NASA TECHNICAL REPORT

NASA TR R-327

C.

NASA TR R-327

PRINCIPLES OF OPTICAL DATA PROCESSING FOR ENGINEERS

by A. R. Shulman

Goddard Space Flight Center
Greenbelt, Md.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • MAY 1970

PRINCIPLES OF OPTICAL DATA PROCESSING FOR ENGINEERS

By A. R. Shulman

Goddard Space Flight Center
Greenbelt, Md. 20771

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – CFSTI price $3.00

ABSTRACT

This document is written primarily for engineers as a self-teaching text on optical data processing. Fundamentals are reviewed and expanded upon to give a clear understanding and working knowledge of the entire subject, including: optical spectrum analysis, optical correlation, photographic film characteristics, and holography. In addition, this document introduces the use of mathematics to describe the various optical operations, thus forming a background for understanding more advanced works in the field.

CONTENTS

CONTENTS (Continued)

PRINCIPLES OF OPTICAL DATA PROCESSING FOR ENGINEERS

by
A. R. Shulman
Goddard Space Flight Center

INTRODUCTION

This document is intended primarily for engineers who desire to obtain a basic understanding of optical data processing techniques. The reader is assumed to have only a basic engineering background. This document introduces the basic principles of optics necessary for the development of optical data processing techniques. Particular care has been taken to give the reader a clear understanding of the various principles by making this a document of a self-contained and self-teaching nature. To this end, no appendices or references are used and this document can be read from cover to cover by anyone with a basic engineering background.

It is believed that this document is the only one of its kind. It is intended to explain quickly and easily the basic principles of optical data processing. The need for such a document becomes apparent when one attempts to obtain a working knowledge of optical data processing. There is a current abundance of literature on all phases of the subject, but almost all of this literature is written at an advanced level (such as for theoretical physicists) with the assumption that all of the basic principles are familiar to the reader. For an engineer interested in gaining an understanding of a new field, this treatment presents a formidable problem. This document, therefore, attempts to fill the void which presently exists.

A practical treatment is used to explain basic principles, and various techniques are shown for solving problems. Many examples are used to assure thorough understanding of all topics and to insure retention of the information presented. The major part of the document is written at a level such that with a minimum technical background the basic ideas and principles of optical data processing can be understood. The remaining small portion involves mathematical discussions which are not necessary for basic understanding but will be useful as an introduction for those interested in pursuing the subject in more detail.

This document actually performs a dual purpose. First, it offers a basic understanding of the principles involved in optical data processing. Second, for those interested in obtaining a working knowledge, this document should bring the reader up to a level somewhat compatible with current technical literature. This document also indicates some of the many disciplines which must be mastered before a complete working knowledge can be achieved. For example, optics, photography, mathematics, electronics, and ultrasonics are all applied in optical data processing. The areas related to optical data processing are so varied and numerous that not all could be covered in this document (e.g. photochromic films presently appear to offer a means for producing erasable photographic images, similar to a magnetic tape recorder, and these films are not discussed herein).

CHARACTERISTICS OF LIGHT

Light is defined as *electromagnetic radiation* capable of inducing visual sensations in the eye. Like any electromagnetic radiation, light can be represented by two mutually perpendicular vibrations (electric and magnetic). In addition to being perpendicular to each other, these vibrations (field vectors) are also perpendicular to the direction of propagation. In modern theory it is an accepted fact that light energy is emitted (and absorbed) by atoms in discrete units or quanta as postulated by Einstein. This leads to the paradox that light has characteristics which exhibit both particle-like and wave-like behavior. When light interacts with particles, it can be treated as particles; and when light interacts with light, it can be treated as waves. Since optical data processing primarily involves the interaction of light with light, only the wave properties need be considered here. The quantum characteristics of light must be considered when dealing with the details of optical sources (primarily lasers) and photographic emulsions. Although these areas are important, they are somewhat beyond the scope of this treatise, which is concerned with the basic principles of optical data processing.

When two light waves intersect at a point, the magnitude of the light at that point is given by a linear combination of the two waves. Beyond the point of intersection, the light waves continue in the original directions of propagation as though there had been no intersection. The additive properties of two waves at a point are shown in Figure 1. Here we assume that the waves are of the same frequency. It is seen that the result of adding two waves at a point is dependent upon the phase difference between the two waves. The resultant sum is a minimum (zero for equal amplitudes) when the phase difference is 180° (opposite phase); maximum when the phase difference is zero (in phase); and takes values between the maximum and minimum for phase differences between 0° and 180°.

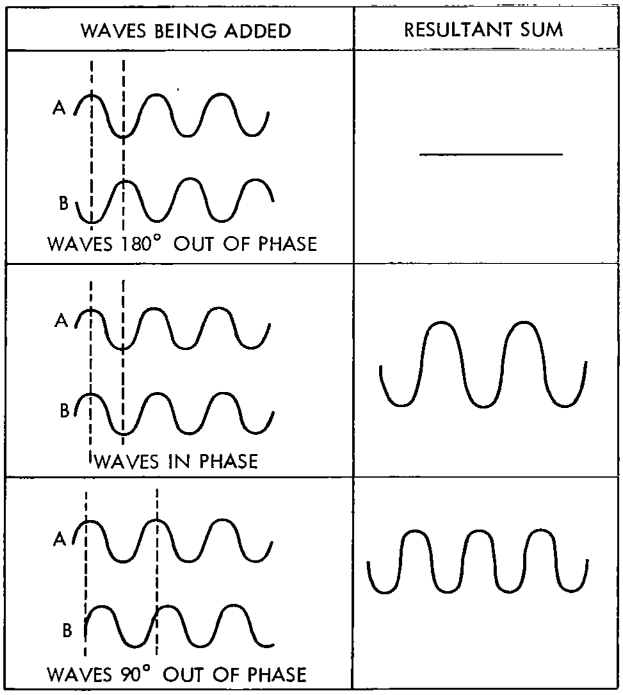
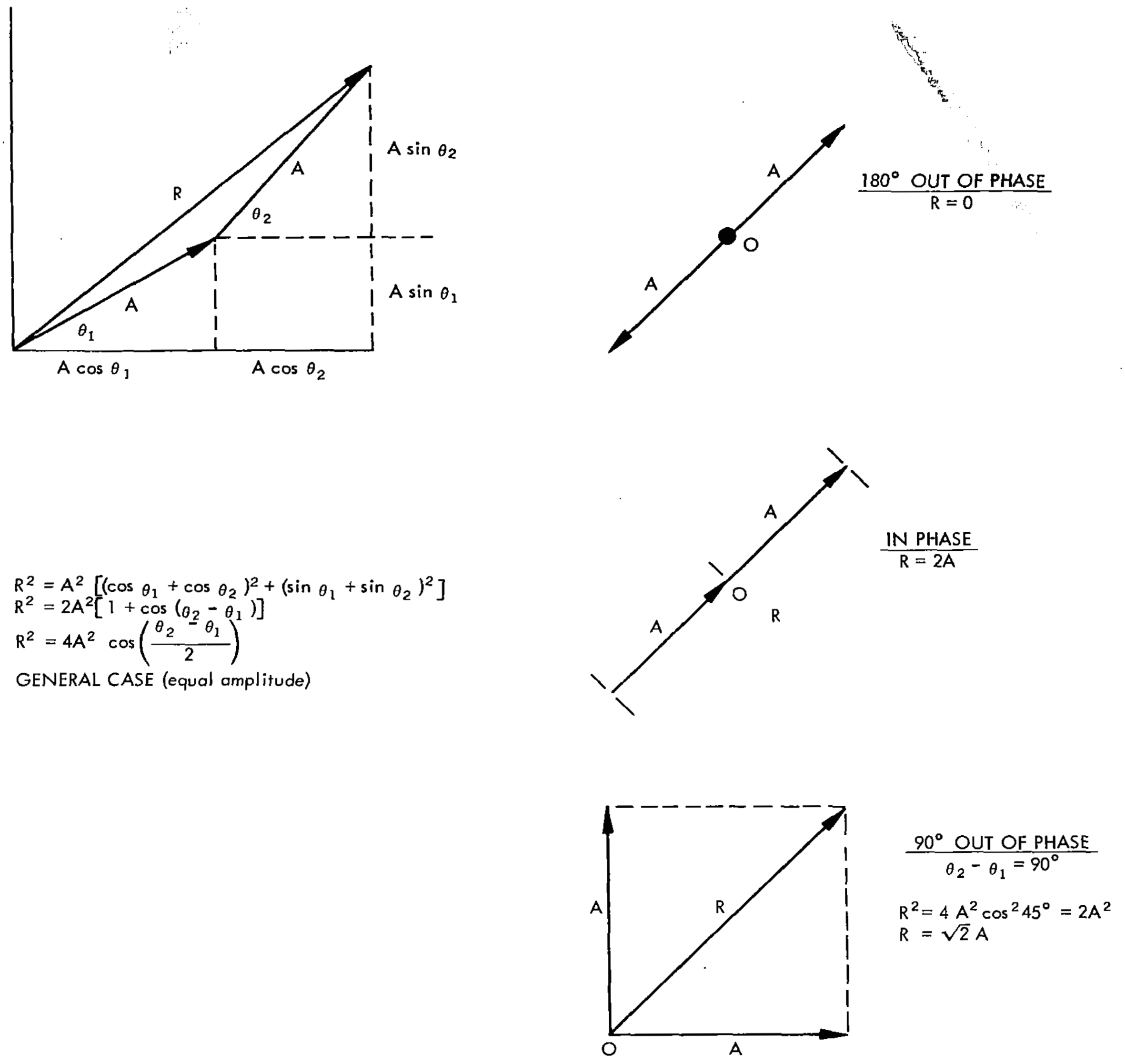


Figure 1—Addition of two waves.

For the investigation of the quantitative properties of wave addition, the use of phasor (vector) notation is convenient. Figure 2 shows the familiar method for addition of phasors. The phasor method is applied to the three examples of Figure 1. In these examples the two waves were assumed to have equal amplitudes.



$$R^2 = A^2 \left[(\cos\theta_1 + \cos\theta_2)^2 + (\sin\theta_1 + \sin\theta_2)^2\right]$$
$$R^2 = 2A^2\left[1 + \cos(\theta_2 - \theta_1)\right]$$
$$R^2 = 4A^2 \cos\left(\frac{\theta_2 - \theta_1}{2}\right)$$

GENERAL CASE (equal amplitude)

180° OUT OF PHASE
$R = 0$

IN PHASE
$R = 2A$

90° OUT OF PHASE
$\theta_2 - \theta_1 = 90°$

$R^2 = 4\,A^2 \cos^2 45° = 2A^2$
$R = \sqrt{2}\,A$

Figure 2—Phasor addition of complex amplitudes.

It is implicit in the drawing of a phasor diagram that all phasors are of the same frequency. This vector addition gives the desired quantitative results without the point-by-point addition necessary in Figure 1.

Intensity and Amplitude

The brightness of light in a beam of light can be denoted by the intensity, which is proportional to the square of the amplitude. Mathematically it is expressed as

$$I = KA^2 .$$

For our purposes, the constant K can be assumed equal to 1. The equation can then be written as

$$I = A^2 .$$

In the case of several waves at a point, the combined intensity is equal to the square of the resultant amplitude. For example, the general case shown in Figure 2 gives the result of adding two waves of equal amplitude and frequency. The resultant amplitude is given by

$$R = 2A \cos \left(\frac{\theta_2 - \theta_1}{2}\right) ,$$

while the resultant intensity is given by

$$I = R^2 = 4A^2 \cos^2 \left(\frac{\theta_2 - \theta_1}{2}\right) .$$

For two waves in phase this gives

$$I = 2^2 A^2 .$$

Repeated use of vector addition for n waves of equal amplitude (A) and the same frequency gives

$$I = n^2 A^2 . \qquad \text{(n waves in phase)}$$

For the general case in which the n waves (equal amplitude and frequency) have fixed phase relations between one another, the resultant intensity will have some value in the range from zero to $n^2 A^2$. That is,

$$0 \leq 1 \leq n^2 A^2 . \qquad \text{(n waves with constant phase relations)}$$

The upper limit $n^2 A^2$ is seen to be the special case in which all n waves are in phase. In all other cases—that is, where the n waves have fixed phase differences between one another but are not all in phase—the resultant intensity will be less than $n^2 A^2$.

The dependence of intensity on the phase relation of the combining waves is due to the $\cos \left[(\theta_2 - \theta_1)\right]/2$ term. If two waves have a random phase relation, this term must be averaged over the range 0° to 360°. For two random waves the intensity is found to be $I = 2A^2$. Repeating for n random waves gives

$$I = n A^2 . \qquad \text{(n waves with random phase relations)}$$

As an example of the intensity relations given above, consider two horns which produce sound of the same frequency and the same amplitude. The resultant sound intensity produced by these two horns together would normally be

$$I = nA^2 = 2A^2 .$$

The resultant intensity is given by the random case above because two horns normally produce sound waves which have no fixed phase relation with each other. If extra care were taken in the design of the two horns, it might be possible to construct them so that the sound produced by each was exactly the same and exactly in phase. The resultant sound intensity produced by these two special horns would be

$$I = n^2 A^2 = 4A^2 .$$

In each case a single horn produces the same amplitude of sound. It is significant to note that two horns of the usual type will produce sound twice as intense $(I = 2A^2)$ as that of a single horn, and two special horns producing sound in phase with each other will produce sound four times as intense $(I = 4A^2)$ as that of a single horn. Since the amplitude of the sound is the same in each case, there will be no appreciable difference in loudness. However, for the same power input, the resultant sound intensity of the special case (in phase) will be twice that of the usual case (random phase). This fact implies that, although there is no significant difference in loudness, the sound of the special horns can be heard at a greater distance than that of the usual horns.

Interference

From the intensity relations discussed above, it is apparent that the intensity resulting from the addition of n waves of identical frequency but with random phase relations is the sum of the intensities of the waves taken separately. However, if the n waves have a constant phase relation with respect to one another, the resultant intensity can have values between 0 and $n^2 A^2$ depending on the phase difference. This modification in intensity due to the phase of the light waves is called interference. If the combining of waves results in an intensity greater than that expected of the waves acting separately, the interference is constructive. If the resultant intensity is less than that expected of the waves acting separately, the interference is destructive. This modification of intensity due to phase can be made to produce interference patterns when certain types of light beams are combined. These interference fringe patterns consist of regions of high intensity and regions of low intensity.

Light Beams and Sources

Electromagnetic radiation, such as light, passing through a point can be represented by a wave as described above. Consider a thin transparent plate placed in the path of a light beam. The light passing through each point of the transparent plate can be represented by a wave. Since there are an infinite number of points on the surface of the plate, an infinite number of waves are required to represent the light beam. Each of these waves is traveling in the same direction as the beam. The amplitude of each wave corresponds to the amplitude of light at the point through which the wave passes. In describing light beams, it is usually sufficient to consider only a few sample waves. Figure 3 shows sample wave trains in a monochromatic light beam. The waves shown are not continuous. Since light energy is emitted in discrete units, the waves will be finite and therefore discontinuous. Normally such wave trains last for approximately 10^{-8} seconds. In Figure 3 the wave trains are shown to be all in phase. This is not normally the case; usually there is no fixed relation between one atom emitting light and another atom emitting light. There is therefore no fixed relationship between one light train and another. The wave trains in Figure 3 are shown to be all of the same frequency. This again is not the usual case. Most light sources, such as a tungsten lamp, emit light of many frequencies (polychromatic). Figure 4 shows sample wave trains of a polychromatic light beam which are more or less the usual case. Some light sources, such as mercury vapor lamps, do emit nearly monochromatic light. In these cases, other frequencies are present but emission at the characteristic frequency is predominant.

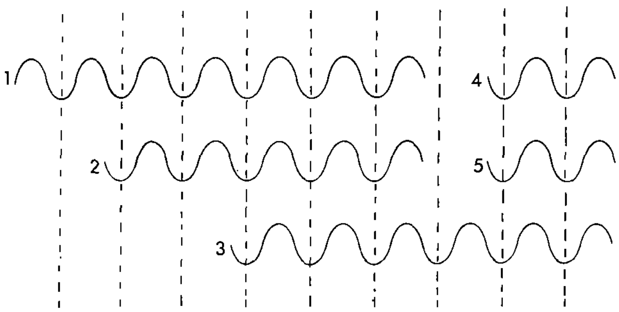


Figure 3—Monochromatic light beam—sample wave trains.

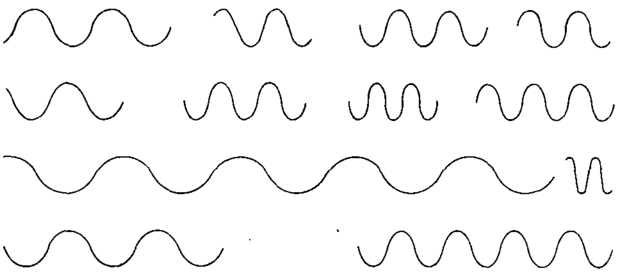


Figure 4—Polychromatic light beam—sample wave trains.

Figure 5 shows sample wave trains which are of the same frequency, in phase, and continuous. Figure 5 can represent a light beam from a laser. In a laser, the emission of light from individual atoms is controlled. In a laser, also, the emission field from an excited molecule stimulates other molecules to emit light of the same wavelength and with such phase that the emission field increases. This process occurs repeatedly and the light radiated from the laser is monochromatic, continuous, and to a high degree uniform in phase.

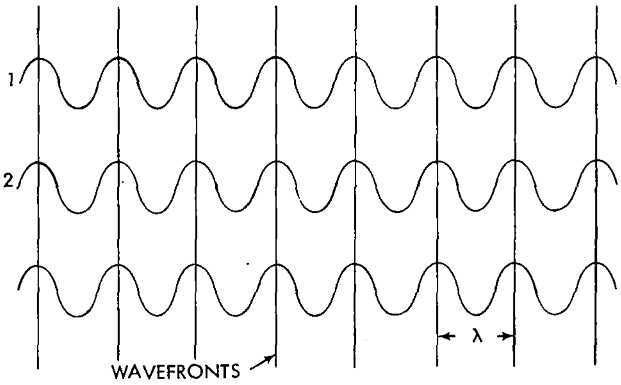
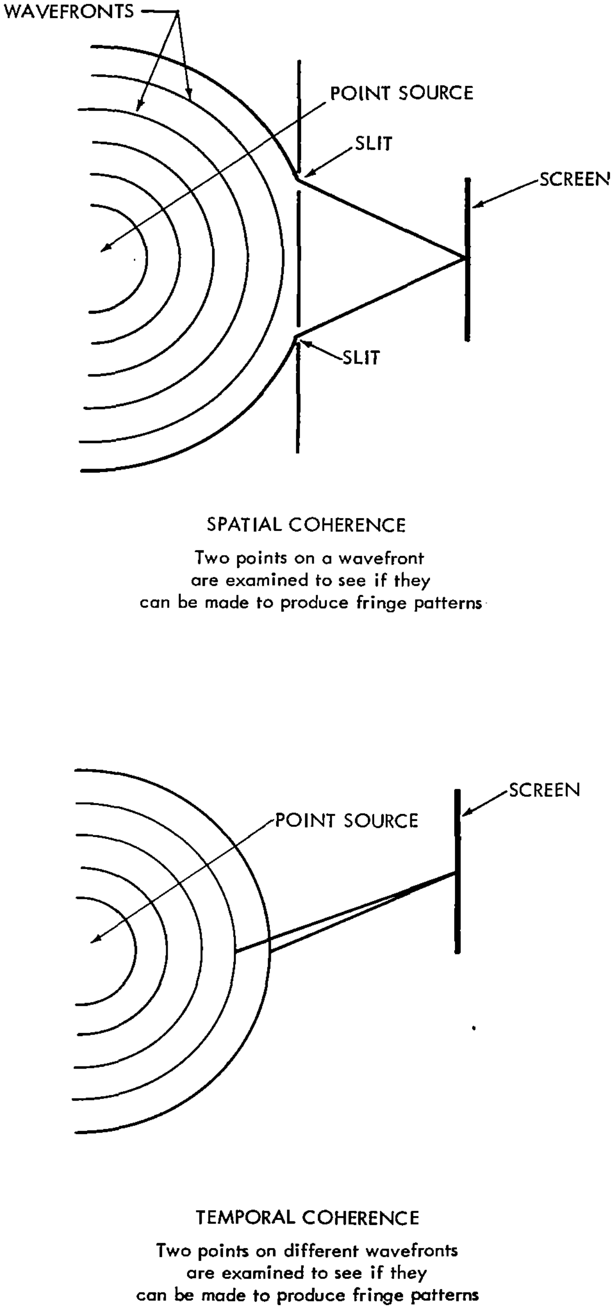


Figure 5—Monochromatic coherent light beam—sample wave trains.

The sample waves shown in Figures 3, 4, and 5 were drawn with equal amplitudes for convenience only. Amplitude variations are present in practical cases. However, instantaneous values of amplitude cannot be measured directly, and the variations of light amplitude with respect to time are neglected in optical data processing applications.

Figure 5 also shows constant-phase wavefronts. Wavefronts are surfaces representing constant phase. Usually the wavefronts are chosen to correspond to the points of maximum amplitude, as shown in Figure 5. The light waves at all points on a wavefront are in phase. In the case of a light beam from a laser, the wavefronts are approximately plane wavefronts, as shown in Figure 5. A point source produces spherical wavefronts. A wavefront moves in the direction of the light waves it represents. For example, if the light waves are propagating to the right in Figure 5, the wavefronts also move to the right. Light that is composed of plane wavefronts is called collimated light.



SPATIAL COHERENCE

Two points on a wavefront are examined to see if they can be made to produce fringe patterns

TEMPORAL COHERENCE

Two points on different wavefronts are examined to see if they can be made to produce fringe patterns

Figure 6—Comparison between spatial and temporal coherence.

Coherent Light Beams

If two light beams can be combined to form interference patterns, they are said to be mutually coherent beams. Light beams from two different sources (except lasers) cannot be combined to produce interference patterns, since the light-wave trains emitted are independent and of random phases. Splitting a monochromatic light beam to produce two beams will give two light beams which are mutually coherent. This is possible because any discontinuities in one beam will appear also in the other beam.

The term "coherence" implies phase correlation between points in a light beam. There are two types of coherence, namely, spatial and temporal coherence. Figure 6 shows a comparison between spatial and temporal coherence. The sources (point sources) used in Figure 6 produce spherical wavefronts.

Spatial Coherence

Spatial coherence is a measure of the phase correlation between two points on the same wavefront. Spatial coherence can be measured by passing a wavefront through two slits, as shown in Figure 6. If the light coming through the slits forms an interference pattern, the light source is said to display spatial coherence. The contrast of the fringe pattern is a measure of the spatial coherence. A 100-percent spatially coherent light beam will produce fringe patterns which vary in intensity from zero to maximum intensity (100-percent contrast). A non-coherent beam will not form a fringe pattern. Therefore the contrast of a fringe pattern of a non-coherent beam of light is zero (actually no fringe pattern).

Since it is an established fact that electromagnetic radiation is a continuous phenomenon (no abrupt discontinuities), there is always some spatial coherence between two closely spaced points in a light beam. Hence the cross-section of a monochromatic light beam passed through a pinhole will be nearly 100-percent spatially coherent. For this reason, early studies of interference patterns were made by passing a monochromatic light beam through a pinhole to obtain spatially coherent light.

Laser light beams have wave trains which are all in phase. When any two points on a wavefront in a laser beam are tested for spatial coherence, very high-contrast patterns are produced. These high-contrast patterns indicate the expected high degree of spatial coherence in a laser beam.

Temporal Coherence

As shown in Figure 6, the degree of similarity between wavefronts separated in time is a measure of temporal coherence. The experiment to determine temporal coherence is usually based on the interference of two successive wavefronts of light with each other.

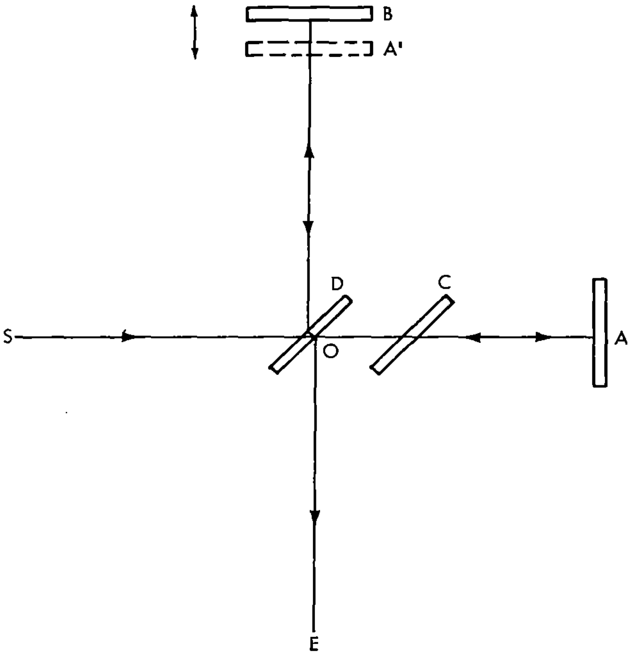


Figure 7—Michelson Interferometer.

The Michelson Interferometer is a device which can be used to determine the degree of temporal coherence of essentially monochromatic light sources. The Michelson Interferometer (Figure 7) causes light from a source S to be split into two beams at point O by the action of the lower surface of glass plate D (which is half-silvered). One of these light beams is formed by light passing through the silvered back surface of plate D. This beam also passes through plate C and falls on the front surface of mirror A. The portion of light reflected by the back surface of plate D forms the second beam, which falls on the front surface of mirror B. The reflected beams from the two mirrors (A and B) recombine at the half-silvered surface of D and enter a detector at point E).

The beam which goes to mirror B passes through plate D three times, while the beam going to mirror A passes through plate D only once. The compensating plate C is usually inserted in a light beam going to mirror A in order to make the arrangements symmetrical. Mirror B is mounted on a slide which can be moved by a micrometer screw in the directions shown by the arrows.

An observer at point E sees the image of mirror A in mirror D at A'. That is, light from mirror A appears to the observer to be originating at A'. By moving mirror B with the micrometer screw, variations in intensity can be produced by interfering the light from mirrors A and B which is going to the observer at E. The difference in path lengths AO and BO permits different wavefronts from the original light source to interfere with each other.

The contrast of the intensity variations was defined by Michelson as $(I_{Max} - I_{Min})/(I_{Max} + I_{Min})$, where: I_{Max} is the maximum intensity, produced when the light beams from mirrors A and B interfere constructively, and I_{Min} is the minimum intensity, produced when the light beams from mirrors A and B interfere destructively.

The resulting interference variations from a Michelson Interferometer are an indication of the temporal coherence of the light source. A plot of the contrast of the intensity variations as a function of the position of moving mirror B is a curve of temporal coherence of the light source. The temporal coherence of a light source is defined as the distance between the half-power points on the temporal coherence curve. The more nearly monochromatic the light source, the higher the contrast of the fringes and the greater the temporal coherence. As an aside, it is interesting to note that the temporal coherence curve is (in mathematical terms) the autocorrelation function of the light source.

Polarized Light

It has already been stated that light is transmitted like any other electromagnetic radiation. The only defined restriction on electromagnetic radiation is that the two mutually perpendicular vibrations (electric and magnetic) must also be perpendicular to the direction of propagation. Ordinary light (such as that received from the sun) can have electric vibrations in random directions perpendicular to the direction of propagation. When the direction of the electric field vibrations is not random but can be specified somewhat regularly, the light is said to be *polarized* with respect to the electric field vibrations. Some examples of polarization are illustrated in Figure 8.

