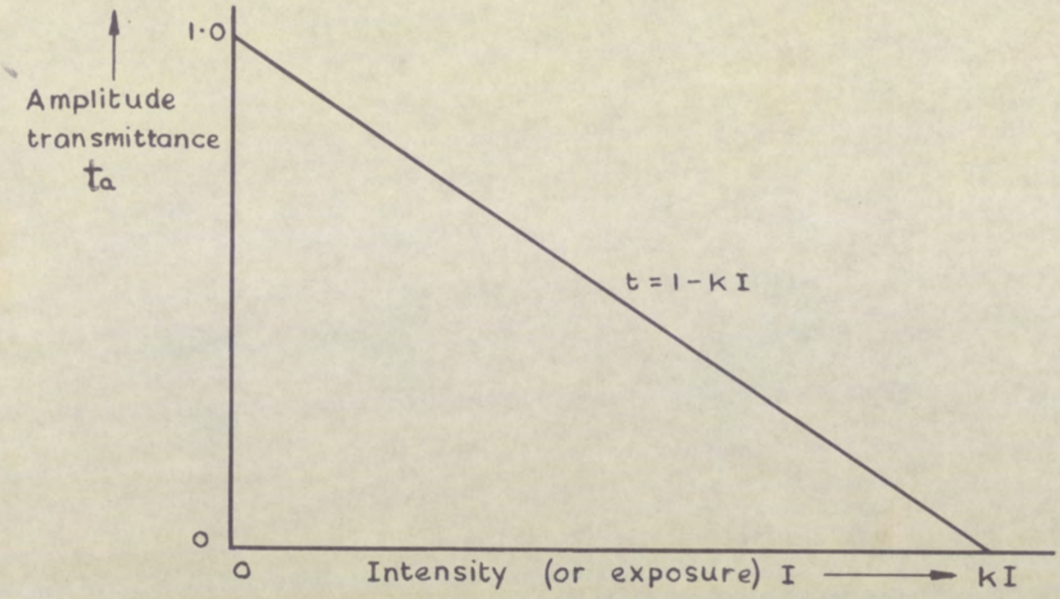
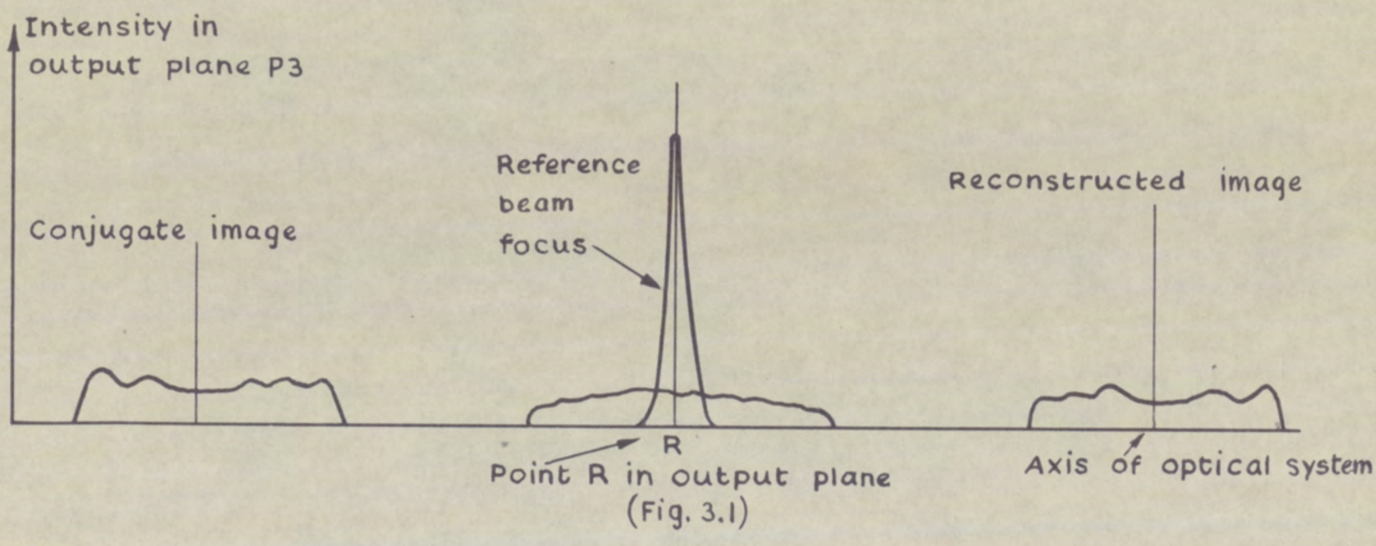


Fig. 3.2 Linear photographic characteristic



Correlation of signal (complex conjugate) and reference function convolved with reference function

ie: $r(x,y) * \left[r(x,y) \circledast s(x,y)\right]$

Autocorrelation of reference function plus autocorrelation of signal convolved with reference function

ie: $r(x,y) * \left[r(x,y) \circledast r(x,y)\right] + r(x,y) * \left[s(x,y) \circledast s(x,y)\right]$

Correlation of signal and reference function convolved with reference function

ie: $r(x,y) * \left[r(x,y) \circledast s(x,y)\right]$

Fig. 3.3 Impulse response of Fourier Transform hologram

There are two important results in obtaining 3.2. One is that the function $S(p, q)$ has been recorded as well as its complex conjugate $S^*(p, q)$, the second is that both these functions have been recorded as intensity distributions in the fourier transform plane.

It is required that the amplitude of light leaving the plate should be proportional to the amplitude of the function which has been recorded, ie: to $S^*(p, q)$. Since the function has been recorded as an intensity distribution it is required that the amplitude transmittance of the plate be proportional to the incident intensity. Thus ideally, $t_A = kI$ where t_A = amplitude transmittance, and k is a constant.

For a single step photographic process (ie: for a negative) there is a decrease in transmission with intensity, and the transmission curve could then be represented by

$t_A = 1 - kI$ if proportionality is to be conserved.

Here the maximum value of $t_A = 1$ and k is a constant. A curve of this form is shown in Fig 3.2

The hologram transmission $H(p, q)$ would then be given by

$$H(p, q) = 1 - k\left[|A_r|^2 + |S(p, q)^2| + A_r S^*(p, q)\, e^{-j\alpha p} + A_r S(p, q) e^{+j\alpha p}\right] \quad \ldots\ldots \quad 3.3$$

This expression may be rewritten, giving

$$H(p, q) = 1 - kA_r^2 - kA_s^2(p, q) - 2\, kA_rA_s(p, q) \operatorname{Cos}(\alpha p - \phi(p, q)) \quad \ldots\ldots \quad 3.4$$

where A_r is the amplitude of the reference wave and $A_s(p, q)$ is the amplitude of the signal transform junction ie $S(p, q) = A(p, q)e^{-j\phi(p, q)}$.

The last expression for $H(p, q)$ shows that the hologram contains a constant term $1 - kA_r^2$ whose transmittance is varied by the term $kA_s^2(p, q)$. In addition to this the hologram transmittance is modulated by a cosine term of fundamental frequency α and of amplitude $2kA_rA_s(p, q)$. The amplitude of the modulation thus varies with $A_s(p, q)$ and also the frequency, since the term $(\alpha p - \phi(p, q))$ is present. It is the cosine term which carries the basic information concerning the signal function.

The response of the hologram may be examined by reference to equation 3.3. For the present we consider the impulse response of the hologram, obtained by illuminating the hologram with the reference wave $A_r e^{-j\alpha p}$. The result of this is the product:

$$A_r e^{-j\alpha p}. H(p, q) = A_r e^{-j\alpha p} - kA_r^3 e^{-j\alpha p} - k|S(p, q)|^2 A_r e^{-j\alpha p}$$
$$- kA_r^2 S^*(p, q) e^{-2j\alpha p} - kA_r^2 S(p, q). \qquad \ldots\ldots \; 3.5$$

The impulse response of the hologram thus contains three groups of terms, those with exponentials of $-j\alpha p$, those with $-2j\alpha p$ and those with zero.

The terms containing $\exp(-j\alpha p)$ appear in the path of the reference beam and are focussed by the second lens L2 to appear at the point R in output plane P3. (Fig 3.1). These terms are:-

$$A_r - kA_r^3 - k|S(p, q)|^2 A_r \qquad \ldots\ldots \; 3.6$$

The first two terms are constant and form a bright point of light at R, the image of the original reference beam source. The third term is $kA_s^2(p, q).A_r$ and gives a convolution pattern centred on the point R. The convolution appears because the lens L2 takes the transform of the product $A_s^2(p, q)Ar$.

The term with $\exp(-2j\alpha p)$ is diffracted through angle 2θ relative to the hologram and appears below R. It consists of the function $kA_r^2 S^*(p, q)$, and since A_r is constant contains the complex conjugate of the signal, $s^*(-x, -y)$. The image will be reversed because of the transformation.

The last term, containing zero exponential, appears on axis, and consists of the function $A_r^2 S(p, q)$. Again A_r is constant and so this is simply the original signal function.

In this discussion we have assumed that the convolution of A_r^2 and $S(p, q)$ to be equal to $S(p, q)$. This implies that A_r^2 is a delta function, which in practice is only true for an extremely small point source for the reference wave.

It would be more precise to express the term $A_r^2\ S(p, q)$ as a triple convolution when it is transformed;

ie: $$A_r^2\ S(p, q) \text{ is strictly } A_r e^{-jp\alpha}\left(A_r e^{+j\alpha p} \cdot S(p, q)\right) \quad .... \quad 3.7$$

Denoting the reference wave by R(p, q) this is:-

$$R(p, q) \cdot \left(R^*(p, q)\ .\ S(p, q)\right) \quad .... \quad 3.8$$

Under transformation 3.8 becomes:-

$$r(x, y) \ * \ \left[r^*(x, y) \ \circledast \ s(x, y)\right] \quad .... \quad 3.9$$

where * denotes convolution and $\circledast$ correlation.

Equation 3.9 shows that the image seen on axis is strictly the correlation of reference source and signal, convolved with the reference source. (Note, since $R^*(p, q)$ is complex, under transformation the convolution becomes a correlation).

Only in the case when r(x, y) is very small (ie: a 'point' source) will the image resemble exactly the original object.

It is also interesting to note that the third term of equation 3.6 is strictly

$$k\left[S(p, q)\ .\ S^*(p, q)\right] .\ R(p, q) \quad .... \quad 3.10$$

and under transformation this is

$$k\left\{\left[S(x, y) \ \circledast \ S^*(x, y)\right] * r(x, y)\right. \quad .... \quad 3.11$$

ie: the term is on autocorrelation of the signal function, convolved with the reference point source.

The impulse response of the hologram may be depicted in diagram form as shown in Fig 3.3.

3.2.3 Correlation response of hologram

The previous section has shown how a hologram can record the function $S^*(p, q)$. The response of a hologram to an input signal f(x, y), whose transform is F(p, q) is now considered.

It is only necessary to be concerned with the term containing $S^*(p, q)$. This is, assuming the reference point source to be a delta function;

$$A_r \, S^*(p, q) \, e^{-j\alpha p}$$

where A_r is a constant term, and may be neglected.

Multiplication of the hologram with the function $F(p, q)$ gives

$$F(p, q) \quad S^*(p, q) \qquad \ldots\ldots \quad 3.12$$

where the exponential is ignored since it is understood that this term will appear off-axis by amount θ^o.

Under transformation 3.12 becomes

$$\iint_{-\infty}^{\infty} F(p. q) \, . \, S^*(p. q) \, e^{-j(px + qy)} \, dp \, dq \qquad \ldots\ldots \quad 3.13$$

Since $F(p. q)$ and $S(p, q)$ are themselves transforms this is:-

$$\iint f(x, y) \, . \, s^*(x + u, y + v) \, dx \, dy \; = \; r(u, v) \qquad \ldots\ldots \quad 3.14$$

The equation 3.14 is the required cross correlation integral between input $f(x, y)$ and the known signal (or pattern) $s(x, y)$. The result of correlation is displayed in the output plane P3, by the function $r(u, v)$, where u, v measure the degree of shift introduced by the correlation.

When the input contains the signal $s(x, y)$ then 3.14 becomes an auto-correlation integral, and gives a very high response in the output plane.

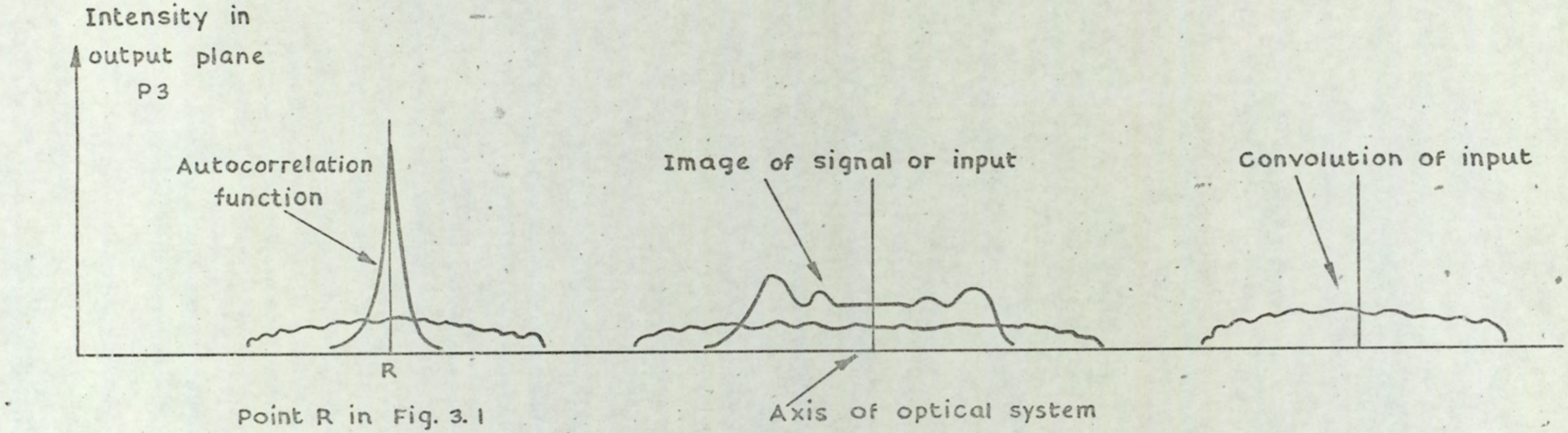
The precise mechanism of this may be seen as follows:-

Let $f(x, y) = s(x, y)$ and 3.14 becomes:

$$r(u, v) = \iint s(x, y) \, . \, s^*(x + u, y + v) \, dx \, dy$$

$$= \iint |s(x, y)|^2 \, dx \, dy \qquad \ldots\ldots \quad 3.15$$

Since $|s(x, y)|^2$ is the signal energy, and the integral is over all x, y, then all the signal energy is concentrated on the point u, v, giving a very high intensity.



Cross correlation of input and signal

ie: $f(x,y) \circledast s(x,y)$

Produces autocorrelation peak when $s(x,y) \circledast s(x,y)$

Convolution of input with autocorrelation of reference beam, plus convolution with autocorrelation of signal

ie: $f(x,y) * \left[r(x,y) \circledast r(x,y)\right]$

$+ \; f(x,y) * \left[s(x,y) \circledast s(x,y)\right]$

Convolution of input and signal

$f(x\,y) * s(x\,y)$

N.B Strictly

$f(x\,y) * \left[r^{*}(x\,y) * s(x\,y)\right]$

N.B. Strictly $f(x,y) * \left[r(x,y) * \overset{*}{s}(x,y)\right]$

Fig. 3.4 Response of hologram to an input

The response of the hologram for the other terms can be easily evaluated;

on axis one has the convolution of the input with the constant reference beam term $k|A_r|^2$ and also with the autocorrelation of the signal $|S(p, q)|^2$.

On the opposite side of the axis the input is convolved with the original signal function (not its conjugate).

The output plane, for the hologram response to an input f(x, y) may be shown in a similar way to Fig 3.3. in Fig 3.4.

The separation of the three terms depends on the value assigned to in the reference beam function, in the making of the hologram.

For an integrating lens (L2) of focal length F, the separation of the centres of the output functions is given by b where

$$b = F \operatorname{Tan} \theta$$

$$= F \operatorname{Tan} \frac{\lambda\alpha}{2\pi}$$ and α is the phase factor of the response wave.

If the spacing of the interference fringes on the hologram is d, the frequency of spacing is $\frac{1}{d}$ and

$$\frac{1}{d} = \frac{\operatorname{Sin}\theta}{\lambda}$$ where θ is angle of reference beam incidence.

Taking $\operatorname{Tan} \theta \approx \operatorname{Sin}\theta$

$$b = \lambda F. \frac{1}{d} \qquad \ldots\ldots \qquad 3.16$$

Equation 3.16 shows that the separation of terms, b, is directly proportional to wavelength, focal length and carrier frequency of hologram.

The actual widths of the terms are given by summing the lengths of the functions involved in convolution or correlation processes.

Let the width of the signal be l and that of the input be L.

The central term is of width L + 2l, the correlation term is of width L + l and the convolution term L + l.

To avoid overlap of these terms the value of b must be given by

$$b \geq L + \frac{3}{2} l.$$

ie: $\lambda F. \frac{1}{d} \geq L + \frac{3}{2} l$ to avoid overlap 3.17

Provided equation 3.17 is maintained, the three terms will be separately resolved, and the hologram can be used to 'search' a whole input plane for a given signal or pattern.

It is important to realise that the foregoing arguments are based on the critical assumption that the hologram transmission is proportional to the incident intensity. The fact that this is true for only a limited range of intensities and that the relative values of $|A_r|$ and $|S(p, q)|^2$ have a pronounced effect on the response mean that theoretical arguments are only very approximate.

3.3 Hologram as a matched filter

It was shown in chapter 1 that the optimum response of a signal detector in the presence of noise, was obtained when a matched filter of the form

$$H(p, q) = \frac{S^*(p, q)}{S_n(p, q)} \text{ was used.}$$

In this section we consider how the noise spectral density may be recorded, together with the conjugate of the signal spectrum.

The function $\frac{1}{S_n(p, q)}$ can be recorded simply by exposing the noise spectrum to a photographic plate, which will record the spectral density. Its amplitude transmittance will be inversely proportional to the spectral density provided that the plate is processed so that the slope of the density/log exposure curve is equal to two (ie: a negative). [This aspect will be discussed later.]

The processing of the photographic plate for noise is different to that for the hologram, although in practice it is possible to compromise and record both by successive exposures on the same plate before processing. Alternatively, two plates could be exposed, and placed in tandem.

The filter function will then be

$$H(p, q) = \frac{1-k\,[\,|A_r|^2 + |S(p, q)|^2 + A_r S^*(p, q)e^{-j\alpha p} + A_r S(p, q)e^{+j\alpha p}\,]}{S_n(p, q)} \qquad \ldots\ldots \quad 3.18$$

where the hologram transmission given by equation 3.3 is used.

Since we are only concerned with the complex conjugate of the signal spectrum, the term of interest is:-

$$\frac{kA_r\, S^*(p, q)e^{-j\alpha p}}{S_n(p, q)}$$

In the case where the input is F(p, q), the hologram response is:-

$$H(p, q) . F(p. q) = \frac{k.\ F(p, q).\ A_r .\ S^*(p, q)e^{-j\alpha p}}{S_n(p, q)}$$

$$= \frac{k_1\ F(p, q)\ S^*(p, q)e^{-j\alpha p}}{S_n(p, q)} \qquad \ldots\ldots \quad 3.19$$

In general, the input function f(x, y) will be composed of noise n(x, y) and there may be a signal s(x, y) present.

ie: $f(x, y) = n(x, y) + s(x, y)$

Under transformation this becomes

$F(p, q) = N(p, q) + S(p, q)$

The design of the filter has been to maximise the peak signal energy to mean square noise energy.

Taking the signal term, the hologram response is

$$\frac{k_1\ S(p, q)\ S^*(p, q)e^{-j\alpha p}}{S_n(p, q)} \qquad \ldots\ldots \quad 3.20$$

($S_n(p, q)$ = noise spectral density)

and under transformation this becomes;

$$k_1 \iint_{-\infty}^{\infty} \frac{|S(p, q)|^2}{S_n(p, q)}\, dpdq \qquad \ldots\ldots \quad 3.21$$

The square of 3.21 gives the peak signal energy in the output plane: ie

Peak signal energy is: $$\left|\, k_1 \iint_{-\infty}^{\infty} \frac{|S(p, q)|^2}{S_n(p, q)}\, dpdq \,\right|^2 \qquad \ldots\ldots \quad 3.22$$

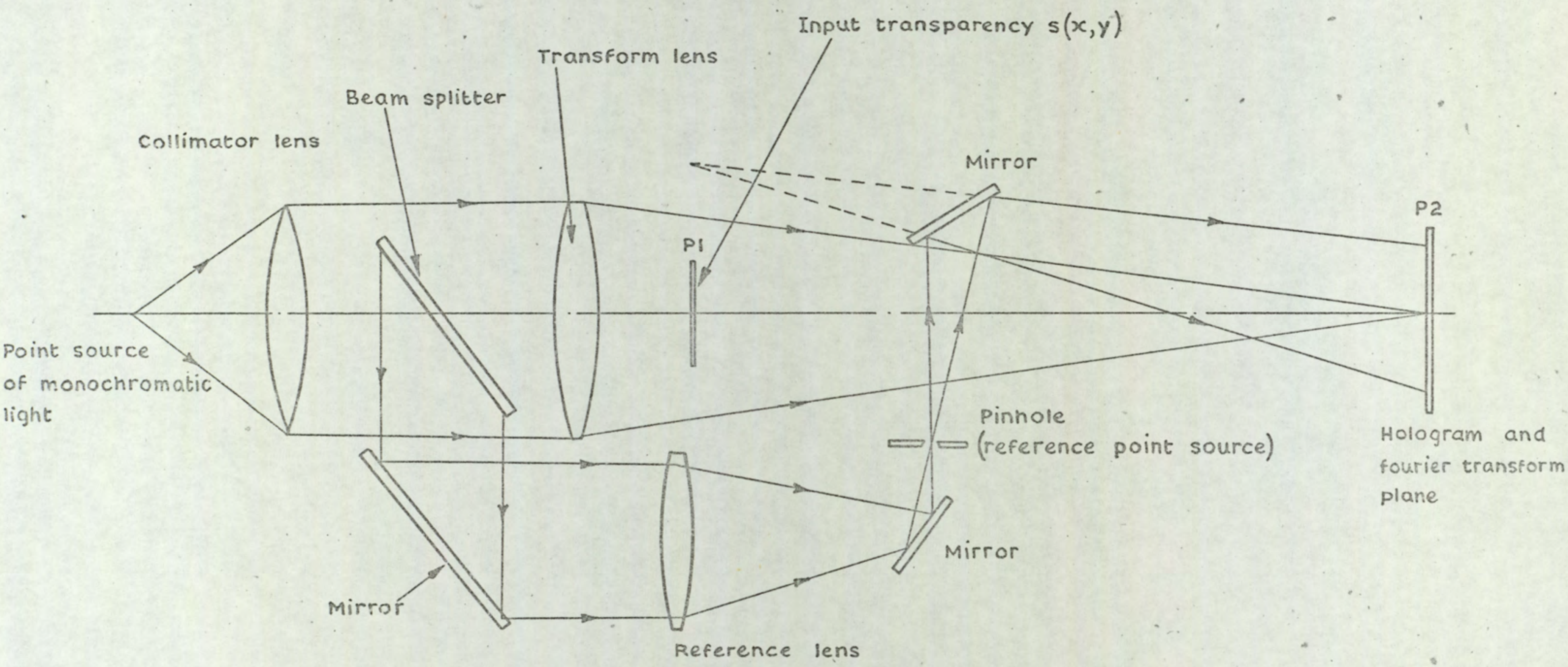


Fig.3.5 Interferometer for making fourier transform holograms (after vander lugt)

The mean square noise is given by

$$|H(p, q)|^2 S_n(p, q) = \frac{|S(p, q)|^2}{S_n(p, q)}$$

which under transformation is:-

$$\iint_{-\infty}^{\infty} \frac{|S(p, q)|^2}{S_n(p, q)} \, dpdq \qquad \ldots\ldots \qquad 3.23$$

The ratio of peak signal to mean square noise energy is equation 3.22 divided by 3.23 which is

$$k_1 \iint \frac{|S(p, q)|^2}{S_n(p, q)} \, dpdq \qquad \ldots\ldots \qquad 3.24$$

Here we see that a point (u, v) in the output plane will contain a signal whose ratio of energy to mean noise energy is given by 3.24.

When the correlation of noise and filter is considered, the "signal to noise" ratio is much lower because the correlation of noise and signal is lower than the correlation of signal with itself.

3.4 <u>Practical Considerations</u>

The previous sections have shown how a hologram may be used to record a complex function and thus show how a filter of the form $\frac{S^*(p,q)}{S_n(p,q)}$ can be synthesised.

With these techniques, and using a coherent optical system, pattern recognition tasks become capable of solution.

An optical arrangement suitable for generating the filter has been described by Vander Lugt[1,2]. This arrangement is derived from a Mach Zehnder interferometer, and is shown in Fig 3.5.

In this arrangement, the reference and signal beam both originate at a point source, being separated by a beam splitter. The pinhole obstructs all but the centre part of the source image, so that a clean perfectly spherical wave is obtained.

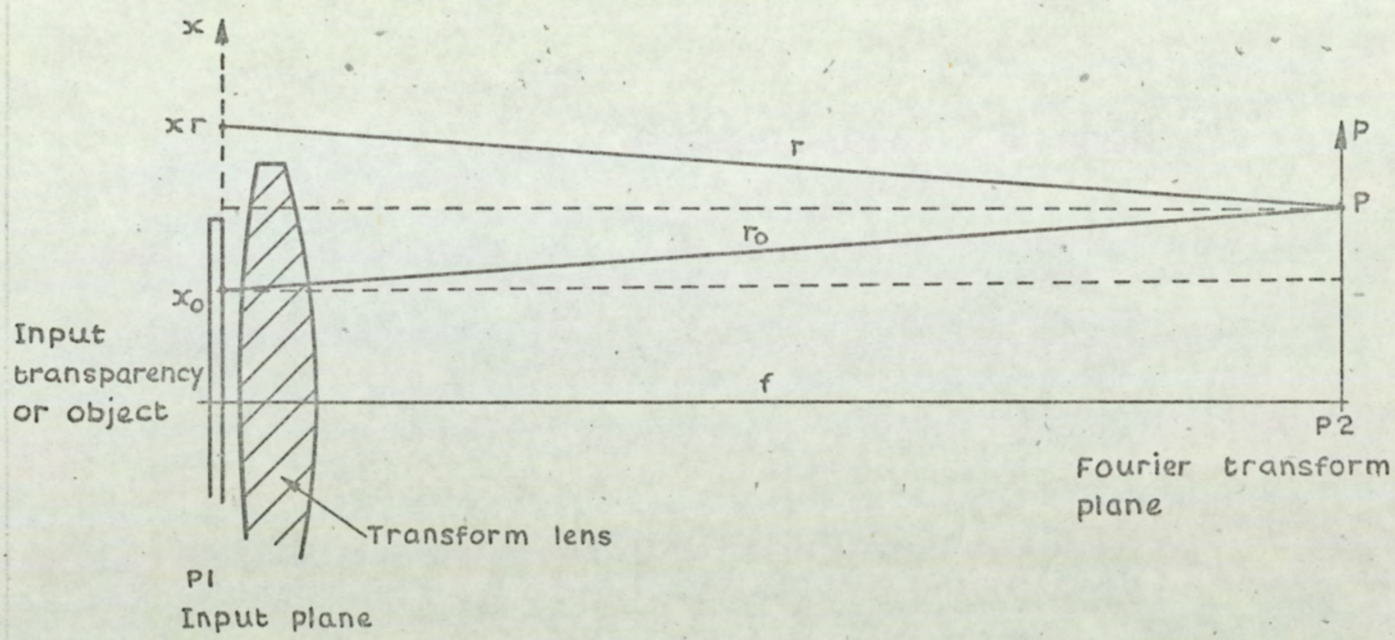


Fig. 3.6 Correction of spherical phase factor

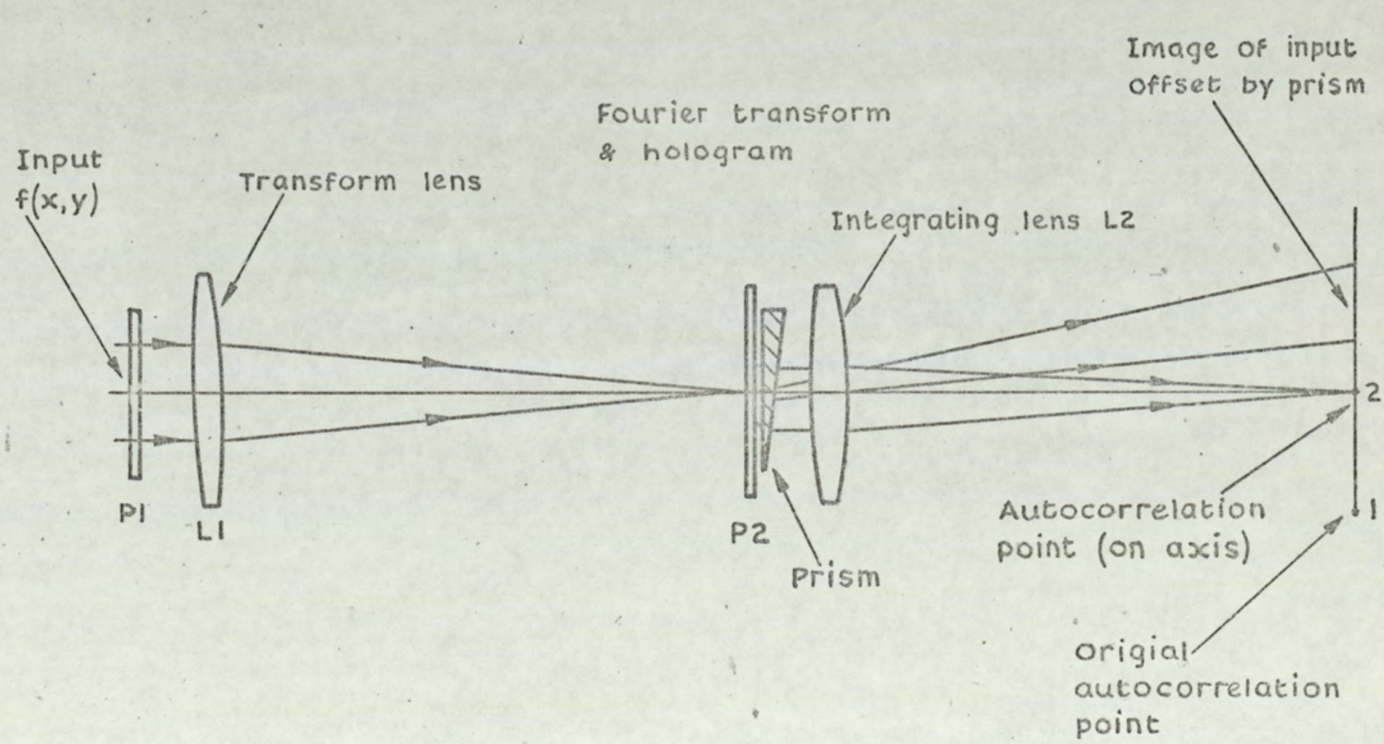


Fig. 3.7 Practical optical correlator

The signal function s(x, y) is contained in the converging signal beam, and its fourier transform is formed in the hologram plane P2. The transform will contain a spherical phase factor, since the signal is not in the front focal plane of the transform lens. This does not destroy the fourier transforming properties of the system, and only means that the output plane is not formed in the focal plane of the lens following the hologram, but some distance further along the axis.

The reference point source (or its image) is placed in the same plane as the signal function (ie: P1) so that the spherical phase factor in the transform is corrected. This is easily seen since the reference wave phase may be represented by $A_r\, e^{-ikr}$, in Fig 3.6, and the object wave, for a point x_o in the object is $A_o\, e^{-ikr_o}$ at the point Pin plane P2.

Now

$$A_r e^{-ihr} = A_r e^{\frac{-ik}{2f}(x_r - p)^2}$$

$$\text{and} \qquad A_o e^{-ikr_o} = A_o e^{\frac{-ik}{2f}(p - x_o)^2} \qquad \ldots\ldots \qquad 3.25$$

The hologram records $|A_r e^{-ikr} + A_o e^{-ikr_o}|^2$ which becomes, using 3.25

$$|A_r|^2 + |A_o|^2 + A_r A_o e^{\frac{-ik}{2f}(x_r - p)^2}\, e^{\frac{+ik}{2f}(p - x_o)^2}$$

$$+ A_r A_o e^{\frac{+ik}{2f}(xr - p)^2}\, e^{\frac{-ik}{2f}(p - x_o)^2} \qquad \ldots\ldots \qquad 3.26$$

The exponential terms of 3.26 evaluate to

$$\exp\left[\frac{-ik}{2f}\left[x_r^2 - 2x_r p + p^2\right] + \left[-p^2 + 2x_o p - x_o^2\right]\right]$$

$$\text{and} \quad \exp\left[\frac{+ik}{2f}\left[x_r^2 - 2x_r p + p^2\right] + \left[-p^2 + 2x_o p - x_o^2\right]\right]$$

The terms containing x_r^2 and x_o^2 are constant for the two points under consideration, while the p^2 terms cancel. The exponentials are then

$$\exp\left\{\frac{-ik}{2f}\left(-2x_r p + 2x_o p\right)\right\} = \exp\left\{\frac{-ikp}{f}\left(x_o - x_r\right)\right\}$$

$$\ldots\ldots \qquad 3.27$$

$$\text{and} \quad \exp\left\{\frac{+ik}{2f}\left(-2x_r p + 2x_o p\right)\right\} = \exp\left\{\frac{+ikp}{f}\left(x_o - x_r\right)\right\}$$

The exponential terms are now linear in p, and this means that the hologram interference fringes will be described by a cosine term

$$\cos \frac{k}{f}p \quad (x_o - x_r)$$

These are straight fringes, of constant period $\frac{k(x_o - x_r)}{f}$, and thus show that the hologram has no focussing power. Straight line fringes can cause a wave front to be deviated or diffracted, but they cannot alter its curvature. The presence of curved fringes or fringes of varying spacing in a hologram is considered undesirable because the hologram would then act like a lens, and since the field angle over which the hologram operates may be large, this could give aberrations and space invariant performance.[2]

Placing the signal function in the converging signal beam allows the scale of the transform to be varied; moving the plane P1 towards P2 reduces the size of the transform, moving P1 away from P2 increases the size. This facility is useful when patterns of the same shape but over a variation of sizes must be examined.

When the hologram is made it can be used in a conventional optical processing system of the type shown in Fig 3.1. Normally, the autocorrelation function would appear off axis, because of the carrier fringes in the hologram, and this means that the aperture of lens L2, following the hologram must be increased to accept this function.

The demands on L2 can be lowered by moving this lens up to the hologram, since its transforming properties will still be retained, apart from a spherical phase factor. This is shown in Fig 3.7.

It is often necessary to consider the orientation of the pattern, and since the transform of the signal is unlikely to be symmetrical about the axis, it will be necessary for either the input or the hologram to be rotated. In the event of the latter suggestion, a prism placed immediately behind the hologram, as in Fig 3.7, will bring the diffracted correlation beam back on axis, and it will only be necessary to look for autocorrelation in the region of point 2, in plane P3.

The detection of the autocorrelation may be done visually, photographically or by photo-electric means. The latter seem particularly suited, since an intensity threshold can be set to enable true autocorrelation peaks to be distinguished from false alarms.

Although a source with high temporal coherence is required to construct the hologram, this is not essential when the hologram is used in the correlating system. However, it is essential that light with extremely good monochromacitity is used, and for this reason, a laser virtually always used in coherent optical systems. The laser also has the advantage that the light is emitted as a thin parallel beam, which can be convenient when setting up the optical system, as shown in Chapter 4.

3.5 Research considerations

There are several openings for investigation in the optical correlator which has just been described. Very broadly there seem to be two aspects of an investigation;

1) Improvements in the design of the optical system, with particular reference to the specific problem on hand.
2) Investigation of the limitations of the basic principles of the system, particularly where assumptions have been made in theory, which are not met in practice.

Most of the research work has been concerned with an investigation of the limitations as described under 2). In the course of this work, modifications to the design of the optical system (from that shown in Fig 3.5 and 3.7) were made, and also in the design of specific items of apparatus.

The main topics chosen for investigate are as follows:-

a) Influence of photographic non-linearity on system response. The assumption that hologram amplitude transmittance is proportional to incident intensity is only true for a limited range of intensities, and this will affect not only the recording of the complex functions but also of the cosine term of the carrier frequency fringes.

b) Other effects resulting from use of photographic material as a spatial filter. There is not only a variation in amplitude transmittance across the hologram, but also a variation in thickness, caused mainly by tanning action of the developer. The effect of this, and the effect of grain on the output must be considered.

c) Influence of hologram or spatial filter location. The necessity for removing the photographic plate for processing and replacing it in the optical system means that small displacements of the filter might have to be tolerated. Some theoretical work on this has already been done by Vander Lugts, but the change in appearance of the autocorrelation function as the hologram is shifted has not been studied experimentally.

d) Space invariance of hologram response over input plane. The filter must be capable of giving a constant response to a signal, irrespective of where the signal is situated in the input plane. In practice, the filter must form autocorrelation peaks of equal intensity over the extent of the autocorrelation plane.

e) Space-invariant properties of optical system. This problem is closely related to d) above. The signal function may appear anywhere in the input plane. Ideally, the only change which should occur at the fourier plane as this signal function moves, is that of phase, and this should vary in proportion with the signal position. In practice lens aberrations, particularly spherical aberration, degrade and displace the fourier transform, especially at the edges of the input field. This problem would occur not only with the transform lens but also with the following integrating lens. This latter must cover a field equal in size to the input plane, even though the use of a prism removes the need to resolve over the angle of carrier frequency fringes; the aberrations of the prism must then be considered.

f) Ratio of reference beam and signal beam intensities. This problem will be closely related to a), since it is the maximum values of $|A_r|^2$ and $|A_s|^2$ (for reference and signal beams respectively) which determine the bias level of the carrier fringes and hence the degree of non-linearity in the recording.

g) Influence of reference beam characteristics. One of the most important demands is that the reference beam point source should be small enough to approximate to a delta function. Strictly, the output from the hologram is

$$f(x, y) \ * \ (s^*(-x, -y) \ * \ r(x, y))$$

where $r(x, y)$ represents the complex amplitude distribution across the source. If this $r(x, y)$ is a delta function, then the required correlation

$f(x, y) \circledast s^*(x, y)$ is produced.

Another important characteristic is the degree of coherence between the object and reference beams. Perfect coherence is theoretically impossible, and the contrast of the fringes has to be modified by a partial coherence factor. The effect of this is to reduce the fringe contrast, making the output weak. The precise effect will be discussed in connection with item a).

h) Geometry of optical system. This strictly refers to the investigation of carrier frequency, or reference beam angle. However, the theory seems to be quite valid in establishing the minimum carrier frequency necessary for avoiding overlap in the output plane. There is little point in having a high carrier frequency, except that it may be desirable to have a minimum number of fringes to "resolve" parts of the transform.

i) Approach to ideal matched filter response. The matched filter gives the optimum response to a signal in noise, but this performance may be degraded by the limitations mentioned in items a) h) above. In addition, the problem of subtracting the mean energy levels ought to be considered, since it was shown that correlation does not, by itself, give a correct indication, especially when one signal is stronger than another. This point has been mentioned (Section 1.5), but could be investigated experimentally.

j) Multiplexed filters. The ability of holograms to record complex functions can be extended to other kinds of filters apart from matched filters. In particular, one can record several kinds of function on the single hologram, which is called multiplexing. Such a hologram can be used to identify any one of a set of patterns, and has obvious applications in character recognition. The multiplexed filter will no longer be a matched filter in the sense of optimum signal/noise ratio, and there will be a possibility of false correlation.

It would be impossible for all the above points to be investigated in detail within a reasonable time. However, most emphasis has been placed on items a) and c) which are considered to be most important in affecting system response. The report continues by discussing the design of the optical system and apparatus, followed by chapters discussing various experiments on the items mentioned earlier.

References to Chapter 3

1.	Vander Lugt	IEEE Trans IT-10 (April 1964) p.139
2.	Vander Lugt	App. Optics 5, 11 Nov 1966 p.1760
3.	Gabor	Proc. Roy. Soc. A197 (1949)
4.	Gabor	Proc. Roy. Soc. B, Vol 64, p.449 1951
5.	Leith and Upatnieks	J.O.S.A. 52, 10, Oct. 1962. p.1123
6.	" "	J.O.S.A. 53, 12. Dec. 1963. p.1377
7.	" "	J.O.S.A. 54, 11. Nov. 1964. p.1295
8.	R. J. Perris	Research Report 65-IC2-OPTIC-R2/1965/ Westinghouse R & D Center
9.	R. W. Meier	J.O.S.A. 56, 2 Feb. 1966. p.219
10.	D. G. Falconer	Phot.Sc. & Eng. 10, 3, May 1966, p.133
11.	J.B.De Vellis, et al	J.O.S.A. 56, 4, April 1966, p.423
12.	J. A. Armstrong	I.B.M. Journal, May 1965, p.171

Chapter 4 Design of Optical System and Apparatus.

4.1 Introduction - design considerations

The design of the experimental optical system will take into account the theoretical requirements, and attempt to meet these within practical limits. These limits will be based on present-day technology, financial restrictions and on available facilities. To some extent, the final system will be a compromise, aiming for the ideal system within the limits stated. The limits may be regarded as parameters which the design must take into account, along with the basic theoretical requirements. These are discussed in turn below:

4.1.1 Theoretical Considerations

The system must meet the demands of the theory discussed in chapters 1, 2 and 3. In some cases there may be advantages in having two systems, one for making the complex filter, and the other for using the filter in correlation processes. The advantages are that the demands of one system would not have to be applied to the other, giving more flexibility in design. In this work, however, a single system is used for both making the filter and for correlating it with an input. This too,has advantages, one of these is that irregularities in the system itself which are recorded together with the required transform when making the filter, will also be present when the filter is subsequently used on a given input and will be cancelled.

The basic optical system will be a combination of interferometer and optical transformer. The two parts may be considered separately:-

a) Consider transformer:- In order that the lenses may form a transform in which the effects of spherical aberration are negligible, it is necessary to restrict the relative aperture of the system to f/10 or smaller. The focal lengths of the lenses will then be determined by the linear apertures required, ie by the dimensions of the input format. To a certain extent, the choice of input format is arbitrary, but it was noted that a large format

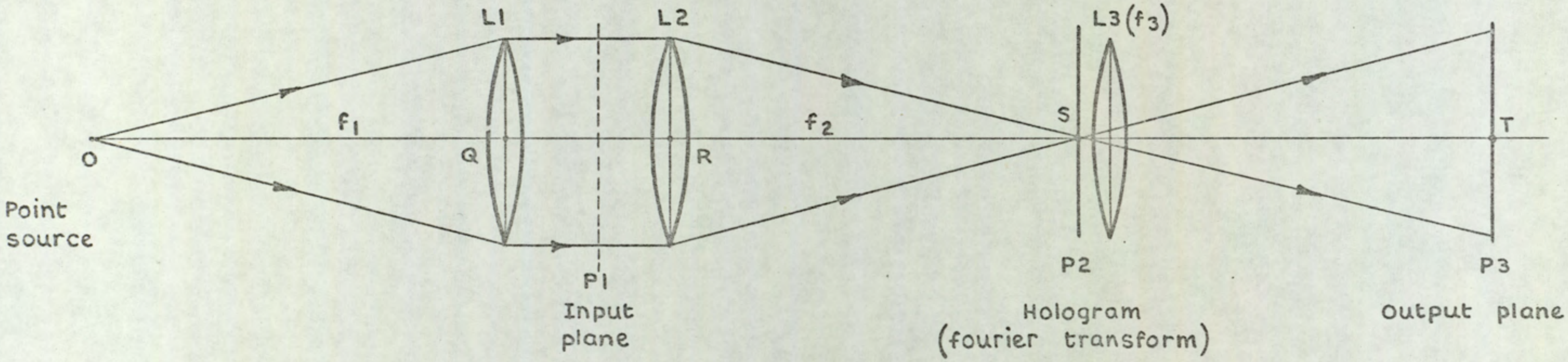


Fig.4.1 Coherent optical system

would demand large and expensive lenses. The format was therefore kept small, to a maximum of 6 x 6 cms square. This means that the focal lengths would be of the order of 60 cms. It was necessary to check at this stage, that the fourier transform dimensions were of reasonable size. The highest spatial frequency likely to be dealt with is 100 lines/mm, and with a focal length of 60 cms, this produces a spectrum some 3.6 cms off axis. The hologram would thus be at least 7.0 cms wide; a not unreasonable size for handling and processing, and also comparing well with available photographic plate sizes.

Assuming that lenses of 60 cms focal length are chosen, the overall working length of the system can be obtained. In order to keep the overall length short, and to retain the maximum space bandwidth product the design was based in part on that of Vander Lugt[1] Fig 3.4), and in part on authors' earlier work on holography and character recognition[2]. This design is shown in Fig 4.1.

The collimating lens (L1) needs to be of performance equal to the transform lens L2, particularly on axis, and so L1 and L2 could be identical. Thus $f_1 = f_2$. The distance between L1 and L2, in which the input would be placed needs to be large enough to allow several components to be inserted if necessary, for example, liquid refractive index matching cells, amplitude masks and neutral density filters. A convenient distance is about 20 cms. The lens L3 has to form the inverse transform in the output plane P3. Since the lens L2 forms a virtual image of the plane P1 at a finite distance from L3, the focal length of the latter must be less than that of L2 otherwise P3 would be at ∞ or virtual. The minimum separation between P1 and P3 occurs when L3 has focal length $f^2/2$, ie: $4f_3$ = RT. This also has the advantage of 1:1 magnification, and the scale of the correlation plane is equal to that of the input plane. The output plane will then be a distance f_2 away from L3.

The overall length of the system, from 0 to T is therefore $60 + 20 + 60 + 60$ cms ≈ 200 cms.

Discussion of the optical components themselves will be considered later, for the present we are concerned with the overall geometry of the system. The components need to be rigidly aligned with respect to each other, and calls for the use of an optical bench.

b) Interferometer

That part of the transformer from 0 to S (Fig 4.1) can be regarded as being one arm of the interferometer. The other arm forming the reference beam, will ideally be of equal optical path length, and the two beams will have originated from some common light source to the left of 0, the "point source" for input beam. In theory, it would be possible to split the input beam after it emerged from 0 and use one of the components as the reference beam. This was shown in Vander Lugt's system in Fig 3.4. However, the inclusion of a beam splitter after the input beam source means that this beam will be degraded by any imperfections of the beam splitter and there will be scattering by dust on its surface. It was considered that a better system would be obtained by splitting the beam before it reached 0. This will add an additional length to the system, of perhaps 10-20 cms to allow for beam splitters and neutral density filters.

The common source of light for the interferometer must be monochromatic and have a high spatial and temporal coherence. The temporal coherence is necessary to accommodate for differences in path length in the arms of the interferometer whilst the spatial coherence maintains a constant phase between all points of the incident wavefront.

As previously mentioned, the ideal source is the laser, of which the gas laser producing a stable, continuous beam of monochromatic radiation is particularly suitable. With gas lasers the output power seems in practice to increase with laser cavity length, and a moderately powered laser of say, 10 mW could be expected to be about 100 cm in length.

The overall length of the system would thus be 200 + 15 + 100 = 315cms, ie: transformer length plus interferometer length.

There will also be very high demands on the stability of the optical components in the interferometer, because the maximum tolerance on movement of the carrier fringes is of the order of one quarter of a fringe width. This works out at about 0.6μ for a reference beam angle of 10°, although Rogers(3) has shown that the exact tolerance for fourier transform is difficult to calculate. Clearly, a very rigid mounting must be provided for all components of the interferometer. The mounting must ensure that all the components of the interferometer remain fixed relative to each other, and that external vibrations are reduced to a minimum.

4.1.2 Practical Requirements

At this stage we can consider some of the practical requirements, which are set by the demands of experimental work. The requirements are:-

a) Flexibility; the system will be required to perform several tasks, depending on the nature of work on hand. It must be capable of being modified easily, and allow different features of the system to be examined. This latter point may be shown by noting that the hologram output is confined not only to one point in the correlation plane, but covers the whole plane from one side to the other. The distribution of energy in this plane depends on whether the impulse or correlation response is being obtained, and it will be desirable to observe both in the course of experimental work.

b) Ease of adjustment. The majority of optical systems can be very rapidly aligned on an optical bench. This facility is needed particularly if the experimental arrangement is to be frequently changed.

c) Cost and availability. Every experimental system is restricted by cost and availability. One has to consider apparatus which is cheaply or rapidly available, as well as existing apparatus, used perhaps in earlier projects.

d) Environment The manufacture of holograms and the operation of a coherent optical system call for an environment which is free from temperature fluctuations and from vibration. The isolated location of the laboratories in which all this work was carried out was such that the nearest source of vibrations from heavy plant and machinery was over 1 mile distant. Although an air compressor in the laboratory was occasionally run in connection with the OTF equipment, its use was intermittent, and it was possible to arrange for hologram exposure and sensitive measurements to be made when the compressor was idle.

The laboratory itself was heated by electric radiators, which were operated by a room thermostat set to 20°C. Temperature fluctuations were thereby reduced to less than $\pm 2^{\circ}$C, over the course of several months, and for shorter periods, the temperature could be regarded as constant.

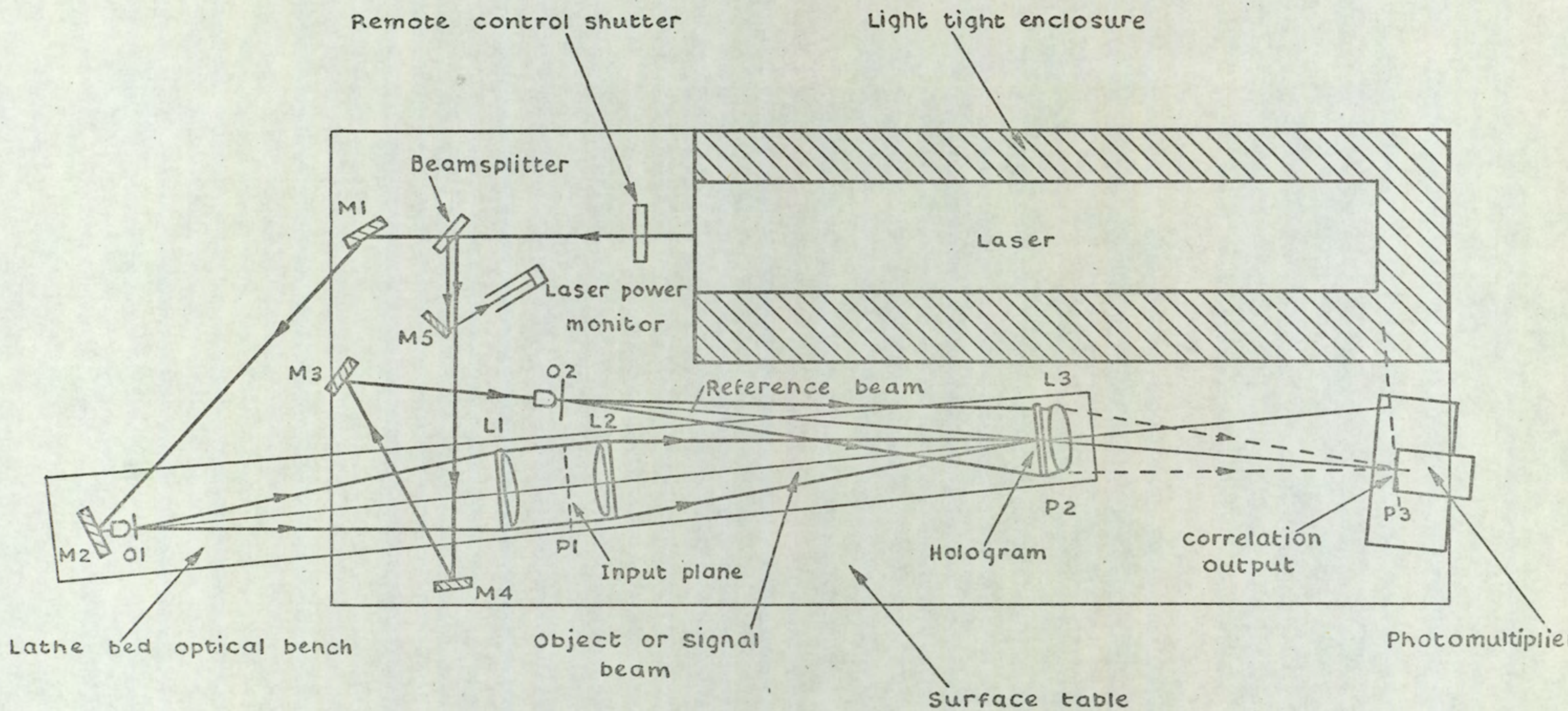
No forced air ventilation was provided in order that air turbulence be kept to a minimum. Draughts from windows were reduced by the use of an X-ray blind, which also enabled the room to be kept permanently darkened. A variety of lighting was available, ranging from fluorescent to darkroom safelight, so that photo electric recordings could be made in relatively dim conditions.

4.2 Basic Design

4.2.1 Existing Equipment

As a starting point the existing equipment was considered. This consisted of:-

1) Steel surface table, 760mm x 2445mm (8' x 3'), providing a very rigid and heavy working surface.



M1 – M5, Plain mirrors

O1, O2, Beam expanders (microscope objectives)

L1, L2, L3, Telescope doublets

Fig. 4.2 Layout of optical system

2) Lathe-bed type optical bench, of length 2000 mm, including bench carriers which were designed to take optical components mounted on pillars 12.7mm ($\frac{1}{2}$ inch) in diameter.

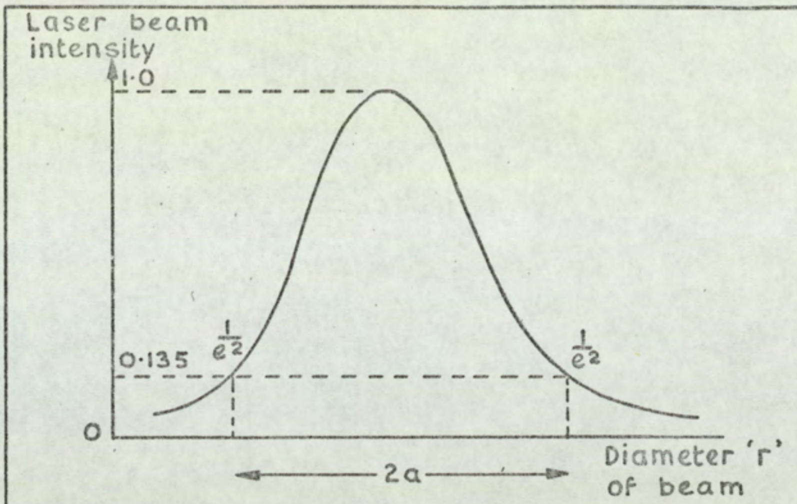
Both of these components proved to be ideal for experimental work; the massive construction of both optical bench and surface table meant that vibrations did not give trouble, and that components remained rigidly in position. The steel surface table was particularly useful in that magnetic clamps could be used; the use of these will be seen later. Because of the vibration free location of the laboratory, it was considered unnecessary to fit anti-vibration mountings to the feet of the steel table. In the event, no trouble was ever experienced from this cause.

The lathe-bed optical bench was used on the surface table, with one of its ends overhanging, giving a total working length (bench + table) of 2770 mm.

In the previous sections (4.1.1) we have seen that an overall length of 3150mm was required. In order to keep the system mounted entirely on one surface, the optical path was folded, so that the light was reflected through 180°.

<u>4.2.2 Optical Path Geometry</u>

Fig 4.2 shows the arrangement of optical paths in the interferometer and transformer. The optical bench was tilted with respect to the surface table to permit scanning of the autocorrelation plane P3. The reference beam was provided by use of a beam splitter (mounted on the steel table with a magnetic clamp), and was folded by plain mirrors M3 and M4 before being expanded by a microscope objective O2. The folding of the reference beam was necessary in order to keep the different paths of the interferometer equal in path length, and hence give good contrast fringes at the hologram. The mirrors M1 and M4 were mounted on the steel table with magnetic clamps;



Laser beam intensity profile given by

$$I(r) = \frac{2P_0}{\pi a^2} e^{-2r^2/a^2}$$

where a is radius at $\frac{1}{e^2}$ or 0·135 x maximum intensity.

Beam diameter = 2a = 2 mm.

Fig. 4.3 (Derived from Ref. 4)

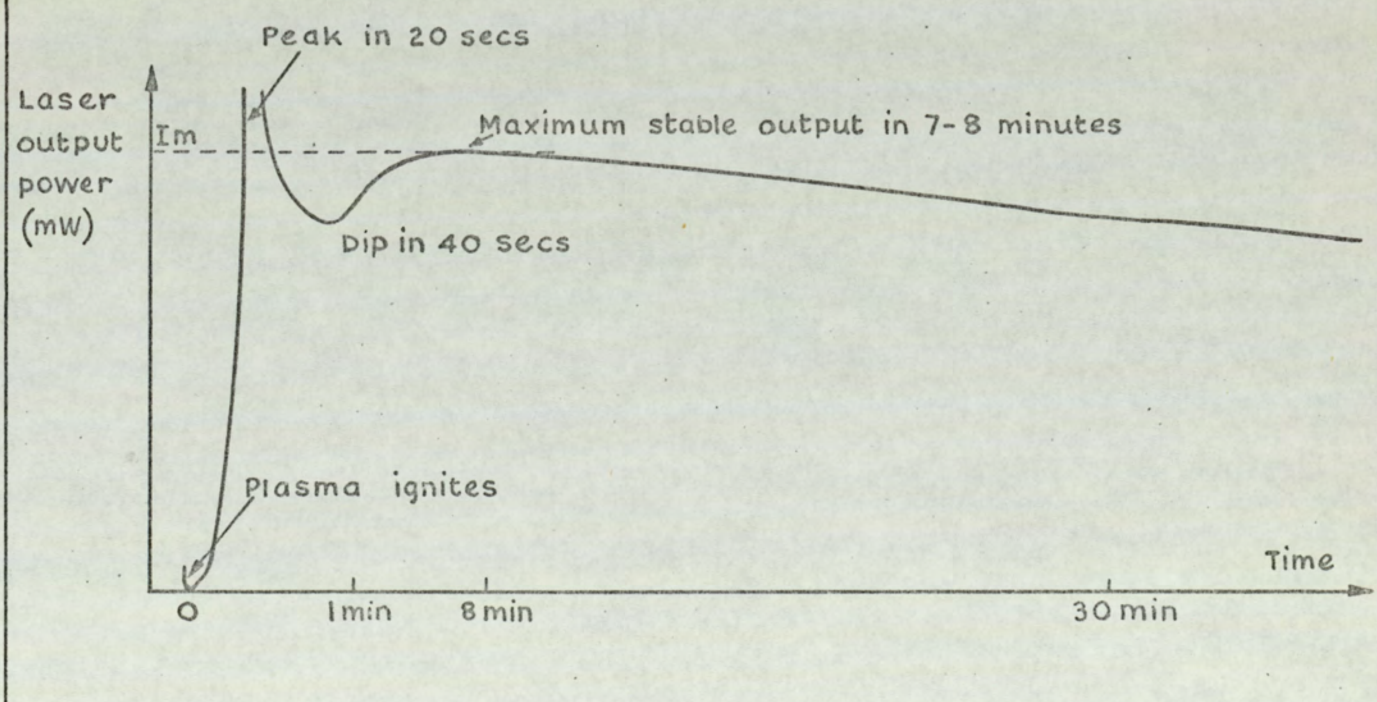


Fig. 4.4 Variation of laser power with time

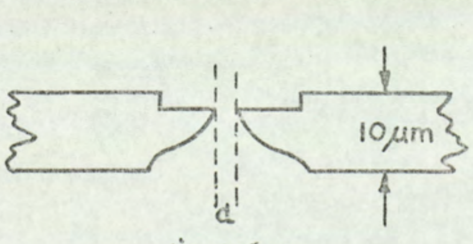


Fig. 4.5 Cross section of pinhole for spatial filtering

M3 and M2 were mounted on the optical bench, whilst objective O2 was mounted on a small triangular optical bench, resting on the surface table.

This arrangement permitted lenses of focal length ≈ 600 mm to be used in the transformer in place of L1 and L2, and also enabled a lens of focal length ≈ 300 mm (L3) to form the correlation plane inside the surface table length. The correlation plane could be examined visually or photographically or by photomultiplier, all apparatus being mounted on the clear area of surface table at the end of the optical bench.

There was some 2000 mm of space available for a laser (6-7 feet) which meant that fairly long cavity lengths could be considered.

4.2.3 Laser

The use of a Helium-Neon gas laser in previous work on holography and character recognition had shown that whilst these light sources were in some respects ideally suited for use in coherent optical systems, their power outputs were rather low, and this could be troublesome if large input formats have to be considered. For this reason, a high power continuous wave gas laser was obtained (Spectra Physics Model 125). This laser gave an output in excess of 50mW at a wavelength of 632.8 nm. The output was continuous and uniphase, ie: the laser could be operated in the TEM_{00} mode, giving a single output beam with an approximate gaussian distribution of intensity across it. The beam diameter was given as 2.0 mm between $\frac{1}{e}$ 2 intensity points (Fig 4.3) and the beam divergence as 0.7 m rad according to manufacturers data.

The overall resonator length was 1800 mm, which meant that the laser fitted well into the space provided on the surface table. The laser was fairly heavy (90 lb) and required no clamping to the surface table. It rested on large brass blocks to raise the laser beam to a convenient working height (330 mm above surface table).

An important requirement of the laser is that its output power and wavefront intensity profile be constant during the period of experiment. This is particularly important during exposure of the hologram, where the intensity must be carefully controlled.

The output power and wavefront intensity profile depend to a large extent on the state of 'tuning' of the laser. This refers to the alignment of the end reflectors of the resonating cavity, optimum power being obtained when the reflectors were perfectly aligned. This also gave the best beam intensity profile (see Fig 4.3), in some cases, misaligned reflectors caused a different profile to be formed. In practice it was found that the laser beam intensity profile differed considerably at times from the Gaussian distribution, and frequently contained irregularities in the form of relatively dark blotches which could be reduced by adjustment of the end reflectors. The output power of the laser was recorded over several periods of up to eight hours in length. The initial power variation after switching on is shown in Fig 4.4.

The initial variation of output power occurs in the first minute, followed by a slow warm upto maximum stable power in about 7-8 minutes. A period of 10 minutes was usually allowed for the laser to warm up. The power output usually decreased very slowly after this point, reaching a 10% reduction after 2-3 hours. The power could be increased by 5-10% at this point by tuning, which implied that temperature changes in the laser caused the mirrors to become misaligned.

In addition to this, variations in output power of the order of 1% occurred, which was superimposed on an apparently slow drift. On occasions a large, higher frequency oscillation of power occurred, of amplitude 5% and period 15-30 secs. This was usually caused by instability in the exciting unit, and could be rectified by adjustment of a potentiometer setting.

Plate 1. Mirror mounting

Plate 2. Beam splitter mounting

The laser was excited by both DC and RF power supplies; the DC excitation provided most of the power, whilst the RF added about 20% power and tended to suppress plasma noise. Measurement of output power showed that the power variations observed above were also obtained with DC excitation alone and that there was a tendency to a smoother variation of power with time.

Study of the laser beam intensity profile was made after the collimating stage and will be discussed later.

[N.B. Towards the end of the research programme the laser was serviced and fitted with a new plasma tube, having various technical improvements. The effect of these were to enable 70-80 mW to be obtained continuously after about 60 minutes warm up time, with virtually no further variation in power. The laser beam intensity profile was also found to be improved, the dark blotches mentioned earlier being absent.]

4.2.4 Optical Components

The optical components of the interferometer and transformer can be divided into lenses and mirrors;

a) Mirrors and beamsplitters All the plain mirrors M1-M4 shown in Fig 4.2 were surface aluminised, optically worked glass, flat to $\lambda/4$. The overall size was 38mm dia ($1\frac{1}{2}$") and each mirror was mounted in an aluminium alloy block fitted with three point mounting of which two points were adjustable, enabling the reflected beam to be easily pointed in any given direction. Each assembly was mounted on stainless steel pillars 13.5 mm dia (0.538") and is shown in plate I.

The beam splitter consisted of a circular parallel glass plate 25.4 mm in diameter and 6 mm thick, with optical surfaces worked flat to $\lambda/4$ and parallel to 3 secs of arc. The beam splitter was mounted in a brass block attached to a stainless steel stem so that it could be notated about a vertical axis. Approximate pointing of reflected beam was done using the adjustment on the magnetic clamp holding the beamsplitter stem. The beam-aplitter is shown in Plate II.

b) Lenses The objectives 01 and 02 used for expanding the laser beam were microscope objectives. In the case of 01, the objective was a Cooke metallurgical objective of 15 x power. This objective was used because it was corrected for infinite conjugate, and would therefore bring a parallel beam to a focus with minimum aberratic spread. With 15 x power the 2 mm diameter beam was diverged sufficiently to cover an input of 6 cm square at 60 cm without too much intensity variation. The reference objective 02 was considerably weaker, since although the distance to the hologram was still 60 cm, (as for 01) the area to be illuminated was considerably smaller. In practice, a variety of lenses were used, ranging from 76 mm focal length to 8 mm focal length, depending on the experiment.

Both objectives were used with spatial filters to reduce the irregular intensity variations in object and reference beams caused by dust, scratches, etc, on preceding optical components. The spatial filters are of the type referred to in Chapter 2 (section 2.6.1), consisting of a circular hole drilled in thin opaque sheet. The diameter of the hole was selected to match the size of the Airy disc formed at the focus of each objective.

This task was complicated by the fact that the laser beam was Gaussian in shape, and so the intensity at the focus was also Gaussian, with a radius $a_f = \frac{\lambda}{\pi a} \cdot f$ (4) where a is the original laser beam diameter, and is defined by points where the intensity is $\frac{1}{e}2$ (or 0.135) times the central intensity and f is the focal length of the beam expanding objective. The size of filter used was a compromise between allowing enough light to be transmitted and restricting the irregular noise terms, which lay outside the radius a_f. The reference (4) gives a table of filter diameters for different focal lengths and input beam diameters. This table suggests a filter of about 15μ diameter for the object beam objective, and in practice a filter of 10μ diameter gave good results. [N.B. The value of a_f here

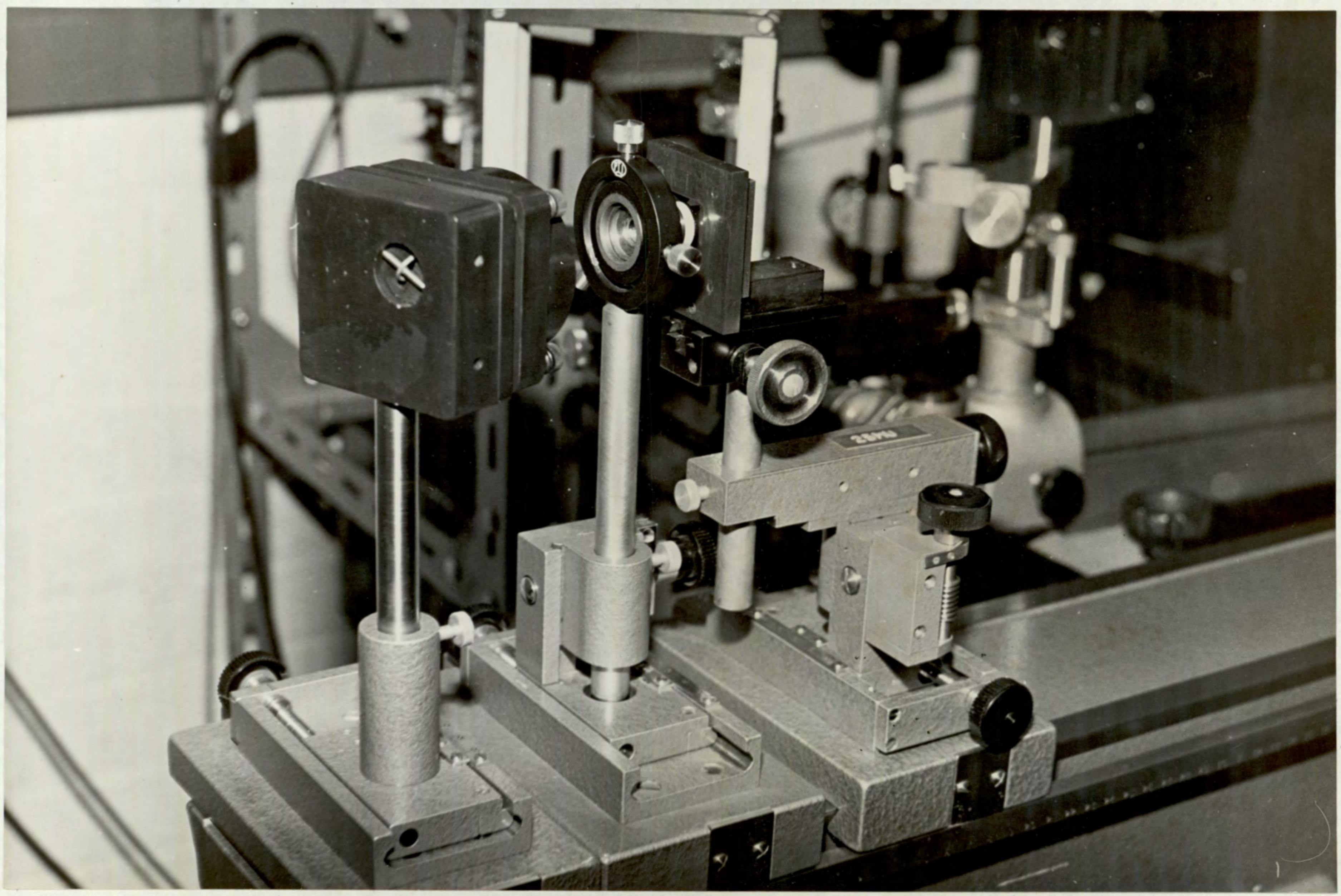


Plate III Object beam expander and spatial filter mounting

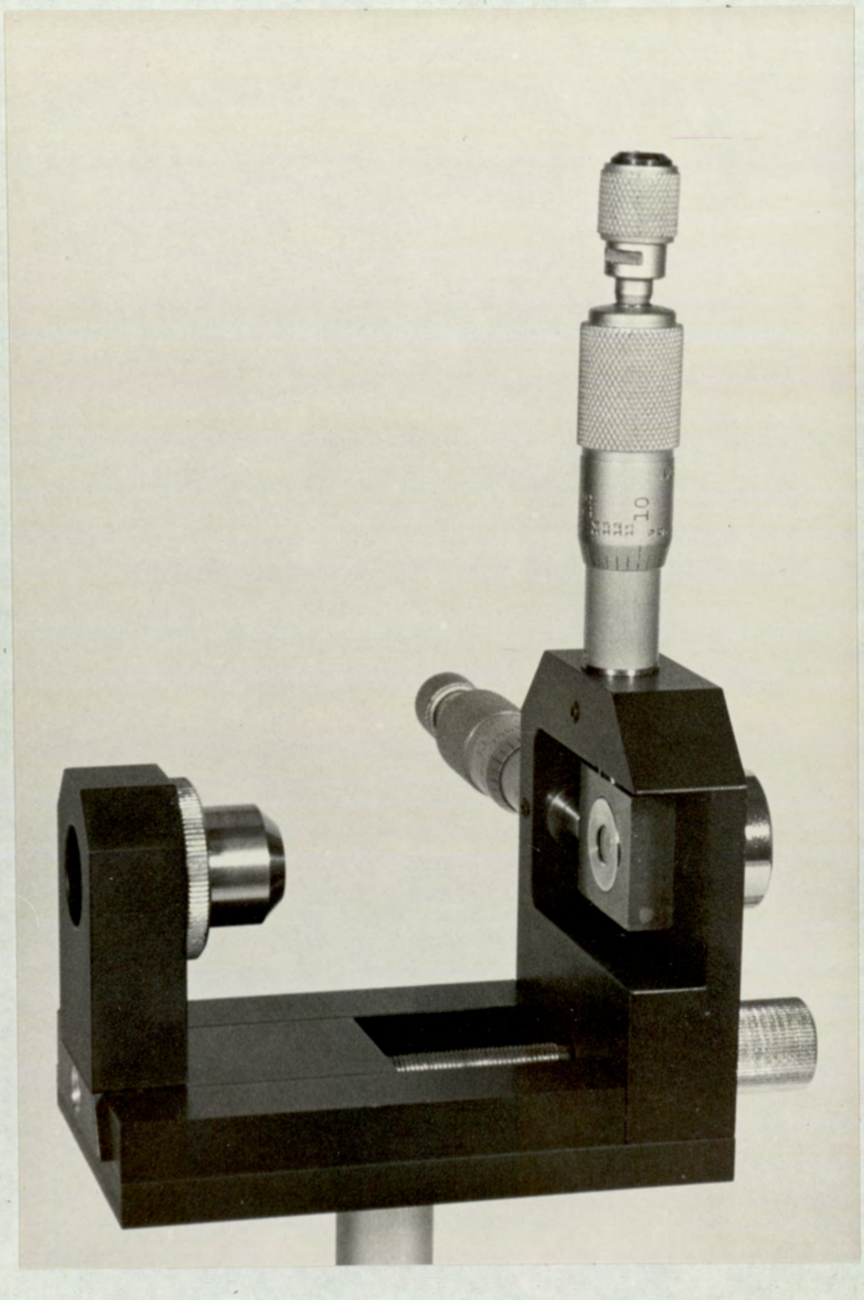


Plate IV Reference beam expander and spatial filter
[Courtesy of Sciencifica & Cook Electronics Ltd]

gave a diameter of 7.5μ, and the reference suggested filters of twice this diameter.] The alignment of the filter was extremely critical, and some modification was made to the existing vertical movement of the optical bench carriers to enable fine adjustments to be made. A standard transverse slide attachment was added to give movement in the axial direction, as shown in Plate III.

The filter itself was one of a set of precision pinholes supplied by Ealing Optics Ltd. These were made in Nickel Stock, 10μ thick and mounted as shown. The cross section of a pinhole is shown in Fig 4.5, based on Manufacturers' data.

The spatial filter for the reference beam was again a purchased item, except that here the lens and filter were mounted on one assembly for ease of adjustment. The filter was mounted on a magnetic block, which clamped the filter to the xy face of the filter body. The position of the block was adjusted by means of micrometer screws, and the objective was focussed by an axial screw passing through the body. The spatial filter was about 30μ in diameter, being sufficiently small for the longer focal length lenses. The arrangement is shown in Plate IV.

The collimator and transform lens L1 and L2 have very similar tasks, although L2 was required to give good performance over the range of spatial frequencies likely to be encountered, whereas L1 was required to image an axial point source. Since the fields over which both lenses were to work were small, and since spherical aberration is likely to be the most important aberration, the task called for telescope objective type lenses. The lenses were in fact uncemented doublets, each having a nominal focal length of 670 mm and a diameter of 63 mm. They were supplied by Broadhurst Clarkson & Co, and were corrected for use in telescope systems. This suggested that performance on axis was likely to be very good, whilst performance a few degrees off axis was likely to be poor. This was confirmed by examination, as shown later.

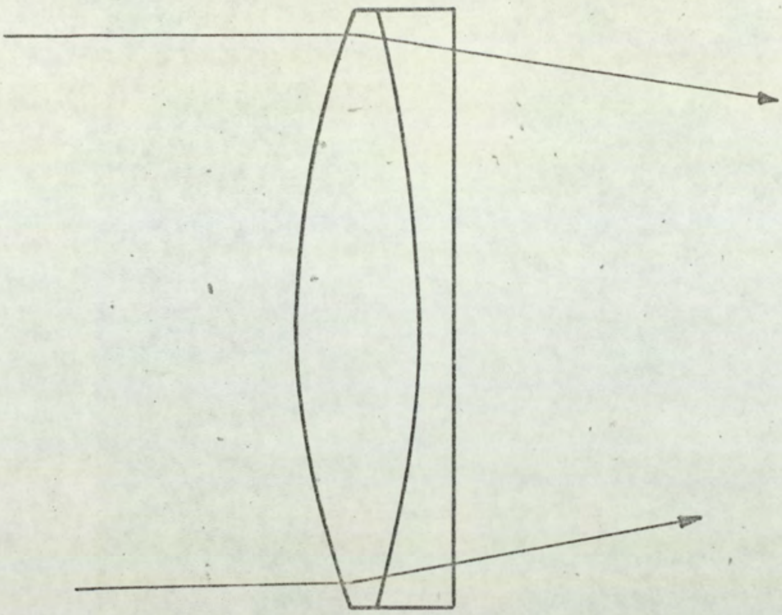


Fig.4.6 Construction of telescope doublet lens (L1 & L2) focal length 66·19 & 66·34 cms

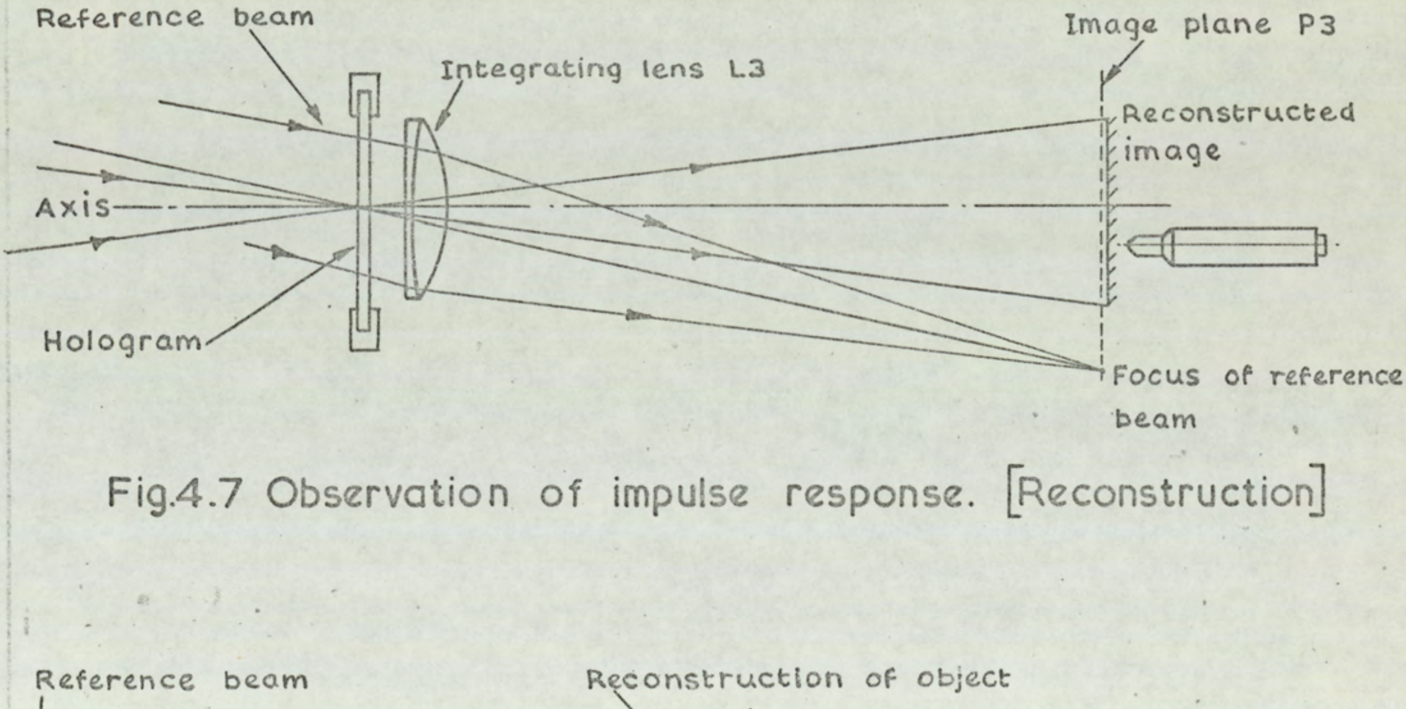


Fig.4.7 Observation of impulse response. [Reconstruction]

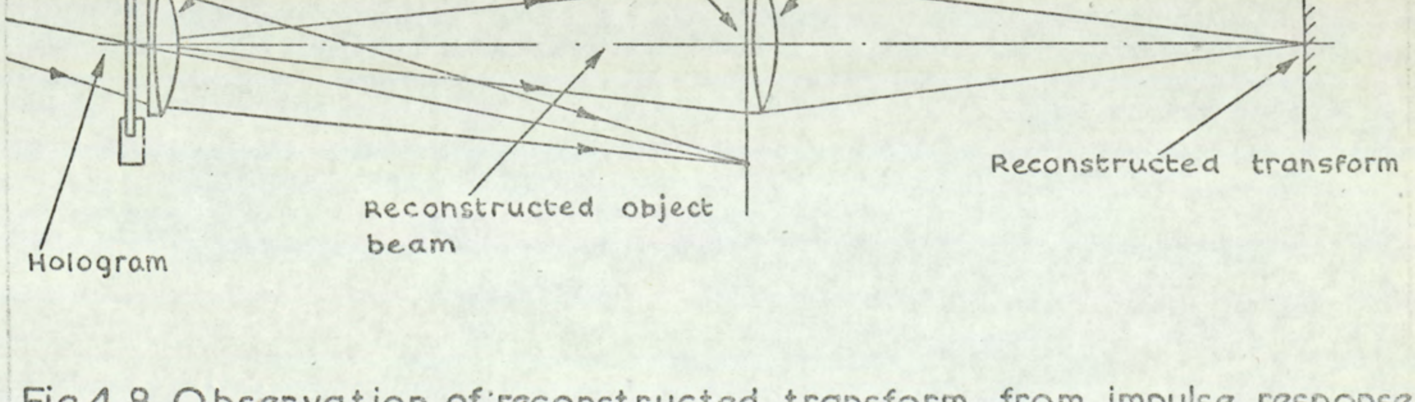


Fig.4.8 Observation of reconstructed transform from impulse response

The construction of the lens is shown in Fig4.6; the lenses were mounted with their curved faces towards each other.

The integrating lens L3 had to image the reference point source (or its reconstructed image, in the case of autocorrelation) and had to work at finite conjugates. The lens also had to cover a field at least equal to that subtended by the input plane at the lens. On these particular grounds a copying lens would be ideal, but the focal lengths of available lenses tended to be too short, giving a small correlation function and output plane.

As an alternative a Zeiss Tessar lens of 19 cm focal length and f/3.5 aperture has been used. This lens gave good correction over a wide field, and for the conjugates mentioned, but had the disadvantage that its several components caused multiple reflections of the laser beam and tended to be confused with the required output. Both the copying lens and the Tessar have the disadvantage of multiple element lenses in that scattering by dust particles on the numerous surfaces will obscure detail in the reconstructed image or autocorrelation plane. For this reason most of the lenses in the optical system were kept as simple as possible.

Finally a lens was chosen which although not strictly ideal, has given good results. This was a cemented doublet telescope objective of focal length 35.2 cms and 5.4 cms in diameter. Such a lens was of sufficiently large aperture to collect light from a fourier transform hologram extending to $\sim$ 60 lines/mm. Although this was lower than the previous figure of 100 lines/mm, it still included most of the information normally contained in photographic transparencies likely to be used. The lens was intended for telescope systems and consequently gave good imagery when the reference source was close to its axis, and so the lens was always used with its axis aligned with that of the reference beam, as shown in Fig 4.2. Its construction was similar to the uncemented doublet, and its curved face was turned towards the longer conjugate, ie: the reference source.

The collimator and transformer lenses were mounted in lens holders which could be rotated about the two axes passing through the vertical lens plane. In this way the lenses could be aligned with the source and with each other. The integrating lens was mounted in a holder attached to the hologram plate holder, discussed below.

4.2.5 Hologram Holder

One of the most important components in the optical system was the hologram holder. The holograms were recorded on photographic plate, and before being used in the system, must be processed and dried. This required that the plate be capable of being removed and replaced in the system in exactly the same location as that used for exposure.

Previous work at the RAE on holography had shown that a good solution was to use stainless steel frames, into which the photographic plate could be inserted. The frames were themselves kinematically located in a large robust holder mounted to the optical bench; this permitted the plates to be removed and replaced whenever necessary.

For practical reasons, stainless steel frames were made, to accept photographic plates measuring 40 mm x 66 mm, with an unobstructed aperture of 54 mm x 33 mm. Now this restricts the spatial frequency recorded by the hologram to 16.5 mm or 39 lines/mm. This decreases the range of frequencies accepted still more than the intergrating lens, but for experimental purposes, was considered adequate.

The plate was clamped in the frame by pins on either side the aperture, the pins being depressed by small thumb wheels. The frame was fitted with spring loaded ball bearings for positive location in the vee grooves of the plate holder itself.

The holder consisted of two components, one being a fixed brass body mounted on a stainless steel stem, the other a steel holder which was free to rotate within the brass body. Adjusting screws were provided for locking and gradual rotational movement. The steel holder contained vee grooves for

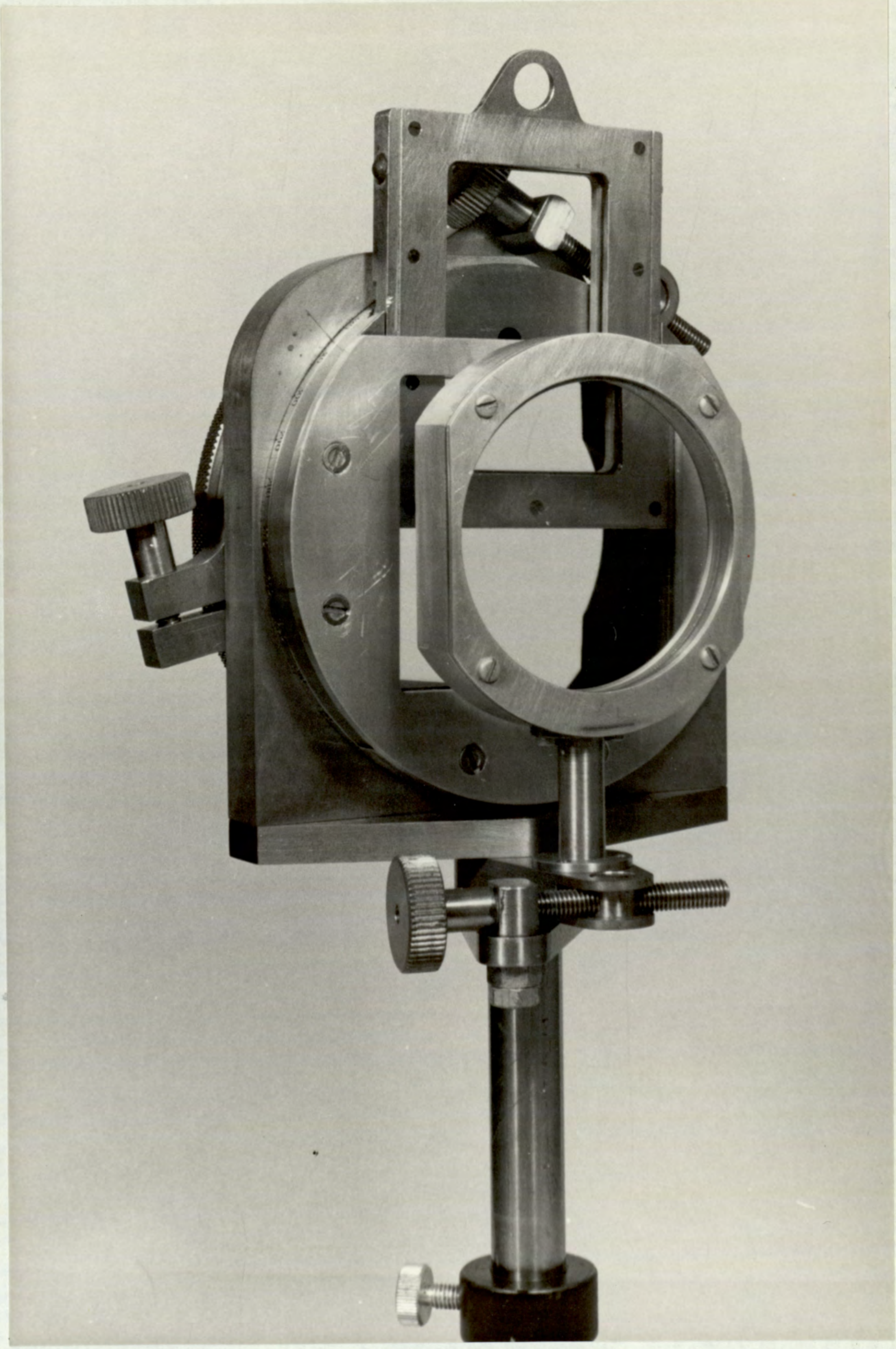


Plate V Hologram plate holder and stainless steel frame – viewed from autocorrelation plane accurate location of the stainless steel frame in 'x' and 'z' directions, whilst the bottom of the holder located the frame in the 'y' direction.

The body of the holder also carried a bracket on which was mounted the integrating lens holder. This was arranged to rotate about a vertical axis passing through the front face of the photographic plate, and enabled the lens to be rotated about the hologram plane. The complete holder and frame is shown in Plate V. The inside faces of the holder were painted matt black to eliminate reflections.

4.2.6 Input devices

Two types of input were used; one type consisted of simple apertures cut in metal sheet, such as circular holes, squares or slits. The other type consisted of apertures or images recorded on photographic plate.

a) Simple Apertures: As remarked, these were often simply holes cut in the metal sheet, but it was also found that adjustable slits and rectangular apertures were of great use. These apertures were supported in standard optical bench fittings; the adjustable slits being standard optical accessories.

b) Photographic Inputs: With these inputs the image was recorded on photographic plate, and the effect of the film and glass must be considered, in addition to the density of the recorded image. Small variations in the thickness of the photographic emulsion were found to form unsymmetrical fourier transforms, an example is shown in Chapter 6, together with the original input transparency.

These variations can be eliminated by immersing the transparency in a refractive index matching liquid, so that only density variations are effective.

Previous experience with fourier holograms had shown that a suitable liquid was Decahydronapthalene, or "Decalin", which has a refractive index of 1.48. This liquid was volatile and had low viscosity, so that air bubbles did not form, as was the case when liquid paraffin (refractive index = 1.44) was used.

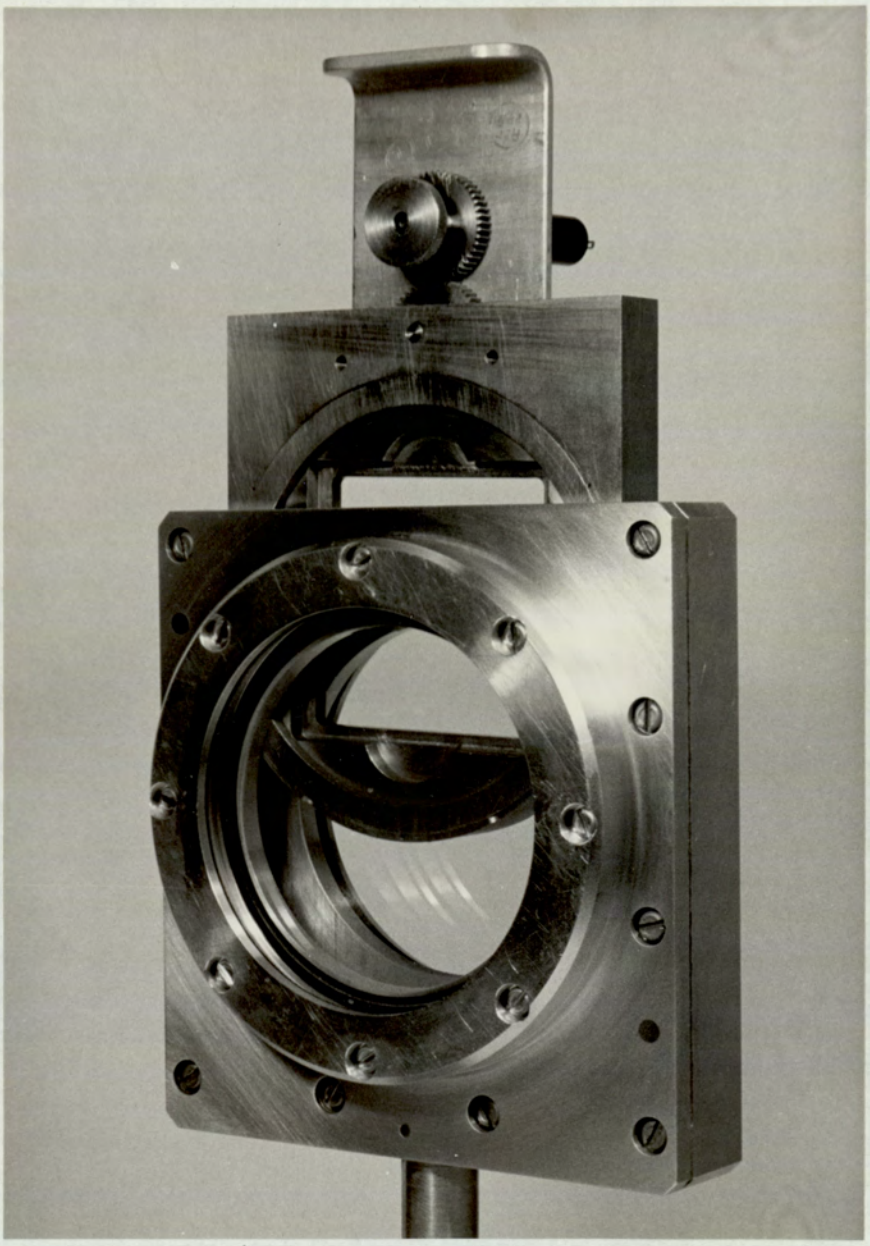


Plate VI Liquid cell and transparency carrier
(for refractive index matching)

A liquid cell was designed, having circular glass windows 75.3 mm diameter by 9.5 mm thick. The windows were optically worked to $\lambda/4$ on each face, and parallel to 3 secs of arc. A check on a Twyman Green interferometer confirmed that they were within the former specification. The windows were held to the body of the cell by clamping rings, the liquid being sealed by 'O' rings pressing against the glass windows. The cell itself was made of aluminium alloy. The photographic transparency was inserted into the cell by means of a carrier, which permitted plates or film of up to 6 x 6 cm square to be used. The carrier contained a rotating component, on which the transparency was held, so that the input could be rotated from outside the cell. A helical potentiometer was attached to the carrier so that rotation of the input produced a continuously varying voltage, enabling the position of the input to be monitored. The liquid cell and carrier are shown in Plate VI. The effect of the complete liquid gate upon the wavefront shape was measured using a shearing interferometer and will be discussed later.

4.3 Apparatus for observing hologram response

There were two situations under which the hologram output could be observed, one of these being when the hologram was illuminated by the reference beam, the other when the hologram was illuminated by the input. The latter was the normal mode of operating the system.

Both situations could be used to investigate the performance of the hologram, and of the system as a whole. Apparatus had been designed which permitted either response to be studied, the main requirement being that it had to be rapidly and easily changed from one set up to the other.

4.3.1 Impulse Response

The impulse response was obtained when the hologram was illuminated by the reference beam, the reference point source corresponding to a delta function.

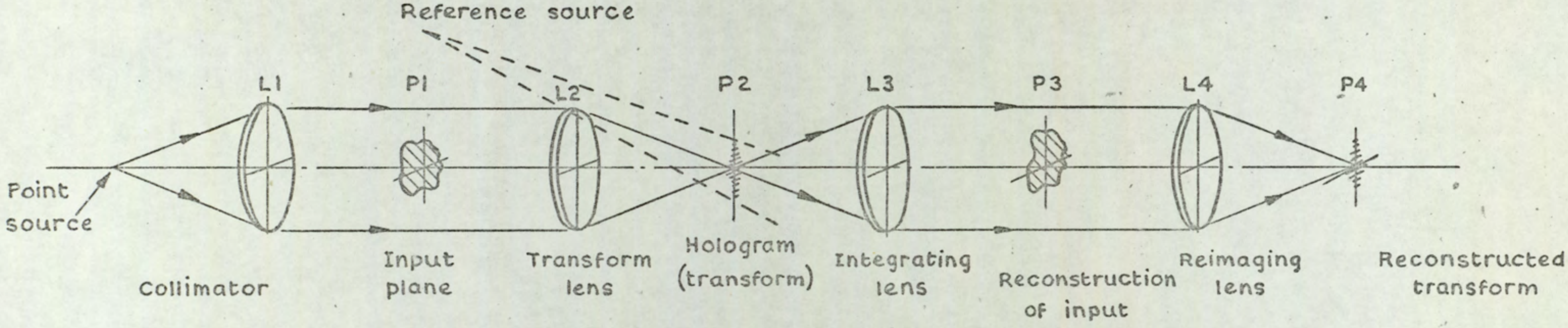


Fig. 4. 9 Observation of reconstructed transform (schematic)

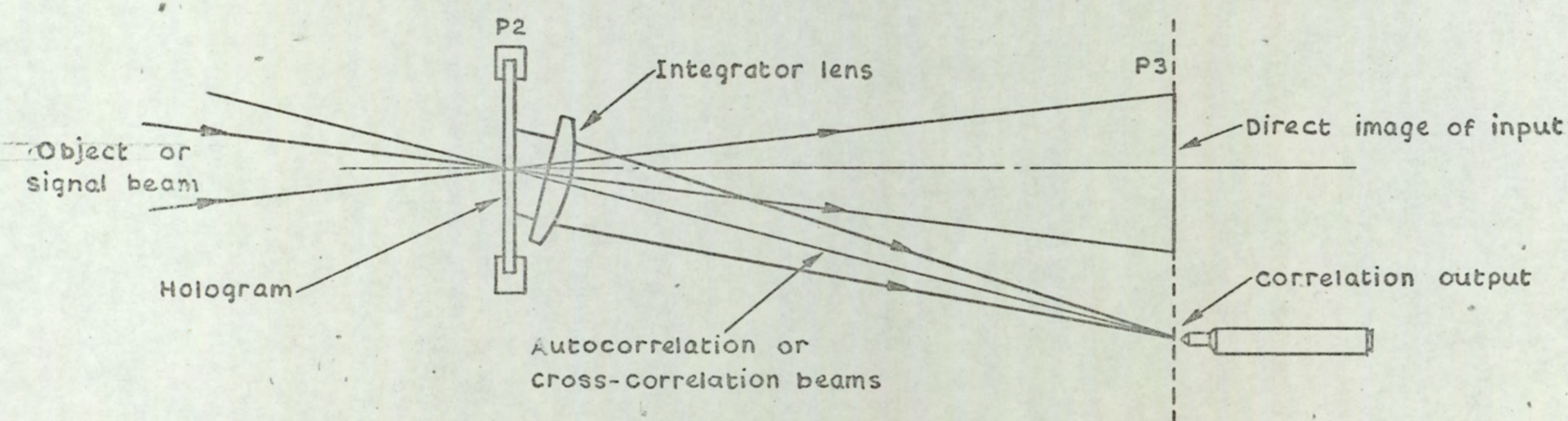


Fig. 4. 10 Observation of correlation response

Three terms could be seen, as shown in Chapter 3 (section 3.2.2), the image of the function being formed on axis. It was this image which was of interest, since it contained the information which had been actually recorded in the hologram. Comparison of this image with the original input allowed the effect of errors in the transforming and recording system to be observed.

The image itself was observed visually, with a low powered microscope, or by recording it photographically. This latter had the advantage that the speckle pattern formed by coherent light was not so confusing. The optical arrangement is shown in Fig 4.7.

In some cases it was desirable to examine the precise effect of the photographic process when recording the fourier transform. This examination was made by forming an image of the fourier transform, using only the reconstructed beam. The optical arrangement is shown in Fig 4.8.

The reconstructed image of the input was formed in the usual way, by illuminating the hologram with the reference beam alone, ie: impulse response. The reconstructed beam was collected by a lens L4, placed in the image plane, and formed a convergent beam which was focussed in plane P4. (see fig 4.8). This focus was conjugate with the hologram or transform plane, and if the lens L4 were sufficiently large, would correspond to the image of the hologram. However, the lens L4 accepted only the reconstructed beam, and so fine detail of the hologram was not imaged. The image corresponded to the fourier transform of the reconstructed image, and was called the "reconstructed transform".

The reconstructed transform differed from the original transform in that the effects of the photographic process would be present, and comparison between the two transforms would enable these effects to be studied.

The way in which the reconstructed image was formed can be seen by reference to Fig 4.9, which shows the whole of the optical system.

The figure has been 'expanded' to make the arrangement easier to understand. Note that the image in P3 and transform in P4 could be formed by the object beam alone (no hologram) or by the hologram's impulse response.

The reconstructed transform was studied by microscope or by photography, and in some cases observed by use of a scanning photomultiplier, which measured the intensity variation across the transform plane P4. This device will be described in detail later.

4.3.2 Correlation response

The correlation response was observed when the hologram was illuminated with the fourier transform of an input transparency. The transparency could be identical to that used in the making of the hologram, or it could be a transparency which was to be correlated with the hologram. The response in the first case would correspond to the autocorrelation function, in the second case to a cross correlation function. If the transparency was not identical to that used in making the hologram but contained a similar pattern, then the cross correlation function will contain an autocorrelation.

The autocorrelation response corresponded to reconstruction of the reference beam and would be formed where the reference beam focus was formed, as shown in Fig 4.7. To avoid the effects of off-axis aberrations, the integrating lens was rotated about the hologram plane, and brought into line with the reference beam. The correlation plane was observed using a microscope or by use of a camera. Fig 4.10 shows the optical arrangement.

Again, it was sometimes necessary to observe the autocorrelation or cross correlation response using a photo multiplier scanning unit. [N.B. The presence of off axis aberrations in the first order image could be avoided by using a prism of shallow angle to refract the 1st order beam along the optic axis. However, the method used here gives greater flexibility and permits different reference beam angles to be used.]

4.3.3 Optical Apparatus for Impulse response

The optical arrangement shown in Figs 4.8 and 4.9 demand the use of an additional lens, which was required to image the transform at preferably 1:1 magnification. It must have a sufficiently large aperture to collect all light contained in the reconstructed image. The ideal lens here would again be a copying lens, but in practice suffered from having either too short a focal length or too many optical surfaces. In this work, a lens similar to that used for the integrating lens was used, being a cemented doublet of 352 mm focal length and 54.0 mm aperture. Since the lens was used over a very limited field (in practice about 5°) the disadvantage of using a telescope doublet was reduced, whilst retaining the advantage of a limited number of reflecting surfaces.

The lens itself was mounted on a magnetic clamp, and fastened rigidly to the surface table. In later experiments, an additional optical bench was used, being supported partly by the surface table and partly by a heavy frame. The demands for rigidity and freedom from vibration were not so stringent as for the first part of the system, where the hologram was to be made.

The optical bench permitted other optical equipment to be mounted, such as viewing microscope and camera.

4.3.4 Photomultiplier Scanning Device

As remarked in section 4.3.1 and 4.3.2, it was sometimes necessary to measure the reconstructed transform or correlation function, by recording the variation in intensity according to position along the appropriate image plane.

This was done using a photomultiplier scanning unit, which permitted a record of the intensity variations to be made for reference.

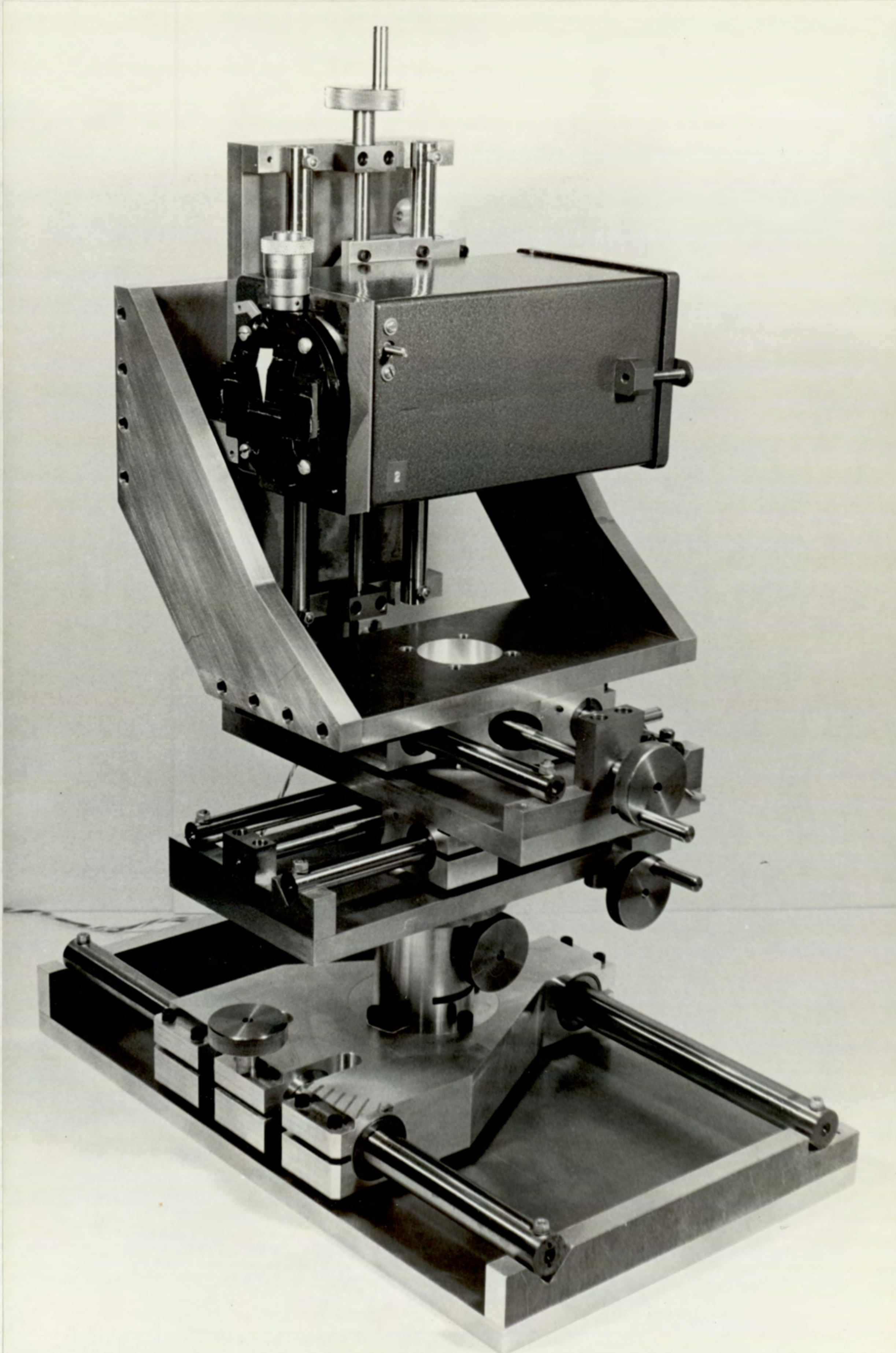


Plate VII. Photomultiplier scanning unit

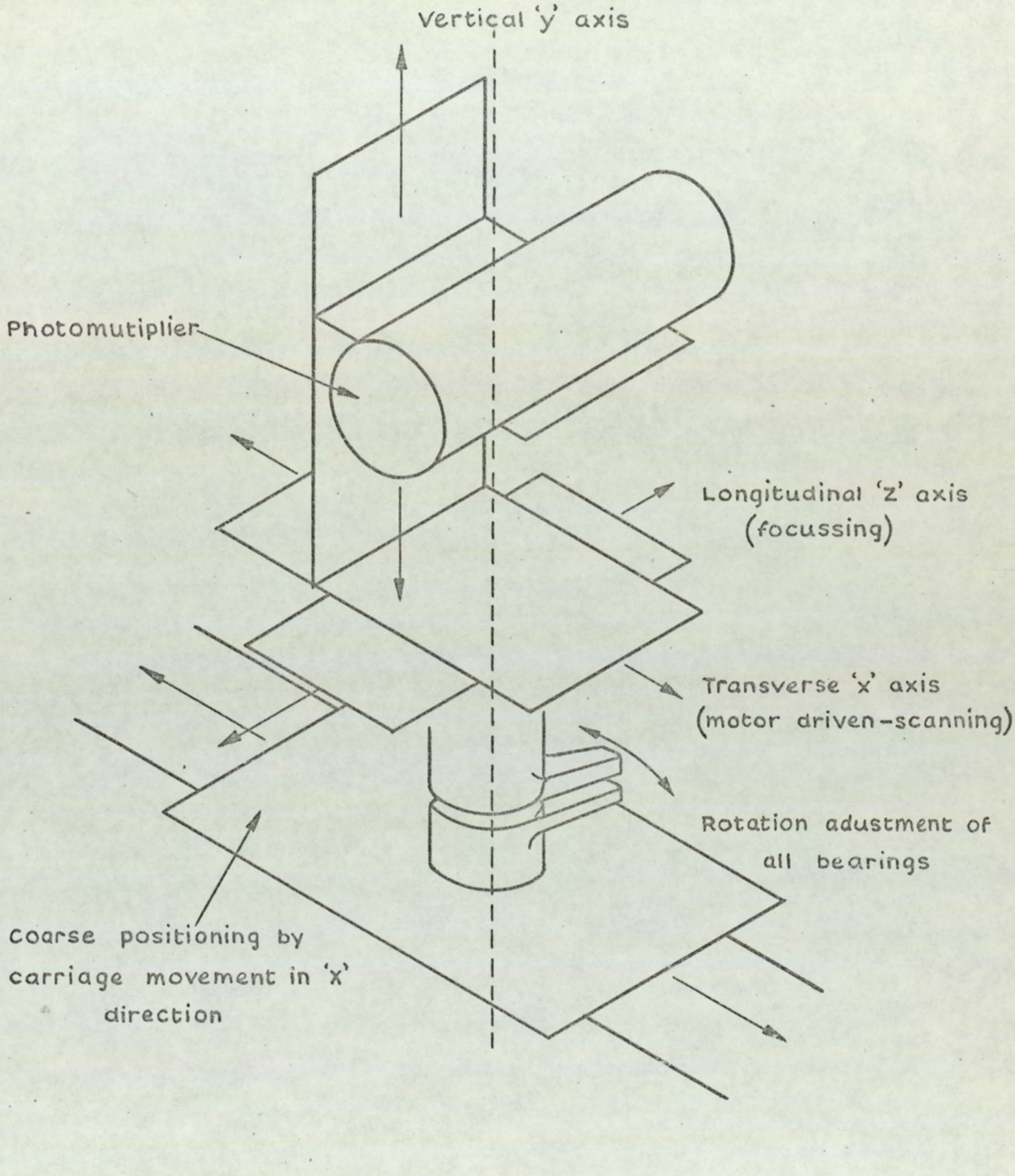


Fig. 4. 11 Photomultiplier scanning unit, showing arrangement of bearings

The scanning unit is shown in Plate VII. It consisted of three sets of linear bearings, mounted at right angles to each other to provide movement in x, y and z directions. A bracket attached to the vertical bearing carried a box containing the photomultiplier.

The lower bearing was attached to a pillar which was imbedded in a large block. This block was part of a large linear bearing and provided coarse positioning of the photomultiplier, in the 'x' or transverse direction. An arm projecting from the pillar and attached to the lower linear bearing permitted the whole assembly of bearings and photomultiplier to be rotated through $\pm$ 10°. The vertical and focussing bearings were fitted with hand driven screws; the transverse bearing was fitted with a screw, but also had a small d.c. electric motor fitted to one end, to give a constant scanning speed.

The arrangement of bearings is shown in Fig 4.11. The bearings were cylindrical bars of steel, mounted on 'dural' plates; most of the remainder of the unit was constructed of dural.

The first three bearings can be regarded as forming a set of co-ordinates, and these could be rotated or moved bodily by the rotation and coarse positioning adjustments provided on the base of the unit.

In this way the photomultiplier could be made to look at any part of an x, y plane, and be accurately focussed on this plane.

The photomultiplier was chosen for its spectral sensitivity, low dark current and high gain. It was an EMI type 9558B, having a tri-alkali (Sb Na KCs) S20 cathode, producing 150 μA/lm. The spectral sensitivity of this cathode extended well into the 0.8μ spectrum from ~0.3μ, and so was ideally suited to detecting light at 0.6328μ.

The photomultiplier was housed in a light tight box, mounted on the scanning unit. The box, also contained the dynode resistor chain, the circuit of which is shown in Fig 4.12 . The front of the box was adapted

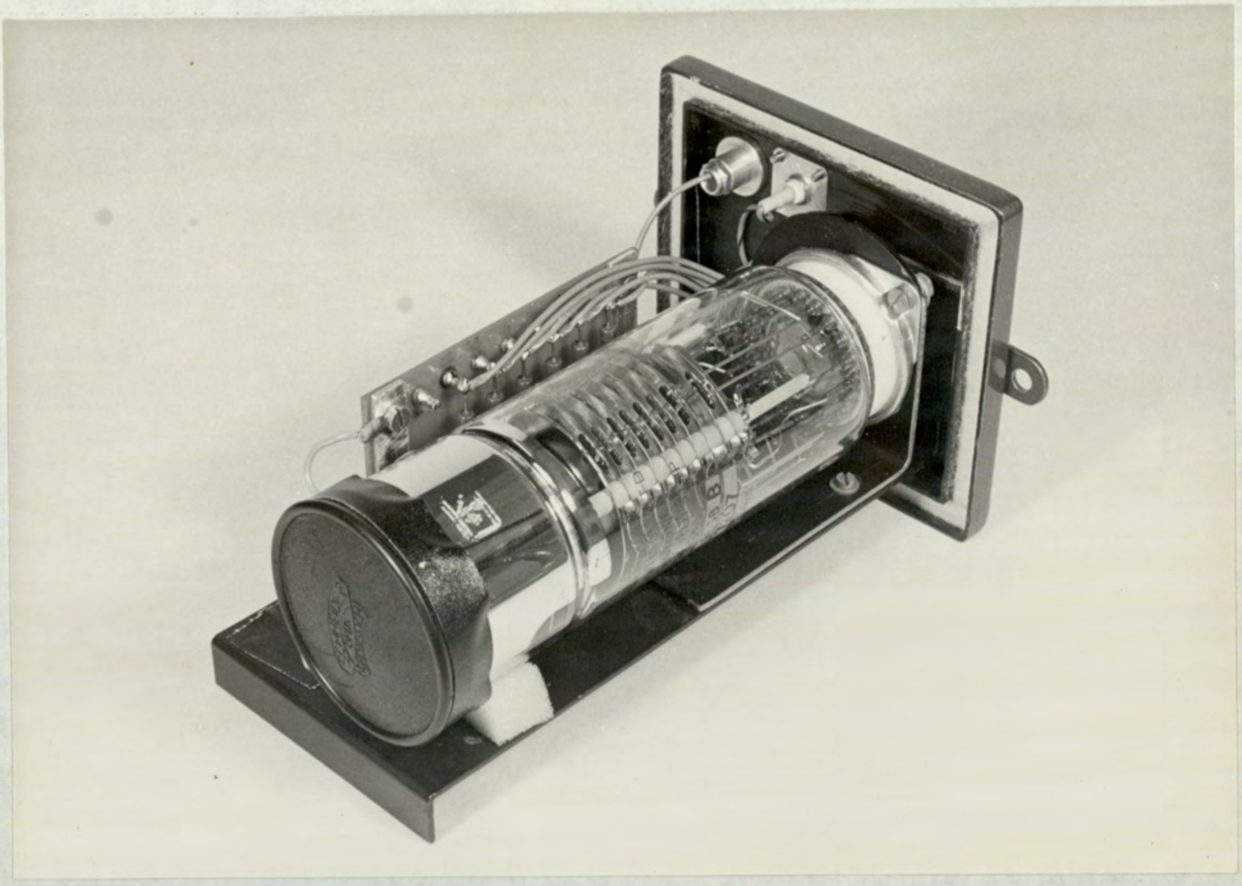


Plate VIII Photomultiplier and chassis

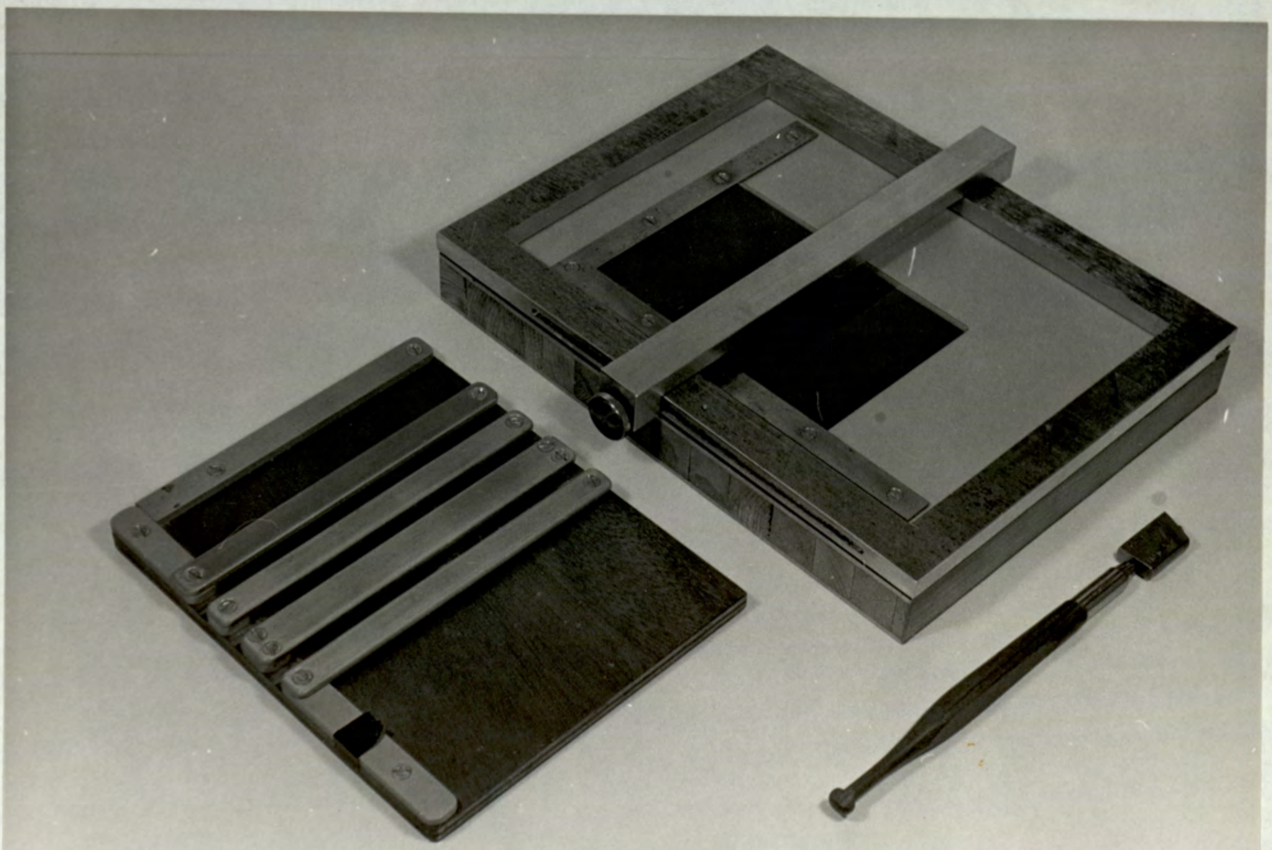


Plate IX Plate-cutting jigs for use in total darkness

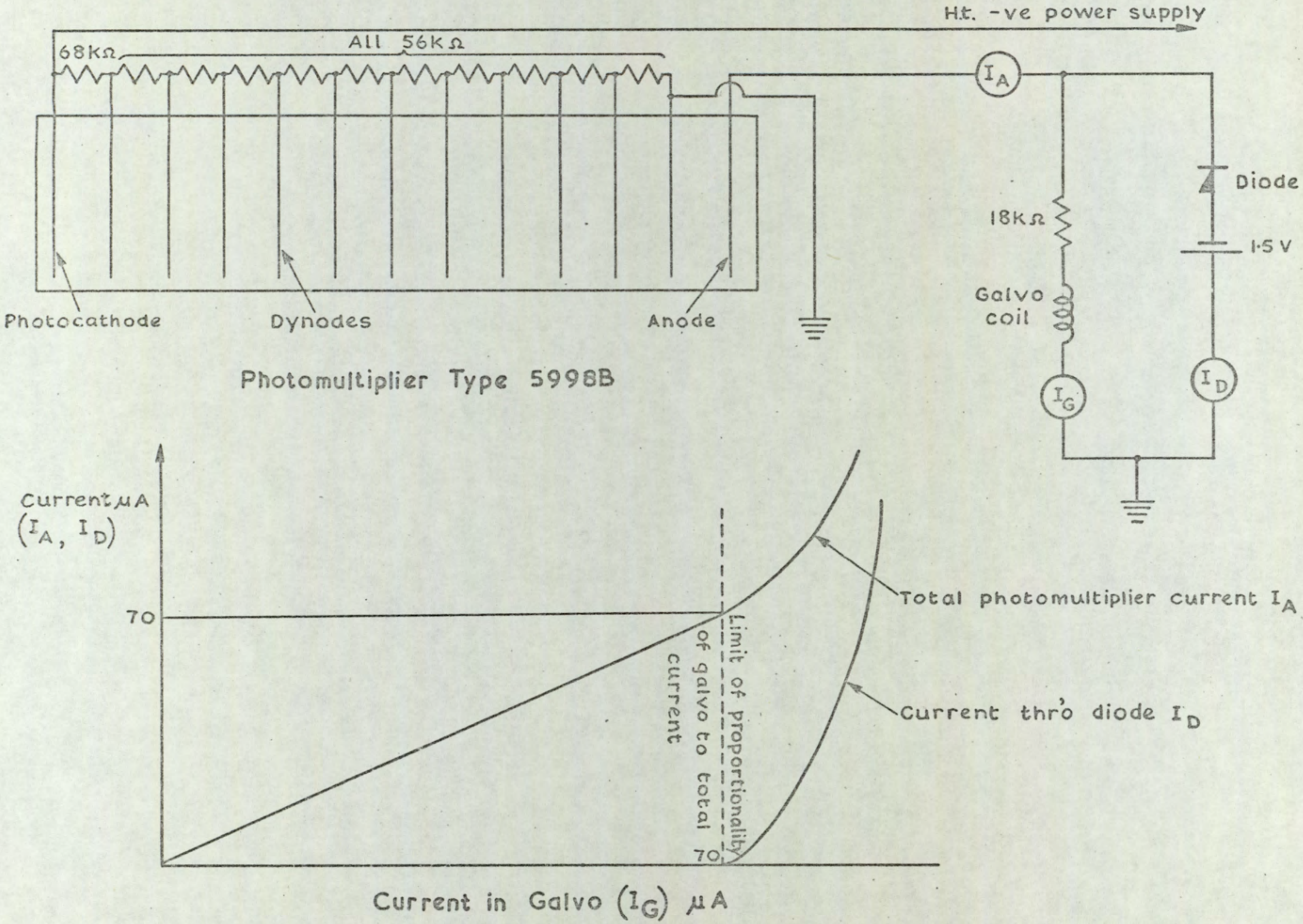


Fig. 4.12 Photomultiplier circuitry

for carrying a precision mechanical slit, although in some experiments precision pinholes described in section 4.2.4 were mounted. Two cables were taken from the rear of the box, which carried the photomultiplier HT voltage and output signal. The rear of the box formed part of a chassis, on which the photomultiplier and dynode resistors were mounted. The chassis could be withdrawn whenever required, without disturbing the rest of the scanning unit. This is shown in Plate VIII.

The h.t. voltage was supplied by a Brandenburg power unit, capable of giving $\pm$ 2000 v d.c., the voltage being taken to the dynode chain by co-axial cable. The signal was taken from the photomultiplier box by co-axial cable, to reduce interference from external electrical devices. The signal could be observed in a variety of ways, the most frequently used method being to observe the deflection of a galvanometer spot. An ultra-violet recording galvanometer was used, having the advantage that an extremely sensitive meter was combined with a recorder, The circuit shown to the right of Fig 4.12 was designed to protect the galvanometers, which could easily be damaged by relatively high current. The diode passed current in excess of 70μA, limiting the galvanometer reading to a maximum deflection proportional to this current. The variation of I_A and I_D with I_G is shown in the graph, inset in Fig 4.12.

In addition to the U.V. recorder, the signal could be accurately measured on a digital voltmeter. Since the voltmeter had its own high impedance ($\approx$ 1MΩ), no anode load was required, and the photomultiplier output was taken directly to the instrument.

Another instrument which was sometimes used, was an X-Y plotter. With this instrument it was possible to show the variation (for instance) of the autocorrelation peak intensity as the input function was rotated. The rotation of the input was converted to a linearly varying potential (section 4.2.6) and hence to the 'X' axis of the plotter.

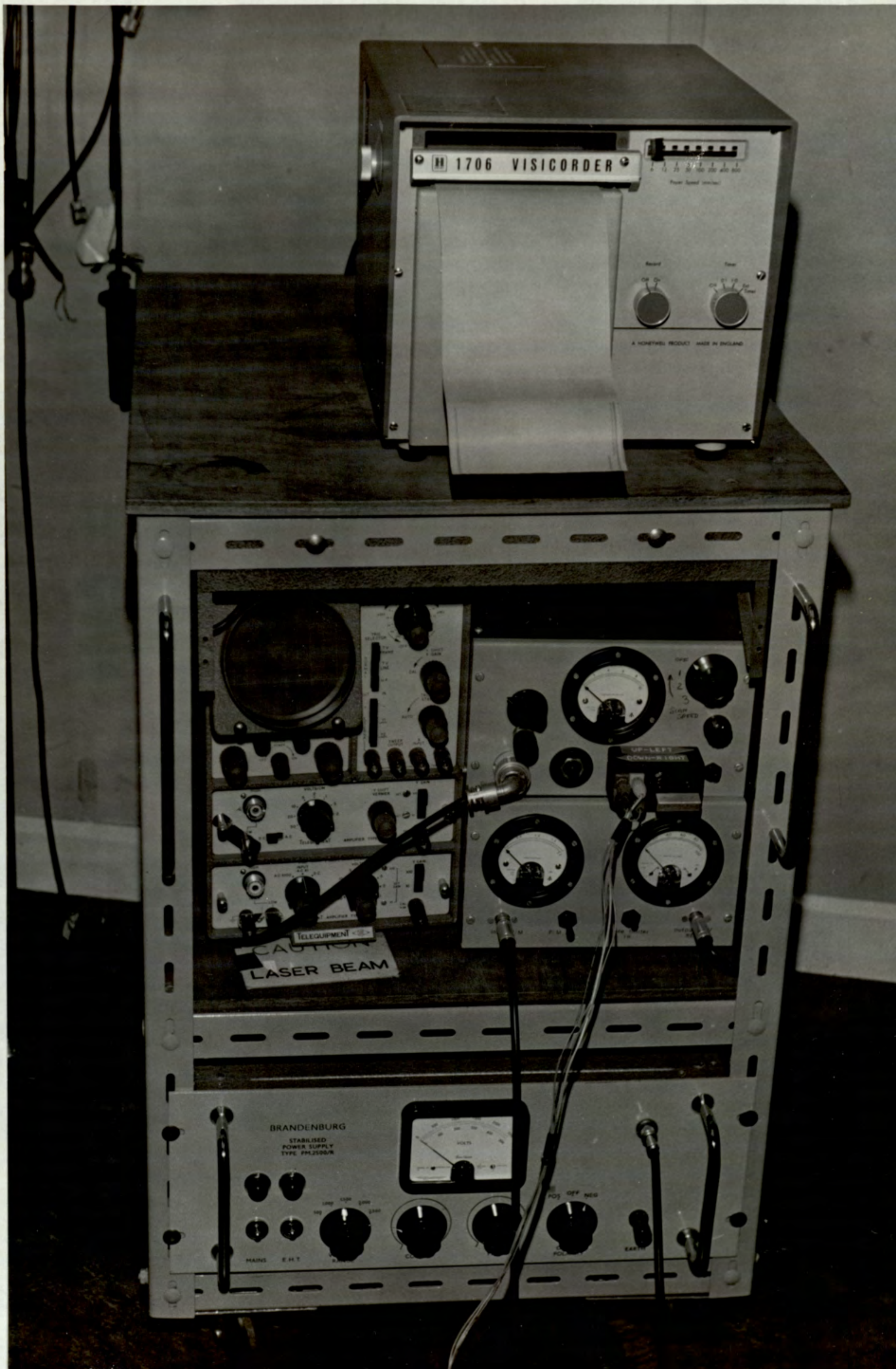


Plate X Recording apparatus

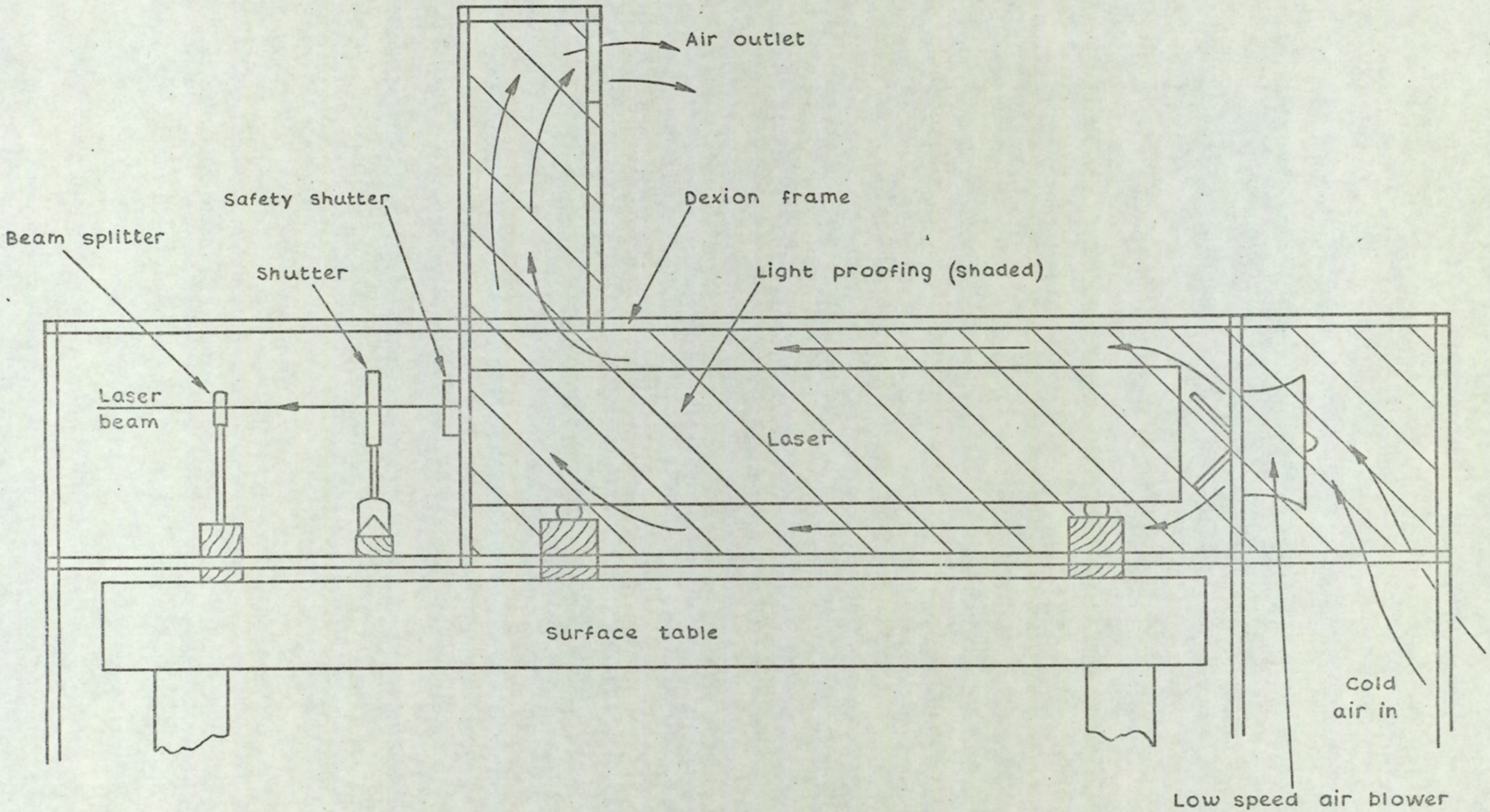


Fig. 4. 13 Light proofing and cooling of laser

A general view of the recording apparatus is given in Plate X.

4.4 Notes on operating techniques

4.4.1 Preparation of photographic plates

The photographic material used in virtually all experiments was Kodak 649F Spectrographic Plate. This material was capable of resolving to 2000 cycles/mm but was also sensitive to the red spectrum, and thus suited for this work. In fact the contrast of the recorded image decreased as the spatial frequency is increased, but at frequencies up to 200 cpmm it was still very nearly unity for this emulsion.

The plates were supplied in standard sizes of $6\frac{3}{4}$" x $4\frac{1}{2}$" (171.5mm x 114.2mm) and had to be cut to a suitable size for the plate holder. This was done by designing a cutting jig, so that the plates could be scored and parted in total darkness. The jig is shown in Plate IX. Each plate was placed in a cellophane envelope after cutting, and kept in a light tight box until required.

4.4.2 Exposing of Holograms

The plates were loaded into plate holders, and exposed in total darkness. Although a camera and shutter could have been mounted on the bench, this would have introduced vibration problems. The shutter was therefore sited some distance away from the optical components and mounted on a dexion frame. This frame also carried the light proofing which surrounded the laser, necessary because of the large amount of diffuse radiation emitted in all directions from the sides of the laser plasma tube. Fig 4.13 shows the layout of the frame, laser and shutter. It also shows the low speed air blower, which kept the laser cool, and the arrangement of light baffles.

The arrangement of blower, and light baffles prevented unwanted light from escaping but also provided a steady cooling stream of air for the laser. The exit of this stream was at the top of a "chimney", where least interference with the optical paths would be caused.

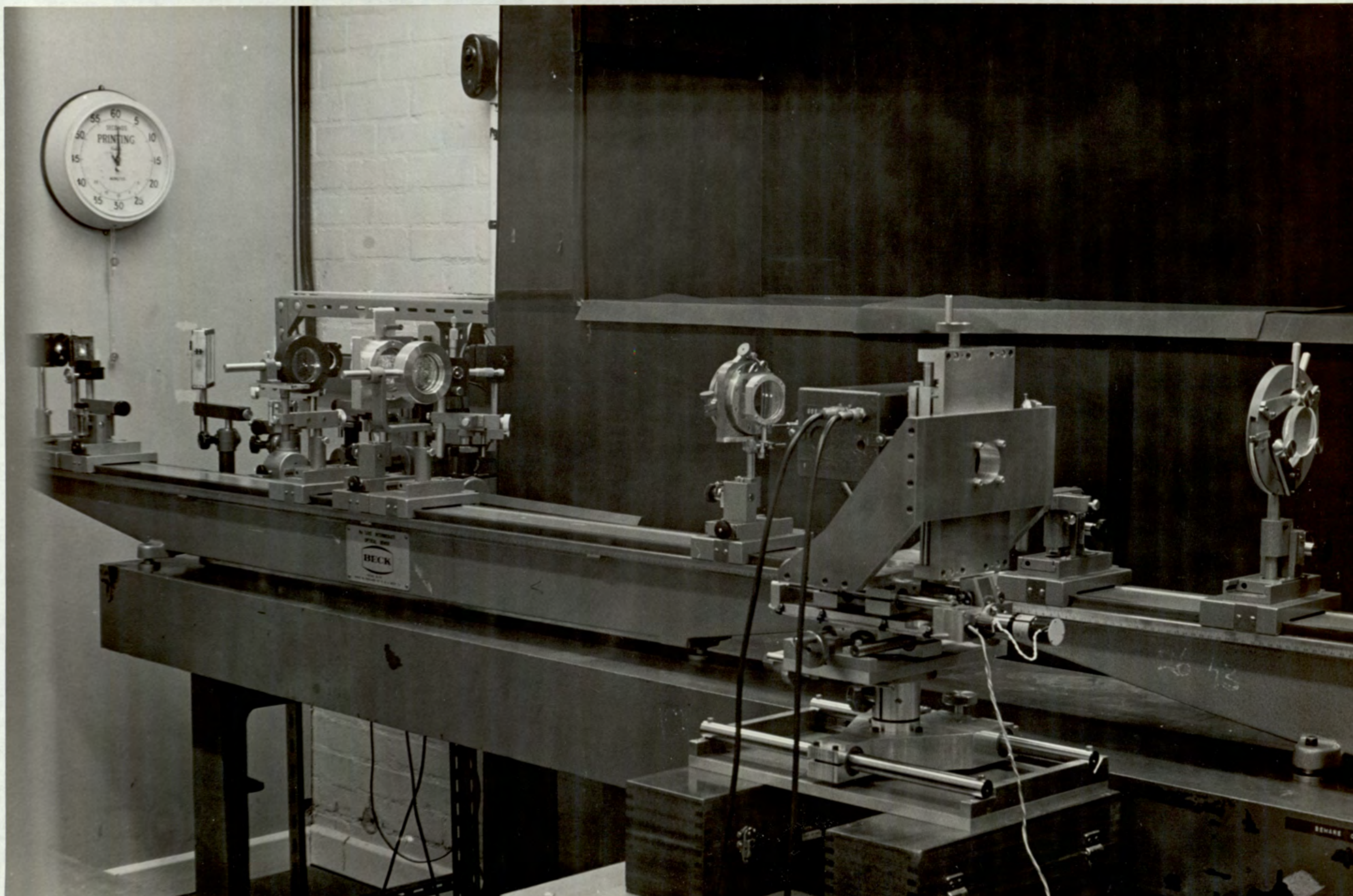


Plate $\underline{\overline{XI}}$ General view of apparatus, showing light proofing.

The light shutter was mounted on a triangular section optical bench on the dexion frame, and activated by a solenoid, allowing remote control operation. This was essential if air turbulence and vibrations were to be minimised. By mounting the shutter and solenoid on the frame, a source of vibration was removed from the surface table. No part of the frame rested on the table, it being designed to stand astride the table and on the floor.

The plates could thus be loaded into the holder and exposed in totally dark ambient conditions, and then replaced into a light tight box prior to development. A general view of the equipment, showing the light proofing, is shown in Plate XI.

The main disadvantage with this arrangement is that the laser was not easily accessible for adjustment, except at either end, where the tuning controls were situated. In later work, the laser was situated in an adjoining laboratory, with the laser beam transmitted through the partition wall.

4.4.3 Hologram Processing

The plates were processed in a well-equipped darkroom, using standardized procedures. Each plate was developed, rinsed in clean water and fixed for 5 minutes in Amfix. They were then washed for 30 minutes in running water, and dried in still air. Great care had to be taken to avoid dust and grit settling on the damp emulsion, and for this reason,the water was filtered, and no fans or blowers used for drying.

The details of the development stage will be discussed in the next chapter; for all plates D19 developer was used, under fixed temperature, time and concentration conditions. A fresh dish of developer was made up for each plate, about 100cc of solution being used.

Before returning the processed plates to the optical system, they were examined under a microscope to check that hologram "carrier" fringes had been recorded. Normally, a 10X or 20X objective was used, in conjunction with a X10 eyepiece. Details of the fourier transform could be studied in addition to the interference fringes. It was considered that a microscope was an indispensable part of the equipment needed for research work. Examples of hologram photographs taken through a microscope will be given later.

4.4.4 Safety Aspects

The use of continuous wave gas lasers of output power as high as 80mW presented a significant possibility of eye damage to the user, and to any other personnel who may have occasion to enter the laboratory.

This risk was minimised by two actions;

(i) Erection of laser warning signs over the doorway into the laboratory which became illuminated whenever the laser was in operation.

(ii) Fitting of a Safety Shutter to the output end of the laser. This shutter consisted of a neutral density filter of approximately 2.0 density, being carried in an arm which could be retracted from the beam by a solenoid. The solenoid was activated by remote control, using a long length of cable fitted with a pistol grip and push button.

The filter was sufficient to reduce the total output power of the laser by $^1/_{100}$ th (ie 0.8 mW), which was considered relatively harmless(5) (6). At the same time, sufficient power was transmitted to permit preliminary alignment of optical components. The filter was withdrawn whenever exposures or photoelectric measurements were made; having made sure that no stray reflections could cause eye damage, or that no other personnel were present.

Great care had to be used whenever visual examination of fourier transforms or hologram response was needed, and in these cases the laser beam intensity was further reduced by inserting neutral density filters into the object beam, before the beam expander and spatial filter.

4.5 Performance of Optical System and Components

4.5.1 Optical components

Observations on the performance of each of the doublet lenses (ie by OTF) were not found to be particularly useful compared with measurements on the optical system as a whole; nevertheless an attempt was made to examine the

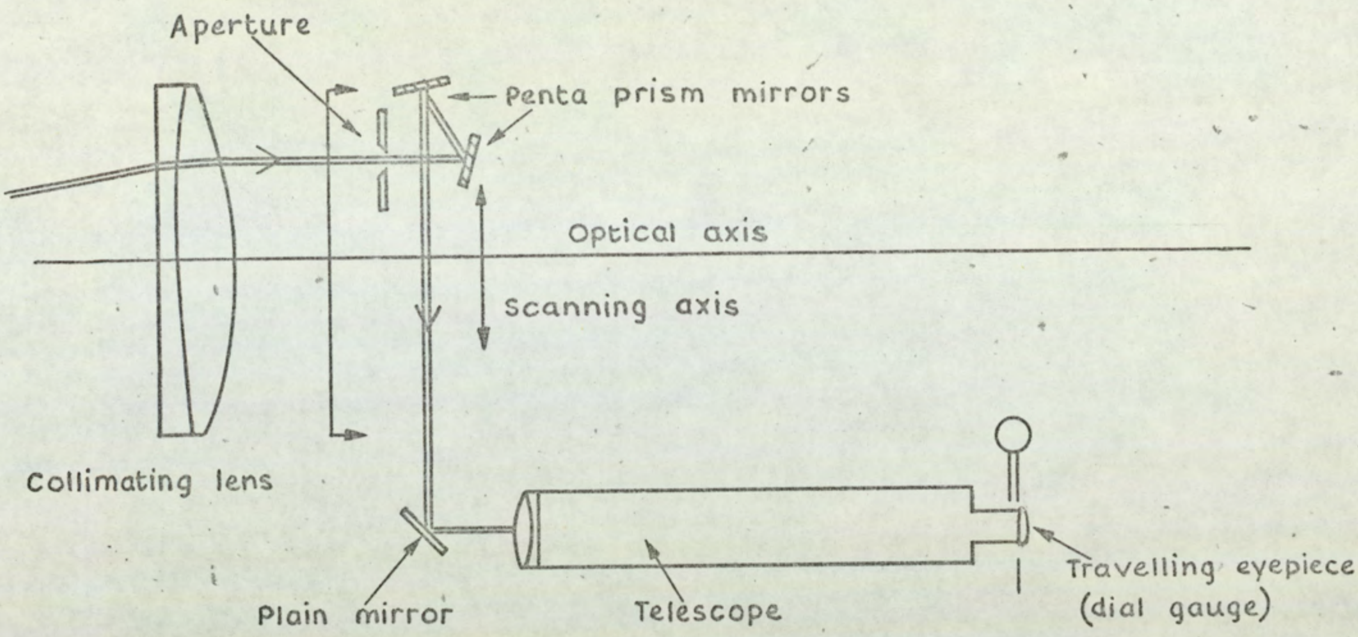


Fig. 4. 14 Measurement of collimated beam aberrations

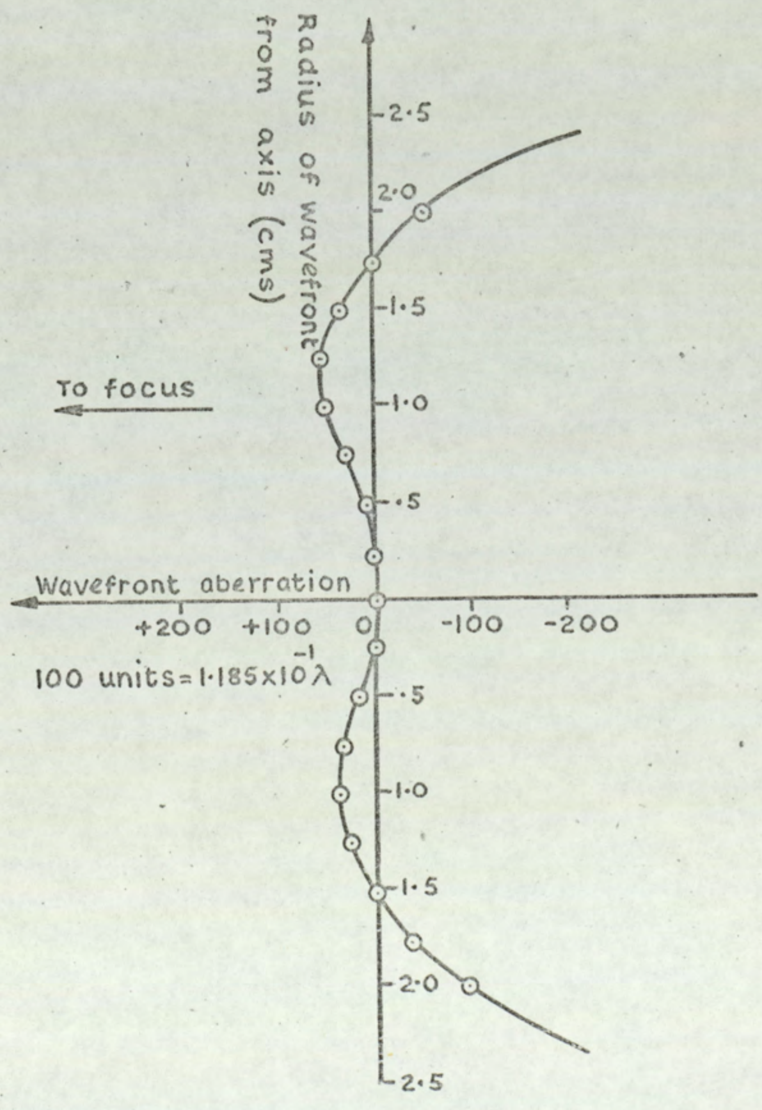


Fig. 4. 15 Collimated beam wavefront

aberrations of the collinated beam, which would indicate the quality of the collimating lens.

The collimated beam was studied by measuring the deviation of a thin pencil of light reflected by a penta prism as it was traversed across the beam. The deviation was measured by observing the shift of focus of the pencil in the telescope fitted with a sensitive travelling eyepiece, enabling very small deviations to be measured. (See Fig 4.14).

An ideal collimated beam would produce no deviation of the pencil at all, assuming no errors in the slideway of the penta prism carriage. In fact, the original telescope, adapted from an autocollimator, proved to be too insensitive for this purpose, and a telescope was assembled using a simple 0.75 dioptre lens and a microscope. Movement of the focus of the pencil was measured by means of a dial gauge attached to the microscope. In this way, angles of 7.5×10^{-7} radian (0.15 secs) could be measured. A table of deviations of the light pencil corresponding to different positions of the penta prism across the aperture was obtained. In practice a small aperture was attached to the penta prism to select a pencil of light from the collimated beam. By summing the deviations in the total of results a curve showing the wavefront shape was derived; this was corrected for tilt and for focussing errors, and is shown in Fig 4.15. The wavefront aberration for a collimated beam diameter of 3 cm is 0.06λ, indicating that the wavefront is well corrected for an axial beam. The aberration appears to become severe towards the edges of the beam however.

The focussing error in the wavefront corresponds to the collimator being in an incorrect position relative to the source (ie the focus of the beam expanding objective); the error being calculated to be approximately 5 mm. Such an error was not considered too troublesome since it simply means that the input transparency will be illuminated by a slightly converging beam and will introduce very small changes of size in the fourier transform if the transparency were moved backwards or forwards.

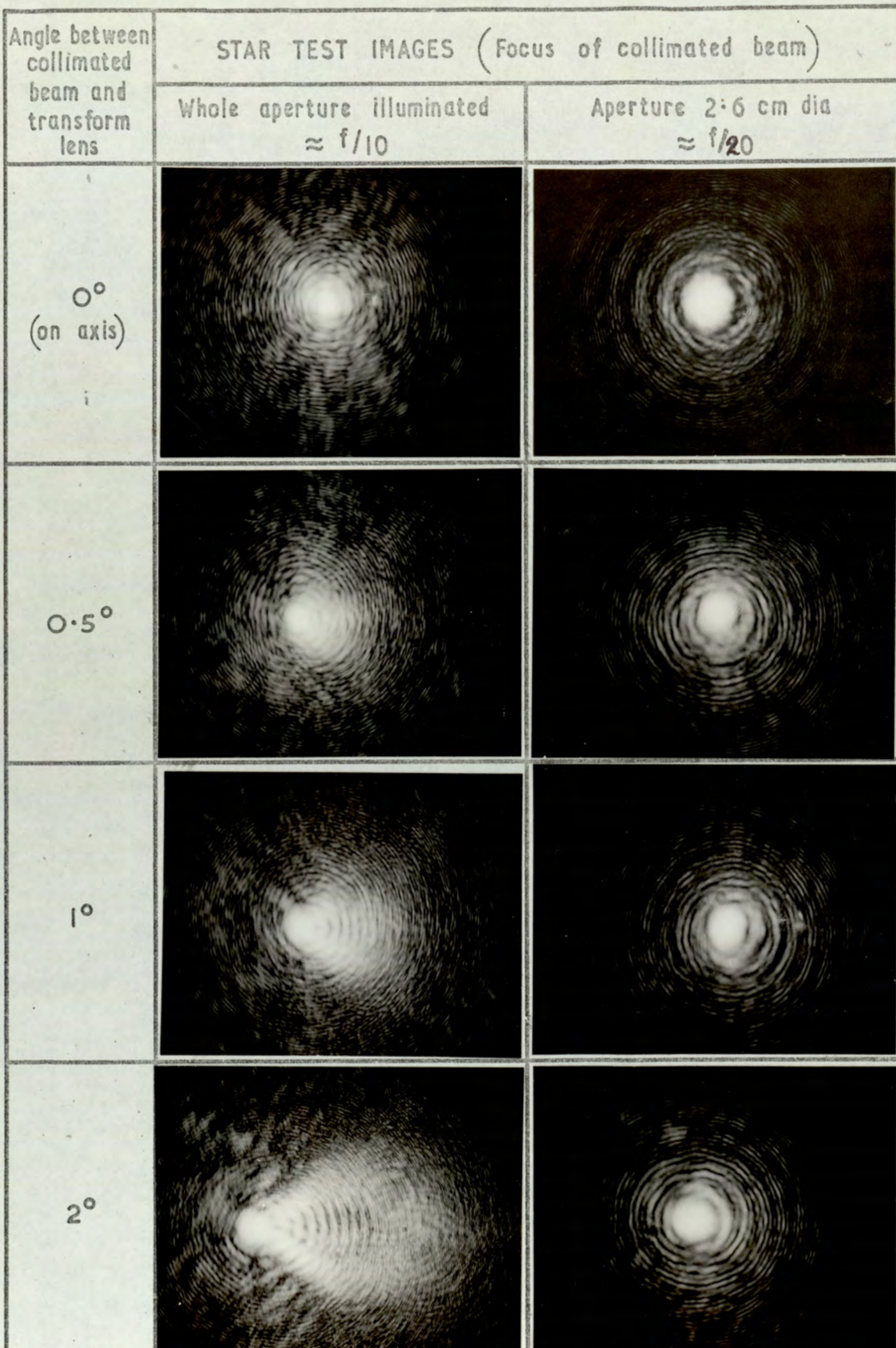
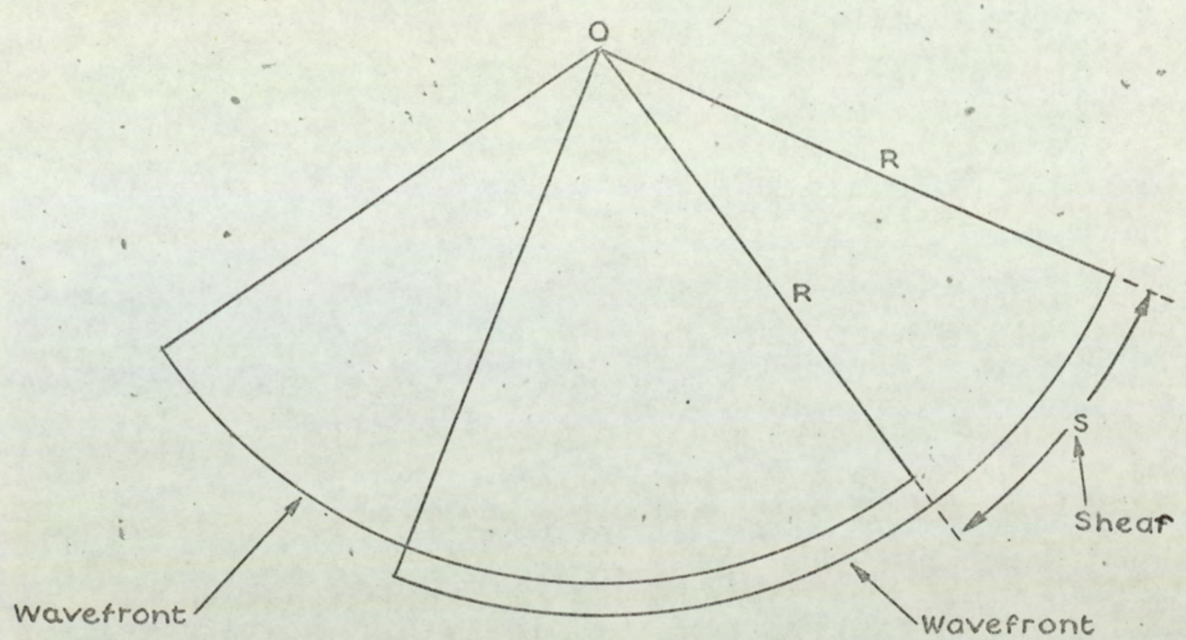
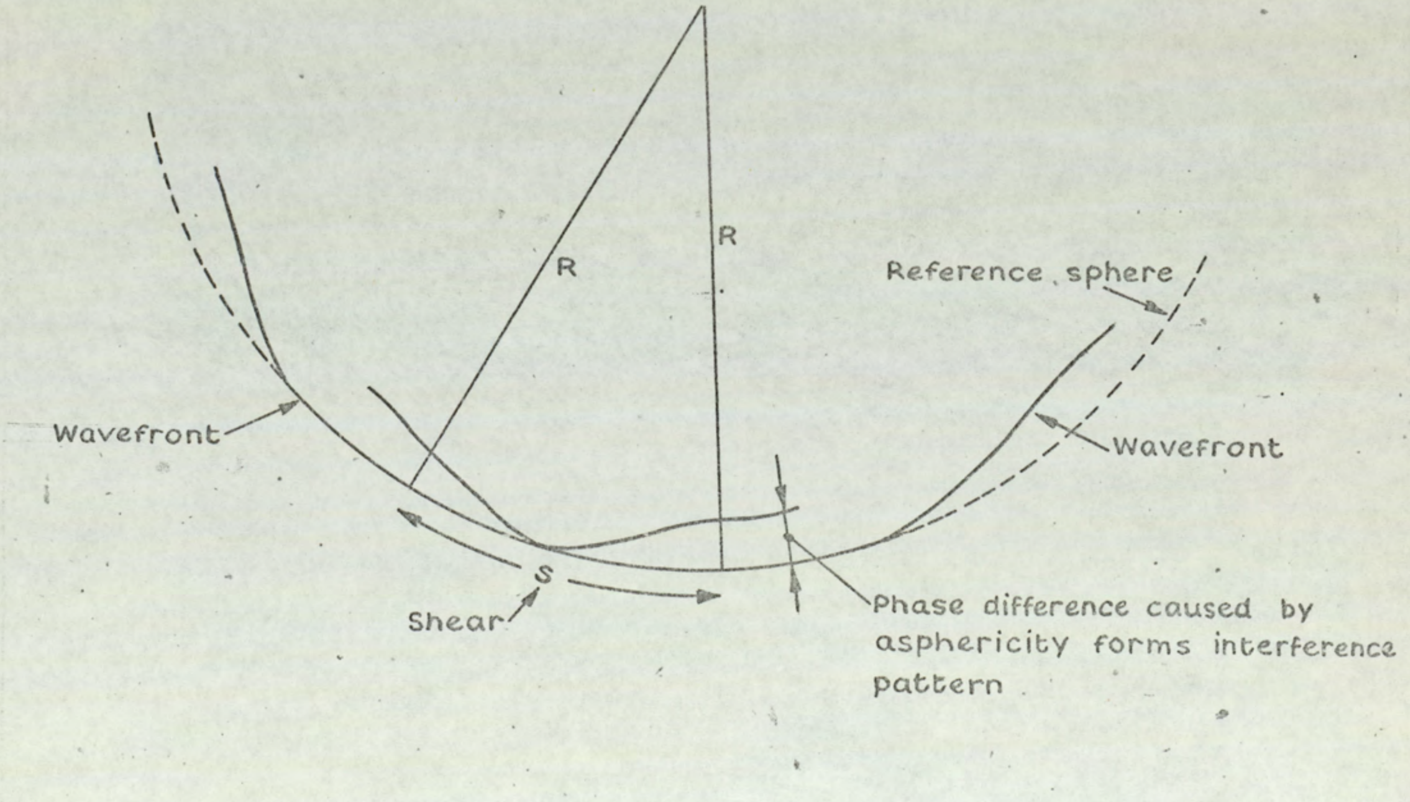


Fig 4.16 Performance of transform lens – star test



Wavefront shearing of spherical wavefronts



Wavefront shearing of aspheric wavefronts

Fig. 4. 17

Because of the limited sensitivity obtained in the foregoing measurement, further investigations were confined to the fourier transform plane of the complete optical system.