In Figure 8 the diagrams on the left represent the envelope of the electric field vectors at an instant of time. A line drawn from a point on the envelope perpendicular to the Z-axis represents the electric field at the point of intersection on the Z-axis. The distance from the Z-axis to the envelope (length of line perpendicular to Z-axis) represents the amplitude of the electric field, and the direction of the line represents the direction of the electric field. The diagrams on the right demonstrate the particular characteristics of the types of polarization illustrated. These diagrams illustrate the electric field characteristics in a plane perpendicular to a point on the Z-axis.

The first diagram in Figure 8 shows a vertical-plane-polarized wave (also called linear polarized). In this wave, the electric field vibrations are confined to the YZ plane. The diagram on the

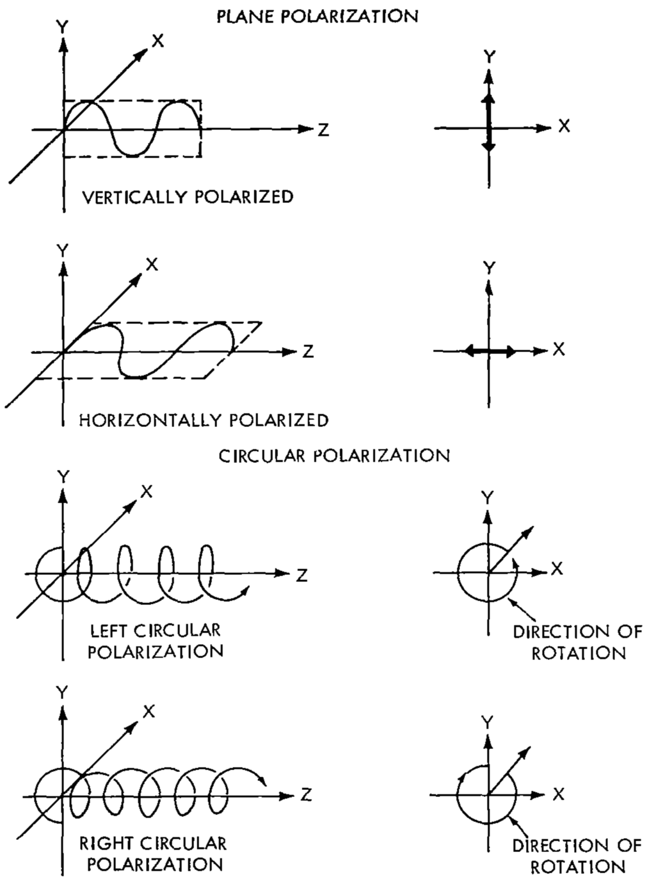


Figure 8—Plane and circular polarization.

left illustrates that the electric field varies in amplitude with respect to Z. The diagram on the right indicates that the amplitudes are restricted to up-and-down variations in the YZ plane.

The second diagram in Figure 8 shows a horizontal-plane-polarized wave. This wave has characteristics identical to those of the vertically polarized waves except that the vibrations are restricted to the XZ plane. It is to be noted that, as shown in these two examples of "plane polarization," a plane-polarized wave varies in amplitude only. The plane determined by the Z-axis (direction of propagation) and the direction of the electric field is called the plane of polarization.

The third and fourth diagrams in Figure 8 illustrate the two types of circular polarization. The envelopes can be traced out by rotating a fixed-length line about the Z-axis while simultaneously moving along the Z-axis. Since the rotation can be in either of two directions, we can distinguish between right- and left-circular polarization. The diagrams on the right indicate that at any given point on the Z-axis the electric field is constant in amplitude but its direction changes as the vector rotates about the Z-axis. The tip of the vector sweeps out a circle as it is rotated. This characteristic of the electric field gives rise to the descriptive term "circular polarization."

It is to be noted that the two types of polarization shown in Figure 8 are the special cases of amplitude variation only or direction variation only. If both the amplitude and direction change, elliptical curves will be swept out by the electric field vector at a point on the Z-axis. This is called elliptical polarization. It should be noted that plane and circular polarization are special cases of elliptical polarization.

There are many methods of producing polarized light, but probably the most important in optical data processing is the laser. The laser produces essentially a highly collimated, monochromatic, coherent, and *polarized* light beam. The laser produces linear polarized light (represented by either diagram 1 or 2 in Figure 8) and a monochromatic coherent collimated light beam (Figure 5).

A light beam is said to be polarized when the electric field vibrations at all points in the beam have the same polarization. When the individual waves in the light beam do not have the same polarization, the light beam is said to be unpolarized. Figure 9 shows sample representations of the waves in an unpolarized light beam.

It is possible to mix polarized light with unpolarized light and have a resultant light beam partially polarized. The degree of polarization will be dependent upon the proportion of polarized wave trains to unpolarized wave trains in a light beam. Degree of polarization can be defined as

$$P = \frac{I_{Max} - I_{Min}}{I_{Max} + I_{Min}} ,$$

where I_{Max} and I_{Min} are maximum and minimum intensities of light passed by a rotating polarizer.

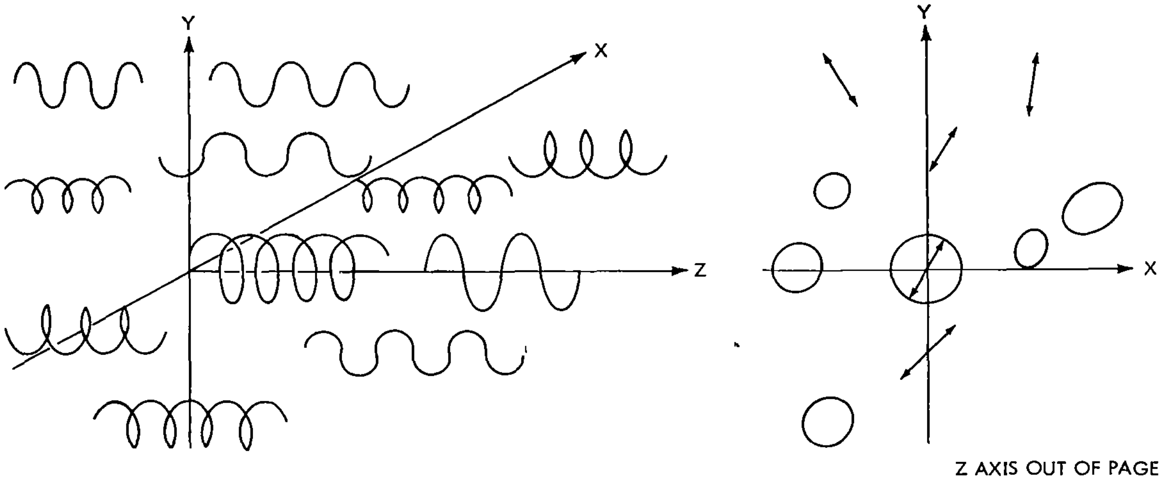
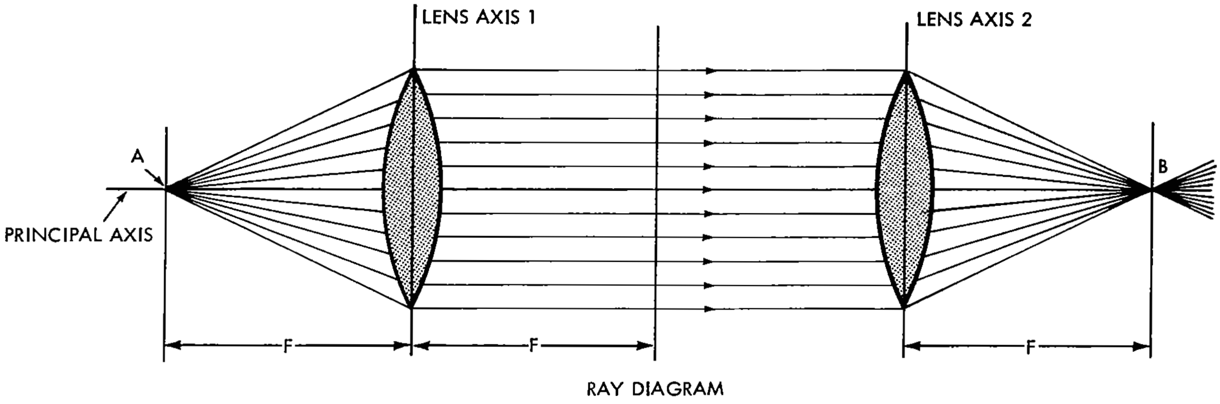


Figure 9—Unpolarized light beam.

Lens Fundamentals

Figure 10 shows two representations of the same phenomenon. The upper diagram of Figure 10 is called a ray diagram. A ray is a line drawn in the direction of wavefront propagation. In this diagram there is a point source of light at point A. When point A is at the focal point to the left of the lens, as shown in Figure 10, the rays of light emerging on the right side of the lens will be parallel and travel horizontally. These parallel rays can be focused to point B by a second lens as shown.

The lower portion of Figure 10 shows wavefronts radiating out as concentric circles from point source A (which is at the focal point of the lens). The lens impedes the wavefront propagation through it sufficiently so that upon emerging the wavefronts become plane wavefronts travelling horizontally to the right. These plane waves are refracted by a second lens and converge to the focal point B of the second lens, as shown. The results obtained from a ray diagram are the same as those obtained from a wavefront diagram, but in practice the ray diagrams are



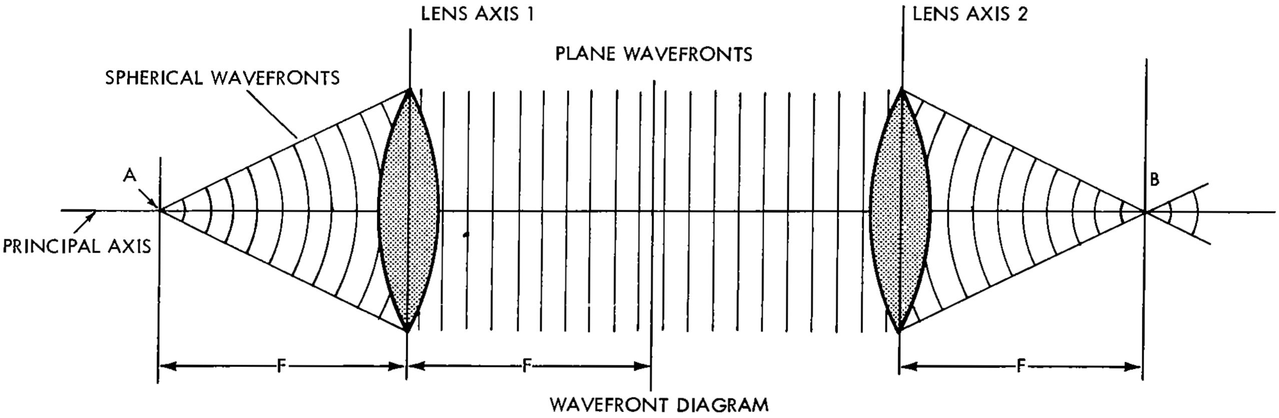


Figure 10—Ray and wavefront diagrams.

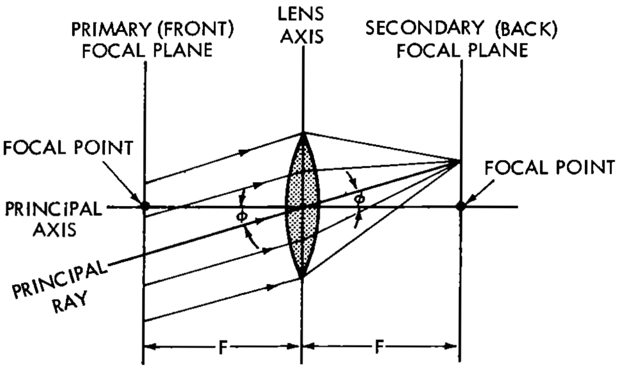


Figure 11—Incident parallel rays focused in secondary focal plane.

usually more practical than the wavefront-type diagram.

Figure 11 introduces some lens terminology. The *principal axis* is the line joining the center of curvature of the two faces of the lens. The *focal points* of a thin lens may be taken to coincide with the centers of curvature of the lens. A *focal plane* is a plane perpendicular to the principal axis at a focal point. The focal plane contains the images of objects which are an infinite distance from the lens (incident rays parallel). The *principal ray* is a ray passing through the lens without deviation (zero refraction). Incident parallel rays making an angle ϕ with the principal axis will converge to a point in the back focal plane of a lens, as shown in Figure 11. The point in the back focal plane to which the rays converge is determined by the angle ϕ.

Diffraction

When waves pass through an aperture or pass the edge of an obstacle, they spread beyond the limits of the geometric shadow. This phenomenon is known as diffraction. Huygens' principle states that each point on a wavefront can be considered as a new source of a wave. With this in mind, the diffraction phenomena can easily be explained. Figure 12 shows a portion of the light from a monochromatic point source being blocked by a baffle which has in it a small aperture. By Huygens' principle, the small aperture can be considered as a new light source.

T. Young performed an experiment similar to that shown schematically in Figure 13. In his experiment Young passed sunlight through a pinhole such as A in Figure 13. By Huygens' principle this pinhole acts as a new point source, and from our discussion of spatial coherence it is apparent that this new point source (Pinhole A) radiates light waves with a high degree of spatial coherence. By using a pinhole, Young was therefore able to produce a point source with a high degree of spatial coherence. At a considerable distance from A, the spatially coherent light is passed through two more pinholes (B and C). In the previous discussion of spatial coherence it was stated that a coherent light beam can be split into two beams

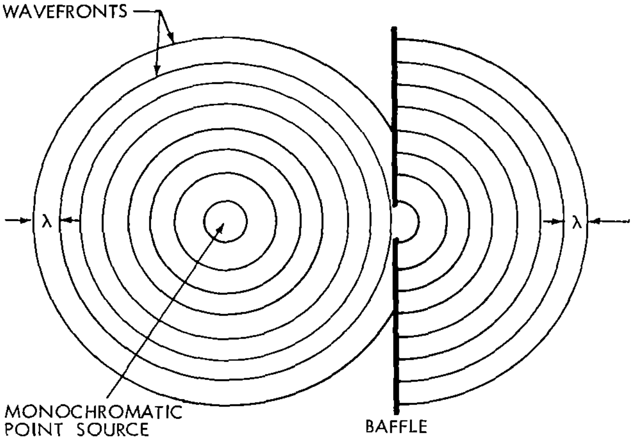


Figure 12—Diffraction of light at a small aperture.

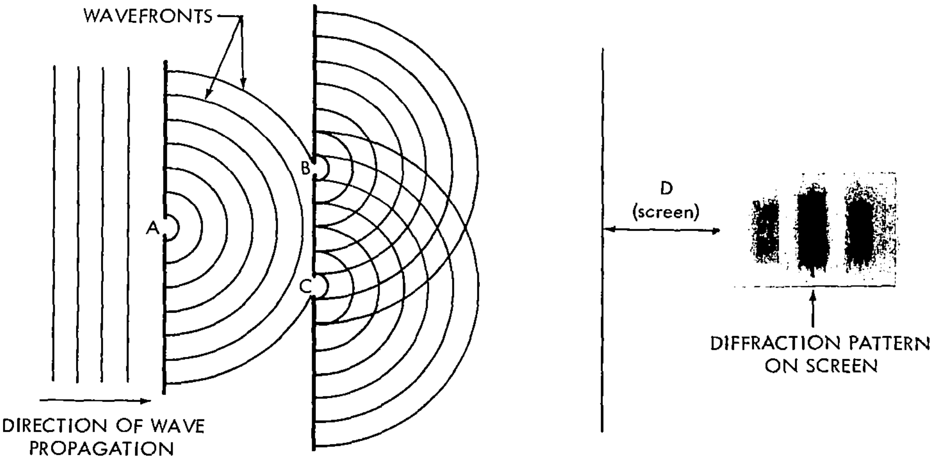


Figure 13—Experimental arrangement for Young's experiment.

which can be made to interfere with each other. In effect this is done by the pinholes B and C. Thus, by using pinholes in the arrangement shown in Figure 13, Young was able to produce two mutually coherent (spatially) light sources for his investigation of interference patterns.

When discussing Young's experiment, it is convenient to replace the pinholes by narrow slits and use a monochromatic source of light. The narrow slits will produce cylindrical wavefronts instead of the spherical wavefront produced by the pinholes. Figure 13 can be used to represent the cylindrical wavefronts emerging from the slits. In either case the wavefronts from the sources B and C will interfere and form an interference pattern consisting of parallel dark and light straight-line bands (shown in Figure 13).

Diffraction from Multiple Slits

The basic operation of an optical processor is easily understood after considering a light beam with plane wavefronts which is diffracted by equally spaced slits. When a plane wavefront from a monochromatic coherent light source passes through a series of slits (Figure 14), each slit will act as a new source of a cylindrical wave in accordance with Huygens' principle. The successive wavefronts spread out on the right side of the slits as shown. As shown in Figure 13, the wavefronts from adjacent slits interfere constructively and destructively with each other.

Zero-Order Wavefront Formation

Figure 14 shows the construction of zero-order wavefronts. These wavefronts are formed by planes parallel to the plane of the slits and tangent to the cylindrical wavefronts of the slits. The zero-order wavefronts travel in the same direction and have the same spacing (one wavelength, λ) as the original plane wavefronts. In Figure 14 it can be seen that as the cylindrical wavefronts travel out (in the direction perpendicular to the slit plane) they become more and more plane, as shown by the darkened portions of the dotted lines representing the wavefronts. These plane wavefronts are called the zero-order wavefronts.

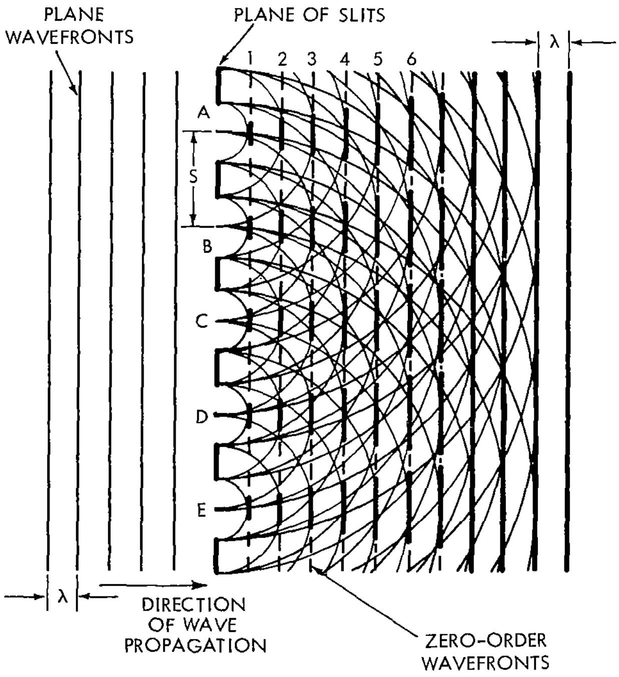


Figure 14—Formation of zero-order wavefronts.

First-Order Diffraction Wavefront Formation

Figure 15 shows the formation of first-order diffraction wavefronts. The cylindrical wavefronts also merge at large distances from the slits into plane waves traveling at an angle upward or downward from the horizontal, as shown. The first-order up wavefronts are formed by planes tangent to the mth wavefront of slit A, the (m + 1) wavefront of slit B, (m + 2) wavefront of slit C, etc. Similarly the first-order down wavefronts are formed by planes tangent to the mth wavefront of slit A, the (m - 1) wavefront of slit B, (m - 2) wavefront of slit C, etc. By diffraction theory and trigonometry, it can be shown that the sine of the angle of the direction of propagation is related to the wavelength of the light and the spacing between the slits S in accordance with the formula

$$\sin\theta = \frac{\lambda}{S},$$

where

λ = wavelength of light,

S = slit spacing,

θ = angle of wavefront propagation from horizontal.

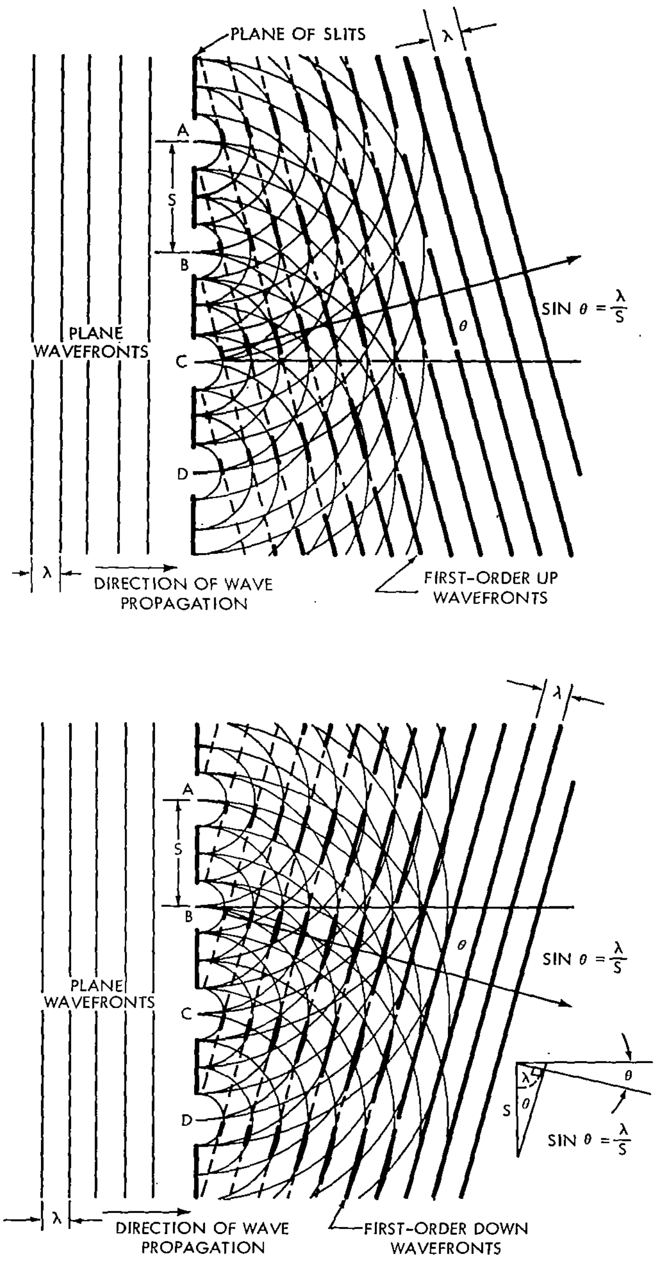

PLANE OF SLITS
λ
A
S
B
C
D
PLANE WAVEFRONTS
θ
SIN θ = λ/S
λ
DIRECTION OF WAVE PROPAGATION
FIRST-ORDER UP WAVEFRONTS
λ
A
S
B
C
D
PLANE WAVEFRONTS
θ
SIN θ = λ/S
λ
S
θ
θ
SIN θ = λ/S
λ
DIRECTION OF WAVE PROPAGATION
FIRST-ORDER DOWN WAVEFRONTS

Figure 15—Formation of first-order wavefronts.

Second-Order Diffraction Wavefront Formation

Figure 16 shows the formation of second-order diffraction wavefronts. Second-order wavefronts are formed by planes tangent to the mth wavefront of A, the (m ± 2) wavefront of B, (m ± 4) wavefront of C, etc. The cylindrical wavefronts merge at large distances from the slits into plane waves as shown. The angle from the horizontal at which the second-order wavefronts travel is given by

$$\sin\theta = \frac{2\lambda}{S},$$

where

λ = wavelength of light,

S = slit spacing,

θ = angle of wavefront propagation from horizontal.

nth-Order Diffraction Wavefront Formation

From the above discussion it can be seen that there are 3rd-order diffraction wavefronts, 4th-order . . . to n orders of diffraction wavefronts, each existing simultaneously. The angles from the horizontal at which the nth-order diffraction wavefront is propagating are given by the formula, $\sin\theta = n\lambda/S$, where n is the order of the diffraction wavefront.

It has been shown how plane waves are diffracted by multiple slits to form the various orders of wavefronts. These diffraction wavefronts are mutually coherent (originate from the same source) and can interfere with one another. Consider a screen parallel to the zero-order wavefronts. Each order of diffraction wavefronts would illuminate the screen uniformly if it were the only order present. However, when two or more wavefronts (of different order) are present, they will not have the same relative phase with respect to each other at every point on the screen. Since the relative phase relation determines how the wavefronts combine as described earlier, a variation in light intensity will occur on the screen. In the case of slits, this interference pattern will consist of light and dark stripes. In effect, the same pattern

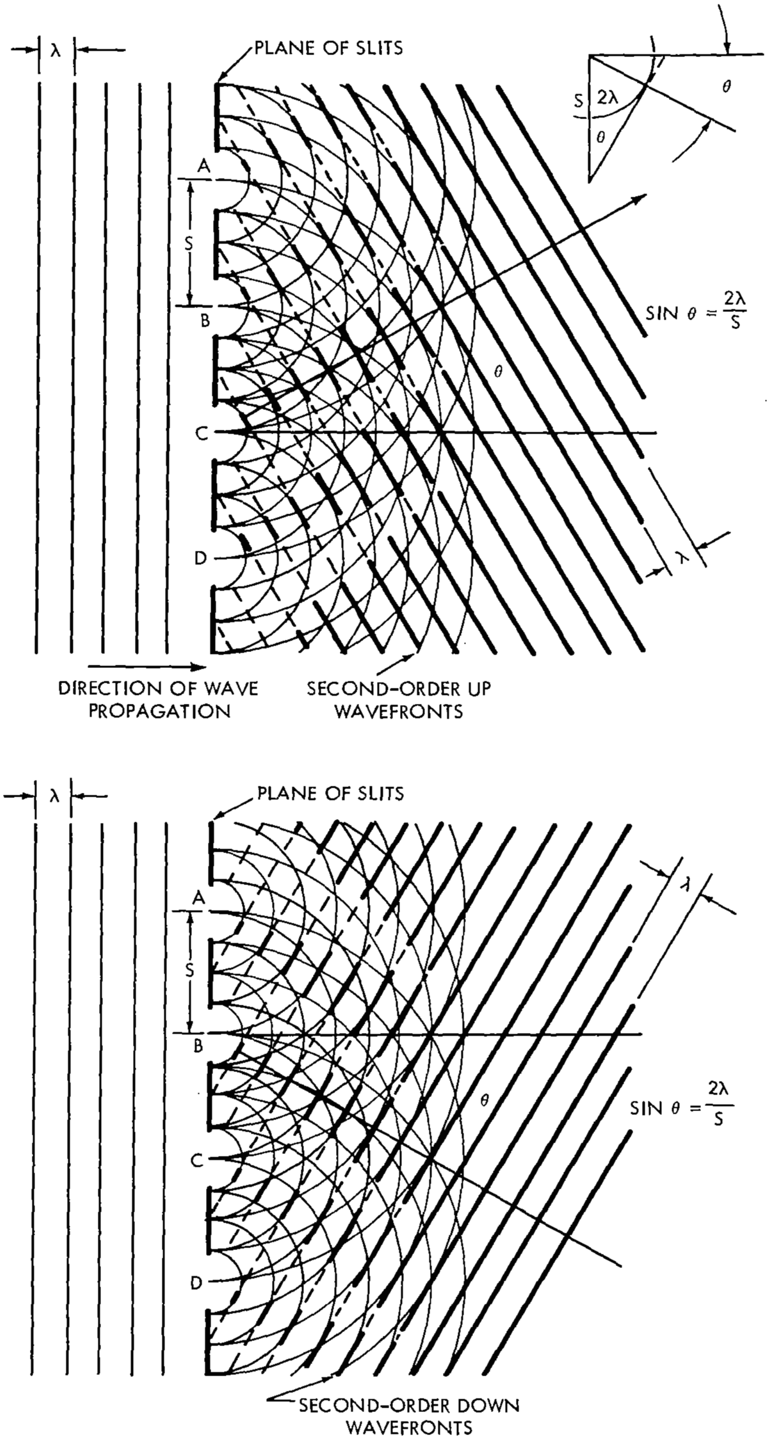

λ
PLANE OF SLITS
A
S
B
C
D
S
2λ
θ
θ
θ
SIN θ = 2λ/S
λ
DIRECTION OF WAVE PROPAGATION
SECOND-ORDER UP WAVEFRONTS
λ
PLANE OF SLITS
A
S
B
C
D
θ
λ
SIN θ = 2λ/S
SECOND-ORDER DOWN WAVEFRONTS

Figure 16—Formation of second-order wavefronts.

could be produced by using separate collimated light sources for each order of diffraction wavefronts. These separate sources would have to be mutually coherent and incident at the same angle as the corresponding order of diffraction wavefronts. Since this interference pattern is formed by wavefronts created by diffraction, it is called a diffraction pattern.