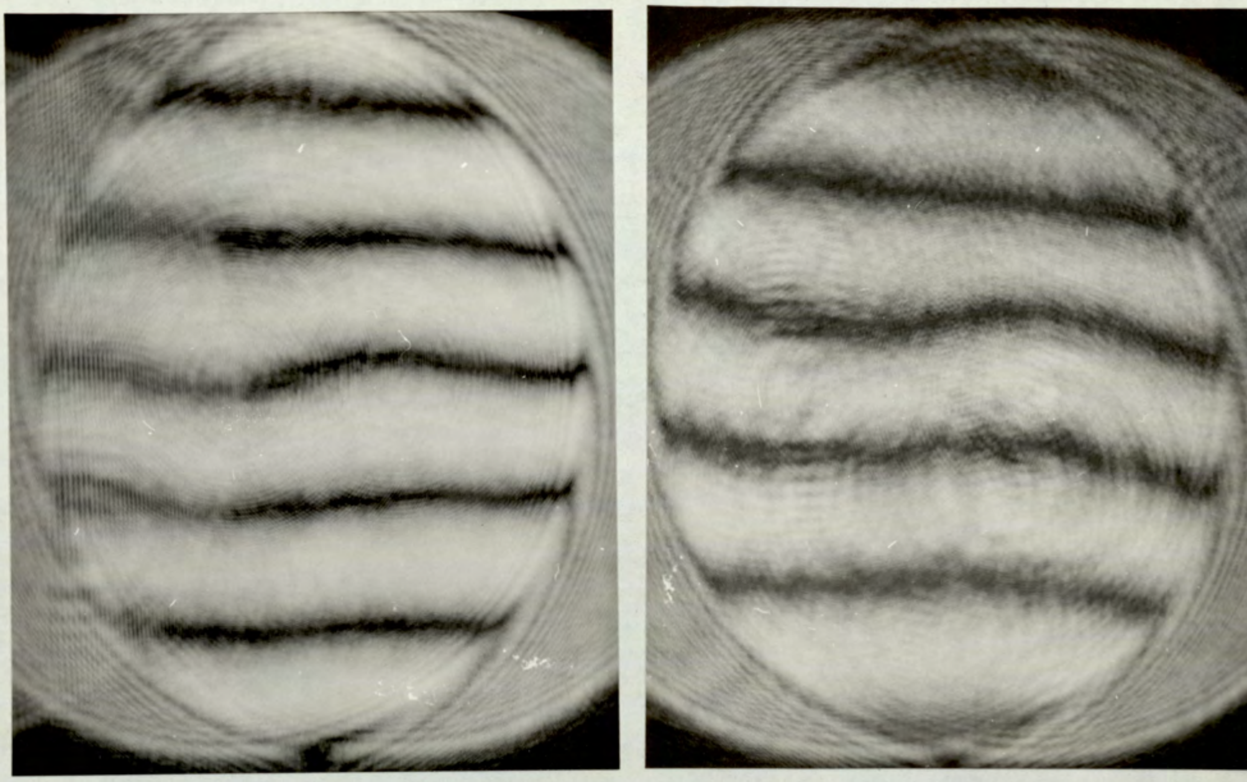
4.5.2. Performance of the Optical transforming system

The wavefront emerging from the transforming lens will contain the aberrations of both collimator and transform lens, and of any transparent object placed in the input plane. The aberrations of this wavefront may be examined by observing the fourier transform itself (ie: equivalent to the "star"test) or by measuring the asphericity of the wavefront with a wavefront shearing interferometer.

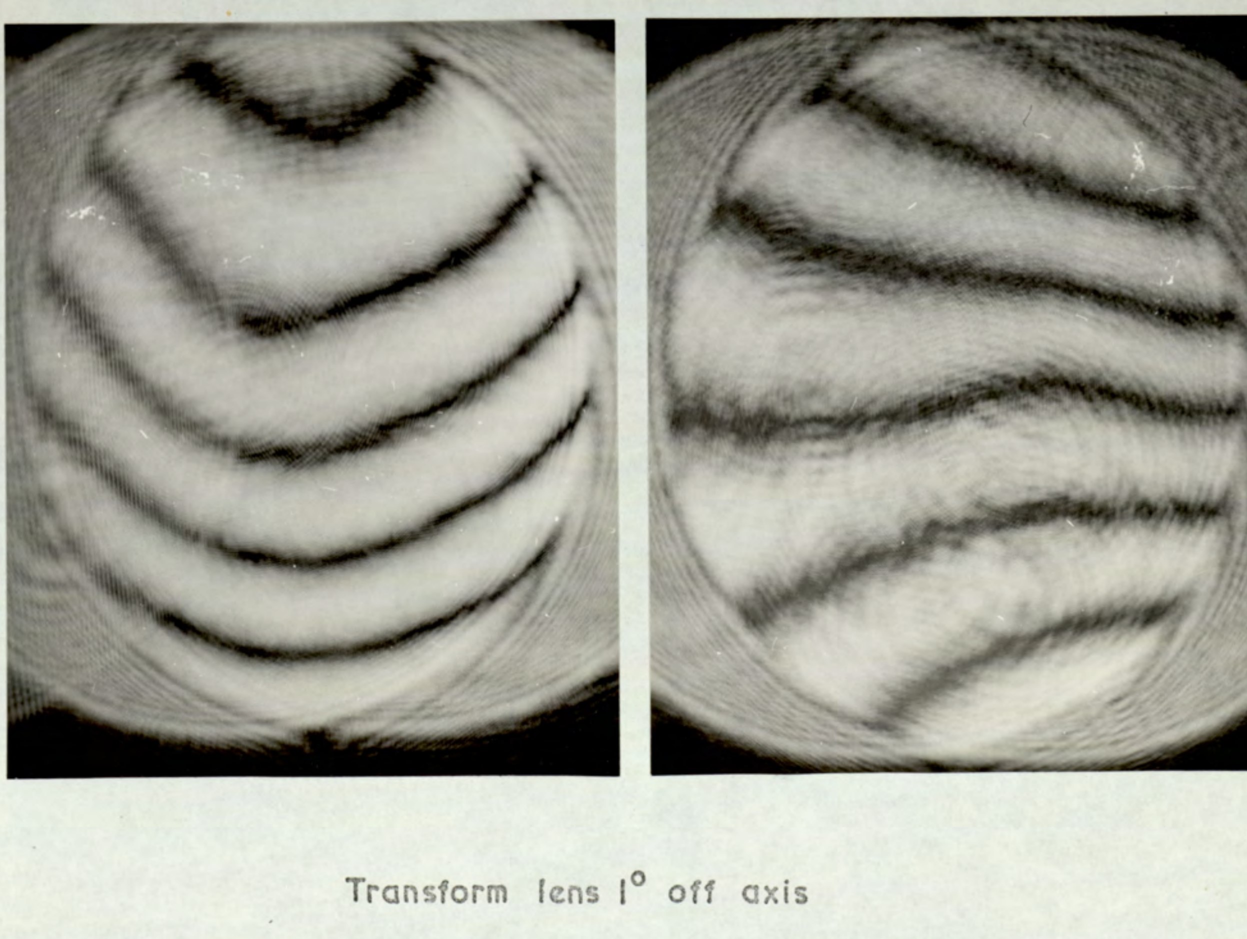
The first method does show the rapid degradation of the wavefront for different apertures of the transform lens and for small angular rotation of this lens. Fig 4.16 shows several photographs of the transform, recorded with the transform lens axis at 0°, 0.5°, 1° and 2° to the optical axis and for different apertures. It is difficult to obtain quantitative assessment of the wavefront using this method, and therefore the wavefront shearing method was used.

A practical wavefront shearing interferometer was described by Bates[7], in which the wavefront was divided by a beam splitter and then recombined, but with one wavefront displaced relative to the other, giving an effective "shear" of the wavefront. If the original wavefront were spherical, and if the shearing were made along an arc equal in curvature to that of the wavefront then no changes are observed in the recombined wavefronts. (see Fig 4.17). However, an aspheric wavefront when sheared in a similar way will produce phase differences between the two wavefronts and these will cause fringes to appear in the recombined wavefronts; these fringes correspond to interference between one wavefront and its' sheared component. Since the phase differences are likely to be small in well corrected optical systems the interference pattern will be difficult to interpret, particularly for aberrations of less than one wavelength. By introducing an amount of tilt

Tangential shear Sagittal shear



Transform lens on axis



Transform lens 1° off axis

Fig. 4.18 Typical shearing interferograms for transform lens into one sheared wavefront with respect to the other, a regular pattern of interference fringes can be formed, which, in the case of perfectly spherical wavefronts will be straight and parallel. The presence of asphericity in the wavefront will cause the tilt fringes to be deviated, and by measurement of the deviation of these fringes the wavefront aberrations can be calculated. The method of reduction has been described by Marchant[8,9] whose papers also describe the principle of the wavefront shearing interferometer in some detail.

An alternative design of interferometer using the same principle has been described by Drew[10] and by Brown[11]. The latter design has been manufactured by Sir Howard Grubb Parsons & Co, and this design (Type WS2) was used in the subsequent study. This instrument is extremely simple to use, and may be easily mounted on an optical bench at the focus of the wavefront to be examined. [Note: since the instrument can be used with a wide variety of wavefront curvatures (up to $^F/3$), the placing of the interferometer at the focus of the wavefront ensures that the shearing curve is approximately equal to the curvature of the wavefront. The method of reduction removes any residual focussing errors.] The interference pattern was studied for a variety of conditions (ie: transform lens off axis, presence of liquid gate, etc) the patterns being recorded on photographic film and printed. Typical interferograms are shown in Fig 4.18, for the transform lens on axis and 1^o off axis, with the shear direction in both tangential and sagittal planes. The distortion of the fringes, indicating the presence of wavefront aberration is quite evident.

The use of the shearing interferometer in both tangential and sagittal planes requires that the instrument can be rotated through 90^o, without any displacement from its original transverse plane. This is particularly important in the case of astigmatism, where astigmatic aberrations are detected as changes in wavefronts curvature in between the sagittal and tangential planes. The rotation of the interferometer was accomplished by

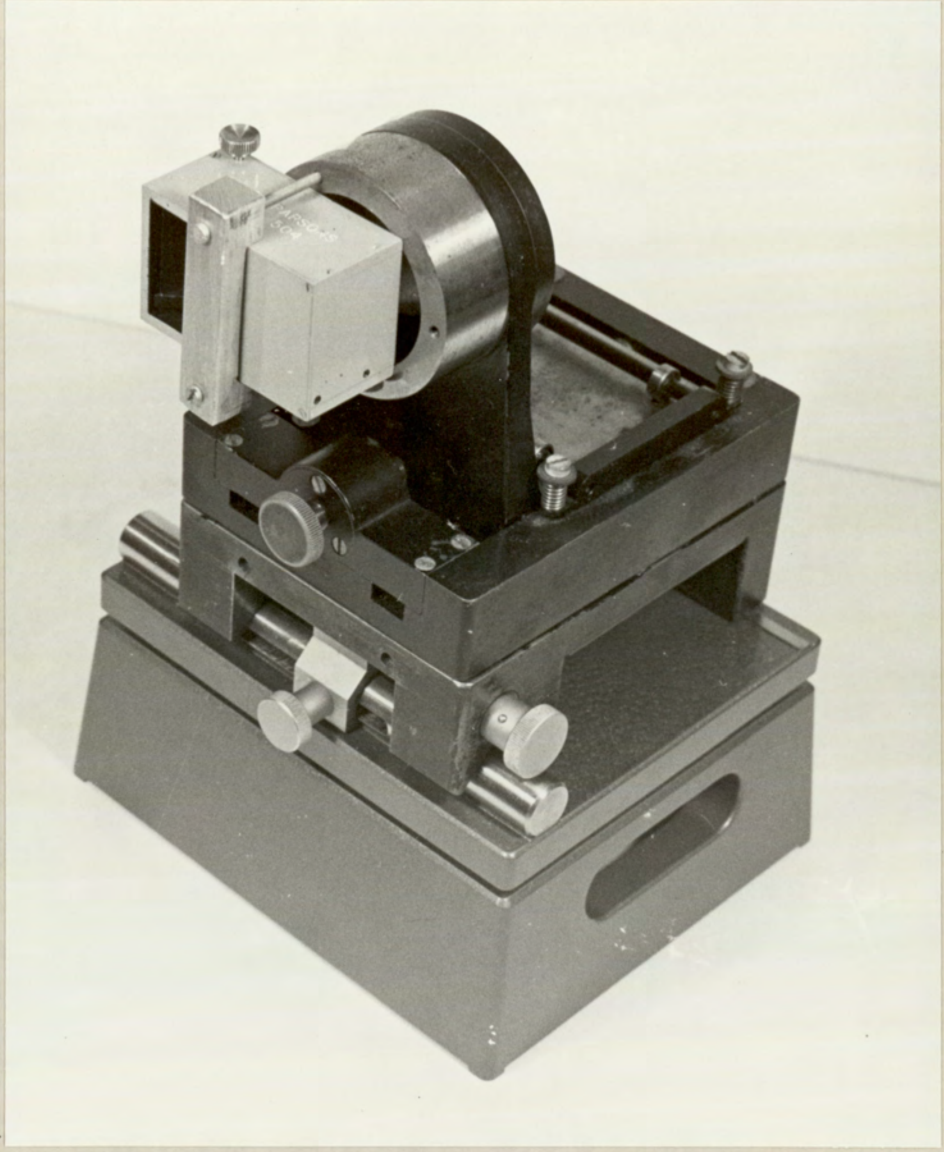


Fig.4.19 Shearing interferometer and mounting

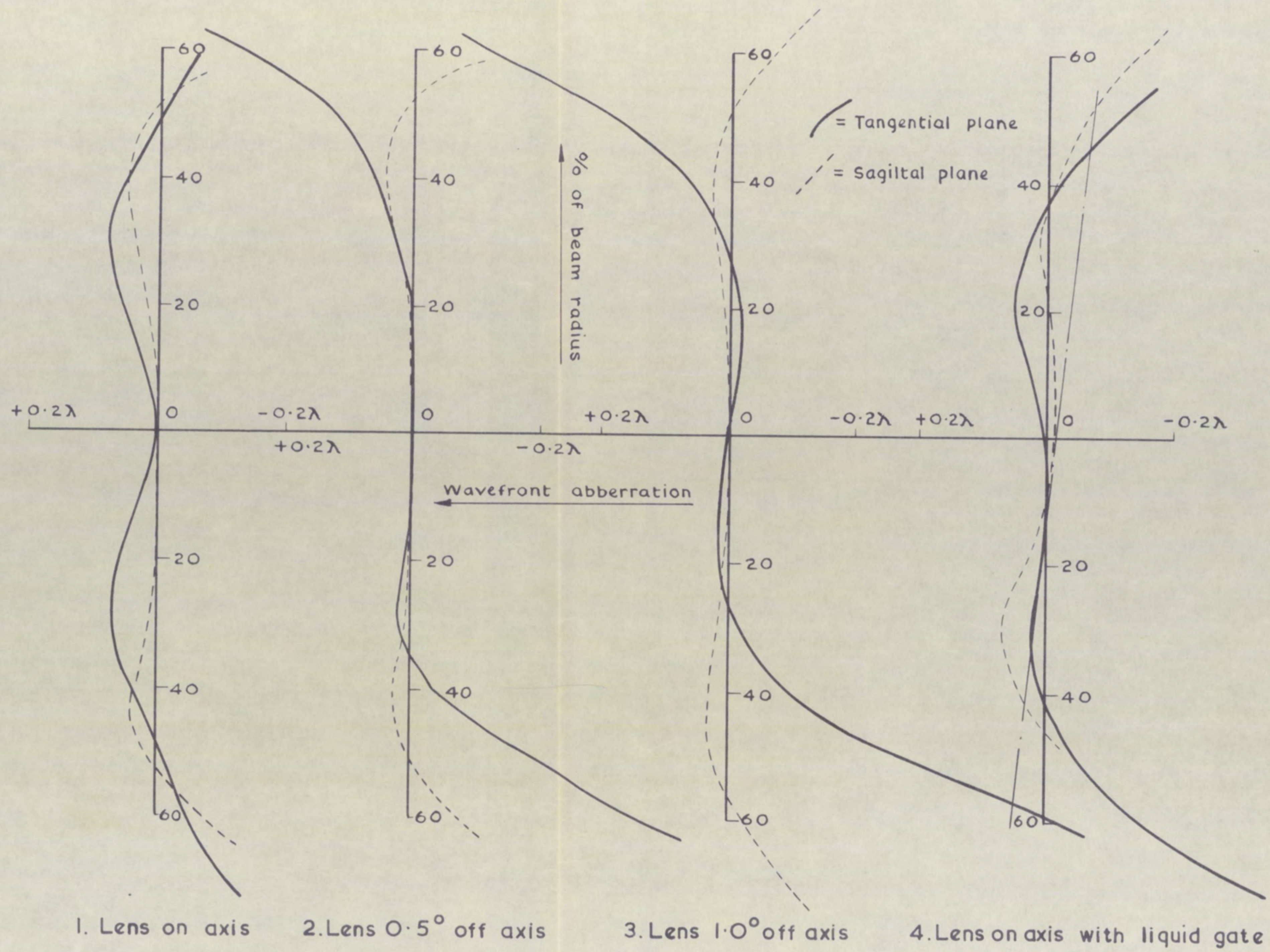


Fig. 4.20 Wavefronts derived by shearing interferometry

mounting it to a circular block, which could be rotated in a massive plain bearing. The bearing itself was mounted on a carriage, which could be moved to focus the interferometer, and also to investigate different wavefront curvatures. The position of the carriage was monitored by a dial test gauge, reading to 0.002 mms. The interferometer and mounting are shown in Fig 4.19.

In the subsequent tests, the rotation of the transform lens was achieved using the adjustment provided in the lens holder. A further dial test gauge was used to determine the position of the lens rim and hence the angular rotation.

The individual tests will be discussed separately, referring to the wavefront shapes shown in Fig 4.20.

(i) Transform Lens on Axis

With the transform lens on axis both the tangential and sagittal planes show the typical shape of wavefronts having correction for spherical aberration; the correction in this case being obtained by use of doublet lenses. The maximum wavefront aberration from a best fit spherical reference surface was found to be 0.08 λ for the tangential plane and 0.06λ for the sagittal plane, for a diameter of 1.35 cms (equivalent to an f/24 cone). Beyond this radius the aberration increases very rapidly in the opposite direction, ie: away from the focus.

(ii) Transform lens 0.5° off axis

The sagittal wavefront shows virtually no change in this case, as expected for a plane containing the principal ray. Here the maximum wavefront aberration is 0.04λ for a radius of 1.35 cms. The variations in wavefront aberrations are attributed to errors in the recording of the interferograms (particularly the effect of air turbulence) and in the process of reduction.

The tangential wavefront shows the curve expected of a wavefront suffering from coma; the deviation of the wavefront from 0.54 on radius (f/60) being quite severe. Over a radius of 1.35 cm, as before, the aberration was found to be 0.33 λ.

(iii) Transform lens 1.0° off axis

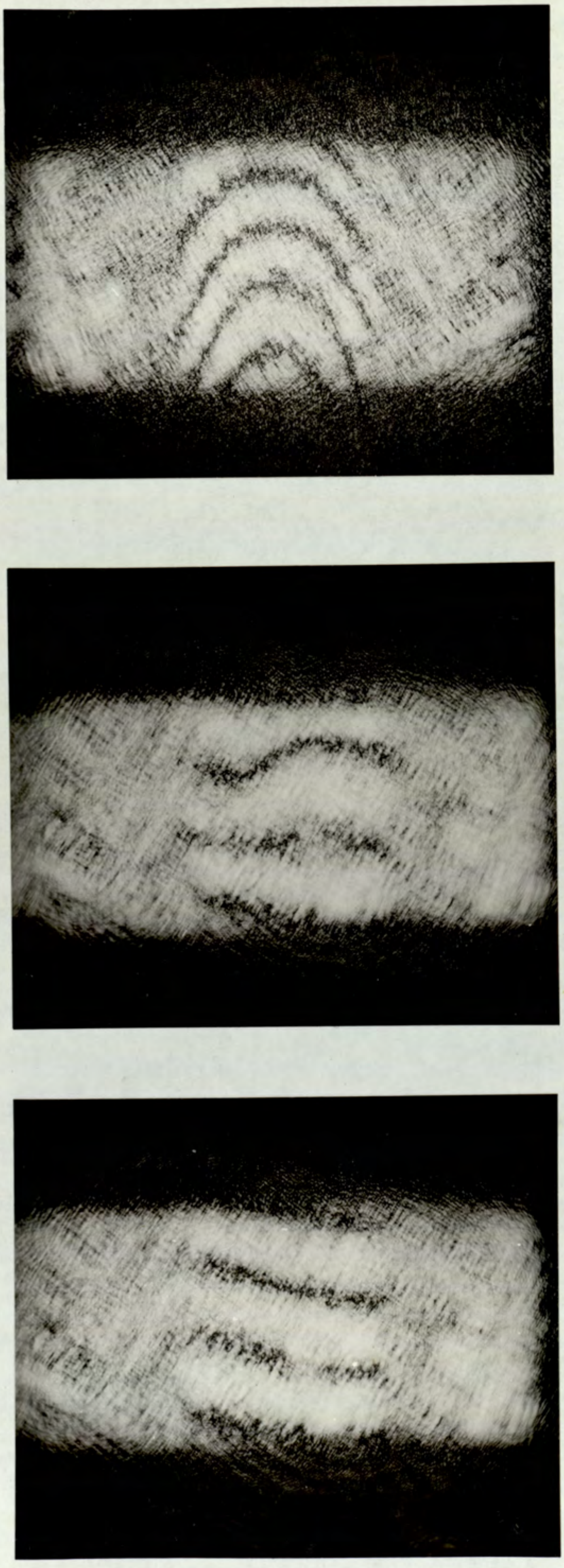
Again, the sagittal wavefront shows virtually no change, as expected. The aberration was measured as 0.03 λ for a radius of 1.35 cm. The tangential wavefront shows a more severe coma aberration than before, of approximately 0.5 λ at 1.53 cm radius.

These curves clearly show the marked degradation of the wavefront at relatively small angles to the optical axis. However, the degradation has been shown to be very small provided that only narrow cones are considered, of the order of f/60. This corresponds to the use of small apertures of linear size 5mm approximately. Whilst this would be practical in the construction of fourier transform holograms, it would be a severe restriction if the total area of the input data were restricted to such a cone. The use of larger apertures must however, consider the effect of aberrations, particularly with regard to which spatial frequencies (and which angular deviations) are most important. Practical investigations will be discussed later in this report.

(iv) Effect of Liquid gate

The effect of the liquid gate described earlier in section 4.2.6 was observed by obtaining shearing interferograms with the liquid cell inserted between collimator and transform lens. In the first test, no transparency was used, the liquid gate containing only Decalin liquid.

The tangential and sagittal curves, each referred to their best reference sphere, are broadly similar. There appears to be a slight increase in the wavefront aberration over 1.35 cm radius which was measured as 0.17 - 0.10 λ in the tangential plane and 0.05 λ in the sagittal plane. Of more concern was the fact that the best fit reference sphere for the sagittal wavefront



Transparency in air
no index matching

Transparency in liquid gate
[immediately after inserting
transparency]

Transparency in liquid gate
[half hour after inserting
transparency]

Fig. 4.21 Shearing interferograms showing removal of phase defects by liquid gate was not the same as that of the tangential wavefront; and when the former was referred to the tangential focus, it appeared that small amount of astigmatism was detected. It was concluded that the liquid cell behaved as a very weak lens (the change of focus in the tangential plane was found to be 0.4 mm) which was distorted, producing a small degree of astigmatism. The distortion was small enough to be neglected however, and was thought to be caused by clamping the glass windows to the liquid gate body. The error in focussing (ie: 0.4 mm) was corrected in practice by movement of the hologram holder along the optic axis. The increase in spherical aberration observed earlier (ie: of about 0.1λ) was not considered high enough to be troublesome for this work, but indicated that a more precise liquid gate would be needed for exact measurements.

The usefulness of the liquid cell in removing phase variations from transparencies was also demonstrated using the shearing interferometer. Fig 4.21 shows photographs of the shearing interferogram with a transparency introduced between the collimator and the transformer.

The curved fringes of the transparency-in-air interferogram indicate the presence of coma, and suggest that the plate is wedge shaped. When the transparency is placed in liquid the curved fringes are straightened, as shown by the centre and lower photographs. The act of inserting the transparency causes temperature differences in the liquid, explaining the difference between the two lower photographs, which were taken half an hour apart. The lower photograph shows some residual curvature, indicating that the glass of the transparency was not completely matched in refractive index by the liquid; there is, however a very noticeable improvement.

Returning to Fig 16, the appearance of the Airy ring pattern at different apertures shows how spherical aberration may be eliminated by using a small aperture. In particular, the ring pattern for the lens on axis, with an aperture of f35 is virtually aberration free. This is in agreement with the interferometer data, since for f35 the radius is 0.95 cm, giving a maximum aberration of 0.08λ. At higher apertures the effect of spherical aberration becomes more pronounced, and may be seen in both Fig 16 and in the interferogra results.

The appearance of the Airy ring pattern at full aperture indicated that astigmatism was present, which since the lens was on axis, was thought to be caused by misalignment of the uncemented components. Attempts at realignment did not eliminate the defect, and it was concluded that the outermost parts of the lenses contained some figuring error. This error was not found for apertures below f/20.

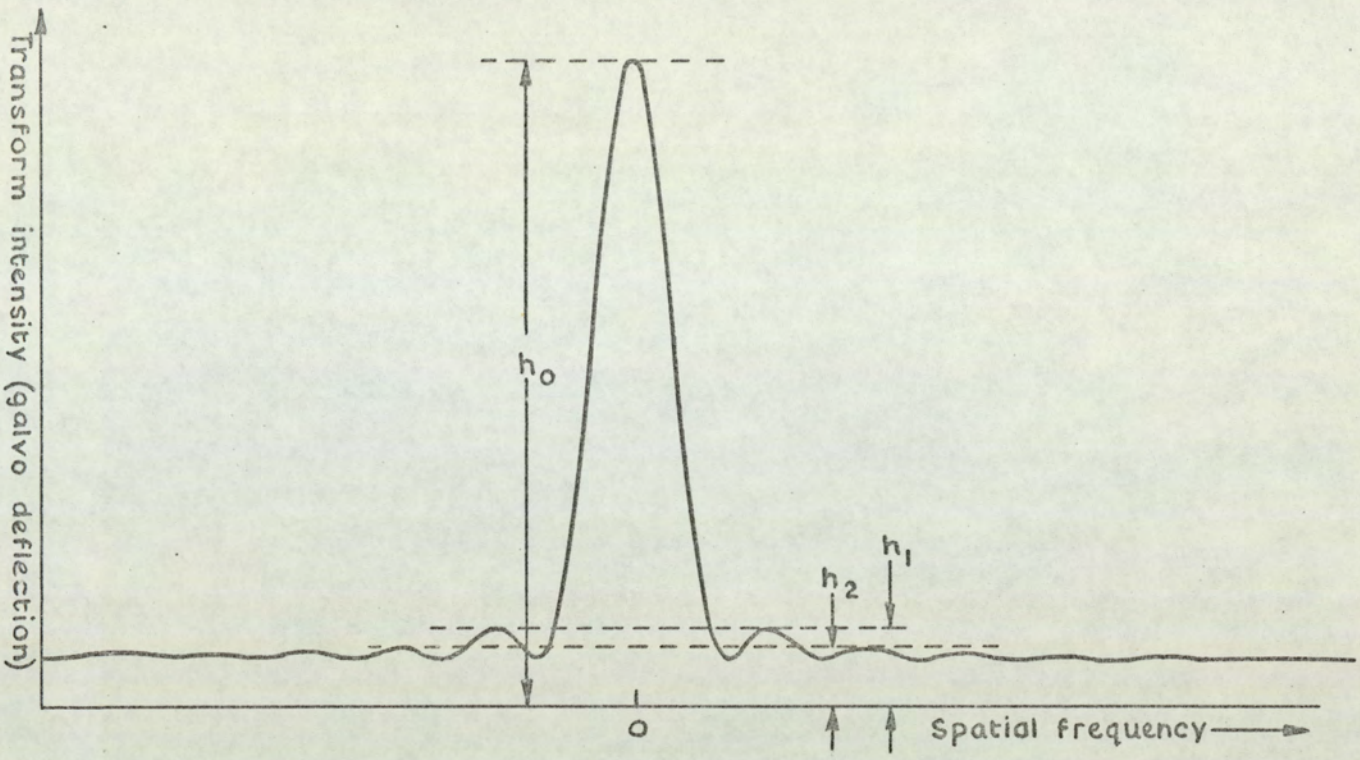
4.5.3 Fourier Transform measurements

The distribution of intensity in the fourier transform was measured, using the scanning unit, partly to check the operation of the apparatus, and partly to confirm that a fourier transform was formed with reasonable accuracy. The method is described here, partly because it illustrates the procedure used when making measurements with the optical system.

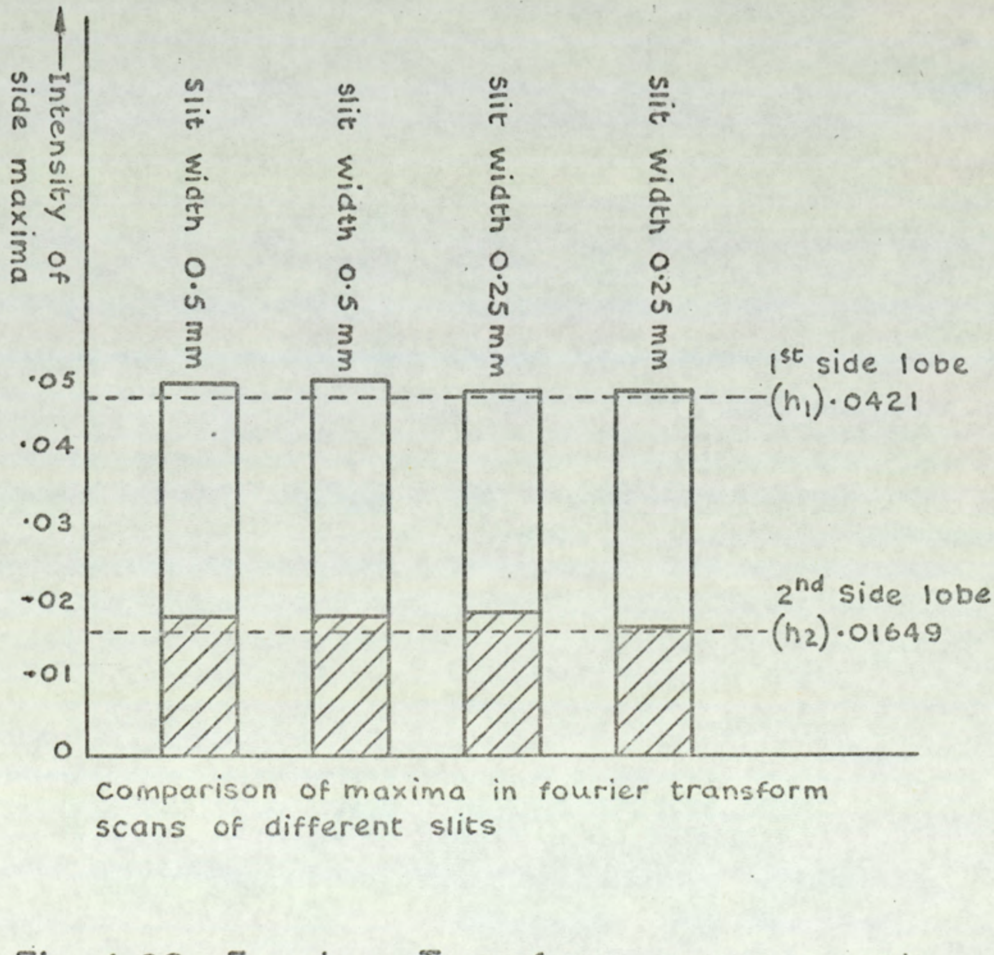
The input function, used for generating a fourier transform was a Hilger variable slit, having micrometer adjustment to 0.005 mm. Its fourier transform was scanned using a photomultiplier scanning unit, fitted with a photographic slit of 0.06 mm width. The scanning unit was different from that described earlier, in that only one movement of the photomultiplier box was permitted, this being motor driven. A 100 lines/inch grating connected to the moving carriage supplied a sinusoidally varying signal to the recorder for measuring linear displacement. The photomultiplier output was displayed on the UV recorder, together with the 100 cycles per inch signal.

The operating procedure was as follows:

1) Laser switched on for 30 minutes, then tuned for maximum power, using laser power monitor (See Fig 4.2)
2) Object beam expander and pinhole adjusted to give as uniform illumination over collimating lens as possible, although at best the distribution would be Gaussian.



Scan of Fourier Transform of slit 0·25mm wide



Comparison of maxima in fourier transform scans of different slits

Fig. 4.22 Fourier Transform measurements

3) Collimator and Transform lens system checked to give good Airy disc and ring pattern at focus of Transform lens, and that no coma or Astigmatism were present

4) Introduce stop in input plane to reduce spherical aberration

5) Observe Airy pattern at focus with microscope; leaving the microscope fixed, adjust a horizontal pointer so that its tip is just in focus with the microscope, ie: pointer tip is coincident with the focus of lens

6) Remove microscope and gently bring photomultiplier up to pointer, and let pointer tip just touch the scanning slit. The pointer is then removed, leaving slit in focal plane

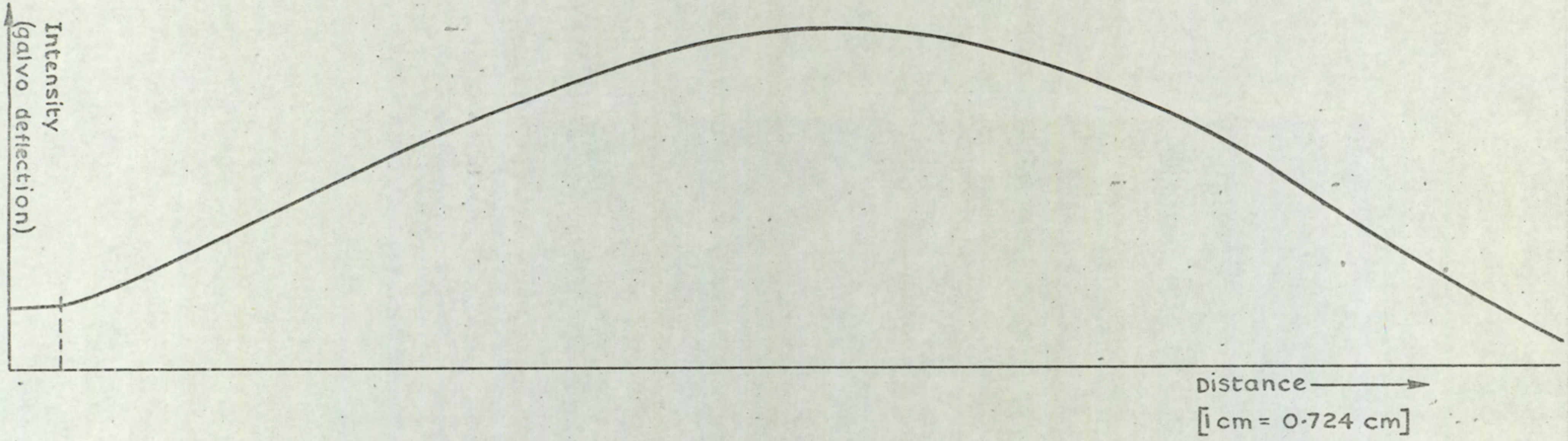
7) Insert input function, or slit in this case, into the input plane, check appearance of fourier transform on white card

8) Scan fourier transform with photomultiplier, adjusting the gain so that the maximum intensity of the transform is equal to full scale deflection of the recorder.

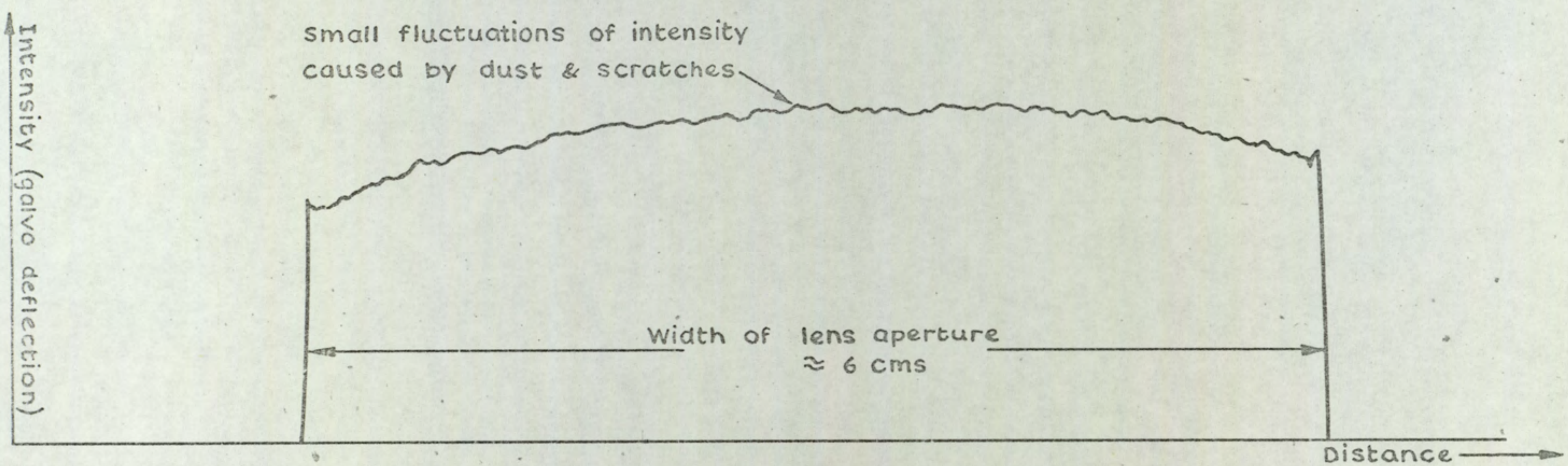
9) Make two recordings of the fourier transform for each slit width.

The results of measurements derived from the recordings are given in Fig 4.22, together with a typical UV recording. It was found that the opposite halves of the transform agreed very closely with each other, indicating that no unsymmetrical aberrations were present. The chart shows that all the values of diffracted maxima, when normalised so that zero maxima equalled, unity were high. This suggests that the normalising coefficient was too low. This could have been caused by non linearity in the recording process, or by the possibility that light was being scattered into the diffracted orders. In general, there is agreement to within ≈10%. It is interesting to note that the magnitude of error is similar for both sets of maxima, indicating that a constant factor was responsible.

The distribution of intensity in the fourier Transform will be discussed later in more detail, in connection with the analysis of hologram response.



Intensity distribution across expanded laser beam, filtered but not collimated



Intensity distribution across collimated laser beam

Fig. 4. 23

4.5.5 Laser beam intensity distribution

In addition to demanding that the wavefront from the collimator be perfectly planar or spherical, it was also required that the wavefront was uniform in intensity. However, as stated in Section 4.2.3, the intensity distribution of the laser beam was gaussian, and there will be a variation in intensity across the input plane.

The intensity distribution in the input plane was studied by making UV recordings of the photomultiplier output as it was moved across the plane. If the laser beam were truly gaussian, and if the collimating lens, and beam expander were correctly aligned, one would expect a curved distribution of intensity, the curve being part of the gaussian curve. In practice, it was found that the laser tended to give a non-gaussian distribution of intensity, having a "dip" in the centre of the input plane. This dip could only be eliminated by keeping the laser mirrors perfectly aligned and by using a 10 μ spatial filter at the beam expander. The filter was extremely sensitive to movement, and great care had to be taken that it was not disturbed in the course of experiment. Small variations in the filter position would cause the intensity distribution to change abruptly.

With this arrangement, some good intensity distributions were obtained, as shown by the recordings in Fig 4.23. The first shows the light distrubution from the beam expander and spatial filter, without the collimating lens. There are no variations caused by scattering or aberrations, and the curve is uniformerly gaussian. The second curve, showing the collimated beam, also shows the gaussian curve. There are small variations caused by dust particles and inclusions in the lens, which cannot be removed by spatial filtering. This was the justification for keeping the number of optical components to a minimum, and thus reducing uneven intensity distributions.

The intensity distribution was not symmetrical, the maximum intensity being some 20 mm off axis. The variation in intensity in this case was of the order of 20%, although an on-axis beam would produce a variation slightly less than this. The variation could only be reduced at the expense of using a higher powered beam expanding objective, and losing more light. It was decided to tolerate the 20% intensity variation, and to allow for this where necessary in results.

4.5.6 Conclusions

The analysis of the optical system in the previous sections has shown that the theoretical requirements are met only for very limited apertures. The development of a system which meets these requirements over a wide aperture must include figuring of the lenses to reduce spherical aberration, masking to maintain constant intensity distribution across input plane and very precise alignment. The transform and integrating lenses would also be corrected for fields of up to 10° and for the relevant conjugates.

Such a system would be expensive, and time consuming to produce: Provided that the aberrations and irregularities of the system described are known and allowed for, it can be used with confidence, at least in research work.

In investigations where the input is confined to the centre of the input plane, and is only of say, a few mm in diameter or length, then the most severe imperfections will be the off axis aberrations of the transform lens in the higher frequencies of the fourier transform. These imperfections will be present in the transform irrespective of the input, but provided that the effect on the transform relations $F(p,q)$. $H(p.q)$ is negligible, then the system will continue to work satisfactorily. This question is however, one of those which could be answered in the research work.

In this connection, the design of an optical system having minimum of aberrations and a short overall length was discussed by Blandford [12]. In this design, it was reasoned that the optical system should be symmetrical and be of a "double telephoto" type, also that the condition that no field curvature be present suggests a Gaussian design, all of which indicate a three or four lens system. The advantage of using four lenses was indicated, since the additional surfaces allowed extra parameters to be introduced which were useful in the control of aberrations.

The resultant system was about 1 metre long, and consisted of two doublets and two plano-concave lenses in a symmetrical arrangement. The maximum spatial frequency transmitted by this system, using light of 0.6328 μ m wavelength, was 47 cycles/mm. The aberration at the edge of the exit pupil, on axis was found to be $\lambda/8$, and shows good correction compared with the system used here.

It is concluded that the optical system described in this chapter whilst containing many inherent imperfections and aberrations, satisfies the theoretical requirements sufficiently well, particularly on axis, for it to be used as an optical correlator.

References Ch 4

1 Vander Lugt, Applied Optics 5, 11, November 1966

2 Marchant, Keyte, et al, RAE Technical Report TR67031 January 1967

3 Rogers, J Sci Instrum., 43, 1966 p 677-684

4 Spectra Physics Data Sheet for Model 332 Spatial filter

5 Ministry of Aviation, Laser Systems: Code of practice. 1965

6 Electronic Engineering Association; A General guide to the safe use of lasers. September 1966

7 Bates, Proc Phys Soc (London) 59 p 940 (1947)

8 Marchant, RAE Tech Memo Ph 262. September 1959

9 Marchant, RAE Tech Report 65012. February 1965

10 Drew, Proc Phys Soc (London) B, 64, p 1005

11 Brown, J Sci Inst Vol 32, p 137-139, April 1955

12 Blandford, AGARD Proceedings No 50, September 1969

Chapter 5

Photographic Process

5.1 Introduction

In chapter 3 it was assumed that the relationship between amplitude transmittance and intensity could be represented by $t_A = 1 - kI$ where I = incident intensity, and I contains the signal amplitude in the form $I = |A_R|^2 + |S(p,q)|^2 = A_R S^*(p,q)e^{-j\alpha p} + A_R S^*(p,q)e^{+j\alpha p}$ from equation 3.2. In practice, however, the relationship $t_A = 1 - kI$ is true only over a very small range of intensities, and since fourier transforms generally contain a wide range of intensity, non-linearities will occur. The purpose of this chapter is to examine the effect of these non-linearities on the hologram response and also to consider the effect of thickness variations in the emulsion.

It is important to realise the distinction between the photographic process as applied to holography, and the process as applied to conventional photography. In the latter, the signal is recorded as a variation in the density of silver grains in the photographic emulsion, and so a linear relationship between density ($\log \frac{1}{\text{Intensity}}$) and log exposure would be desirable. This relationship is found by plotting the Hurter and Driffield curve of log exposure versus density, a typical curve being shown in Fig 5.1, for Kodak 649F emulsion, which was commonly used for recording holograms.

Over the linear part of the curve, the relation

$$D = \gamma \log E \qquad \ldots\ldots \quad 5.1$$

is true, where E = exposure, and D = density. The coefficient γ, describes the slope of the curve,

Eqtn 5.1 may be written

$$\log \frac{1}{T_I} = \gamma \log I \qquad \ldots\ldots \quad 5.2$$

where T_I = transmitted intensity

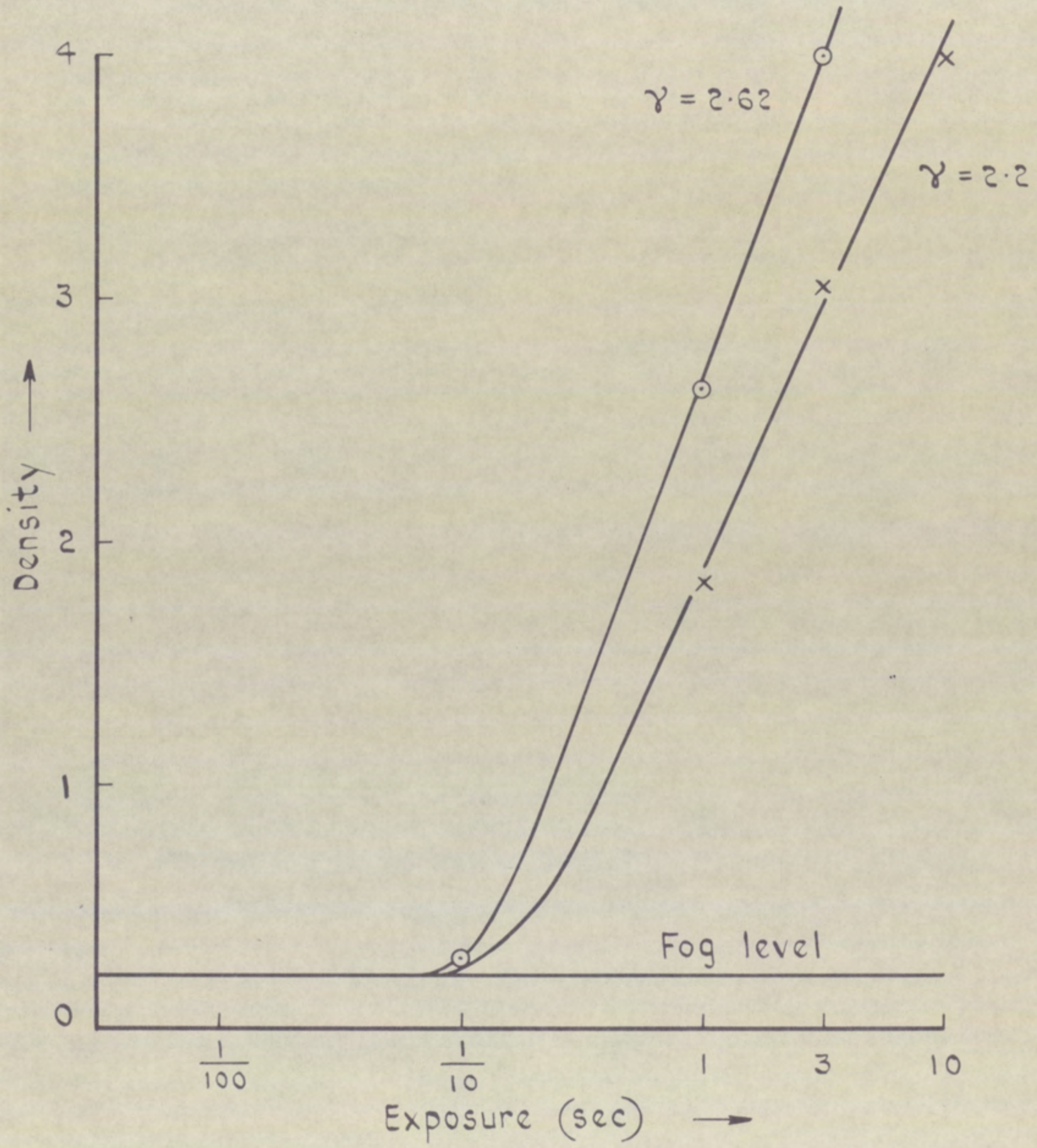


Fig. 5.1 H + D curves for 649F plates

and taking $T_I = t_A^2$ where t_A = amplitude transmittance,

$$\log \frac{1}{t_A}2 = \gamma \log I \qquad \ldots\ldots \quad 5.3$$

Evaluation of 5.3 gives $\frac{1}{t_A}2 = I^\gamma$

$$t_A^2 = I^{-\gamma}$$

$$t_A = I^{-\frac{\gamma}{2}} \qquad \ldots\ldots \quad 5.4$$

If the processing is adjusted so that $\gamma = -2$, then

$$t_A = I$$

and the amplitude transmittance is proportional to the intensity, as required.

Setting $\gamma = -2$ implies that a positive recording has been made, whereas holograms can work equally well as a negative recording, provided that the holograms are made using the oblique reference beam method of Leith and Upatnieks. For holograms made in the original Gabor method, it was shown[1] that setting $\gamma = -2$ gave best contrast in the reconstructed image. Stroke[2] has shown that neither the sign nor the actual value assigned to γ is significant in the recording process. Changing the sign of γ from -ve to +ve merely means that the hologram carrier fringes are shifted in phase by 180°, whilst the value of γ only determines the transmittances of each term in the recorded function. Stroke based his argument on the assumption that the reference beam intensity greatly exceeded that of the object beam and obtained the transmittance as:

$$t_A = 2\,A_o^2 - \gamma A(x)^2 - \gamma AoA(x)e^{j\phi(x) + j\gamma x} - \gamma AoA(x)e^{-i\phi(x) - i\gamma x}$$

where x is a distance in the hologram plane and A_o and Ax are the reference and signal beam amplitudes respectively.

In fourier transform holography it is likely that there will be cases where the object beam intensity may be equal to or greater than the reference beam intensity, and Stroke's arguments would no longer be valid. In addition, it is not reasonable to assume that the relation $D = \gamma \log E$ is true over the working range of the hologram. This may be seen by considering the values of amplitude transmittance over the linear portion of the curve in Fig 5.1. The lowest value of density is about 1.0, corresponding to an intensity transmittance of $\frac{1}{10}$. The amplitude transmittance is then 0.316, and other points on the linear part of the curve will have lower values still. This means that the hologram working range would be restricted to transmittances of 0 to 0.32, and would be inefficient. As the hologram is used as a transparency it is desirable to keep the transmittance as high as possible, and this means working in the "toe" of the Density/log Exposure curve. For this reason, it is more convenient to consider the amplitude transmittance/exposure relationship.

5.2 Relation between amplitude transmittance and exposure - experimental work

The relation between amplitude transmittance and exposure was studied experimentally, partly to examine the extent of a linear relationship, if any, and partly to find the effect of processing on the shape of the curve. Similar work was reported by West and Archer[3] at about the same time as this study was undertaken.

The relation was found by exposing strips of Kodak 649F Spectrographic plate to a range of intensities on different parts of the plate. The intensity was varied by inserting neutral density filters into the laser beam; the exact transmission of the filter having been measured beforehand. The strips of exposed plate were processed using D19 developer, under controlled conditions. Although the strength of developer and the development time were variables, all other factors were held constant. Thus, developer was made using distilled H_2O, and maintained at 20°C

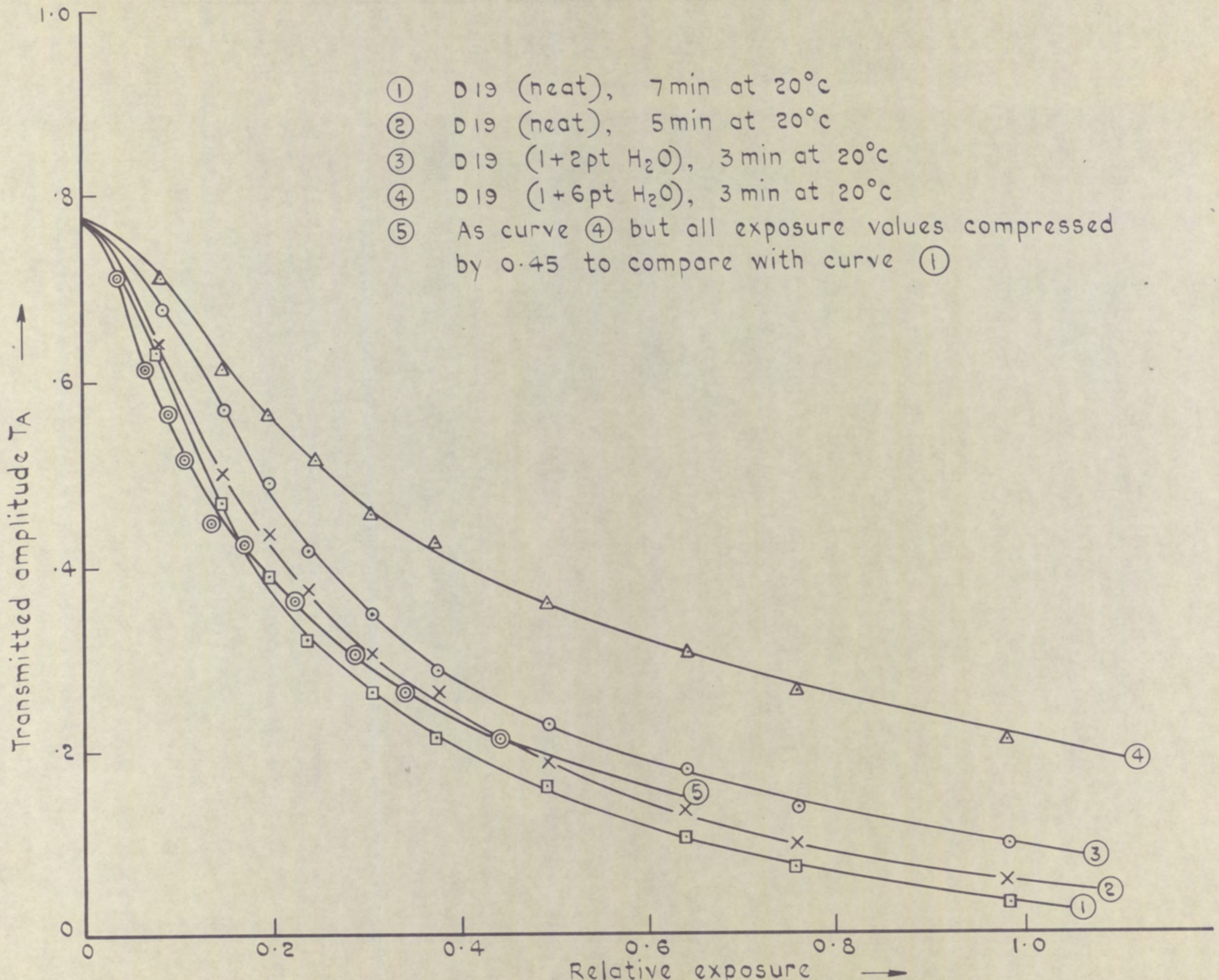


Fig. 5.2 T_A/exp curves for 649F plates

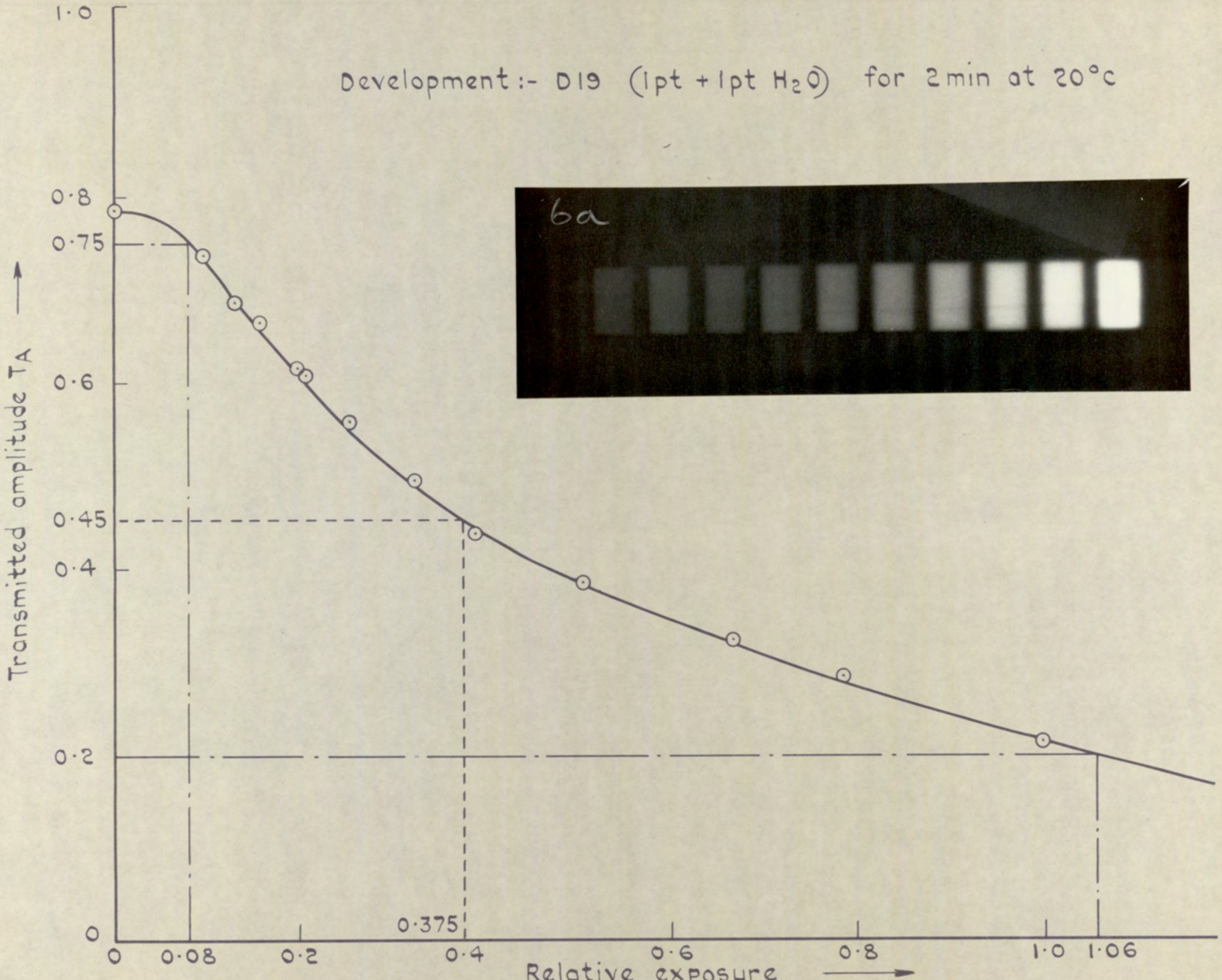


Fig. 5.3 T_A / exp curve for 649F (standard processing)

($\pm \frac{1}{4}^{o}C$), development was followed by rinsing in a stop bath comprising 5% Acetic acid (by volume) and water, and fixed for 2.5 minutes in Amfix ultra rapid fixer. The plates were washed for 10 minutes in running water; no further processing was used to remove the weak purple dye common to these plates.

The intensity transmission of each exposed area of the plate was measured using a Baldwin photometer, and the amplitude transmittance calculated from these figures, allowing a curve of amplitude transmittance versus exposure to be plotted.

Fig 5.2 shows a set of curves covering a range of development conditions from 7 minutes in concentrated D19 to 3 minutes in diluted (1 pt + 6 pts water) D19. Variation between these two extremes is not large, the main effect of weakening the developer being to flatten the curve slightly. Apart from this, the overall shapes of the curves are very similar, but one might perhaps select the lower curves as being most useful since they cover the widest range of amplitude transmittance.

Fig 5.3 shows the amplitude transmittance exposure curve according to development conditions which were adopted as standard for the following work; ie D19 developer mixed with distilled water in ratio 1:1 by volume, and developed for 2 minutes. Also shown is a print of the plate used to obtain the curve, illustrating the small variation in density over the range of intensities given.

The curve of Fig 5.3 shows that there is no linear region, although small variations in intensity will produce linear variation in transmittance between 0.2 and 0.6 transmittance values. The linearity of the curve improves as higher intensities are considered; this corresponds with moving on to the linear part of the H+D curve.

One of the significant results is that as intensities are increased, the variation in amplitude transmittance becomes small, until eventually no variation in amplitude transmittance occurs. The significance of this will be observed later.

5.3 Theoretical model of transmittance exposure curve

Methods of predicting the hologram response on the basis of transmission/exposure curves obtained in section 5.2 are now considered. The main difficulty is in obtaining an accurate model which fits the observed shape of the t_A/exp curves. The procedure adopted by several authors is to obtain an expression relating the amplitude transmittance and exposure in a general form, and then to use a mathematical approximation for expressing the non-linear relationship more precisely.

For example; Bryngdahl and Lohmann[4] used a third order polynomical approximation for the Kodak 649F emulsion, of the form

$$t(E) = \sum_{n=0}^{n=3} t_n E^n$$

where $t(E)$ is the amplitude transmittance at exposure E.

Application of this approximation to the case of the Fraunhofer (or Fourier transform) hologram showed that non-linearities produced higher order correlations and convolutions, although these were not accompanied by experimental results.

An analysis of photographic non-linearities by Friesein and Zelenka[5] used a ν^{th} law approximation for the non-linearity, after using Kozna's model of non-linear recording. This model assumes that the non-linearity acts only on the signal part of the incident intensity, as shown in Fig 5.4.

The non-linearity $g(x)$ was assumed to have a transformation $G(\xi) = \int_0^\infty g(x)e^{-\xi x}\,dx$, and the output from the non-linearity was represented by $\nu(x)$ where

$$\nu(x) = \frac{1}{2\pi j} \quad G(\xi) \exp^{[\text{signal function}]} d\xi$$

It was shown that $\nu(x)$ could be evaluated to an integral containing $G(\xi)$ and the exponentials in the form of a Bessel function of first kind. The presence of the Bessel function indicated that there would be several

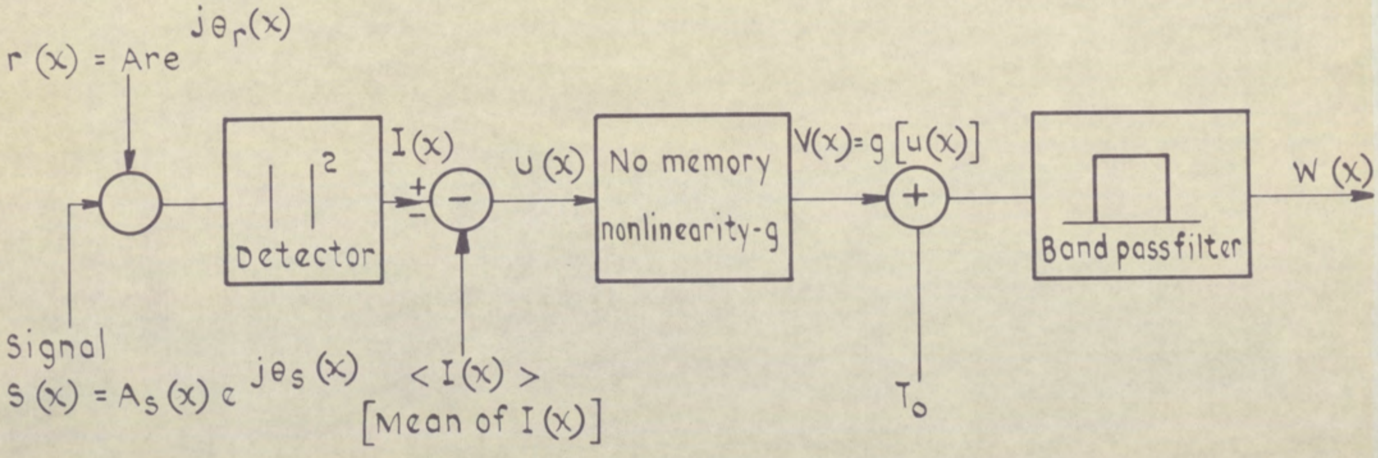


Fig. 5.4 Friesem & Zelenka's model of nonlinear recording

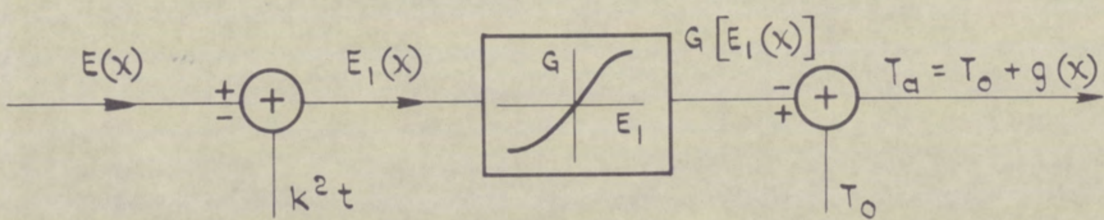


Fig. 5.5 Kozma's model for analysing film nonlinearities

Fig. 5.6 Effect of nonlinear recording of carrier fringes

orders of reconstructions from the hologram, but in their analysis, Friesim and Zelenka assumed these orders to have been removed by spatial filtering, as shown in Fig 5.4. The first order image was then evaluated using a ν^{th} law approximation such as $g(x) = x|x|^{\nu-1}$ for $v \geq o$. From their results, it was deduced that extra images would be seen in the neighbourhood of the true reconstruction, and these were predicted in the case of an object consisting of two point sources.

The initial approach used by Friesem and Zelenka was originally adopted by Kozma[6], who has contributed much useful work on the subject of non-linear photographic recording. In this paper[6] however, after deriving the transmittance in terms of a Bessel function and the non-linearity G(w), an error function limiter was used to express the non-linearity more specifically. The form of the error function limiter was given as

$$G(\zeta) = \frac{2L}{\delta(2\pi)^{\frac{1}{2}}} \int_0^{\zeta} e^{-\lambda^2/2\delta^2} \, d\chi$$

and is shown in Fig 5.5. This function was substituted for the non-linearity in the expression obtained for the transmittance, and it was again shown that several higher orders of reconstruction were produced, which could be neglected provided that they did not overlap the first order reconstruction. Further analysis showed that the first order reconstruction contained an amplitude distortion term which depended on the severity of non-linearity, and that for small signal amplitudes, the distortion is reduced. It was shown that small departures from linearity could have two effects, one being to widen the spatial frequency spectrum and hence possibly cause overlap between the zero and first order images, the other being to modify the amplitude of the first order, which in spatial filtering applications would cause a reduction in efficiency of the filter. For large departures from linearity complete amplitude distortion occurs (ie no amplitude information is recorded), and only phase information remains in the hologram. For the case of matched filtering, the latter degree of non-linearity was discussed in an earlier paper by Kozma and Kelly[7] where the severe non-linearity was represented by a hologram whose amplitude transmittance had only two values -1 and +1, -1 being represented by a clear region, +1 by an opaque region. The results from such a hologram were that a 1st order image or autocorrelation function peak appeared flanked by several secondary maxima. There was also a certain amount of loss in efficiency and very low signal to noise ratios would have to be treated using an amplitude matched filter.

In a more recent paper[8] Kozma analysed the film non-linearities more thoroughly. In this paper the incident intensity was expressed in terms of a function$^{(B(x))}$, which describes the normalised deviation from the mean of the intensity produced at the hologram by light scattered from the object, ie:

$$B(x) = \frac{a^2(x) - \langle a^2(x)\rangle}{\langle a^2(x)\rangle} \quad \text{where } a^2(x) \text{ is the intensity at } x$$

In the case where the beam ratio is large, it was shown that the amplitude transmittance was linearly proportional to the recorded intensity, and that faithful reproductions of the original object would be produced.

However, where the beam ratio is not sufficiently large, the linear approximation is not valid, and the transmittance has to be expressed as a function of exposure ie:

$$t_a(E') = t_b + F(E)$$

Where t_b is the transmittance at the bias point on the t_A/E curve. The function $F(E)$ was then treated in a similar way to that used in Kozma's earlier paper[6], from which it was deduced that non-linearities of the first order term depended on the magnitude of $B(x)$ given above; for large values of $B(x)$ additional "intermodulation" terms were predicted in the first order image. The existence of higher order images was also predicted from these results, but provided that they are separated by having a high spatial carrier frequency, they can be neglected.

The non-linearity was also treated using Taylor series to expand the F(E) function; from this the transmittance was obtained as a series of polynomials involving B(x) and constants depending on the series expansion of the t_A/E curve and on the beam ratio. This argument depended on the assumption that F(E) could be represented by a power series,

$F(E) = \sum_{v=1}^{\infty} s_v E^v$ derivatives of which were used in the Taylors expansion.

The range of exposure over which the power series was considered valid was calculated, using statistical methods. From this series the coefficients of the polynomials were calculated and their relationship to the beam ratio shown; the decrease of these coefficients as the beam ratio was increased was particularly marked.

The effect of the non-linearity was shown to produce additional images alongside the undistorted image and the appearance of the additional images depends on the type of object; for a diffusely illuminated object produces a random spread of light about each image point, whereas a periodic signal generates intermodulation products which are themselves periodic.

Other papers have been published[9,10] which also discuss the photographic process, although the non-linearities were not discussed in detail.

The main conclusions to be drawn from the theoretical arguments are that two kinds of effects of non-linearity are produced; one being the production of higher order images than the first, the other being the production of images surrounding or overlapping the first order images. The first need not give any trouble since the higher order images can be removed by spatial filtering, provided they are separated by using a sufficiently high spatial carrier frequency. These higher orders can be thought of as arising from distortion of the hologram fringes themselves rather than of the signal. This may be seen in the following way:-

Let the amplitude transmittance be represented by:-

$$t(E) = t_B + F(E)$$

Where t_B is some constant and $F(E)$ is the transmittance at exposure E. The function $F(E)$ can be represented in a number of ways; in particular, it can be expanded into a power series of the form:-

$$F(E) = A_o + A_1E + A_2E^2 + A_3E^3 \quad + \ldots\ldots \text{ (Maclaurin's Theorem)}$$

the value of E in each term is given by:-

$$E = t\left[|A_r|^2 + |A_s(p)|^2 + 2A_rA_s(p) \text{ Cos } [\propto p - \phi(p)] \right]$$

Where A_r is the reference beam amplitude, $A_s(p)$ is the object beam (fourier transform) amplitude and Cos $[\propto p - \phi(p)]$ represents the sinusoidal variation in intensity caused by interference. Neglecting the influence of $|A_s(p)|^2$, it can be seen that there will be terms containing Cos $[\propto p - \phi(p)]$, $\text{Cos}^2[\propto p - \phi(p)]$, etc, in the expansion of $F(E)$. Since $\text{Cos}^2 [\propto p - \phi(p)] = \frac{1 - \text{Cos}^2 [\propto p - \phi(p)]}{2}$ the angle through which the non-linear terms are diffracted is represented by $2(\propto p - \phi(p))$, $3(\propto p - \phi(p))$ etc, (the actual angles being found by putting $\propto = \frac{2\pi\theta}{\lambda}$ where θ is the angle of incidence of reference beam to hologram). Thus four, six, eight, etc, diffracted orders would be expected, depending on the degree of non-linearity.

Although such orders can be eliminated in optical processing systems simply by spatial filtering (as described in Chapter 2) it is of interest to examine an example of this non-linearity. The example was originally reported by the author[11] in connection with character recognition, and shows the response of a hologram made from the fourier transform of an "S". This is shown in Fig 5.6. The upper line of images shows the impulse response of the hologram, and consists of the reference beam focus, flanked by the reconstructions of the letter "S". Further out still are two more images of the reference beam focus, one on each side. In the original experiment, a further image could be seen, corresponding to the image of the letter "S" again. Note that the side images of the reference beam focus have another term associated with them which consists of the autocorrelation of the letter "S". This is not seen in the central term because of the large amount of scattered light.

The lower row of images shows the response when the hologram is illuminated by the transform of the letter "S". Either side the undiffracted image is the autocorrelation (on left) and autoconvolution (on right). These latter two functions appear similar because there is two fold rotational symmetry for the letter "S", and hence it's convolution and correlation patterns are identical. Further out from these two images are faint images of the original letter "S" caused by non-linearity, one on each side of the pattern. Note that the images are in 'line' with the corresponding series of images formed by the reference beam. This suggests that the output of the hologram is the same irrespective of whether the reference beam or object beam is used, the only change is in the position of the zero order image. This is only strictly true, however for the centre pair of images in each response when there is no non-linearity present, and is confirmed by the theoretical expressions for these images, obtained in Chapter 3. A mathematical explanation for the appearance of these images is given in Appendix A.

5.4 Effect of Non-linear recording on first order images

Although many of the papers discussed in the previous section have discussed the effect of photographic non-linearity on the first order image the nature of the effect, in the case of hologram filters for optical processing, was not fully examined. In this section the probable result of non-linear effects are considered, and experimental results presented in the following section. Some of this work has already been reported by the author[12].

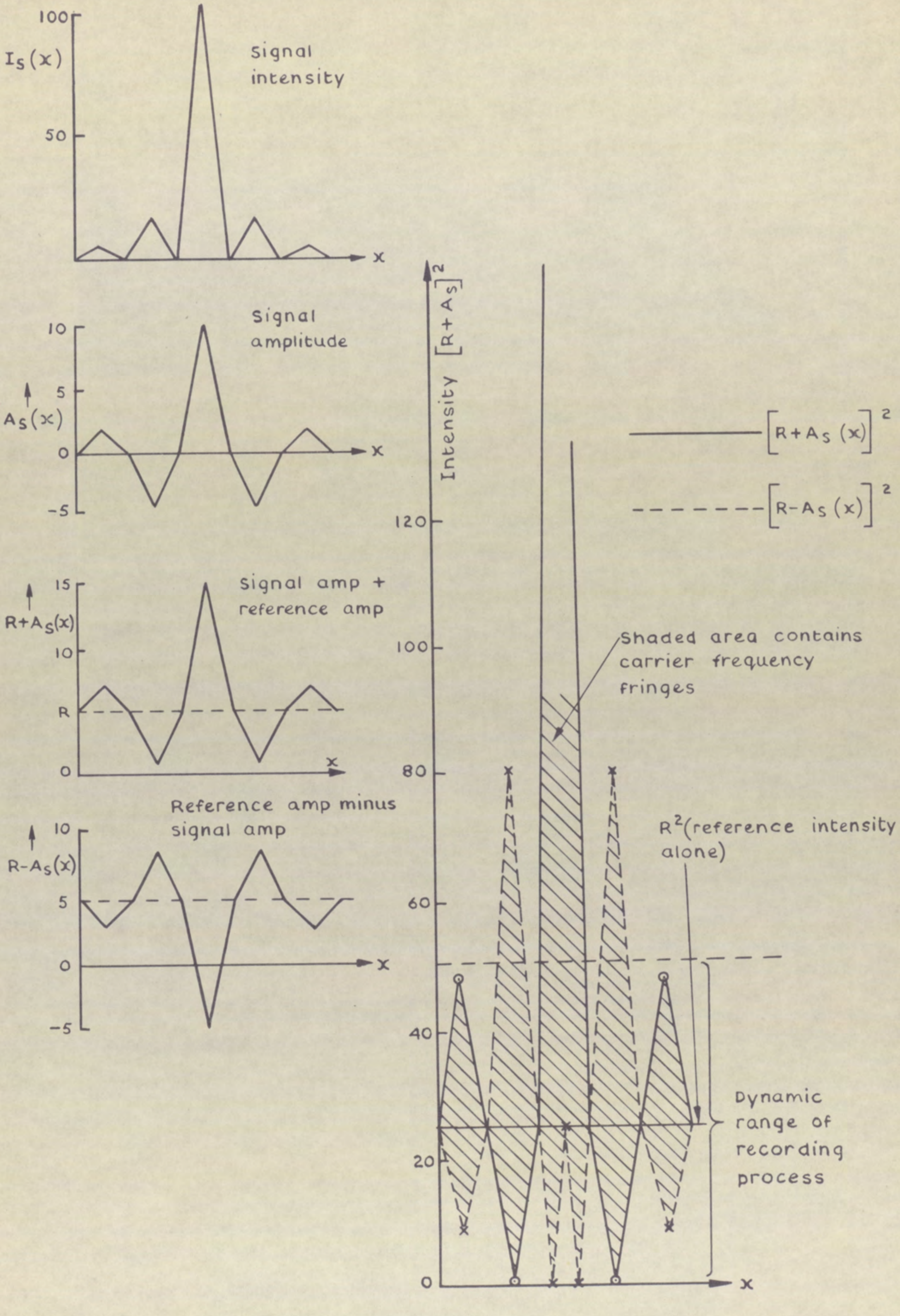


Fig. 5. 7a Formation of hologram carrier fringes

Before looking at the actual recording process, it is convenient to consider how the intensity distribution in the hologram plane is obtained. In particular, one is interested in the shape of the envelope containing the peaks of interference or carrier fringes. Since faithful recording of these fringes ensures an accurate recording of the signal transform.

The formation of the carrier fringe envelope is shown diagrammatically in Fig 5.7(a). The signal transform is assumed to have an intensity distribution $I_s(x)$, and a corresponding amplitude distribution $A_s(x)$. It is also assumed that a reference beam of amplitude R, is added to the signal transform. In the formation of the carrier fringes, there will be amplitude maxima wherever the signal and reference beam are in phase, and the amplitude distribution describing the envelope containing these maxima is given by $[R + A_s(x)]$, and is shown in Fig 5.7(a). Where the reference and signal are completely out of phase will be represented by the envelope containing the minima of the interference fringes. This will be given by $[R - A_s(x)]$ and is also shown in Fig 5.7(a). By taking the squares of these two functions, ie: $[R + A_s(x)]^2$ and $[R - A_s(x)]^2$ one obtains the envelope containing the intensity distribution of the carrier fringes, shown on the right of Fig 5.7(a). The curve for $[R + A_s(x)]^2$ is similar in shape to that of the original signal $(I_s(x))$ but the curve for $[R - A_s(x)]^2$ appears to be distorted. However, if these fringes were recorded correctly, with particular regard to intensity, the original signal could be restored. It may be readily seen that if the dynamic range of the recording process is restricted to a range intensity around R^2, as shown on Fig 5.7(a) then the centre part of the signal transform will be severely distorted. For the less intense peaks of the signal transform there will merely be a flattening of the centre of the peaks, while for higher intensities there could be a decrease in intensity at the centre. This distortion can become quite severe, as shown later in this chapter.

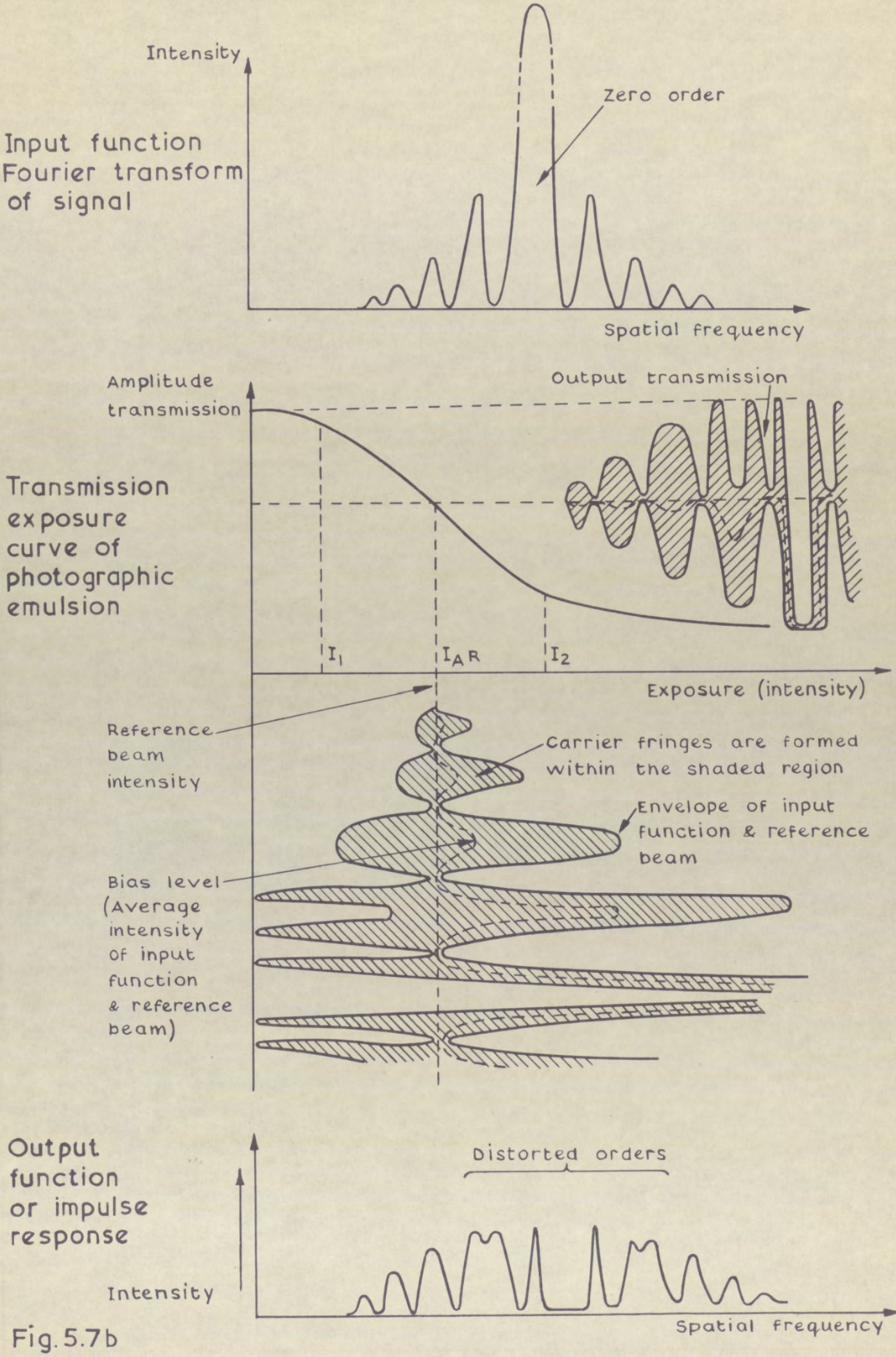

Intensity
Zero order
Input function
Fourier transform
of signal
Spatial frequency
Amplitude
transmission
Output transmission
Transmission
exposure
curve of
photographic
emulsion
I1
IAR
I2
Exposure (intensity)
Reference
beam
intensity
Carrier fringes are formed
within the shaded region
Envelope of input
function & reference
beam
Bias level
(Average
intensity
of input
function
& reference
beam)
Output
function
or impulse
response
Distorted orders
Intensity
Spatial frequency

Fig. 5.7b

For the mechanism of the recording process it is convenient to refer to the characteristic curve of amplitude transmittance versus exposure, shown in Fig 5.7(b). Assuming that the reference beam intensity is represented by I_R, the carrier fringes will be formed about this point. For a fourier transform, the intensity is typically of the form shown in the upper part of the diagram, being very intense in the centre (zero spatial frequency) and relatively weak in the wings (high spatial frequencies). When this function is added to the reference beam interference occurs and the resulting intensity distribution is shown at the bottom of the t_A/Exp curve, being derived as shown in Fig 5.7(b). The fringes do not oscillate with equal amplitude about I_R because the mean intensity of the fourier transform itself, has to be included. Instead, the fringes oscillate about the mean intensity $(I_R + \bar{I}_s)$ where $\bar{I}_s$ is the average intensity of the fourier transform. Returning to the notation used in chapter 3, $\bar{I}_s$ is equal to $|A_s(p)|^2$, that is, the signal amplitude squared. The mean intensity $[I_R + |A_s(p)|^2]$ is therefore a function of spatial frequency (p) and fluctuates according to the fourier transform intensity. It is this variation of the bias point which complicates the non-linear analysis, since otherwise the bias point could be fixed at some relatively linear portion of the t_A/Exp curve.

The shaded envelope to the bottom of the t_A/Exp curve indicates the region containing carrier fringes; if this envelope were exactly reproduced then no distortion would occur in the hologram transmittance.

The variation in transmittance is shown to the right of the curve. The effect of the t_A/Exp curve is to suppress variations of high intensity so that some carrier fringes are not recorded, or else very weakly recorded. It is the distortion of these carrier fringes which gives rise to the non-linear effects observed in the hologram response. [Note; the distortion referred to here, is not distortion of the shape of the carrier fringes, since this produces non-linear effects discussed earlier. Instead it is distortion of the envelope containing these fringes.] The impulse response of the hologram, assuming a distorted transmission similar to that shown in the diagram, is shown at the bottom of Fig 5.7(b). The absence and attenuation of the carrier frequency fringes has caused some terms of the transform to vanish, while others have been deformed, particularly at the centre of the term, where the highest intensities are located.

The effect of non-linear recording is therefore to suppress the highest intensities in the recorded transform, the degree of attenuation increasing with increasing intensity, until no information at all is recorded. This has the effect of dividing or splitting some of the terms of the transform which then produce diffraction effects at higher angles than those expected from the original transform. This in turn causes extra images to appear in the neighbourhood of the linearlly reconstructed image.

The effect of non-linear amplitude distortion is also related to the fact that the intensities which are heavily suppressed are also those which occur at the lowest spatial frequencies in the majority of cases. This has several results, an important one being that the zero spatial frequency is almost completely attenuated. This can be of good use, since it corresponds to subtracting the mean or dc levels in the signal and input, and is desirable in pattern recognition (as shown in Chapter 1). To a certain extent, the amplitude distortion also corresponds with the recording of the noise spectral density, which was also mentioned in Chapter 1. This correspondence is inprecise, however, since the transform is not usually a smoothly decreasing intensity function, but contains many peaks as in Fig 5.7. Nevertheless, the suppression of low frequencies does improve the filters performance to some extent, particularly where the higher spatial frequencies contain most of the differences between the pattern being sought and other inputs.

The suppression of low frequency terms in the hologram may also be regarded as a differentiation of the signal and input, this aspect having been reported by Lowenthal and Belvaux.[13,14] In these papers, the differentiation could be performed by using a filter of the form

$$u^2 \, S^*(u)$$

where $S^*(u)$ is the complex conjugate of the fourier transform of the signal, and is produced by holographic means, while u^2 is a parabolic transmission function, being zero in the centre of the transform, (u is spatial frequency variable). It was shown that the application of a hologram of this type acted on the high frequencies of the object spectrum and hence gave better discrimination. A hologram of the form $u^2 \, S^*(u)$ could be made approximately by overexposing the lower frequencies in the manner described earlier in this section. The u^2 function would be better made as a separate exposure however, if the non-linear effects were to be kept to a minimum.

The non-linear effects are produced by the curvature of the t_A/Exp curve at high intensities. If the intensity of the signal were reduced so that the bias point $[I_R + |A_s(p)|]^2$ always lay near the midpoint of the t_A/Exp curve, then this distortion would vanish. Although this can be done in practice by using neutral density filters, it also means that the weak high spatial frequency parts of the transform are not recorded.

The modulation transfer function of the emulsion was approximately unity at the carrier frequency concerned (from manufacturers data) and may thus be ignored. The contrast of the recorded carrier fringes will then depend only on the contrast of the incident light and if this contrast is too low it appears that the signal diffracted by these recorded fringes is too weak to be detected.

The emulsion has the effect of a low intensity limit as well as an upper intensity limit and means that the hologram will record a range of intensities which will correspond to a spatial frequency bandwidth of the fourier transform.

The width of the spatial frequency band will vary depending on the strength of the object signal, since if the latter is high, then the emulsion will record a wide range of relatively low intensities at high spatial frequencies. Conversely, if the signal intensity is low, then the emulsion records only the small range of frequencies at relatively higher intensity. This follows from the distribution of energy in the fourier transform, shown in Fig 5.14 for a rectangular aperture.

The effect of non-linear recording on the autocorrelation response of the hologram will therefore depend on exposure and on the distribution of spatial frequencies in the transform. These relationships are studied in the next section.

5.5 Experimental study of non-linear effects

5.5.1 Comparison of actual and recorded transforms

The apparatus used for comparing the transforms was described in Chapter 4, section 4.3. The comparison was made by recording the original fourier transform in the hologram plane using the photo-electric scanning device (Section 4.3.4) and comparing this with the record of the fourier transform when scanned in plane P4 (Fig 4.8). This comparison allowed the effect of the optical system to be examined. The object in this study was a rectangular slit, having a clear aperture. The reasons for choosing this object were that it's width and length could be easily adjusted and that no phase variations were present across the slit aperture, as possibly the case with photographically recorded objects. Other shapes of aperture were considered, in particular, the circular aperture. This latter was not used here, because the intensity varies

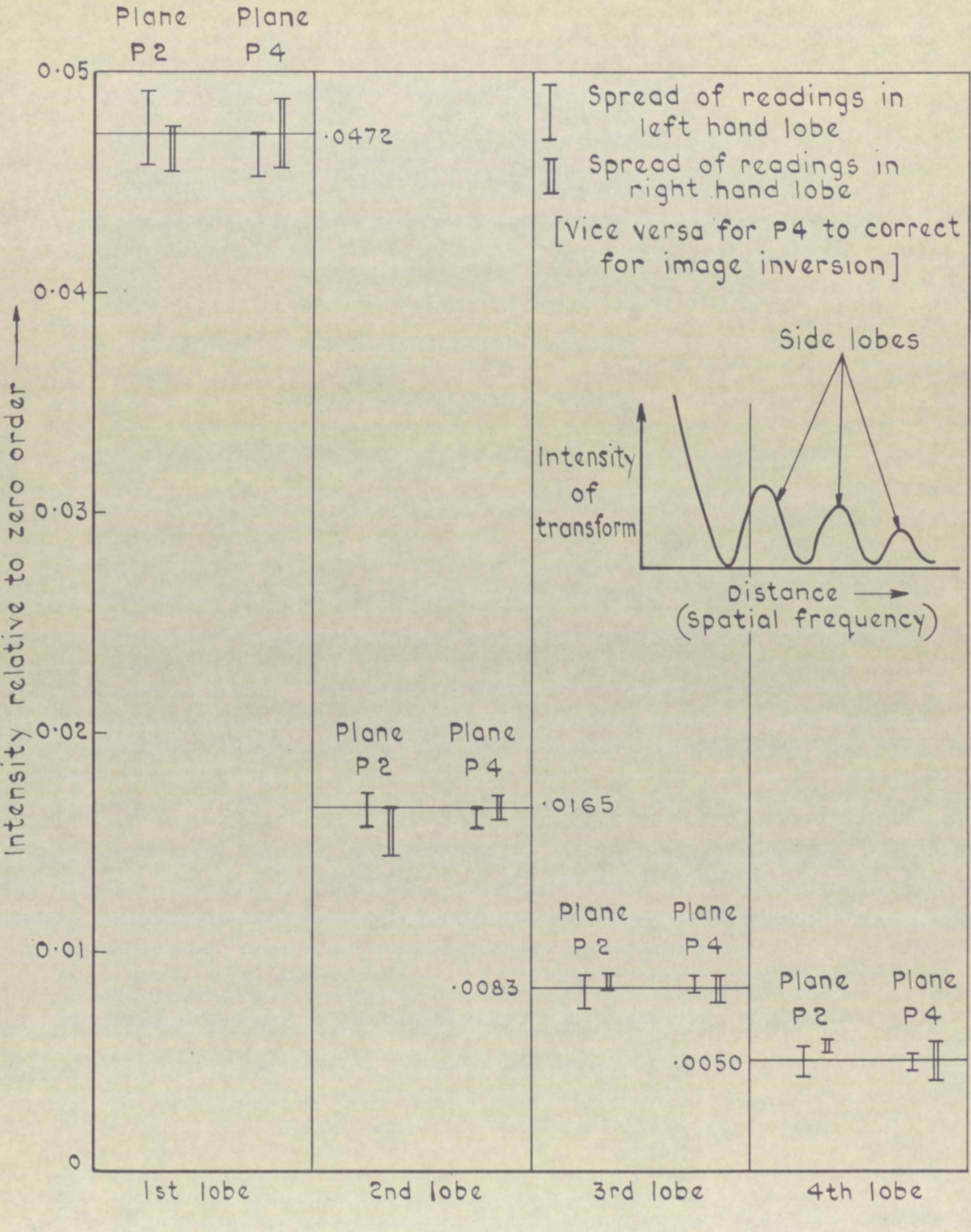


Fig. 5.8 Measurement of Fourier transform intensities for planes P2 and P4

along the radius of the transform, and means that small vertical displacements of the photomultiplier scanning unit whilst scanning in a horizontal line, could give erroneous results. On the other hand, the transform of a rectangular aperture contains intensity variations in the horizontal direction which are constant for any positions along the vertical direction and the small vertical displacements when scanning can be tolerated.

The output current of the photomultiplier was measured using a digital instrument. The photomultiplier slit was set at approximately 5μm, and actually measured as 6.53μm, whilst the object aperture width was 0.287cms, so that the diffracted lobes of the transform were approximately 150μm apart, giving a scanning slit resolution of about $\frac{1}{23}$rd the width of fourier transform lobes.

Fig 5.8 shows the results obtained by measuring the maxima of the first four side lobes of the diffraction pattern for both the actual transform (in plane P2) and the "re-imaged" transform (in plane P4). The results are presented as being values relative to the zero order, but in fact they are only relative to each other, the zero frequency intensity not being recorded because it was too intense. The figures shown were related to theoretical values by multiplying by a normalising factor. This factor was simply the ratio of total of theoretical values to the total of measured values, as shown below:-

The theoretical values are[15]:-

$$\left.\begin{array}{l}\text{1st side lobe maxima } 0.0472\\ \text{2nd side lobe maxima } 0.0165\\ \text{3rd side lobe maxima } 0.0083\\ \text{4th side lobe maxima } 0.0050\end{array}\right\}\quad \begin{array}{c}\text{Central intensity}\\ = 1.0000\end{array}$$

The total is 0.0770, and the normalising factor for each reading is therefore $\left[\dfrac{0.0770}{\text{Total of all four results}}\right]$

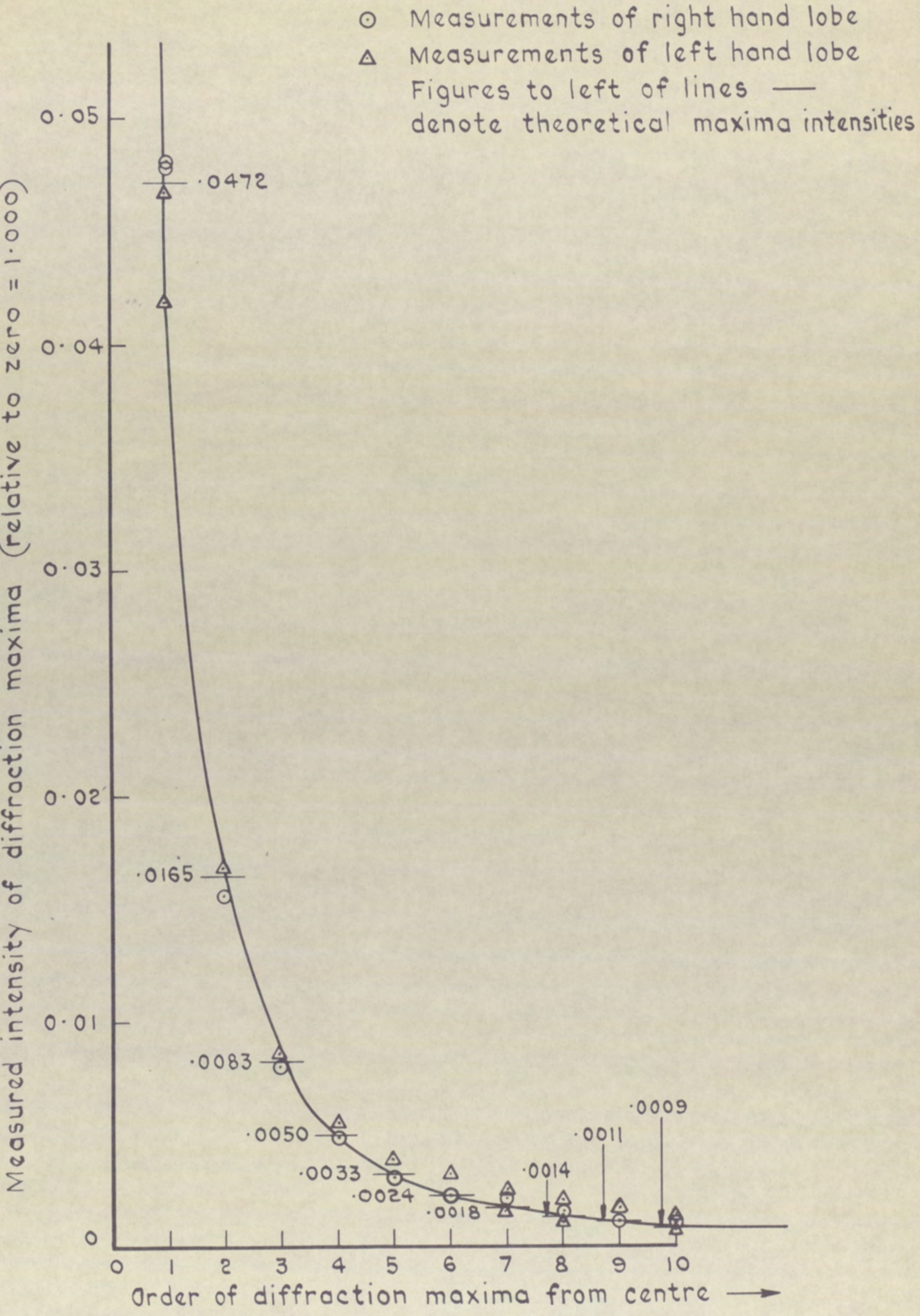


Fig.5.9 Measurement of reimaged transform (plane P 4) and compared with theoretical values

The table below is included to illustrate the normalisation:-

	1st Max	2nd Max	3rd Max	4th Max	Total
(Actual reading (μA)	5.88	1.970	1.083	0.696	9.629
Normalised	0.0471	0.0158	0.00867	0.00558	0.0770

The results are shown in Fig 5.8 for both halves of the transform and for both planes P2 (hologram plane) and P4 (re-imaged transform). Although there is considerable scatter of results, there is fairly good agreement between theoretical and measured results. There was no significant difference between plane P2 and plane P4, except that the greatest scatter sppears to be on the same side of the zero order in both transforms, irrespective of the image inversion. It was concluded that within the limits of experimental accuracy, that both transforms were identical.

Further comparison of the re-imaged transform with theoretical results was then made, over a wider range of spatial frequencies. The readings at each diffraction maximum were measured as before, with a digital voltmeter, and normalised in the same way, using the total of theoretical values of diffraction maxima intensity. These were calculated, using a simple computer programme which was designed to find the intensities of successive points in the rectangular aperture diffraction pattern. The intensities of the first 50 maxima were found, and using these results and the total 0.088030 for the first 10 maxima, the results shown in Fig 5.9 were obtained.

Again, there is some scatter of measured results about the theoretical values, but in general the agreement is fairly good. One set of readings for the left hand lobe is in error to a large extent, otherwise most readings are in agreement with theory to within 10%.

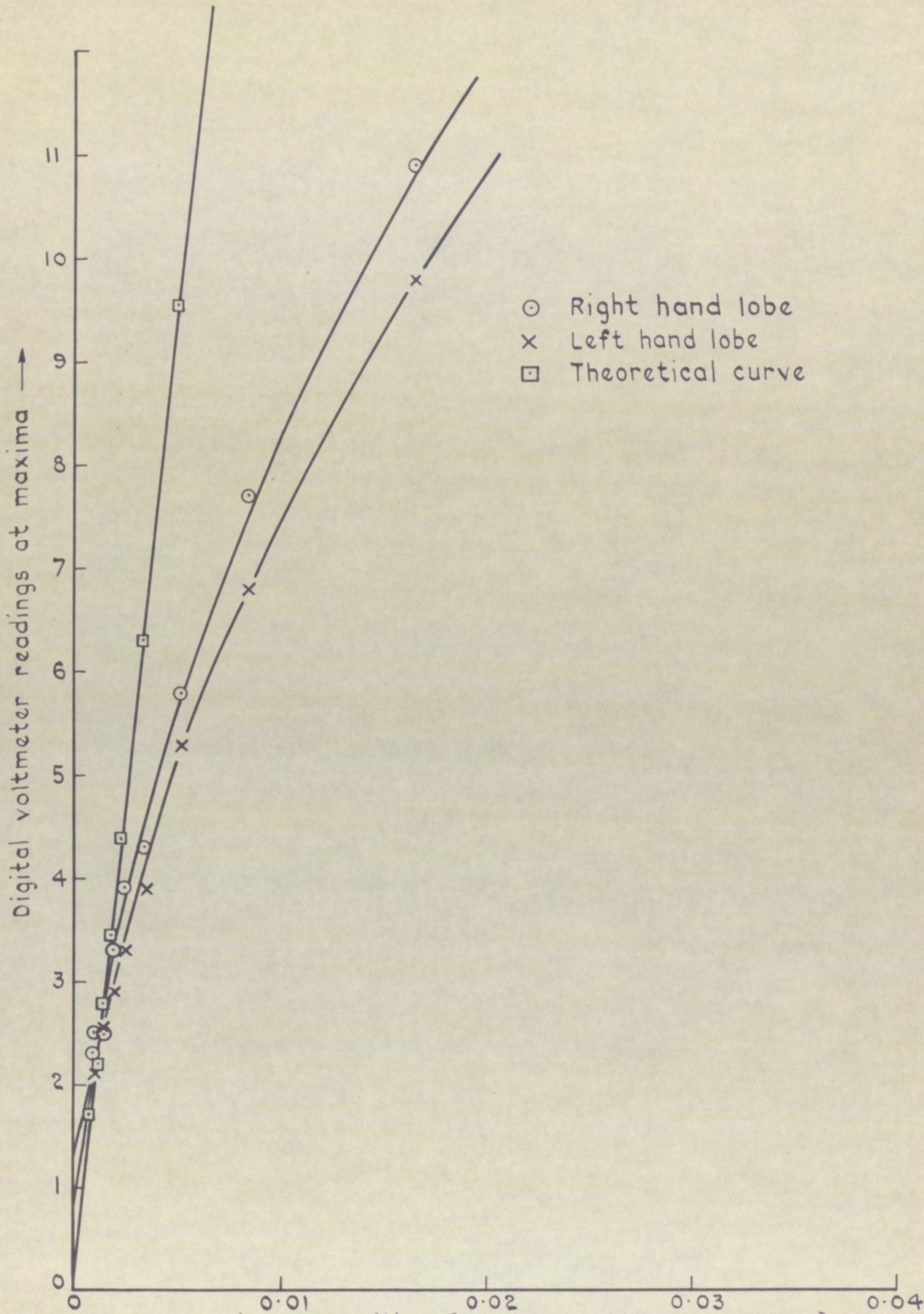


Fig. 5.10 Effect of nonlinear recording on intensities of maxima in reimaged transform

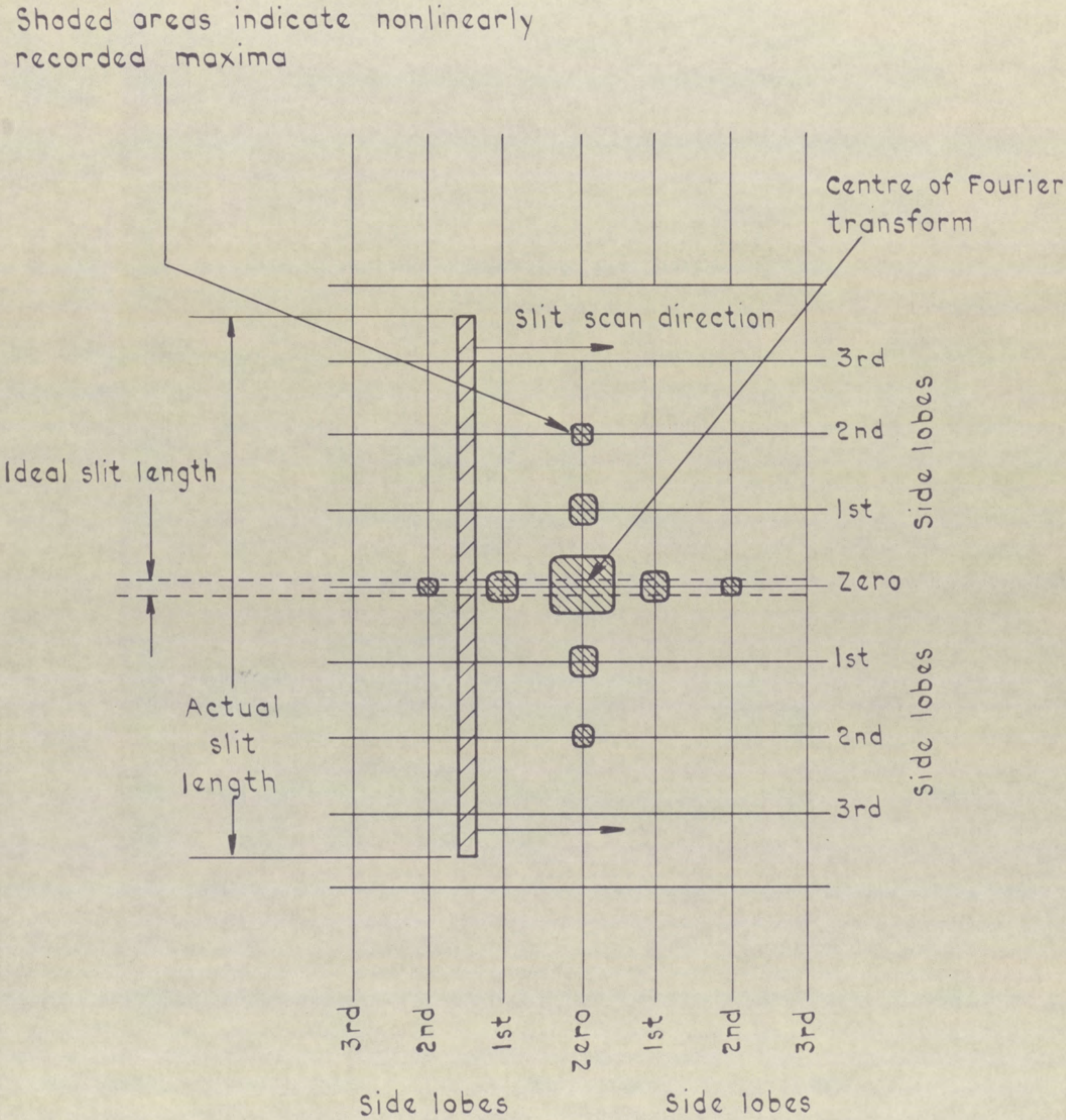


Fig. 5.11 Plan of reconstructed transform showing integrating effect of long slit

Having checked that the correct distribution of intensity was being obtained in planes P2 and P4, and the probable extent of error; the distribution of intensity in the reconstructed transform from a hologram was measured. Several holograms were made, each using the experimental arrangement shown in Fig 4.2. Measurements were made on the hologram whose reference beam density corresponded to an amplitude transmittance of 0.4. The object beam exposure was sufficient to overexpose the zero and first order lobe, so that some non-linearity could be expected. The carrier fringes were found to be 4.3μm apart so that there would be about 35 fringes for each side lobe of the diffraction pattern. The intensities at the maxima of the reconstructed transform are shown in Fig 5.10, with the theoretical values as an abscissa.

The curve for the right hand lobe is slightly higher than the left lobe, but this could be the effect of laser intensity variations during measurement. The curves are in general agreement as to the overall shape, and show a marked discrepancy when compared with the ideal curve, which was plotted using a constant factor derived from the readings for the 5th-10th maxima. If no non-linearity were present, then the readings obtained for the hologram would lie on a straight line, showing proportionality with the theoretical intensities. As expected, the degree of non-linearity increases as the intensity is increased, and one would expect the experimental curve to take on a negative slope, falling to zero when saturation of the plate occurs.

The disadvantage of this measurement was that intensity variations were not confined to the one direction in the transform, but varied in two orthogonal directions, corresponding with the edges of the rectangular aperture. This meant that non-linearities would occur in both directions, although the results were regarded as applying to only one direction. This is illustrated in Fig 5.11. The length of the slit was such that it integrated over a large number of maxima in the vertical 'q' direction.

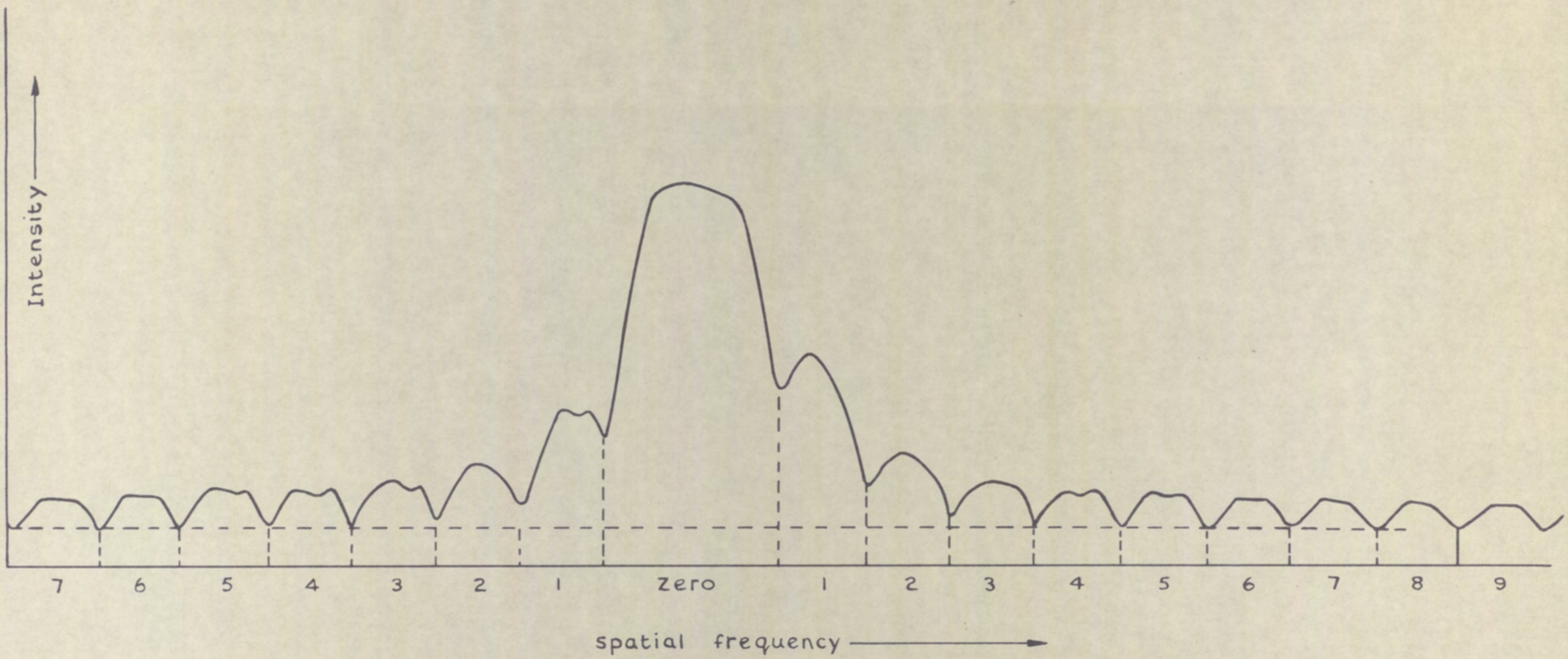
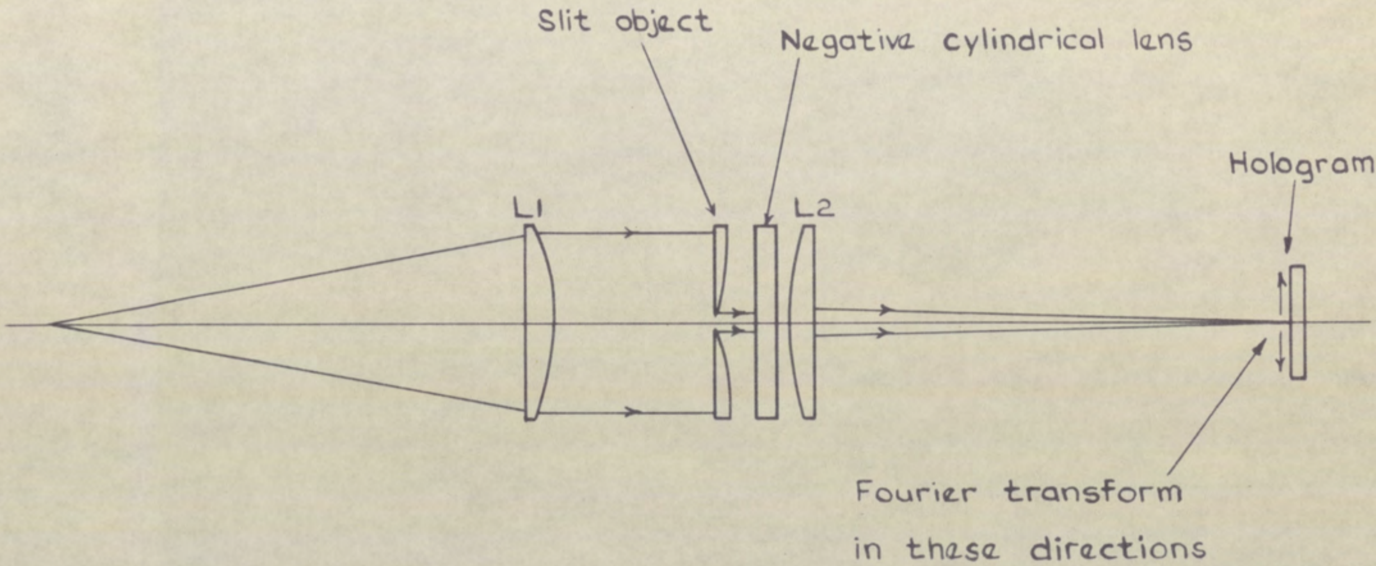
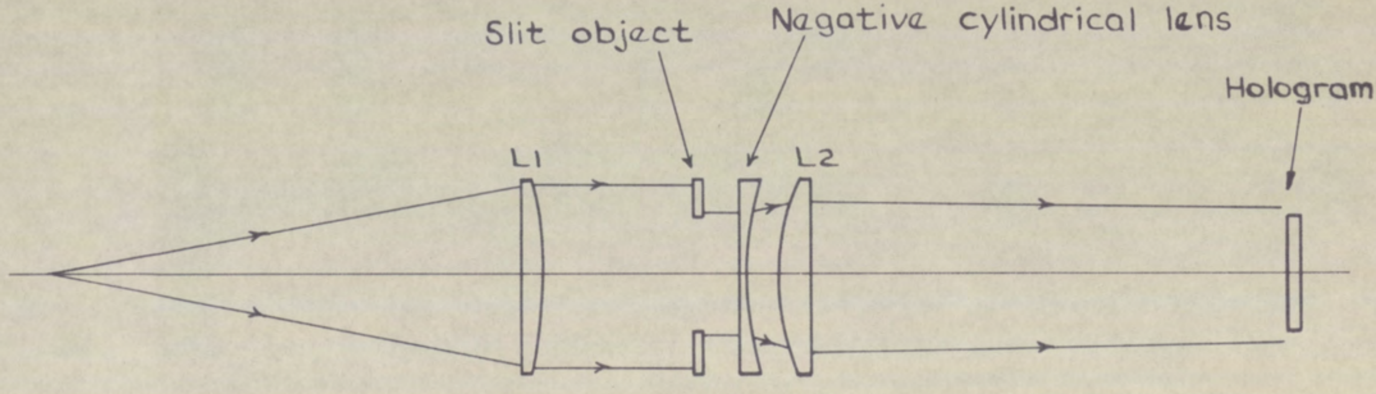


Fig. 5. 12 Intensity distribution in reconstructed transform of rectangular aperture



Plan view



Side view

Fig.5. 13 Optical arrangement for producing one dimensional fourier transforms

As a result, the decrease in intensity of the non-linearly recorded maxima were partly compensated by the intensities of linearly recorded maxima on either side in the vertical direction. The effect of this is shown in Fig 5.12, where the intensity distribution is shown for holograms where first three maxima were completely overexposed. The flattening of the higher orders 6-9 is expected, and one can see the dip in intensity in the centre of the orders 4, 5 also as predicted by non-linear recording (see Fig 5.7). However, this dip vanishes in the central orders, zero, 1, 2 and 3, because it is compensated by the integrating effect of the slit.

There are several ways of overcoming this difficulty; the obvious step of reducing the slit length to a pinhole was not chosen because

a) it would be difficult to ensure that one tracked the same line of maxima without moving in the orthogonal direction, and

b) there would be a reduction in signal strength which would require amplification and introduce noise problems.

The alternative is to eliminate intensity variations in the orthogonal ("q") direction by using a cylindrical lens to form a transform whose intensity varied only along the one direction. The results of this method are discussed in the following section.

5.5.2 Comparison of transforms using a one dimensional transform

The apparatus used for preparing one dimensional transforms is shown in Fig 5.13. Of the cylindrical lenses available best results were obtained using a negative lens in conjunction with the original spherical lens. The object for this experiment was a slit 2.3mm wide by 15mm long, and the fourier transform consisted of equally spaced bright lines corresponding to the diffracted orders of the transform. The intensity distribution in the actual transform is shown in Fig 5.14; opposite side lobes were not exactly equal in intensity, but the distribution of intensity in the maxima was found to be in reasonable agreement with theory. The differences in

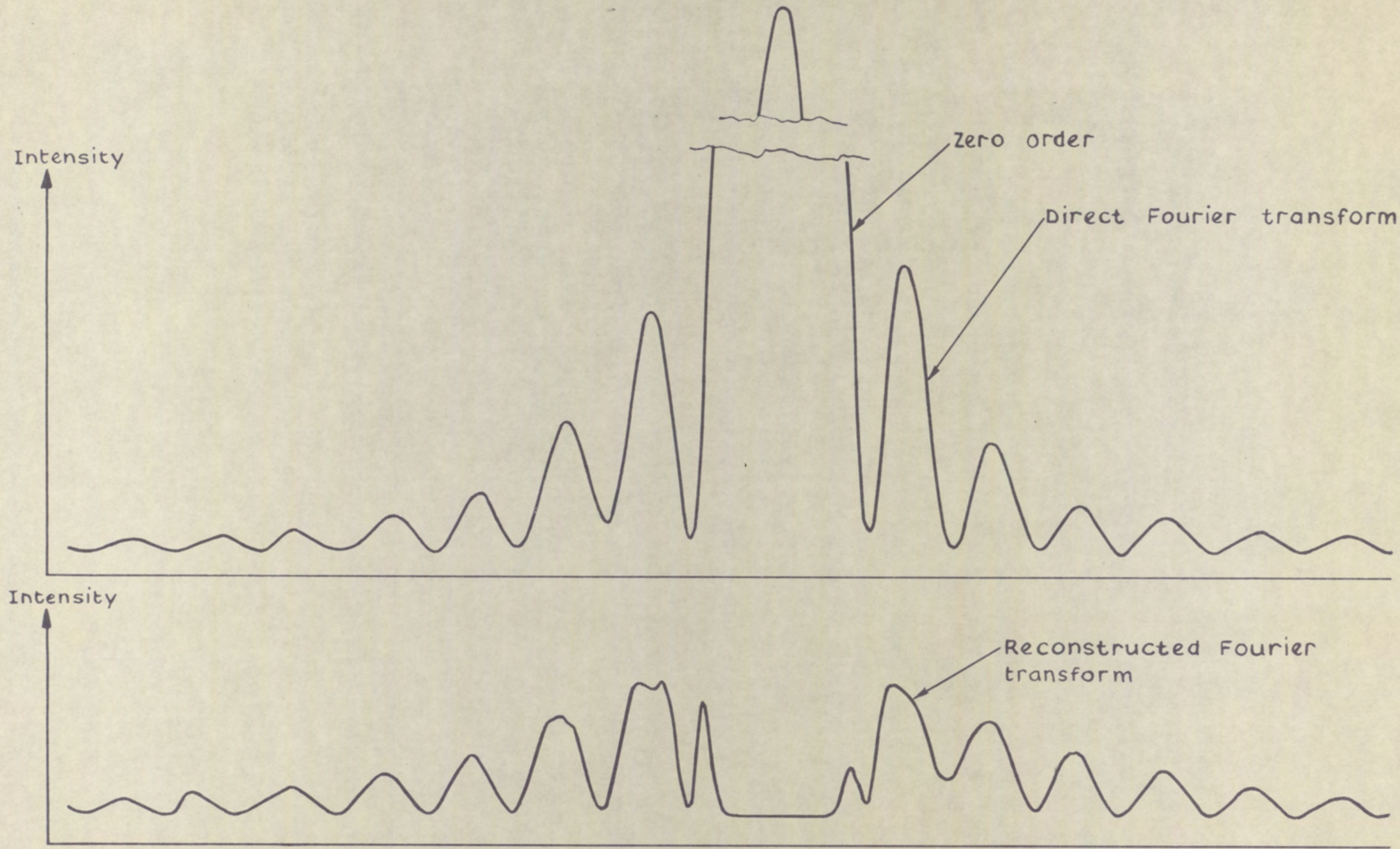


Fig. 5. 14 Comparison between impulse response of slit hologram and direct transform

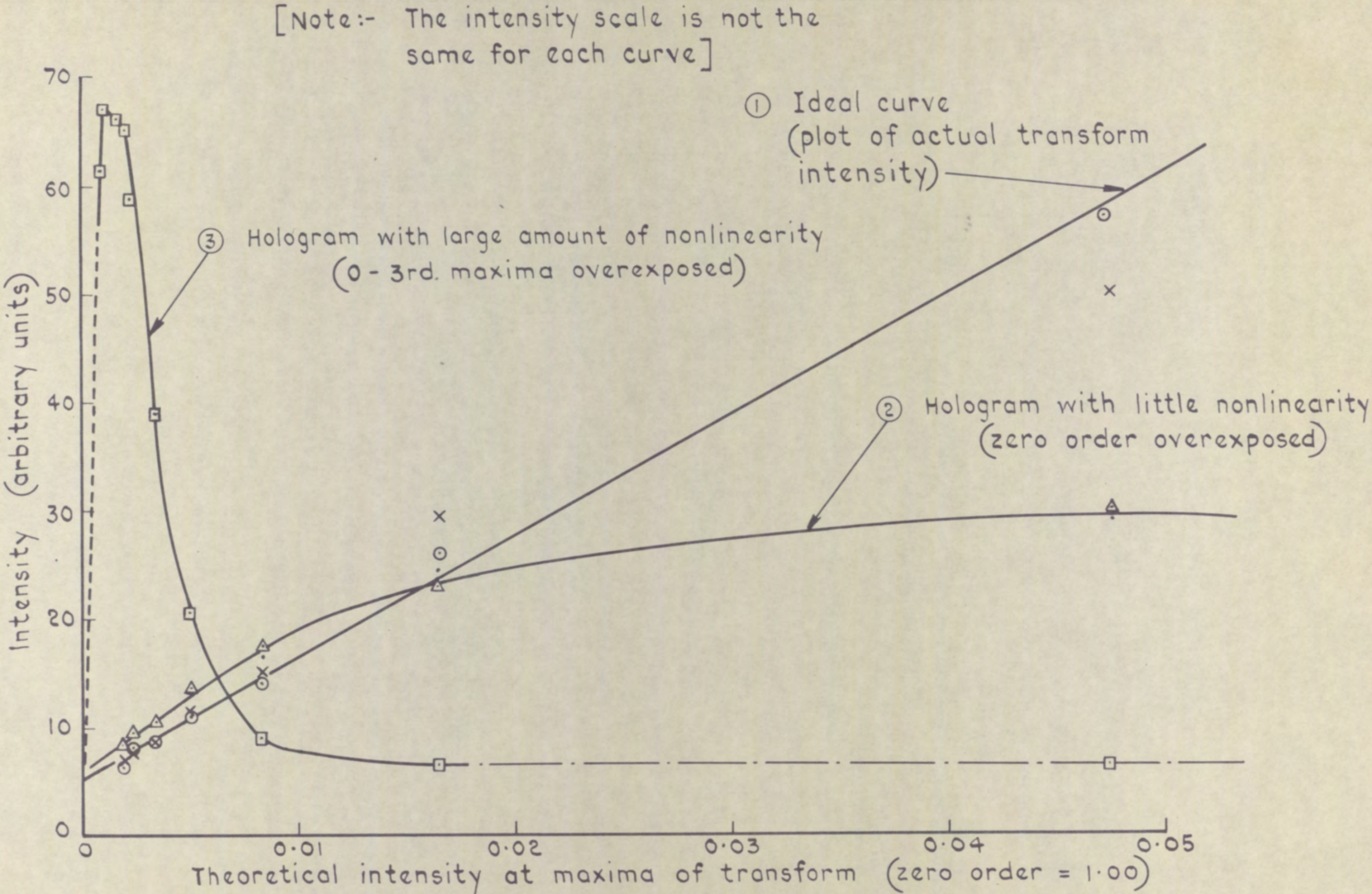


Fig.5.15 Nonlinear response of one dimensional transform holograms

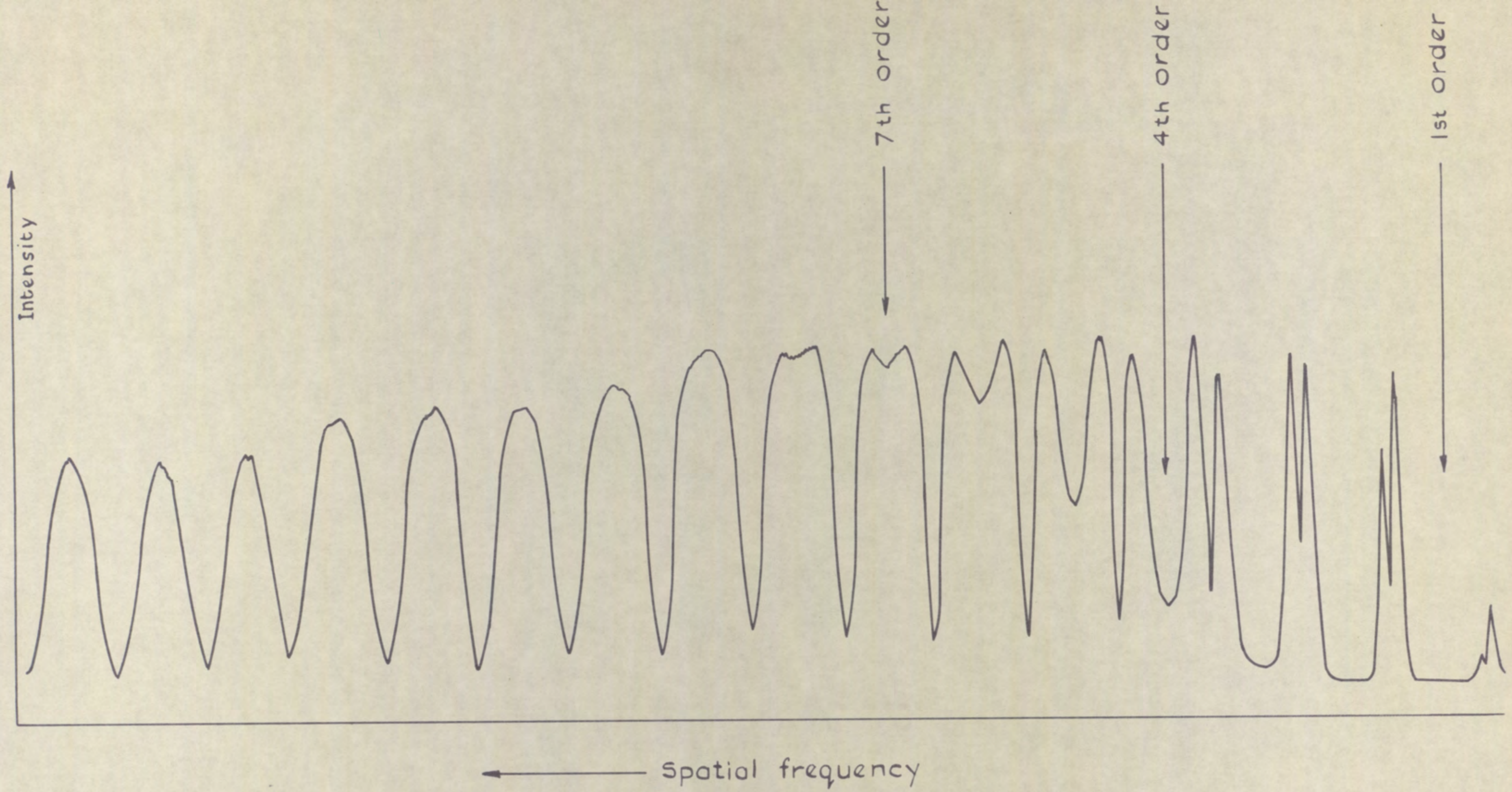
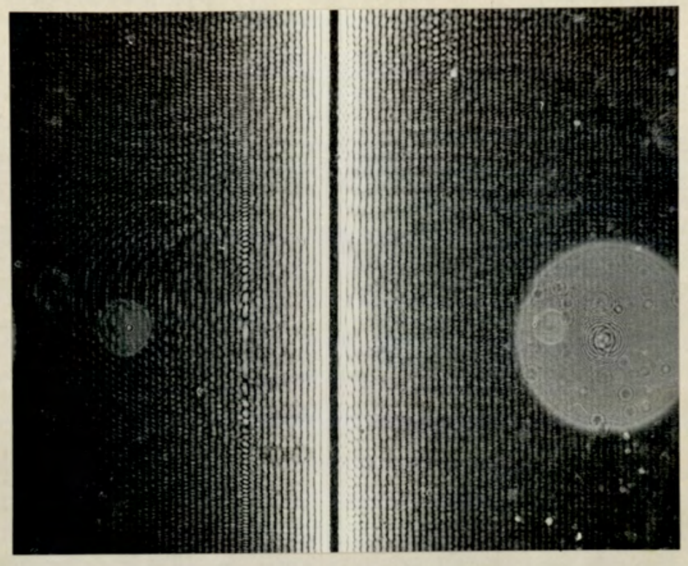


Fig. 5.16 Impulse response of slit hologram intensity distribution in reconstructed transform showing distorted low frequencies

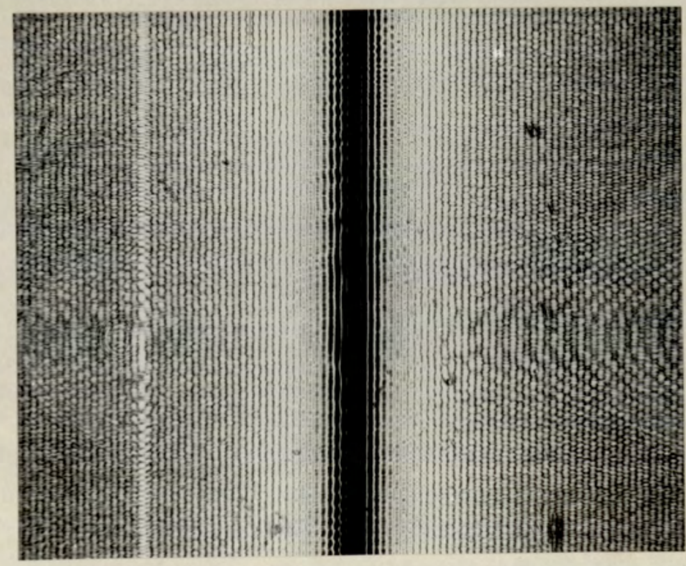
intensity were attributed to errors of form of the -ve cylindrical lens, and to errors in alignment.

The intensity of the reconstructed transform for a hologram made using the one dimensional arrangement is shown in Fig 5.14. The exposure of the transform was such that the zero order was overexposed and also part of the 1st and 2nd orders. This is seen in the intensity distribution; the 1st and 2nd maxima show dips where the maxima should occur, and there is zero intensity in the centre of the zeroth order. When these peak values are plotted on a graph versus the theoretical maxima intensities, the non-linearity becomes apparent, as shown in Fig 5.15, curve 2. If the hologram is made with a higher transform exposure, then more maxima become distorted, as shown in the recording of Fig 5.16. Here the 0th, 1st, 2nd and 3rd maxima are not recorded, whilst maxima out to 8th are distorted. The corresponding curve is shown on Fig 5.15 (curve 3), and the effect of non-linearity looks particularly severe. Ideally, the maxima of both curves should be equal, and the difference here is caused by differences in the photomultiplier voltage and by differences in laser beam intensity. Note that the hologram of curve 3 received a higher intensity object beam than that of curve 2, consequently the slope of curve 3 should be greater than that of curve 2, as shown in the extreme left of the diagram, where non-linear effects are not present.

An interesting observation which emerges from these curves follows the limiting intensity at which non-linear recording starts to occur. For curve 3 this is ≈0.001 on the relative intensity scale, whereas for curve 2 the limiting intensity is 0.008. The actual difference in object intensity of the two holograms was 1:60. This difference refers to the intensity difference, but the hologram reproduces the amplitude of the transform, which for these two cases is a difference of approximately 1:8. This is in agreement with the results shown on the graph, where the difference in limiting intensities is again 1:8. Note that this

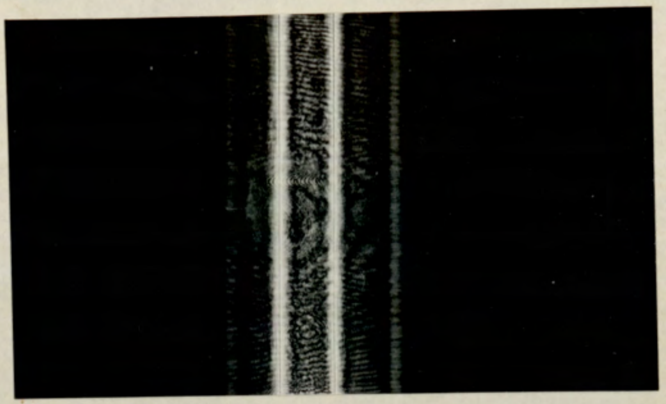


Low exposure

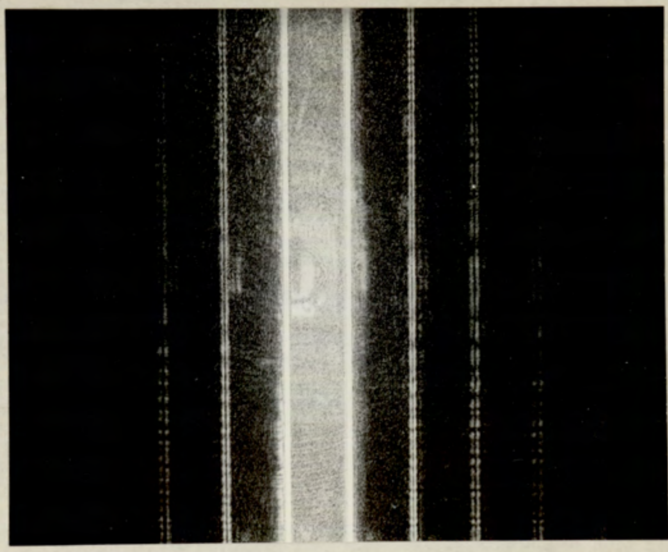


High exposure

Fig.5.17 Reconstructed transforms from one dimensional Fourier transform holograms



Low exposure



High exposure

Fig.5.18 Reconstructed image (impulse response) of one dimensional Fourier transform holograms observation extends to the point where non-linearity starts, at this point the effect of the shift of bias point on the t_A/Exp curve caused by the object beam intensity (or amplitude squared term) becomes pronounced, as shown in section 5.4.

Fig 5.17 is included to show the appearance of the two reconstructed transforms mentioned in the preceding paragraphs. The absence of the overexposed orders is indicated by the black strips down the centre of the transform, whilst higher frequencies are correctly recorded. The difference between exposures, both in amount of overexposure and in the extent to which higher spatial frequencies are recorded, can be seen.

5.5.3 Effect of non-linear recording on the autocorrelation and impulse response of one dimensional hologram

Having observed the immediate effect of non-linear recording on the fourier transform itself, the effects on the impulse and autocorrelation responses are now considered. Attention is again restricted to one dimensional transforms for the present. Most of the study was made by photography, although intensity profiles were made when necessary. Fig 5.18 shows the reconstructed images for the two holograms discussed earlier.

The main effect of overexposure in the zero and low spatial frequencies has been to remove the light distribution from all parts of the image except at the edges with which high spatial frequencies are associated. The centre of the image is not completely dark, however, and contains some fairly coarse fringes which are believed to be caused by phase variations across the aperture of the object slit, caused probably by the cylindrical lens. These phase variations are of low spatial frequency, and appear mainly on the response of the low exposure hologram. Their presence indicates how phase variations can be made visible, and is identical in principle to the method of phase contrast using central dark ground illumination.

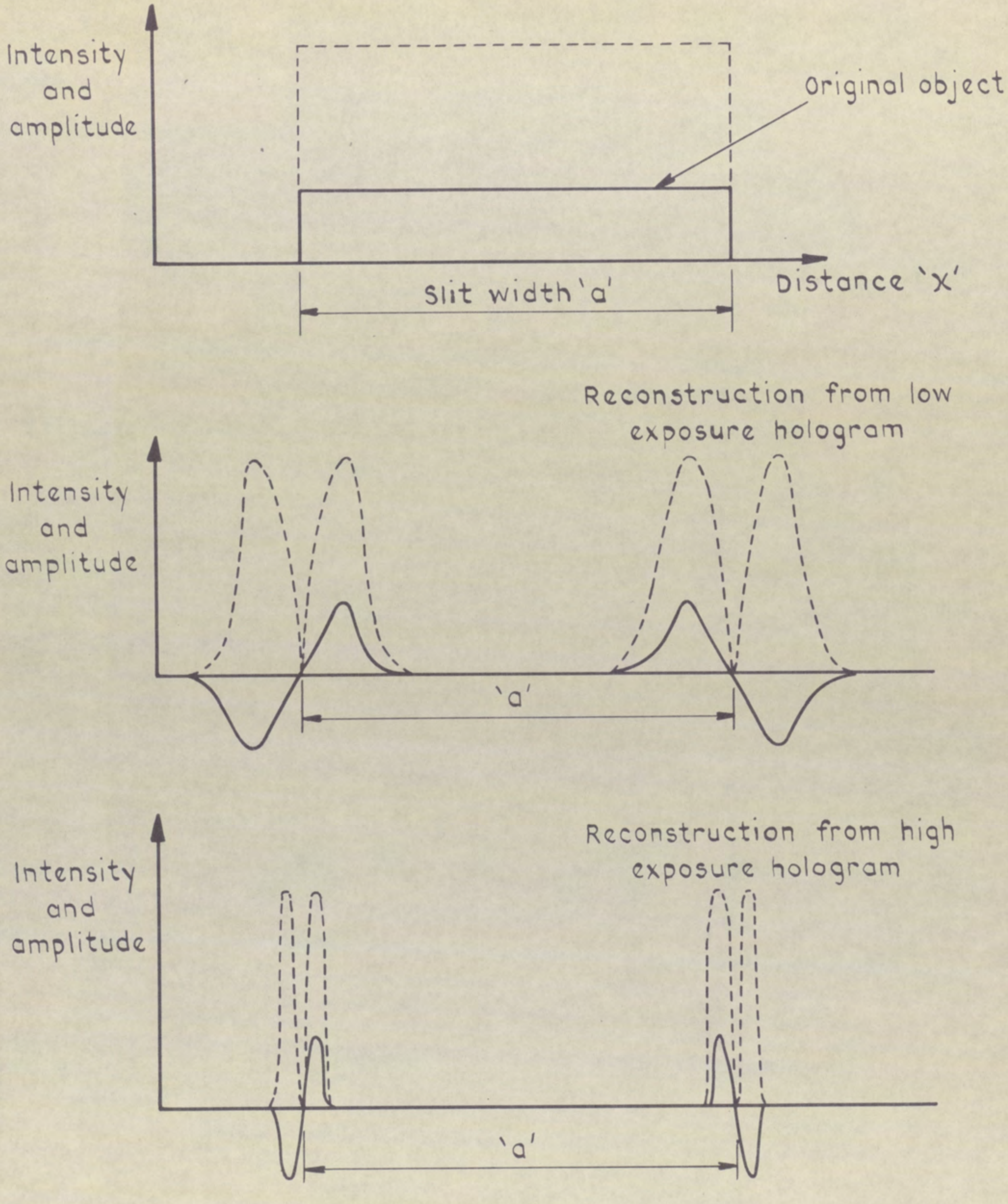


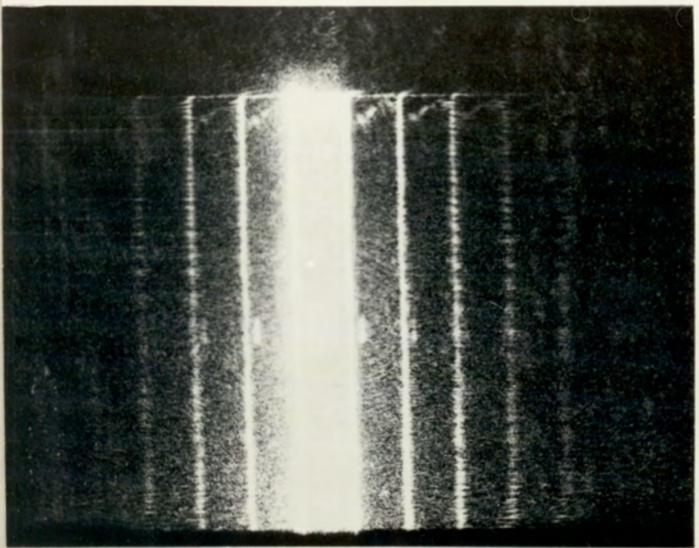
Fig. 5.19 Amplitude and intensity profiles for reconstructed images

The edges of the object slit are indicated by fine black lines, surrounded on each side by a bright strip of light. The appearance of the image is very similar to those reported by Birch[16], where the low spatial frequency was removed by means of an opaque stop.

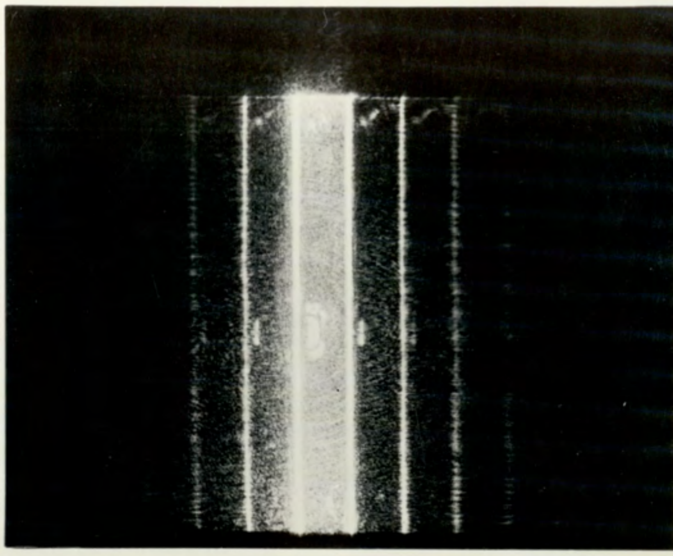
The appearance of theedge may be explained by reference to Fig 5.19. There is a very rapid change of amplitude at the edge of the reconstruction, passing from positive (on the inside of the aperture) to negative (on the outside). The zero amplitude corresponds to the aperture edge, and is responsible for the fine dark line seen in the reconstructions. The rate of change of amplitude is determined initially by the spatial frequency content of the object structure, and then by the number of spatial frequencies recorded by the hologram. In the low exposure case, relatively few frequencies were recorded but even so, the line is quite narrow. In the high exposure case, a large band of spatial frequencies are recorded, and both the line and the bright strips either side are narrow. This variation in the appearance of spatially filtered images was also reported by Birch.

Another consequence of non-linear recording is indicated by the appearance of false images to either side of the reconstruction. These have a similar appearance to the edge image just discussed. The presence of more false images for the higher exposure hologram suggests that non-linearity is responsible and that the effects are more severe for the high exposure case.

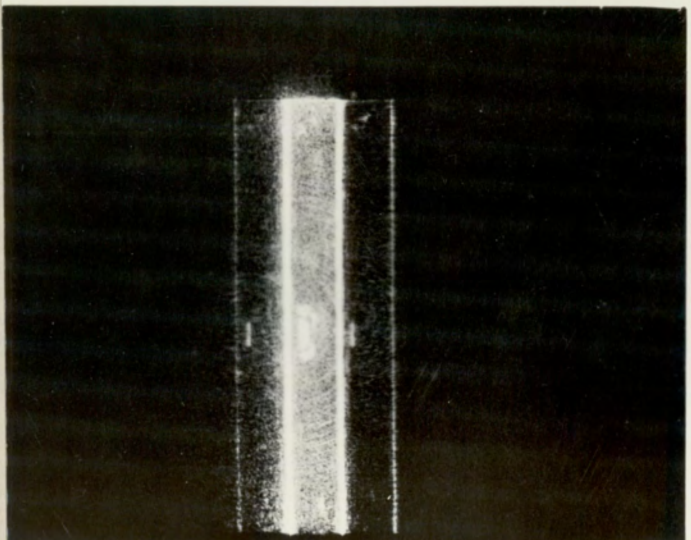
The appearance of these images is explained by referring to the reconstructed transform intensity distribution shown in Fig 5.16. The effect of non-linear recording has caused the higher intensities to be suppressed so that single side lobes become split into two parts. These secondary peaks can be regarded as forming an image in the same way that the individual lobes of the original transform contribute to the image



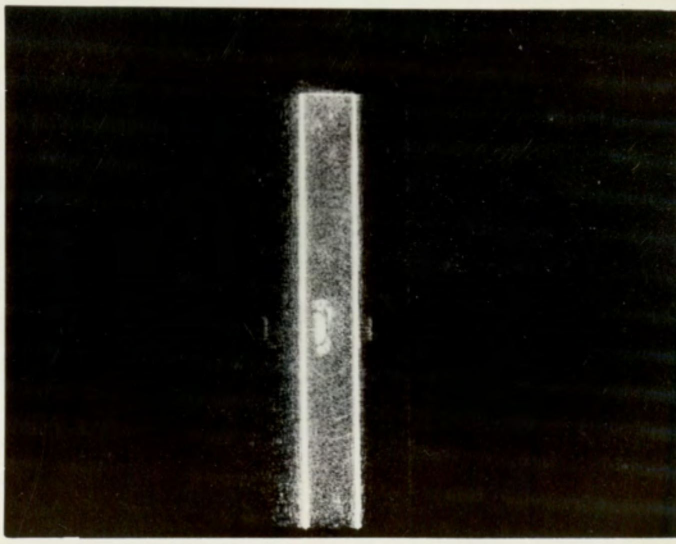
Half hologram obstructed



Half and 0 → 2nd orders obstructed



Half and 0 → 7th orders obstructed



Half and 0 → 12th orders obstructed

Impulse response of slit hologram, showing removal of non-linear effects

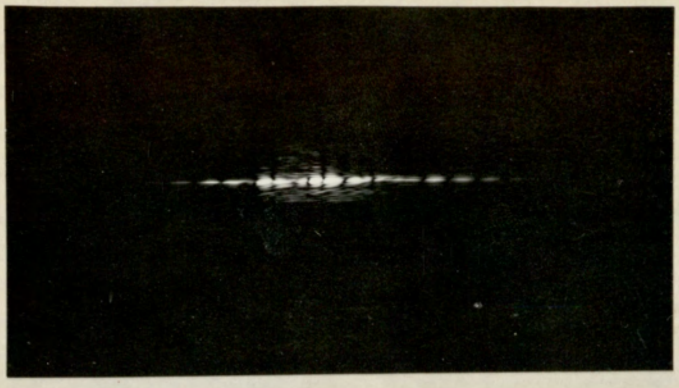
Fig. 5.20

of the object slit, as shown by the Abbe theory. Since the spacing of these secondary peaks is much smaller than the spacing of the original peaks, it follows that the non-linear images will be larger than the reconstructed image, and that several such images will be formed, depending on the number and spacing of the secondary peaks. In Fig 5.14, the reconstructed transform of the low exposure hologram is shown, with little if any, splitting of the side lobes. As a result, few false images are present in the reconstructed image.

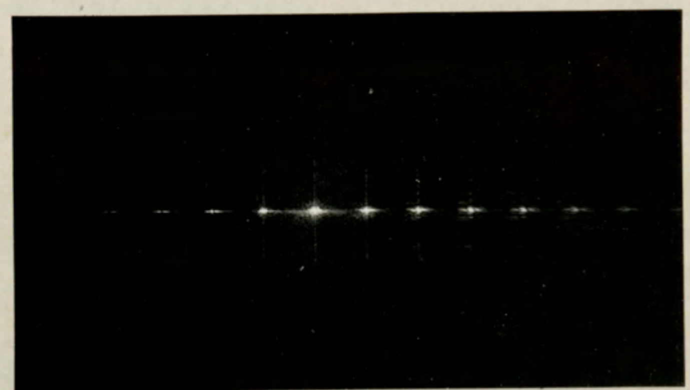
It follows from the previous argument, that the false images furthest removed from the true reconstruction are caused by secondary peaks which lie closest together, and these in turn are the distorted lobes in the low spatial frequencies. Thus if an opaque obstruction is placed over the lower spatial frequencies of the hologram, then the outermost false images will disappear, and that as the width of the obstruction is increased to obscure higher frequencies, then the false images will gradually vanish, working towards the reconstruction. This was demonstrated by experiment, and is shown in Fig 5.20, for the high exposure hologram. In the actual experiment, an opaque plate was used to obscure half the fourier transform hologram; and the illustrations were taken as this plate was moved steadily across the remaining half of the transform. The effect of using only half the transform is to remove the fine dark line marking the aperture edge, whilst the non-linearities remain unaffected. It can be seen that removal of successively higher orders of distorted maxima cause the non-linear false images to disappear.

In addition to the non-linear false images, there was also a considerable amount of scattered light, associated with the more intense parts of the image. This is caused by the graininess of the emulsion, each grain tending to scatter the light in a diffuse way. Apart from this, no other effects of non-linearity were observed.

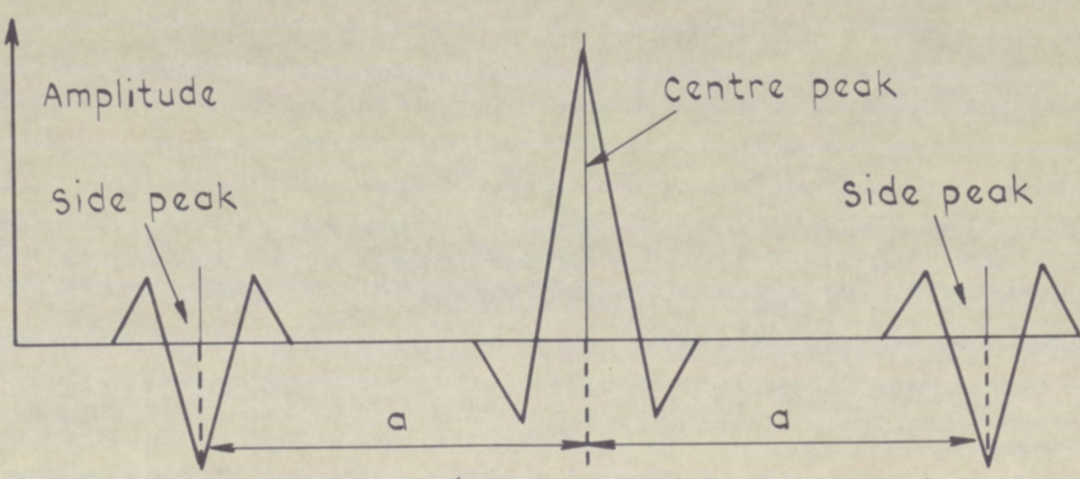
Low
exposure



High
exposure



Autocorrelation functions of one dimensional Fourier transform holograms



Autocorrelation function : linear recording but zero frequency removed

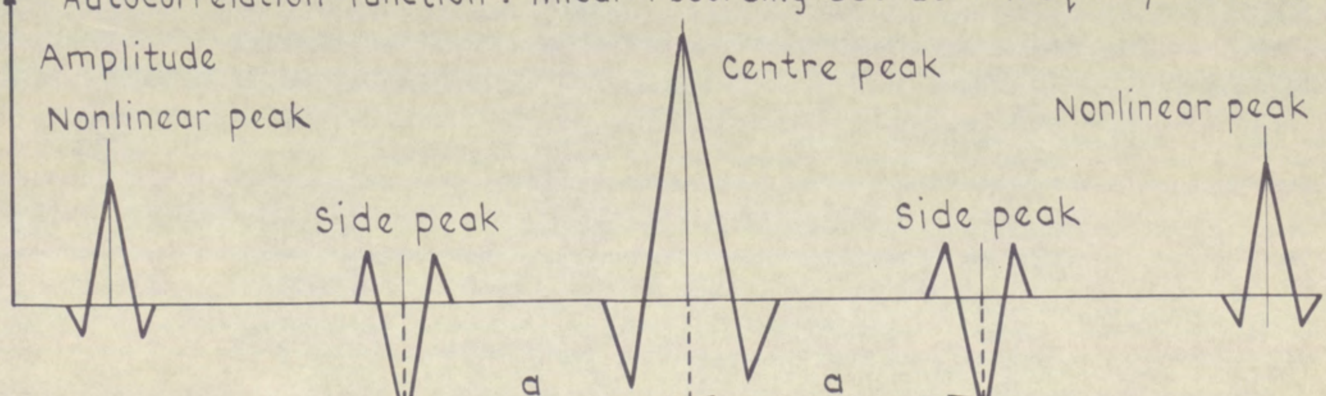


Fig. 5.21 Autocorrelation function : nonlinear recording

According to Jull[17] there are three types of effects from using photographic emulsions; (i) intermodulation noise arising from film non-linearity, (ii) film grain noise, and (iii) surface relief noise. The latter effect is not considered for the present, it is caused in the chemical processing of holograms and can be removed by refractive index matching. The two remaining effects of photographic recording have been observed; in particular the presence of intermodulation noise is indicated here by the false images seen in Fig 5.20.