It can also be shown how the orders of wavefronts caused by the diffraction of light through slits can be focused by a lens. It was shown (Figures 14, 15, and 16) that orders of diffraction wavefronts will be set up when plane wavefronts illuminate a set of slits. The directions of propagation of the zero-order wavefronts and of the first-order up and down wavefronts are shown in Figure 17. These directions are indicated by rays and can be compared to Figures 14, 15, and 16. The various orders of diffraction wavefronts produced by the slits are brought to focus as shown in Figures 10 and 11. To be more specific, the zero-order wavefront is brought to focus in the focal plane in a manner similar to that shown for Lens 2 in Figure 10, where parallel horizontal rays incident upon Lens 2 are shown to be brought to a focus at the back focal point B. The zero-order diffraction wavefronts can be represented by parallel horizontal rays and therefore can be brought to a focus at the back focal point of a lens. This focusing of zero-order wavefronts is illustrated in Figure 17. The first-order up and down wavefronts focus in the focal plane in the manner similar to that shown in Figure 11. Figure 17, therefore, combines in one diagram the various effects that have been discussed. It is to be noted that when the diffraction pattern is imaged in the back focal plane of a lens, the imaged pattern consists of points of light on a line perpendicular to the direction of the slits.

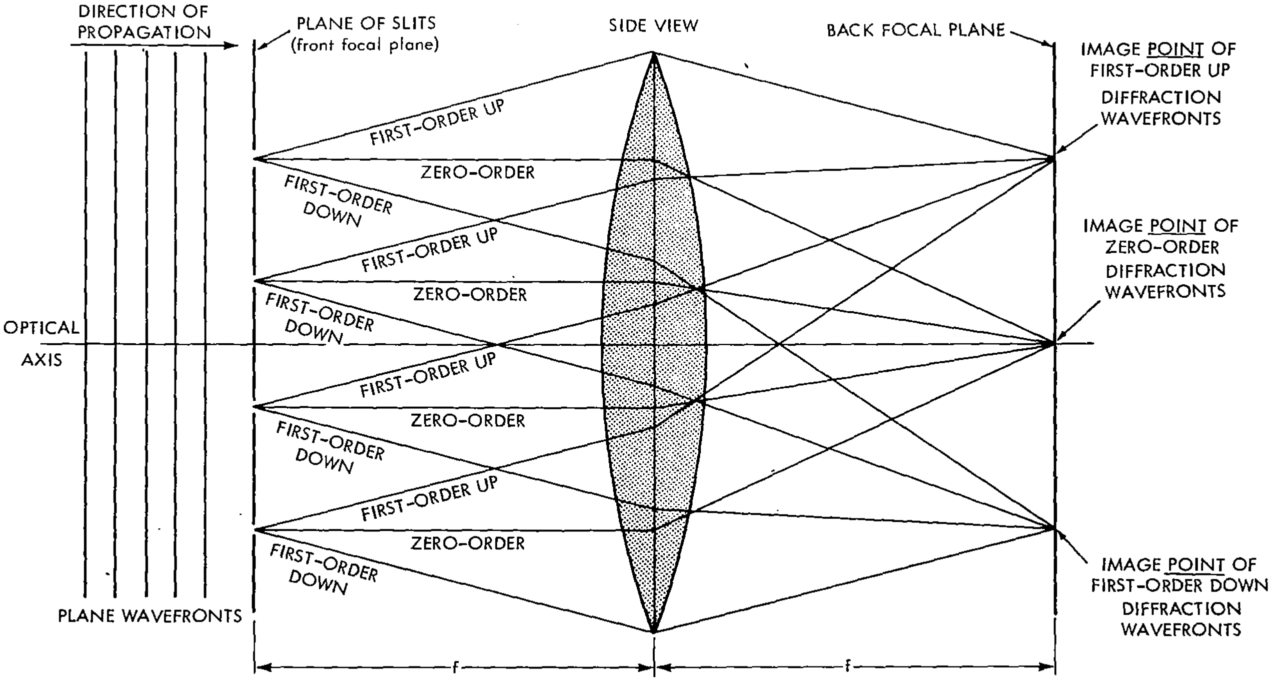

DIRECTION OF PROPAGATION
PLANE OF SLITS (front focal plane)
SIDE VIEW
BACK FOCAL PLANE
IMAGE POINT OF FIRST-ORDER UP DIFFRACTION WAVEFRONTS
FIRST-ORDER UP
ZERO-ORDER
FIRST-ORDER DOWN
IMAGE POINT OF ZERO-ORDER DIFFRACTION WAVEFRONTS
OPTICAL AXIS
IMAGE POINT OF FIRST-ORDER DOWN DIFFRACTION WAVEFRONTS
PLANE WAVEFRONTS
f
f

Figure 17—Focusing of diffraction wavefronts.

The principle shown in Figure 17 can be interpreted as a form of frequency analysis (spectrum analysis). For example, consider the first-order diffraction wavefronts. The distance of the first-order image points (formed by the lens) from the principal axis is inversely proportional to the slit spacing. If the space between slits is considered as one period (or one spatial wavelength), the distance of the first-order points from the principal axis can then be considered proportional to the spatial frequency of the slits. This interpretation will be discussed in more detail later and, in fact, will be used as the basic principle in most applications of optical data processing methods.

Zone Plate

A zone plate is a screen made up of transparent and opaque zones. Figure 18 shows a typical zone plate. We have seen from Figure 10 how parallel wavefronts are refracted by a lens into spherical wavefronts whose center is at the focal point of the lens. The zone plate can act on parallel wavefronts much the same as a lens does. The wavefronts reaching the focal point are wavefronts of light which have the same phase, i.e. the light waves reaching the focal point are all in phase.

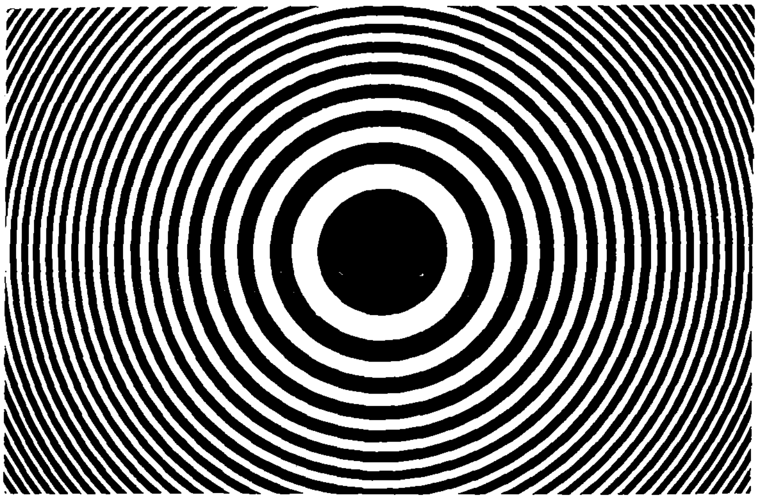


Figure 18—Zone plate.

A zone plate is constructed by drawing concentric circles with radii proportional to the square root of natural numbers. Calling the center circle zone 1, each of the succeeding larger bands is numbered in order. If the odd (or even) bands are colored so as to be opaque and the even (or odd) zones are made transparent, the screen thus formed will transform a plane wavefront into a spherical wavefront whose center of curvature is the focal point. The focal length of such a zone plate is

$$F = \frac{R_n^2}{n\lambda},$$

where

F = focal length of the zone plate

R_n = radius of the nth zone,

n = number of zones,

λ = wavelength of the light used.

The opaque areas block any light which would produce out-of-phase waves at the focus. Since all light reaching the focal point is in-phase, the intensity of light at the focal point is greater with the zone plate present than without it. This increase in intensity happens at the focal point even though portions of the light are blocked by the opaque areas.

Another way of looking at a zone plate is to consider it as slits. We have seen how slits can diffract light, causing wavefronts of different orders to be formed. A zone plate can be considered to act in very much the same way as the slits, except in this case the slits are not straight but curved. The curved slits cause spherical wavefronts to be formed, instead of the plane wavefronts as shown in Figures 14 through 16. Figure 17 shows the direction in which the straight slits cause light to be diffracted. The fact that in a zone plate the slits are circular causes the diffraction pattern to be circular, resulting in the formation of spherical wavefronts.

As an aside, it would appear that additional intensity could be obtained at the focal point if phase correction were used instead of blocking the light in opaque areas. In other words, increased intensity could be obtained at the focal point if the light in opaque areas were phase-shifted instead of completely blocked.

Color

A great many studies have attempted to determine the mechanics of how human beings perceive color. These studies have formed a basis of our present color photography techniques. In order to understand color photography, it is necessary to understand the results of thest studies.

Initial studies relating to color perception determined that the visible spectrum includes the wavelength range from approximately .4 microns to approximately .7 microns. (Refer to Table 1 for conversion factors.)

Table 1

Conversion Factors.

MULTIPLY NUMBER OF TO OBTAIN NUMBER OF BY	Ångstroms	Millimicrons	Microns	Millimeters	Centimeters
Ångstroms	1	10	10^{4}	10^{7}	10^{8}
Millimicrons	10^{-1}	1	10^{3}	10^{6}	10^{7}
Microns	10^{-4}	10^{-3}	1	10^{3}	10^{4}
Millimeters	10^{-7}	10^{-6}	10^{-3}	1	10
Centimeters	10^{-8}	10^{-7}	10^{-4}	10^{-1}	1

It is well known that sunlight or white light can be broken down into the spectral colors it contains (e.g. rainbows, oil films on water, soap bubbles). One of the most familiar methods of breaking white light into its spectrum is the use of a prism. A prism consists of two surfaces included at some angle so that the refraction produced at the first surface is not canceled by the second surface but is further increased. It is this process which increases the chromatic dispersion so that the spectrum can be seen. Color is determined by the frequency of vibration (or the associated wavelength) of the light. White light is composed of many wavelengths of light blended together. These wavelengths give rise to an infinite number of hues which make up the spectrum. These hues are usually grouped together broadly into six principal colors: red, orange, yellow, green, blue, violet. The angle of refraction of a prism varies with wavelength, shorter wavelengths being refracted more. No definite relationship covering all prisms exists between the wavelengths and the refraction angles. This means that prisms made of different substances will spread out the component colors of the spectrum to somewhat different extents. The normal eye can discern differences in wavelengths of .005 microns. Differences in hues do exist between colors whose wavelengths differ by less than .005 microns, but it is not possible for the eye to distinguish this difference. Each hue is determined by a different frequency or wavelength of light. It would seem, therefore, that in order for the eye to determine the color it would need a receptor in its sensitive area for each frequency of light that it can discern. This would mean that the eye would have in its sensitive area an infinite number of light-sensitive elements, each of which would be sensitive to light of but a single wavelength. This concept must be rejected as unrealistic.

A more acceptable concept which explains how color is detected by the human eye is to accept a relationship between the eye and the brain. The sensitive receptors which make up the retina of the eye consist of two main types, rods and cones. Light falling on these receptors produces electrical impulses which are transmitted along the optic nerves to the brain. It is the brain which interprets these many electrical impulses into the sensation of sight.

The receptors can be considered to form three separate systems, each responding to one-third of the spectrum, i.e.

System 1 responds to violet and blue,

System 2 responds to green and yellow,

System 3 responds to orange and red.

Each of these systems has a broad response, i.e. the response of each system overlaps the adjacent system. Yellow light will excite the green-yellow system and also the orange-red system. It is the brain that interprets the electrical signals from each of the systems as a single response of yellow. It is possible to excite the green-yellow system with a green color and the orange-red system with a red so that the combination of the two colors is interpreted by the brain as yellow. The brain therefore can interpret true yellow impulses as yellow or can interpret a combination of green and red as yellow. These facts have been verified experimentally.

Experiment has shown that the simple mathematical equation shown below relates any four distinct colors as determined by the eye:

$$s(B) + t(C) + u(D) \equiv r(A) ,$$

where A, B, C, and D are four colors; r, s, t, and u are constants relating to units or amounts of each color; and "≡" is to be read "matches." The only restriction on the use of this equation is that no one of the four colors should be matchable by a mixture of any other two. This relation can be written as the algebraic equation

$$s + t + u = r$$

For example, if r units of color A are placed into one-half of a chromometric field of a color-matching instrument, this color can be matched by:

s units of color B, plus

t units of color C, plus

u units of color D.

Should any coefficient(s) in the equation be negative, the affected color(s) are added to color A and the mixture is then matched by the remaining colors.

When two colors are mixed together, the eye distinguishes this mixture as a third color. It cannot discern either of the two original colors in the third. The eye, therefore, cannot tell whether a given color is a sensation produced from a monochromatic light source (light of a single wavelength) or whether it is produced by a mixture of colors. Physical color-matching equipment operates differently from the eye in that the light is dispersed into a spectrum. The component colors are then added together to determine the resulting hue. It is clear, therefore, that with regard to human perception it is necessary to take into account the physiology of what the eye sees when determining the color of a light. In color matching, as far as the eye is concerned, it is immaterial whether a monochromatic light or a color blend is used as a reference. The results will be the same, since the eye cannot distinguish between them.

A better understanding of how the eye distinguishes between colors is obtained by an attempt to classify different colors into groups. Up to this point we have been separating the colors in accordance with their hue. Suppose, for example, we have a large number of colored cards which we want to classify into groups. We would normally place all the reds in one pile, the greens in another pile, and so on. This is straightforward and is usually the normal grouping that one thinks of when attempting to classify colors. However, it is not the only grouping that is possible. For example, it is possible to compare a red and a green. We could take a particular red card and look at its redness, or intensity, and then take a green card which closely approximates the intensity that appears on the red card. That is to say, the intensity of the red can be matched with the intensity of the green. We are thus classifying the intensities or brightnesses of different colors. Another type of color classification can be made in terms of saturation. That is, we can take a red which looks truly red and compare it with others which may not look so red. A red mixed with white would still be red but would have a somewhat paler appearance.

As we previously pointed out, a monochromatic light can be matched by a mixture of three independent colors. Three colors are independent when any one cannot be matched by any combination of the other two. It is possible, therefore, to make a chart which relates the quantity and proportions of three given independent colors required to match a given monochromatic light. When three independent colors are chosen, we consider these to be the primary colors of our system. It should be noted here that for any specific choice of real primaries there will be a few colors that cannot be matched. This fact was observed earlier as the possibility that one of the coefficients in the color equation can be negative. The selection of three colors as primaries is arbitrary (i.e. choosing the colors represented by B, C, D in the above equation to match a color A is arbitrary). Having chosen one set of primary colors, it is possible to determine a second set of primaries from the first. The reason for this is that the specific values of the coefficients determined from the first set of primaries can be used to relate the second set of primaries. Once the relationship between the two sets of primaries is known, it is possible to convert one into the other. The International Committee on Illumination has determined the coefficients of a hypothetical set of primaries to match the given wavelengths of monochromatic light. Selected values are tabulated in Table 2. X, Y, and Z were chosen so that positive amounts of each could be mixed to match any given color regardless of hue or saturation. Thus they do not correspond to physical colors. They can be used in color equations, however, and the results can then be translated into a set of physical primaries in the manner described above. X, Y, and Z have the form of supersaturated colors (i.e. their saturation is greater than 100 percent). Their use offers two advantages. First, negative amounts of "color" are avoided in the equations. Second, Y has the property that it is also a direct quantitative measure of the brightness of the resultant color of a mix of these three.

Table 2

Hypothetical Color Coefficients.

Wavelength (millimicrons)	X	Y	Z
400	.014	.000	.068
450	.336	.038	1.772
500	.005	.323	.272
550	.433	.995	.009
600	1.062	.631	.001
650	.284	.107	.000
700	.011	.004	.000

The color equation implies the relation

$$R = X + Y + Z ,$$

where R is the coefficient of the color being matched. For example, .082 parts of a color with a wavelength of 400 millimicrons can be matched by .014 parts of X, .000 parts of Y, and .068 parts of Z, giving .082 parts of the color R.

Dividing the above equation by R gives

$$1 = \frac{X}{R} + \frac{Y}{R} + \frac{Z}{R} .$$

Let $x = X/R$, $y = Y/R$, and $z = Z/R$. Then

$$x + y + z = 1 .$$

Here we see that only two of the coefficients x, y, z are independent; the knowledge of any two coefficients will determine the third. Let us take x and y as the independent coefficients. Using the values of X, Y, and Z given in Table 2, the corresponding values for x and y can be determined by

$$x = \frac{X}{R} = \frac{X}{X + Y + Z} \quad \text{and} \quad y = \frac{Y}{X + Y + Z} .$$

The values for x and y determined in this way are tabulated in Table 3. The graph shown in Figure 19 is a plot of the values given in the above table. This plot is usually called a chromaticity diagram because it indicates the dominant hue and the saturation of any given color. The chromaticity diagram is extremely valuable because it can graphically show all the known facts concerning mixing of colors. The outer perimeter of the curve depicts colors of complete purity or saturation. The wavelengths of these 100-percent pure colors are indicated on the perimeter of the curve. Colors which are not spectrally pure are located inside this curve. One such point is indicated by A. This point represents a mixture of equal parts of the three primary colors, i.e. $x = y = z = .33$. The result of this mixture will appear to the eye as white light. Pure white, theoretically, should indeed be an equal mixture of three primary colors. Point B was selected such that its coordinates are (.2, .5). To obtain the dominant hue of Point B, a line is drawn between points A and B (point A corresponding to white light). This line is extended to intersect the spectrally pure curve at Point C. Point C represents a wavelength of 510 millimicrons. This monochromatic light (wavelength = 510 millimicrons) is the dominant hue of the mixture represented by Point B. The spectral purity of the color at Point B will be represented by the ratio between the lines AB and AC or the ratio of 1 to 2.4. Therefore, the spectral purity of the Point B mixture is approximately 42 percent. Another

Table 3

Independent Coefficients.

Wavelength (millimicrons)	x	y
400	.17	.00
450	.16	.02
500	.01	.54
550	.30	.69
600	.63	.37
650	.73	.27
700	.74	.27

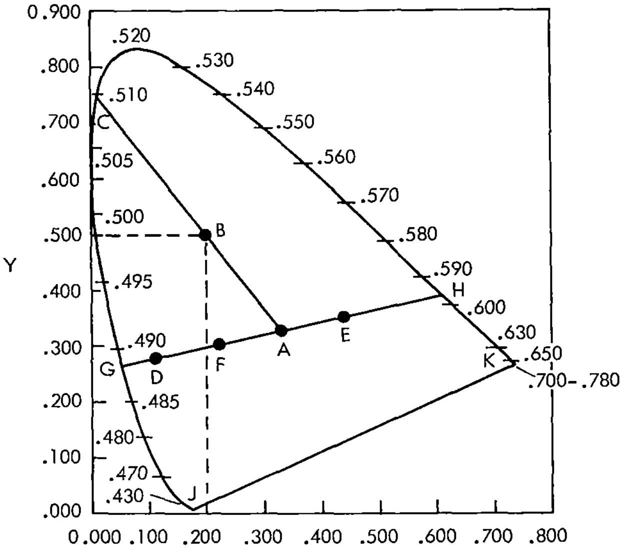


Figure 19—Chromaticity diagram.

way of stating this is to say that taking 2.4 parts of normal white light and one part of monochromatic light of wavelength 510 millimicrons will result in a color match denoted by Point B. The effects of mixing colors can also be demonstrated by the chromaticity diagram. Two colors, D and E, are to be mixed in the proportion of two parts D to one part of E. To determine the result, connect Points D and E by a straight line and divide this line in the ratio of 2 to 1. It is fairly obvious that the resulting mixture will be dominated by the color D, rather than E. Point F divides line DE in the ratio of 2 to 1 (i.e. 2 DF = FE) and, since two parts of D were used point F is situated on line DE closer to Point D than to E. The result of mixing colors D and E will be a color lying on a line joining D and E. In this particular case, the line joining D and E went also through Point A. Thus another proportion of colors D and E could be mixed to give the appearance of white light. It is also clear from the chromaticity chart that an infinite number of pairs of points could have been selected such that a line joining them would pass through Point A. The results of mixing any of these especially chosen colors (in the proper ratio for each case) will appear to be white light. Colors which have this property are called complementary colors. There is obviously an infinite number of complementary colors. The ling joining D and E extended also intersects the 100-percent purity curve at 488 and 595 millimicrons. This means that, when equal quantities of monochromatic light of these wavelengths are mixed, the resulting *sensation* will be white light (i.e. since AH = AG, equal parts of the two colors add to give the sensation of white light). It is possible to determine all of the complementary pairs of colors by using the method above. Thus the sensation of white light can be produced by two monochromatic sources of proper wavelengths, although white light from the sun normally contains most wavelengths.

Not all of the colors to which the eye is sensitive appear in a natural spectrum of white light. Mixtures of red and blue are not present in a natural spectrum. Mixtures of these colors give rise to the pinks and violets. These pinks and violets, which the eye can see as distinct colors, have no counterpart in the natural spectrum, and no spectral matches of these colors are possible. Link JK in Figure 19 represents these pinks and violets.

PRINCIPLES OF OPTICAL DATA PROCESSING

Fourier Transform by Diffraction

Optical data processors have been developed by taking advantage of a unique property of diffraction. Only diffraction from discrete slits has been considered up to this point. In such cases, the fraction of incident light (monochromatic and coherent) passed at a point in the object plane (the plane of the slits) is either 1 (slit) or zero (no slit). In general, a diffraction pattern is produced when the amplitude or phase of the incident coherent light is varied by the object plane. That is, the amplitude (and/or phase) of the incident light passing through the object plane is not the same for every point in the object plane.

To simplify further discussion, only one dimension of the object plane will be considered. Figure 20 shows a basic diffraction system. The object plane (one dimension) is described by the coordinate x, and the back focal plane (one dimension) is described by the coordinate y. The transmission of light through the object plane can be expressed as a transmission function F(x), where

$$F(x) = V(x)\, e^{j\phi(x)} .$$

At any point x the fraction of the incident-light amplitude that will pass through the object plane is V(x). The value of the fraction V will vary as a function of the local (x) of the point in the object plane being considered. Differences in the fraction of light amplitude passing through the object plane at different points can be caused by varying transparency in the object plane. The exponential term $e^{j\phi(x)}$ accounts for any phase change dependent on the coordinate x. A phase shift of this type can be caused by a transparent plate (e.g. glass) of varying thickness.

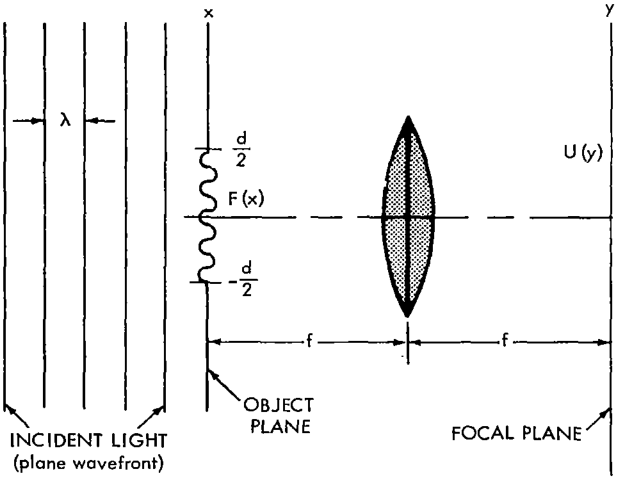


Figure 20—Optical Fourier integrator.

As an example, if the complex amplitude of the incident light is $Ae^{j\theta}$, the light appearing at the exit surface of the object plane will be given by the product of the incident amplitude and the transmission function:

$$F(x)\, Ae^{j\theta} = A\, V(x)\, e^{j[\theta+\phi(x)]} .$$

This example shows that V(x) is the fraction of incident amplitude transmitted as a function of x, and $\phi(x)$ is the phase shift introduced as a function of x. In the special case of discrete slits, these functions would have the values

$$V(x) = \begin{cases} 1 & \text{when x is in a slit} \\ 0 & \text{when x is not in a slit} \end{cases}$$

$$\phi(x) = 0 \quad \text{(No phase shift introduced by slit)}$$

When the object plane is located in (i.e. coincides with) the front focal plane of a lens, application of Huygens' principle gives the complex amplitude in the back focal plane by the equation

$$U(y) = \int F(x)\, e^{-jk(yx/f)}\, dx ,$$

where

U(y) = complex amplitude in the back focal plane as a function of the coordinate y,

$F(x)$ = transmission function of the object in the front focal plane as a function of the coordinate x,

$k = 2\pi/\lambda$, λ = wavelength of incident light,

f = focal length of the lens.

This equation has the form of a Fourier transform and can be interpreted by the statement: The complex amplitude in the back focal plane is determined by the Fourier transform of the transmission function in the front focal plane. As shown in Figure 20, an aperture is usually present to define the spatial limit of $F(x)$. For an aperture width d and $x = 0$ at the center (Figure 20, $F(x) = 0$ for $|x| > d/2$. This limit imposed by the aperture allows the integral to be written as

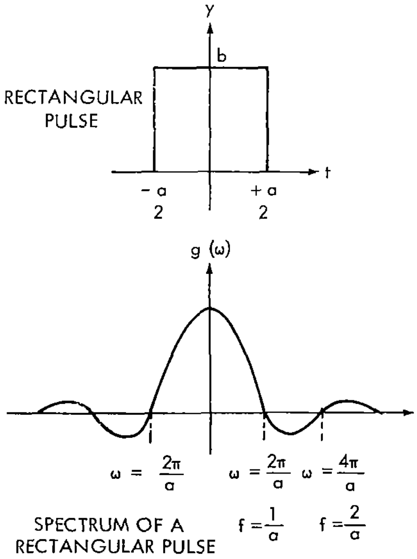
$$U(y) = \int_{-d/2}^{d/2} F(x)\, e^{-jk(yx/f)}\, dx \, ,$$

since the integrand is zero for $|x| > d/2$.

It is apparent that this unique property of diffraction offers the possibility of performing a Fourier transform on any function that is represented by the transmission function. Thus it is possible to process data optically by introducing signals as transmission functions and performing any necessary operations on the Fourier transforms.