The effect of non-linear response on the autocorrelation function of the object slit is now considered. The functions were observed for the two holograms discussed earlier using the arrangement shown in Chapter 4, and are shown in Fig 5.21.

There are several points of interest;

(i) Although the autocorrelation function of a slit (a rectangular aperture) is strictly pyramid shaped, the recording of the fourier transform in one direction means that the autocorrelation also appears only in one direction. Since the overall aperture of the hologram in the orthogonal direction is relatively large, the width of the function in this direction will be very small, thus it appears as a line of intensity. This follows from fourier transform theory, since the autocorrelation is the transform of the hologram plane.

(ii) The absence of the zero and low spatial frequencies in both holograms means that triangular shape of the autocorrelation function (for one dimension) is lost. Instead, there are three distinct peaks of intensity. The formation of these peaks can be understood by referring to Figs 5.18 and 5.19. The reconstruction of the aperture contains two parts, corresponding to the edges of the original object slit. When the hologram is correlated with the object, the two edges "match", producing the central peak of intensity in the autocorrelation function. The action of the optical system allows the hologram to be scanned over the object slit, so that the right hand edge of the object is correlated with the left hand edge of the hologram image, and vice versa. Thus two additional peaks of intensity are formed, one each side the central peak. The spacing of these peaks is equal to the width of the object slit. Since opposite edges are being correlated, the correlation function will be negative in amplitude. This is shown in the lower half of Fig 5.21.

The individual peaks in the autocorrelation also have a fine structure associated with them. This takes the form of two small peaks one each side the central peak, but with opposite sign of amplitude. They are caused by the fine structure at the edge of the reconstructed image, ie by the reversal of amplitude which also causes a zero of intensity to appear. All of these peaks appear in the autocorrelation as a row of points, the amplitude differences not being evident.

(iii) The splitting of high intensity lobes in the transform by non-linear recording caused additional peaks of intensity to appear in the autocorrelation function, as shown in Fig 5.21. These peaks appear outside the side peaks previously mentioned, and their number appears to depend on the severity of non-linear recording. Thus more non-linear peaks are seen for the high exposure hologram than for the low exposure hologram. Their spacing was observed to be constant and equal to the slit width. The intensity of the peaks diminishes steadily progressing away from the centre of the function, in addition the structure associated with each peak seems to become less fine.

The appearance of these additional peaks (apart from those caused by the absence of the zero frequency term) can be explained by referring again to the reconstructed images of Figs 5.18 and 5.19. One can consider the non-linear reconstruction to be scanned over the object in much the same way as the true reconstruction is scanned, thus producing peaks of intensity whenever the non-linear edge image and an edge of the object

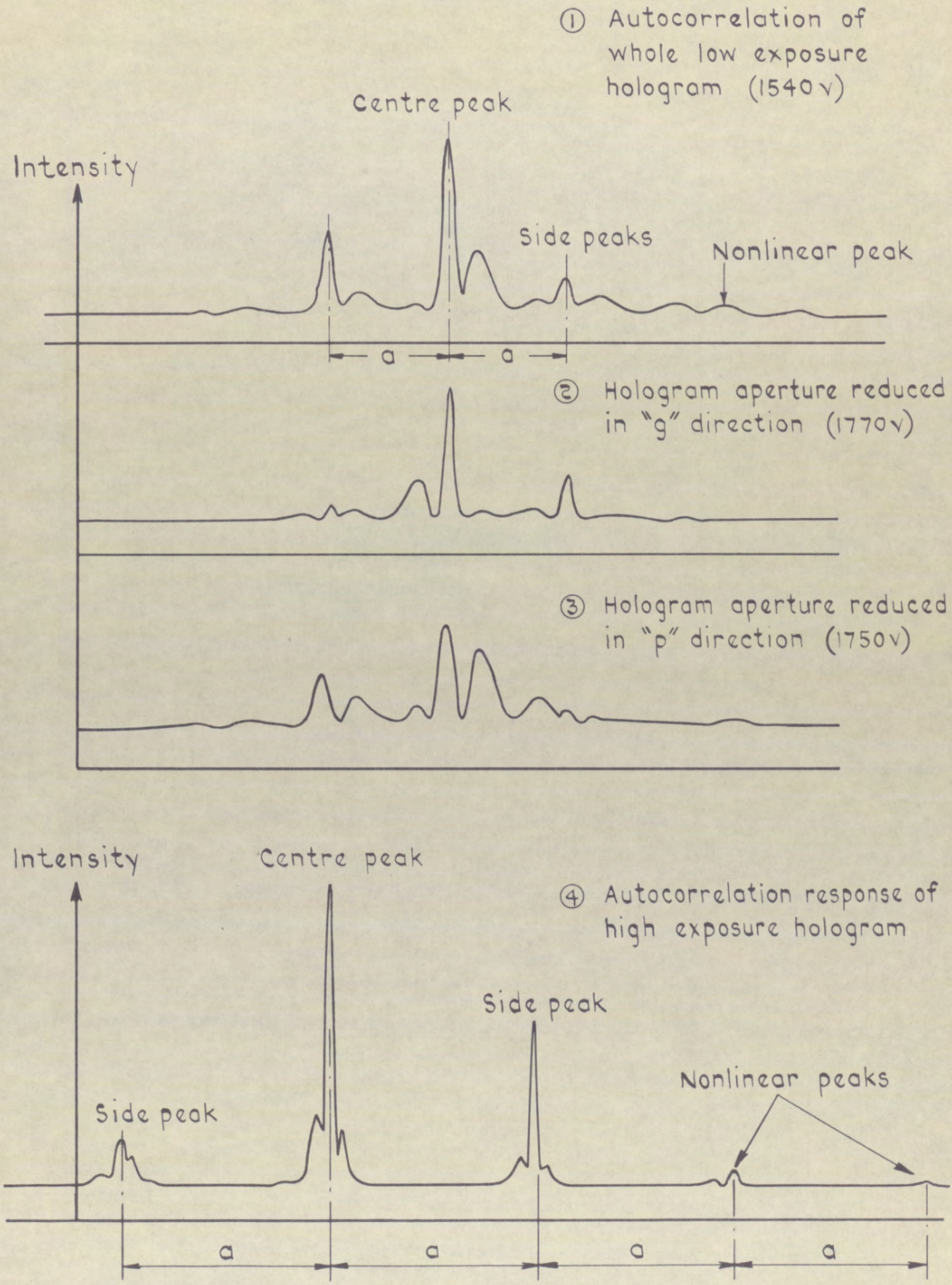


Fig. 5.22 Intensity profile of autocorrelation response

are matched. Since the non-linear peaks are approximately equally spaced, two images will be in alignment with the object at the same time, producing two autocorrelation peaks in the same position but of different sign. The sign of the resulting peak will depend on which of the two peaks is the stronger. This explanation is only satisfactory in the sense that a physical explanation is given; it would be better to have a mathematical description of the process. The description could assume that the recorded transform, instead of being $F(p)$ say, was represented by $F(p) + \delta[F(p)]$ implying that the function $F(p)$ is present but that it is modified by some other function, δ, which varies as the amplitude of $F(p)$. The result of autocorrelation would then be written as:-

$$F(p) * \{F(p) + \delta[F(p)]\}^*$$

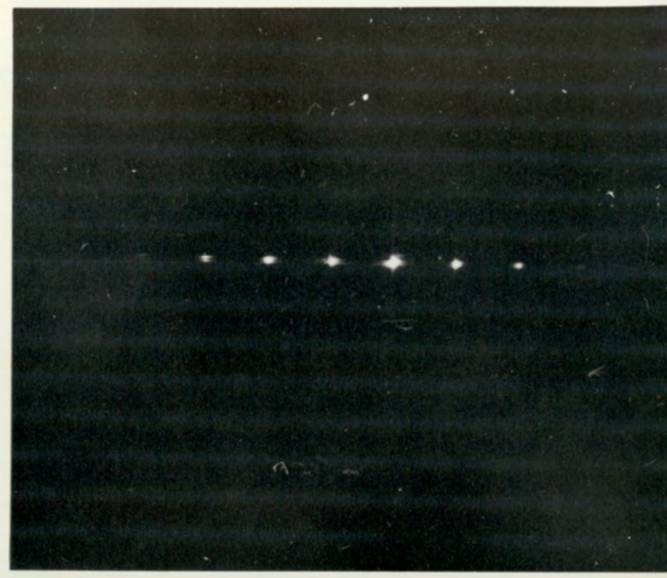
giving:-

$$F(p) * F(p)^* + F(p) * \delta[F(p)]^*$$

Thus the expected autocorrelation appears, plus additional terms depending on the form of $\delta[F(p)]$.

The distribution of intensity in the autocorrelation functions is shown in Fig 5.22. The first three curves show the autocorrelation intensity for the low exposure hologram, whilst curve 4 is of the high exposure hologram. The reduction in width of the peaks in curve 4 has been caused by the recording of high spatial frequencies, which represent a very rapid change of amplitude in the object and hence produce sharp autocorrelation peaks. The non-linear peaks in both holograms can be seen, although they are considerably weaker than the central peak. Curves 2 and 3 show the effect of reducing the hologram aperture in first the "q" and then the "p" directions; these directions referring to the coordinates of the fourier transform. In curve 2, there is no change in shape of the autocorrelation since there is no change in the fourier transform in this direction. The overall intensity is reduced, and a higher photomultiplier voltage is required to retain the recording signal. In curve 3, the effect of reducing the hologram aperture in the

Half hologram obstructed



Half and 0 → 2nd orders obstructed

Half and 0 → 7th orders obstructed

Half and 0 → 12th orders obstructed

Autocorrelation response of slit hologram showing removal of nonlinear effects

Fig. 5.23

"p" direction is marked since the transform contains amplitude variations in this direction. The relative intensities of the peaks have altered, in particular the peaks having negative amplitude being increased considerably. A logical explanation of this effect is that the removal of contributions from the higher spatial frequencies means that the lower spatial frequencies contribute a higher proportion of the autocorrelation energy. The image corresponding to a fourier transform with only the lower spatial frequencies present is more complicated than those shown earlier (Figs 5.18, 5.19). In the case of only the two first side lobes being present one finds that the image has a grating like appearance with several maxima and zeros. The result of correlating such an image with the original object would produce several peaks where intensities were more equal, as shown in curve 3 (Fig 5.22). The variation in structure of a slit image with the amount of fourier transform transmitted is shown in Appendix B.

In correspondence with the observation made for the impulse response, one would expect the non-linear peaks in the autocorrelation function to disappear as the distorted parts of the transform were covered. This was observed, using the same method as reported earlier, and the results are shown in Fig 5.23. As the distorted orders are covered, so the non-linear terms vanish, and eventually one is left with the three peaks corresponding to a linear recording, but with the zero spatial frequencies removed. Note the absence of fine detail in the autocorrelation peaks; this was caused by half the hologram being covered, thus eliminating the fine structure in the same way as for the impulse response.

5.5.4 Effect of non-linear recording for two dimensional transform holograms

In the previous section the effects of non-linear recording were particularly vivid because of the use of a one dimensional transform. When conventional two dimensional transforms are considered, the

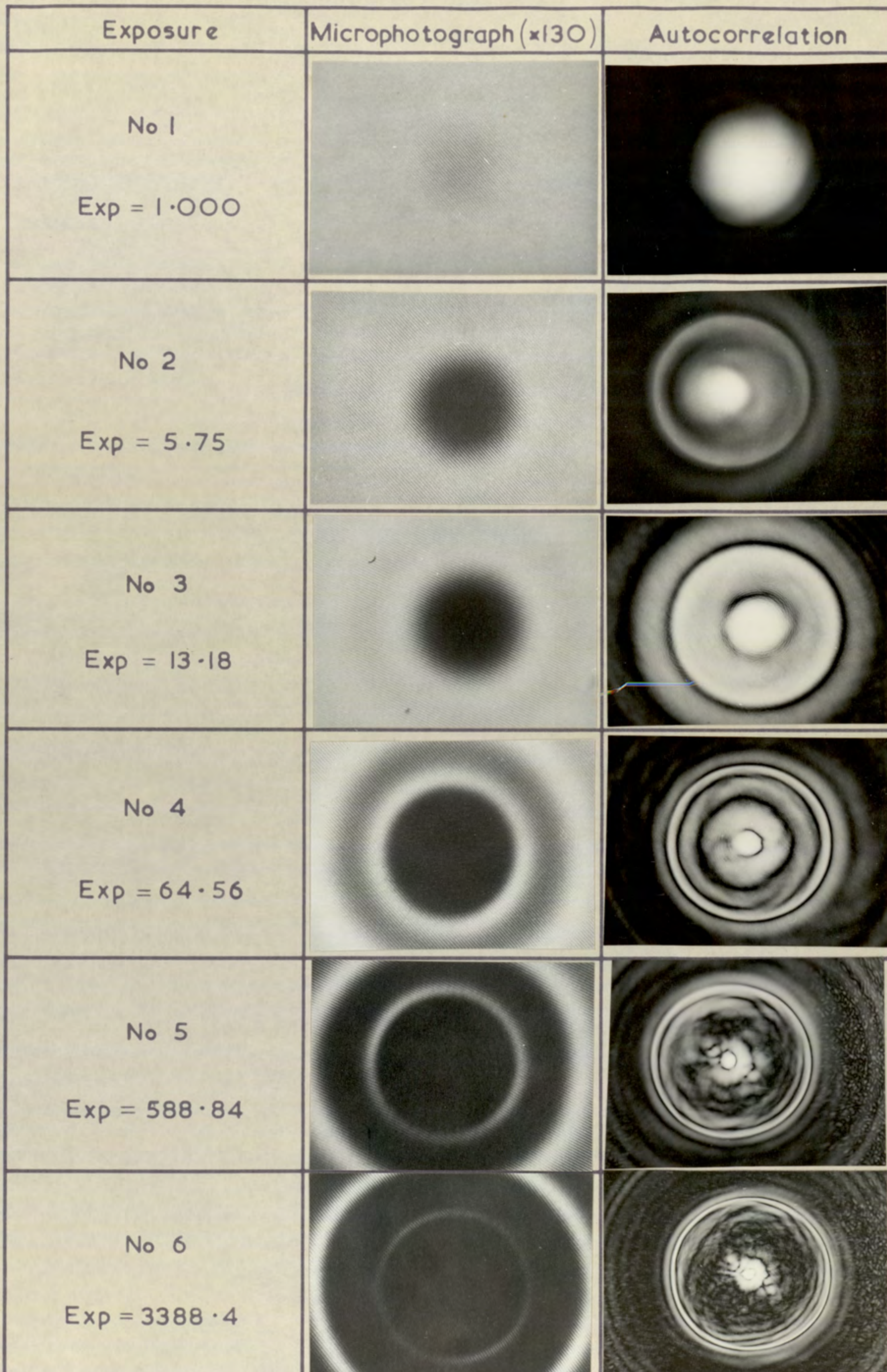


Fig. 5.24 Microstructure and autocorrelation response of circular aperture hologram showing effect of exposure

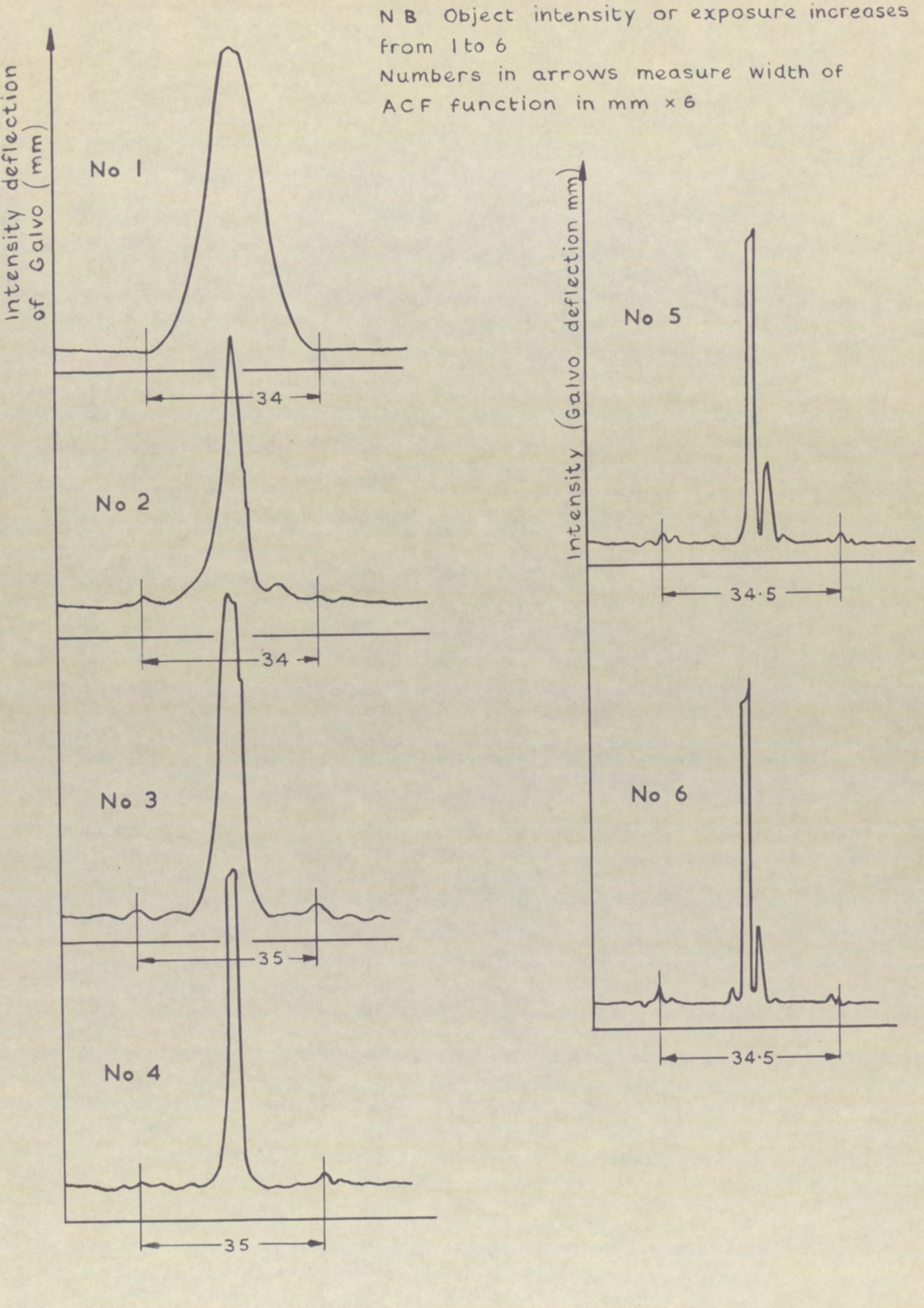


Fig. 5. 25 Intensity distribution in ACF of different holograms

non-linearities are not so apparent. Nevertheless, they may be observed, particularly for simple object shapes, such as a rectangular or circular aperture; in this section we consider the response of the latter.

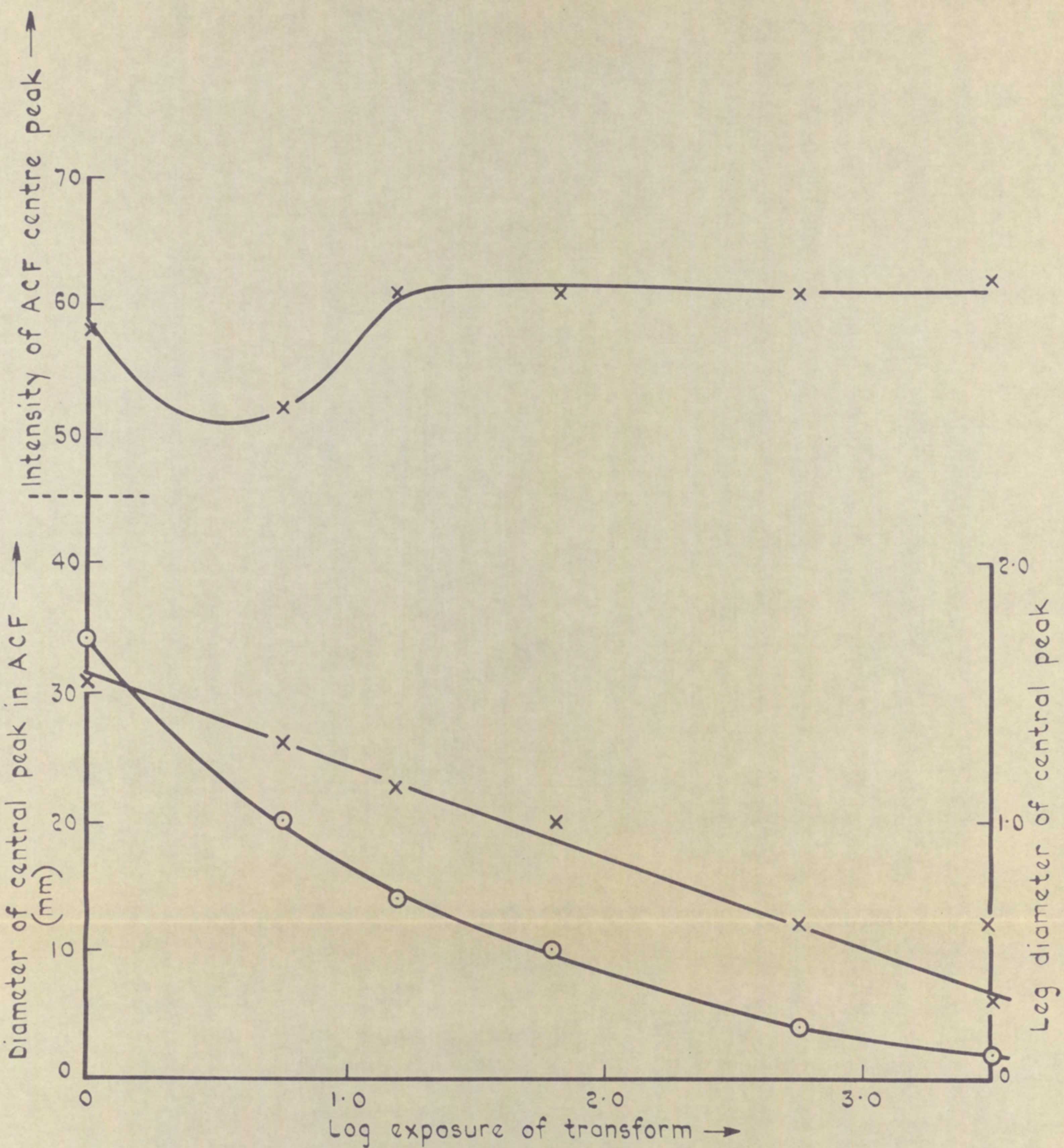
Several holograms were made of a circular aperture object measuring 6mm diameter. The exposure of the fourier transform was varied over a range 1 to 3400, the reference beam meanwhile being kept constant. The autocorrelation response and microstructure of each hologram were photographed and are shown in Fig 5.24.

The impulse responses are not shown because they were too weak to be recorded.

In general, the change in appearance of the autocorrelation function with exposure is in agreement with the results obtained earlier.

In the case of No 1, where all the centre of the transform has been recorded, the autocorrelation function has the expected circular shape, and is twice the diameter of the original aperture. As the exposure is increased, so the autocorrelation function shrinks; leaving a ring marking the edge of the original function. The cross section intensity profile of the autocorrelation function of any of the responses 2-6 would resemble, in general, the intensity profiles obtained earlier. For the high exposures, the autocorrelation function has a very fine central peak, surrounded by a ring of much lower intensity, and about 1/5th the radius of the expected autocorrelation function. Beyond the set of rings marking the edge of the expected function lie additional rings which are caused by non-linear recording, and may be explained in a similar way to the effects observed with one dimensional transforms.

The variation in intensity of the autocorrelation function was investigated using the photomultiplier scanning unit and the ultra-violet recorder. The results are shown in Fig 5.25. There is a similarity between these profiles and those obtained for the one dimensional transform. In addition to the variation in width of the base of the



Intensity of ACF centre peak
70
60
50
40
30
20
10
0
Diameter of central peak in ACF (mm)
2·0
1·0
0
Log diameter of central peak
0
1·0
2·0
3·0
Log exposure of transform

Fig. 5.26

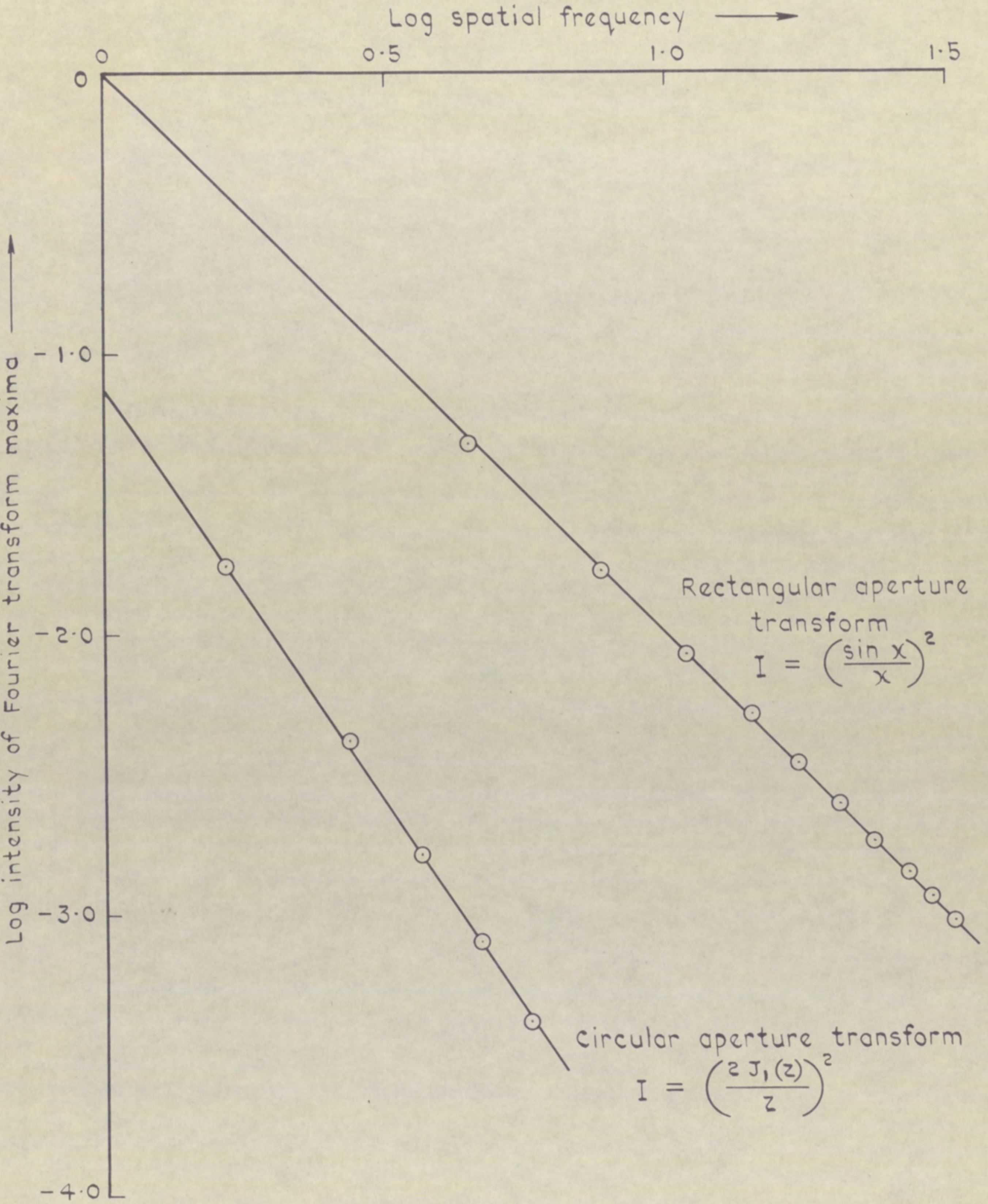


Fig. 5.27 Variation of intensity of maxima with spatial frequency on log – log scales

central peak there is a variation in intensity; as the exposure is increased the intensity falls because the centre of the transform is attenuated by overexposure. There is a minimum in intensity variation after which the intensity rises, because the portion which has been darkened remains almost constant whilst the area of the hologram increases rapidly. According to the results, the rise in intensity ceases and no further variation seems to occur; in fact one would expect the intensity to slowly decrease as the hologram overexposure increases.

The variation in intensity and in base width of the autocorrelation peak are shown in Fig 5.26 plotted against the log exposure. The variation in base width, when plotted on a log/log scale, appears to be a straight line. This relationship was confirmed by the following argument.

Consider the overall variation of intensity in the fourier transform, neglecting the zeros. This variation can be seen by plotting the intensities of successive maxima versus spatial frequency, and is shown in Fig 5.27 for a log/log scale. The relationship for this scale is linear and indicates that

$$\log I = - K_1 \log (p)$$

where I is intensity, k is a constant and p is spatial frequency at the peak intensities.

This may be written in an alternative form:-

$$K_1 \log p = - A + \log E$$

where E is the exposure needed to give saturation of the emulsion on the photographic plate for different values of p. [It may be represented simply by drawing the curve of 5.26 with a positive slope]. There is an inverse relationship between the spatial frequency p, and the diameter of the autocorrelation function 'd', because of the transformation between these two planes. Thus we can write:-

$$\log p = - k_2 \log d$$

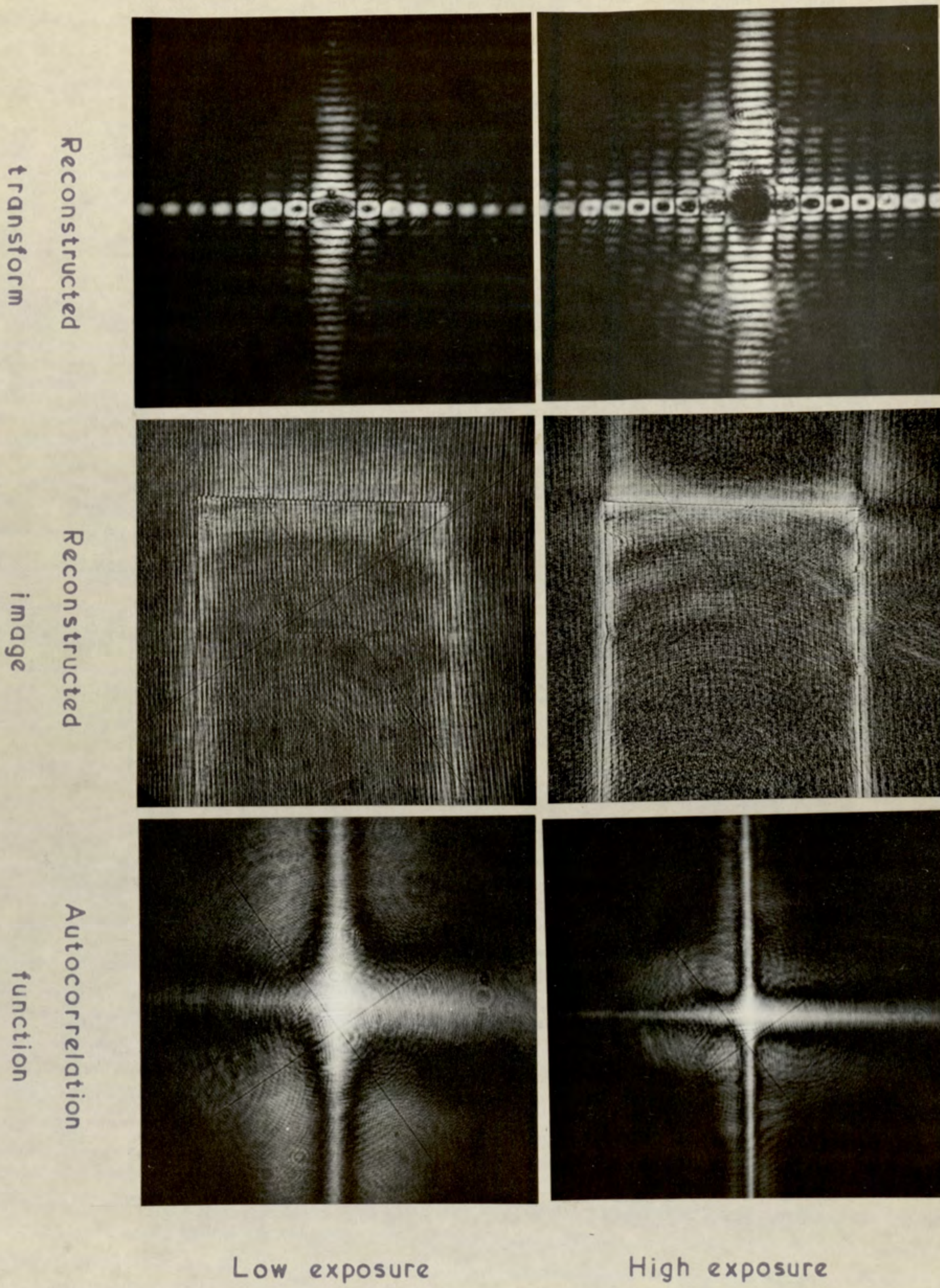


Fig. 5.28 Response of rectangular aperture hologram

Substituting in the equation for p and E,

$$k_1 \; [- k_2 \log d] = - A + \log E$$

or

$$- k_3 \log d = - A + \log E$$

This is a straight line of negative slope k_3 and intercept A as shown experimentally in Fig 5.26. This result is important, because it indicates that the diameter of the central peak of the autocorrelation function is related to the lowest spatial frequency recorded in the hologram. The linear logarithmic relationship with exposure is only as accurate as the logarithmic relation between the fourier transform intensity maxima and spatial frequency. This seems to be closely by linear for the rectangular aperture, but for the circular aperture the approximation is less accurate.

These observations were confirmed by results obtained with rectangular aperture holograms. In these experiments the circular aperture was replaced with a rectangular aperture measuring 10mm by 4mm, and its fourier transform formed by the spherical lens used for the circular aperture experiment. Again, two different transform exposures were given, differing by 1:4.5. The autocorrelation, impulse response and reconstructed transform for both exposures are shown in Fig 5.28.

In the reconstructed transforms the area of overexposure can be seen as dark patches within the transform, the difference in size of the patches being caused by difference in exposure. The impulse response of the high exposure hologram contains the black line profile of the aperture, but this is not so apparent in the low exposure hologram. In addition, there is a broadening of the bright band surrounding the edge profile. More striking is the appearance of fine vertical fringes running across both holograms, but considerably stronger in the case of the low exposure hologram. The origin of these fringes is not known for certain, but it is thought that they were caused by an additional background wave due to secondary reflections at a lens surface, since the fringes are curved. The difference in brilliance is attributed to the difference in density of the hologram. The fringe pattern does serve to indicate the difference in phase between light inside and outside the profile of the aperture, and confirms the statements made earlier in this chaper, referring to the direction of amplitude of light in the reconstruction. The presence of the black line in the image enables small imperfections in the aperture to be clearly seen, the original aperture image being shown for comparison.

The two dimensional autocorrelation function of a rectangular aperture would normally have a rectangular base, four times the area of the aperture. The intensity distribution would rise from zero at the edges of the rectangle to a maximum in the centre; along the axes of the rectangle the distribution would be linear, and a triangular shaped intensity function would be obtained.

The autocorrelation thus appears as a cross with one arm thicker in section than the other. The effect of increase in hologram exposure is to reduce the thickness of both arms in proportion making the whole function finer. In between the arms of the cross are faint light patches which are caused by non-linear recording. These patches correspond to the correlation of the broad light bands either side the reconstruction profile, with the original aperture. As in the case with the impulse response, the presence of interference fringes indicates the change in phase between the faint patches of light and the main part of the autocorrelation. [The fringes here are thought to be caused by dust particles.]

No false images were observed in these last experiments, either in the reconstruction or autocorrelation function. The reason for this was that a) the amount of overexposure was relatively low, compared with other holograms, and b) the size of the hologram was small, so that the area of overexposed hologram contributing to any non-linear images would be very small. In later experiments, non-linear false images from rectangular aperture holograms were detected.

5.6 Phase Variations in Holograms

5.6.1 The action of the photographic process produces variations in emulsion thickness as well as density when the emulsion has been exposed and processed. The thickness variation occurs because of the removal of unexposed silver salt from the emulsion during the process, leaving an emulsion whose thickness varies with emulsion density. The variation in thickness is likely to be severe where large changes of intensity are expected, as in fourier transform holography.

The variation in emulsion thickness will cause phase modulation of the transmitted light, and this may distort the output of the system; it could also be argued that the phase modulation would sometimes reinforce the diffraction occurring at density variations, and that a stronger output would be obtained. Holograms using only the phase variation in bleached photographic emulsions[18], dichromated gelatin[19] and thermographic emulsion[20] have been reported. Here, we are concerned with the effect of phase variations in the photographic emulsion produced as a consequence of normal photographic processing, rather than by bleaching, as in ref 18.

The effect of phase variations has been studied by Leith[21], who examined the effect of phase variations on the spectra of a sinusoidal target, using a liquid of matching refracture index to remove thickness variations. In general, it was found that phase variations shifted the energy content in the spectrum from low to high spatial frequencies. The higher order terms were regarded as harmonic distortion, which disappeared when the phase variations were removed by the matching liquid. The effect was found to be more pronounced in coarse grained than in fine grained emulsions. It was also observed that the intensities of some of the lower spatial frequencies were considerably reduced when index matching was used; showing that phase modulation was the dominant factor.

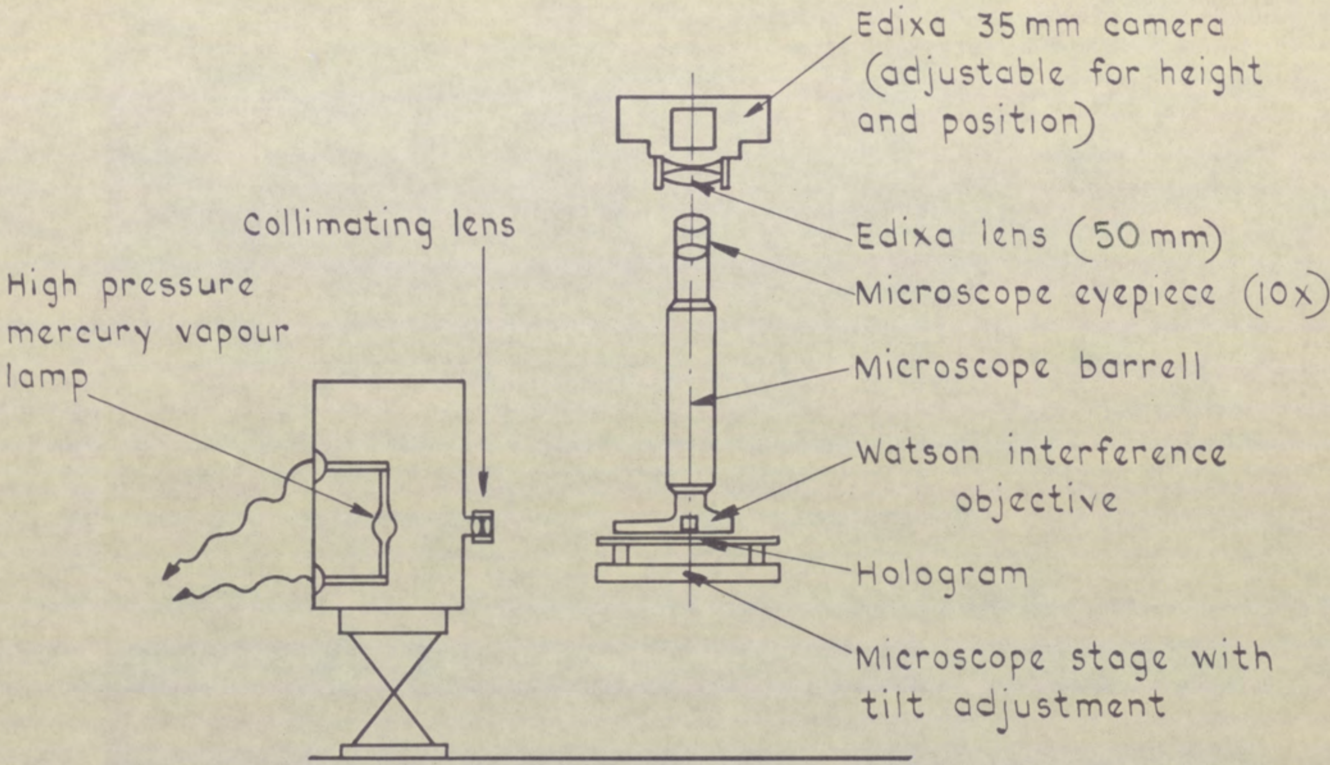
In a comprehensive report, Smith[22] examined the effect of different parameters, such as spatial frequency, intensity variation, emulsion thickness, development process and emulsion type on the phase variations or relief image. His main conclusion was that the height of the relief image depended primarily on the degree of tanning of the emulsion, rather than on the removal of unwanted silver halide. The tanning action is complex, but was briefly described by Smith as the result of forming insoluble and mechanically strong molecular bonds in the emulsion, giving rise to differntial swelling and shrinkage, and causing the tanned areas of the image to stand in relief above the unexposed areas of the emulsion. It was found that the height of relief image depended not only on the quantity of silver reduced, but also on the spatial frequency content of the image; there being a maximum relief image at one spatial frequency for any one image contrast. This spatial frequency appeared to vary inversely according to emulsion thickness; in particular, for 649F emulsion, the spatial frequency was about 10 lines/mm. Further work in this field was also reported by Hannes[23] who worked with Agfa-Gevaert emulsions, and where results agree substantially with those of Smith.

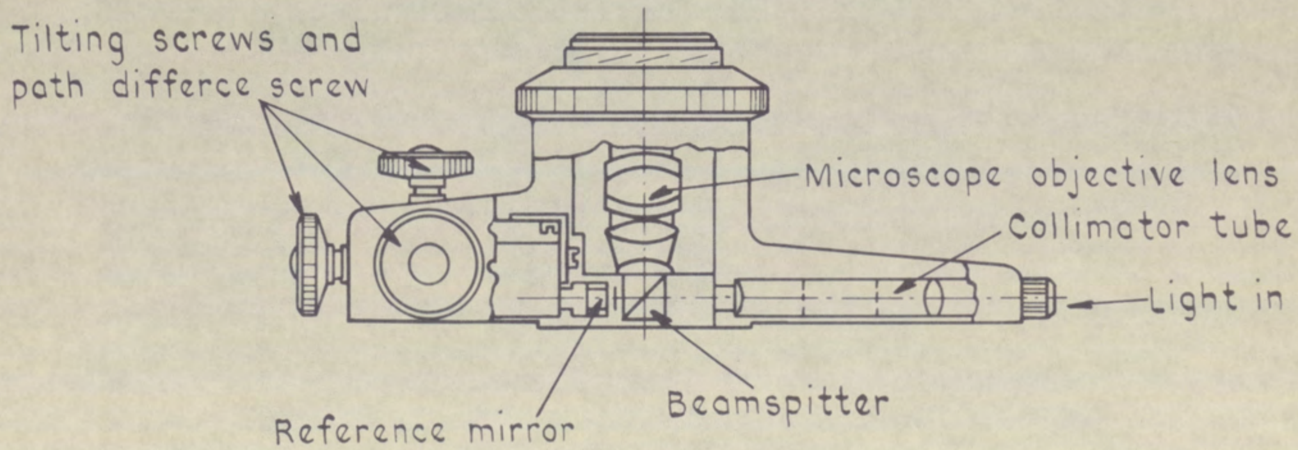
The effect of film substrate thickness variations as distinct from emulsion thickness variations was studied by Ingalls[24], who discussed the use of a liquid gate of correct refractive index in removing these variations.

It appeared that the exact effect of hologram phase variations upon the autocorrelation and impulse response had not been studied experimentally, particularly for fourier transform holograms, and therefore some work into this topic was undertaken.

5.6.2 Micro Interferograms of Fourier Transform Holograms

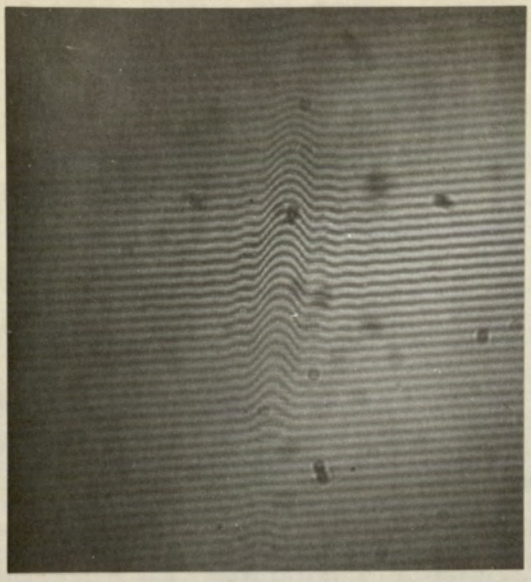
It is of interest to consider first the actual thickness variations in a fourier transform hologram, after processing. These variations were



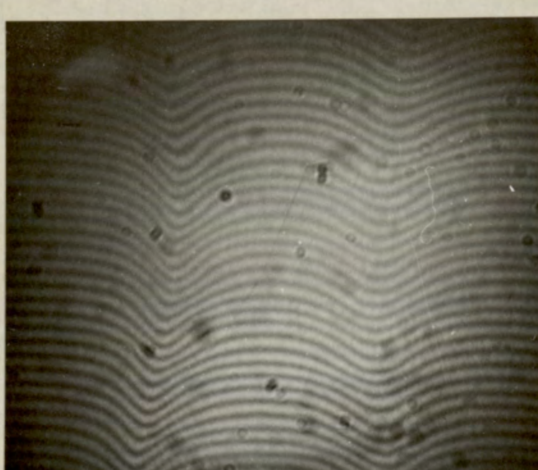


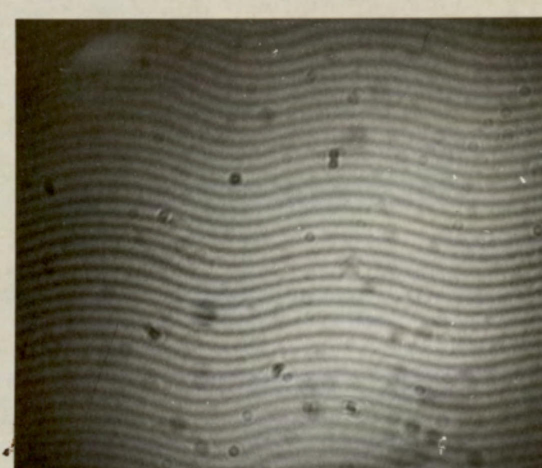
Watson 16mm interference objective

Fig. 5.29 Arrangement for interference microscopy

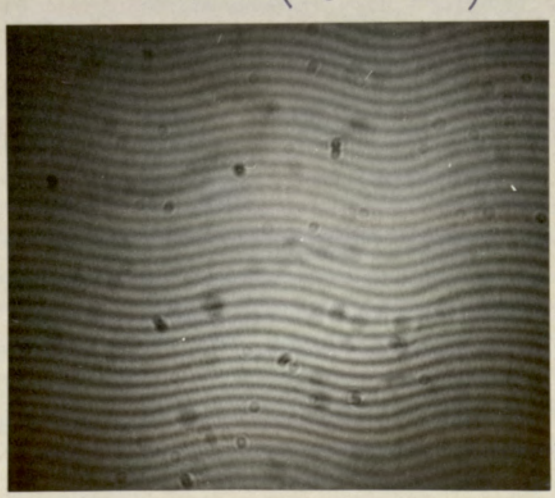


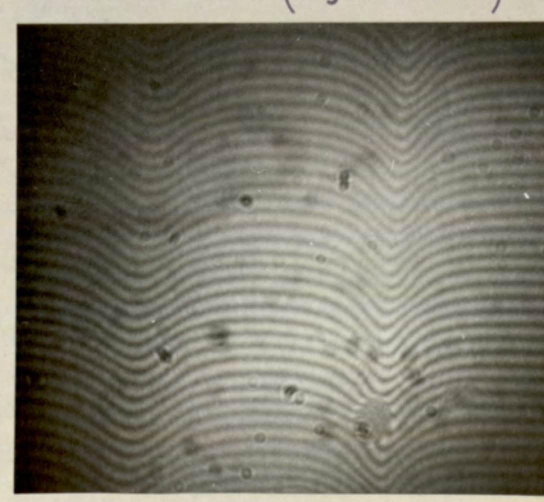
a Microinterferogram of rectangular aperture hologram





b Ist side lobe (right hand) e 4th side lobe (right hand)





c 4th side lobe (left hand) d Ist side lobe (left hand)

b → e Interferograms of one dimensional transform hologram

Fig. 5.30 Microinterferograms of holograms all about 200 x magnification studied by examining the holograms using a Watson Interference Microscope Objective of 16mm focal length, giving an overall magnification of x100 when used with a x10 eyepiece. Micro-interferograms were recorded using a 35mm camera over the microscope eyepiece, as shown in Fig 5.29.

These interferograms are shown in Fig 5.30 for a hologram of a rectangular aperture$^{(a)}$ and for the one dimensional hologram of a slit$^{(b-e)}$. The rectangular aperture hologram shows a strong relief image at the zero order, of height $2\frac{1}{2}$ fringes or $\frac{5}{4}\lambda$, or, since mercury light was used, $\approx 0.8\mu$. The relief images caused by the 1st order maxima can also be seen, although these are much smaller. Relief images caused by the carrier fringes of the hologram were not seen, although these could be resolved by the microscope system. The reason they are not seen is because the relief image would only be strong when there is a large contrast in the carrier fringes, and this will occur whenever the fourier transform intensity equals the reference beam intensity, as for example, on the slopes of the side lobes of the transform. Another reason is that the carrier fringe spatial frequency is of the order of 200 lines/mm, and the spatial frequency at which the surface relief effect is strongest is 10 lines/mm, according to Smith, so that relief images would not be pronounced at the carrier fringe frequency.

The interferograms of the one dimensional hologram show the variation in surface relief for the lowest orders, and higher orders, and between both halves of the transform. There appears to be little difference between the two halves of the transform, confirming that equal intensities of light produce an equal amount of thickness variation. The interferogram of the lowest orders show the flat topped relief of the zero order, with a rapid thickness variation to the 1st zero or minimum. For higher orders, the thickness variations become less rapid, and in the 4th to 5th orders, there is a smooth, almost sinusoidal thickness variation, with a maximum relief of only $\frac{\lambda}{2}$ compared with $\frac{3\lambda}{2}$ for the zero order.

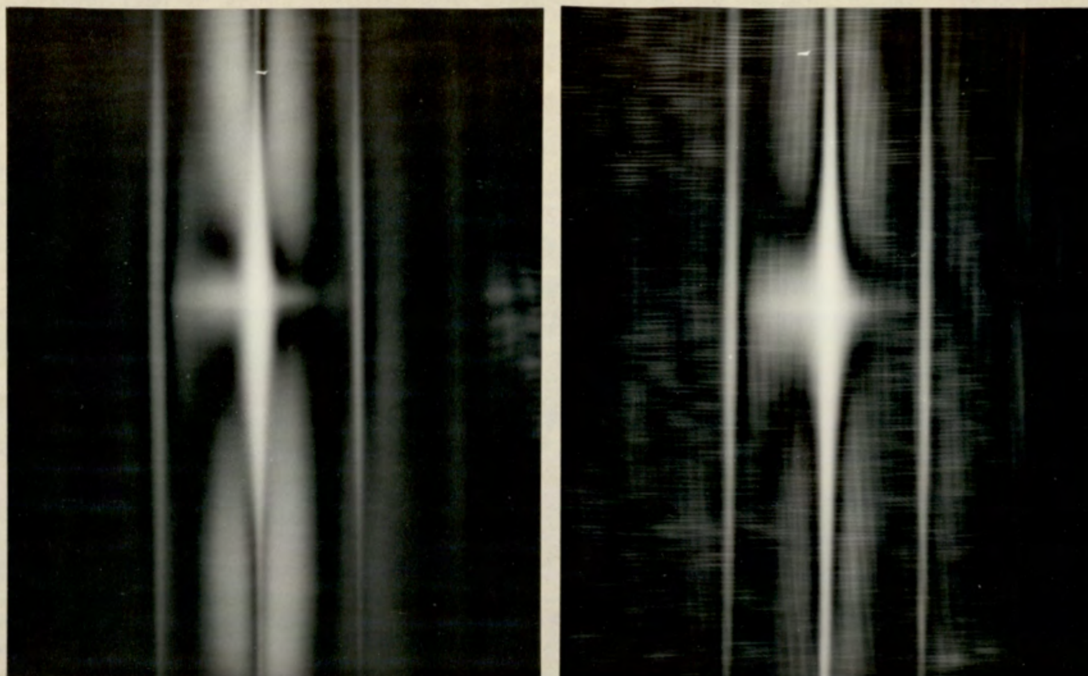
These interferograms confirm that thickness variations exist, and that they are most pronounced in the region of maximum intensity. The highest relief was found to be $\sim\frac{3\lambda}{2}$, or 3 fringe widths, and this was the same for both holograms, as expected, since the maximum density will be constant for a given emulsion.

5.6.3 Effect of thickness variations upon the autocorrelation and impulse responses

The effect of thickness variations on the hologram response was studied by immersing the hologram in a liquid gate containing liquid of matching refractive index. The construction of the liquid gate was described in Chapter 4, section 4.2.6. One of the effects of using a liquid gate is that the position of the focal plane will be moved slightly, and that there will be a small scale change in the size of the fourier transform. The focal plane shift can be corrected by focussing, but the scale change implies that the hologram must be immersed in the liquid during exposure. It was thought that this should be avoided if possible, and the magnitude of the scale change was calculated to examine its significance.

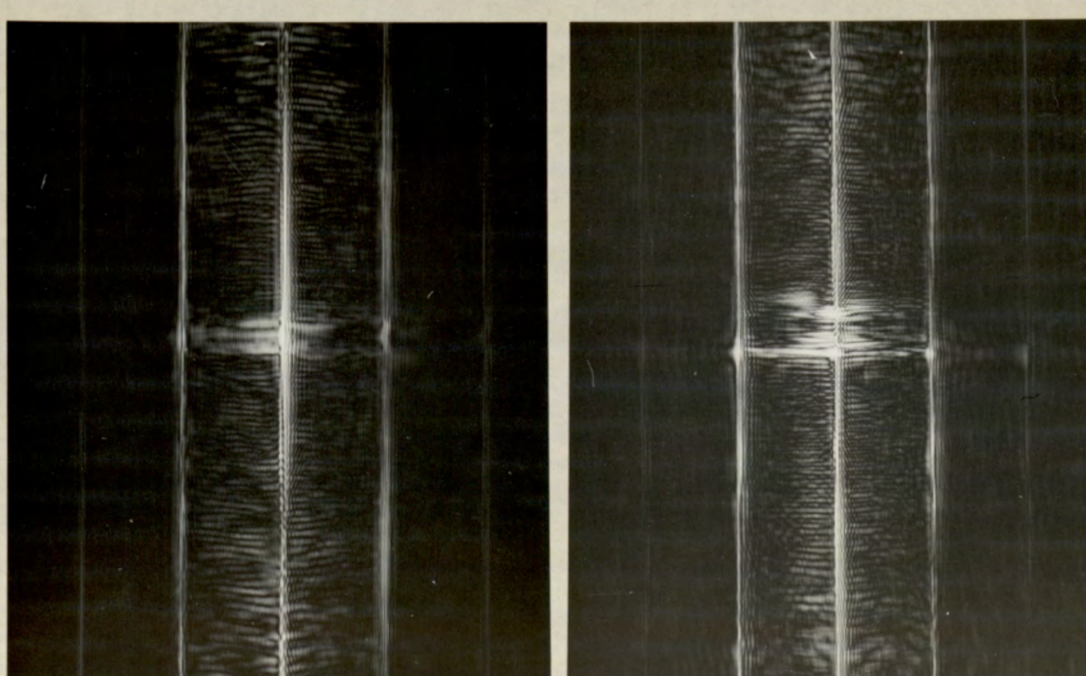
The change in relative position of two points 1cm apart in the transform, one point being on axis, was calculated for an object 10cm wide and for a lens of focal length 100cm. It was assumed that the cone of rays intercepted the liquid gate 2cm before its focus. The relative position of the two points was than found to be 0.99999cm, ie a change of 1pt in 100,000. Since the change in size would be so small, it was considered justifiable to expose the hologram in air, and then immerse it in liquid after processing.

The hologram used for this experiment was that shown in the previous section, of a rectangular aperture, in which only the zero order was completely overexposed. This hologram gave a strong autocorrelation and



In air　　　　In liquid gate

Low exposure hologram



In air　　　　In liquid gate

High exposure hologram

Fig. 5.31 Effect of refractive index matching on autocorrelation response of rectangular aperture hologram

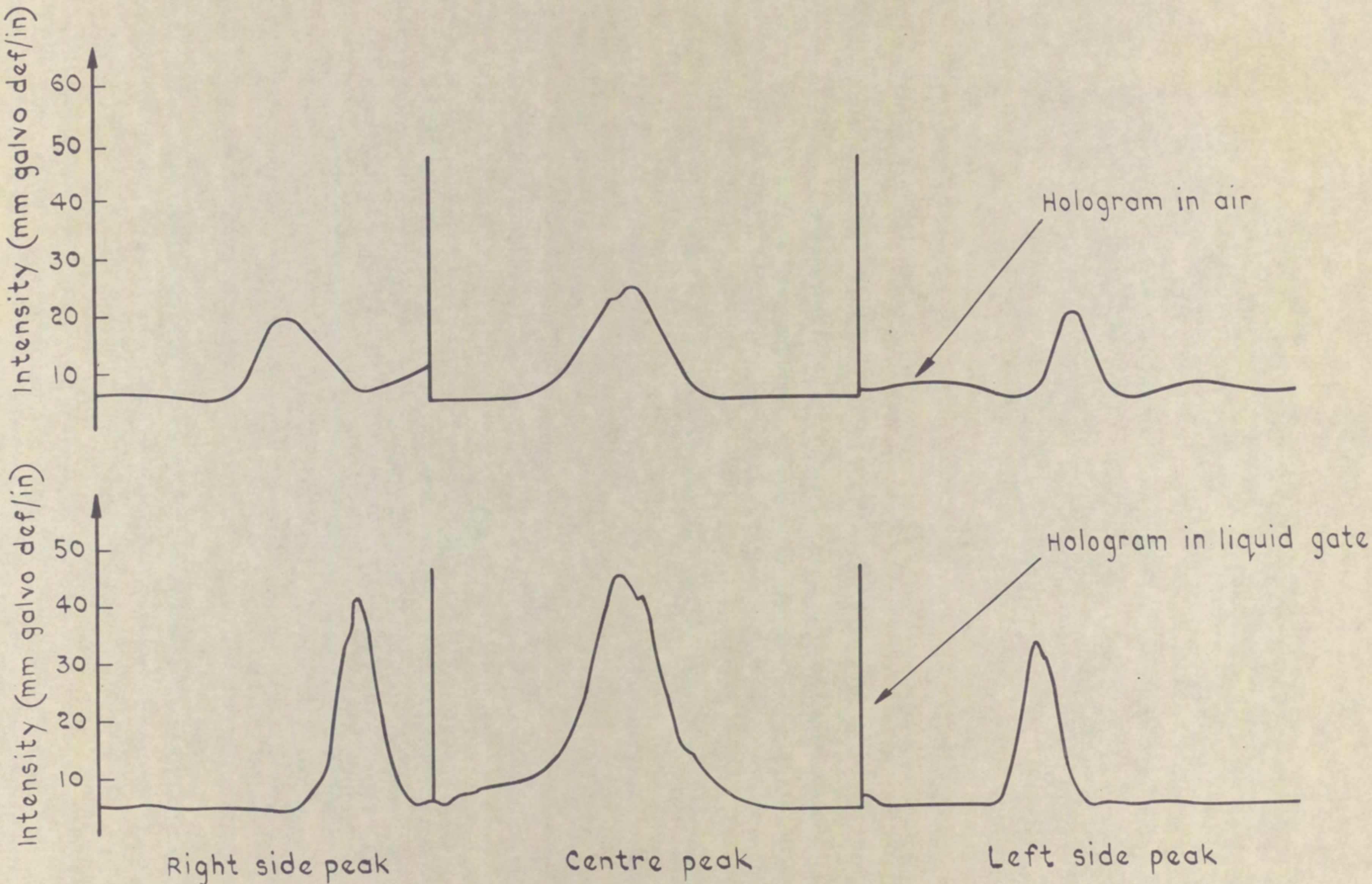


Fig. 5.32 Effect of liquid gate on autocorrelation intensity profile
(intensities of centre peaks multiplied by 0·091)

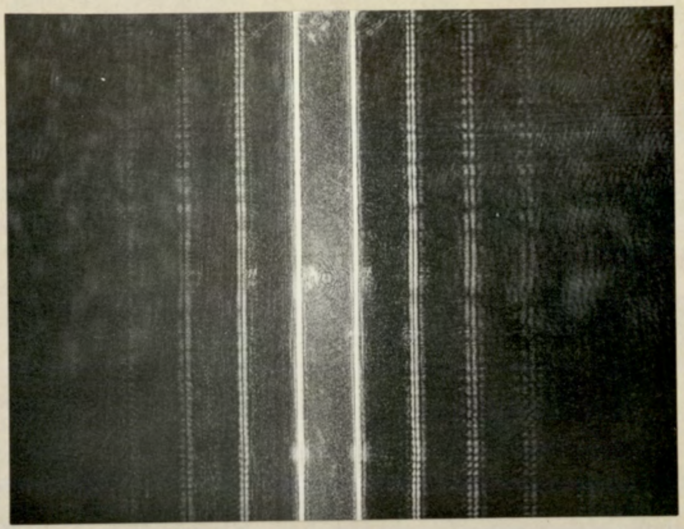
was therefore suitable for photometric measurements. In addition another hologram of the same aperture but with higher exposure was examined. The autocorrelation responses of these holograms were photographed with holograms in air and immersed in the liquid gate. The results are shown in Fig 5.31.

The photographic evidence does not show any major change in the appearance of the autocorrelation function. For the high exposure hologram, there appears to be a slight reduction in the non-linear terms, and the intensity at the centre of the autocorrelation function seems to be reduced. In the low exposure hologram the function appears to be sharper, with both the centre and first peaks of the function appearing brighter. The diffuse patches of light either side the central term appear to be less pronounced when the liquid gate is used.

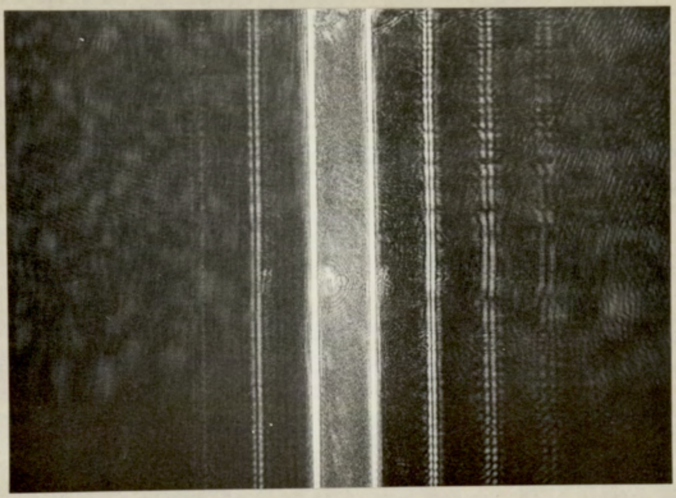
The photographic and visual observations of autocorrelation response did not give any significant results from using a liquid gate. The intensity distribution in the autocorrelation function was therefore measured, for the low exposure hologram, with and without index matching liquid. Examples of the photomultiplier recordings obtained are shown in Fig 5.32, where the increase in intensity caused by index matching is apparent for all three peaks of the autocorrelation function. This result was confirmed by taking several recordings, which are summarised in the following table.

<u>Effect of Index Matching on Autocorrelation function intensities</u>

Hologram Condition	Intensity of Autocorrelation peak (mm galvo deflection)			Photo multiplier Voltage
	Right Peak	Centre	Left Peak	
In Air	-	15	-	600v
"	13	214	15	800v
"	9	151	9	800v
In Liquid	-	36	-	600v
"	-	38	-	600v
"	37	460	29	800v
"	30	169	14	800v
"	42	561	25	800v



In air



In liquid gate

Fig. 5.33 Effect of refractive index matching on impulse response of one dimensional transform hologram

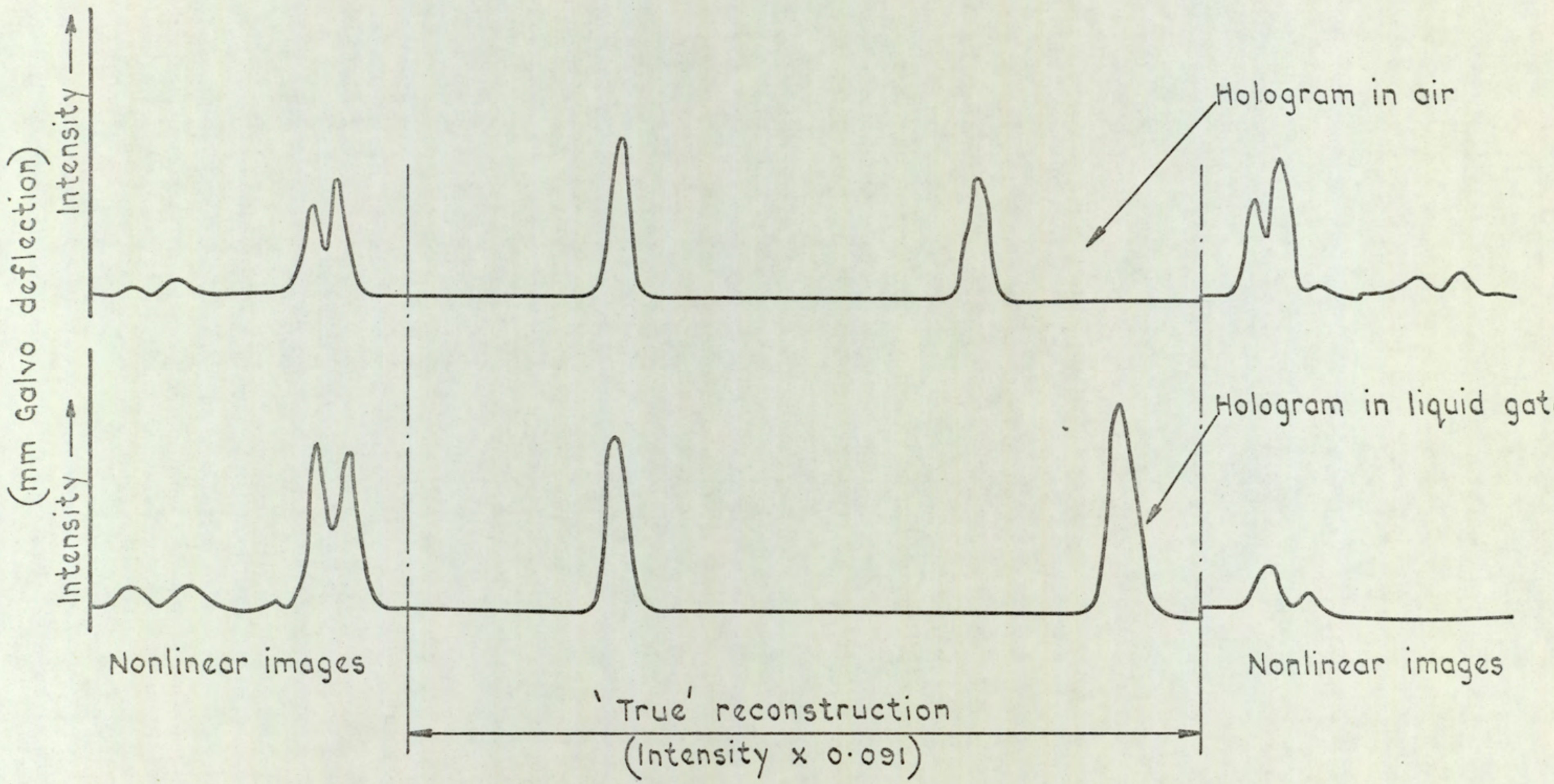


Fig. 5.34 Impulse response of one dimensional hologram in air and in liquid gate

In general, the effect of the liquid gate is to increase the intensities of both the central and side peaks by a factor of between 2 and 3. Variations in the results shown are fairly typical of intensity measurements on the autocorrelation function, and are thought to have been caused by air turbulence and vibrations. The laser beam intensity was fairly constant in this test, varying by no more than 3%.

These observations appear to be in agreement with those of Leith, since the effect of index matching is to concentrate the signal energy into the central and side peaks of the correlation function, these corresponding to the lower spatial frequencies of Leiths' sinusoidal target.

The effect of index matching on the non-linear terms of the autocorrelation was not observed because these terms were very weak in intensity. An alternative to this is to observe the impulse response of a one dimensional hologram, where the non-linear terms are very pronounced. Fig 5.33 shows the impulse response of a one dimensional hologram of a slit, with the hologram in air and in liquid. Although the non-linear terms are still present when the hologram is index matched, these terms appear to have been reduced in intensity, particularly those on the left hand side.

As with the autocorrelation function, the intensity distribution of the impulse response was measured, and examples are given in Fig 5.34. The index matching does not appear to have modified the centre part of the response, and only one of the non-linear terms has been reduced; this corresponds with the photographs shown in Fig 5.33. This measurement was repeated several times, and the results are summarised in the following table:-

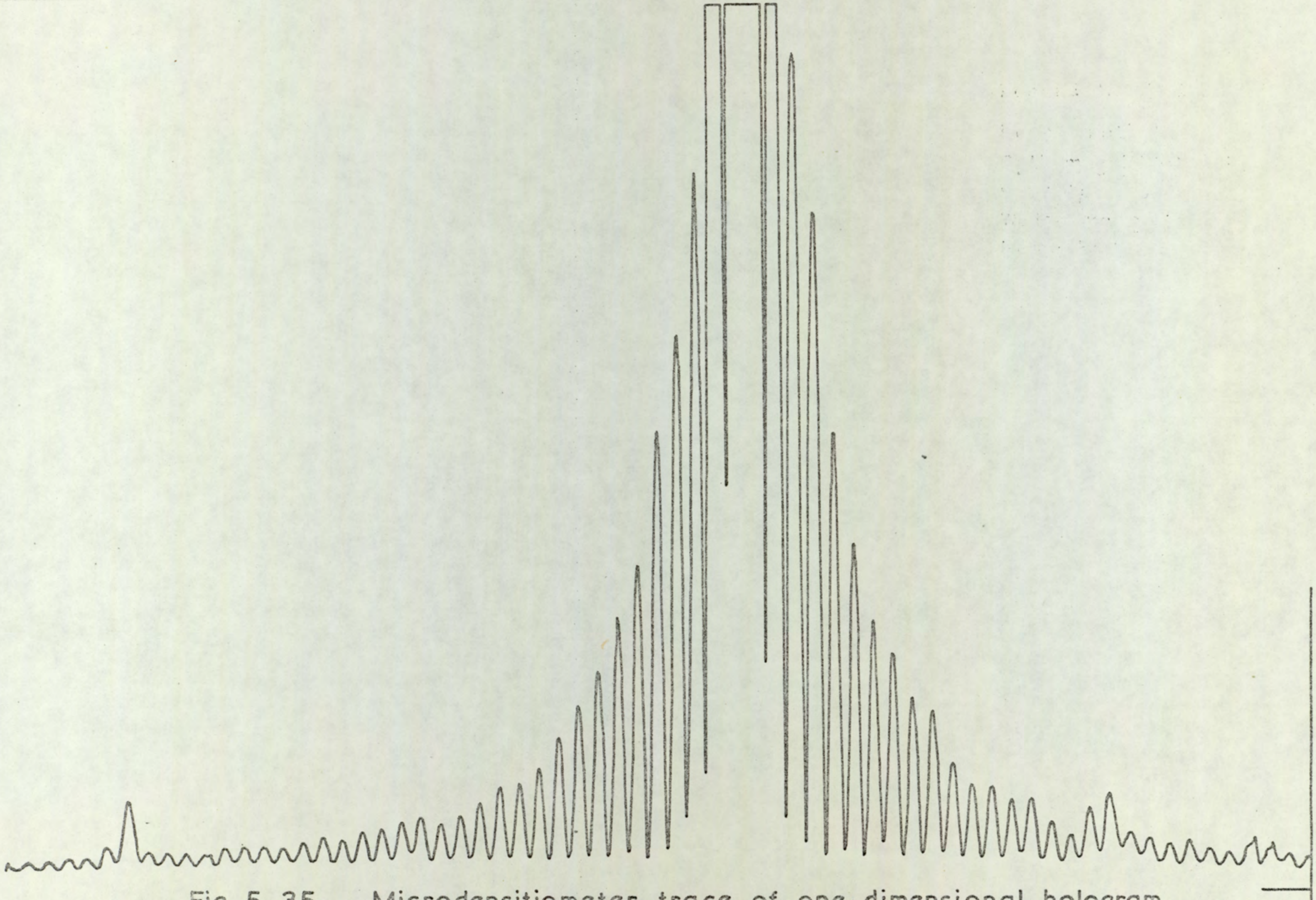


Fig. 5. 35 Microdensitiometer trace of one dimensional hologram

Hologram Condition	Non-linear Right hand image	Centre Image		Non-linear Left hand image
		Right	Left	
In Air	27/30	374	443	23/18
In Air	23/16	340	443	18/14
In empty liquid cell	28/20	340	391	24/13
In empty liquid cell	27/18	340	374	24/12
In liquid	24/14	391	545	9/4
In liquid	26/25	476	579	7/3
In liquid	21/22	340	426	7/3

The results are all expressed as mm of galvo deflection for a constant photomultiplier voltage. Values for the centre images were obtained by using a neutral density filter to attenuate the photomultiplier signal to within the galvanometer range and correcting for this in the table.

The results show that the liquid has apprently affected only one side of the reconstructed image, giving a reduction in intensiry of the non-linear image and an increase in the adjacent central image. There is a much smaller variation on the opposite side, where again, there is a slight decrease in intensity of the non-linear terms, and an increase in intensity of the central term.

The difference between the two halves of the reconstruction is not fully understood, but it is thought that there may have been assymetry in the incident transform or in the reference beam intensity, causing one half of the hologram to have more surface relief than the other half. In general, however, the results show that the index matching causes non-linear terms to be reduced, and allows more energy to fall into the true reconstructed image.

The reason for an unsymmetrical reconstruction obtained with the hologram in index matching liquid was examined by measuring the density variation across the hologram with a microdensitometer. This variation is shown in Fig 5.35, where the microdensitometer scanning slit was 5μ wide. Owing to the very wide range of densities formed in the hologram, accurate measurement of the first and central orders was not possible since the densiometer range was from 0 to 4. Nevertheless, it appears that the density variation either side the zero order is not symmetrical, particularly for the first three orders of the transform. This variation was probably caused by uneven intensity distribution across the object slit. The unsymmetrical appearance of the reconstruction is explained since the unsymmetrical density variations will be accompanied by approximately proportional phase variations.

5.6.4 Conclusion

The presence of phase variations in hologram filters caused by chemical action on the emulsion has been observed. The effect of removing these variations is to improve the filter response, by two or three hundred percent (in terms of autocorrelation peak intensity). Phase variations may be removed by using a liquid cell or gate containing liquid of matching refractive index. Within limits, a hologram may be exposed in air, and placed in a liquid gate after processing without much error.

There was no attempt to look for refractive index variations within the emulsion or within the substrate, as these would not be affected by use of the liquid cell. From the evidence of this work on surface phase variations, it seems likely that internal phase variations will also tend to degrade the autocorrelation response and if present should also be reduced.

5.7 Summary of photographic process limitations and effects

The substance of this chapter may be summarised as follows:-

1) The function to be recorded consists of interference fringes superimposed on a bias level which varies as the amplitude of the fringes varies. Because of the characteristics of the photographic emulsion (ie the shape of the t_A/Exp curve) the increases in bias level of the

fringes can cause the latter to be unrecorded, since there will be no variation in transmittance corresponding with the variation in intensity of the fringes.

2) The highest intensities in the fourier transform are located at the centre; ie at the zero and low spatial frequencies. Absence of hologram fringes in this region caused by overexposure, means that the hologram behaves as if it were a high pass spatial filter, transmitting only the high spatial frequencies.

In some ways, this attenuation of low spatial frequencies is a good thing, since it improves the discrimination of the spatial filter and hence it's performance, eg Binns, Dickinson and Watrasiewiz[25]. This may be understood by noting that different patterns will tend to differ most in the high spatial frequencies, and so it would be desirable to weight these frequencies, at the expense of the lower frequencies. In many cases, the overexposure may be regarded as recording the noise function $\frac{1}{S_{n(p,q)}}$ mentioned in chapter 1, since the function $S_n\,(p,q)$ will tend to be very high for low values of p and q except in the case of 'white' noise.

Another important result of overexposure is that the mean levels of the signal and input are effectively subtracted, since these appear at the zero spatial frequency. This is required by the theoretical discussion of pattern recognition (Chapter 1) and corrects for the input signals which are different in shape but give a higher cross correlation than the autocorrelation.

3) The absence of the zero and low spatial frequencies means that the reconstructed image contains only the profiles of edges where the spatial frequencies have been recorded. The autocorrelation function is split into three parts, each part corresponding to the correlation of the object with the edges of the reconstructed image.

The centre part of the autocorrelation function varies in both intensity and in width as the hologram exposure is varied; increase in exposure causes the width (or diameter) of the central peak to decrease. The variation appears to be approximately logarithmic, but this would depend on the intensity distribution in the signal spectrum. The autocorrelation function formed by non-linear recording has the advantage that a signal of relatively broad spatial dimensions can be precisely located to within a fraction of a millimetre.

4) Saturation of the photographic emulsion by the high intensities of side lobes in the transform causes a distorted fourier transform to be recorded, containing additional intensity maxima of different spacing to those in the original transform. The effect of these is to cause non-linear images to appear in both the reconstruction and the autocorrelation function.

5) The curvature of the t_A/Exp curve means that in some cases the interference fringes of the hologram are non-linearly recorded and this causes the appearance of more than one order of diffracted images to appear. These latter images are 2nd or 3rd order diffraction effects, and need not interfere with the required 1st order, unless the hologram carrier frequency is too low.

The effects of non-linear recording judged to be most important are those mentioned in paragraphs 2 and 4. The subsequent chapters will discuss methods of removing the non-linear terms of 4, whilst retaining the desirable features of paragraph 2.

6) Apart from non-linear intensity response, the photographic process causes phase variations to be introduced into the emulsion. These phase variations take the form of a surface relief effect and are most pronounced where the intensity of exposure is high. In fourier transform holograms the low spatial frequencies are nearly always very intense, and phase variations will therefore be found in this region. Their effect is to

dissipate light away from the focussed image, whether this is the reconstructed image or the autocorrelation function. The phase variations may be reduced by using a liquid cell containing liquid of appropriate refractive index, such as Decalin. This technique has been shown to improve autocorrelation response by 200-300%. There will also be a slight reduction in intensity of non-linear images.

Whilst much of the work in this chapter was being carried out, similar work was being done in the United States of America and in France. Recently, two papers have been published which are particularly relevant.

The behaviour of hologram filters as an example of spatial frequency band pass filters was discussed by Bulabois, Caron and Vienot[26] in a continuation of their study[27] on holography and spatial filtering. In this paper[26] the performance of hologram spatial filters of a rectangular aperture was studied in terms of the hologram reference beam exposure and the carrier fringe contrast. The performance of the filter was defined in terms of acuity (ratio of intensity of autocorrelation to intensity of a particular cross correlation) and of finesse (ratio of intensity of autocorrelation to width at mid height). Both the acuity and finesse were found to increase as the carrier fringe contrast was increased, and the latter was higher for high exposures than for low exposures, in agreement with the results of this chapter (section 5.5.4). The control of fringe contrast and exposure as a means of governing the acuity and finesse were discussed; in particular, the problem of reducing the selectivity (ie reducing the spatial frequency pass band of the filter) whilst maintaining a high acuity ratio was considered for pattern recognition problems where the signal does not differ greatly from it's environment.

The other paper which is considered relevant is that published by Ansley, Peters and Cassidy[28]. This paper discusses the shape of the autocorrelation function of a rectangular aperture as a function of the non-linear characteristics of the photographic emulsion. The authors measured the autocorrelation intensity profile for a variety of signal to reference beam intensity ratios and compared these with theoretical curves obtained by computation of data derived from the transmission/exposure curve of the emulsion. The triangular autocorrelation function of a rectangular aperture was only obtained when the signal to reference ratio was 0.286 (measured at the zero frequency of the signal spectrum, but the intensity of the autocorrelation increased to a maximum when the signal to reference ratio was unity. Further increases in this ratio were not observed, but this work does appear to be complementary with the work reported in this chapter.

Reference to Chapter 5

1. Gabor: Proc Roy Soc A197 (1949)

2. Stroke: Introd to Coherent Optics and Holography (Academic Press 1966)

3. P West, P R Archer: Sira Report, R408, June 1968

4. Bryngdahl and Lohmann: JOSA 58, 10. Oct 68. p1325

5. Friesem and Zelenka: App Optics 6 10. Oct 1967 p1755

6. A Kozma: JOSA Vol 56, 4. April 1966 p428

7. Kozma and Kelly: App. Optics 4, 4, April 1965 p387

8. Kozma: Optica Acta, 15, 6, 1968 p527-551

9. Van Ligten: JOSA, 56 1. Jan 1966 p1

10. D G Falconer: Phot Sc & Eng. 10, 3. May-June 1966

11. A C Marchant, G E Keyte et al: RAE Tech Report No 67031

12. G E Keyte: XVIIth AGARD Avionics Panel Tech Symp. Sept/Oct 1969 Norway

13. Lowenthal and Belvaux: Optica Acta 14, 3, (1967), 245-258 RAE Library Trans 1265

14. Lowenthal and Belvaux: Rev d'Optique 46, 1, (1967) 1-29 RAE Library Trans 1278

15. Jenkins and White: Ch 15, p302, 3rd Ed McGraw Hill 1957

16. K G Birch: Optica Acta 1968, vol 15, No 2 p113-127

17. G W Jull: XVIIth AGARD AV Panel Tech Symp 1969

18. John N Latta: Applied Optics, 7, 12, Dec 1968

19. T A Shankoff: Applied Optics, 7, 10, Oct 1968

20. Urbach and Meier: Applied Optics, 5, 4, April 1966

21. Leith: Photographic Science & Engineering, 6, 2. March/April 1962

22. H M Smith: J.O.S.A. 58, 4. April 1968

23. H Hannes: Optik 26 (1967/68) No 4, 363-380 RAE Library Translation 1323

24. A L Ingalls: Photographic Science and Engineering 4, 3, May/June 1960

25. Binns et al: Applied Optics, 7, 6, June 1968, p1047

26. Bulabois, Caron, Vienot: Optics Technology 1, 4, August 1969, p191-195

27. Vienot, Bulabois: Optica Acta, 14, 1, 1967, 57-70

28. Ansley, Peters and Cassidy: XVIIth AGARD Avionics Panel Tech Symposium (Sept/Oct 1969)

Chapter 6

The Optical Correlator

In this chapter experiments are described which investigate the use of complex spatial filters in the complete optical system. The effect of optical system aberrations upon the hologram response are examined, and also practical problems concerned with use of photographic transparency input data.

6.1 Effect of optical system aberrations on hologram response

6.1.1 Introduction

When the optical correlator system is used to search a given input format or transparency, it is required that the autocorrelation function of the signal or pattern be the same intensity and shape irrespective of the signal position in the input format.

There are several effects which could cause the autocorrelation function to change, depending on the signal position, and some of these are tabulated below:-

1) Variation in intensity of the collimated beam across the input plane. Causes reduction in intensity of autocorrelation pattern.

2) Aberrations of the collimator beam.
3) Aberrations of the transform lens.

Both these effects can cause movement of the fourier transform and degrade it, producing a decrease in autocorrelation intensity and possibly a change in its shape (Fig 6.1).

4) Variation in hologram response at different signal beam angles. Could cause reduction in autocorrelation intensity.

5) Aberrations of integrator lens. Will degrade autocorrelation point and cause reduction in intensity and perhaps change of shape. (Fig 6.1).

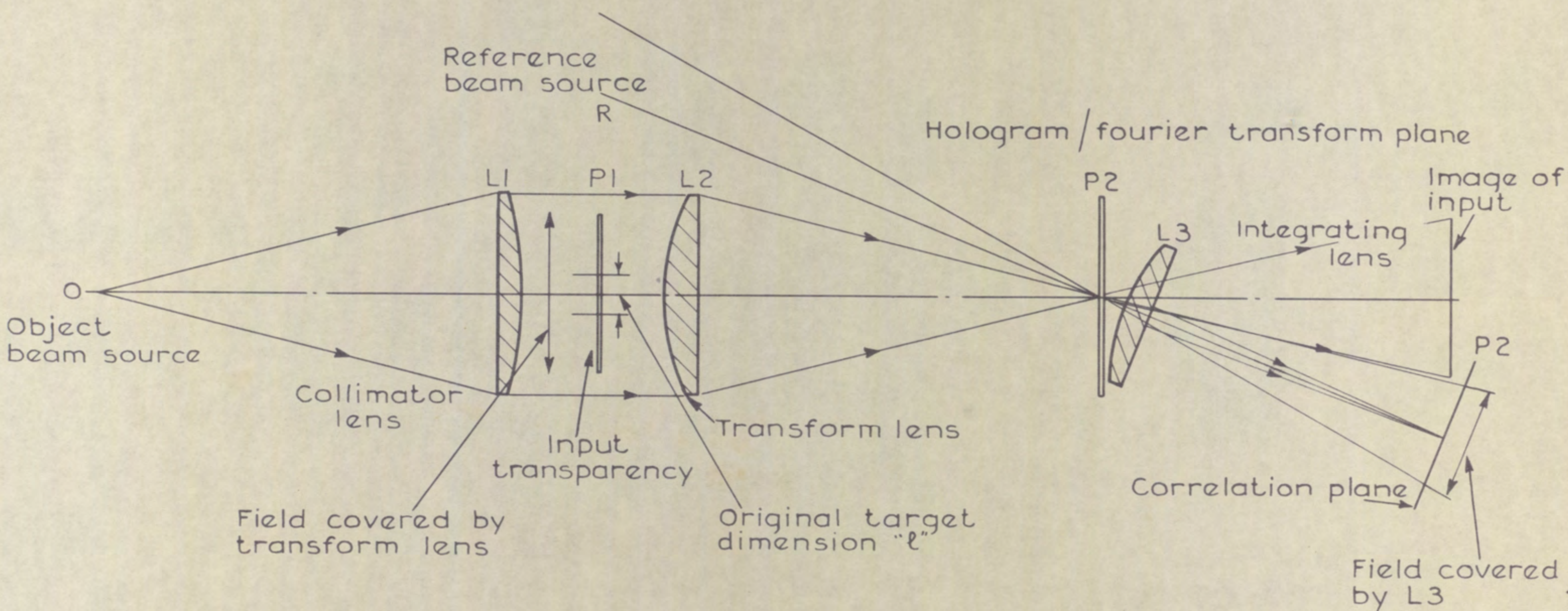


Fig. 6.1 Optical correlator - showing fields to be covered

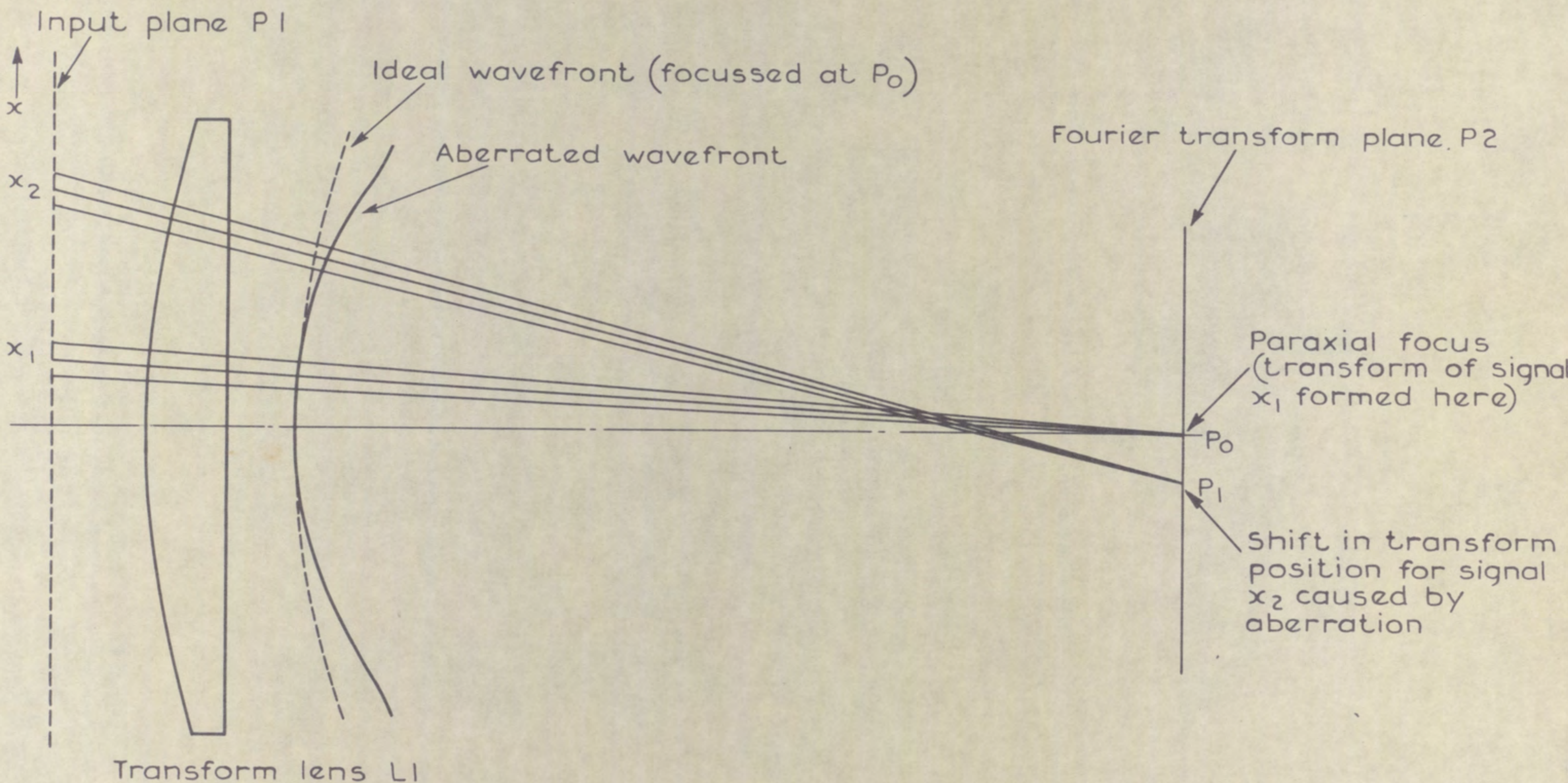


Fig. 6.2 Effect of spherical aberration on transform position

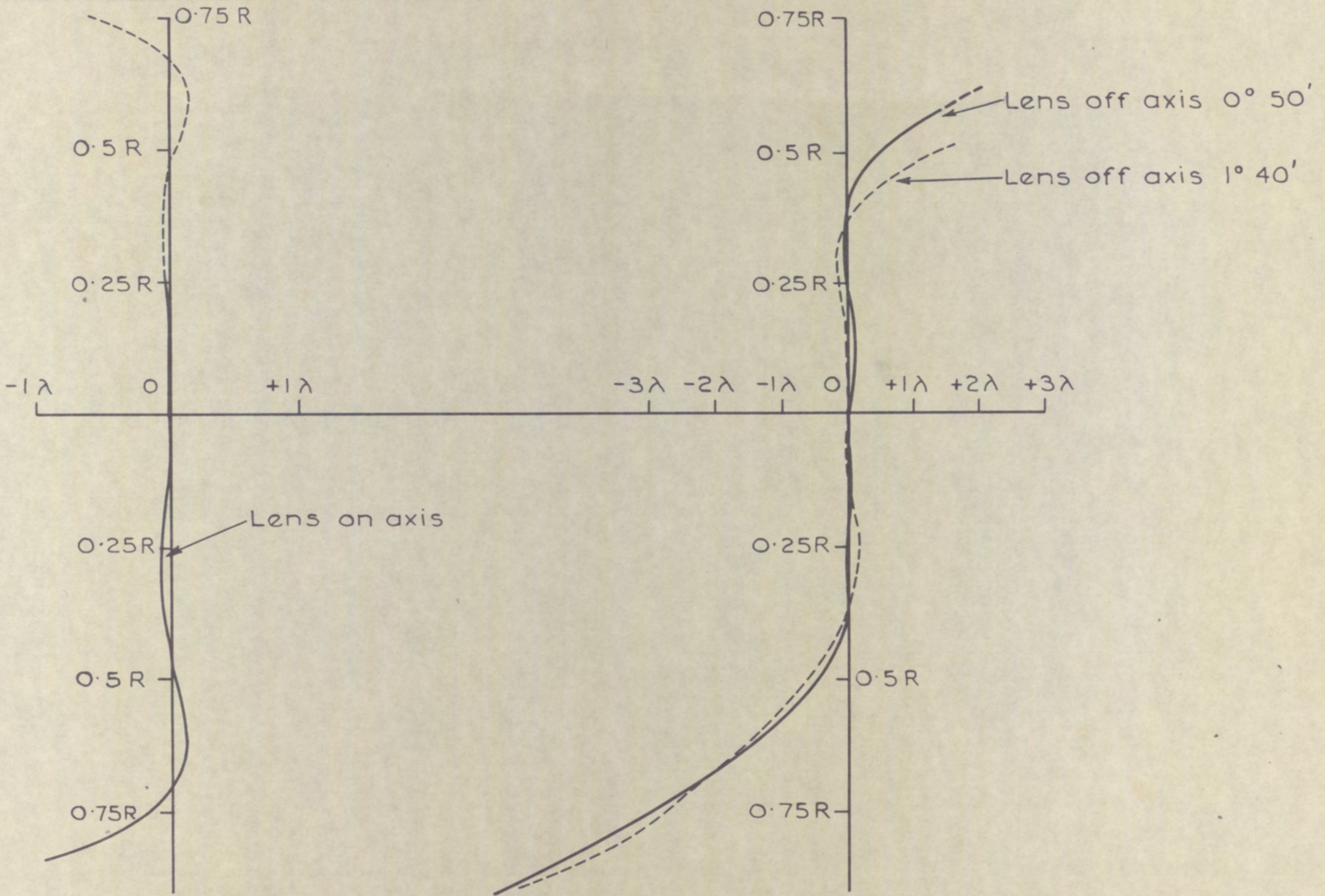


Fig. 6.3 Wavefront aberration curves for integrating lens (L3)

Of these five factors, the fourth, is considered negligible because of the narrow angle subtended by the input format at the hologram (4^{o}).

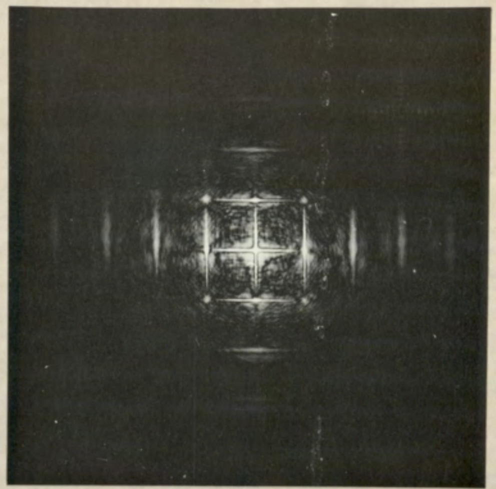
Tha variation in collimated beam intensity can be measured and used to correct the autocorrelation function (ACF) intensity, assuming that the intensity distribution remains unchanged during the course of the experiment. The variation in intensity caused by the aberrations of both the transform and integrating lens ought to be constant for any given system, and in theory could be allowed for when observing the ACF response. For relatively small patterns or signals (compared with the transform lens aperture), the presence of spherical aberration will cause a lateral shift in position of the transform, which could be corrected by movement of the hologram. This is shown in Fig 6.2.

6.1.2 Aberrations of Integrator Lens

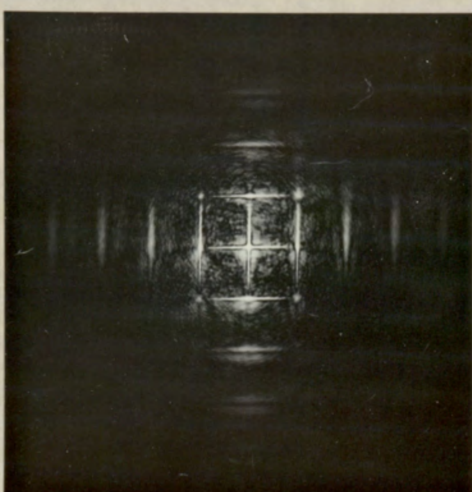
Before investigating the effect of collimator and transform lens aberrations it was necessary to consider the effect of integrator lens aberrations. As stated in Chapter 4, a doublet lens was chosen here, to minimise the number of surfaces which could scatter light, and it was offset at an angle equal to the reference beam angle to reduce the effect of off axis aberrations. Such a lens will only give good resolution for points close to its optic axis, and this will limit the field which can be usefully employed in the input plane.

The performance of the integrator lens was measured in a similar way to that used for the transform lens (described in Chapter 4), ie: by using a shearing interferometer. Fig 6.3 shows tangential wavefronts derived from shearing interferograms for an object point 0^{o}, 0^{o} 50' and 1^{o} 40' off axis. The wavefront for 0^{o} shows the usual overcorrection for spherical aberration at three quarters of the aperture, whilst the wavefronts for both 0^{o} 50' and 1^{o} 40' show coma above one quarter of the aperture radius. [N.B. In these measurements the object point was

Aperture on axis



7mm Off axis



17mm Off axis

26mm Off axis

Aperture 26mm off axis but hologram moved to correct for transform displacement

Fig. 6.4 Autocorrelation response of rectangular aperture hologram arranged to be at finite conjugates, to correspond with the position of the reference source.

It is evident from these measurements, that high spatial frequencies in the correlating process will suffer more from the effect of integrator lens aberrations than low spatial frequencies. Taking a quarter of the radius of the integrator lens as a limit below which off axis aberrations were not present, one finds that this limit corresponds to a spatial frequency of approximately 15 lines/mm.

The off-axis aberrations will become more severe for object points which are even further off-axis, but the angle of $1^{\circ}\ 40'$ corresponds roughly to maximum semi angle subtended by the input field at the integrator lens, ie: approximately 2°.

In view of the off axis performance of the doublet lens it was decided to replace the latter with a well corrected lens whilst studying the effect of transform lens aberrations. The lens chosen was a 25 cm f/8 Anastigmat, giving good performance over the required field, with a maximum useful aperture of 3 cms, compared with 2.0 cms of the doublet lens. (Although the aperture of 3 cms is still low compared with the maximum diameter of the doublet lens, it is equal to the width of the hologram plate holder, and a larger aperture would be unnecessary).

6.1.3 Variation of ACF with object position - rectangular aperture

The variation in ACF with object position was first studied using a rectangular aperture as an object or signal. The rest of the lens aperture was obscured using an opaque screen. The ACF was observed by recording the function on 35mm film. A selection of recordings are given in Fig 6.4, showing the appearance of the ACF with the aperture in the centre of the lens, 7mm, 17mm and 25.8mm off axis. Also shown is the change in ACF when the hologram is moved horizontally to correct for spherical aberration, the movement being measured as 0.037mm in opposition to the movement of the aperture, and in agreement with Fig 6.2.

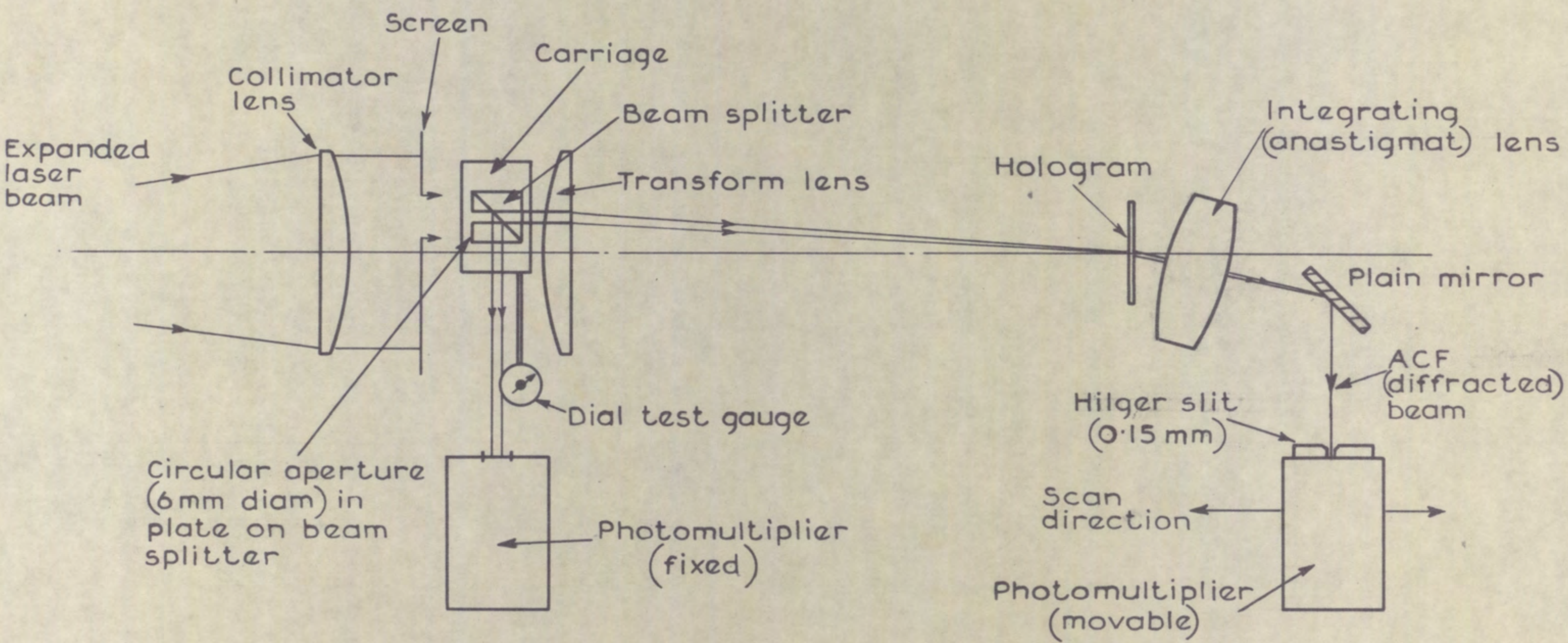


Fig. 6. 5 Measurement of object beam / ACF intensity relationship

A study of the recordings made at 1mm intervals revealed that there was no appreciable visual change until the aperture was 17 mm off axis (ie: roughly $\frac{2}{3}$rds lens aperture). Small changes in the distribution of light in the ACF pattern could be seen, although the pattern as a whole was substantially unchanged. After 22mm off axis, the change in the ACF became more pronounced. The central cross became weak whilst the side peaks either side the central peak remained unaltered. Since the exposure and processing was kept constant the number of nonlinear terms seen for each recording was found to decrease from 4 with the aperture in the centre, to 2 with aperture at the edge of the lens.

In general, however, the ACF maintained a rectangular shape over the entire range of aperture positions, and it was thought that the most significant variation would probably be in intensity rather than in shape. These measurements are described in the following section.

6.1.4 Variation of ACF with object position - circular aperture

A circular aperture was used in the following section because the energy in the transform is more uniformly distributed, and it was considered that aberrations acting in one direction would have a less unbalanced effect. A beam splitter was also used in this work to obtain simultaneous measurements of laser beam intensity and autocorrelation function intensity, as shown in Fig 6.5.

The circular aperture (6mm diam) was drilled in a brass plate and attached to the front face of a beam splitter. In this way the aberrations of the beam splitter would be kept constant. The reflected beam was measured by a photomultiplier, directly coupled to the ultra-violet recorder. The autocorrelation function was measured with another photomultiplier, fitted with an adjustable slit and its output connected, via amplifier and 6 Kc/s filter to another channel of the recorder. Thus a simultaneous

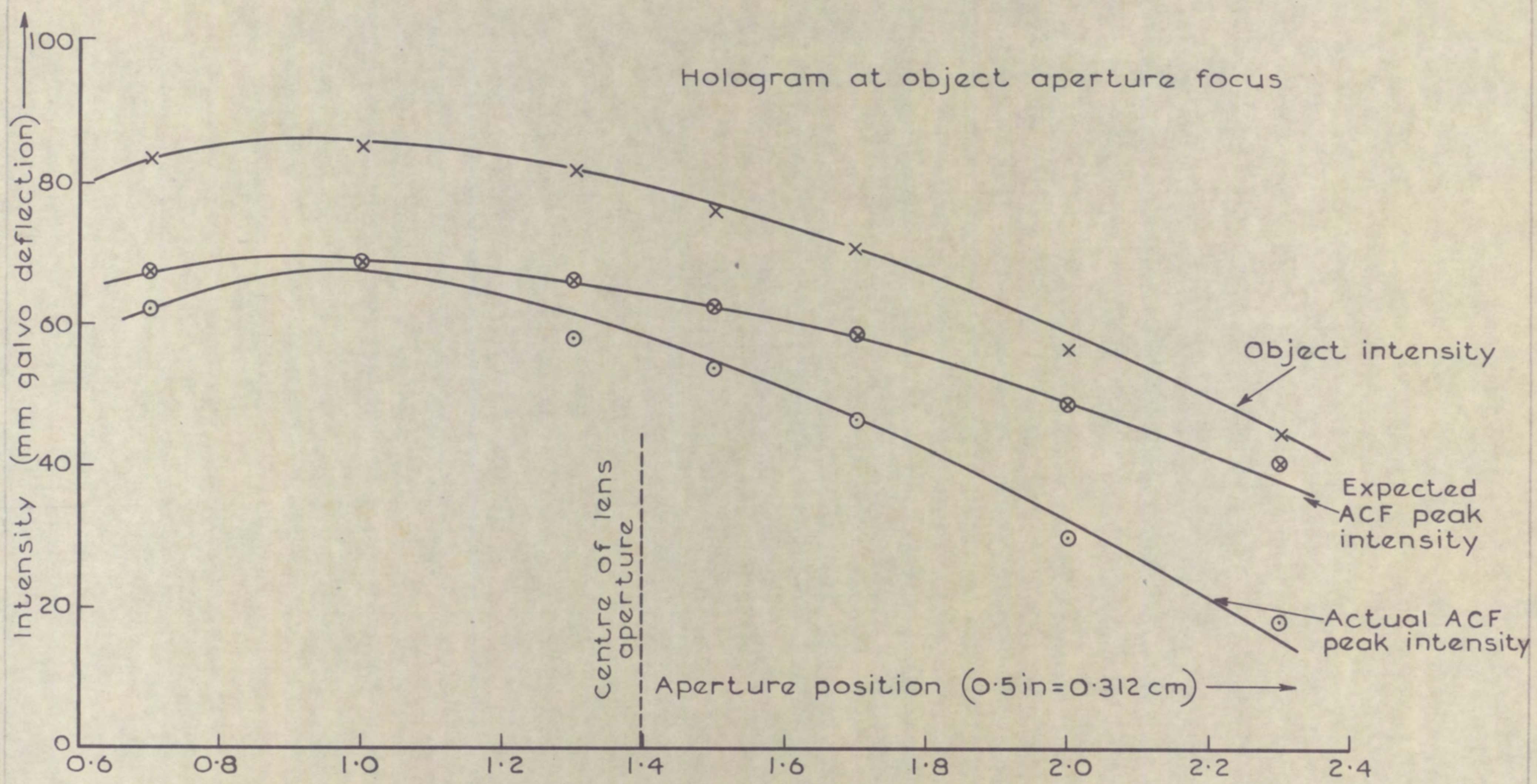


Fig. 6.6 Autocorrelation peak intensity and object intensity v aperture position

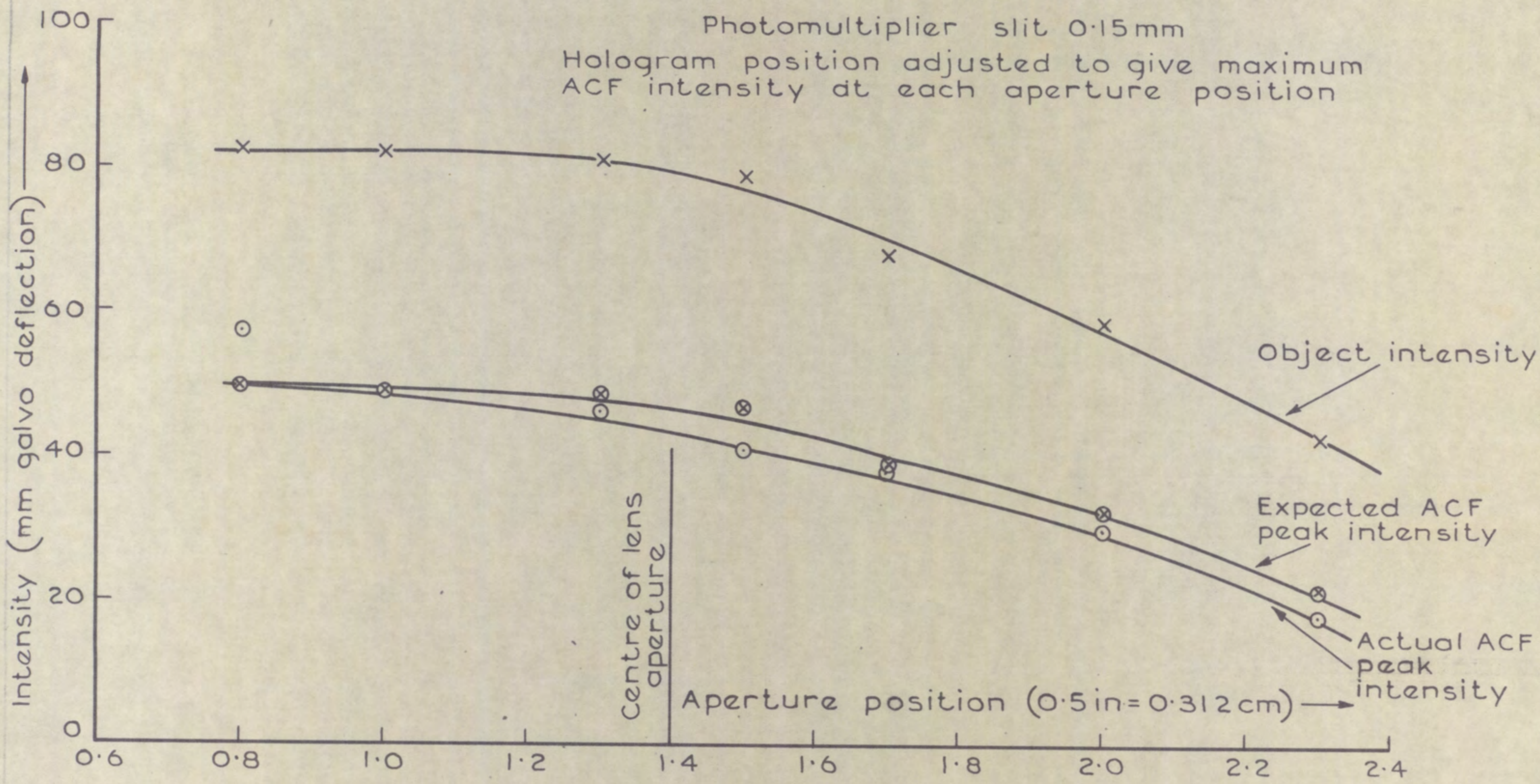


Fig. 6.7 Autocorrelation peak intensity and object intensity v aperture position

record could be obtained of the object intensity and of the autocorrelation function. (Note; although the laser was modulated at 6 Kc/s, only 10-20% of the laser beam intensity was thus modulated, and the first photomultiplier measured the dc component only.)

The position of the circular aperture was monitored by a dial test gauge (as before) and to extend the range of movement, a linear potentiometer was also attached.

The response of the two photomultipliers was related by varying the object beam intensity with neutral density filters. (A circular aperture hologram with its centre and 1st ring over-exposed was used in this calibration). Both photomultipliers had a linear response, and by measuring the slopes of the two lines obtained, changes in object intensity could be related to changes in ACF intensity.

The variation in object beam and ACF peak intensity together with the expected curve for ACF intensity (based on object beam variation only) is shown in Fig 6.6. The most intense part of the object beam was off axis relative to the transform lens, caused by a small displacement of the spatial filter and beam expander.

The autocorrelation intensity is below that which would occur if intensity variations of the object beam alone were present. This suggests that aberrations in the transform lens are effective. Since spherical aberration effectively moves the transform laterally, it may be corrected by moving the hologram in the appropriate direction, and aligning it with the transform. This was accomplished in a subsequent experiment using a small plane mirror immediately behind the hologram and observing both hologram and transform with a microscope (2" objective). In this way the curves of Fig 6.7 were obtained. These curves show fair agreement between the actual and predicted autocorrelation intensities.

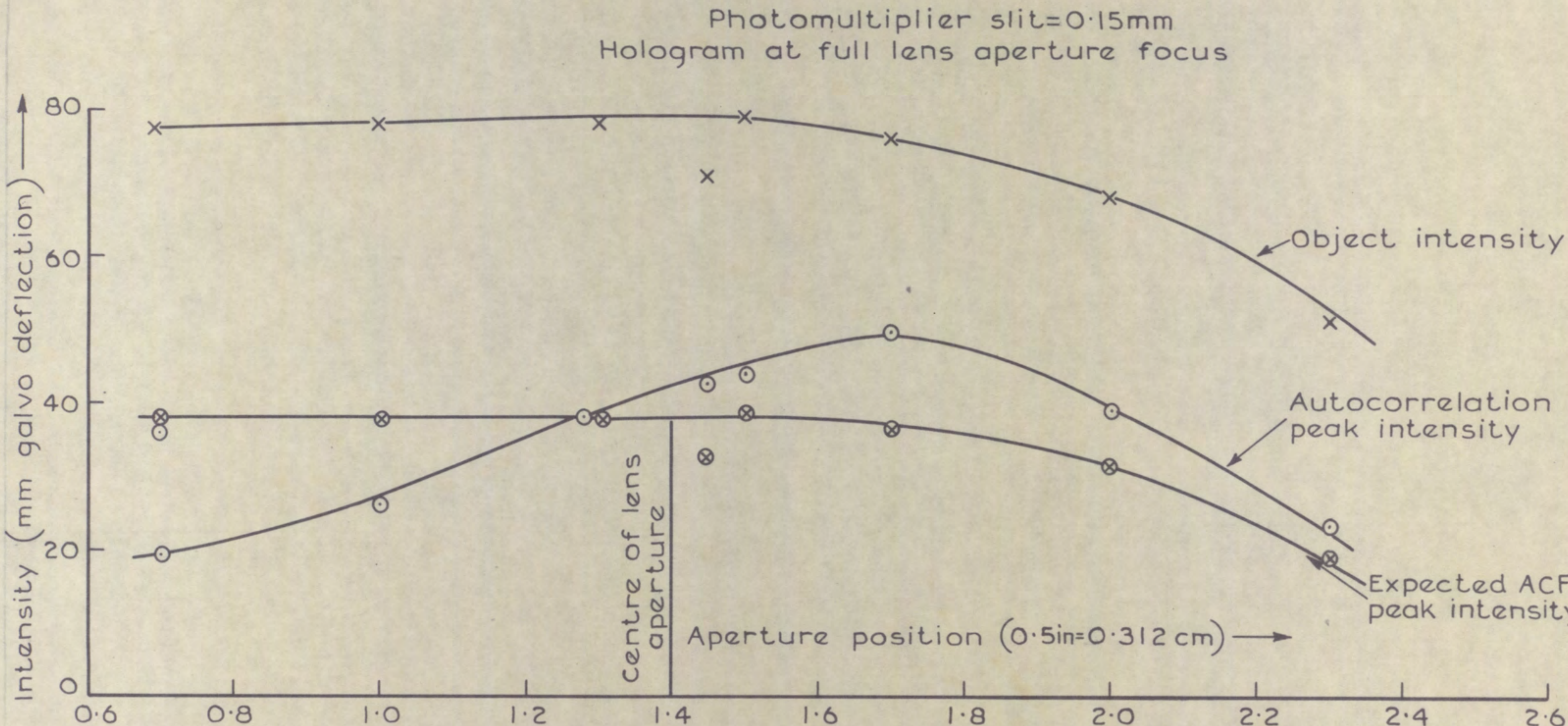


Fig. 6.8 Autocorrelation peak intensity and object intensity v aperture position

It was found that much of the transform movement was caused by incorrect positioning of the hologram in the focal plane of the transform lens. In the foregoing experiments the hologram was aligned with the fourier transform of the circular aperture, which effectively placed the hologram in the paraxial focus, since the object aperture was considerably smaller than the total lens aperture.

The measurements were therefore repeated, except that now the hologram was focussed with the transform lens at full aperture.

Fig 6.8 shows the relation between ACF intensity and aperture position with the hologram at the full aperture focus. There is a marked change in the appearance of the ACF intensity curve, although the object intensity distribution is very similar. Good agreement between the expected ACF intensity and that actually measured was not obtained; this could be obtained by lateral adjustment of the hologram position, and indicated that misalignment of the transform and the hologram was still responsible for the difference in results.

6.1.5 Conclusions

Although no definite relation between spherical aberration and ACF intensity for simple apertures could be established, no further work in this direction was carried out because of insufficient time, and because results obtained indicated that more attention be paid to precise alignment of optical equipment. There is evidence that work needs to be done in this direction, perhaps as a specific project. Such work must examine not only the alignment of optical components, but also the accuracy of the optical bench and bench fittings.

Some general conclusions may be drawn from the work described in this section.

The first is that the integrator lens must be chosen to give good performance over the object field. The effect of using a lens having poor resolution over even a narrow field angle was found to be quite severe. The disadvantage in using a more well corrected lens is that the extra components and surfaces can cause multiple reflections and scattering of light. It ought to be possible to design a lens specifically for the task, having a minimum of components and surfaces, but giving good resolution over the field.

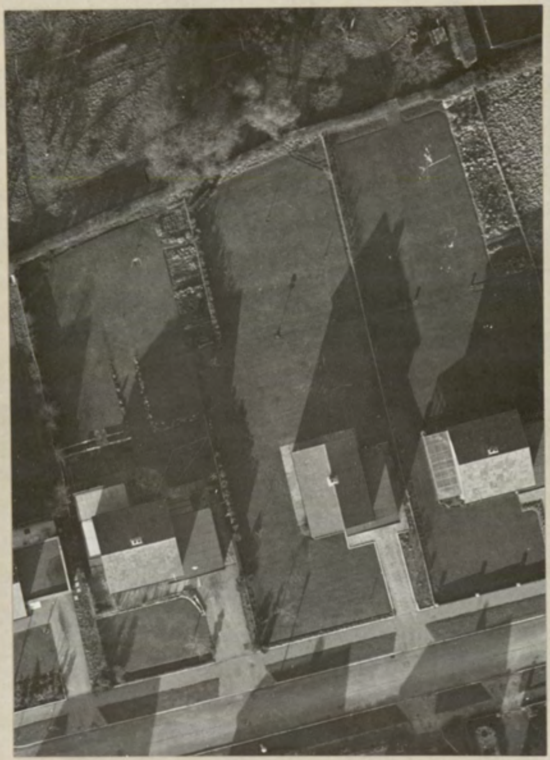
The most significant cause of variation in ACF intensity seems to be movement of the transform, caused by spherical aberration or by incorrect focussing. Realignment of the hologram restored the ACF intensity to its original strength (allowing a correction for the object beam intensity). The focussing of the hologram was critical; to avoid errors in focussing and to minimise translation caused by spherical aberration it was necessary to focus on the transform of the whole input plane, rather than of the small aperture. The effect of spherical aberration could be removed by designing a lens corrected for the purpose, or by use of a corrector plate. No other serious variations appeared to arise during translation of the aperture in the input plane. In the case of the rectangular aperture, the general shape of the correlation pattern remained unaltered, apart from a decrease in the number of non-linear terms, and a change in the relative brightness of the central and side peaks of the autocorrelation function.

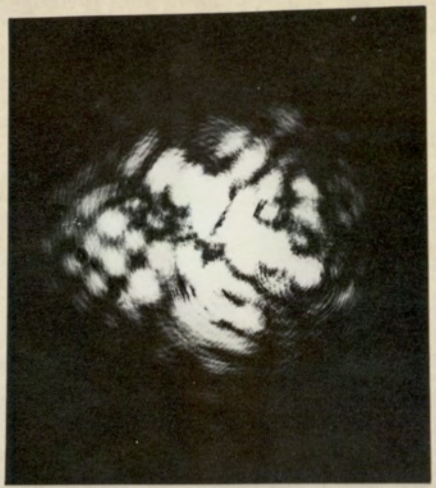
6.2 Use of Photographic Transparencies as input media

6.2.1 Preliminary study of use of transparency inputs

The work described in this section is concerned with the problem of recognising patterns in photographic transparencies, with particular reference to applications in aerial reconnaissance.

Original air photograph transparency





1 Transparency in air, no liquid gate in system

2 Transparency in air liquid gate in system

3 Transparency in liquid gate

4 Transparency replaced by uniform density film

Fig. 6.9 Effect of liquid gate on fourier transform of transparency

As a starting point, it was decided to look for differences in the fourier transforms of different aerial photographs. The reason for this approach is that aerial photographs can be of low contrast, and their transforms thus tend to be weak, and perhaps easily confused with scattered light and film grain noise. The examination of the transforms of different photographs would indicate which details would be recorded strongly in fourier transform holograms.

A set of 35mm photographic transparencies of aerial scenes was obtained, and the fourier transform of each transparency recorded using a microscope camera behind the fourier transform plane. In this first experiment, no liquid gate was used, since the gate described in Chapter 4 was still under construction. Attempts to correlate the appearance of the fourier transforms with their parent transparency were unsuccessful. The lack of correlation between features in the transparency and in the transform was attributed to a) insufficiently high exposure to record faint details of the fourier transforms, and b) the presence of phase or film thickness variations in the transparencies which cause the appearance of unsymmetrical fourier transforms.

A very simple liquid gate was constructed, using two sheets of glass 6mm thick, sandwiching an aluminium alloy frame, shaped to hold the transparency frame. The cell was sealed using araldite adhesive, and contained "Decalin" as the refractive index matching liquid. Using this liquid gate or cell, the transform of one of the original air photograph transparencies was re-examined. Four results are shown in Fig 6.9, showing the fourier transform for each of the following cases:-

(1) Transparency in air, and no liquid cell in optical system,

(2) Transparency in air, but with liquid cell in optical system,

(3) Transparency in liquid cell,

(4) Transparency replaced by 35mm film of uniform density.

All photographs received identical exposure and development.

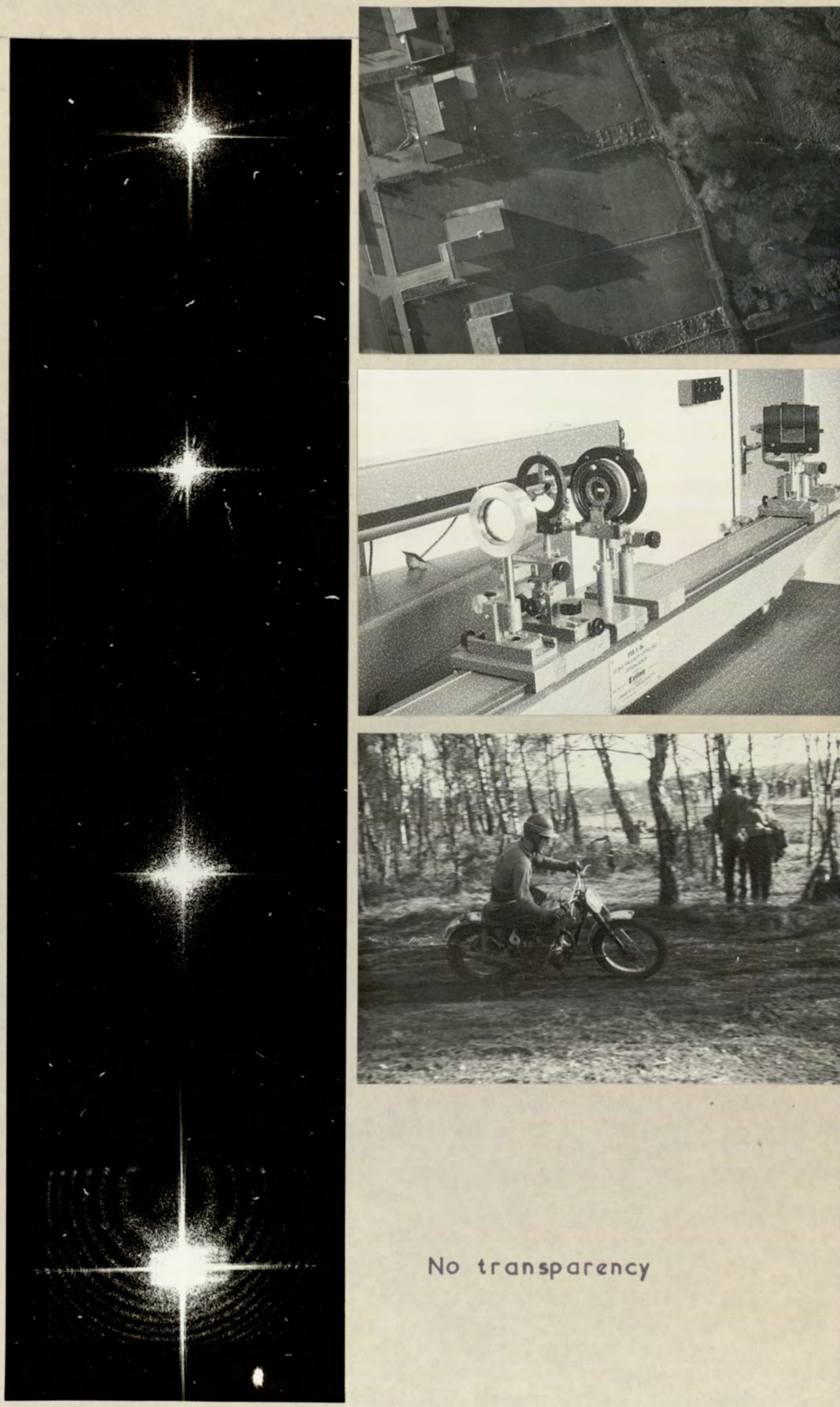


Fig. 6.10 Appearance of fourier transforms of different transparencies

The most striking result is the difference between numbers (2) and (3), showing the removal of film thickness effects. The transform of (3) consists of a cross, surrounded by a concentric ring pattern. The ring pattern was caused by diffraction by the circular aperture placed in front of the liquid cell, of about 15mm diameter, with the intention of reducing the strong diffraction pattern from the 35mm frame of the transparency. The cross appears to be part of the actual transparency transform since the latter contains strong features at right angles. The fourth photograph also shows a faint cross, whereas here there was no transparency to be transformed. This suggested that there may be astigmatism in the optical system, perhaps introduced by the liquid cell.

Further attempts were made to identify the fourier transform of the transparency by recording the transform directly on photographic plate and by using a rectangular aperture instead of the circular aperture. In the former case, no conclusive evidence of fourier transform was found, even though a range of exposure from 1 to 1000 was given. The circular ring pattern caused by using a circular aperture seemed to be very prominent and could well obscure relatively faint detail.

In a subsequent series of tests, recordings were made of the transforms several transparencies using the 35mm frame as the maximum aperture. Although this produced a strong diffraction pattern, the diffracted energy was concentrated in two narrow lines, leaving a relatively large area of fourier transform space in which the transparency transform could be seen. These results are shown in Fig 6.10, together with their parent transparencies. In all cases, the liquid cell was used. In the first two transparencies details can be seen of the fourier transform which can be correlated with details in the original transparency. The absence of prominent lines in the third transparency produces a more symmetrical

1. Transparency in liquid gate

2. Uniform film in liquid gate

3. No film, liquid gate alone with rectangular aperture

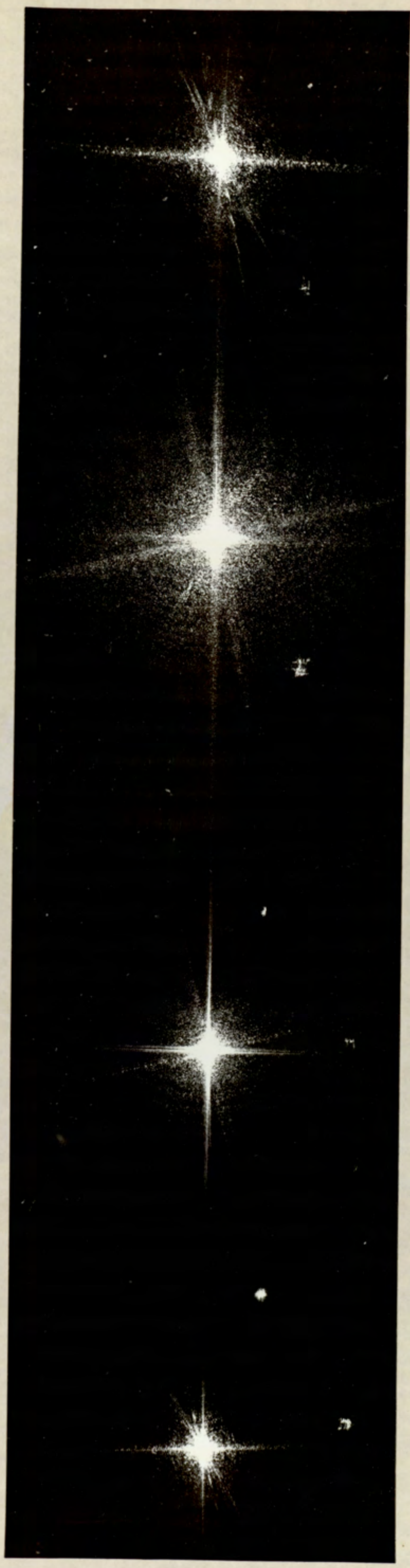
4. No film or gate, rectangular aperture only

Fig. 6.11 Appearance of fourier transform with different input conditions distribution of energy in the fourier transform. The last result, for which no transparency was present, (an empty 35mm frame was used) shows a large amount of scattered light in the transform plane, and the appearance of rectangular shapes, caused by multiple reflections at the air/glass surfaces in the optical system.

These results were successful in showing the presence of the fourier transforms of different transparencies but also indicated that the transform plane contained a considerable amount of scattered light which could not be attributed to an input transparency. The scattered light was thought to originate either at the liquid cell, or at the grains of a photographic input transparency. This point was investigated by recording the fourier transform with and without film and liquid cell in the input stage. The film was a uniformly exposed frame of 35mm film contained in a rectangular 35mm mount. (The mount was placed in the input stage in all these exposures.) Some compensation was necessary to allow for the different transmission of each type of input; these are listed below:-

1.	Object transparency + liquid cell	0.2 ND
2.	Uniform density film + liquid cell	No filters
3.	No film, with liquid cell	1.5 ND
4.	No liquid cell; ie: rectangular aperture alone.	1.6 ND

Also shown are the neutral density filters used with each input to compensate for differences in transmission. The corresponding fourier transforms are shown in Fig 6.11 The most significant observation is the difference between the 3rd and 4th exposures, when the liquid cell was removed indicating that much of the scattering originated at the liquid or at the cell. The difference between the 2nd and 3rd exposures was not pronounced, indicating that there is relatively little scattering from the film itself. The first exposure shows the fourier transform of the object



1. Lab. scene transparency
no liquid gate

2. Air photo transparency
no liquid gate

3. Air photo transparency
in liquid gate

4. Lab. scene transparency
in liquid gate

Fig. 6.12 Effect of liquid gate on fourier transforms of transparencies transparency which contains prominent diffraction lines as well as a concentration of light about the centre. These results suggest that the liquid cell is responsible for much of the scattered light in the transform, and that the transparency itself contributes energy in a similar way, ie: some of the transform energy appears as if it had been scattered. (This point will be discussed in detail later.)

The effect on the transform of varying the density of the uniform density film was examined, but no significant variation in the appearance of the transform was observed.

Finally, a further examination of the effect of the liquid gate upon the transforms of different transparencies was made; the results being shown in Fig 6.12. In this experiment, the transforms were recorded with both transparencies immersed in the liquid cell, and with both transparencies outside the cell.

The results show that when the transparency was immersed in the cell, there is a reduction in both the prominent diffraction lives and in the "scattered" light. The effect of the liquid cell was to remove film and emulsion thickness variations and since the latter will accompany the density variations in the emulsion (caused by the tanning effect), one would expect prominent features of the transform to be reduced. Scratches and marks on the emulsion will also be matched by the liquid, and this reduces the general scatter of light about the transform. These effects are more marked for the 2nd and 3rd photographs, and suggest that the transparency in the 1st and 4th photographs is less scratched.

General conclusions which could be made at this stage were:-

1. Use of a liquid cell or gate enables the thickness variations in a film or transparency input to be removed. Although this is accompanied by some reduction in the intensity of the required transform, (because of

tanning action) the overall result is improved because the transform is that due only to the density variations of the transparency.

2. The liquid cell causes some of the light to be scattered, giving a grainy appearance to the recorded transform. This is not present when the cell is removed, neither does it appear to be caused by film granularity. The graininess does not seem to be too obstrusive, and features in the required transform may still be discerned.

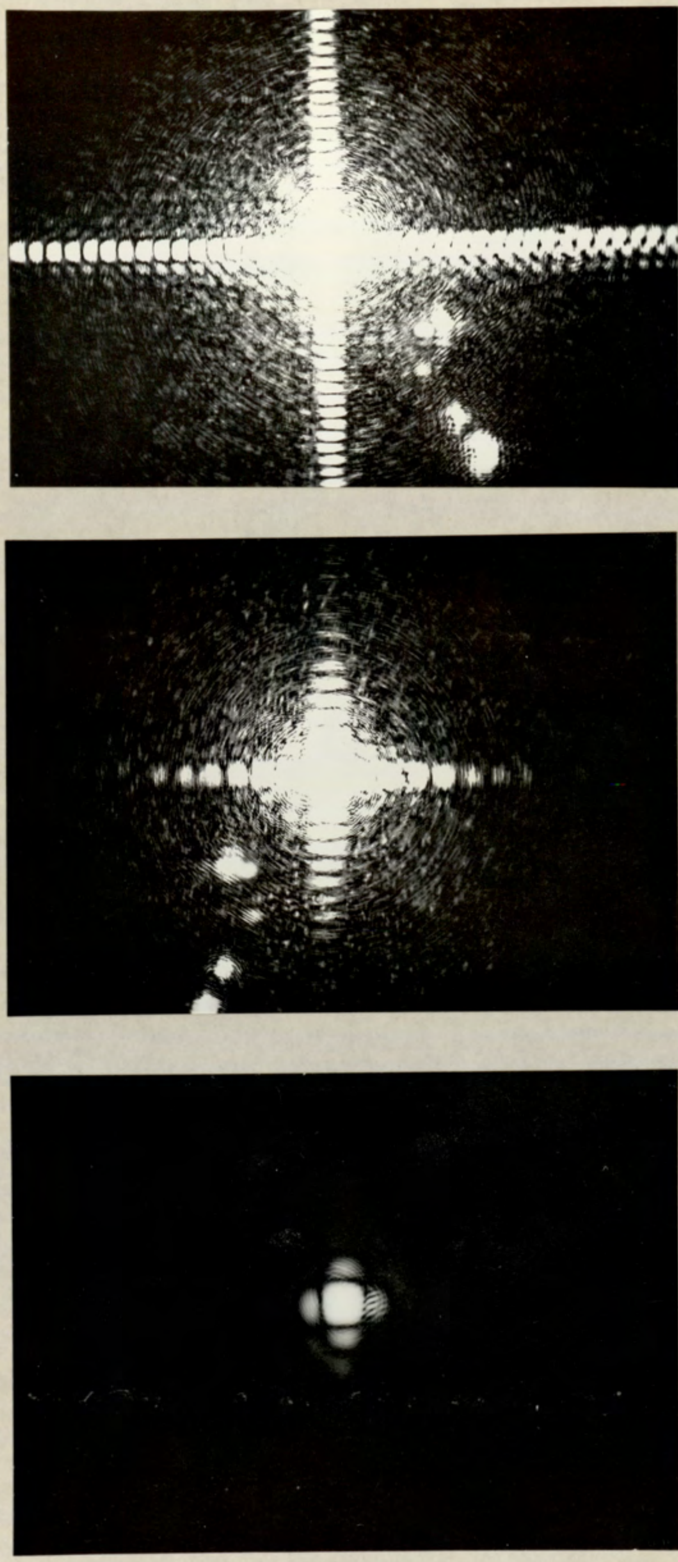
3. There is a reduction in scattered light when the liquid cell is used which seems to be caused by matching of scratches and blemishes on the emulsion.

4. The use of a rectangular aperture enables the fourier transforms of transparencies to be seen quite clearly, although this would be undesirable in holograms because of the strong diffraction pattern which is produced.

<u>6.2.2 Use of Apodized Masks with transparencies</u>

Although the use of a rectangular aperture or frame with the transparency enabled the fourier transform of the latter to be seen, the presence of the transform of the aperture itself is undesirable. The aperture transform is very distinctive, compared with the transparency, and there would be strong correlation between rectangular apertures, irrespective of the details on the transparencies contained within.

A way of overcoming this problem is to reduce the intensity of the rectangular aperture transform by reducing the rate of change of amplitude transmission at the edge of the aperture; ie: by using a graded boundary, or apodized mask. The basis of this solution is simply that the fourier transform will contain high spatial frequencies depending on the rate at which the amplitude is changed at the aperture edge; if the rate is reduced then the transform will contain relatively low spatial frequencies. To obtain a minimum of diffraction by the aperture edge it is necessary that



Rectangular aperture
ungraded [sharp] edge

Rectangular aperture
graded [apodized] edge

Small aperture with same
amount of grading

Fig. 6.13 Effect of apodized masks on fourier transform of rectangular aperture the amplitude be changed in a definite way: for example, a gaussian distribution of amplitude across the aperture would produce a gaussian distribution of amplitude in the fourier transform. Similarly, a sinusoidal variation will produce a transform having only one spatial frequency in the appropriate direction.

It is not easy in practice to produce amplitude variations according to a prescribed formula, but quite good masks can be made by defocussing the sharp edges of an aperture. Fig 6.13 shows the fourier transforms of a rectangular aperture, one with sharp edges, the other being a transparency of the same aperture but with the edges defocussed. The reduction of intensities in the high spatial frequencies is marked; only seven side lobes can be seen either side the central lobe of the transform.

If the edge is defocussed still more, it would be possible to obtain a further reduction in the high spatial frequency content of the transform. This could be achieved only at the expense of reducing the clear area of the mask; in the limit, the "ideal" mask would have maximum transmission only in the centre, falling away gradually to zero on either side. This would produce a minimum of diffraction in the transform, but would also be inefficient as a mask for the input transparency. To some extent, there is a compromise between the amount of blurring or grading of the edge, and the amount of clear transparency required.

If the amount of blurring is kept constant, but the overall area of the mask reduced, then the intensity of higher orders in the diffraction pattern becomes smaller, as shown by the last illustration in Fig 6.13. Such a mask, having a small transparent area, with a relatively large degree of blurring at the edges, would be ideal for selecting target areas on larger transparencies, from which the hologram filter would be made.

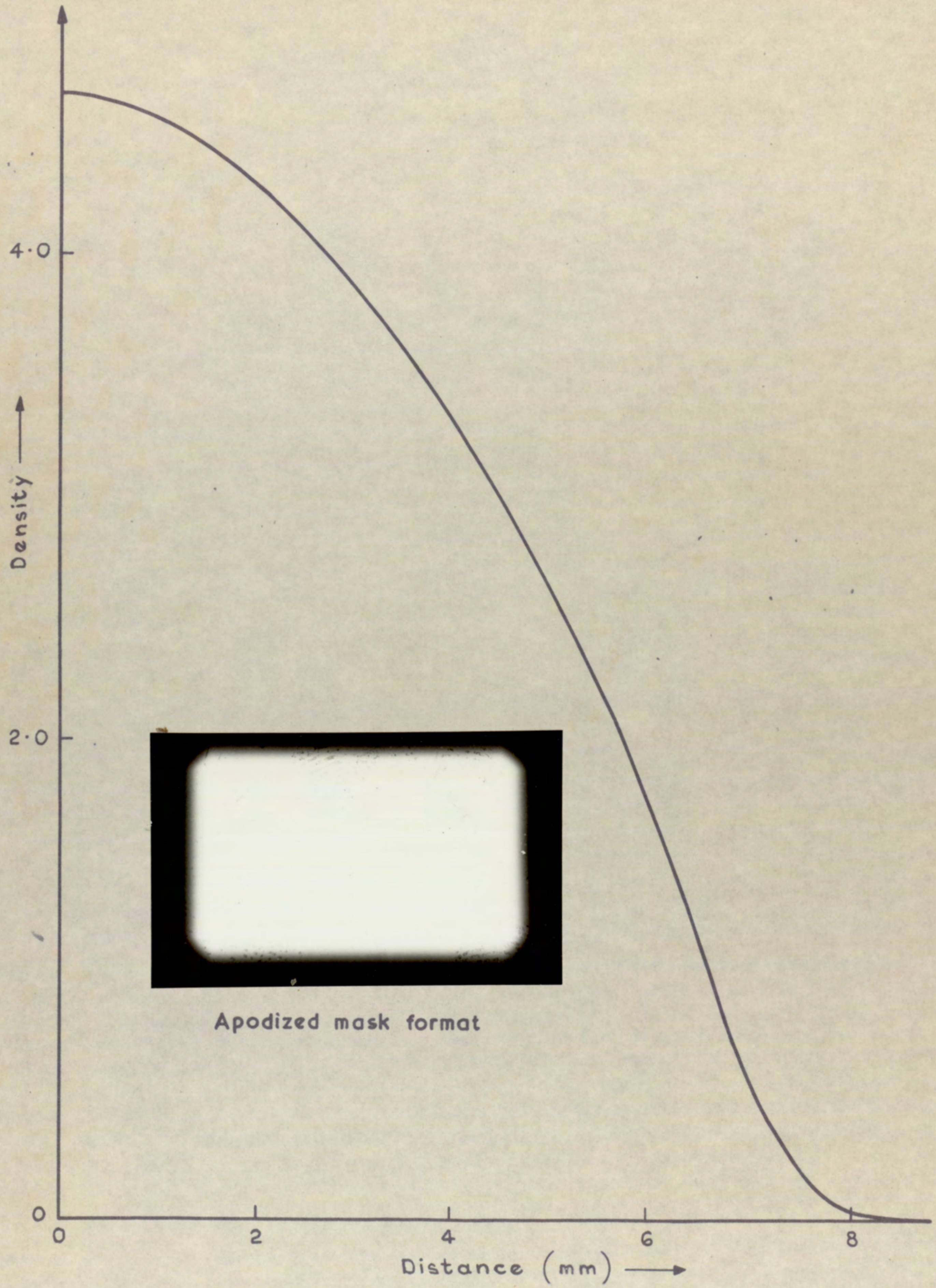
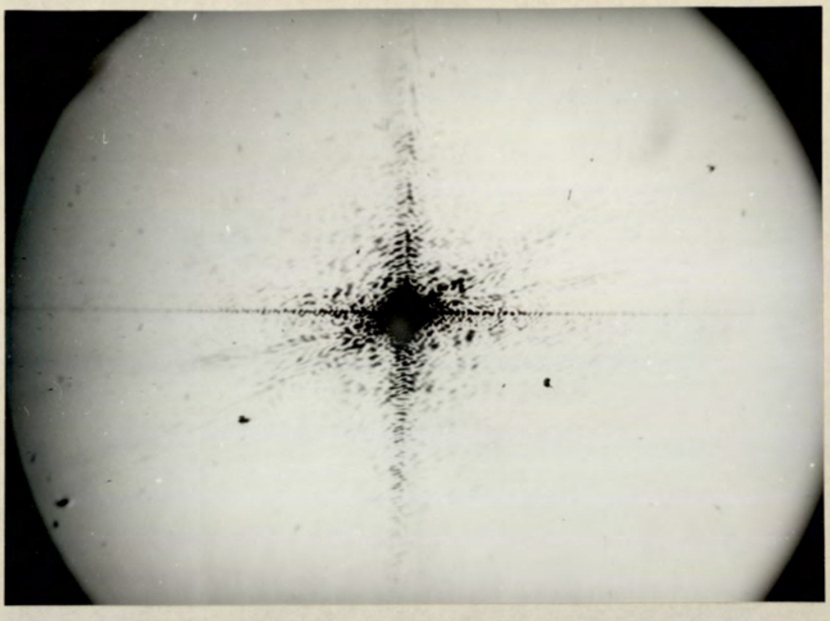
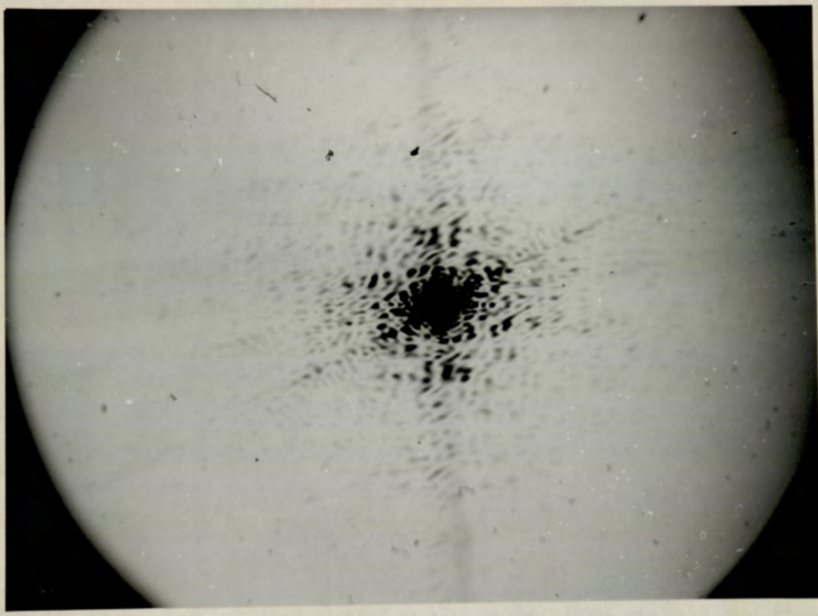


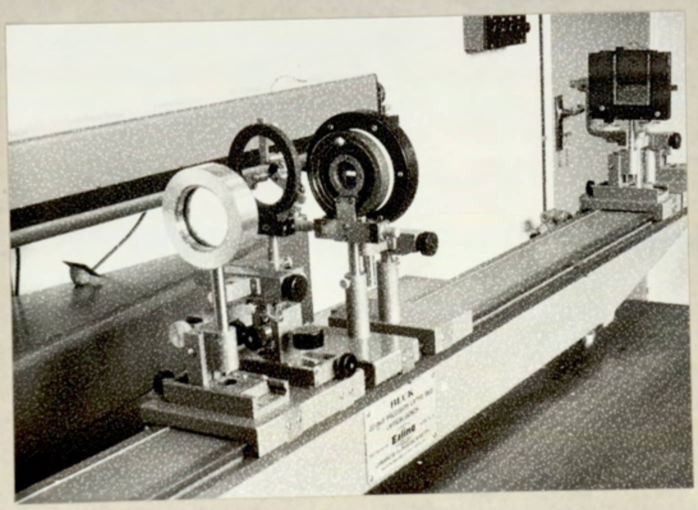
Fig. 6.14 Density variation for apodized mask



Hologram of un-apodized transparency



Hologram of apodized transparency



Original transparency used in holograms above

Fig. 6.15 Holograms with apodized masks.

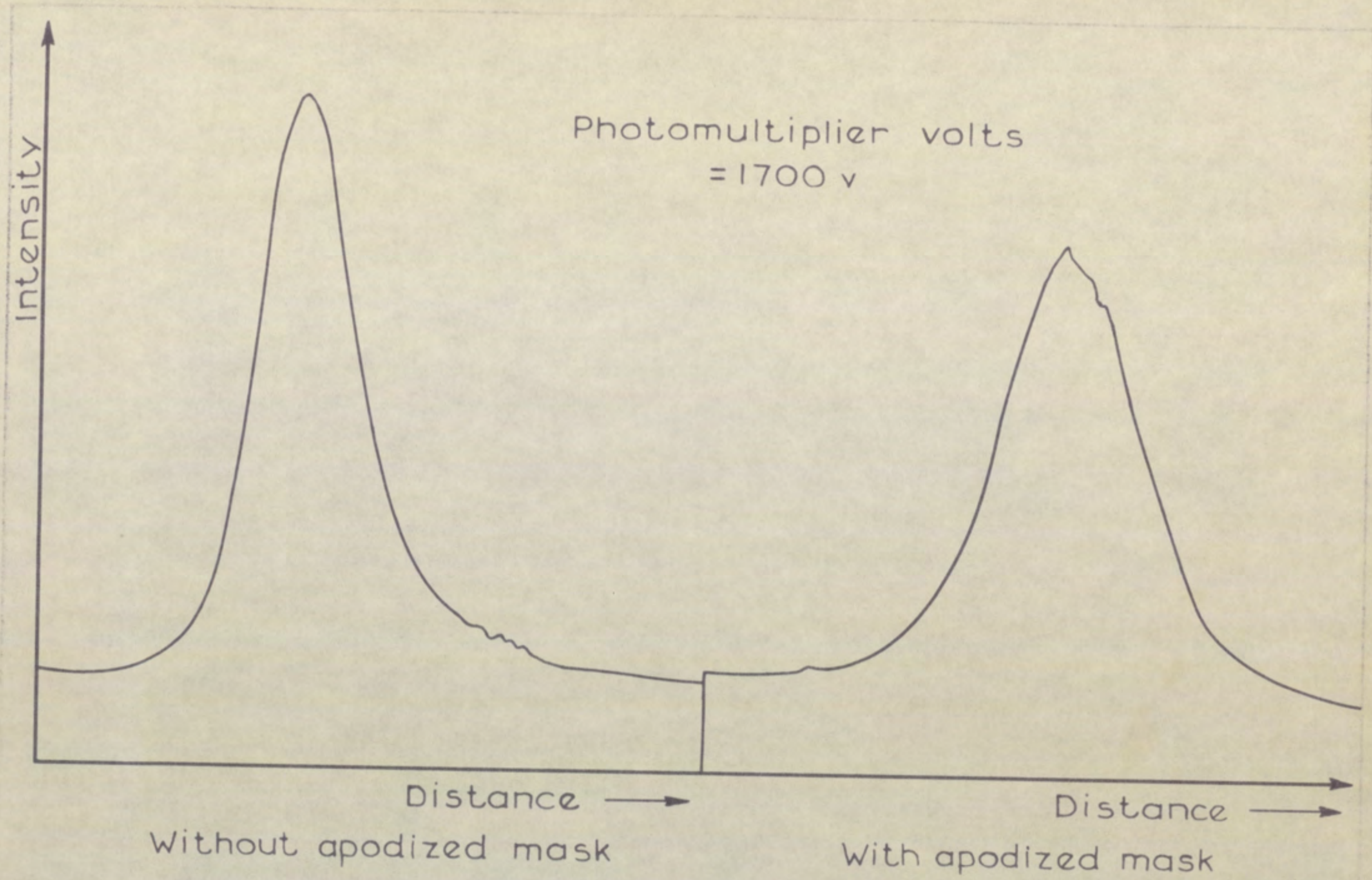
The actual variation of density across a typical apodized edge is shown in Fig 6.14, and inset, the complete mask. The actual variation of density is somewhat arbitrary; the slower the rate of change of density, the smaller the diffraction in its transform. If the masks are recorded as photographic transparencies it is necessary to use them in a liquid cell so that thickness variations in the film or emulsion are absorbed. It is possible to prepare a transparency containing the information which is to be recorded on the hologram, and to incorporate an apodized mask with the same transparency. This method has the advantage of simplicity, and reduces the number of glass surfaces in the system which could cause light scattering and degrade the transform.

In some cases it is desirable to have the transparency and mask separate, and for the latter to be capable of being manipulated into any given position. This will allow different areas of the transparency to be precisely selected for making a hologram. The mask must still be contained within a refractive index matching cell, but since the mask does not have to be changed in the cell, there is no need for the elaborate liquid gate described in Chapter 4. Instead, the mask may be sandwiched between two optically flat pieces of glass, using Canada Balsam cement or "SIRA" mountant between the mask and the glass to absorb thickness variations. Examples of the use of both of these techniques will be shown later.

6.2.3 Preparation of holograms from transparency inputs

Holograms of transparency inputs were made in the same way as those described in Chapter 5. To examine the effect of the apodized mask two holograms were prepared, one using an apodized mask, and the other without.