Fourier Transform

At this point a short review of Fourier integrals as applied to optics in one and two dimensions might be helpful. Figure 21 represents a common problem in electrical engineering, i.e. given a



$$G(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} f(t)\, e^{-j\omega t}\, dt$$

$$= \frac{1}{2\pi}\int_{a/2}^{a/2} (b)\, e^{-j\omega t}\, dt$$

$$= -\frac{b}{2\pi\, j\omega}\left[e^{-j\omega t}\right]\frac{a/2}{-a/2}$$

$$= \frac{b}{\pi\omega}\;\frac{e^{j\omega t/2} - e^{-j\omega t}}{2j} = \frac{b}{\pi\omega}\ \mathrm{SIN}\ \omega\frac{a}{2}$$

$$= \frac{ba}{2\pi}\;\frac{\mathrm{SIN}\,\frac{\omega a}{2}}{\frac{\omega a}{2}}$$

Figure 21—A rectangular pulse and its spectrum.

waveform such as the rectangular pulse shown in Figure 21, find the "spectrum" of its sinusoidal components. The spectrum shown is a plot of the amplitude of each frequency component required to produce the given rectangular pulse. It should be noticed that the continuous spectrum curve implies that all frequencies are present in the rectangular pulse. This is in contrast to the discrete frequency spectrum of a continuous periodic waveform.

The frequency spectrum of a given waveform is found by taking the Fourier transform of the function describing the waveform. This transform is the process of multiplying by an exponential factor and integrating with respect to the variable of the given waveform. The example of Figure 21 includes the mathematics used to obtain the spectrum. The limits of integration are defined as $+\infty$ and $-\infty$; however, in the example the integration is actually performed between the limits $-a/2$ to $+a/2$, since the integrand is zero for all other values.

The example of Figure 21 was expressed as a function of time (t). This is the usual case in electrical engineering, where voltage varies as a function of time. In Figure 22 the same rectangular pulse is shown as a function of the spatial dimension x. Considering the rectangular aperture shown in Figure 22, it can be seen that the rectangular pulse can be used to represent the amplitude of light plotted against the horizontal dimension x. If the light incident upon the aperture is constant in amplitude and phase, the light passed by the aperture along a horizontal line will be given by the rectangular pulse. That is, no light is passed beyond the dimension of the aperture and the incident-light amplitude has a constant value across the aperture as given by a rectangular waveform. In terms of a one-dimensional transmission function as described in the previous section, this aperture is represented by

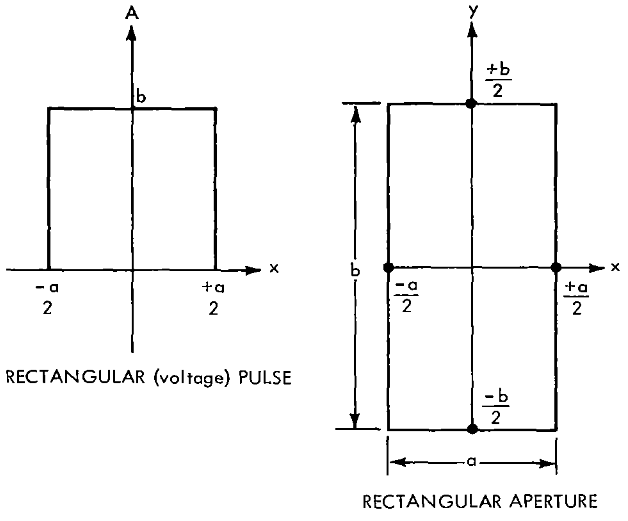


Figure 22—Rectangular pulse—rectangular aperture.

$$F(x) = \begin{matrix} 1 \text{ when } x < \frac{a}{2} \\ \\ 0 \text{ when } x > \frac{a}{2} \end{matrix} .$$

According to the principles of the preceding subsection, a lens will produce a Fourier transform of this function, as follows:

$$G(\omega) = \int_{-a/2}^{a/2} F(x)\, e^{-j\omega x}\, dx .$$

In this application, the ω term represents a spatial dimension in the transform plane. Comparing this transform equation with that of the voltage waveform in Figure 21, it is apparent that the two cases are identical except for the interpretation of the variables. It may be noted that the $1/2\pi$ factor is not present in the above expression. This constant factor involves only a scale change and is therefore of no importance.

Further consideration of the rectangular aperture of Figure 22 leads to a two-dimensional Fourier transform. Just as the light passing through the aperture is represented by a rectangular pulse with respect to x in the horizontal direction, along a vertical line the light can be represented by a similar rectangular pulse with respect to y. This rectangular pulse can be analyzed independently from that representing the horizontal case. This would represent another one-dimensional transform identical in form to the one given above. However, we can also consider both functions at the same time by using a two-dimensional transmission function given as

$$F(x, y) = \begin{matrix} 1 & \text{when} & x < \frac{a}{2} & \text{and} & y < \frac{b}{2}, \\ 0 & \text{when} & x > \frac{a}{2} & \text{or} & y > \frac{b}{2}. \end{matrix}$$

The Fourier transform is then given by a double integral

$$G(\omega_x, \omega_y) = \int_{-b/2}^{b/2} \int_{-a/2}^{a/2} F(x, y)\, e^{-j(\omega_x x + \omega_y y)}\, dx\, dy \ .$$

It is important to note that the two-dimensional transform produces a two-dimensional spectrum in terms of the spatial terms ω_x and ω_y.

The capability of taking a two-dimensional Fourier transform is an advantage which optical systems have over the usual electronic system. An extension of processing to two dimensions provides an increase in the amount of information which can be processed at one time. In a specific application one of the dimensions can be used for signal information while the other dimension is used to provide multichannel operation.

Details of Fourier Transform by Diffraction

The Fourier transform characteristic of diffraction is applied extensively in optical processing systems. Therefore, the limitations and requirements for producing an optical Fourier transform should be clearly understood. The following derivation should clarify the restrictions imposed to obtain an optical Fourier transform.

Consider the optical system shown in Figure 23. Collimated (plane wavefronts) monochromatic light is incident from the left and is represented by parallel rays. A thin lens of focal length f is mounted with its axis parallel to the direction of the incident light rays. At a distance g in front of the lens there is an opaque screen, P, with a rectangular aperture. The dimensions of the aperture are d in the plane of the paper and b perpendicular to the plane of the paper. We will derive the expression for an optical Fourier transform in only one dimension. The second dimension can be considered independently and similar results would be obtained. When the Fourier transforms of both dimensions are considered simultaneously, a double integration (transform) is required. The two-dimensional, or double, Fourier transform was discussed for a special case in the preceding subsection.

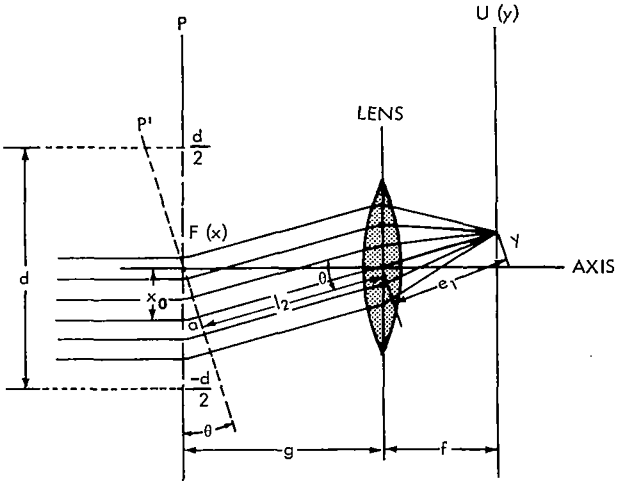


Figure 23—Optical Fourier transform system.

In order to assure that light passing through the aperture will not be outside the lens field, it is necessary that the aperture dimension d be small compared to the dimensions of the system (but large compared to a light wavelength λ). A film with a transmission function $F(x)$ recorded along the x coordinate is inserted in the aperture. The collimated light incident upon the aperture can be represented by a constant amplitude A. The incident-light amplitude is varying at a particular frequency with respect to time. This variation with respect to time is represented by an exponential factor $e^{j(\omega t+\alpha)}$, where ω is the angular frequency and α is the initial phase constant. Since this exponential term will appear in all amplitude terms unchanged, it can be neglected as a common factor. It must be kept in mind that, although the incident-light amplitude A is constant with respect to spatial coordinates of the aperture, all amplitudes are actually synchronously varying with respect to time. By previous definition, the light amplitude leaving the film with transmission function $F(x)$ will be given as the product of the incident-light amplitude and the transmission function, or

$$U(x) = A\, e^{j(\omega t+\alpha)}\, F(x) \; .$$

Eliminating the time-dependent exponential $e^{j(\omega t+\alpha)}$, the light amplitude leaving the film can be expressed

$$U(x) = A F(x) \; .$$

Starting with this amplitude and applying Huygens' principle, it is possible to determine the amplitude and phase of the light converging to a point y in the back focal plane of the lens. A rigorous derivation of the expression describing the amplitude and phase of the light in the back focal plane is beyond the scope of this report. The following descriptive development, however, should clarify the physical significance of the resulting equation.

The amplitude of the light leaving the film plane will be, as above,

$$U(x) = A F(x) \; .$$

If we consider a small interval dx about the point x, the contribution of this interval to the light amplitude beyond the film can be written

$$A F(x) \, dx \; .$$

The energy of electromagnetic radiation (light) traveling through space is inversely proportional to the square of the distance traveled. Since the energy is proportional to the square of the amplitude, the amplitude will be inversely proportional to the distance traveled. Thus for a distance of travel r from the point x in the film plane, the amplitude will be

$$\frac{A}{r} F(x) \, dx \; .$$

The fact that low-frequency (long-wavelength) radiation does not travel as far as high-frequency (short-wavelength) indicates that a factor of $1/\lambda$ (λ = wavelength) is required:

$$\frac{A}{\lambda r} F(x) \, dx \; .$$

There is also an effect on amplitude caused by the fact that not all of the light travels parallel to the optical axis. This effect is seen in the light pattern produced by a pinhole where the pattern is brightest at the points directly in line with the pinhole. An obliquity factor $(1 + \cos\theta)/2$ accounts for this amplitude variation:

$$\frac{1 + \cos\theta}{2} \frac{A}{\lambda r} F(x) \, dx \; .$$

A phase shift due to traveling a path length r is given by a factor e^{-jkr}, where k is the wave number $(2\pi/\lambda)$. A -90° phase shift inherent in the application of Huygens' principle is represented by a factor $1/j$. Applying these phase shifts to our expression for amplitude produces

$$\frac{A}{j\lambda} \frac{1 + \cos\theta}{2} \frac{1}{r} F(x) \, e^{-jkr} \, dx \; .$$

If r is defined as the distance from a point x (in the film plane) to a point y (in the back focal plane of the lens) the light amplitude at y contributed only by the interval dx in the neighborhood of x is

$$dU(y) = \frac{A}{j\lambda} \frac{1 + \cos\theta}{2} \frac{1}{r} F(x) \, e^{-jkr} \, dx \; .$$

The total light amplitude at a point y due to the sum of the contributions from all x is given by integration (sum of an infinite number of infinitesimal dx). The resultant equation for the light amplitude at a point y is

$$U(y) = \frac{A}{j\lambda} \int_{-d/2}^{d/2} \left(\frac{1 + \cos\theta}{2}\right) \frac{1}{r} F(x)\, e^{-jkr}\, dx \ ,$$

where

λ = wavelength of the incident light,

r = optical path length from point x in plane P to point y in the back focal plane,

$k = 2\pi/\lambda$ = wave number,

θ = angle between diffracted wavefront and plane P (see Figure 23).

This equation can be simplified as follows:

1. The factor $1/j$ contributes a constant phase shift and may be neglected.
2. The factor A/λ is a constant and may be neglected.
3. $(1 + \cos\theta)/2$ is called the obliquity factor. In a practical case θ is very small. Therefore $\cos\theta \approx 1$ and $(1 + \cos\theta)/2 \approx 1$.
4. The term $1/r$ represents the effect of distance upon the amplitude of the light wave. For different points in the back focal plane (different y coordinates), r will vary by only a few wavelengths. Since r is usually of the order of 10^5 wavelengths, $1/r$ may be considered a constant and may be neglected.

The simplified equation will then be

$$U(y) = \int_{-d/2}^{d/2} F(x)\, e^{-jkr}\, dx \ .$$

Now consider the plane wavefront P′ (Figure 23) which is perpendicular to the parallel rays that are focused to the point y by the lens. The optical distance from the plane P′ to y is the same along all rays perpendicular to P′. The rays having the shorter geometrical paths travel the longer paths in the lens. Since light travels slower in glass (because of greater refraction index), their time of travel will be greater. Because of this effect, the overall optical path lengths (defined as the distance traveled in air in an equal time interval) are the same for all of the rays for the case mentioned. Since the optical path lengths of the parallel rays are equal, the consideration of any one ray will not affect the generality of the derivation. Consider the ray, originating at x_0, which passes through the center of the lens and is therefore undeviated. The distance from P′ to y along this path is

$$c = \ell_1 + \ell_2 \ .$$

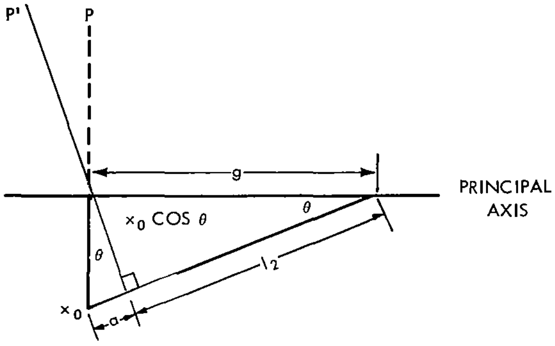


Figure 24—Geometry for calculation of ℓ_2.

From Figure 23 it is seen that

$$\ell_1 = \sqrt{f^2 + y^2} = f\sqrt{1 + \left(\frac{y}{f}\right)^2} .$$

Referring to the expanded view shown in Figure 24, it is apparent that

$$\ell_2 = \sqrt{g^2 - x_0^2 \cos^2\theta} .$$

For small θ, $\cos\theta \approx 1$, and from Figure 23 we have

$$\frac{x_0}{g} = \frac{y}{f} , \quad \text{or} \quad x_0 = g\frac{y}{f} .$$

Substituting this expression for x_0 in the equation giving ℓ_2 gives

$$\ell_2 = \sqrt{g^2 - g^2\left(\frac{y}{f}\right)^2} = g\sqrt{1 - \left(\frac{y}{f}\right)^2} .$$

For the practical case where $y \ll f$, the square root factors can be expressed (with good approximation) as the first two terms of their respective binomial expansions. The binomial expansion of $(1 \pm Z^2)^{1/2}$ is given by the formula

$$\left(1 \pm Z^2\right)^{1/2} \;\square\; 1 \pm \frac{Z^2}{2} - \frac{Z^4}{4} \pm \cdots .$$

The first two terms in the binomial expansion of ℓ_1 and ℓ_2 give

$$\ell_1 = f\left[1 + \frac{1}{2}\left(\frac{y}{f}\right)^2\right] , \quad \text{and} \quad \ell_2 = g\left[1 - \frac{1}{2}\left(\frac{y}{f}\right)^2\right] .$$

Therefore, the optical path length c from P' to y along any of the rays perpendicular to P' is

$$c = \ell_1 + \ell_2 = f + g + \left(1 - \frac{g}{f}\right)\frac{y^2}{2f} .$$

The distance a, along the ray under consideration, from the plane P' to the plane P is

$$a = x_0 \sin\theta .$$

However, for small θ, $\sin\theta \approx \tan\theta = y/f$, which gives

$$a = \frac{x_0 y}{f} .$$

The distance r from the point x_0 in the plane P to a point y in the back focal plane is then the sum of the distances c and a

$$r = c + a = (f + g) + \left[\frac{1}{2f}\left(1 - \frac{g}{f}\right)\right] y^2 + \frac{x_0 y}{f} .$$

This can be written

$$r = A + By^2 + \frac{x_0 y}{f} ,$$

where

$A = f + g =$ constant,

$B = \frac{1}{2f}\left(1 - \frac{g}{f}\right) =$ constant.

Since the optical path lengths are equal, r can be expressed in terms of x rather than x_0:

$$r = A + By^2 + \frac{xy}{f} .$$

We can now substitute this expression for r in the equation

$$U(y) = \int_{-d/2}^{d/2} F(x)\, e^{-jkr}\, dx ,$$

obtaining

$$U(y) = \int_{-d/2}^{d/2} F(x)\, e^{-jk(A+By^2+xy/f)}\, dx .$$

Now the exponential term can be written as the product of three exponential terms:

$$U(y) = \int_{-d/2}^{d/2} F(x)\, e^{-jkA}\, e^{-jkBy^2}\, e^{-jkxy/f}\, dx .$$

The exponential factor e^{-jkA} is independent of both x and y. Therefore, it represents a constant phase shift. This constant phase can be neglected, since we are interested only in the *relative* phase difference between various points in the back focal plane. The expression for the light amplitude in the back focal plane has now been reduced to

$$U(y) = e^{-jkBy^2} \int_{-a/2}^{a/2} F(x)\, e^{-jkyx/f}\, dx \;.$$

The integral is identical in form to the Fourier transform. The coordinate x corresponds to the conventional time coordinate, and the coordinate y corresponds to the frequency domain. The Fourier transform is generally a complex quantity with both amplitude and phase. The exponential factor e^{-jkBy^2} introduces phase shifts dependent upon the y coordinate. Unless this term is eliminated, the relative phase of the light in the back focal plane will not correspond to that of the desired Fourier transform.

In order to eliminate this phase error term, the exponent must be zero for all values of y. The necessary condition is, therefore, $B = 0$, or

$$B = \frac{1}{2f}\left(1 - \frac{g}{f}\right) = 0 \;.$$

Simplifying this expression yields the necessary condition,

$$g = f \;.$$

Therefore, when the plane P coincides with the front focal plane of the lens, the Fourier transform of a transmission function placed in the plane P will be given by the light amplitude and phase in the back focal plane.

It is important to note the effects of the location of the aperture plane. The relative amplitude and phase corresponding to a Fourier transform are produced only when the aperture plane coincides with the front focal plane of the lens. This is a subtle point, since visual observations do not detect differences that are due to phase. The eye (or a photocell) responds to intensity of light, which is not affected by changes in phase. In applications where the transform is the end result, the phase differences are not important, since all readout devices are intensity-sensitive. However, in applications where further operations are performed on the complex amplitude representing a Fourier transform, the phase differences must be correct. In these cases the aperture plane must coincide with the front focal plane to avoid phase errors.

Intensity-Amplitude Complication

It has been stated that optical data processing is based on the diffraction principle, which produces a Fourier transform of the transmission function. This Fourier transform and transmissio function are based on *light amplitude.* In practice the complex amplitude of light cannot be measured.

The characteristics of physical photo detectors (e.g. film, human eye, photocell) are such that they can sense only the *intensity* of light (amplitude squared) and cannot sense the amplitude of light directly. Measured values of light are therefore expressed in terms of intensity, rather than of amplitude. For optical data processing, it is necessary to consider how a desired signal can be expressed as an amplitude of light. Since only intensity of light can be measured, the square root of the measured intensity (which gives amplitude) should be the desired signal. The following discussion will clarify this process.

As an example, assume a desired signal is to be the sine wave given by $y = A \sin x$ as shown in Figure 25 (A). It is desired that the amplitude of the light is to vary in accordance with this equation. The sine wave has negative values on alternate half-cycles. It is not possible to have negative values of light amplitude. There is either no light (zero amplitude) or there is some light (positive amplitude). It should be noted here that the transmission function is never negative. In order to represent this signal (sine wave) as a variation in light amplitude, it is necessary to introduce a bias to the amplitude, as shown in Figure 25 (B). It is seen that adding a bias of $+A$ to the original wave eliminates negative values. Should a photocell be used to monitor a light signal of the form shown in Figure 25 (B), it will not produce a sine wave as an output. The photocell responding to light intensity will produce an output which will be of the form in Figure 25 (C), which is the square of Figure 25 (B). In other words, if a photocell is used to monitor a light signal of *amplitude* $y = A + A \sin x$, its output will correspond to the intensity of this light and will be $y = (A + A \sin x)^2$. In order to find the amplitude of the input light signal, it is necessary to take the square root of the photocell output.

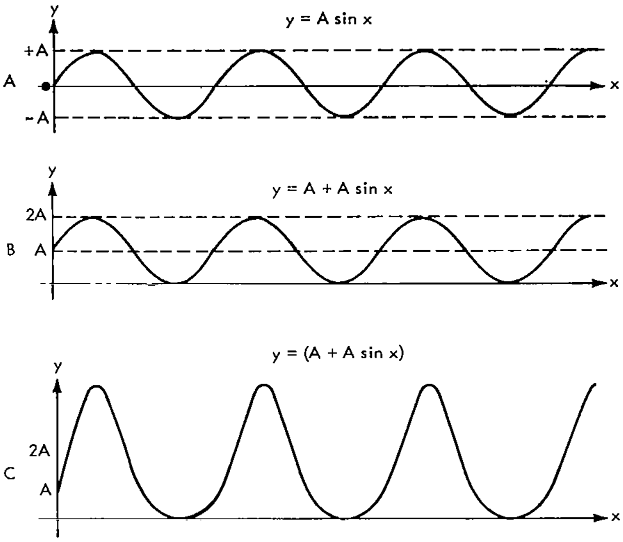


Figure 25—Amplitude, bias, and intensity.

To summarize, it has been shown that a desired sine wave amplitude variation must be modified by the addition of a bias to form an amplitude variation which does not require a negative light amplitude. It was further pointed out that it is necessary to take the square root of a photocell (or any other photo sensor) output to determine the amplitude of the light input, since only light intensity can be sensed.

Once again consider the light intensity given by $y = (A + A \sin x)^2$. A voltage waveform $y = (A + A \sin x)^2$ applied to the vertical deflection input circuit of an oscilloscope would be seen on the cathode ray tube face as shown in Figure 26 (A). If a photocell were to monitor such a presentation, there would be no variation in light intensity corresponding to the waveform being monitored. The

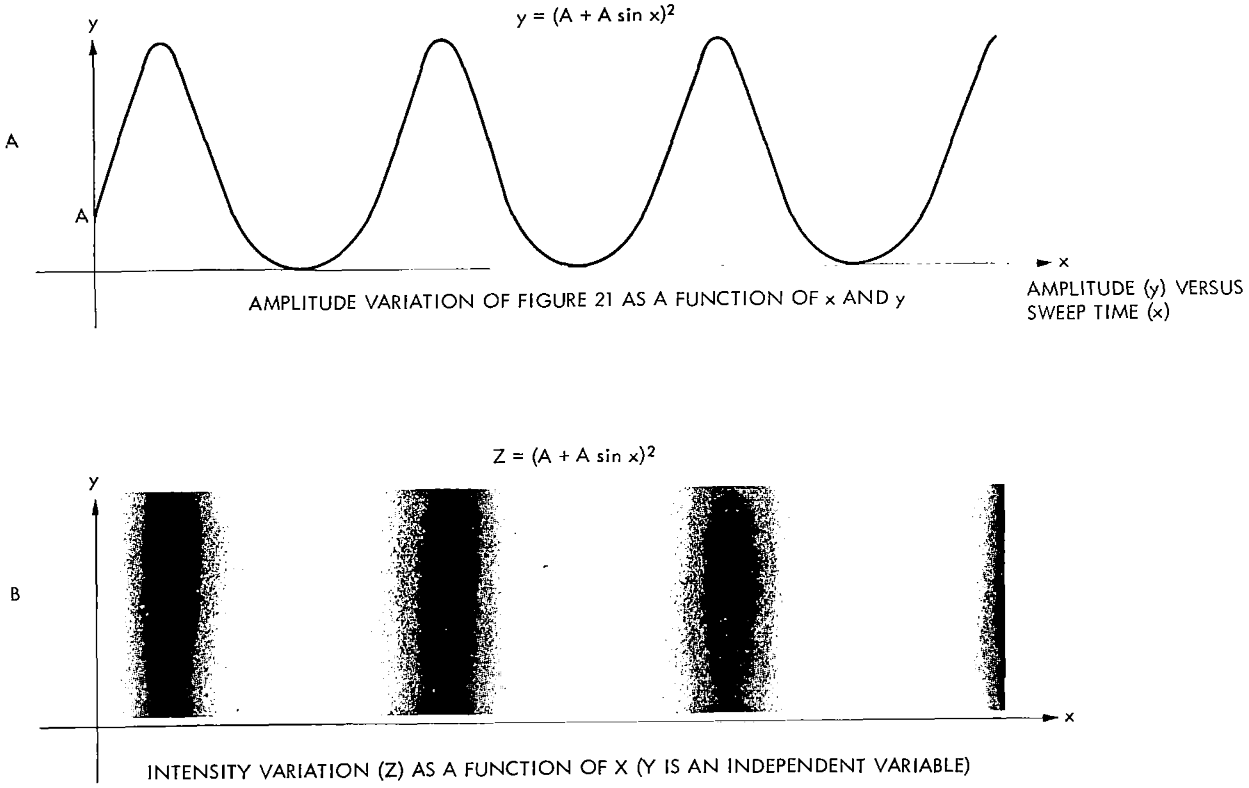


Figure 26—Oscilloscope waveform and intensity displays.

light intensity output from a cathode ray tube is approximately linear with respect to the input voltage on the accelerating grid. Thus, when a voltage waveform $y = (A + A \sin x)^2$ is applied to the accelerator grid (z-axis modulation instead of vertical deflection), the light intensity on the face of the cathode ray tube will be proportional to $(A + A \sin x)^2$. Figure 26 (B) shows the oscilloscope presentation when the applied voltage $y = (A + A \sin x)^2$ appears as a light-intensity variation. Comparing the waveform of Figure 26 (A) with the intensity variation of Figure 26 (B), it is seen that the areas in which the applied signal is maximum correspond to areas of maximum intensity in the display.

The characteristics of transmission functions must also be considered in terms of intensity, since photo sensor characteristics are expressed in terms of intensity. By using Figure 26 and the discussion above, the relation between the transmission function and intensity characteristics is easily shown. The intensity variation obtained by z-axis modulation of an oscilloscope can also be obtained by inserting an appropriate variable-density film in the path of a constant-intensity light beam. The transmittance characteristic of film expressed in terms of intensities is called the transmittance. If the intensity of the incident light is I_0 and the intensity of the light after passing through the film is I_T, the transmittance is given by

$$T = \frac{I_T}{I_0} .$$

The intensity (I_T) of transmitted light corresponding to Figure 27 is of the form $(A+A\sin x)^2$. Let A^2 equal the intensity (I_0) of incident light. Substituting for I_0 and I_T in the equation for transmittance gives

$$T(x) = (1+\sin x)^2 .$$

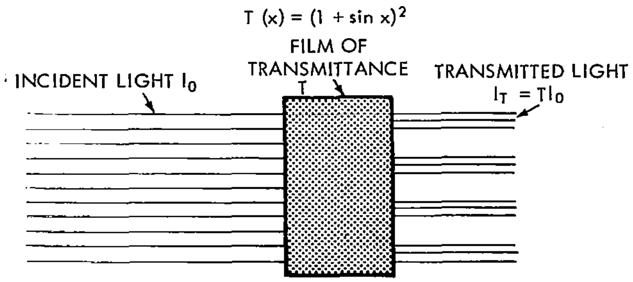


Figure 27—Transmittance of film.

However, the transmission function for this case is given by

$$F(x) = \frac{A_T}{A_0} = \frac{A+A\sin x}{A} = 1 + \sin x ,$$

where

$F(x)$ = Transmission function of film,

A_0 = Amplitude of light incident on film,

A_T = Amplitude of light transmitted through film.

Therefore, the relationship between transmission function, transmittance, and intensity is given by

$$F(x) = \left[T(x)\right]^{1/2} = \left[\frac{I_T}{I_0}\right]^{1/2} .$$

Although this relationship was derived for a specific case, the same result is obtained in the general case, i.e. if the transmission function $F(x)$ corresponding to a desired signal is of the form $A+A\sin x$, as in Figure 25 (B), the light intensity transmitted I_T and transmittance $T(x)$ are of the form $(A+A\sin x)^2$, as in Figure 25 (C).

At this point, the wavelength interpretation of slit spacing can be effectively demonstrated. The similarity between the intensity presentation of Figure 26 (B) and that produced by slits is apparent. The intensity variations can be likened to the effect of producing transparent and opaque areas which as an overall effect will correspond to slits. It is quite apparent how the spacing between the slits corresponds to a wavelength by comparing diagrams A and B of Figure 26. If a film is produced with a transmittance corresponding to Figure 26 (B), the diffraction pattern produced will be similar to that for the corresponding slit spacing. In Figure 17 it has been shown that the zero-order diffraction wavefronts are focused to a point at the back focal point of the lens. This point corresponds to a dc component. The first-order up diffraction wavefronts are focused to a point in the focal plane above the dc component. The spacing of the first-order point from the dc component is inversely proportional to the slit spacing or directly proportional to the spatial frequency. From this simple example it is seen how the frequency of the transmission function is transposed into spatial dimensions. This effect demonstrates the Fourier-transform character of diffraction and will be explained in more detail in a later subsection.

Light Modulators

In order to continue the development of optical processing principles herein, it is necessary to investigate the means available for implementing transmission functions. The most obvious method for varying the transmission of light is to insert film in the path of a coherent light beam having uniform intensity. The optical density of the film will produce variation in the light intensity (and therefore in amplitude). The various types of film can be categorized by the general processes which determine the density of the film.

Photographic Film

"Normal" photographic film, in everyday use for taking snapshots, depends on the light-sensitive characteristics of the silver salts. Since this is the most frequently used film in optical processing, the next subsection will review its characteristics in detail.

Thermoplastic Films

Films which are dependent upon heat sensitivity are called thermoplastic films. Kalvar is a thermoplastic film which is heat-developable. When Kalvar is exposed to ultraviolet light, minute bubbles are formed. These bubbles vary the index of refraction of the plastic traip, causing the photographic image to be formed. Developing consists of applying heat to set the plastic. After development, the formed bubbles are retained and further exposure to ultraviolet light has no effect. The image recorded on Kalvar can be viewed by either reflected or transmitted light. When viewed by transmitted light, the image is a positive. When viewed by reflected light, the image is a negative. Therefore, one has an option of viewing the single exposure as either a positive or a negative.

Another type of thermoplastic film is produced in flat sheets. Upon exposure to heat, the flat sheets distort and form ridges and gullies. Light passing through the distorted film is diffracted by the ridges and gullies and thus the light intensity is varied. A typical use for such a film is for instant replays as used on television for football games. The optical picture is imaged on the film; the heat from this imaged picture produces distortions in the film. After the film "sets," light can be passed through the plastic. The intensity of this light will vary in accordance with the original image which caused the distortions in the plastic film.

Cathode Ray Tube Display

Another type of light modulator is an oscilloscope. An oscilloscope presentation can be in the form of an intensity modulation as previously described. This intensity modulation can be used as a light source for an optical processor. The major disadvantage of an oscilloscope tube as an optical source of signal is that the light intensity is quite low. Several developments are presently under way in this field. One promising development is an oscilloscope tube being produced by RCA. This tube has an endless belt inside the evacuated chamber. The phosphor is coated onto this endless belt. Small magnets are mounted on portions of the endless belt, permitting it to be moved by a magnet driven by a motor external to the evacuated chamber. An electron gun paints the phosphor with the desired signal. Depending on the tube construction, the readout can be produced in two ways:

(a) The light of the source is reflected off the phosphor. The reflected intensity of the light beam is the product of the electronic signal impressed on the phosphor and the incident-light intensity. Erasure of the signal and re-exposure to a new signal are done before the light beam is reached again.

(b) In the second method, the light source is behind the phosphor-coated belt. The light projects through the phosphor, having its intensity varied in accordance with the electronic signal impressed on the phosphor. From these tubes an intensity of light can be obtained several orders of magnitude greater than from conventional CRT's. This increase is due to the fact that the light being used for processing is derived from a separate light source rather than from the phosphor in the tube.

Magneto-Optic Modulator

Another type of light modulator is a magneto-optic light modulator. A magneto-optic light modulator is used to read a recorded magnetic tape optically. This is usually accomplished by passing the magnetic tape with the recorded signal near another magnetic material (such as iron cobalt). When the magnetic tape passes near a film of iron cobalt, it will affect it in accordance with the information on the magnetic tape: If polarized light is projected onto the iron cobalt film, the reflected light will be polarized differently by the iron cobalt in the areas where the signal existed. This type of modulator, in general, is not as satisfactory as other types because of severe light losses which take place upon reflection and because polarized light must be used. With present techniques, the change in polarization of the light due to the magnetic signal is not sufficient for the usual demands of optical data processes.

Ultrasonic Light Modulator

Another type of light modulator is the ultrasonic light modulator cell. An ultrasonic modulator consists of two optically flat glass plates separated by some liquid such as distilled water. With no signal, any light passing through the plates and water will not be disturbed. Perpendicular to the light beam is a small ultrasonic cell which induces sound waves into the water in a direction perpendicular to the light beam. These sound waves in the water are absorbed at the upper end of the cell by an absorber (such as foam rubber) so that no standing waves are set up. As the sound waves pass through the water, compressional waves cause local periodic changes in the density of the water which result in corresponding variations of its index of refraction. The light beam passing through the light modulator will be diffracted in accordance with the changes in the index of refraction caused by the compressional sound waves existing in the water. Although such cells have been used successfully, at present the main disadvantage of ultrasonic light modulators is that only relatively short-duration signals can be used to modulate the light. This adversely affects the information capacity of such systems. Attempts to increase the signal time in such modulators have been unsuccessful for several reasons:

(1) Increasing the length of the modulator also increases the absorption of the signal. The signal may be attenuated before reaching the end of the water tube.

(2) Longer cells require greater ultrasonic power to drive them.

(3) The possibility of multiple reflections increases as the cell length and sonic power increase.

Photographic Film Basics

From the above review of light modulators, it appears that at the present time photographic film is probably the most convenient and versatile medium available. Photographic film offers the possibility of the greatest bandwidth, as well as reasonable capabilities for real-time use. The use of photographic film in optical data processing requires a rather thorough knowledge of the characteristics and limitations of this type of film. For this reason, this section will review some of the detailed characteristics of photographic films.

The term "photographic film" usually refers to a light-sensitive emulsion on a base support. If a plastic base is used to support the emulsion, the end product is called photographic film. When a glass plate is used to support the emulsion, the end product is called a photographic plate. The light-sensitive emulsion is basically a mixture of silver halide crystals in gelatin. The sensitivity of the silver halide crystals can be affected by the addition of various sensitizing agents. The size of the silver halide crystals, which are photosensitive, varies from less than 1 micron to several microns. Usually, the larger the silver halide crystals, the greater their sensitivity to light, i.e. a larger crystal generally requires less light to expose it. An emulsion which contains relatively large crystals is generally called a "fast" emulsion. A "fast" emulsion, however, will lack the capability of registering fine details. The limits of resolution of the emulsion will be determined by the size of the crystals which become exposed. The smaller the crystals, the more detail that can be captured in the film. The smaller the crystals, however, the slower the speed of the film. For a given emulsion there is a definite exposure latitude that is determined by the size of the crystals composing the emulsion. Below a certain level of exposure the smaller crystals will not be exposed. This effect influences the contrast of the film. The photographic emulsion must be properly selected to permit the greatest resolution to be captured on the film as well as the highest contrast ratios. Film must be selected, therefore, to render a final product which registers the resolution required as well as the tonal range. In general, both of these requirements cannot be met fully, and some compromise between them is usually made.

Any light that will produce a photographic effect on silver halide must be absorbed by the sensitive material. When light energy (photons) is absorbed by silver halide crystals, there is no visible change in the appearance of the emulsion. The exposed emulsion, however, contains an invisible *latent image* which can be converted into a visible silver image by the action of the developer.

The developer chemically changes the exposed silver halide crystals to metallic silver, thus making the exposed portions of the emulsion opaque.

The development process is possible because certain chemicals react with the exposed silver halide in preference to that which is unexposed. If the developing process is continued sufficiently long, all the silver halide in the emulsion (unexposed as well as exposed) will be converted to silver. Even during the normal development process, some unexposed silver halide crystals will be reduced to metallic silver. This tends to produce a very slight uniform silver coating over the entire image. This thin coating is generally referred to as fog. Photographic fog is generally considered undesirable, as it adversely affects both resolution and contrast capabilities of the film. When the amount of light absorbed by the silver halide crystals is insufficient to expose them, the developer (initially) has no effect on them. The purpose of a fixing bath is to dissolve out any silver halide crystals remaining in the emulsion.

The steps involved in the normal development process for making a positive picture are as follows:

1. Expose the light-sensitive material.

2. Develop the exposed light-sensitive material to produce a *negative.*

3. Fix the negative by removing any remaining light-sensitive materials to assure its permanence.

4. Wash to remove any harmful residue remaining from the fixing process.

5. Dry the negative.

6. Expose a second photographic-sensitive material through the negative.

7. Develop the newly exposed material to produce a positive.

8. Take the positive through steps 3, 4, and 5 (as was done for the negative) to complete the process.

Another photographic process is the reversal technique. In this process, the exposed film is developed and then bleached. The bleaching removes the exposed metallic silver from the emulsion, leaving the unexposed silver halide crystals. This emulsion is then exposed to light and developed through the normal development process. The resulting image is now a positive instead of a negative. The following step-by-step procedure is used for the reversal process.

Step 1. First Development

The first development is the most critical operation in the reversal process and therefore must be carefully controlled. It is in this step that the negative silver image is formed. (For the remaining steps, timing is not critical as long as the times are long enough to carry out the processes.)

Step 2. Stop

Either a stop bath or running water is used to immediately stop the development action and remove any first developer from the film.

Step 3. Bleach

The negative silver image is dissolved, leaving only unexposed silver halide in the emulsion. (Note: The remaining steps can be accomplished in subdued light without affecting the results.)

Step 4. Rinse

The film is rinsed in water to remove the bleach.

Step 5. Clear

The film is placed into a solution to prevent the later formation of stains which might be an after-effect of the bleach.

Step 6. Rinse

The film is again rinsed to remove the clearing solution.

Step 7. Exposure

The film is exposed to light to expose the remaining silver halide.

Step 8. Second Development

In the second development, the balance of the silver halide is converted to silver, producing a positive silver image. (It has been found advantageous to develop slightly beyond the point when the image appears to have been fully brought out.)

Step 9. Rinse

A stop bath or rinsing in cold water stops the second development.

Step 10. Fix

The film is now fixed and hardened. There should be little or no silver halide remaining for removal by the fixing-bath; however, the hardening of the emulsion seems to be advantageous at this point.

Step 11. Wash

Washing in running water to remove the fixing solution from the emulsion reduces the possibility of strains forming.

Step 12. Photo-Flow

The film is rinsed in a photo-flow solution to prevent spotting during drying. Photoflow is a solution which reduces the surface tension of water, thus preventing drops from forming.

Step 13. Drying

The film is allowed to dry without any wiping.

To review: The first developer brings out the negative image in the usual manner by forming a silver image where the silver halide had been exposed to light. (Step 1 is the only step that permits convenient altering of the final results. The reason is that all of the remaining steps are essentially carried to completion.) This silver image can be removed by bleaching, leaving only unexposed and undeveloped silver halide. When the bleached emulsion is exposed and developed, a positive image results. In other words, wherever there had been a dense deposit of silver in the negative there will be the least amount of silver in the positive. Likewise, where there had been no silver in the negative image there will be dense silver in the positive. The action of the bleach can be considered exactly opposite to that of the fixer in the normal developing process. The bleach removes the silver image without disturbing the undeveloped emulsion, while the fixer removes the undeveloped silver halides without affecting the silver image.

Photographic Films for Optical Processing

A great deal of research has been done on the production of films and the investigation of film characteristics. Unfortunately, most of this research has been directed towards the commercial market and very little towards specific theoretical investigations. The possibility of using films for optical data processing has not been fully investigated. Several unique requirements exist for films to be used in this application. Since the particular characteristics of films advantageous to optical data processing have not been fully explored, it is quite necessary that additional research efforts be directed along these lines.

For optical data processing, it is evident that high-resolution film is necessary to obtain a large time bandwidth. In general, the speed of the film does not appear to be as important as its resolution. It is felt that delays caused by long exposure time can be either tolerated or overcome by using greater light intensities. Long exposures are not considered critical, since delays in processing the films will more than overshadow any time required for exposure. Therefore, research on films should be directed towards obtaining the maximum possible resolution. This is not the usual case, for the commercial film market requires not only high resolution but also fast exposure speeds. Generally, the research work done has been a compromise between these two requirements (somewhat favoring film speed over resolution). In present optical data processing there is no need for compromise. The strict requirement of high resolution must be met regardless of exposure speeds. Certain commerical films can be used in optical data processing, but it is quite apparent that these films do not offer the maximum potential available.

As a prelude to research on films, a study was made to determine which areas offered the most potential for improvement. It was found that in 1908 Dr. G. Lippmann won a Nobel Prize for the development of a particular film emulsion. This emulsion appears to have the basic characteristics necessary for optical data processing, principally high resolution. Indeed, Dr. Lippmann's emulsion has so high a resolution that it is even possible to capture the separation between particular wavelengths of light. This ability to capture and record information as small as the wavelength of light allows this emulsion to be used for making color pictures. It was for the development of this color film that Dr. Lippmann was awarded the Nobel Prize.

Dr. Lippmann's color process requires that a fine-grained panchromatic emulsion (sensitive to all colors) be placed in contact with a pool of mercury. Collimated light is projected through the emulsion normal to the plane of the emulsion. Upon reaching the mercury, the light is reflected back on the same path it had just taken. The incident and reflected light sets up interference patterns within the emulsion. At points where the incident and reflected waves cancel each other, the film will not be exposed. At points where the incident and reflected waves reinforce each other, exposure in the film occurs. This action can be visualized as standing waves being set up within the film emulsion. As the incident wave proceeds through the emulsion, the reflected wave will either aid or cancel it, causing periodic layers or areas in the emulsion to be exposed. The separation between the exposed layers in the emulsion will be one-half wavelength. It would appear desirable to have a number of layers exposed within the emulsion in order to reproduce satisfactorily.

A Lippmann plate is viewed by reflected light. That is, light is projected onto the plate at an angle so that it is reflected by the plate. Interference patterns are formed, since the incident light is reflected by the different layers. These interference patterns produce an exact replica of the original imaging light, both in form and in color. It is apparent that the Lippmann emulsion not only provides the high resolution required for optical processing but adds a new dimension of color. This additional capability allows the film to be used as a storage medium as well as a light modulator. Therefore, further investigation of the Lippmann emulsion is very much justified and necessary.

Initial attempts to produce the Lippmann emulsion at Goddard were partially successful. However, the basic goals of this study have not been fully achieved. The control of the many variables necessary to achieve reproducible results with any emulsion is difficult. In order to achieve the maximum capability from any film, a proper match between exposure and developing is required. At Goddard, the additional opportunity of controlling the process of making the Lippmann emulsion is possible. Thus several new parameters have been added. For example, it has been found that slight changes in certain ingredients of the emulsion will dramatically change the spectral sensitivity of the film. In addition, slight changes in the time of mixing various ingredients will dramatically affect the grain size and the speed of the film. Although these processes are somewhat critical, it is felt that they can be standardized and that with proper equipment, the reproducibility of results can be assured.

In his work, Dr. Lippmann used a physical developer instead of the usual chemical developer. The basic difference between a physical developer and a chemical developer is that a physical developer supplies most of the silver for forming the negative image, whereas in the normal chemical development the emulsion is the sole source of silver. That is, with a physical developer the solution is put into contact with the exposed film and silver from the developer solution is deposited on areas of the film which have been acted upon by light. The grain of an image formed by physical development can be finer than that formed by chemical development.

Both chemical and physical developers have been used on limited samples of the Lippmann emulsion produced at Goddard. No observable differences were found between the results obtained using these two methods. Limited samples of normal films were developed chemically and physically. The results of comparison were not conclusive, although it appeared that the physically developed samples were slightly superior. The purpose of these initial tests was not to establish the superiority of one type of developer over the other, but to develop standardized procedures so that the effects of varying parameters could be studied systematically. The technique for reversing film was also studied using the normal reversal procedures outlined in the previous subsection.

Initial efforts to standardize procedures using commercial films have been successful. Future efforts will be directed to similar standardization of procedures for the Lippmann emulsion. As stated previously, the Lippmann emulsion is exposed so that the image produced is in the form of layers. The layers should be as distinct as possible and as numerous as possible. In a very thick emulsion this may present a problem. For example, as the incident light passes through a thick emulsion, the light will be attenuated and upon reflection there may not be enough energy remaining to produce clear interference patterns with the incident light. It is also necessary to consider the problems involved in the development of a thick emulsion. To completely develop the emulsion the developer must penetrate the emulsion to reach all the silver salts. It is quite apparent that the underlayers of the emulsion will be acted upon at a later time than the top layers because of the time required for the developer to penetrate the emulsion. In the future, when the problems of producing thick emulsions which will not attenuate the light have been solved, the problem of obtaining uniform developer action must be considered. One solution might be to cool the developer to a point where it is no longer active and allow it to soak into the film until penetration is complete. At this time the film and developer can be heated rapidly by electronic induction heating. This procedure may be possible, but at the present time the investigation of this technique cannot be considered.

As mentioned above, attempts to produce very thick Lippmann emulsion films at Goddard have been successful. However, upon exposure the color rendition was not entirely satisfactory. It is evident that more control is necessary to obtain uniformity and standardization of the process for coating the emulsion as well as of the techniques for exposing and processing this type of emulsion.

The importance of the Lippmann emulsion in optical data processing cannot be overemphasized. It is important to realize that not only is the resolution more than adequate, but the possibility of color adds immeasurably to the potential applications in data processing. Consider, for example, a spot on the film which is to be used as a memory source. This point could be exposed to contain a desired bit of information such as a "1" or a "0." In addition to the "1" or "0" there is the possibility of storing an array of light frequencies (or colors) in this one spot.

When white light is projected onto an exposed Lippmann emulsion, the reflected light corresponds to the original frequency (color) of the light which exposed the emulsion. The frequencies of light reflected from a point on a Lippmann emulsion will be the same as those that exposed the point. It is possible, therefore, to expose a single point to three different frequencies of light and after development this point will reflect all three of the original colors. By selective filtering of the reflected light it is possible to determine whether any specified color is present. The presence of these three colors at one point could represent a specific number of information bits. The use of this characteristic of the Lippmann emulsion as a computer memory (or storage device) seems within the realm of feasibility.

Bragg Effect

It was pointed out in the above subsection that, when white light is projected onto an exposed Lippmann emulsion, the wavelength (color) of the reflected light will correspond to the wavelength (color) of the original exposing light. This color characteristic of the Lippmann process is explained by the Bragg effect, which is given by the equation

$$2d \cos \theta = n\lambda ,$$

where

d = spacing of reflecting surfaces (exposed layers),

θ = angle of incident and reflected light with respect to the normal to the surface,

λ = wavelength of incident and reflected light,

n = integer.

This equation states the necessary condition for constructive interference of reflected light from multiple layers. As shown in Figure 28, the effects of refraction at the surface of the emulsion are neglected and the angle of reflection is assumed equal to the angle of incidence. In our discussions we will consider only the value of $n = 1$, so that the Bragg effect will be written as

$$2d \cos \theta = \lambda .$$

Our discussion will also be restricted to the case of a Lippmann emulsion exposed by collimated monochromatic light incident normal to the emulsion surface. The exposed layers in this case will be parallel to the emulsion surface and will have spacing $d = \lambda_E/2$ (λ_E = wavelength of exposing light). Substituting this expression for d in the Bragg equation (for $n = 1$) gives

$$\lambda_E \cos \theta = \lambda .$$

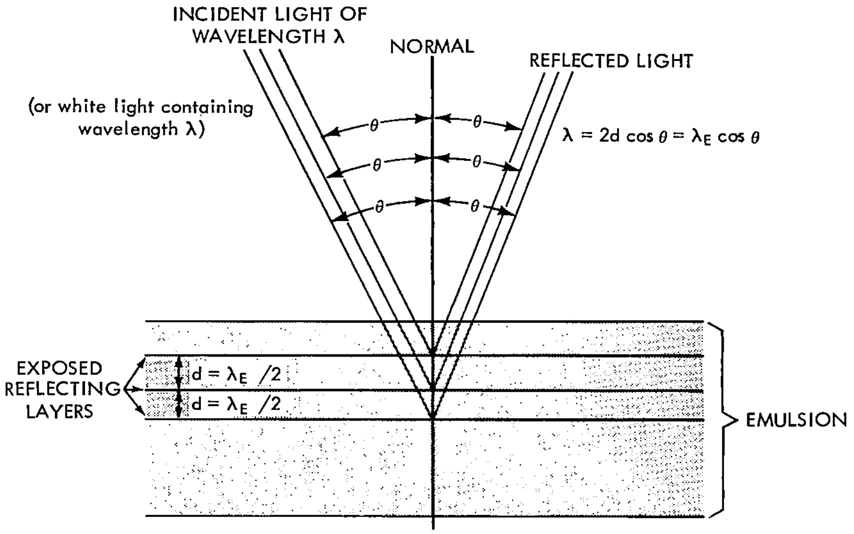

INCIDENT LIGHT OF WAVELENGTH λ
(or white light containing wavelength λ)
NORMAL
REFLECTED LIGHT
λ = 2d cos θ = λE cos θ
θ
EXPOSED REFLECTING LAYERS
d = λE /2
d = λE /2
EMULSION

Figure 28—Bragg effect.

One interpretation of this equation is that, for an emulsion exposed by light of wavelength λ_E to reflect light incident at an angle θ, the incident light must have wavelength λ as given by the equation. For example, consider white light (almost all frequencies) incident normal to the surface of the emulsion. In this example $\theta = 0$ and the wavelength of reflected light is

$$\lambda = \lambda_E \cos 0^\circ = \lambda_E .$$

This result corresponds to the statements made above that when white light is projected onto an exposed Lippmann emulsion the wavelength λ (color) of the reflected light will correspond to the wavelength λ_E (color) of the original exposing light. It is important to note that this statement is valid only when the white light is incident upon the emulsion at the same angle as the exposing light.

Bragg's equation can be rearranged to give

$$\lambda_E = \frac{\lambda}{\cos \theta} .$$

The application of this form of Bragg's equation can be explained by considering the case of white light incident at some angle θ unequal to zero. The light wavelength λ reflected at the angle θ can be detected. Using the detected value of λ and the known value of θ, the above equation gives the wavelength of the light used to expose the emulsion.

In practice, detecting the wavelength of the reflected light may be difficult. An alternative application of the above expression is the use of a monochromatic light source of known wavelength λ. By varying the angle of incidence and detecting at what angle θ reflection occurs, the value of the exposing wavelength λ_E can be determined from the equation $\lambda_E = \lambda/\cos \theta$. When the wavelength

λ of the monochromatic light source is known and the angle θ can be measured, the two values necessary to determine λ_E can be substituted into the equation.

The principles discussed above can be applied to information storage using a Lippmann emulsion. Consider a Lippmann emulsion exposed by three separate wavelengths λ_{E1}, λ_{E2}, and λ_{E3}. Each of these wavelengths can be treated as three independent cases by Bragg's equation. When an exposed emulsion is scanned by a monochromatic light beam of wavelength λ, reflection will occur for three angles θ_1, θ_2, and θ_3 given by the three equations

$$\cos\theta_1 = \frac{\lambda}{\lambda_{E1}},$$

$$\cos\theta_2 = \frac{\lambda}{\lambda_{E2}},$$

$$\cos\theta_3 = \frac{\lambda}{\lambda_{E3}}.$$

If three detectors (e.g. photocells) are arranged so that each one would detect reflected light at one of the angles θ_1, θ_2, or θ_3, an output from a given detector would indicate that the respective wavelength was used to expose the area being examined. In this configuration the incident monochromatic light beam can be moved to vary the incident angle (serial readout), or three separate light beams incident at the appropriate angles (parallel readout) can be used. The presence or absence of a particular wavelength can be used to represent an information bit of "1" or "0." The multiple-color possibilities could be used for separate bits, e.g. the three color cases described above could be used for three-bit codes.

The color principles discussed in the examples above point out some of the potential applications for the Lippmann emulsion as a storage medium. It is for these potential capabilities that the Lippmann emulsion is considered to be important for future data processing considerations.

Photographic Techniques for Optical Processing

A photographic *negative* is produced by exposing, developing, and fixing the photographic film. A photographic *positive* can be made from the negative by direct contact printing. Contact printing is accomplished by placing the developed negative over and in direct contact with an unexposed film and exposing the unexposed film by projecting light through the negative. The ability of developed photographic images to block the passage of light can be used as a measure of the photographic quality of the reproduced image. The ratio of the intensity (I_T) of the light transmitted through the film to the intensity (I_0) of the incident light is a measure of the photographic film's ability to transmit light. This ratio (I_T/I_0) has previously been defined as the transmittance (T). The ratio of incident light to transmitted light is defined as opacity (I_0/I_T). The common logarithm of the opacity is the optical density (or density). The equation defining the density is

$$\text{Density} = \log_{10} \frac{I_0}{I_T} = -\log_{10} T .$$

Now assume that it is possible to produce photographic film such that only a single layer of metallic silver salts is developed. When this developed film is placed in front of a light source, it will reduce the intensity of the incident light (I_0) by the factor T_1 (i.e. $I_T = T_1 I_0$). If a second identical film were placed over the first, the intensity of the transmitted light would be

$$I_T = (T_1)^2 I_0 .$$

When n identical layers are used, the intensity of the transmitted light will be

$$I_T = (T_1)^n I_0 .$$

Following this development, assume a photographic emulsion is made up of n different layers of metallic silver, each having transmittance T_1 and corresponding density D_1 as given by the above equation. The total transmittance is then

$$T = I_T/I_0 = (T_1)^n$$

and the total density is

$$D = -\log_{10} T = -\log_{10} \left[(T_1)^n\right] = n\left(-\log_{10} T_1\right) = n D_1 .$$

It is evident from this last equation that the density is proportional to the number of layers (n) in the emulsion.

A photographic emulsion that is exposed to a constant intensity of light will have a developed image whose density will increase with increasing exposure time (within limits). The relationship between the density of the developed image and the exposure time is usually indicated by a graph called the characteristic curve. Figure 29 shows a typical characteristic curve for a photographic emulsion. The characteristic curve is obtained by plotting the density versus the common logarithm of the exposure. The exposure is determined by the product of the intensity of the light (I_0) and the time that the film is exposed to this light (t). From this we can see that if the exposure time is kept constant and the intensity of the exposing light is increased, the density of the developed image will be increased. The characteristic curve is the S-type commonly found in engineering. In Figure 30, the lower limit of the curve indicates the fog level. The photographic emulsion will not increase in density for exposure values less than "a." When the photographic film is exposed to

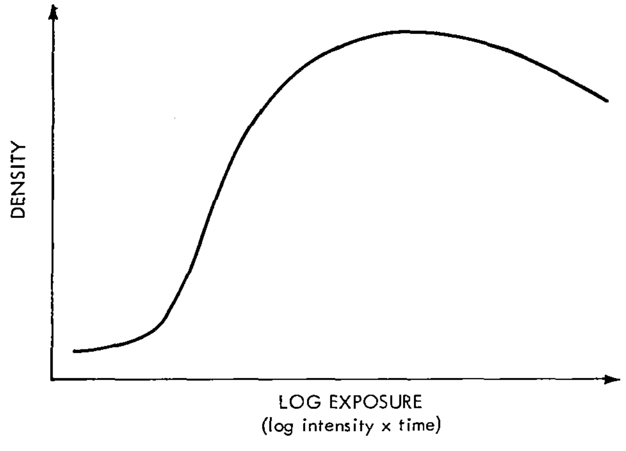


Figure 29—Typical characteristic curve for a photographic emulsion.