The holograms themselves are shown in Fig 6.15 together with the object transparency. The effect of the apodized mask in reducing the unwanted diffraction pattern of the rectangular aperture can be seen. Note that when these holograms were made the path lengths of the hologram interferometer were inadvertently made unequal. As a result, the carrier



Autocorrelation scans of holograms of transparencies

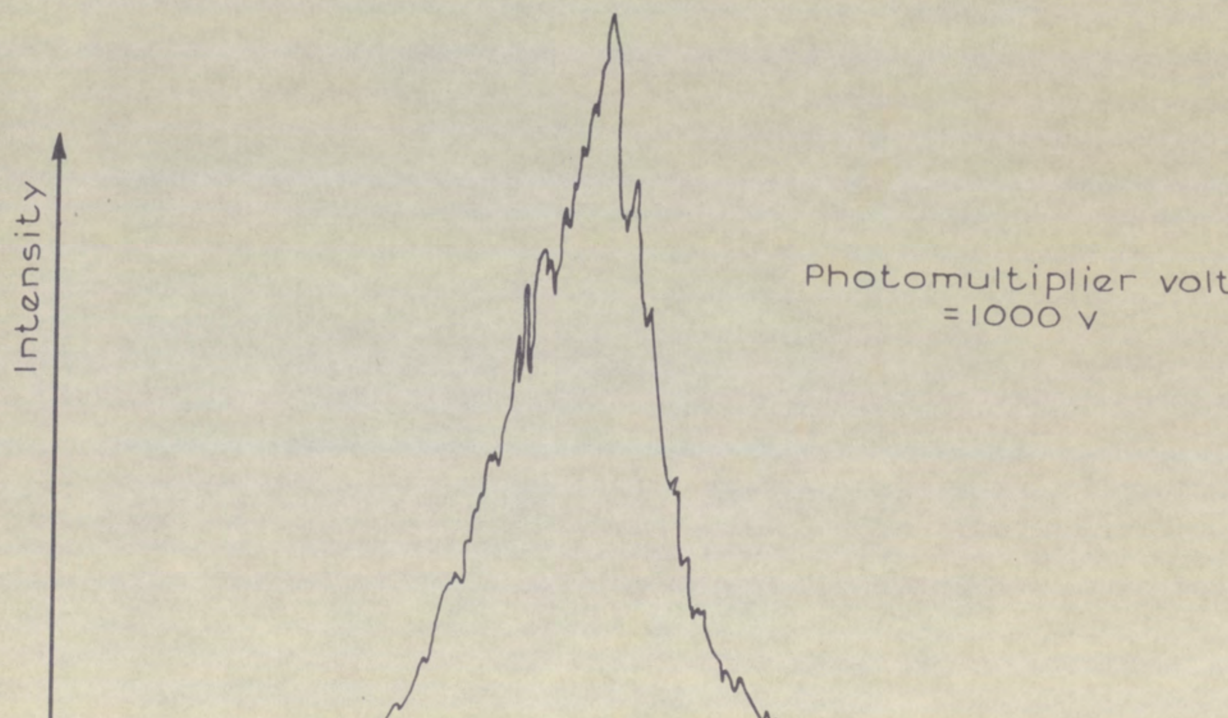
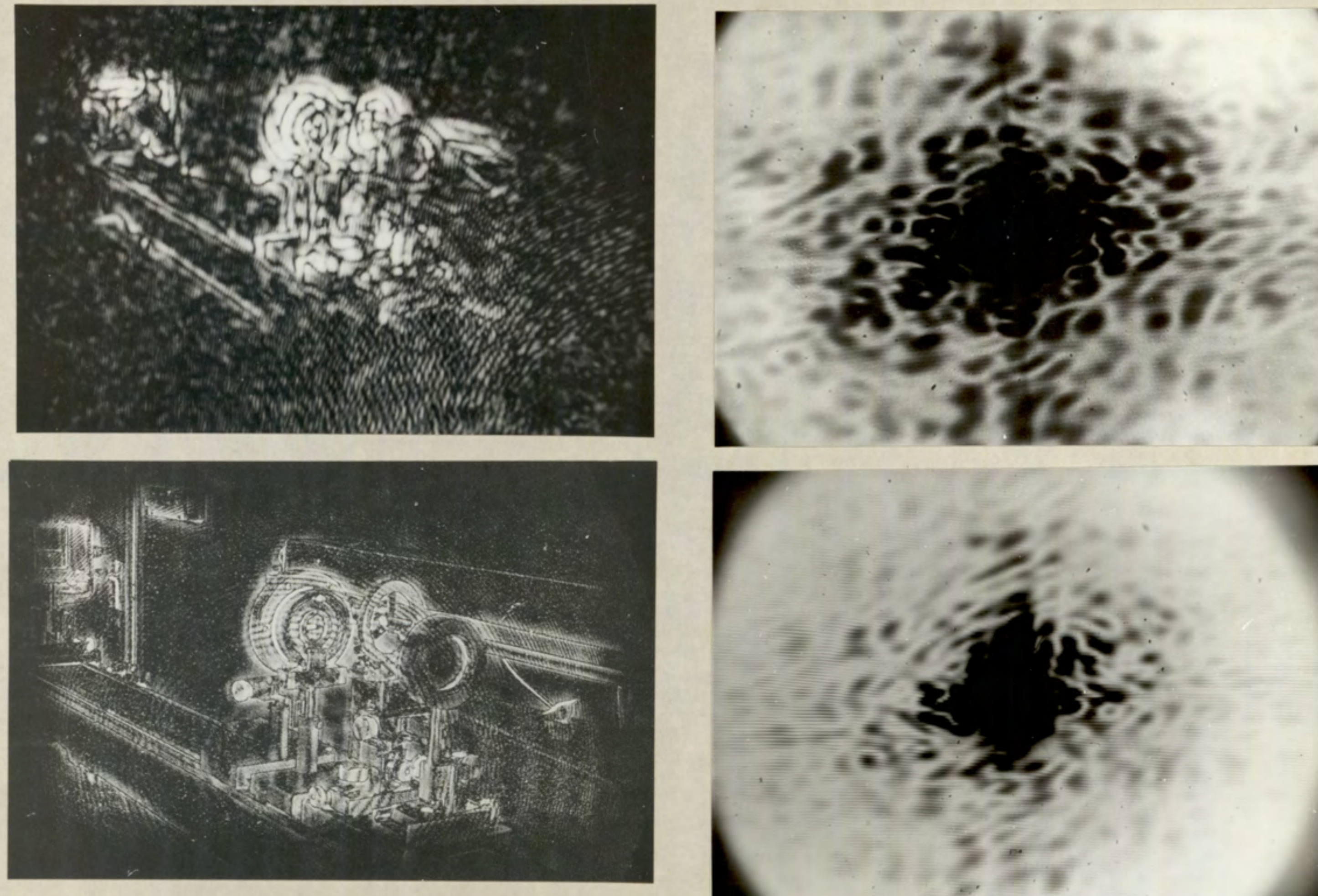


Fig. 6.16 Autocorrelation function of apodized transparency hologram with maximum contrast carrier fringes

Hologram with low fringe contrast [Unequal interferometer path length]

Hologram with good fringe contrast [Equal interferometer path length]



Reconstructed image

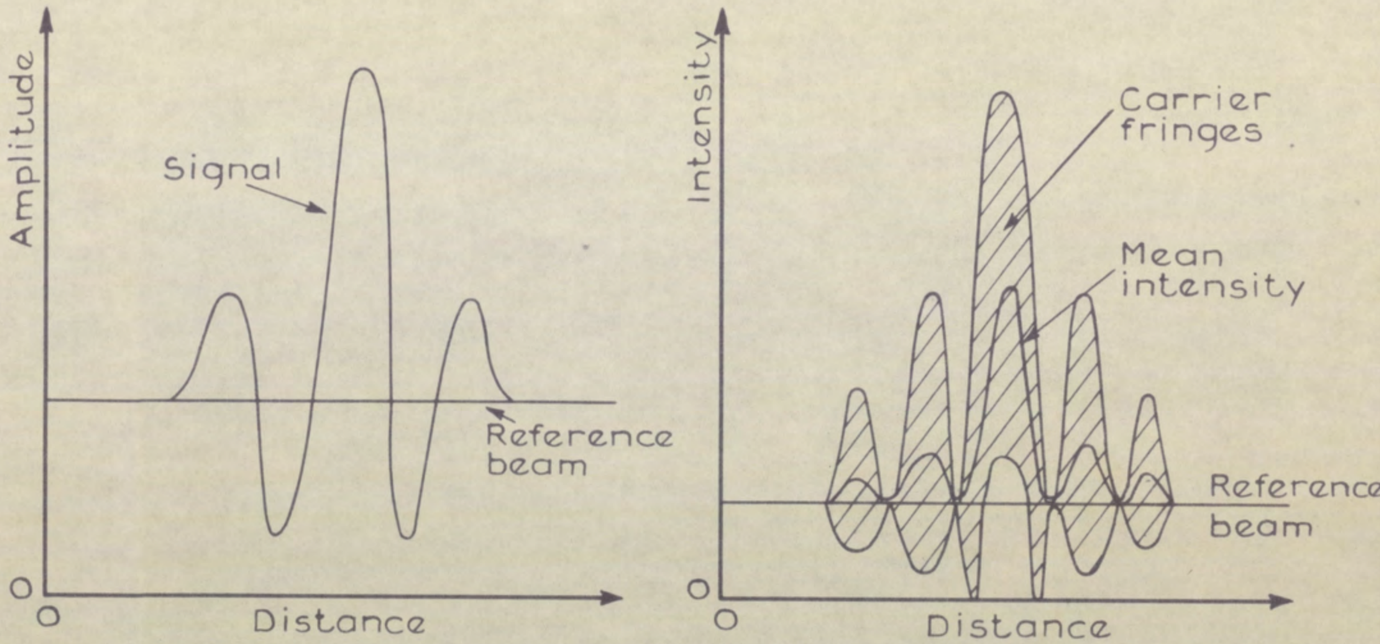
Appearance of hologram (x 100)

Fig. 6.17 Effect of path length and fringe contrast fringes are not as contrasty as might be expected; the effect of this will be discussed later. Both holograms could be used as a filter; the autocorrelation functions were measured and are shown in Fig 6.16. The peak for the unapodized transparency is higher and sharper, this is caused by correlation between the rectangular aperture and the hologram of its diffraction pattern. When the pattern is reduced by blurring the edges, a broader autocorrelation peak is produced. Both holograms were overexposed in the centre of the transform, and the autocorrelation functions were formed mainly by high spatial frequencies, and were therefore relatively sharp considering the dimensions of the object.

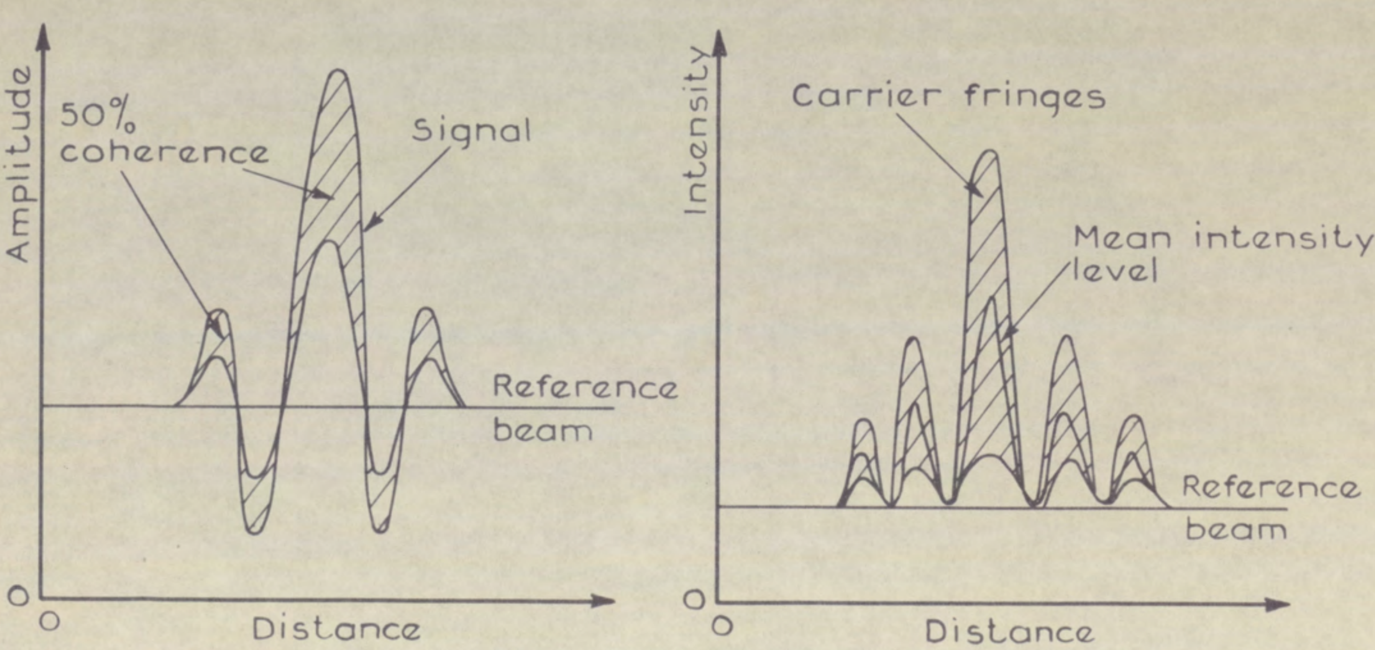
Another hologram was prepared having this time ensured that the interferometer had equal path lengths. The object was one of the apodized transparencies used earlier. The autocorrelation function was measured and is also shown on Fig 6.16; the shape is roughly similar to the previous result using an apodized mask, but because the contrast of the carrier fringes is very much improved the overall intensity of the autocorrelation response has increased. This is indicated by the fact that a lower photo-multiplier voltage was used to obtain a galvanometer deflection of the same order as those obtained previously.

The effect of improving hologram carrier fringe contrast not only increases the intensity of the reconstruction; it also increases the bandwidth of spatial frequencies recorded. This may be seen by reference to Fig 6.17, showing reconstructions of the original transparency for the two holograms discussed earlier.

The reconstruction from the hologram with poor contrast carrier fringes shows a lack of fine detail, only the broad outlines of the object transparency being formed. There is also a lack of half tone detail, indicating that low spatial frequencies are also unrecorded. Both fine



Hologram fringes with almost 100% coherence between beams



Hologram fringes with 50% coherence between beams

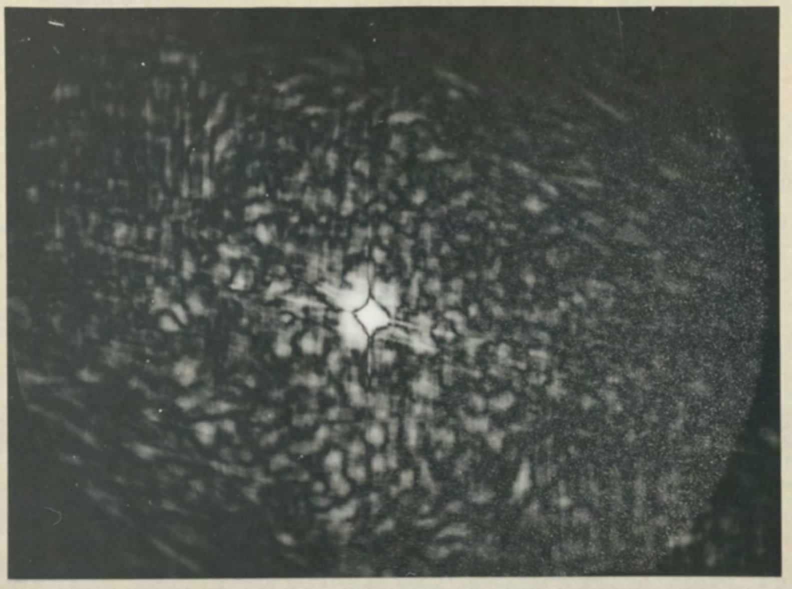
Fig. 6.18

detail and half tones are present in the reconstruction from a hologram with good contrast fringes, the difference being very pronounced. Fig 6.17 also shows microphotographs of the same holograms at about X120 magnification. There is a considerable difference in contrast of the carrier fringes, the over-exposure of the hologram is roughly equal in both cases.

The mechanism of hologram carrier fringe formation when the object and reference beams are not completely coherent may be seen by reference to Fig 6.18. Supposing that there is only 50% coherence between object and reference beams; then only half the signal amplitude will interfere with the reference bean. This has two effects. First, the mean intensity level of the interference pattern fluctuates more than in the complete coherence case because 50% of the signal energy is simply added to the reference beam energy. Secondly, the contrast of the fringes will be reduced because only half the signal amplitude is available for interference.

It was shown in Chapter 5 how the variation of the mean intensity level causes the interference fringes to be non-linearly recorded on the fourier transform hologram, and it follows that if the variation in mean intensity is increased, then the recording will become more distorted. Thus the remaining interference fringes will be subject to a disproportionally higher amount of nonlinear recording. The result is that one obtains rather poorer reconstructions and performance than might perhaps be expected.

The reduction in hologram bandwidth follows on two points; the low intensity, high frequency carrier fringes have an even smaller amount of signal energy available for interference and are limited by the sensitivity of the emulsion. In the low frequency fringes, where intensities are relatively high, the reduction of signal energy is not so important as the increase in mean intensity, which causes these fringes to be severely distorted by the photographic recording process.



Autocorrelation response of hologram with good fringe contrast

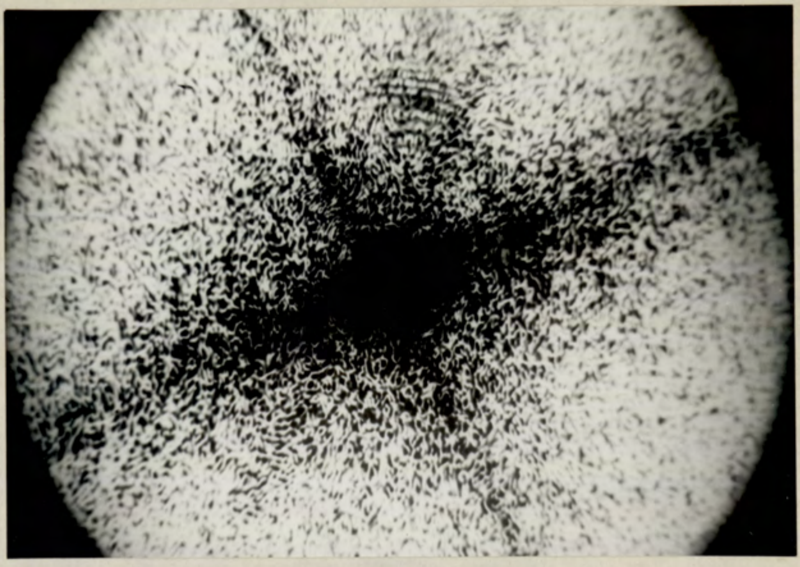
Autoconvolution response

Fig. 6.19 Correlation and convolution functions of hologram shown in Fig. 6.17 (Hologram with good fringe contrast)

The appearance of the autocorrelation function of the hologram with good contrast fringes is shown in Fig 6.19. The function is very small compared with the dimensions of the transparency, this is caused by over-exposure of the hologram in low spatial frequencies. The function has a definite shape; there are two lines, one vertical and one 75° to this, and the ACF is centred at the junction of these lines with a diamond shape. The lines occur because of the strong vertical and diagonal lines of the object transparency (see Fig 6.15), and therefore there is good correlation along these lines. The ACF is surrounded by a dark band with light patches outside this band, in a similar manner to some results observed in holograms of a rectangular aperture. There does not seem to be any evidence of an autocorrelation pattern corresponding to correlation between the rectangular aperture and the hologram. This suggests that the apodized mask has removed a sufficient number of high spatial frequencies from the aperture, and those frequencies which remain are masked by the transform of the transparency itself, or unrecorded due to overexposure of the photographic emulsion.

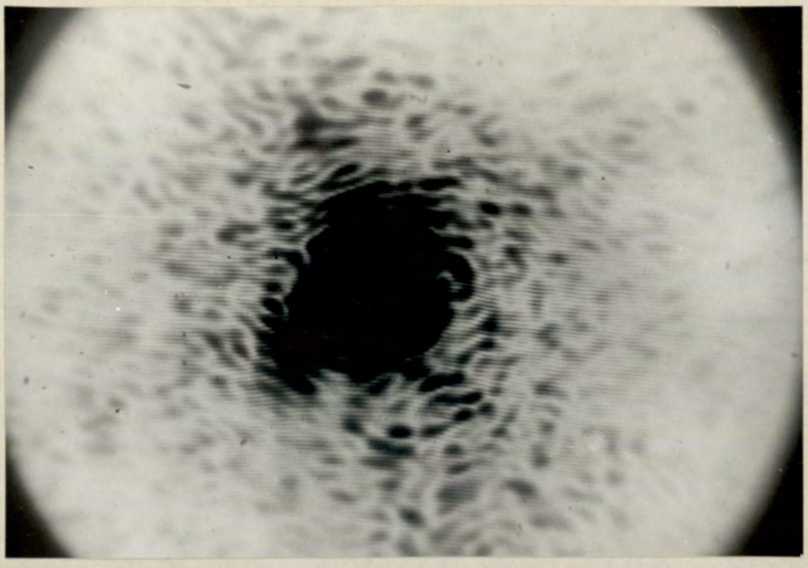
There does not appear to be any indication of non linearities in the ACF of this hologram, other than the fact that the ACF is small; there are no false images, for example. The background of the ACF consists of light and dark patches of roughly equal size; much of this seems to be caused by scatter from the hologram, the size of the patches being determined by the size of the lobes in the fourier transform. These lobes may be seen in the microphotograph of Fig 6.17; their origin and importance will be discussed later. It is of interest to note that the light patches appear in the position of the autoconvolution function, although of course, the autoconvolution itself contains no significant features.

Air photo. Transparency with apodized border



Fourier transform of air photo

Fig. 6. 20 Air photo. Transparency and fourier transform



Microphotograph (x 100) of air photo hologram

Reconstructed transform from above hologram

Fig. 6.21 Microphotograph and reconstructed transform of air photo transparency

6.3 Experiments with air photograph input transparencies

6.3.1 Effect of nonlinear photographic recording of the hologram

Using a photographic transparency as an object, attempts were made to examine the effect of nonlinear recording of the hologram on the auto-convolution response and reconstructed image. The object in these experiments was an air photograph, printed onto a glass slide together with an apodized mask. It is shown in Fig 6.20. The transparency contains a large number of easily recognisable features such as roads and houses of which nearly all are at right angles and form a distinctive fourier transform. This is shown in the lower half of Fig 6.20, where the transform is reproduced with approximately X35 magnification. This transform was recorded in the reconstructed transform plane, ie: in plane P4, (see Fig 4.8, Chapter 4) rather than in the hologram (P2) plane, since it was to be compared with the reconstructed transform from a hologram.

The hologram itself was exposed in the usual way; a one second exposure with standard development of 2 minutes in dilute (1 to 1) D19 developer. A microphotograph of the hologram is shown in Fig 6.21, it can be seen that good contrast carrier fringes were recorded in all parts of the transform except the centre, where the hologram was overexposed. The reconstructed transform is shown in Fig 6.21 also, it can be compared with Fig 6.20, the main difference being the absence of any record in the centre of the transform, due to overexposure. Otherwise the two transforms are very similar, especially in the appearance of a cross-like diffraction pattern and in the overall distribution of energy. It was not possible to correlate individual lobes in each pattern, and it was assumed that their appearance was uniquely determined in each exposure by small variations in hologram position, recording plate position and slight variation in optical component position.

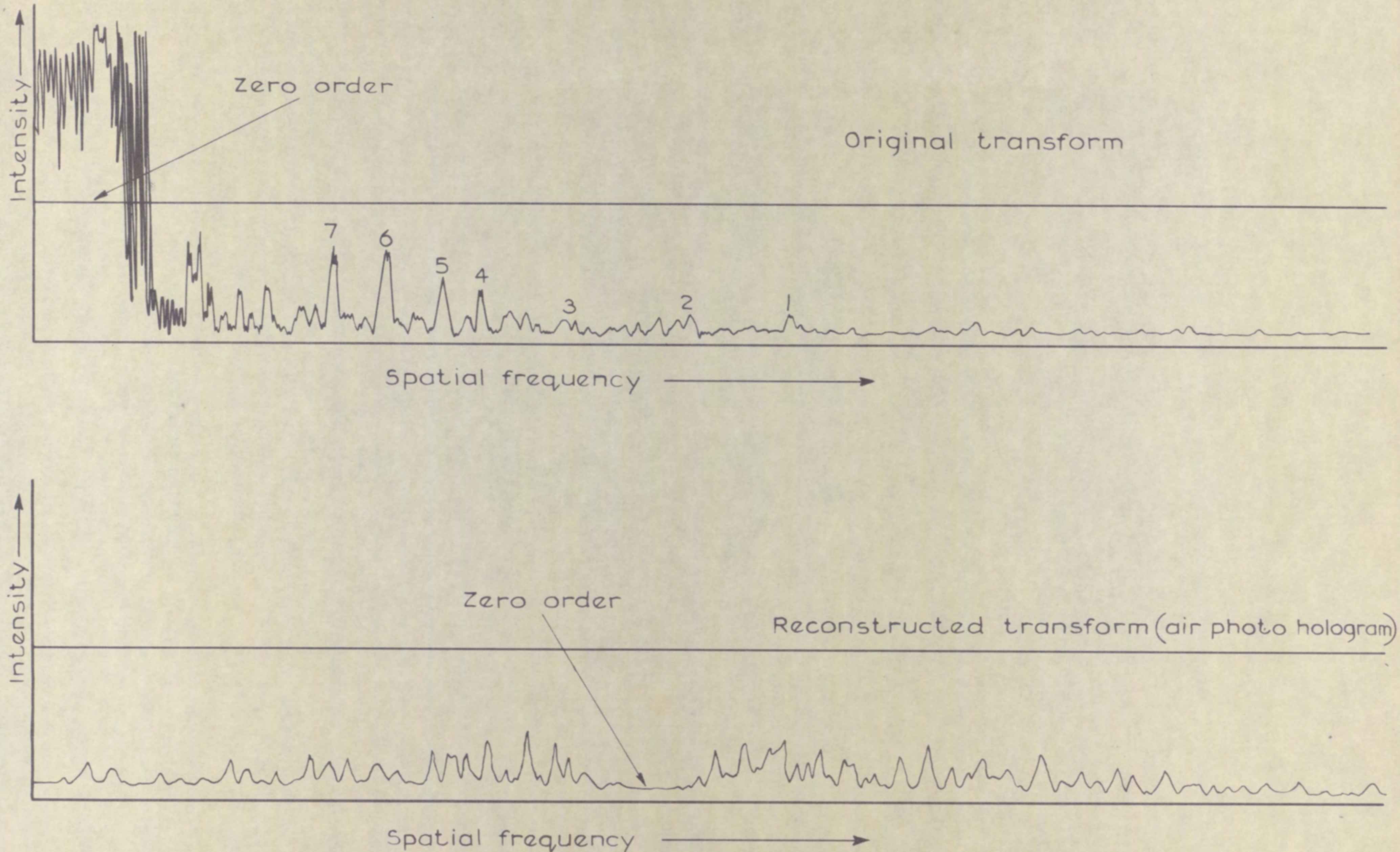

Intensity
Zero order
Original transform
7
6
5
4
3
2
1
Spatial frequency
Intensity
Zero order
Reconstructed transform (air photo hologram)
Spatial frequency

Fig. 6.22

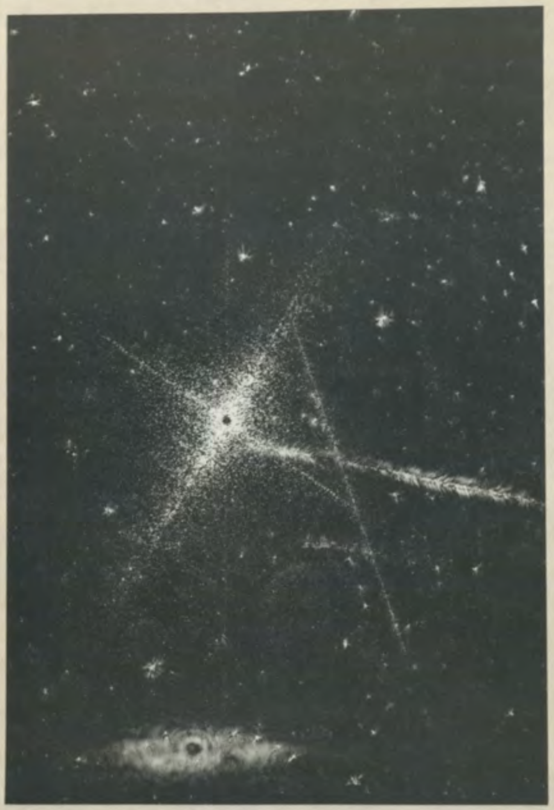
The extent of the unrecorded area of the reconstructed transform was measured as being approximately 0.2mm in diameter. This gave a limit for the lowest spatial frequency recorded, of approximately 0.24 cycles/mm. This may not seem significant when the upper limit is perhaps of the order of 20 cycles/mm but there is a relatively high proportion of energy contained within this area, and both the reconstruction and the autocorrelation function will be weak compared with the total signal energy. The significance of the low frequency cut-off will depend on whether there low frequencies are being searched for in the signal or pattern to be recognised.

The intensity distribution in the reconstructed transform was compared with that of the original transform to examine the possibility of distortion in the low spatial frequencies. The two distributions are shown in Fig 6.22. The reduction in intensity of the low spatial frequencies is very pronounced near zero frequency. Although it was not possible to correlate individual peaks of intensity between the two distributions, it can be seen that there is a reduction in the intensity maxima, indicated by the broken curve. The effect of nonlinear photographic recording seems to compress the recorded intensities into a narrow intensity band.

There seems to be no evidence of splitting of the diffracted lobes in the sense reported earlier, in Chapter 5. The reason for this is that any distortion of the lobes would be difficult to detect in a random distribution of intensity peaks such as those recorded here. Where the distribution of intensity is in a regular pattern, as for a rectangular aperture, then distortion or splitting of the side lobes can be easily seen.

6.3.2 Properties of holograms made with transparencies

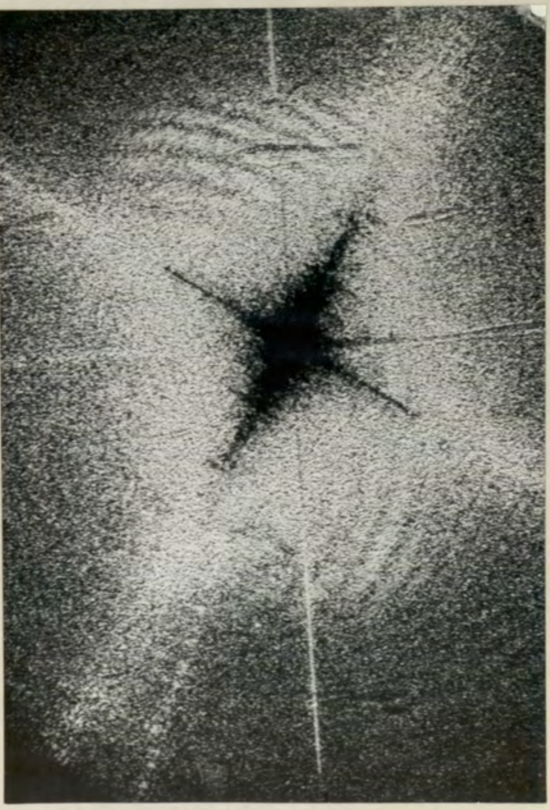
In this section the responses of four holograms are described, each hologram being made with the same transparency as that used in section 6.3.1, but having a range of object beam exposures, the reference beam being



Low object beam intensity (1 unit)

Object beam intensity = 10 unit

Object beam intensity 100 units



Object beam intensity 1000 unit

Fig. 6.23 Reconstructed transforms of air photo holograms at different intensities

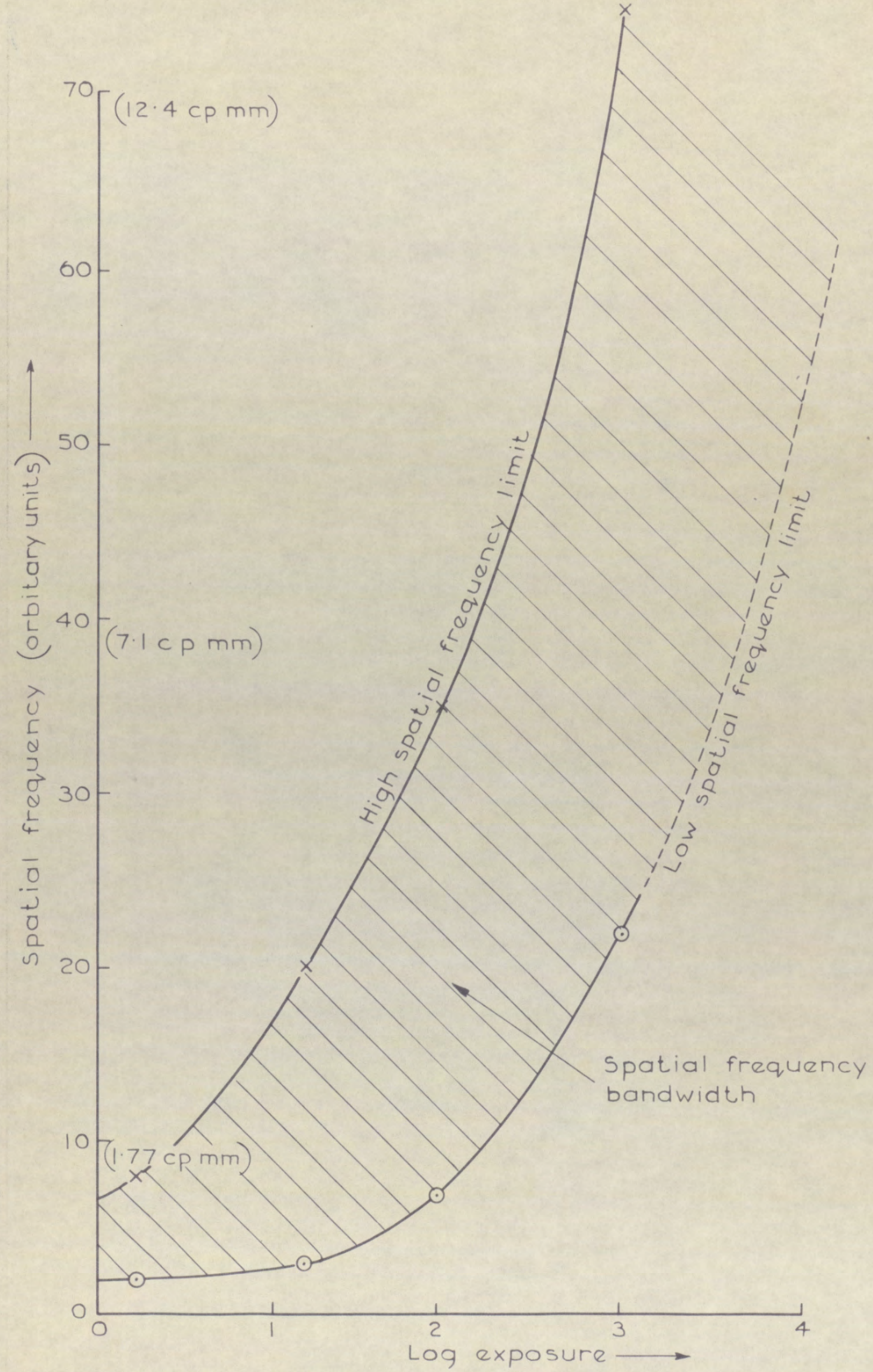
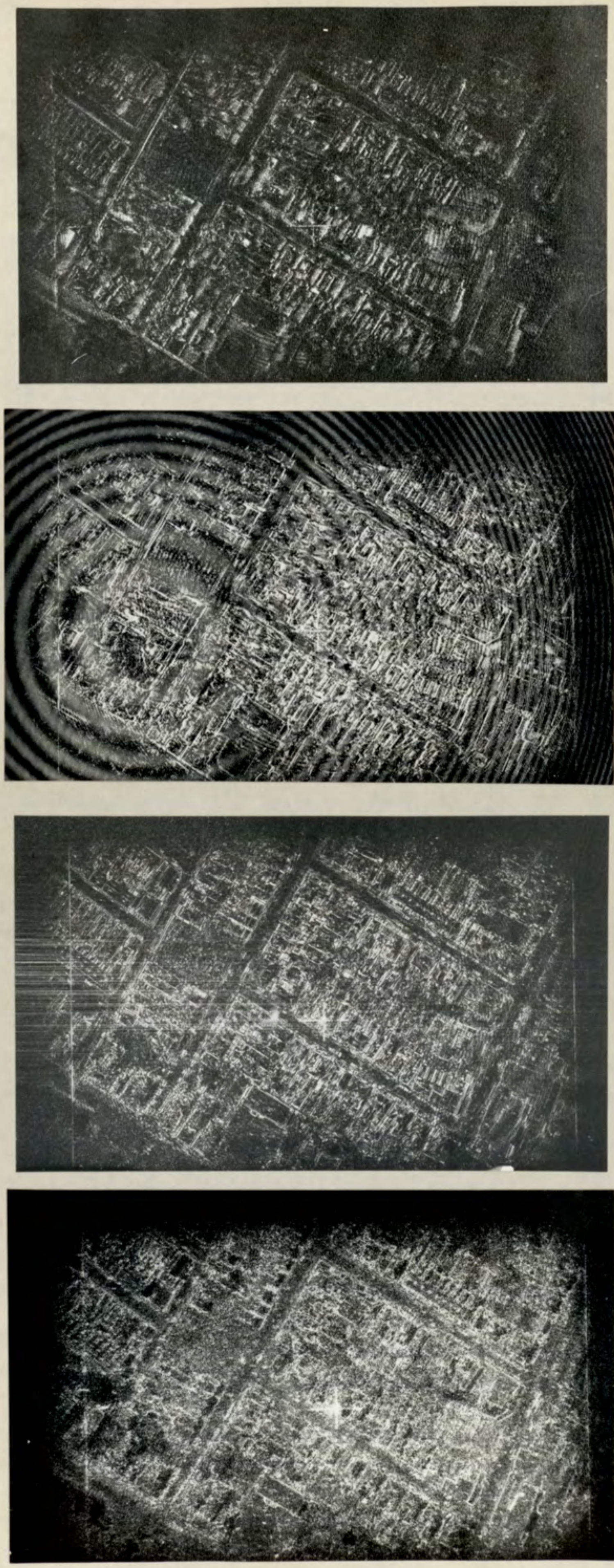


Fig. 6.24 Variation of bandwidth with exposure



Object beam exposure

I Unit

IO Units

IOO Units

IOOO Units

Fig. 6.25 Impulse response of air photo holograms with different object beam intensities constant in all four cases. The various responses are compared below:-

(i) Reconstructed transform (Fig 6.23)

The change in the reconstructed transform with object exposure may be seen in two effects. The first is in the extent of high spatial frequency recording, and the second is in the limit of lowest spatial frequency recording. It is interesting to note that there is little change in the amount of unrecorded transform for low exposures, but when the exposure is high enough to record much of the higher spatial frequencies, then the area of unrecorded transform increases considerably. These illustrations also show the concept of spatial frequency bandwidth, and how the mean spatial frequency and the bandwidth itself vary with exposure. These relationships are shown on Fig 6.24, from which it appears that high bandwidths are obtained with high exposures. If fairly high frequencies (ie: above 20 cp mm) are important, it is clear that a considerable proportion of the lower frequencies would be unrecorded. It is estimated that the maximum spatial frequency bandwidth recorded by the Kodak 649F emulsion, for this type of input, is of the order of 10-15 cycles per mm.

(ii) Impulse response - reconstructed images (Fig 6.25)

The main features (ie: roads and houses) of the object transparency can be seen in all four images, but there is a considerable difference in their appearance. The first (ie: lowest exposure) shows a moderate amount of fine detail, and some half tones are present. There is a considerable difference in detail between this and the next highest exposure in which fine details of houses, plot divisions, etc, can be seen. The half tones are absent, and in some cases the absence of low spatial frequencies has caused the edges of some features to become distorted in a similar way to that observed in the case of a slit object. In the next highest exposure the reconstruction still contains fine detail, but this is not so clear as in the previous illustration, there being more distortion

1 Unit 10 Units

Object beam intensity

100 Units 1000 Units

Object beam intensity

Fig. 6.26 Autocorrelation response of air photo holograms along straight edges and some grain noise has been recorded. The final illustration shows an image which is almost completely swamped by noise, although some features are still recognisable.

It would seem that the selection of optimum exposure must consider the presence of diffraction due to grain noise, since this noise will tend to correlate with itself irrespective of data on the transparency. The reconstructions suggested that though the lower spatial frequencies were absent and there was some distortion of straight edges in the pattern, the overall appearance of the image was unchanged and undistorted.

(iii) Autocorrelation response (Fig 6.26)

Again, there is a marked variation in shape of these functions. The low exposure produces a function (ACF) which is diamond shaped, with axes at right angles and in line with the main features of the transparency. The function itself is broad and shows no fine structure compared with the next highest exposure. The axes of the previous function are stronger and there is some indication of nonlinear recording in the appearance of patches of light outside a dark band which surrounds the function. The next highest exposure contains finer detail; there is some evidence of strong features in between the two axes mentioned earlier; this seems to be caused by a road running at an angle of approximately 30° to other roads in the transparency. The final illustration shows a marked difference; the axes of the previous functions have vanished and there is now a more symmetrical function. This ACF is caused mainly by the recording of grain noise in the hologram and since the latter is symmetrically distributed about the fourier transform, the ACF appears symmetrical about its centre.

It is clear from these results that some care must be exercised in the recording process so that only that information in the transparency is recorded which is required for recognition purposes. This is particularly

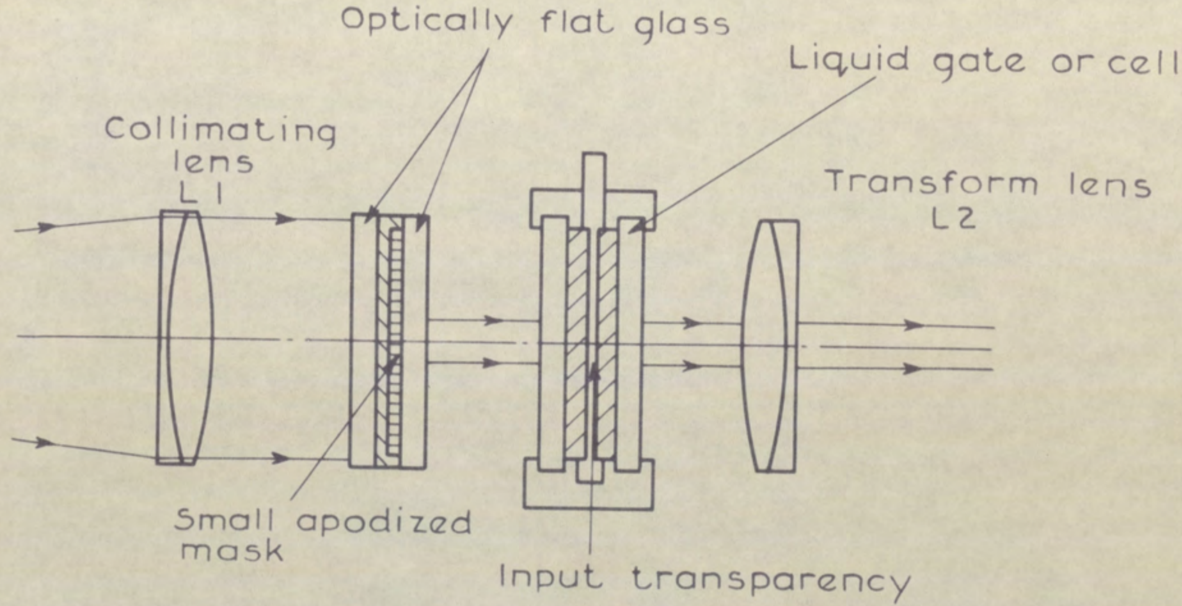
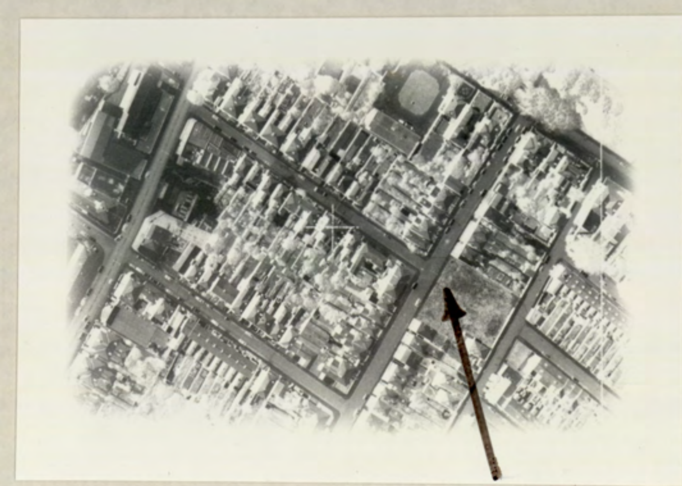


Fig. 6.27a Arrangement of apodized mask and input transparency



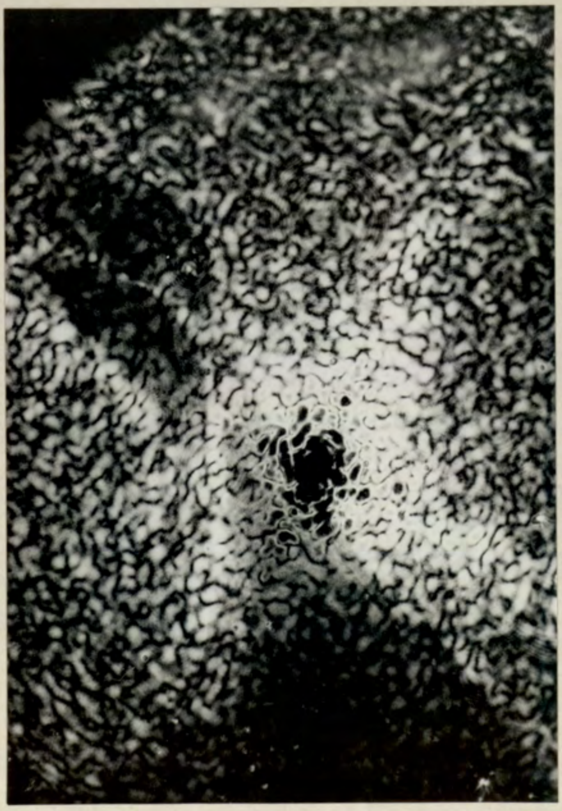
Target area

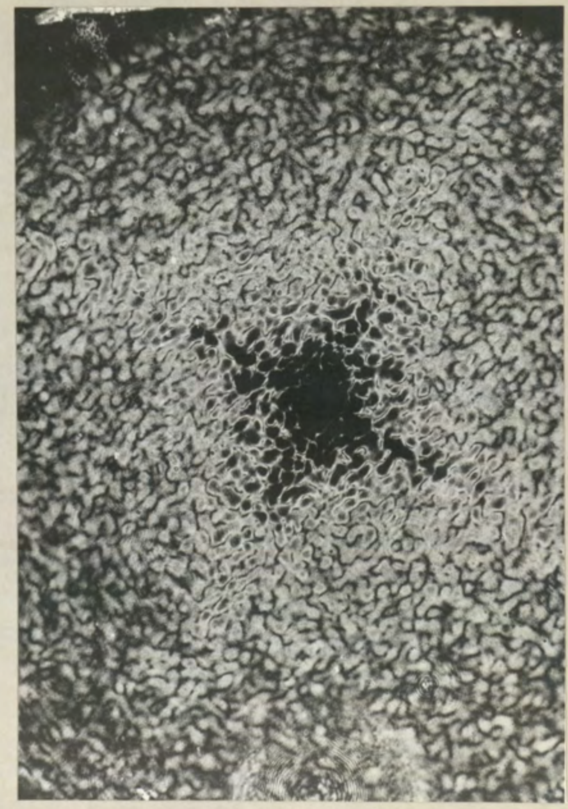
Fig. 6.27b Air photo transparency showing target area

I Unit

IO Units

Object beam intensity





IOO Units

IOOO Units

Object beam intensity

Fig. 6.28 Reconstructed transforms of air photo holograms with different intensities and small target area object

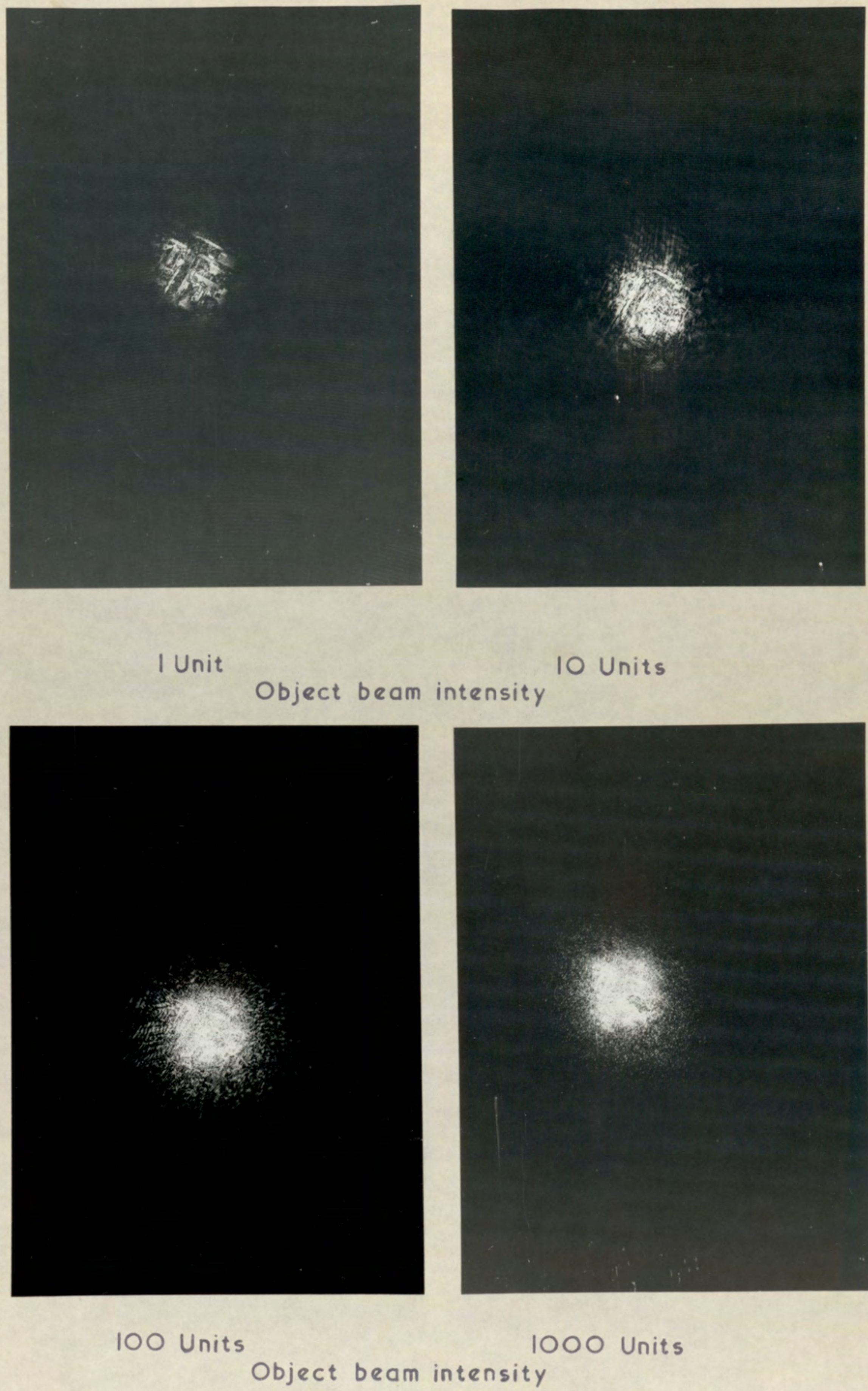


Fig. 6.29 Reconstructions from air photo (target) holograms

important where high spatial frequencies are present in the pattern or signal to be sought; it may be necessary to use an input transparency having very fine grain in some cases.

6.3.3 Response of holograms made with a selected area of transparency

In these tests, the holograms were given a range of object exposures, as before, but the object transparency was selected from the master transparency by using a small apodized mask. The aim was to examine the effect of correlating a hologram of small area with the whole transparency, with particular reference to the influence of emulsion grain in the reduced input aperture. The arrangement of apodized mask and input transparency is shown in Fig 6.27.

(i) Reconstructed Transform (Fig 6.28)

The main difference between these results and those observed earlier is in the size of the lobes or spots in the transform. Here the lobes are larger and more distinct, corresponding with the change in "grain" size when speckle patterns are viewed through a small aperture. (The lobes are formed in a similar way as speckles are formed in a speckle pattern, ie: by interference between light from different elements of a diffusing screen, the screen in this case being the photographic transparency.)

The darkening of the centre of the lobes, leaving a bright fringe around the perimeter corresponds to the nonlinear recording effects observed earlier. It is not likely that their effect will be easily visible since the diffracted energy will be scattered in a random manner about the reconstructed image. The variation in bandwidth of the holograms is again apparent, being very similar to that observed earlier.

(ii) Reconstructed Images (Fig 6.29)

The variation in these images is again similar to those of the previous section, except that the grain noise seems to be more obstrusive.

1 Unit 10 Unit
Object beam intensity

100 Units 1000 Units
Object beam intensity

Fig. 6.30 Cross correlation between air photo transparency and target holograms (see Fig. 6.29)

In the magnified image of the high exposure image it can be seen that features of the original input are present, even though there is considerable grain noise and distortion.

(iii) Cross correlation response with whole of input transparency (Fig 6.30)

In these results, the holograms were correlated with the whole input transparency, ie: with the small apodized mask removed. Ideally, there will be correlation only with the small area of transparency used for making the hologram, the remainder of the correlation plane being dark.

In the first (low exposure) photograph, there is a diffuse patch of light indicating correlation with the "target" area, but there is also some correlation with other parts of the transparency. These correlations are relatively weak, but it is interesting to note how they occur at other road junctions and along the regular features of the transparency.

The additional correlations are reduced in size together with the size of the main correlation spot by increasing exposure, as shown in the adjoining photograph. There is still cross correlation with other areas of the transparency along lines of regular features, although these are more frequent, indicating that perhaps fine details such as corners of houses, etc, are correlated.

In the third illustration the correlation function has become elongated in the directions of a cross and resembles one of the auto-correlations of Fig 6.26. There is much background "noise", and it is difficult to decide if this is due to correlation with different parts of the transparency or if it is caused by scattered light from the hologram. As some of the "noise" shows faint regularities and alignment with regular features it is thought that the former suggestion is probably true. In the final illustration there is virtually no background noise, and the correlation spot is small and sharp. The correlation spot has no cross-like appearance, this apparently caused by the recording of grain noise rather than of transparency details.

Since the film grain is randomly distributed in the input transparency, there will be a point in the fourier spectrum at which the film grain energy is greater than that from the transparency itself, and recordings of higher spatial frequencies will be high in film grain noise content. This is very clear in Fig 6.27, where the high exposure hologram produces a 'grainy' reconstruction.

Since the noise is randomly distributed in the input, the auto-correlation function will be very sharp as in Fig 6.28. It will also be roughly circular in appearance, and will be very precisely located in the correlation plane, since no other part of the input will have the same grain structure as that recorded on the hologram.

It is evident that holograms made with exposures sufficiently high to record grain noise would not be of use, since they will only correlate with the transparency from which they were prepared. By giving the hologram a somewhat smaller exposure, most of the information recorded will be related to the pattern in the transparency, and such holograms should correlate equally well with similar patterns on other transparencies. This was demonstrated by experiment using another copy of the air photograph transparency used earlier. For holograms made with relatively low exposures, autocorrelation peaks were obtained when correlated with the copy, rather than the original.

6.3.4 Discussion concerning the use of photographic input media

It has been shown that the use of photographic input data produces a speckling effect on the fourier transform in much the same way in which a speckle pattern is produced by coherent light reflected or transmitted by a diffuser. The size of the lobes or speckles is inversely related to the size of the aperture containing the input transparency; the formation of the resultant transform may perhaps be best described by convolving the aperture transform with the transform of the transparency.

The formation of these lobes and the effect of film grain noise are directly related to each other. In the last section it was shown that high exposures could lead to excessive film grain noise being recorded, giving correlation only between the hologram and that particular distribution of grains in the emulsion. Comparison of the reconstructed transforms of high exposure holograms in Figs 6.21 and 6.26 show that the lobes in the latter figure are more obstrusive than those of the former, and that this corresponds with the "noisy" appearance of the image in Fig 6.27. In contrast, the reconstruction corresponding to the reconstructed transform of Fig 6.21, shown in Fig 6.23 is much clearer, although some granularity is present. It also follows that the latter hologram could be used successfully in correlation with patterns on other input transparencies and that a relatively small fraction of the autocorrelation energy is due to correlation between the hologram and the film grain structure. Thus, for a given film grain size in the transparency, best results would be obtained using the largest permissible aperture. Conversely, the grain size of the input data should be kept as small as possible. Ideally, the input data would be grainless, using, for example, photochromic film.

6.4 Attempts to reduce nonlinear recording of hologram

6.4.1 Introduction

In this section some experiments are described in which the nonlinear recording of the hologram has been reduced. The procedure adopted here was to reduce the intensity range of the fourier transform, rather than extend the dynamic range of the recording material. Although this meant that the

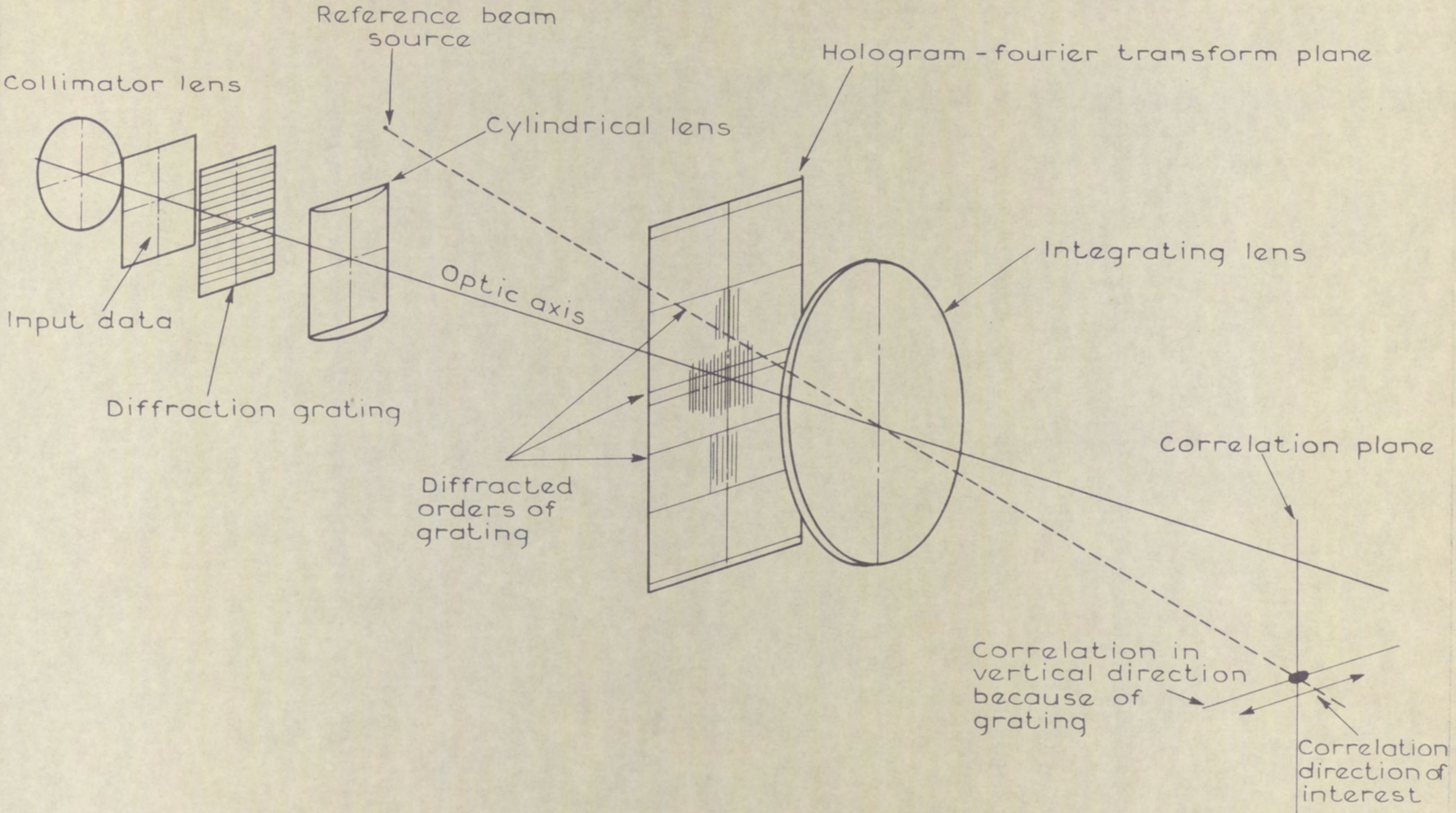

Reference beam source
Hologram - fourier transform plane
Collimator lens
Cylindrical lens
Integrating lens
Optic axis
Input data
Diffraction grating
Correlation plane
Diffracted orders of grating
Correlation in vertical direction because of grating
Correlation direction of interest

Fig. 6.31

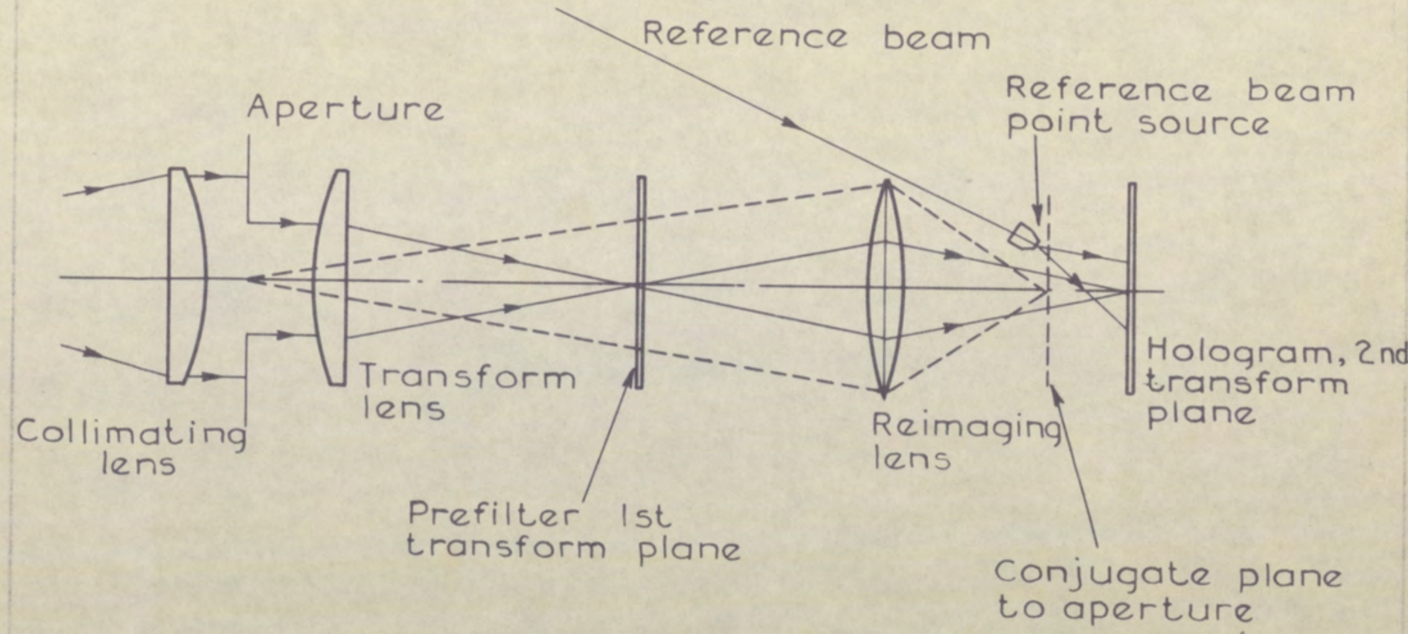


Fig. 6.32 a Use of prefilter with two dimensional transforms

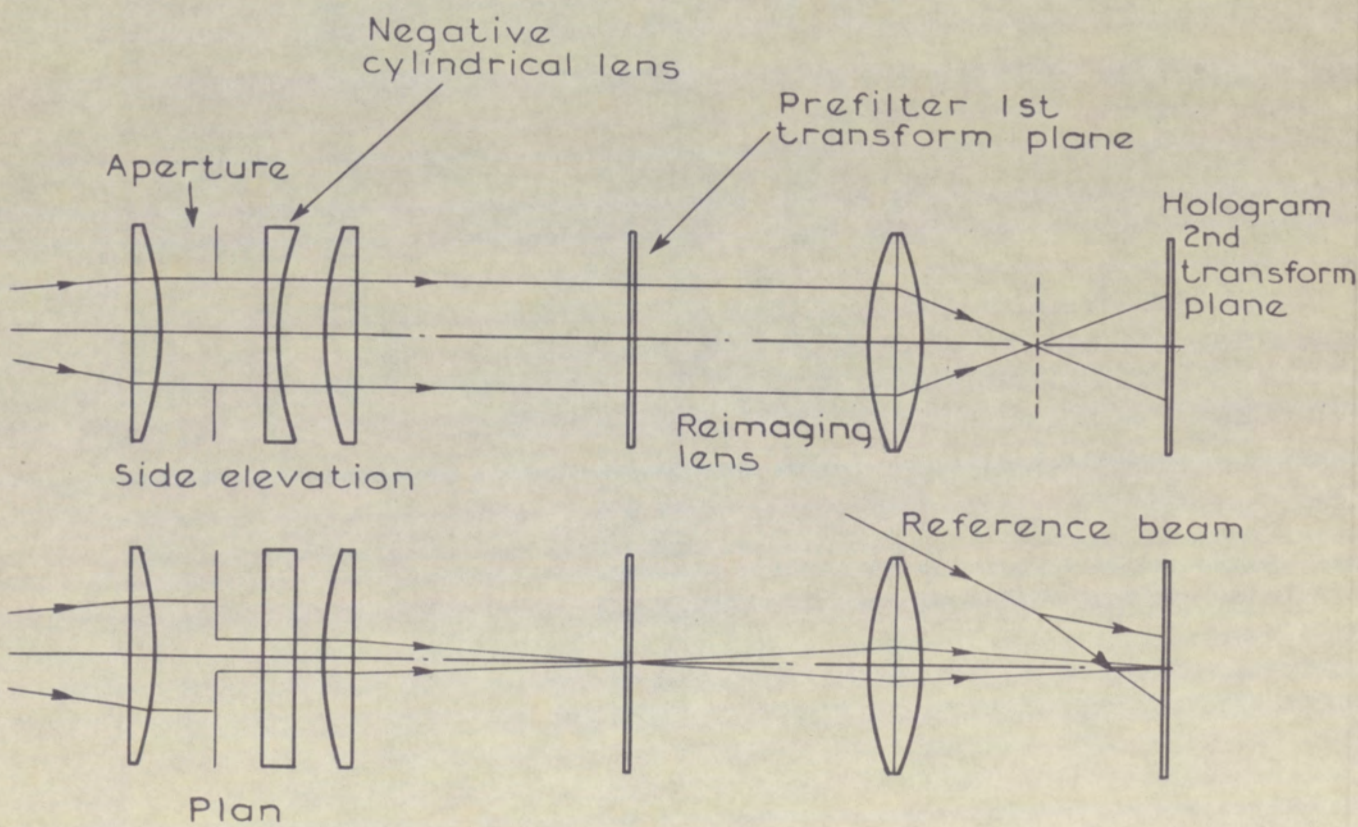


Fig. 6.32 b Use of prefilter with one dimensional transforms

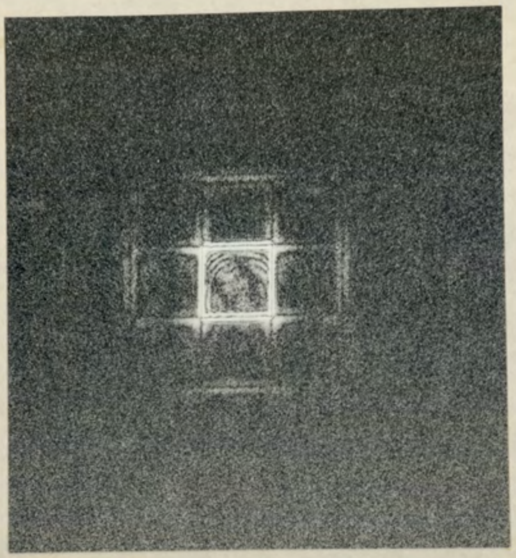
signal would be modified, the aim was to remove nonlinear terms from the autocorrelation and impulse response, rather than to obtain a perfect autocorrelation function or reconstructed image.

Several other methods were considered; for example, in single dimension fourier transform processing (ie: with a cylindrical lens), a diffraction grating could be used in series with the input transparency to form a range of transforms, each having different intensities, and thus enabling any single exposure time to record all parts of the transform. There will, of course, be correlation between the grating and the hologram, but this correlation can be arranged to be in the orthogonal direction to that in which the transform is recorded, and eliminates the disadvantage. This idea is shown in Fig 6.31. Although this idea may be useful for certain types of input data (eg: electronic data), it was not considered suitable for general two dimensional data processing, and the original intensity range reduction method was employed.

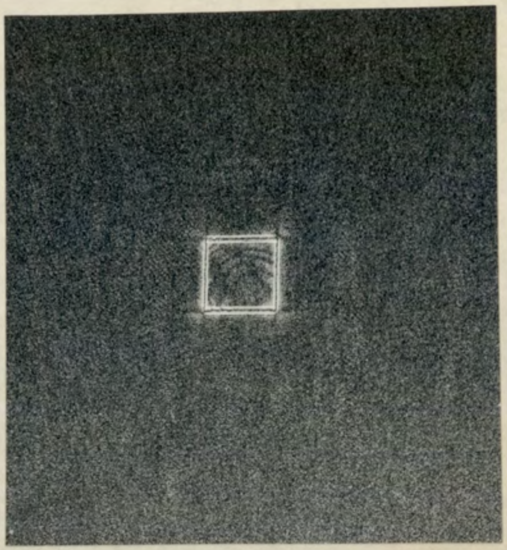
6.4.2 Experiments with two dimensional transforms - rectangular aperture

In these experiments the intensity reduction was accomplished by using a "prefilter" in the fourier transform plane, which absorbed the high intensities, whilst the hologram was recorded in a second fourier transform plane, imaged at 1:1 magnification by a simple doublet lens. The arrangement is shown in Fig 6.32(a). Note that the reference source must be positioned in the plane conjugate to the input plane in order that the hologram has no focal power. Apart from this, the hologram was exposed and processed in the usual way.

The prefilter was made by exposing a photographic plate (Kodak 649F) to the transform in the 1st transform plane, an exposure of 1/10th second being sufficient to record the zero and lower spatial frequencies. It was given a low contrast development, $\frac{1}{2}$ minute in D19 (1pt + 4pts water). The

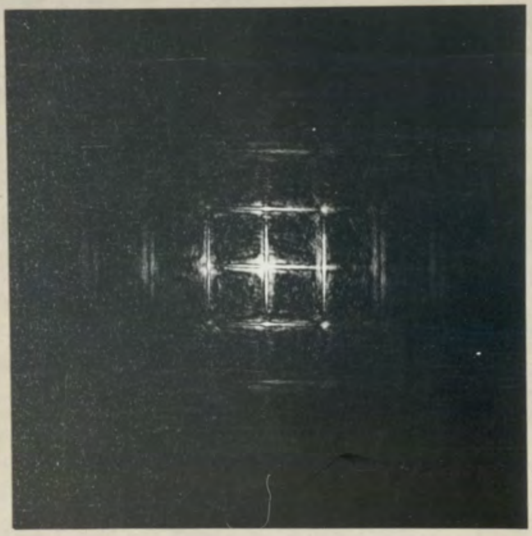


Unfiltered hologram

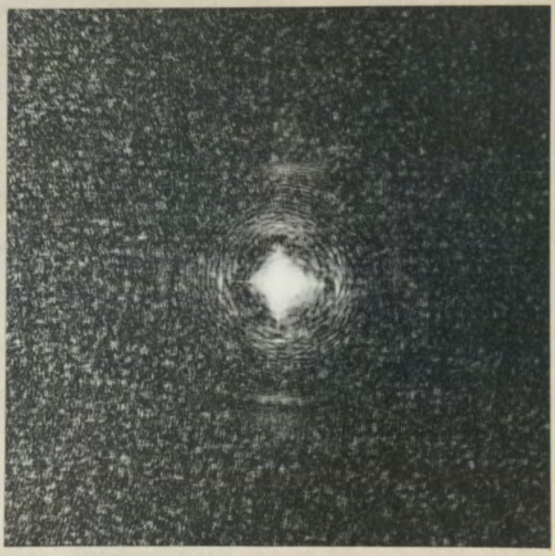


Hologram made with prefilter

Fig. 6.33 Reconstructions of holograms showing effect of prefilter

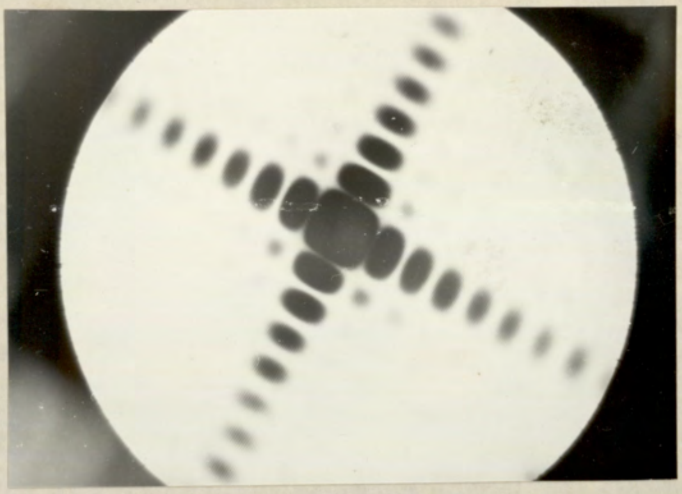


Unfiltered hologram

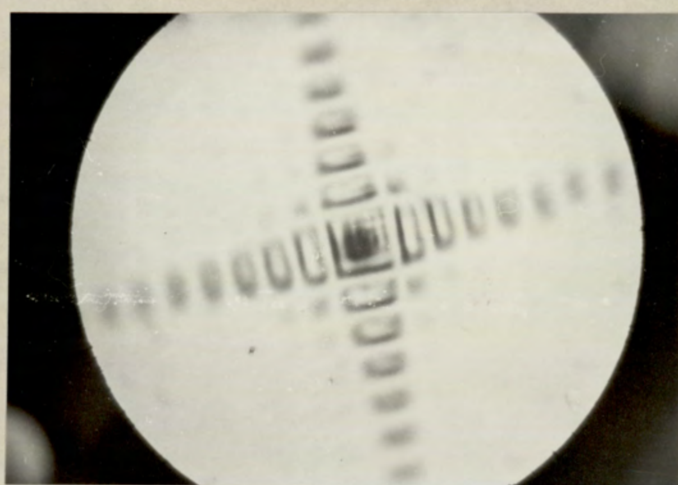


Hologram made with prefilter

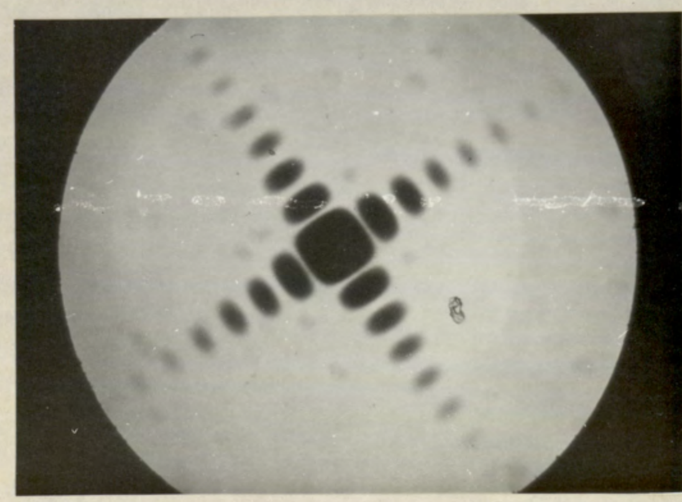
Fig. 6.35 Autocorrelation of holograms showing effect of prefilter



Normal hologram, made without prefilter



Hologram made with prefilter



Prefilter

Fig. 6.34 Holograms made with and without prefilter and appearance of prefilter (all x 60) object in this case was an aperture, 6mm square, this being used for making the filter and the hologram. Two holograms were made, one with and one without the prefilter, and their reconstructed images are shown in Fig 6.33. There is a marked difference between the images; although the direct light (zero frequency) is absent in both, there are no nonlinear terms or distortion in the image of the prefiltered hologram. Microphotographs of the holograms and of the prefilter are shown in Fig 6.34. In these photographs, the action of the filter is demonstrated, it can be seen that some parts of the transform have been attenuated too strongly, resulting in blank areas, in which carrier fringes are almost nonexistent. This lack of recording would be as equally undesirable as the overexposure of the transform in the unfiltered hologram, and the prefilter transmittance must be adjusted to obtain optimum attenuation conditions.

The autocorrelation response of the hologram may be viewed with or without prefilter, and both are shown in Fig 6.35. The effect of leaving the filter in the system when viewing the ACF response is more pronounced with the filtered hologram than with the unfiltered hologram, as expected. The nonlinearities of both unfiltered hologram responses can be seen, and these are clearly reduced when the response of the filtered hologram is considered. In this hologram, the recording of a wide range of low spatial frequencies produces a broad autocorrelation spot, this may be reduced in size and sharpened, by leaving the prefilter in the system.

6.4.3 Experiments with one dimensional transforms - slit aperture

The action of the prefilter in removing nonlinearities was studied by looking at nonlinear terms of one dimensional hologram, which, as shown in Chapter 5, can be very pronounced. The apparatus for this work is shown in Fig 6.32(b). The prefilter was made in a similar way to that used in the previous section, the best exposure being found by trial and error.