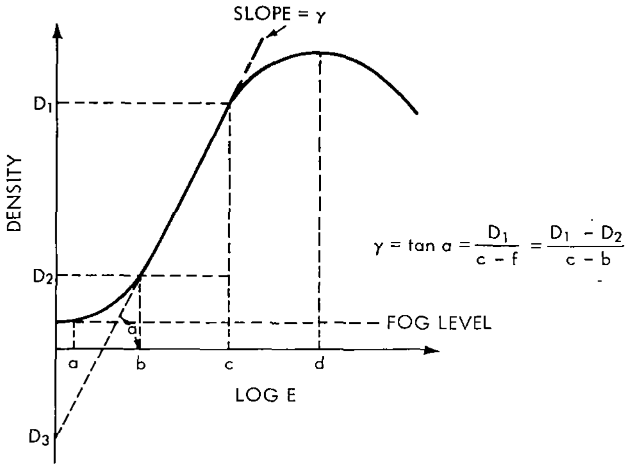


Figure 30—Characteristic curve of a photographic emulsion.

values between "a" and "b," the change in density will be non-linear. This area is usually considered the area of underexposure. Exposure values between "b" and "c" will result in a *linear* increase in density with respect to the logarithm (base 10) of exposure. Exposure in the area "c" to "d" results in a very slight changes in density with exposure (i.e. slight increase in density for increased exposure). This area of the curve is generally considered the area of overexposure. The area of the curve corresponding to an exposure greater than "d" shows a decrease in density with increasing exposure time. This region of the curve is usually associated with what is called solarization of the film.

On the straight-line portion of the curve (i.e. corresponding to exposures between "b" and "c"), the change in density is proportional to the change in the logarithm of exposure. From Figure 30 we can see that if the straight-line portion of the curve is extended it intersects the log E axis at point f. This point f is often used as a measure of the film speed. The slope of this line is given by

$$\gamma = \tan \alpha = \frac{D_1}{c - f} = \frac{D_1 - D_2}{c - b},$$

where γ is the slope of the straight-line portion of the curve.

Note: The angle relation $\alpha = \gamma$ holds only when the two axes (density vs log E) are plotted to the same scale. When the two axes are not to the same scale, $\tan \alpha$ will not equal the slope γ. However, the slope γ will always equal $\Delta D/\Delta \log_{10} E$.

The characteristic curve of photographic emulsions depends not only upon the nature of the emulsion, but also upon the method by which it was developed. Different emulsions will generally have different characteristic curves and therefore will have different γ. Each of these emulsions in turn can have a different γ depending on the development process. The type of developing solutions used, the time of processing, the temperature of the developer, all in turn affect the value of γ. Figure 31 shows a family of characteristic curves for a single type of film for which only the development time has been changed. It can be seen from the plots in Figure 31 that in general the γ will increase as the development time increases.

Linearizing Photographic Film

From the characteristic curves of photographic films we can see that the photographic process is extremely nonlinear. The very fact that the density is plotted against the *logarithm* of the exposure is an indication of this nonlinearity. There are methods by which the photographic process can be linearized. The following discussion will illustrate one of these methods.

Consider the case where *amplitude* variations of light corresponding to a given signal E_s are desired. Initially a photographic film is exposed to a varying light *intensity* corresponding to the signal E_s, as shown in Figure 32. Assuming the exposure of the photographic film is accomplished in the linear region of the characteristic curve shown in Figure 30, the density will be given by the equation

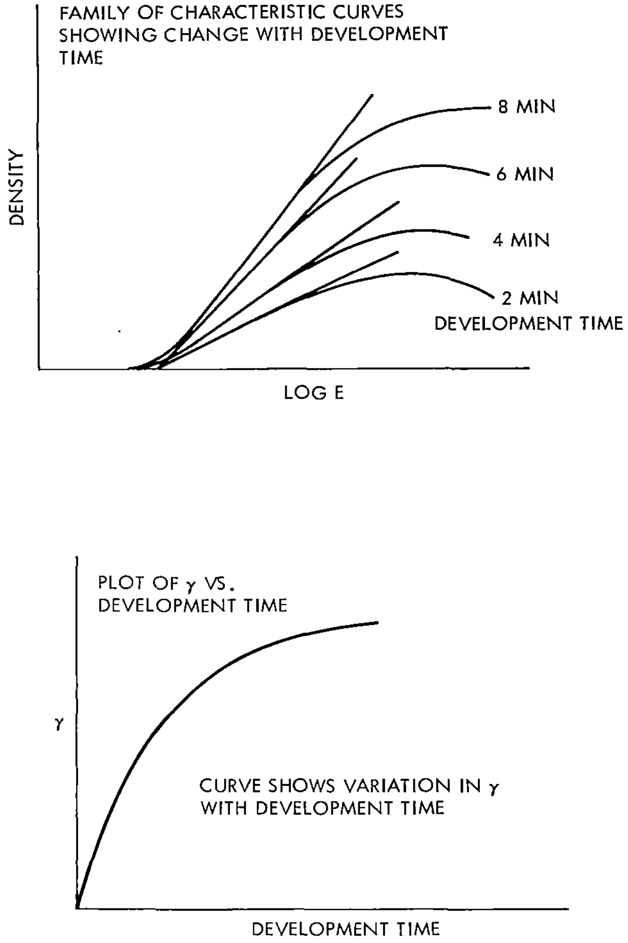


Figure 31—Effect of development time on γ.

$$D_1 = \gamma_1(c-f) = \gamma_1\left(\log_{10}E_c - \log_{10}E_f\right),$$

where

E_c = product of intensity and time (exposure) corresponding to point c.

E_f = product of intensity and time corresponding to point f.

$$D_1 = \gamma_1\left(\log_{10}t_1 I_s - f\right), \qquad (1)$$

where

D_1 = the density of the exposed film,

I_s = the light-intensity variation corresponding to E_s,

t_1 = the exposure time,

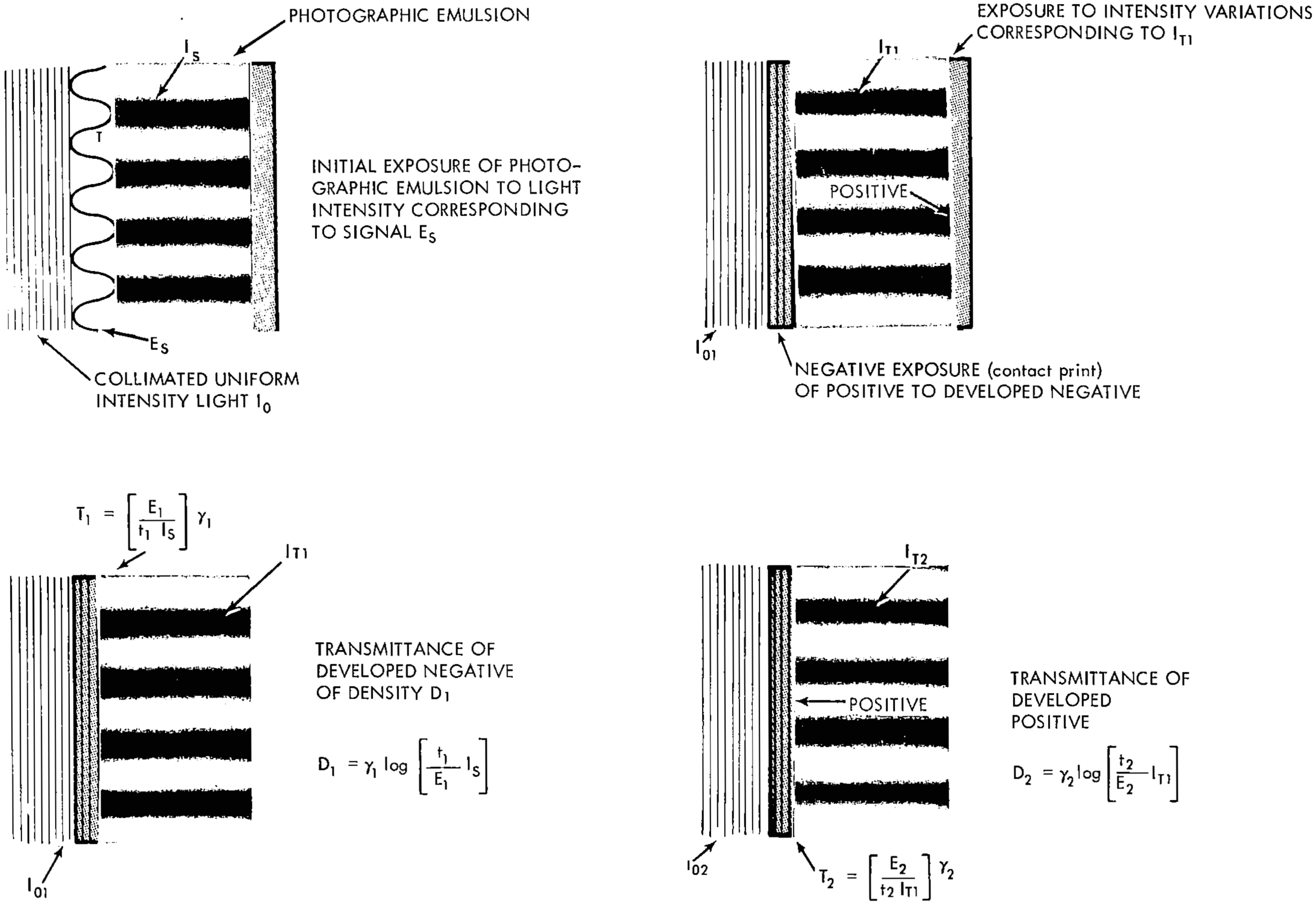

PHOTOGRAPHIC EMULSION
I_S
T
E_S
COLLIMATED UNIFORM INTENSITY LIGHT I_0
INITIAL EXPOSURE OF PHOTOGRAPHIC EMULSION TO LIGHT INTENSITY CORRESPONDING TO SIGNAL E_S
EXPOSURE TO INTENSITY VARIATIONS CORRESPONDING TO I_{T1}
I_{T1}
POSITIVE
I_{01}
NEGATIVE EXPOSURE (contact print) OF POSITIVE TO DEVELOPED NEGATIVE
$T_1 = \left[\frac{E_1}{t_1 I_S}\right]\gamma_1$
I_{T1}
TRANSMITTANCE OF DEVELOPED NEGATIVE OF DENSITY D_1
$D_1 = \gamma_1 \log\left[\frac{t_1}{E_1} I_S\right]$
I_{01}
I_{T2}
POSITIVE
I_{02}
$T_2 = \left[\frac{E_2}{t_2 I_{T1}}\right]\gamma_2$
TRANSMITTANCE OF DEVELOPED POSITIVE
$D_2 = \gamma_2 \log\left[\frac{t_2}{E_2} I_{T1}\right]$

Figure 32—The linearized photographic process.

γ_1 = the slope of the straight-line portion of the characteristic curve as determined by the type of emulsion, type of exposure, and development process,

$f = \log_{10} E_f$ = constant noted above.

Let E_1 be the exposure value corresponding to point f (i.e. $f = \log_{10} E_1$, or $E_f = E_1$). Substituting for f in Equation 1 and rearranging terms gives

$$D_1 = \gamma_1 \log_{10}\left(\frac{t_1}{E_1} I_s\right) . \tag{2}$$

Since $D = \log_{10} T$, the transmittance of this film will be

$$T_1 = 10^{-D_1} = 10^{-\gamma_1 \log_{10}[(t_1/E_1)I_s]} = 10^{\log_{10}[(E_1/t_1)(1/I_s)]^{\gamma_1}} ,$$

or

$$T_1 = \left(\frac{E_1}{t_1 I_s}\right)^{\gamma_1} . \tag{3}$$

It is quite obvious from Equation 3 that the transmittance is not linearly proportional to the amplitude variation of the signal E_s (or I_s). After development, the transmittance of this film can also be determined by

$$T_1 = \frac{I_{T1}}{I_{01}} , \tag{4}$$

where

I_{01} = intensity of uniform incident light,

I_{T1} = intensity of transmitted light.

Rearranging Equation 4 gives the transmitted light intensity as

$$I_{T1} = T_1 I_{01} . \tag{5}$$

Substituting Equation 3 into Equation 5 gives

$$I_{T1} = I_{01} T_1 = I_{01}\left(\frac{E_1}{t_1 I_s}\right)^{\gamma_1} . \tag{6}$$

Thus the intensity (I_{T1}) of the light transmitted through the negative is nonlinear with respect to I_s. For simplification let $k_1 = (E_1/t_1)^{\gamma_1}$ in Equation 6 to produce

$$I_{T1} = k_1 I_{01} I_s^{-\gamma_1} . \tag{7}$$

Now assume the negative is used to expose a second film (positive) by contact printing, as shown in the third diagram of Figure 32. As shown, the light exposing the second film must pass through the initial negative. If the incident light of constant intensity (I_{01}) illuminates the negative (second diagram of Figure 32), the intensity transmitted is I_1. The intensity of the light exposing the second film will be I_{T1}, as defined above. The density of this second film (after development) will be (using the form of Equation 2 and referring to diagram 3 of Figure 32)

$$D_2 = \gamma_2 \log_{10} \left(\frac{t_2}{E_2} I_{T1} \right) . \tag{8}$$

Continuing as above (Equation 3), the transmittance of this second film is

$$T_2 = \left(\frac{E_2}{t_2 I_{T1}} \right)^{\gamma_2} .$$

Letting $k_2 = (E_2/t_2)^{\gamma_2}$, the transmittance can be given as

$$T_2 = k_2 I_{T1}^{-\gamma_2} . \tag{9}$$

When this positive is developed and illuminated by uniform light intensity (I_{02}), the transmitted light intensity (I_{T2}) will be given by (see lower right-hand diagram, Figure 32)

$$I_{T2} = T_2 I_{02} = k_2 I_{02} I_{T1}^{-\gamma_2} . \tag{10}$$

Substituting for I_{T1} as given by Equation 7 gives

$$I_{T2} = k_2 I_{02} (k_1 I_{01})^{-\gamma_2} I_s^{\gamma_1 \gamma_2} . \tag{11}$$

Defining a new constant $K = k_2 I_{02} (k_1 I_{01})^{-\gamma_2}$, Equation 11 becomes

$$I_{T2} = K I_s^{\gamma_1 \gamma_2} . \tag{12}$$

Equation 12 shows that the intensity of light transmitted through the positive is proportional to the original signal light intensity raised to the $\gamma_1 \gamma_2$ power. When $\gamma_1 \gamma_2 = 1$, the intensity of the light transmitted through the positive (illuminated by a uniform light intensity) is proportional to the original light intensity.

The amplitude (A) of the light transmitted through the positive is by definition the square root of the intensity given by Equation 12 or

$$A = \sqrt{I_{T2}} = K_1^{1/2} I_s^{\frac{\gamma_1 \gamma_2}{2}} . \tag{13}$$

When $\gamma_1 \gamma_2 = 1$, the amplitude (A) is proportional to $I_s^{1/2}$. This means that, although the intensity variations in the transmitted light vary as the original signal intensity, the amplitude variations do not.

In an earlier discussion it was pointed out that this problem can be solved by using an original signal intensity which is the square of the desired signal. However, it is apparent from Equation 13 that when $\gamma_1 \gamma_2 = 2$ the *amplitude* of the light transmitted through the positive is proportional to the original signal intensity (desired signal). It is apparent from this discussion that a desired input signal can be represented by the intensity used to expose a film, and if the gamma product is set equal to 2 (i.e. $\gamma_1 \gamma_2 = 2$) the resulting amplitude transmission function will correspond to the desired signal.

Figures 33, 34, and 35 are graphical representations of the mathematics discussed above. It should be noted that the $\gamma_1 \gamma_2$ product is different for each figure, so that a graphical comparison

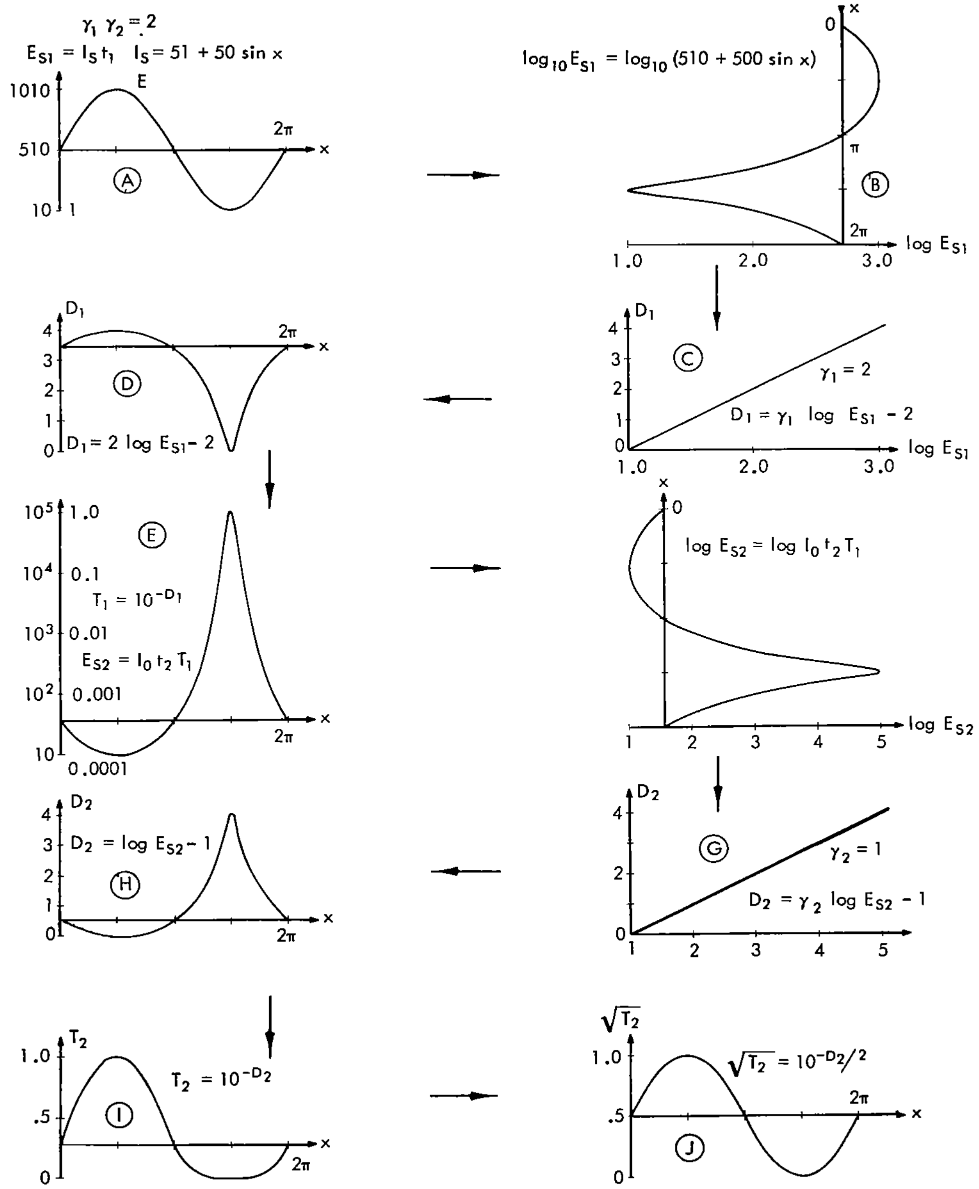


Figure 33—Photographic process with $\gamma_1 \gamma_2 = 2$.

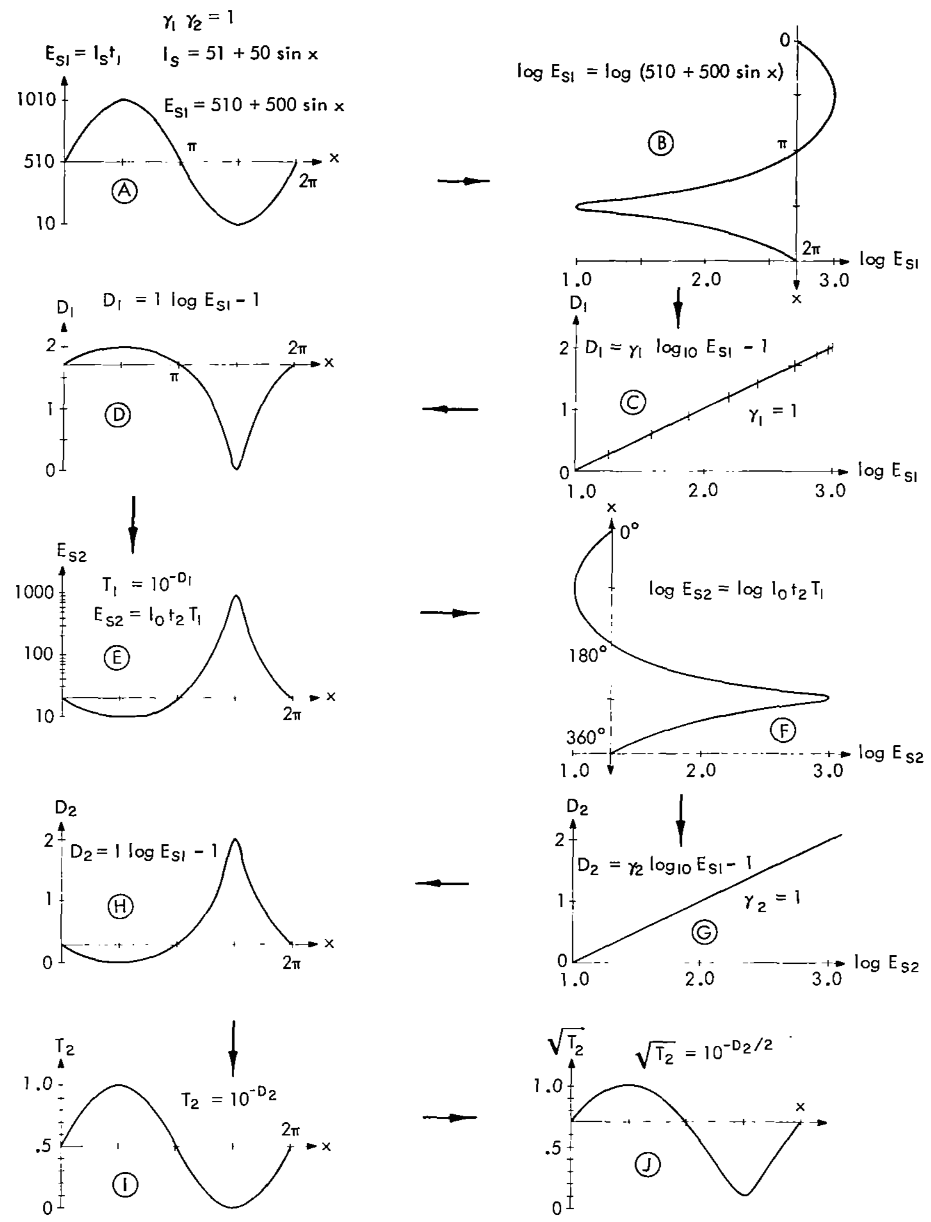


Figure 34—Photographic process with $\gamma_1 \gamma_2 = 1$.

of the effects of this term is possible. Only Figure 33, which shows the case where $\gamma_1 \gamma_2 = 2$, will be explained in detail, since the same explanation applies for all values of $\gamma_1 \gamma_2$.

Diagram A of Figure 33 shows a sine wave which is assumed to be an input signal for an optical processor. As shown earlier, an input signal must be represented by a transmission function for optical processing purposes. For the use of a photographic plate to implement the transmission function, Figure 33 shows the steps required to represent the input signal of diagram A as a transmission function. As in previous examples, only one dimension will be considered (x coordinate), but the basic processes apply to the two-dimensional case (x and y coordinates). Diagram A of

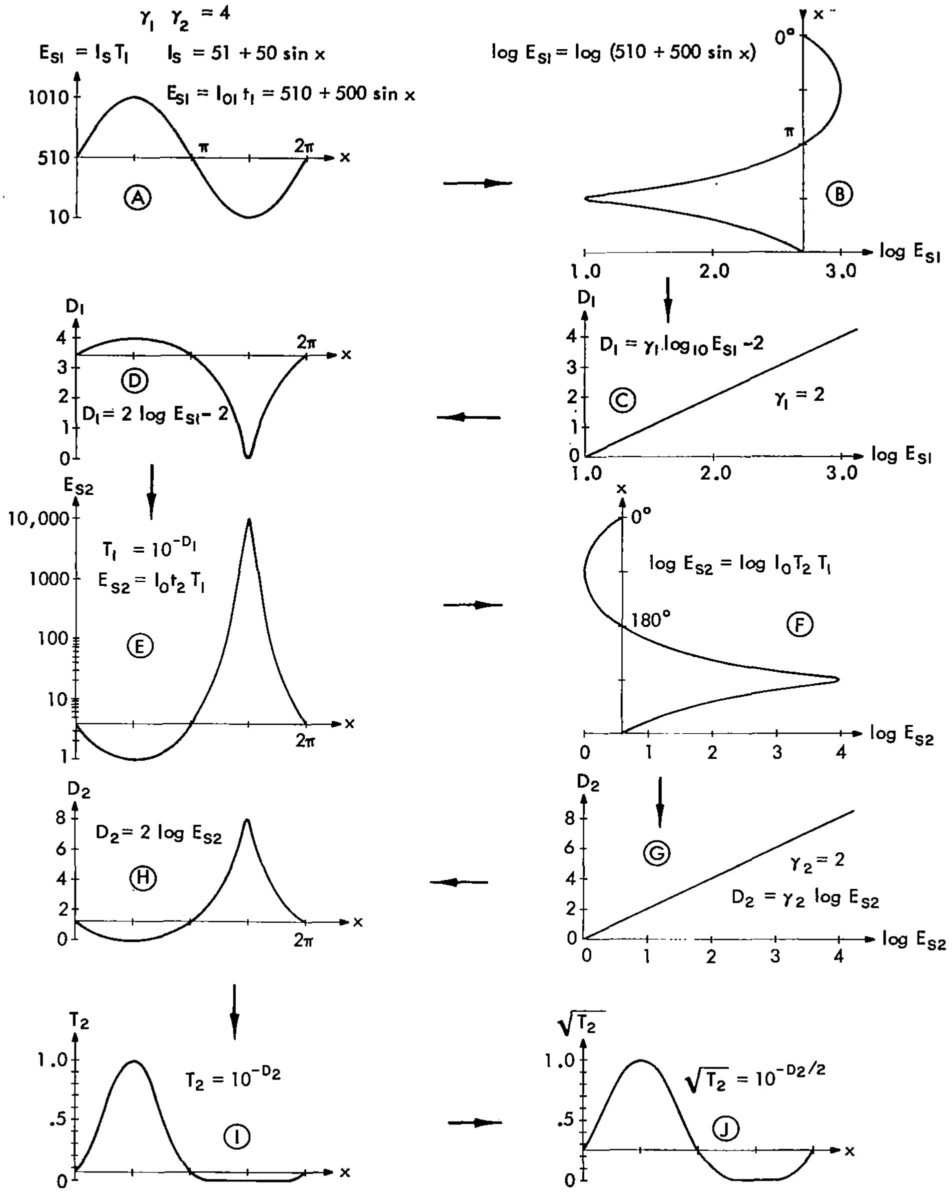


Figure 35—Photographic process with $\gamma_1\ \gamma_2 = 4$.

Figure 33 shows the input signal as a function of x represented by an intensity ($I_s = 51 + 50 \sin x$). This intensity signal is used to expose a photographic plate for 10 seconds ($t_1 = 10$). The exposure of the plate is given by the product $(t_1 I_s)$. Therefore, the exposure (E_{s1}) is given by

$$E_{s1} = t_1 I_s = 510 + 500 \sin x .$$

The waveform for this exposure (E_{s1}) is also given by diagram A, except that now the scale from diagram A has been multiplied by a factor of 10.

Diagram C represents the linear part of the characteristic curve for the film being exposed. It is assumed that the plate is developed so that $\gamma_1 = 2$ and $f = 1$ (see Figure 30), as shown. The characteristic curve is a plot of density versus the logarithm (base 10) of exposure. Therefore, in order to project values of exposure E_{s1} (diagram A) onto the characteristic curve of diagram C, it is necessary to plot $\log_{10} E_{s1}$ as a function of x as shown in diagram B. Using diagrams B and C, it is possible to graphically determine the density as a function of x shown in diagram D. For example, choose a value of x in diagram B, and from this point project a straight line vertically downward to intersect the straight line of diagram C. From this point of intersection, projecting a point horizontally to the chosen x coordinate in diagram D gives the value of the density for that value of x. Repeating this procedure for all values of x will give the complete density curve of diagram D. Diagram D shows that the variation in density is proportional to the common logarithm of the original light intensity (I_s).

Diagram E is a plot of the transmittance (T_1) corresponding to the density shown in diagram D ($T_1 = 10^{-D_1}$). The negative with transmittance (T_1) shown in diagram E can be used to expose a new plate by contact printing. If the exposing light used has uniform intensity I_0 (in this case 10^4), the exposure of the second plate will be given by the relation $E_{s2} = I_0 t_2 T_1$. This exposure E_{s2} is plotted in diagram E for $t_2 = 10$ seconds. The $\log_{10} E_{s2}$ is plotted in diagram F. Assume the second film is developed to a $\gamma_2 = 1$, as shown by the characteristic curve of diagram G. Projecting values for $\log_{10} E_{s2}$ onto the characteristic curve gives the plot of the density shown in diagram H. This procedure is the same as that used above for obtaining diagram D, i.e. the density D_2 shown in diagram H was produced in the second film by an exposure E_{s2} developed to a $\gamma_2 = 1$. The transmittance corresponding to the density of diagram H is shown in diagram I.

It has previously been shown that the transmission function $F(x)$ is equal to the square root of the transmittance. In Figure 33, diagram J shows the transmission function obtained by taking the square root of the transmittance given in diagram I. Comparing the transmission function given by diagram J to the signal expressed as an intensity in diagram A, it is apparent that the transmission function corresponds to the desired input signal. It is important to note that the correspondence between diagram J and A in Figure 33 is due to the gamma product being equal to 2 ($\gamma_1 \gamma_2 = 2$).

Figure 33 shows that when $\gamma_1 \gamma_2 = 2$ the square root of the final transmittance of the film will correspond to the original exposing signal intensity, and therefore the final transmission function $F(x)$ will be equal to the original signal.

Figure 34 shows that when $\gamma_1 \gamma_2 = 1$ the final transmittance of the film will correspond to the original exposing signal intensity, and therefore the final transmission function $F(x)$ will be equal to the square root of the original signal.

Figure 35 shows that when $\gamma_1 \gamma_2 = 4$ neither the final transmittance nor the transmission function of the film will correspond to the original exposing intensity.

Comparing diagram J of Figures 33, 34, and 35 demonstrates the requirement for a product $\gamma_1 \gamma_2 = 2$. Although the same desired input signal was represented by the exposing signal intensity in each case, only when $\gamma_1 \gamma_2 = 2$ is the transmission function of the second film a representation of the desired signal.

OPTICAL DATA PROCESSORS

Spectrum Analyzer

The spectrum of the transmission function of the object located in the front focal plane is produced in the back focal plane of the lens. This spectrum is formed by the Fourier transform characteristic of diffraction. Sinusoidal transmission functions will be used in the examples below to demonstrate the basic principles of an optical spectrum analyzer.

A review of the basic principles involved in focusing the diffraction pattern formed by slits will be helpful in understanding the diffraction patterns formed by sinusoidal transmission functions. Figure 36 shows that the diffraction wavefronts produced by slits can be focused in the back focal plane of a lens. The zero-order wavefronts are focused to a point at the back focal point of the lens. The zero-order image point is usually referred to as the dc component of the diffraction pattern (or spectrum). The two first-order wavefronts will be focused to two points located symmetrically about the zero-order image point. The distance between the zero-order and first-order image points will be proportional to the spatial frequency (1/S) of the slits, where S is the distance between slits. It is important to note that the images formed in the back focal plane are *points* whose locations are dependent upon the spatial frequency of the slits.

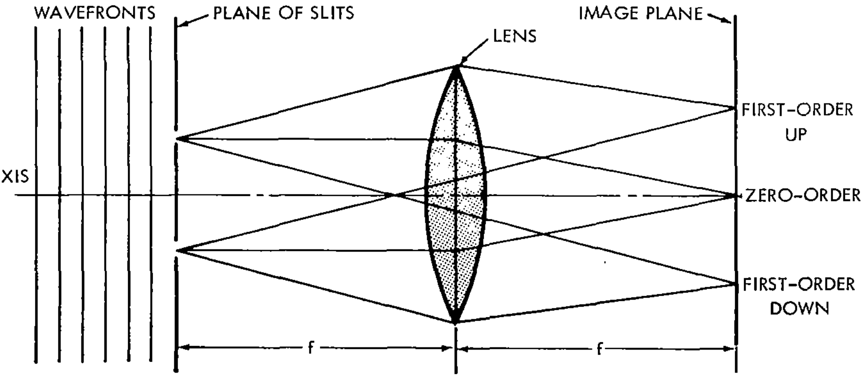


Figure 36—Diffraction pattern focused in focal plane.

In terms of a Fourier spectrum, the diffraction pattern of slits can be easily related to diffraction by sinusoidal transmission functions. The zero-order term, or dc component, corresponds to a bias. In previous discussions it has been shown that a bias must be added to a sine wave to avoid negative values of the transmission function. It is this bias which determines the amplitude of light at the zero-order point. The first-order term of slit diffraction corresponds to the fundamental frequency of the Fourier series which can be used to represent the slits. The wavelength of this fundamental frequency is equal to the distance between slits. The higher-order terms of a slit-diffraction pattern correspond to the higher-frequency components in the Fourier expansion. A sinusoidal transmission function (true single-frequency sine wave) will produce three spectral points. The dc term corresponding to the film bias will appear at the back focal point. Two points located symmetrically about this dc point will correspond to the frequency of the sine wave. The distance of these points from the dc point will increase if the frequency of the transmission function is increased (slit spacing is reduced). These two points correspond to the images of the first-order diffraction wavefronts of slits with spacing equal to the wavelength of the sine wave.

The basic operation of an optical spectrum analyzer will now be demonstrated by examples illustrated in Table 4. It should be kept in mind that the spectral points of a transmission function representing a sine wave (plus bias) correspond to the zero- and first-order diffraction points of slits with spacing equal to the wavelength of the sine wave. Figure 37 shows the normal spectrum analyzer configuration.

Example I

Ia shows a normal oscilloscope presentation of a 100-cps sine wave.

Ib shows the intensity modulation produced when the same 100-cps sine wave is applied to the z-axis of an oscilloscope. The darkened areas correspond to maximum light output.

Ic shows a representation of the transmission function of a film exposed to the intensity modulation given by Ib and processed with a gamma product of 2.

Id shows the spectral image that would appear in the image plane of the spectrum analyzer shown in Figure 37. The film with a transmission function shown in Ic is placed in the object plane (front focal plane) of the lens. It is to be noted that the spectral-image points are formed on a line parallel to the axis of the transmission function.

Example II

Example II shows a 150-cps sine wave represented as the normal oscilloscope pattern and as an intensity modulation. The spectral image formed by the transmission function produced by the intensity modulation is shown in IId. The difference between Examples I and II is effectively a change in scale.

Example III

Example III shows a resultant wave which is formed by the sum of a 100-cps sine wave and a 150-cps sine wave. The intensity modulation of this resultant wave is shown in IIIb, and the transmission function is shown in IIIc.

Each of the sine wave components will produce spectral-image points at different locations in the image plane. These points in the image plane are shown in IIId. The operation of a spectrum

Table 4

Examples Illustrating Diffraction Patterns from Various Slit Orientations.

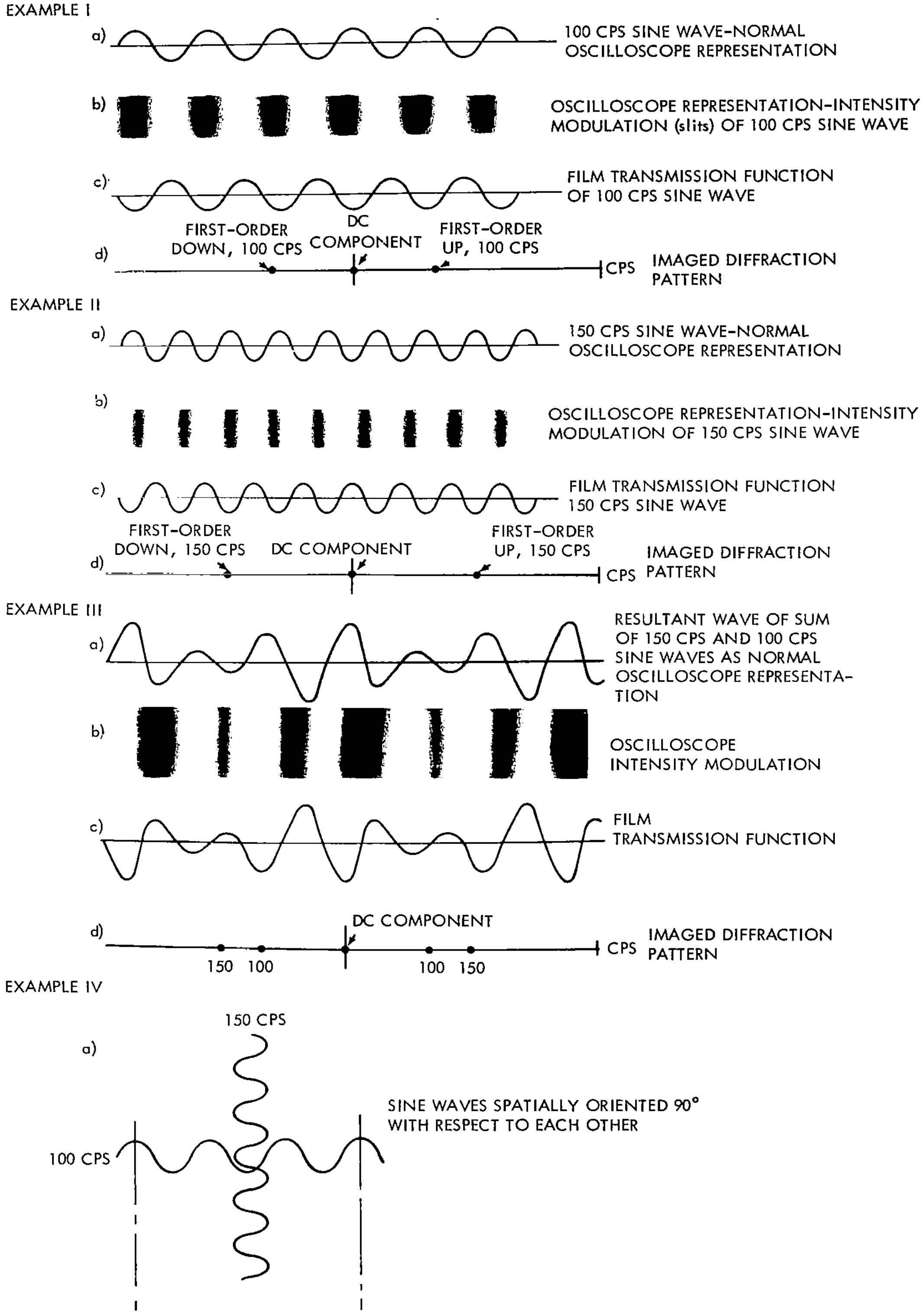
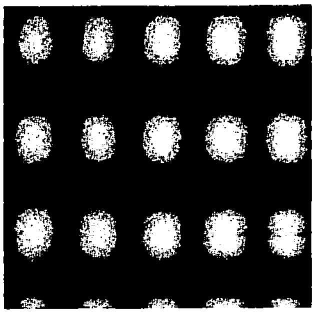

EXAMPLE I
a)
100 CPS SINE WAVE-NORMAL OSCILLOSCOPE REPRESENTATION
b)
OSCILLOSCOPE REPRESENTATION-INTENSITY MODULATION (slits) OF 100 CPS SINE WAVE
c)
FILM TRANSMISSION FUNCTION OF 100 CPS SINE WAVE
d)
FIRST-ORDER DOWN, 100 CPS
DC COMPONENT
FIRST-ORDER UP, 100 CPS
CPS
IMAGED DIFFRACTION PATTERN
EXAMPLE II
a)
150 CPS SINE WAVE-NORMAL OSCILLOSCOPE REPRESENTATION
b)
OSCILLOSCOPE REPRESENTATION-INTENSITY MODULATION OF 150 CPS SINE WAVE
c)
FILM TRANSMISSION FUNCTION 150 CPS SINE WAVE
d)
FIRST-ORDER DOWN, 150 CPS
DC COMPONENT
FIRST-ORDER UP, 150 CPS
CPS
IMAGED DIFFRACTION PATTERN
EXAMPLE III
a)
RESULTANT WAVE OF SUM OF 150 CPS AND 100 CPS SINE WAVES AS NORMAL OSCILLOSCOPE REPRESENTATION
b)
OSCILLOSCOPE INTENSITY MODULATION
c)
FILM TRANSMISSION FUNCTION
d)
DC COMPONENT
150
100
100
150
CPS
IMAGED DIFFRACTION PATTERN
EXAMPLE IV
a)
150 CPS
100 CPS
SINE WAVES SPATIALLY ORIENTED 90° WITH RESPECT TO EACH OTHER

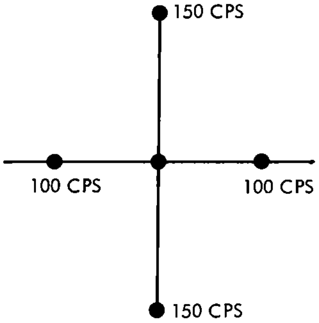
Table 4 (Continued)

b)



INTENSITY MODULATION IN TWO DIMENSIONS

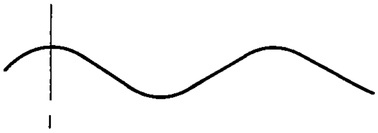
c)



IMAGED DIFFRACTION PATTERN

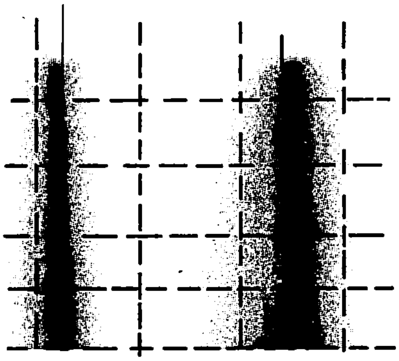
EXAMPLE V

a)



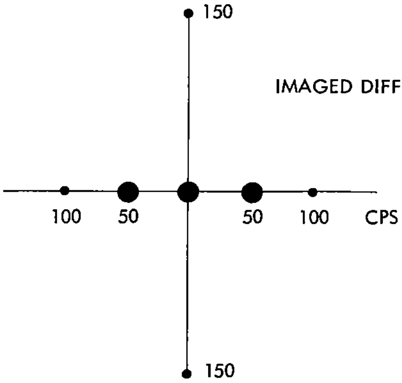
50 CPS
SINE WAVE

b)



50 CPS INTENSITY MODULATION SUPERIMPOSED ON BACKGROUND OF REFERENCE PAPER

c)

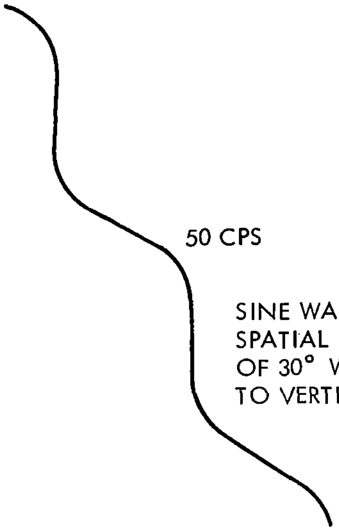


IMAGED DIFFRACTION PATTERN

Table 4 (Continued)

EXAMPLE VI

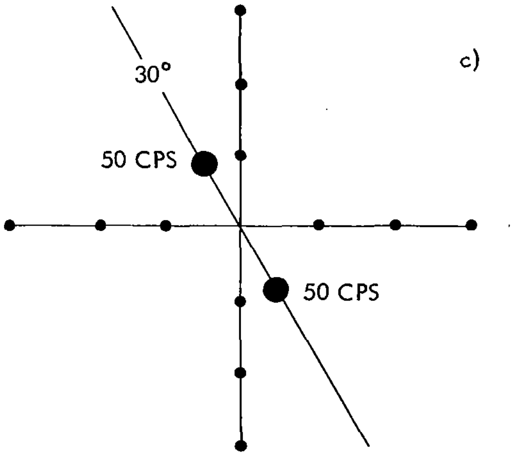
a)



SINE WAVE WITH SPATIAL ORIENTATION OF 30° WITH RESPECT TO VERTICAL

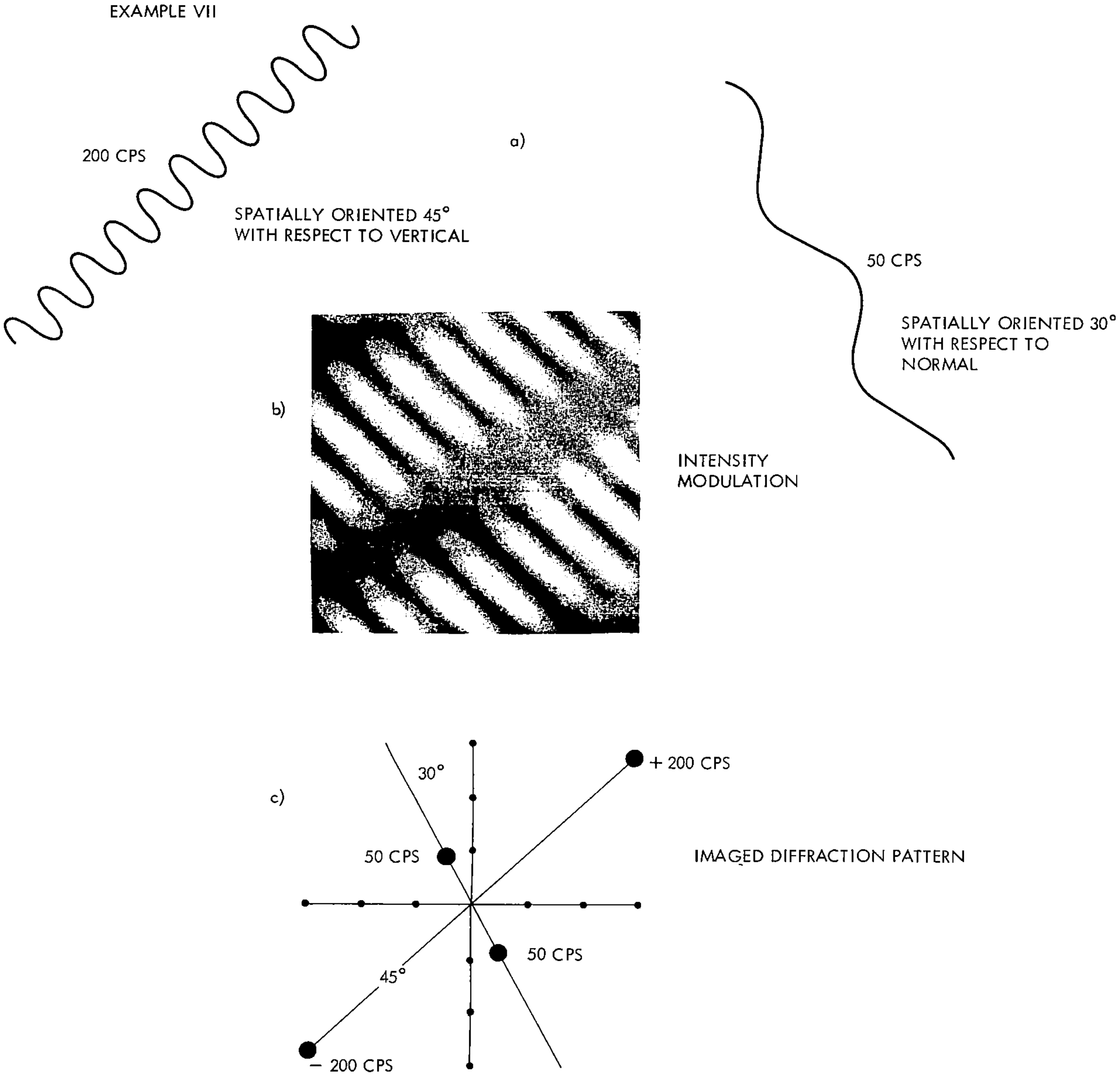
b)

INTENSITY MODULATION SUPERIMPOSED ON REFERENCE GRAPH PAPER



c) IMAGED DIFFRACTION PATTERN

Table 4 (Continued)


EXAMPLE VII
200 CPS
SPATIALLY ORIENTED 45° WITH RESPECT TO VERTICAL
a)
50 CPS
SPATIALLY ORIENTED 30° WITH RESPECT TO NORMAL
b)
INTENSITY MODULATION
c)
30°
+ 200 CPS
50 CPS
IMAGED DIFFRACTION PATTERN
50 CPS
45°
− 200 CPS

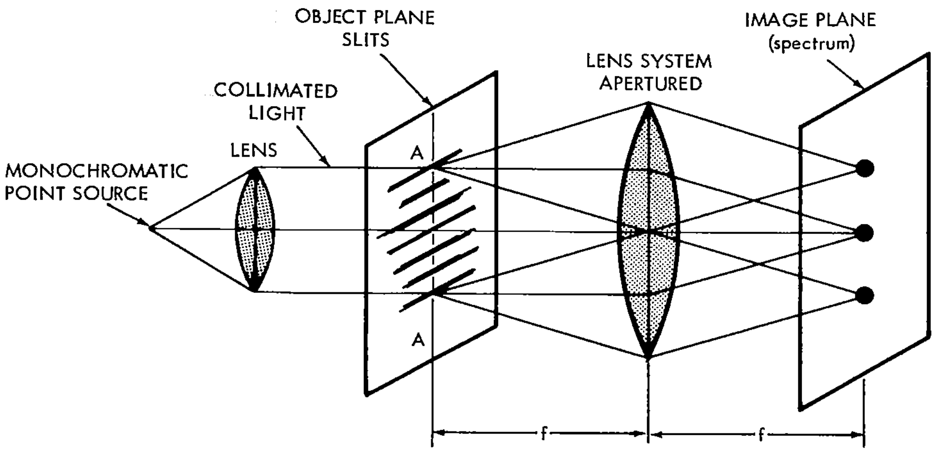


Figure 37—Optical spectrum analyzer.

analyzer is now quite apparent. With suitable calibrations we can determine the frequency components of any desired input signal.

Example IV

Example IV shows a 100-cps sine wave perpendicular to a 150-cps sine wave. The intensity modulation of these two sine waves is shown in IVb. The similarity between the intensity modulation and graph paper is quite apparent. The diffraction pattern of a transmission function corresponding to this intensity modulation is shown in IVc.

This example is significant for two reasons. The transmission function expressed horizontally produces a horizontal diffraction pattern whereas the transmission function expressed vertically produces a vertical diffraction pattern. The similarity of the intensity modulation formed by vertical and horizontal modulations to graph paper suggests a convenient method of calibration.

Example V

Example V shows how graph paper (as the intensity modulation of Example IV) can be used as a background for a new intensity modulation signal which is to be analyzed. This background graph paper forms convenient calibration marks in the diffraction pattern. Vc shows that the diffraction pattern consists of the calibration dots of the background graph paper as well as dots corresponding to the 50-cps signal (Va) which is being analyzed.

Example VI

Example VI shows the same reference background graph paper of Example V with a 50-cps sine wave signal positioned diagonally. The diffraction pattern of this signal is shown in VIc. Note that, since the transmission function was represented along a diagonal, the spectral-image points of this function will also appear along a diagonal line. The diagonal line of the spectral-image points will be at the same angle as the diagonal line along which the transmission function was represented.

Example VII

Example VII shows background calibration graph paper and two signals positioned along diagonal lines. The 200-cps sine wave is along a line going up at an angle of 45° to the right, and the 50-cps sine wave is along a line going down at an angle of 60° to the right. The diffraction patterns formed by these signals are shown in VIIc. Again, it is clear that each diffraction pattern will be imaged on a line at the same angle as the line along which the respective transmission function is represented.

In the above examples it has been shown how an optical spectrum analyzer simultaneously determines the frequency components of any input waveforms. One apparent advantage of optical processing is the simultaneous processing of several signals. The conventional *electronic* spectrum analyzer can operate on only one input signal at a time. An additional factor is the definite time interval required for electronic processing, which can limit applications. The optical spectrum analyzer determines the spectrum instantaneously.

Spectrum Analyzer Performance

The time-bandwidth product can be used as a measure of the storage capacity of a memory device, or as a measure of the information content of a signal. The time-bandwidth product can be defined as the number of cycles of a given frequency which appear in a given time interval. Mathematically the time-bandwidth product is defined by the equation

$$\mathrm{TB} = tf ,$$

where

TB = time-bandwidth product,

t = time interval,

f = frequency.

As an example, consider a signal with a frequency of 1,000 cps and 10 milliseconds time duration. The time-bandwidth product will be

$$\mathrm{TB} = tf = 10 \times 10^{-3} \times 10^{3} = 10 .$$

In optical systems the signals are normally represented on film as a function of distance rather than of time. A scale factor k (length/sec) determines the linear relation between a time-dependent signal and its spatial representation. Using this scale factor, the time t represented by a distance d will be

$$t = \frac{d}{k}, \tag{14}$$

where

d = distance (spatial coordinate),

t = time (time coordinate),

k = distance/unit time.

Likewise a frequency in time f (cycles/second) represented by a spatial frequency f_x (cycles/unit length) will be

$$f = k f_x . \tag{15}$$

Substituting Equations 1 and 2 in the time-bandwidth equation gives

$$TB = ft = k f_x \frac{d}{k} = d f_x .$$

Thus in either space or time representations the time-bandwidth product is given by the product of signal frequency and signal length. This indicates that the time-bandwidth product provides a means for comparing the performance of analogous optical and electronic systems.

The frequencies and time intervals in an electronic system are normally restricted by system parameters. In optical systems the spatial frequencies representable are limited (upper limit) by the resolution of the system, and the spatial length is limited by the system apertures. Determining the respective time-bandwidth products of the analogous systems therefore provides a relative evaluation of the systems.