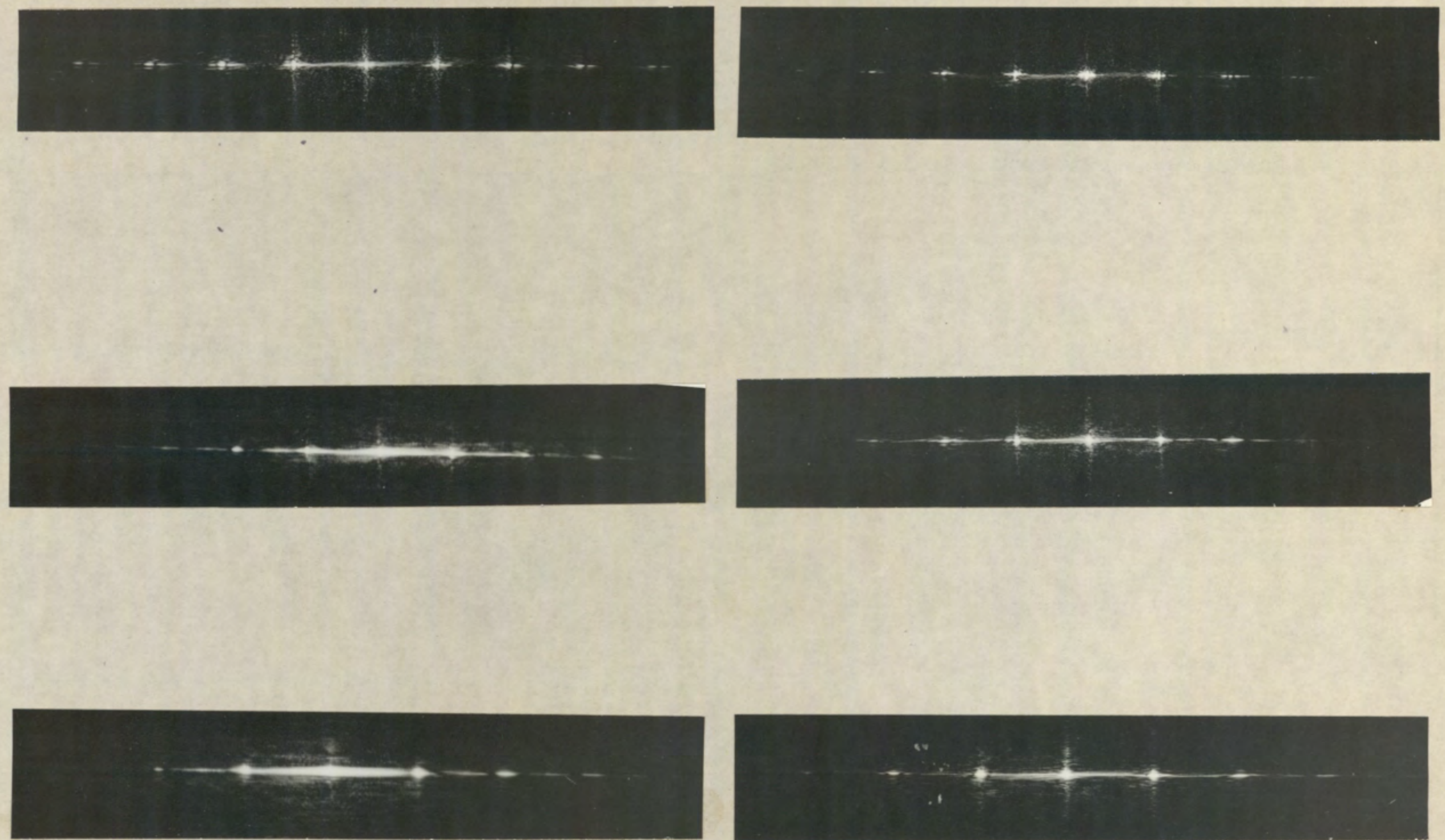
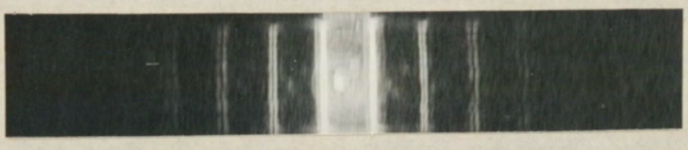
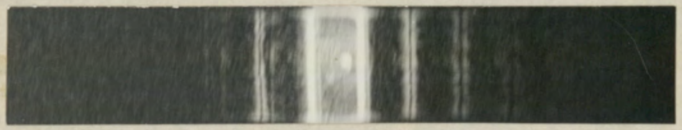


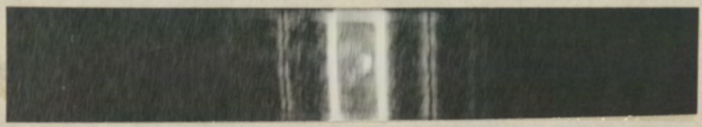
Fig. 6.36 Autocorrelation responses of one dimensional holograms showing effect of prefiltering



No prefilter



Hologram with prefilter in air



Hologram with prefilter in liquid gate

Fig. 6.37 Reconstructions from one dimensioned holograms showing effect of prefilter and filter in liquid gate

Three holograms were made, one without prefilter, one with prefilter and one with the prefilter in a liquid cell, to absorb thickness variations. The impulse responses of all three holograms (ie: the reconstructions) are shown in Fig 6.36. Although the reduction in nonlinear terms is not striking, it may be seen that there is a reduction in nonlinearity when the prefilter is used, and that this result is improved when the filter is immersed in a refractive index matching liquid. The explanation for this latter effect is simply that phase diffraction effects produced by the filter itself tend to degrade the recorded fourier transform by diffracting light away from its normal position in the transform. This is a similar effect to that observed in Chapter 5, when the hologram was immersed in a liquid gate.

The autocorrelation responses of these holograms are shown in Fig 6.37. The first three illustrations show the responses obtained where the prefilter was not in the system (although in the case of the liquid cell hologram, the cell was left in position). The recording of low spatial frequencies and the broadening of the autocorrelation function is in agreement with earlier results, and contrasts strongly with the unfiltered hologram. Again, it appears that the filter and liquid cell combination are most effective in reducing nonlinear terms. The corresponding responses obtained when the filter is in position in the system are given in the lower photographs. Here the difference between responses is not great, although the absence of fine detail in the ACF of the filtered hologram corresponds with recording of low spatial frequencies.

6.4.4 Conclusion

The use of a prefilter has been successfully shown to reduce nonlinear images and terms from the reconstructions and autocorrelation functions of holograms. The reduction was not complete, and needs careful manipulation of exposure and development of the prefilter to produce optimum results. In cases where nonlinear terms could be intolerable and perhaps confused with the true autocorrelation, then the use of prefilters may be justified.

Although the use of such a filter seems to be contradictory with the concept of recording noise, this concept may still be carried out by using a small stop of the appropriate size in the 1st transform plane, the prefilter being used to reduce intensities on the outside of the stop, and hence the nonlinear terms. A better suggestion may be to use an apodized stop, or filter, whose density would vary in a gradual manner, from the centre (zero spatial frequency) to high frequencies. Such a filter could be made by recording an out of focus fourier transform of a circular aperture. Although this filter would not contain density variations matched to those of the transform of the signal transparency, its overall density profile will correspond with that of the signal transform and thus reduce high intensities sufficiently to reduce nonlinear recording. At the same time, the filter would act as a redundancy noise filter, since it would obstruct zero and low spatial frequencies.

Chapter 7 Applications and Conclusion.

7.1.1 Introduction

One of the most promising applications of holographic spatial filters is that of character recognition. In this application the general problem is to translate each of the individual letters or characters of a text into a suitable code which can be retained or stored by a machine, for example, the memory store of a computer. The commercial requirement is that the character reader be fast in operation with very high accuracy in recognition, although allowing as much tolerance as possible on the form of the individual characters.

There are many ways in which the character reading or recognition may be implemented (for example Fischer, Pollock, Raddach and Stevens[1]) most of which involve scanning the character or letter format and observing intensity variations.

In such methods incoherent light is used with its many inherent advantages; a promising approach to character recognition by incoherent fourier transformation was recently described by Leifer, Rogers and Stephens[2].

One of the first papers to discuss character recognition by holographic spatial filtering was that of Vander Lugts, Rotz and Klooster[3]. Here the holographic process was used to enable a matched filter to be made, which would select any one character from a set. The properties of the filter were discussed, for example, the appearance of the auto-correlation response as a function of spatial frequencies recorded, and as a function of signal orientation and signal quality.

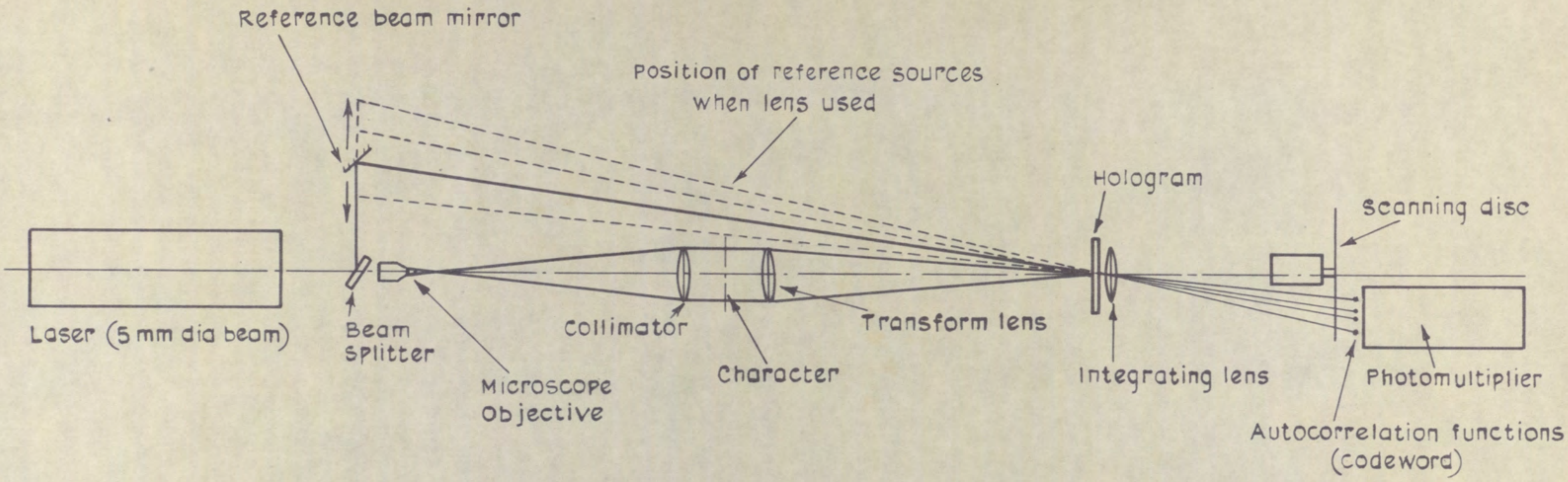
The problem of recognising more than one character was also discussed. The limitations in having one filter for each character were realised, and instead a multiple character filter was proposed, having data concerning all the characters of the set recorded on it. This was accomplished by having a set of numerals which were recorded on the hologram as a complete array. The response of the hologram was then observed, by showing a set of numerals which were displayed in each of the positions occupied by a number in the original display. Points of light, indicating autocorrelation, were obtained whenever a number in the set agreed with the number in the array.

The performance of the filter as a function of its frequency pass band characteristics was emphasised, as well as the difference in relative autocorrelation intensities caused by different signal energies of the numerals. The possibility of applying this method to recognition of certain key words in text was also demonstrated.

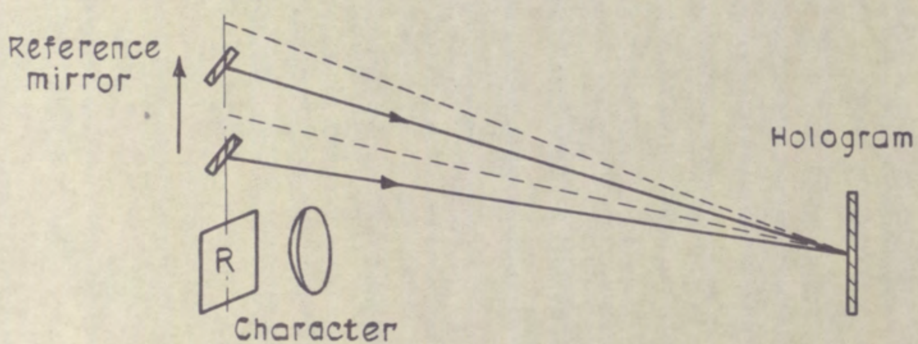
The possibilities of this method will be limited by the signal to noise ratio required for detecting the autocorrelation peaks. This ratio will be decreased if the filter is made to recognise a very large number of patterns or characters because of cross correlation between one pattern and another. In addition the filter performance will be limited by characteristics of the photographic emulsion on which it is stored; scattering of light by the emulsion will decrease the signal/noise ratio. This problem was studied by Burckhardt(4) who calculated that the storage capacity of a hologram filter was reduced from 10^5 to 10^3 for a noisy system, assuming that a signal to noise ratio of 100:1 were demanded.

Character recognition using a similar method to that of Vander Lugt was reported by Binns, Dickinson, Watrosiewitz(5) and also in France, by Lowenthal and Belvaux(6).

A somewhat different approach to character recognition was described by Gabor(7), in which variants of letters could be identified using a single photographic plate. Vander Lugt(8) has pointed out that this method can read only one character at a time (as opposed to his original method) and therefore demands time sharing of the input data. An examination of



a Experimental character reading device



b Coding process

Letter	Code			
R	o	—	o	—
5	o	o	—	—
2	o	—	—	o
0	o	o	o	—

c Codes

Fig. 7.1

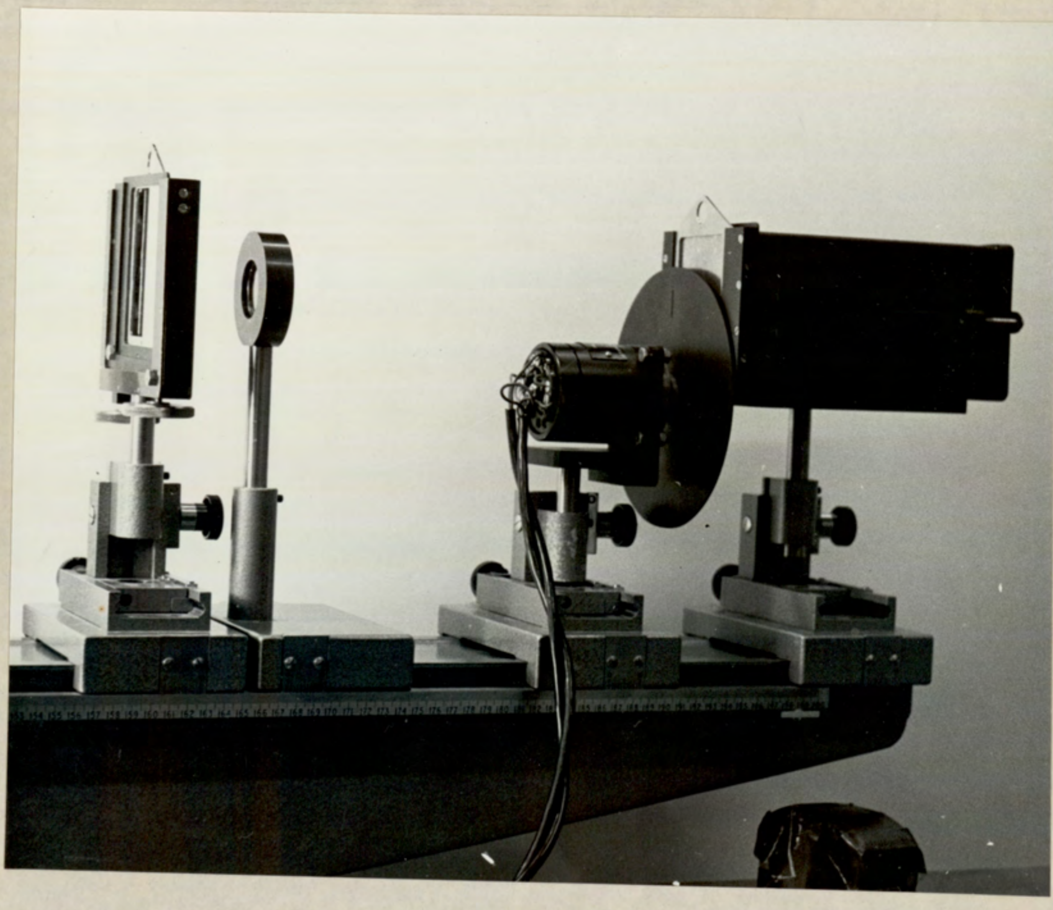


Fig. 7.2 Detector apparatus for character recognition

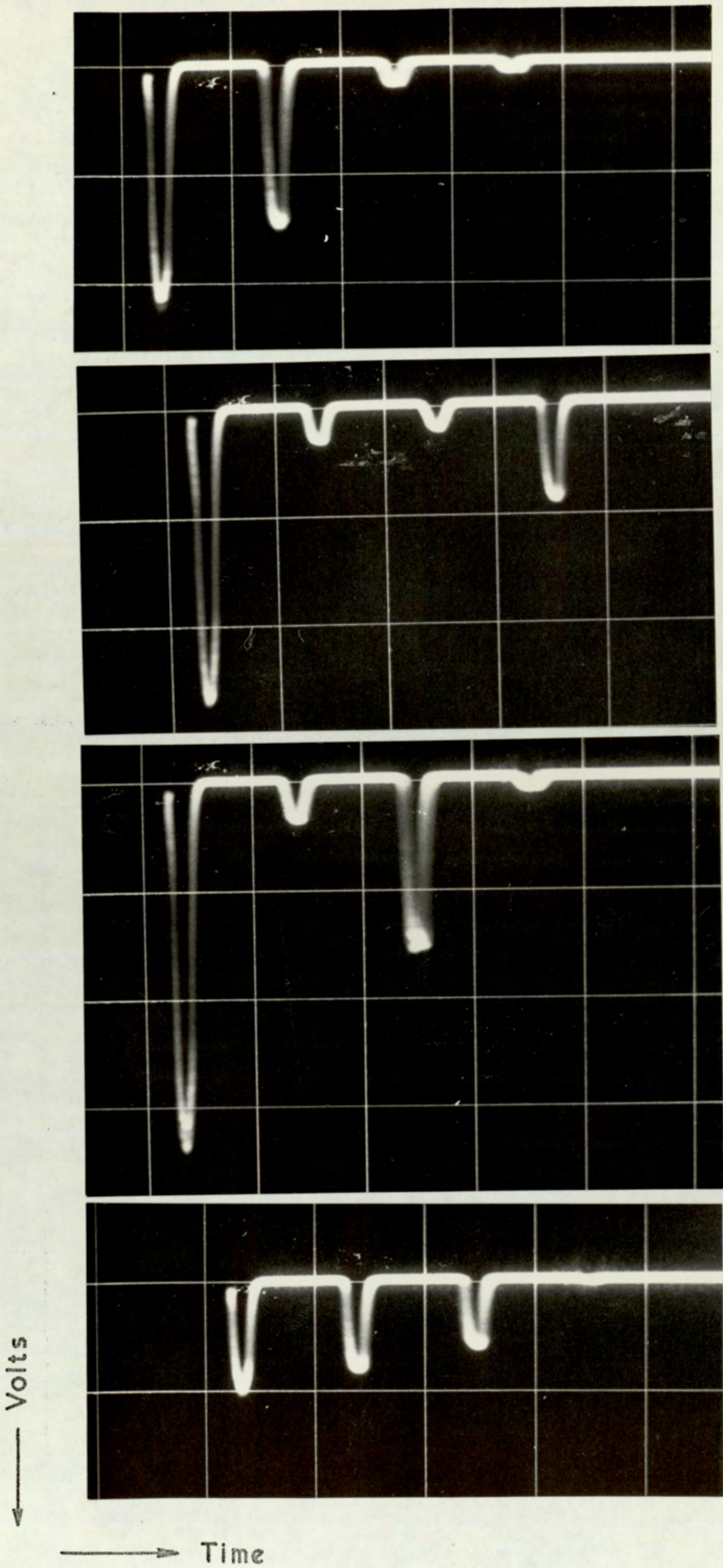
Gabor's proposal was made by Dickinson[9] who showed that the method was fundamentally sound but could have difficulties in practice, especially when extended to large numbers of characters and their variants. Gabor's method was also discussed by Watrasiewicz[10] who studied some of the theoretical requirements, in particular pointing out that characters must have a very narrow autocorrelation if it is to work efficiently.

An experimental investigation in character recognition, based on Gabor's method was carried out by Marchant and Keyte[11,12] which culminated in a pilot character reading device being exhibited at the Physics Exhibition, London 1967. This experimental work is reported in the following section.

7.1.2 Experimental Character Recognition Device

A diagram showing the optical system used in this device is given in Fig 7.1. The hologram was prepared by exposing it to the fourier transforms of each of four characters in turn. For each character there were a number of exposures, the reference beam mirror being moved between each exposure so that each character was recorded together with a number of reference beams. The different angular positions of the reference beams formed a code, which could be altered for different characters. For example, in Fig 7.1(b), letter R was exposed with the reference beam in two positions, as shown, giving the code: 1010. The other three letters were exposed with different codes, as shown in Fig 7.1(c). All the holograms were exposed with the plate in one position, this was then processed and replaced in this position using kinematically located plate holders.

The output from the hologram was detected using a photomultiplier, the actual output or correlation plane being scanned by a rotating slit, to enable each correlation point to be recorded separately. This arrangement is shown in Fig 7.2. The photomultiplier output was coupled to an



Letter S
Code 1100

Letter 2
Code 1001

Letter R
Code 1010

Letter O
Code 1110

Fig. 7.3 Character recognition device - output signals

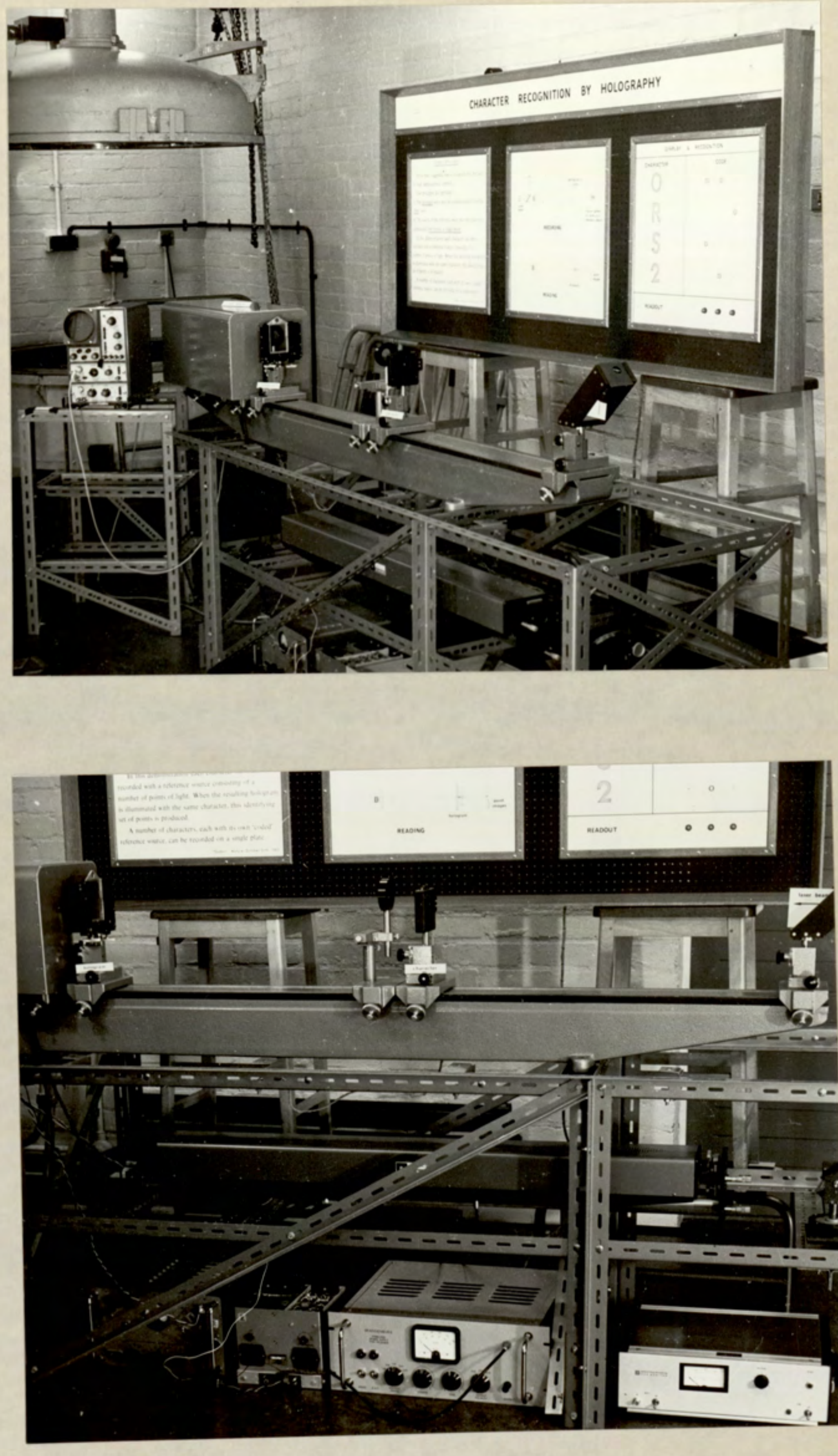


Fig. 7.4 General views of character recognition device

oscilloscope, producing the traces shown in Fig 7.3. Although cross correlation between different letters produced some unwanted correlation peaks, there was sufficient difference in voltage levels of different peaks to enable a threshold to be set, of the order of 1.5 volts, above which a correlation peak may be said to be caused by autocorrelation between the hologram and the letter.

As an additional refinement, (for display at the Physics Exhibition) a memory store was added, which indicated which letter was being displayed on a board at the rear of the device. The complete exhibit is shown in Fig 7.4.

Practical experience with the device showed that it could be readily transported and assembled, and that it could work efficiently under fairly arduous lighting and vibration conditions. The output signal strength was improved by removing the beam expanding microscope objective and collimator shown in Fig 7.1, and allowing the 5mm diameter laser beam to illuminate the character. This worked quite well, although not strictly desirable, since the character was very unevenly illuminated.

7.1.3 Conclusions

The basic principle of Gabor's idea had been successfully demonstrated, but it was felt that the storage of more than 4 letters as fourier transforms in one area of the photographic emulsion would present problems in obtaining the correct code words. With care, perhaps 10 numerals might be stored, although one might have difficulty due to cross-correlation between the numerals 3 and 8; this might be useful in banking and accountancy. The task of reading the complete alphabet is clearly difficult and account must also be taken of the variants (for example, capitals and different type fonts). For this task, it is likely that Vander Lugt's method would be more successful, since the hologram plate is used to better advantage.

7.2 Radar Signal Processing

7.2.1 Introduction

The use of optical systems for processing and filtering of data was discussed at length in a paper by Cutrona, Leith, Polermo and Porcello[13], and also by Cooper[14], with particular reference to radar signal processing. The specific application of optical techniques to radar data processing was discussed by Cutrona, Leith, Porcello and Vivian[15], with reference to the processing of data collected by side ways looking synthetic aperture radars.

A method of optical signal processing for circular array radar was described by Lewis[16]. This method proposed that the reference function, with which the radar signal was to be correlated be made holographically, and is therefore within the context of this report.

7.2.2 Optical signal processing for circular array radars

A convential radar system must search the space about it by scanning with a narrow beam of radio waves, measuring the position of any target by the time taken for the reflected wave to be received and by the azimuth of the transmitted beam. This has the disadvantage that it cannot search any given azimuth continuously, and because of the need to wait for echoes to be received, means that the scanning speed is fairly slow; typically 6 rev/minute. One thus sees a target every 10 seconds, which can be a severe limitation for fast moving targets (eg: aircraft).

A method of overcoming this restriction uses a stationary circular radar aerial array scanning all 360° at once. This would be similar in principle to planar arrays, on which experimental work has been done, but which are not suitable beyond $\pm 40^{\circ}$ of the normal to the array. The circular array could in principle give virtually continuous scans of all azimuths, its continuity being limited only by the pulse repetition frequency of the radar beam.

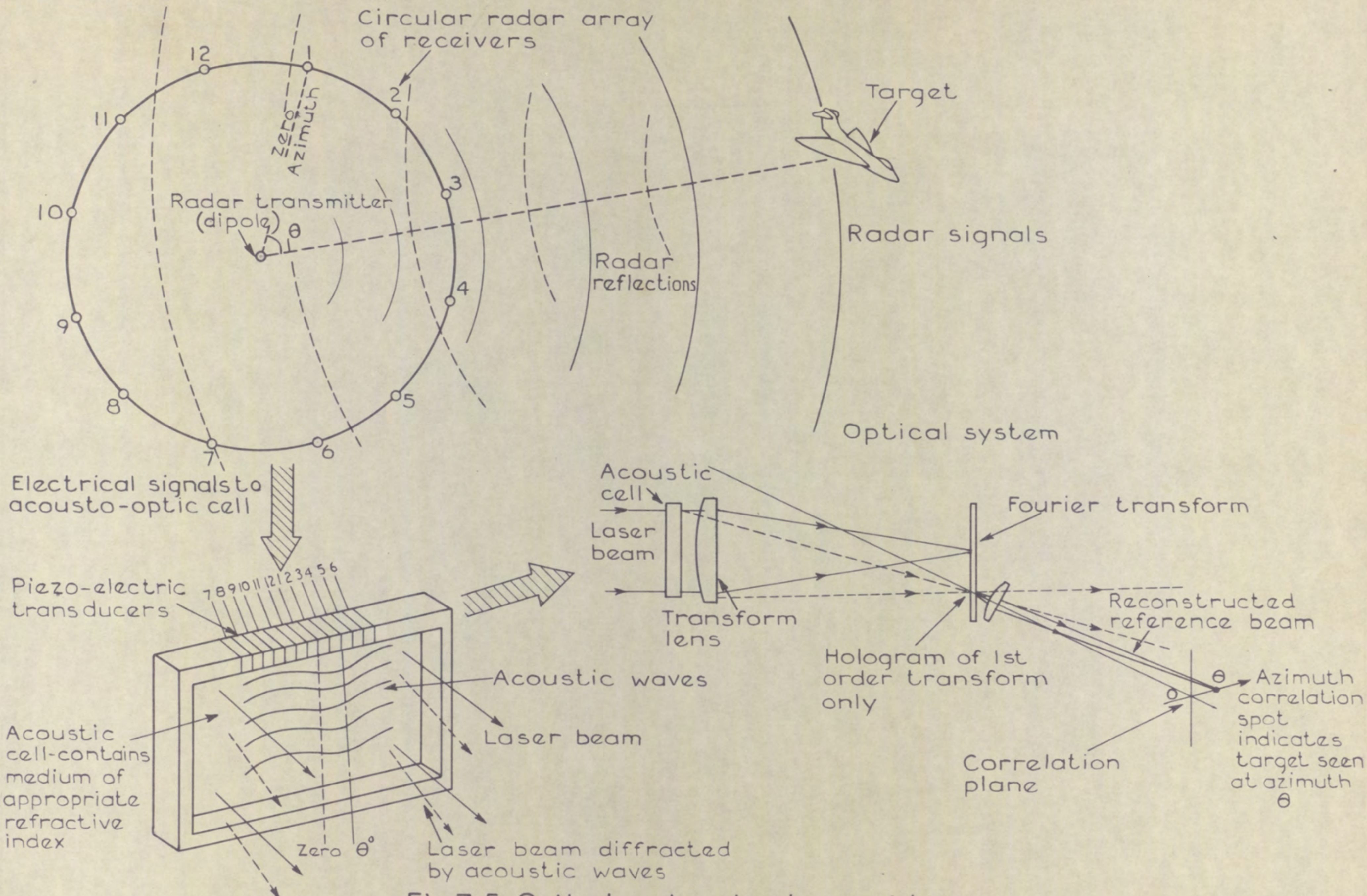


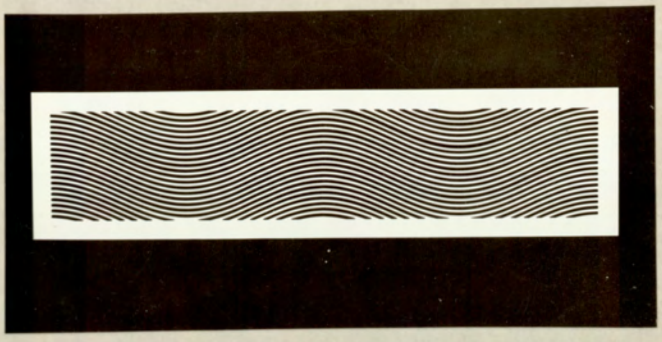
Fig.7.5 Optical radar signal processing

The signals from each aerial in the array must be processed so that the presence of any target and its azimuth relative to the array is obtained.

For any distant target there will be a function $s(\gamma)$ describing the signals at each aerial (γ) in the radar array. For a specific target at azimuth θ(say) there will be a specific function $r(\gamma - \theta)$ describing the radar signals, and to determine the presence of the target it is necessary to correlate these two functions (see Chapter 1). Thus the process must evaluate $\int s(\gamma)\ r(\gamma - \theta)d\gamma$. This will determine whether or not a target is present at azimuth θ; for other azimuths it is necessary that the value of θ be shifted, just as the value of "u" is shifted in an optical correlator. Thus two things have to be performed; i) the operation $\int s(\gamma)\ r(\gamma - \theta)$ and ii) this operation repeated for $0^{\circ} < \theta < 360^{\circ}$. Lewis showed that such an operation can be accomplished fairly easily by optical means, using electronic techniques the process can become very complex and expensive.

7.2.3 <u>Practical optical radar processor</u>

The actual operation is shown in diagrammatic form in Fig 7.5. Signals from each of the receivers in the array are taken to an acoustic-optical cell, where the signals are converted into acoustic waves by piezo-electric transducers. These waves travel downwards through the cell, where they diffract an incident laser beam. The fourier transform of the diffracted beam is formed by a transforming lens, the transform then being multiplied by a fourier transform hologram of the signal function $s(\gamma)$. Correlation between the diffracted beam and the hologram is indicated by a point of light (the reconstructed reference beam) in the correlation plane. In the example shown in Fig 7.5, the target is at azimuth θ, therefore the acoustic wave peaks are displaced horizontally by a linear distance proportional to



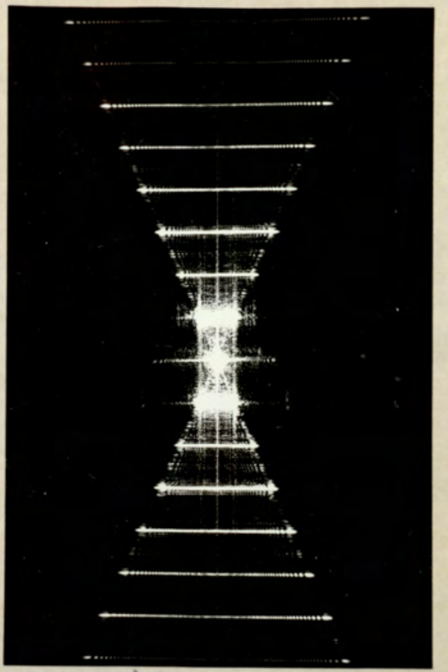


Fig. 7.6 Input mask

Fourier transform

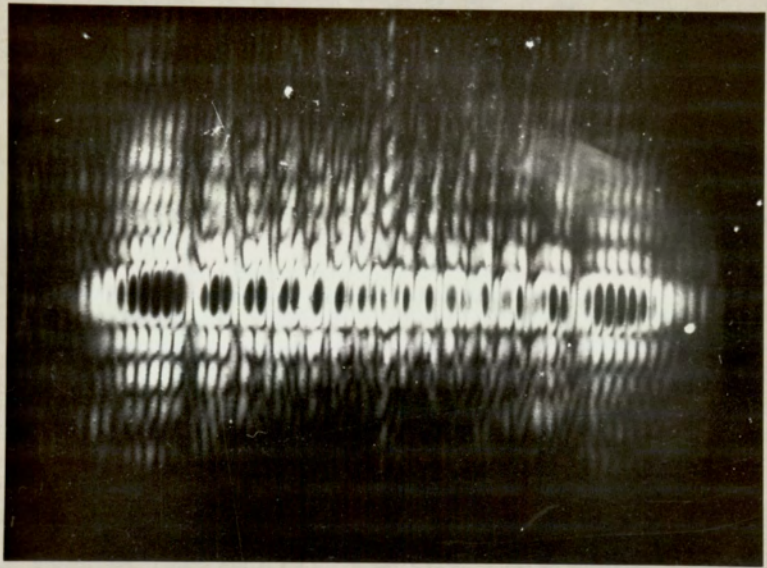


Fig. 7.7 First order transform hologram

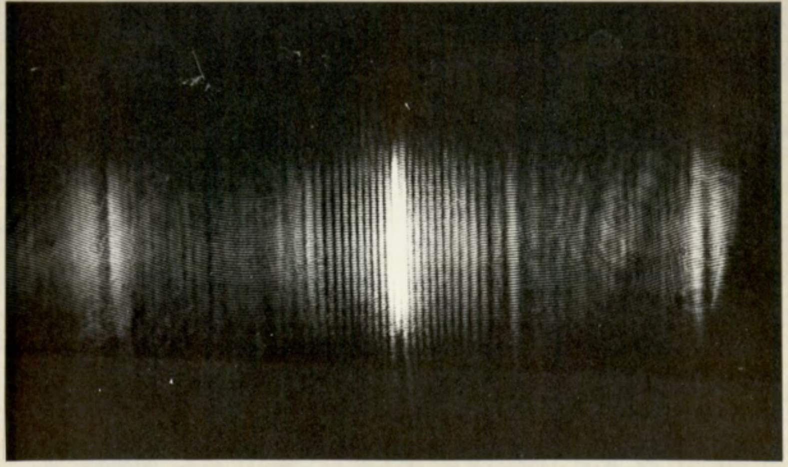


Fig. 7.8 Autocorrelation response

θ, which in turn alters the phase of the diffracted beam, and its fourier transform, causing the correlation spot to be displaced. The displacement of the spot is proportional to azimuth, so that one has both a method of evaluating the basic correlation integral to find if a target is present, and simultaneous information about the azimuth of the target.

In an experimental processor, constructed at RAE by Worrall[17], the acoustic-optical cell was simulated by a transparency consisting of alternate opaque and clear lines. This is shown in Fig 7.6, together with its diffraction pattern or foûrier transform. Since the lines are "square wave" in form, many orders are obtained, although the hologram was made of only the first order of the diffraction pattern, (shown in Fig 7.7) corresponding to the sinusoidal variation of the acoustic waves. The optical system for making the hologram was very similar to that described in Chapter 4, and embodied similar design considerations, (eg: vibration free working surface). The apparatus was tested by looking for the auto-correlation response of the hologram with the original mask; this is shown in Fig 7.8. The correlation peak is indicated by a central vertical line, on either side are two weak lines caused by cross correlation between opposite halves of the input pattern.

To ensure that correlations could be obtained over the full 2π range of azimuths, the input transparency was made of width 4π (ie: twice the width of the normal aperture of the system, see Fig 7.6). The hologram was made by recording the fourier transform of the whole 4π transparency; outputs were obtained by covering the original transparency with an aperture corresponding to a width 2π, and allowing this to correlate with the hologram. The mask could be moved to expose different portions of the transparency, and thus simulate targets at different azimuths.

[The 4π transparency and the area masked off, of width 2π, are shown in Fig 7.6.]

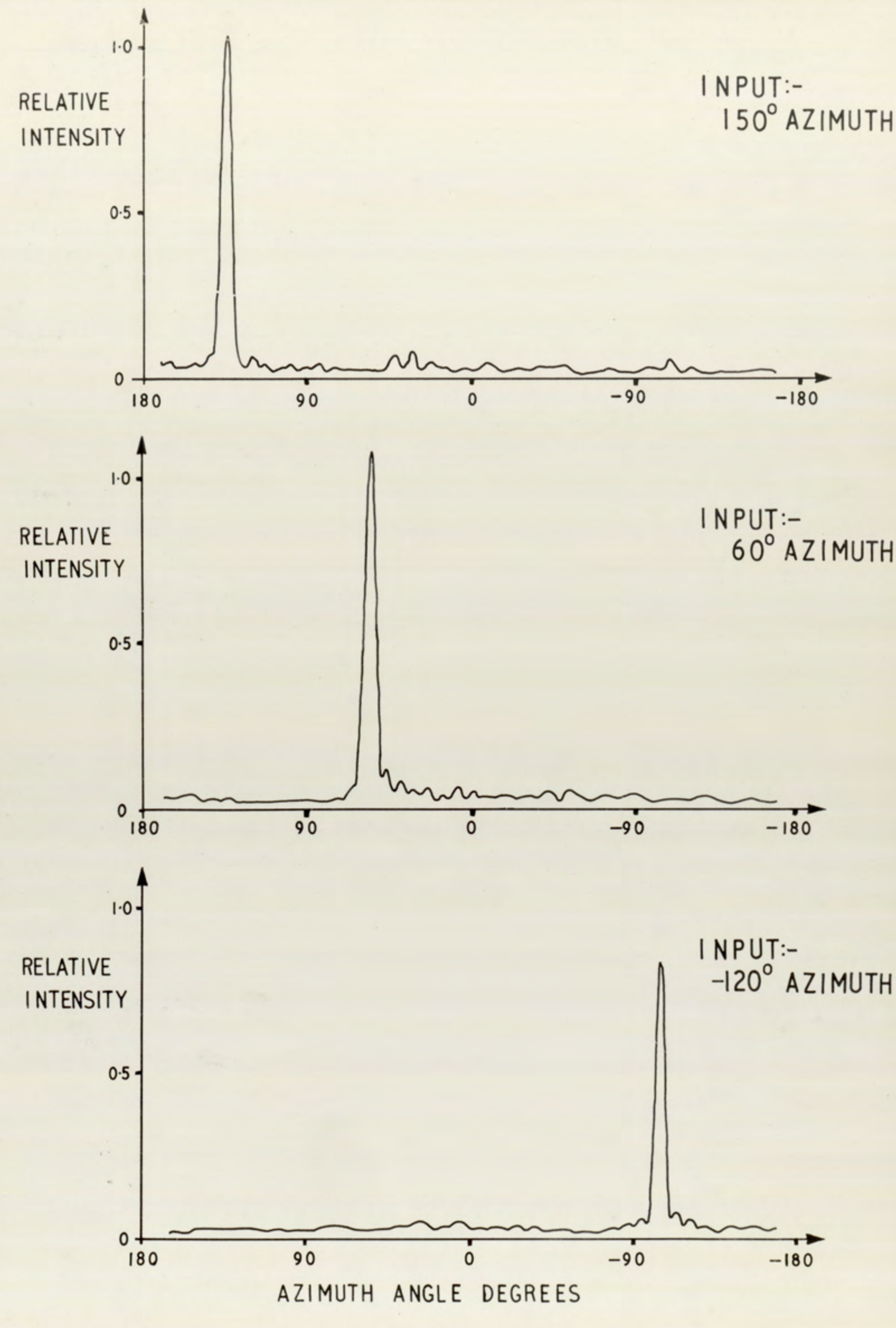


Fig. 7.9 Intensity distribution in correlation plane of radar signal processor for a simulated target at different azimuths

The intensity of the autocorrelation function was measured for different azimuths, the intensity distributions being shown in Fig 7.9. In general, it was found that there was good agreement between calculated intensity distributions and those measured.

Further work is dependent on the development of an acousto-optic cell of sufficient channel capacity and acceptable dimensions.

7.3 Fingerprint recognition

The possibility of using complex spatial filters to measure the correlation between fingerprints was fully explored by BAC.[18] In this work, the aim was to develop a system which would recognise scene of crime fingerprints by correlation with a rolled impression from Police files.

The basic principle of BAC's method is similar to that described in Chapter 3 and 4; a transparency is made of a filed fingerprint, and this is used to prepare a hologram. The hologram is then correlated with other transparencies, containing images of scene of crime prints. The presence of an autocorrelation peak, or bright spot in the correlation plane indicates that the two fingerprints are identical.

Although the idea was basically sound, considerable difficulty was experienced when attempting to correlate different prints of the same finger, due mainly to distortion of the finger when making the impression.

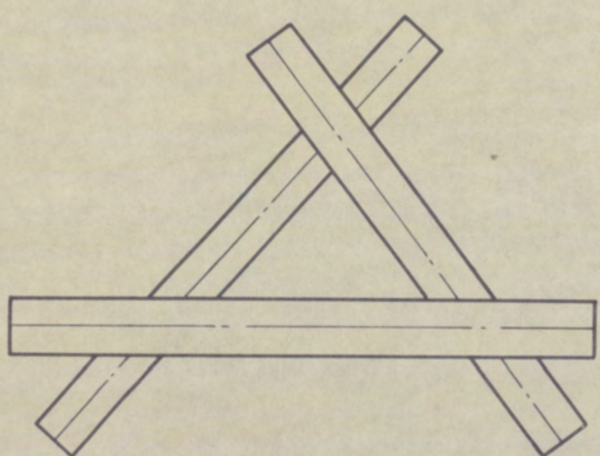
The distortion was carefully studied, and theoretical calculations were made for the correlation output for different types of distortion (based on simple models of finger print structure). In all cases, it was found that correlation output decreased rapidly as the amount of distortion increased. It was found that good discrimination of one fingerprint from a set of prints could be obtained provided that the print being searched for was well formed. Fragmentation of the print, smudging, scale change, skew and rotation orientation all tended to make the autocorrelation output low. Because of the variation in absolute levels of the autocorrelation output a method of determining the signal to noise ratio of the auto-correlation spot was devised. Here the ratio of peak autocorrelation intensity to intensity at some distance from the peak was measured. A variety of procedures for measuring the ratio were explored, one of the most promising being a method of integrating over an area surrounding the autocorrelation peak whilst keeping the peak itself obscured, thus enabling the noise intensity to be found.

Nevertheless, even with signal/noise ratio autocorrelation measurements, it was found that only a limited degradation of prints could be tolerated in order that there would be only one "false alarm" in 500 searched prints. These were considered to be very good prints compared with typical scene of crime prints where severe degradation might be expected.

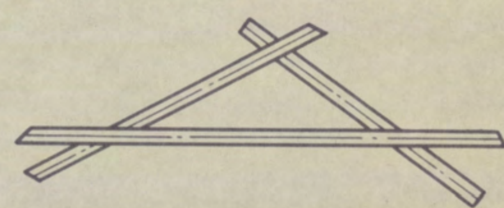
The result of this work concluded that optical matched filtering would be unable to assist in recognition of scene of crime prints, but might be of use in other fields where the quality of prints would always be good, eg: customs or passport control.

7.4 Recognition of Air Photographs

The possibility of recognising targets from aerial photographs is one which has attracted some attention (eg: in the USA) particularly for military applications. Examples used in this report have shown that fourier transform holograms can be made from air photographs, and used to recognise a particular area. However, experience in the use of fourier transform holograms in this work, and on character recognition supports the evidence of BAC's fingerprint report, in which distortion of the pattern being sought is the limiting factor. In aerial photography it is obvious that considerable distortion could be present, particularly in photographs taken from low altitudes at varying angles of incidence. This distortion would be most severe for tall objects, but even relatively flat two dimensional



Runway pattern – overhead
(90°)



Runway pattern – oblique incidence
(25°)

Fig. 7.10

objects will suffer distortion, as shown in Fig 7.10. Unless it is possible to ensure that the target is always viewed from the same azimuth, it is unlikely that optical matched filtering will be of great use. It is possible to overcome some of these limitations by storing information about the target in different azimuths on the one filter (ie: a multiplexed hologram, similar to that used for character recognition). This has the additional risk that the discrimination of the filter is reduced.

In high altitude air photography it seems that there may be more possibility of using optical matched filtering, mainly because the field of view would be narrow and hence introduce less risk of distortion. This advantage could be offset however, by the change in image due to shadows and by lack of contrast caused by scattering, cloud cover, etc. The likelihood of optical techniques being successful in this field would depend very much on the quality of the air reconnaissance system.

7.5 General Summary and Conclusions

7.5.1 Correlation processes using coherent optics

The method of correlating input data with a master pattern using coherent optical systems and holography has been well established by theory and experiment. The maximum intensity of the correlation pattern obtained can be used as an indication of whether the input contains the pattern being sought, the threshold intensity being set by the intensity obtained when an autocorrelation (ie: self-correlation) response is obtained. This means that factors which can cause the absolute intensity level to change must be rigidly controlled, which can be difficult in practice. For example, lack of space invariance within the optical system; laser beam temperal intensity fluctuations, laser beam input plane intensity fluctuations are all factors which cause alteration in correlation peak intensity.

In practice, laser beam intensity fluctuations, whilst increasing the mean levels of the input signal, also increase the amplitude of the signal itself. For example, referring to Fig 1.2 (Chapter 1) the amplitudes of the bar pattern are 3 and 2 for s(x). If the laser beam intensity was increased by 10%, and if the function s(x) is a function of transmission then the amplitudes become: 33 and 22. The diffraction pattern formed by this pattern will depend on the change in amplitude across the input, which can be seen to vary depending on the incident intensity. Thus the autocorrelation peak intensity will not be constant even if the zero spatial frequencies are blocked.

This means that it is difficult to construct efficient optical correlators which use a measurement of absolute correlation intensity, although limited success was obtained using this method for character recognition (section 7.1). Instead, it is necessary to consider using signal to noise intensity measurements, (ie: peak autocorrelation energy, to the mean energy in correlation plane surrounding peak) as demonstrated by BAC (section 7.3). An alternative method would be to monitor the input signal power, and use this to control the gain of the photo-detector measuring the correlation intensity. This latter method would only work for those cases where the autocorrelation peak intensity was greater than the intensity at any other point in the correlation plane, otherwise incorrect results could be obtained. In general, it seems that the method of sampling the noise energy around the autocorrelation peak is the method least likely to give false readings. This is particularly true where a wide variety of input data may be encountered, since then the correlation plane can contain large intensity variations.

One method of simultaneously obtaining both correlation peak intensities and noise intensities is to scan the correlation plane with a plane mirror which contains a pinhole of appropriate diameter in its reflecting surface.

When aligned with the correlation peak, the peak energy is transmitted by the pinhole, where it is detected by a photo detector. The surrounding noise energy is reflected by the mirror and collected by a second photo detector. The noise signal may then be used to determine the threshold, above which a signal is said to be caused by correlation with the pattern being sought.

7.5.2 Nonlinear processing effects

It has been shown that the photographic recording process is incapable of allowing an accurate recording of the fourier transform to be made except at low exposure levels, where there may be some loss of information in high spatial frequencies. The main effects of nonlinear recording are that the autocorrelation function intensity and shape will vary according to exposure, and that there may be false images surrounding either the autocorrelation function or the reconstruction. The former effect indicates that if peak intensity to noise intensity ratio measurements are used as a criterion of autocorrelation, then particular care must be taken in judging the hologram exposure. The study of the variation in autocorrelation intensity distributions also suggests that the method of noise measurement should be selected according to the type of pattern being sought. There is some evidence that the integrated noise measurement suggested by BAC may be useful for a variety of patterns.

7.5.3 Optical system

It was found that aberrations of the optical system could have a marked influence on the autocorrelation measurement. It seems that most of this trouble arises from spherical aberration of the transform lens, and could be corrected by use of a figured corrector plate, or by using specially designed optical components. The tolerance on fourier transform movement will be of the order of microns, and it will be necessary for both focussing and aberration errors to be reduced to this order of movement. There is a need for further information on the effect of aberrations on autocorrelation intensity.

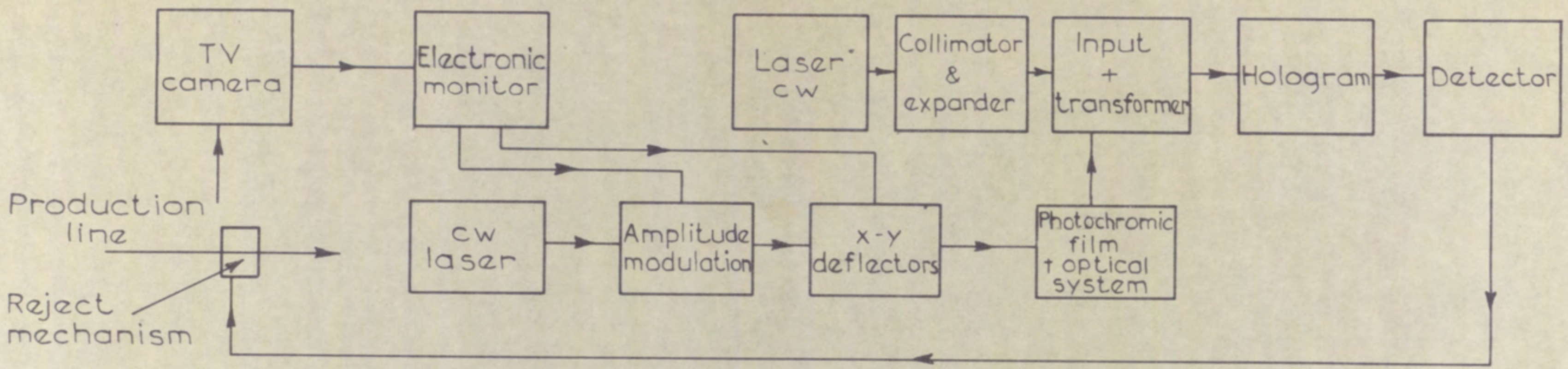
The recording of fourier transform holograms from half tone transparencies does not appear to have presented insoluble difficulties. Using graded boundary (or apodized) masks and liquid refracture index matching cells, the fourier transform of only the amplitude transmittance information can be obtained. The presence of a slightly diffuse medium in the input transparency modulates the transform forming a network of light and dark lobes, whose size depends on the aperture of the apodized mask. The presence of these lobes suggests additional research work, in an attempt to determine their effect on the performance of the correlator, and if necessary, their optimum size.

Some simple experiments have also been made (although not reported here) in which the sensitivity of holograms to patterns of different shape to that recorded were examined. It was not possible to investigate this problem in sufficient detail, but since it is related to the problem of distortion in applications such as finger print recognition, it could prove a particularly useful field for future work. In particular, one would be interested to know the maximum amounts of distortion which can be tolerated for different kinds of patterns.

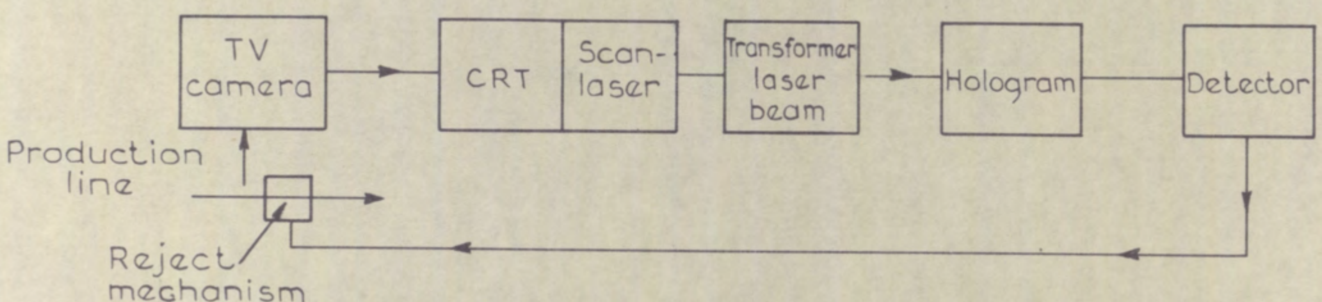
7.5.4 Future applications

In all of the applications so far discussed in this chapter there is the possibility that the pattern being sought will be distorted in some way. Whilst some allowance can be made for distortion, for example, in recognising different type fonts of a single letter, this makes the recognition system less fool proof, and can lead to a greater risk of false alarms. It seems that this particular problem is quite severe in the optical matched filtering process, compared with other recognition systems.

There are applications, however, where the pattern being sought is required to be identical to that recorded on the hologram, and that any distortion or fragmentation in samples of the pattern should be rejected.



Real-time recognition system using photochromic film



Real-time recognition using a scanlaser

Fig. 7.11

It is thouth that the holographic optical correlator would be more suitable for this kind of application, since any errors would tend to be of the "fail-safe" kind, such that only perfect patterns are accepted by the machine. Some applications which come to mind are; rejection of imperfect stamps, banknotes and other mass produced documents which must be rigidly controlled.

With this kind of application comes the problem of feeding the input data into the coherent optical system. Although most work reported in literature and in this report has been concerned with transparency type input data, there are several possibilities by which information from an incoherent optical scanning system can be fed into a coherent system. Examples of two possibilities are given in Fig 7.11. The first, using photochromic film, on which the data is written by an electronically modulated and deflected laser beam uses a concept developed originally for a laser driven display system.[19] The second utilises an idea developed by Pole and described by Baker[20]. The scan laser consists of a CRT whose face acts as a resonator of a laser cavity. This face consists of sandwich of electro optical and birefringent materials which serve to reflect and change the polarisation of the incident laser beam. The change in polarisation stops laser action, and may be cancelled by directing the CRT electron beam to that particular point, at which laser action will commence. Thus control of the CRT enables control of the intensity and direction of the laser beam.

Investigations in these last two directions will call for much effort and expenditure, and it may be more appropriate to use other methods of pattern recognition. The limitation applies only to opaque objects, since any form of transparent object can be easily inserted into a laser beam, and cause a fourier transform to be formed from its phase structure.

Another use of fourier transform holograms is not concerned with optical correlators, but with the recording of phase variations in transparent objects. It has been shown that the recording of over exposed fourier transform holograms enabled a detailed study to be made of amplitude and phase changes in the input data. In particular, the phase structure across an aperture may be clearly revealed when the zero frequency of a hologram is unrecorded, although care must be taken that low frequency information about the phase structure is allowed to be recorded. This method uses the impulse response of the hologram, rather than the auto-correlation response, and is identical in principle to the method of phase contrast used in microscopy. It has the advantage that information can be easily stored and that control of exposure and processing enables certain frequencies to be suppressed whilst others are recorded.

It seems that there may be some use for holographic filters as a research tool in optics, and particularly in biology, where transparent specimens may need to be examined, and any phase variations to be revealed. The edge enhancement of other irregularly shaped objects is a possible application, in radiography for example. In this field however, there are already machines which are capable of looking for density or intensity gradients and giving either contour maps of half tone structure, or enhanced edges; there seems to be little point in developing a holographic system except again, as a general research tool.

This point seems to be supported by the fact that there is no evidence of holographic correlators being used for a particular application, whereas at least one commercial organisation offers a complete holographic correlator system as a general research tool[21].

7.5.5 Acknowledgments

I wish to thank the Ministry of Technology for giving me permission and support to carry out this research. I wish also to thank my supervisors Dr G L Rogers and Mr A C Marchant for their advice, discussions and criticisms. I am grateful for the help given by members of the Optics Section, RAE on a wide variety of technical matters, and to various departments of the RAE who were concerned in the manufacture of equipment, the preparation of some photographs and in the typing of the manuscript.

References: Chapter 7

1. Fischer, etc Symposium on Optical Character Recognition, Washington DC 1962 (Spartan)

2. Leifer, Rogers, Stephens. Optica Acta 16, 5, 1969, 535-553

3. Vander Lugt, Rotz, Klooster. Chapter 7, Optical & Electro Optical Information Processing, MIT press 1964

4. Burckhardt, Applied Optics 6, 8, August 1967, 1359-1366

5. Binns, Dickinson, Watrasiewicz. Applied Optics 7, 6, June 1968

6. Lowenthal, Belvaux. Rev d'Optique 46, No 1. 1967

7. Gabor, Nature Vol 208, Oct 1965, p422

8. Vander Lugt. Optica Acta 15, 1, 1968, 1-33

9. Dickinson, Marconi Review, First Quarter 1967

10. Watrasiewicz. Conference on image detection and Processing, RRE Malvern 1967

11. Marchant, Keyte, et al. RAE Tech Report 67031 January 1967

12. Marchant, Keyte, Conference on image detection and processing, RRE Malvern 1967

13. Cutrona et al IRE Trans.Inf.Th. IT6 3 386 (1960)

14. D C Cooper Rad & E E July 1966

15. Cutrona, Leith, Porcello, Vivian, 9th Symposium AGARD Avionics Panel 1965

16. Lewis RAE Technical Report TR67291 November 1967

17. Worrall, M J, Optica Acta, 1970, Vol 17, No 1, 37-42

18. Binns, Gregory et al. BAC Report K27/30/ELO/412

19. Elliott Brothers Ltd, (Private Communication)

20. Baker, IEEE Spectrum, December 1968 pp 39-50

21. Thomson - CSF Optical Correlator CC 632 Provisional Manual

APPENDIX A

Mathematical explanation of appearance of higher order images in the reconstruction from a fourier transform hologram

In the reconstruction from a fourier transform hologram it was observed that a series of images were formed consisting alternately of an image of the reference source and of the object, ie: source; object; source; object. In general the source image was surrounded by a pattern having the appearance of the autocorrelation pattern of the object. The explanation for these images is given below.

Consider a reference beam $R = A_R e^{-j\alpha p}$ and an object beam $S = A_s(p, q)e^{-j\phi(p, q)}$ where S is the fourier transform of a signal s(x, y). In the hologram plane there is an intensity distribution $[R + S]^2$ or $|R|^2 + |S|^2 + 2A_RA_S \text{ Cos } \theta$ where θ represents the angle $[\alpha p + \phi(p, q)]$ and shows that there is a sinusoidal variation of intensity across the hologram.

Thus the exposure $E = A_R^{\ 2} + A_S^{\ 2} + 2A_RA_S \text{ Cos } \theta$.

Because of nonlinearities in the photographic process, the transmission is not a linear function of E and may be represented by

$T_A(E) = k_o + k_1 E + k_2 E^2 + \quad \ldots\ldots$

For the term containing E there will be a record of $A_R^{\ 2} + A_S^{\ 2} + 2A_RA_S$ Cos and these will be multiplied by A_R (when the hologram is used in the reconstruction process) giving;

$$A_R.A_R^{\ 2} + A_R A_S^{\ 2} + 2A_R.A_RA_S \text{ Cos } \theta.$$

When this equation is inversely transformed by the reintegrating lens we find that on axis there is: $[\delta * \delta * \delta^*] + [\delta * S * S^*]$ where the δ signifies a delta function, (ie: the reference point source) and $*$ indicates convolution. This produces a point source image and the autocorrelation of the object $[S * S^*]$.

Off axis, one obtains $\delta * \delta * S$ and $\delta * \delta * S^*$ which produce the real and inverted images of the object.

For the E^2 term however, one has, $[A_R^2 + A_S^2 + 2A_RA_S \text{ Cos } \theta]^2$ giving;

$A_R^4 + A_S^4 + 4A_R^2A_S^2 \text{ Cos}^2 \theta + 2A_R^2A_S^2 + 2A_R^2.2A_RA_S \text{ Cos } \theta$

$+ 2A_S^2.2A_RA_S \text{ Cos } \theta.$

The $\cos^2 \theta$ term may be expanded giving $2A_R^2A_S^2 - 2A_R^2A_S^2 \text{ Cos } 2\theta.$

Thus on axis, as a result of nonlinearity there are additional images given by:-

$$A_R^5 + A_RA_S^4 + 2A_R \cdot A_R^2A_S^2 + 2A_R \cdot A_R^2A_S^2 .$$

All these images may be shown to produce either a point source image of reference source or an autocorrelation pattern of the object.

In the first order, Cos θ direction there are images

$2A_R^3\ 2 \cdot A_RA_S$ and $2A_RA_S^2 \cdot 2A_RA_S$.

These may be transformed giving; $(\delta * \delta * \delta^*) * \delta^* * S$ and $(\delta * S * S^*) * \delta^* * S$ for the one side and similarly for the opposite side but with complex conjugates of δ and S where appropriate.

These terms correspond to images of the object (S) and to the object convoluted with its autocorrelation function. Thus, in the first order direction, these nonlinear images reinforce that produced by linear recording.

In the second order, ie: Cos 2θ direction, there is an image given by $2A_R^3A_S^2$ or, $\delta * \delta^* * \delta * S^* * S$. This has the appearance of the auto-correlation of the object, so that one apparently sees another image of the reference source in those cases where the object has a sharp autocorrelation function.

There is a similar effect when the hologram is illuminated by the object beam, except that the transmission is multiplied by A_S not A_R, and the position of the on axis image will be shifted by an angle θ.

Thus the second order term becomes $2A_SA_R^2A_S^2$ giving $S * \delta^* * \delta * S * S^*$. This has the appearance of the image of the object multiplied by its auto-correlation, which in general will correspond to the object again. Since the hologram is now illuminated from a different direction, this image is shifted by amount θ, compared with that described earlier. Thus one obtains a series of images of source and object for both object and reference beam, and these images will appear to be in line with one another; thus an 'S' reconstructed by object appears in the same position as an 'S' reconstructed by the reference beam. [See Fig 5.6].

APPENDIX B

Effect of different parts of fourier transform in forming an image of a slit.

This study was undertaken in order to observe the change in appearance of the slit image when various parts of the fourier transform were obstructed. The fourier transform was formed in one dimension only, so that image effects would be easily observed, and also enabled obstruction of different parts of the transform to be carried out.

Obstruction of high orders of the transform was done using a wide slit, whose jaws could be adjusted to cover side lobes of the slit transform from 11th to 1st. The central lobes of the transform were obstructed using rods of metal of different diameters, selected to obstruct the required parts of the transform. In practice it was convenient to use twist drill s tems whose diameter varied from 0.5mm to over 10mm. In order that both means of obstruction could be used to select different parts of the transform, a double diffraction optical transform system was set up, as shown in Fig A1.

Both the rod and slit jaws for obstruction were accurately aligned with the fourier transform by using a low powered microscope. In this way different spatial frequency bandwidths could be selected, with equal bandwidths from both halves of the transform.

The appearance of slit images for different parts of the transform obstructed are shown in Fig A2.

There is a considerable similarity between some of these images and those shown in Chapter 5 (eg: Fig 5.18). This is especially true in cases where the zeroth and first order lobes have been obstructed and many high frequency lobes present. The results also show the variation in appearance of the thin dark line marking the edge of the slit aperture as different spatial frequencies are present in the fourier transform. Generally speaking, when high spatial frequencies are present, there are thin lines,

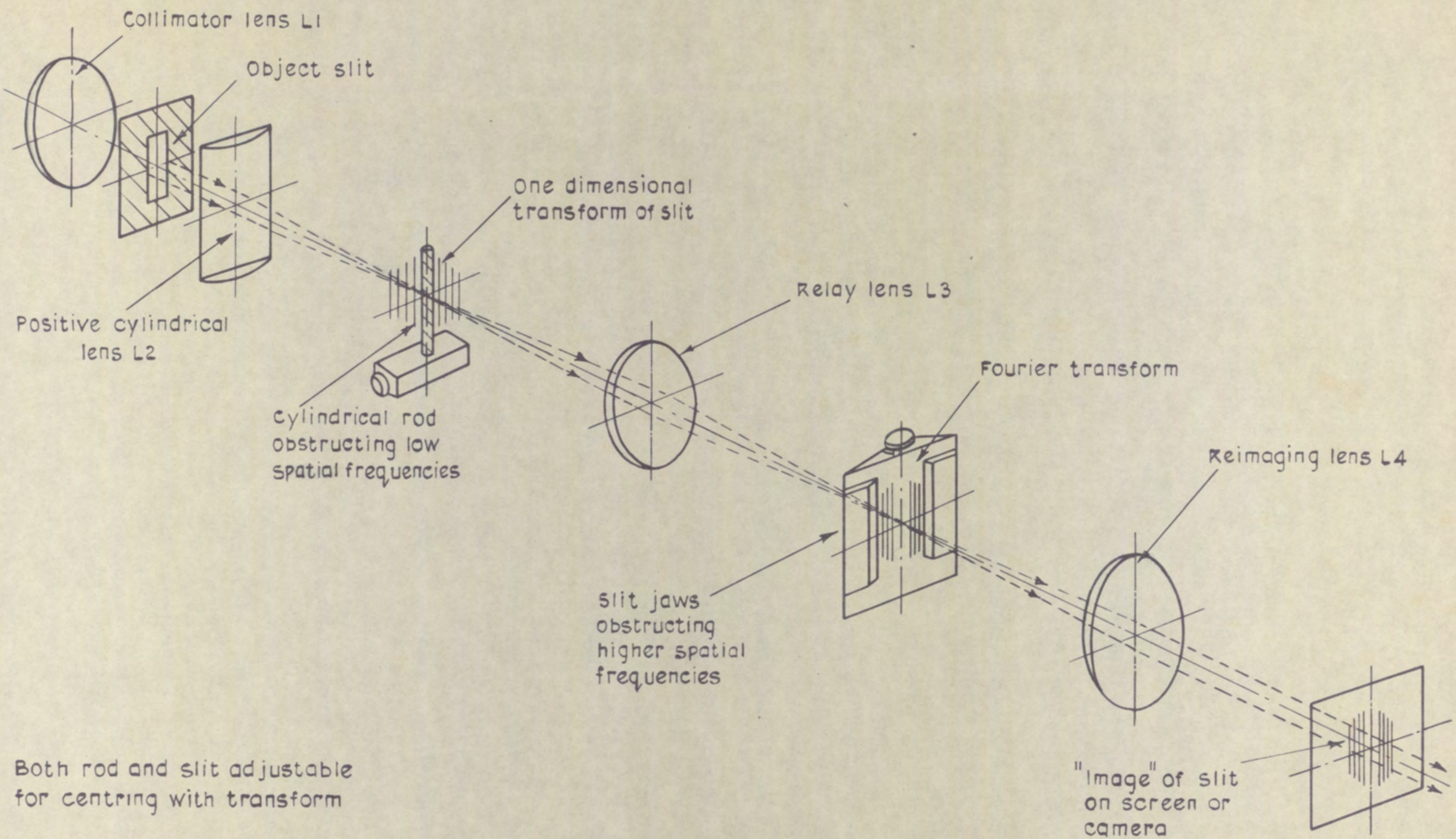


Fig. A1 Observation of slit image with obstructed fourier transform

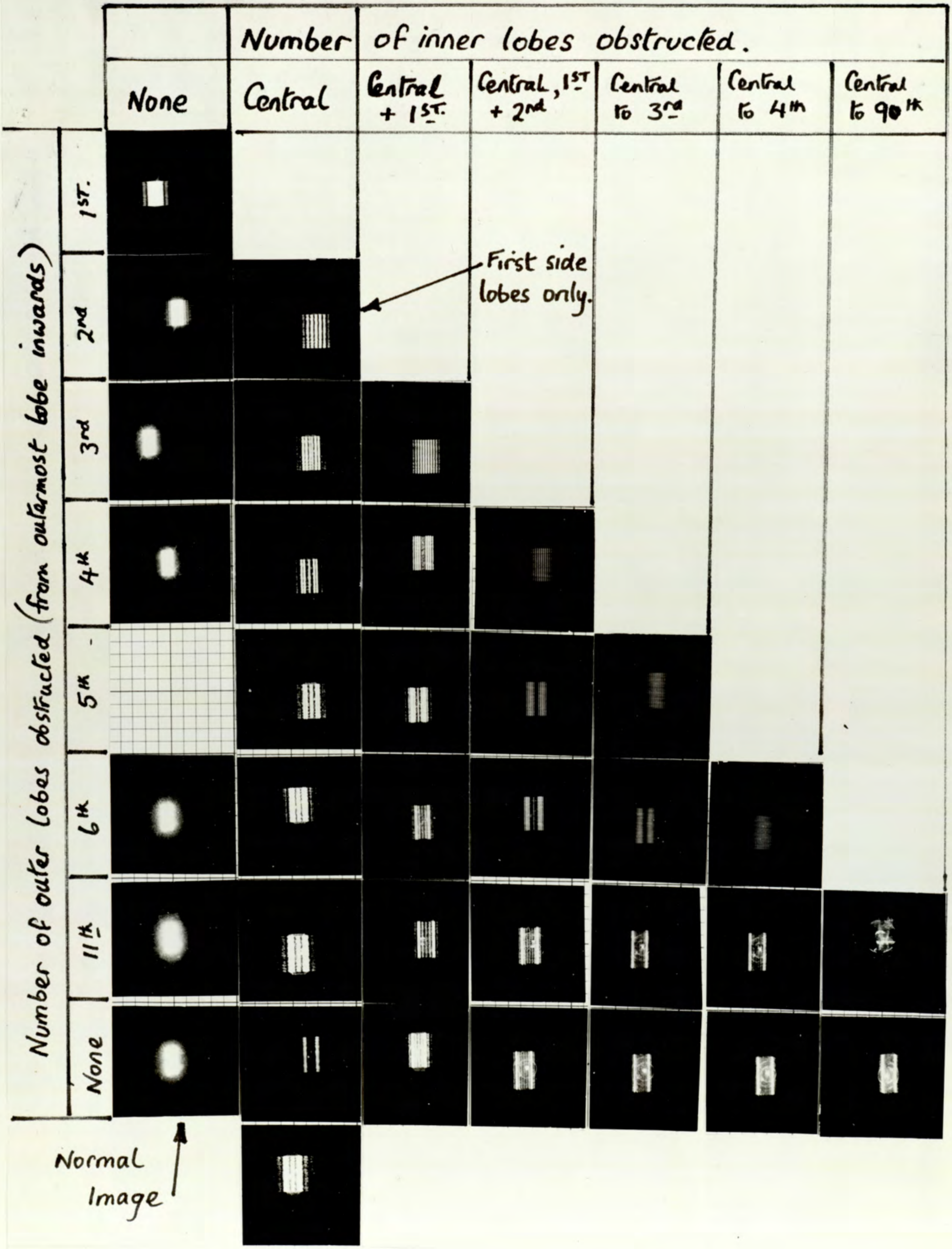
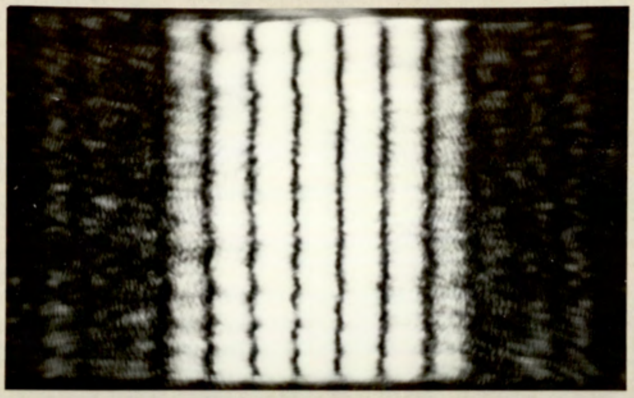
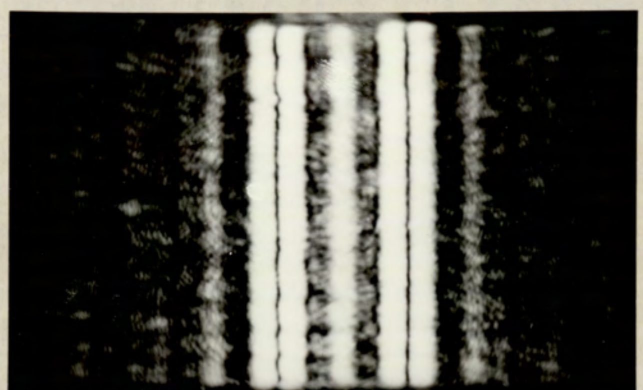


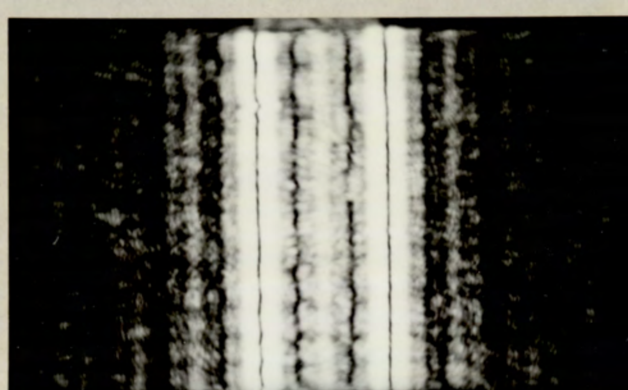
FIG A2. Variation in Image with obstruction of fourier transform



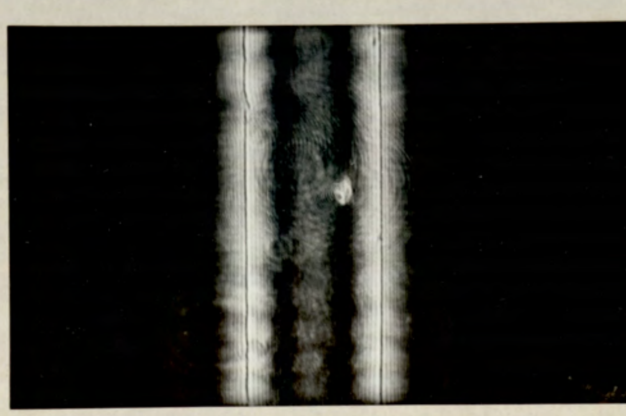
First side lobes only



First, second and third side lobes



First → fifth side lobes



All side lobes except centre

Fig. A3 Appearance of slit image with centre lobe of transform obstructed and with different outer lobes obstructed whilst if only low spatial frequencies are present, there are relatively broad lines. This is shown more clearly in Fig A3. This appears to agree with the observations made of reconstructions from holograms and of the nonlinearly reconstructed terms.

The appearance of the slit images for different parts of the fourier transform is a good illustration of the Abbé theory of microscopic imagery. Of general interest is the appearance of the image when only one pair of lobes in the transform are transmitted; the image consisting of parallel fringes whose spacing decreases as the spatial frequency of the transmitted lobes is increased, (as for the Young double slit).