We will now consider the maximum time-bandwidth product for an optical system. The resolution (R) (normally expressed as lines/unit length) of an optical system specifies the maximum spatial frequency (cycles/unit length) which can be used in the system. Since optical Fourier transformation requires small diffraction angles and the aperture is usually smaller than the lens, it can be assumed that the length of signal will be limited by the aperture, as shown by the dimension A-A in Figure 37. We can now express the maximum time-bandwidth product in terms of resolution (R) and aperture length (d):

$$TB_{MAX} = Rd,$$

where

R = resolution of optical system (film and lens),

d = aperture dimension.

Per common practice, 35 mm has been adopted in our work as a standard size for input signal films. The practical resolution of the film and optical system to be used will be about 35 lines per mm. The 35-mm format limits the effective aperture to 24 × 36 mm. Along the 36-mm dimension, the maximum time-bandwidth product will be

$$TB_{MAX} = Rd = 35 \times 36 = 1,260 \ .$$

As an example, consider a signal 10 milliseconds long which is to be represented in the 36-mm aperture. The maximum frequency that can be represented for this time interval will be

$$f_{MAX} = \frac{TB_{MAX}}{t} = \frac{Rd}{t} = \frac{1,260}{10 \times 10^{-3}} = 126 \text{ K cps} \ .$$

In other words, with the maximum time-bandwidth product equal to 1,260, the maximum frequency (f_{MAX}) and time interval (t) which can be represented spatially in this optical system are specified by

$$f_{MAX}\, t = 1,260 \ .$$

It is important to note that for a specified maximum time-bandwidth product (e.g., $TB_{MAX} = 1,260$) only frequency-time interval products less than the maximum can be handled by the system (i.e. $TB \leqq TB_{MAX}$).

Optical Filtering

Figure 38 shows how the basic principles set forth in a spectrum analyzer can be used to develop an optical filter, by the addition of a single lens which images the spectrum into an image plane. We have seen how lens 1 acts as the spectrum analyzer, producing in its back focal plane (transform plane) the spectrum of the transmission function in its front focal plane (object plane). If the points of light in the transform plane are used as an object in the front focal plane of lens 2, the back focal plane (image plane) of lens 2 will contain the spectrum of the light in the transform plane. Lens 2 thus produces a transform of a transform which is in effect the inverse transform,

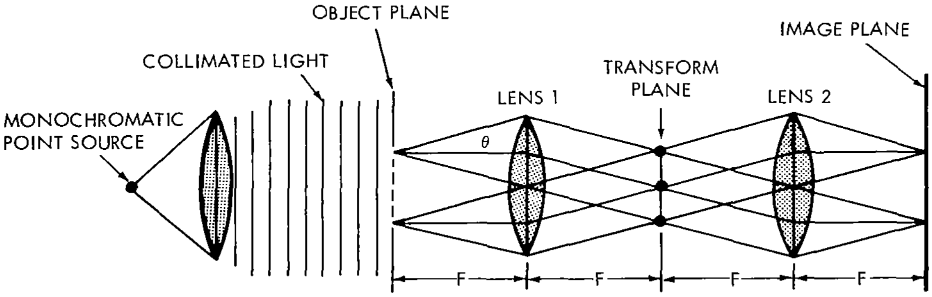
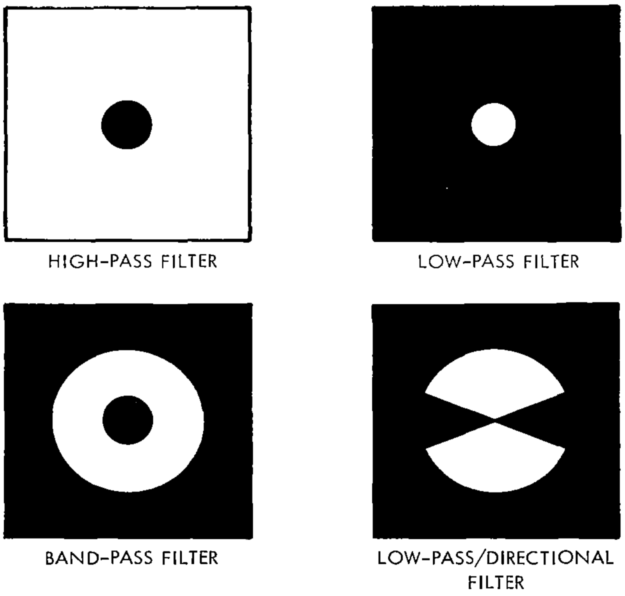


Figure 38—Optical filter.

giving the original transmission function. To be more specific, any object in the object plane of lens 1 will be imaged by the two lenses into the image plane of the system.

We have seen from the examples in Table 4 how sinusoidal transmission functions produce dots in the transform plane. For example, a 100-cps sine wave transmission function will produce a dc dot and two equally spaced dots (corresponding to 100 cps) centered about the dc term. If the light from such a corresponding pair of dots is blocked in the transform plane so that the light passing through them is not allowed to reach lens 2, the light in the image plane will not have variations corresponding to the frequency blocked. This is clarified if we consider Example III in Table 4. In this case we had a transmission function representing the sum of a 100-cps and a 150-cps sine wave. Two pairs of dots were produced in the spectrum. One pair corresponded to the 100-cps sine wave and the other to the 150-cps sine wave. If we block the light from the pair of dots corresponding to 100 cps in the transform plane, the light in the image plane will have only variations corresponding to the 150-cps sine wave. Since no light corresponding to the 100-cps wave is allowed to pass the transform plane, variations corresponding to this wave will not appear in the image plane. We have thus effectively filtered 100 cps from an input signal.

Extension of this basic idea can be used to make optical filters to pass or reject any desired range of spatial frequencies. Not only can a range of spatial frequencies be passed or eliminated, but specific inclinations can also be filtered. A few typical examples of optical filters are shown in Figure 39 and discussed below.


HIGH-PASS FILTER

LOW-PASS FILTER

BAND-PASS FILTER

LOW-PASS/DIRECTIONAL FILTER

Figure 39—Optical filters.

High-Pass Filter

Low frequencies can be removed by placing an obstruction in the transform plane which will block the area around the dc component (low frequencies) while a transparent area permits all light corresponding to frequencies greater than a specified value to be passed to lens 2.

Low-Pass Filter

A low-pass filter can be formed by putting a standard photographic iris into the transform plane. This iris will permit all light corresponding to low frequencies to be passed to lens 2 while blocking the light corresponding to high frequencies.

Band-Pass Filter

A band-pass filter can easily be constructed by blocking the light corresponding to the low frequencies, and permitting light corresponding to a specific range of frequencies to be passed.

Any frequencies above the desired passband can also be eliminated by blocking the light corresponding to these higher frequencies. It can be seen that effectively we can produce a band-pass filter by combining a high-pass filter with a low-pass filter.

Directional Filter

Additional flexibility in optical filtering can be achieved by selecting specifically oriented transmission functions for filtering. By blocking the light corresponding to the orientations, we can in effect select specific orientations for filtering. High- and low-pass filters can be combined with directional filters to allow only light from transmission functions of specific frequency and orientation to pass to lens 2.

Optical Signal Correlator

The mathematical operation of correlation is the comparison of two signals to determine the degree of correspondence between them. For two periodic functions of the same frequency, $f_1(t)$ and $f_2(t)$, the correlation function is expressed by the equation

$$\phi_{12}(\tau) = \frac{1}{T}\int_{-T/2}^{T/2} f_1(t)\, f_2(t+\tau)\, dt \ .$$

When $f_1(t)$ and $f_2(t)$ are not the same function, the correlation function $\phi_{12}(\tau)$ is called a *cross-correlation.* The second subscript always corresponds to the function that is displaced. The order of the subscripts is important, since $\phi_{12}(\tau)$ is not necessarily the same as $\phi_{21}(\tau)$.

For the special case where the two periodic functions are equal, the correlation function can be expressed by

$$\phi_{11}(\tau) = \frac{1}{T}\int_{-T/2}^{T/2} f_1(t)\, f_1(t+\tau)\, dt \ .$$

In this case the correlation function $\phi_{11}(\tau)$ is called an *autocorrelation.* The subscripts 1 1 indicate that an autocorrelation is obtained by comparing a function with itself.

The correlation expressed mathematically as above can be performed optically. An optical signal correlator, as shown in Figure 40, is basically an extension of the principles discussed in the preceding subsection "Optical Filtering." It is quite evident that lenses 1, 2, and 3 in Figure 40 form the basic optical filter of Figure 38. Lens 1 collimates the light from a point source (i.e. lens 1 changes the spherical wavefronts of a point source to plane wavefronts). To produce collimated light, the point source must be at the front focal point of lens 1. The plane wavefronts incident upon the object plane produce a diffraction pattern which is focused by lens 2. When the object plane is

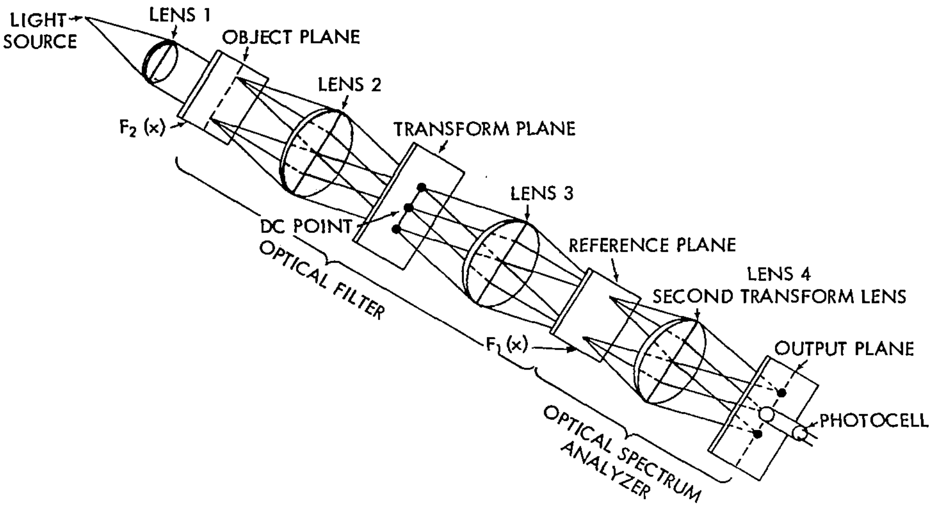


Figure 40—Optical correlator.

in the front focal plane of lens 2, a Fourier transform of the object transmission function is produced in the back focal plane (transform plane). Lens 3 is positioned so that the transform plane is in its front focal plane. The Fourier transform of the Fourier transform in the transform plane is produced in the back focal plane (reference plane) of Lens 3. Effectively this is an inverse Fourier transform which reproduces the original object transmission function. Therefore, lenses 2 and 3 form a projection system which images the transmission function of the object plane onto the reference plane. Up to this point, the operation of the optical correlator is identical to that of an optical filter. As will be shown later, this optical filtering characteristic inherent in an optical correlator can be used to advantage.

Rather than photograph the image in the reference plane, as would be done in an optical filter, the image is used as the incident light on a second transmission function inserted at the reference plane. The amplitude of the light transmitted by the reference plane will be the product of the incident-light amplitude and the transmission function of the reference plane. At this point a mathematical derivation will be useful to demonstrate the properties of the light amplitude transmitted by the reference plane. Let the constant amplitude of the plane waves incident upon the object plane be given by A_0. The light transmitted by the object plane is then the product $A_0 F_1(x)$, where $F_1(x)$ is the transmission function of the object plane. It is this transmitted light at the object plane that is imaged at the reference plane. Therefore the amplitude A of the incident light at the reference plane is given by

$$A = A_0 F_1(x) .$$

Now, as in the case of the object plane, the light transmitted by the reference plane will be the product of the incident-light amplitude $A_0 F_1(x)$ and the transmission function $F_2(x)$ of the reference plane. The expression for the transmitted light from the reference plane then becomes $[A_0 F_1(x) F_2(x)]$. From this last expression it is apparent that the light present beyond the reference plane is the same as that which would be present if the original light (A_0) were incident upon a transmission function given by the product of $F_1(x)F_2(x)$. Now, returning to the correlator shown in Figure 40: lens 4 is located a focal length from the reference plane and therefore produces a Fourier transform of the light transmitted by the reference plane. As was explained in the section on Fourier transforms, the amplitude in the output plane (back focal plane of lens 4) is then given by the Fourier transform of the product $F_1(x)F_2(x)$, or

$$\text{Fourier transform}\left[A_0 F_1(x)F_2(x)\right] = A_0 \int_{x_1}^{x_2} F_1(x)F_2(x)e^{j\omega x}\,dx\ .$$

Provisions can be made to displace the transmission function $F_1(x)$ in the object plane. The displaced function can be expressed as $F_1(x+\tau)$, where τ is the displacement. The amplitude in the output plane is then given by

$$\text{Fourier transform}\left[A_0 F_2(x)F_1(x+\tau)\right] = A_0 \int_{x_1}^{x_2} F_2(x)F_1(x+\tau)e^{j\omega x}\,dx\ .$$

The equation is similar in form to the correlation function

$$\phi_{2\,1}(\tau) = \int_{x_1}^{x_2} F_2(x)F_1(x+\tau)\,dx\ .$$

The constant in front of the integrals can be disregarded, since any difference would only involve a change in scale. The exponential in the Fourier transform equation can be set equal to 1 by setting ω equal to zero. Dropping the constant factor and equating ω to zero gives

$$\text{Fourier transform}\left[F_2(x)F_1(x+\tau)\right]_{\omega=0} = \phi_{2\,1}(\tau) = \int_{x_1}^{x_2} F_2(x)F_1(x+\tau)\,dx\ .$$

Since setting $\omega = 0$ is equivalent to eliminating all output terms except the zero-order term, the above relation amounts to the statement that the zero-order spectral term in the output plane of the optical correlator is the correlation function of the transmission functions in the object and reference planes.

For any particular displacement τ, the zero-order output term gives the value of the correlation function corresponding to the specific value of τ. If the signal film in the object plane of the correlator is continuously displaced, the zero-order term varies as the correlation function. The zero-order term can be monitored by a photocell as shown in Figure 40. Since the amplitude of the zero-order term is the correlation function, the intensity-sensitive photocell output will be the square of the correlation function. It is also possible to record the correlation function on film by replacing the photocell with a recording film. By moving the signal film and recording film synchronously it is possible to record a continuous picture of the correlation function.

To perhaps clarify the correlation operation, we can consider the analogy of sliding a negative over a duplicate negative. This analogy would correspond to the case where $F_1(x) = F_2(x)$ (autocorrelation) in the discussion above. As the negative slides over its duplicate, the amount of light transmitted will vary. The maximum amount of light will be transmitted when the images of the two negatives coincide. This point corresponds to maximum correlation. A similar situation occurs in the optical correlator.

The operation of an optical correlator can be summarized as the cascade operation of an optical filter and an optical spectrum analyzer with a second transmission function inserted in the plane common to the two sections. In other words, an optical correlator can be made by taking the output of an optical filter and using it as the input to an optical spectrum analyzer with a transmission function inserted in the optical filter output plane (spectrum analyzer input plane). The light amplitude at the zero-order spectral point in the correlator output plane (spectrum analyzer output) is a measure of the correlation function between the transmission functions present in the object and reference planes.

In practice the desired correlation is not the correlation of the transmission functions but the correlation of the signals represented by the transmission functions. The difference between these two cases is shown in Figure 41. It is noted that the correlation of the transmission functions includes bias terms in addition to the desired correlation function. For a constant τ the additional terms are, in effect, a bias which determines the dc level of the output-light amplitude. Examination of these undesired terms indicates that a dc stop in the transform plane of the optical filter section will eliminate all but the desired correlation function. Assume $F_2(x)$ is the transmission function inserted in the subsection "Optical Filtering." The light amplitude in the transform (filter) plane will be the transform of $F_2(x)$, or

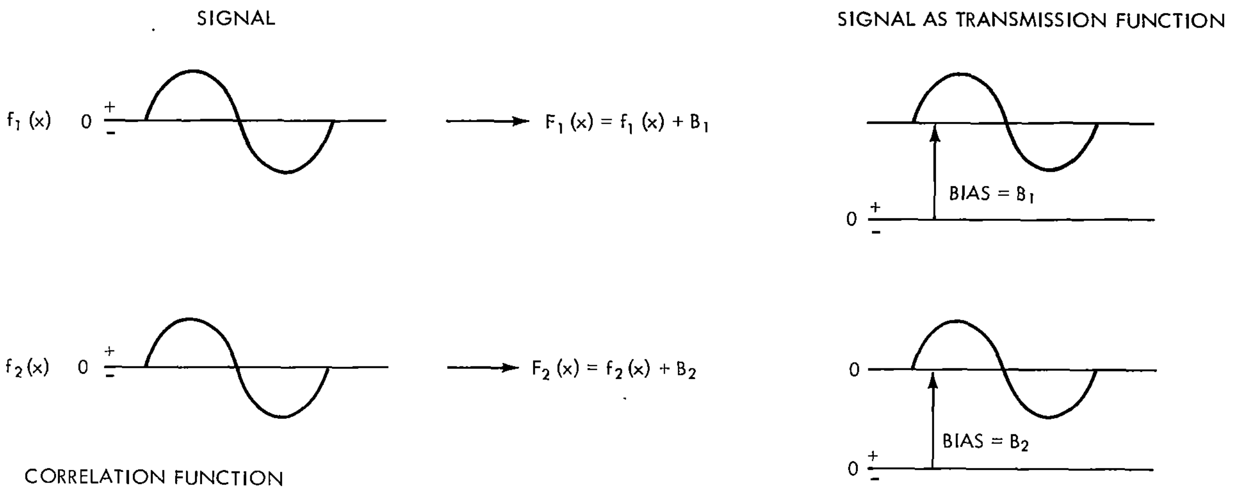
$$\text{Fourier transform}\left[F_2(x)\right] = \int B_2 e^{j(ky/f)x}\,dx + \int f_2(x)\, e^{j(ky/f)x}\,dx \;.$$

The first integral (of $B_2 e^{j(ky/f)x} dx$) is of the form $\sin y/y$, which has its maximum value at $y = 0$ and corresponds to the dc component of $F_2(x)$. The second integral can be written as the sum of two integrals (value at $y = 0$, and value at $y \neq 0$):

$$\int f_2(x)\, e^{j[ky/f]x}\,dx = \left[\int f_2(x)\, e^{j(ky/f)x}\,dx\right]_{y=0} + \left[\int f_2(x)\, e^{j(ky/f)x}\,dx\right]_{y\neq 0} \;,$$

or

$$\int f_2(x)\, e^{j[ky/f]x}\,dx = \left[\int f_2(x)\,dx\right]_{y=0} + \left[\int f_2(x)\, e^{j(ky/f)x}\,dx\right]_{y\neq 0} \;.$$



CORRELATION FUNCTION

$$\phi_{21} = \int_{x_1}^{x_2} f_2(x)\, f_1(x+\tau)\, dx$$

CORRELATION FUNCTION

$$\phi_{21} = \int_{x_1}^{x_2} F_2(x)\, F_1(x+\tau)\, dx = \int_{x_1}^{x_2} \left[f_2(x) + B_2\right] \left[f_1(x+\tau) + B_1\right] dx$$

$$= \int_{x_1}^{x_2} f_2(x)\, f_1(x+\tau)\, dx + B_1 \int_{x_1}^{x_2} \left[f_2(x) + B_2\right] dx + B_2 \int_{x_1}^{x_2} f_1(x+\tau)\, dx$$

$$\phi_{21} = \phi_{21} + B_1 \int_{x_1}^{x_2} f_2(x)\, dx + B_2 \int_{x_1}^{x_2} f_1(x+\tau)\, dx + B_1 B_2 (x_2 + x_1)$$

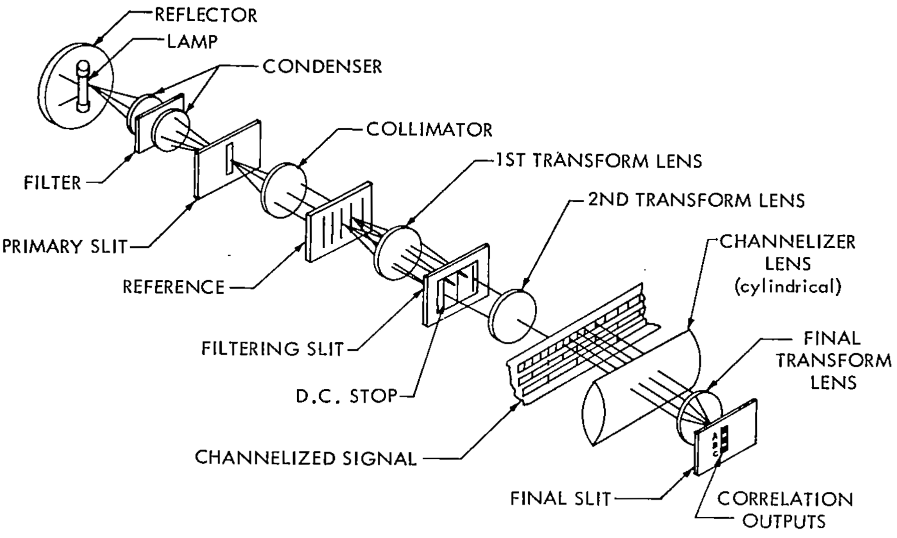
Figure 41—Correlation functions—signal vs transmission function.

From this discussion it is apparent that the factors B_2 and $\int f_2(x)\,dx$ correspond to light which must pass through the zero-order (or dc) point ($y = 0$) in the transform (filter) plane. Inserting a dc stop at this point effectively sets $B_2 = 0$ and $\int f_2(x)\,dx = 0$ beyond the transform (filter) plane. Thus the three undesired terms will be zero, and only the desired correlation function will appear at the correlator output. Similarly, if $F_1(x)$ is inserted in the optical filter, B_1 and $\int f_1(x+\tau)\,dx$ will be set equal to zero beyond the transform (filter) plane. Therefore, by using a dc stop the correlator output will be made equal to the correlation function of the two signals inserted as transmission functions regardless of which signal appears first in the correlator.

A big advantage of optical correlators over the conventional electronic correlator is the ability to handle many input signals simultaneously. A channelized optical correlator is shown in Figure 42. Comparison of the optical correlators of Figure 40 and Figure 42 indicates that the basic difference is the addition of a cylindrical lens for channelized operation.

The details of the channelized section of the optical correlator are shown in Figure 43. We have previously mentioned that a two-dimensional optical process can be treated as two separate one-dimension processes. The channelized section can therefore be represented by the two processes shown as the top and side views in Figure 43. The top view shows that the cylindrical lens

Figure 42—Channelized optical correlator.



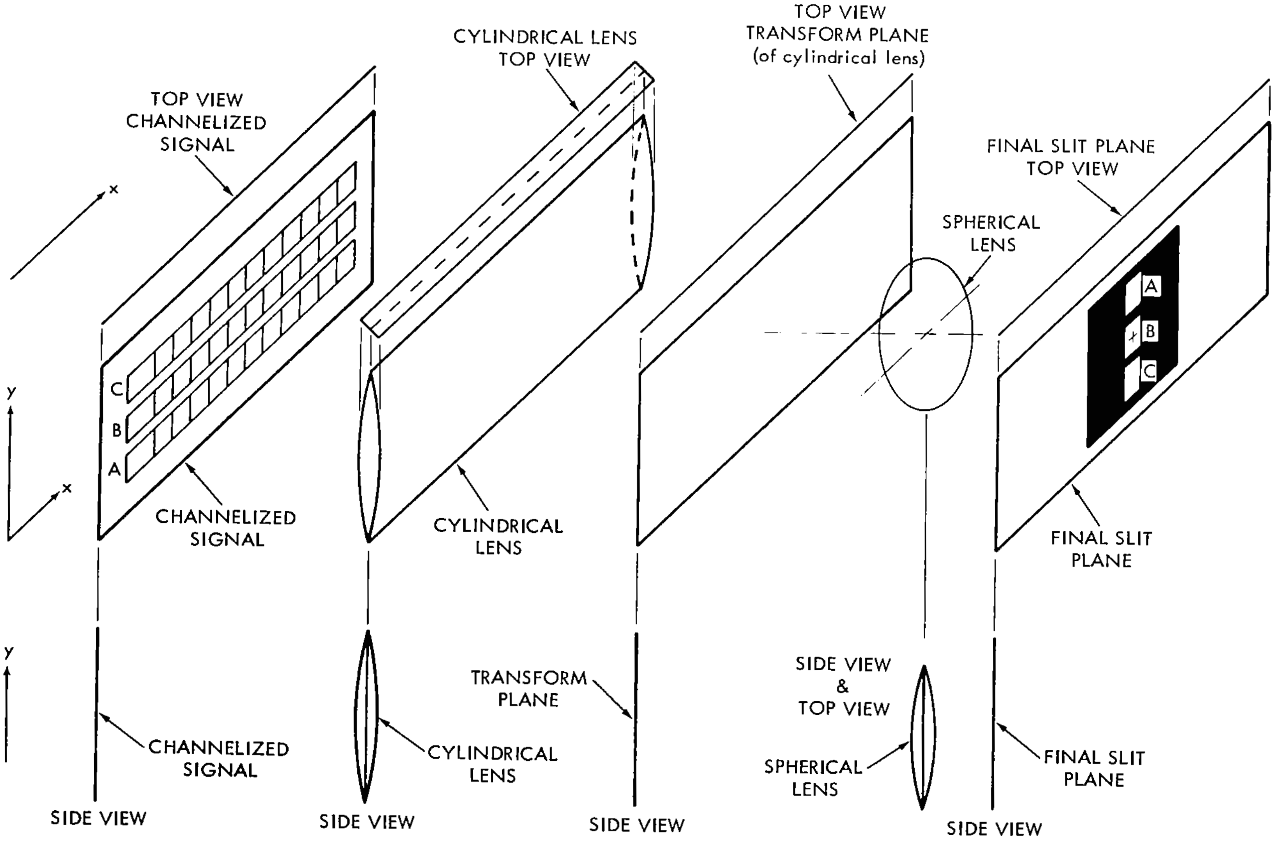


Figure 43—Channelized section of optical correlator.

has no focusing effect (no curvature) for the x coordinate. The top view, therefore, has the configuration of a spectrum analyzer with the spherical lens producing a Fourier transform in the final slit plane. This Fourier transform will correspond to information signal variations along the x direction (i.e. vertical slits). Examination of the side view shows that both the cylindrical and spherical lenses have a focusing action for the y coordinates. The side view has the configuration of an optical filter system which images the information in the y direction (i.e. horizontal channels) into the final slit plane. The back focal plane of the cylindrical lens (transform plane in Figure 43) corresponds to the filter plane of the optical filter configuration discussed in a previous subsection. In effect, the cylindrical lens produces a Fourier transform in the y direction and the spherical lens produces a Fourier transform of the transform (i.e. an inverse transform) which results in an image of the original signal in the y direction.

It has been shown that the Fourier transform of the information (i.e. vertical lines) in each channel is produced, and each channel is imaged, in the final slit plane. The imaging of the horizontal channels assures that the Fourier transform of each channel is obtained without intermixing information between channels.

As previously shown in the mathematical derivation of the correlation process, the correlation function is given by the light amplitude of the zero-order term of the Fourier transform at the final slit. The final slit performs the function of blocking all terms except the zero-order term. As shown in Figures 42 and 43, a separate zero-order term (A, B, C) is obtained for each channel. The amplitude of each zero-order term is the correlation function value for the respective channel.

The channelized correlator shown in Figure 42 and discussed above involves the correlation of many signals with a single reference. This amounts to comparing the channel signals with the reference signal to detect matching signals. In some applications it may be desirable to match one signal with several reference signals. This operation can be realized simply by channelizing the reference film instead of the signal film. It must be kept in mind that the addition of a cylindrical lens is required wherever the channelized light signals are to be processed separately.

It has been shown above that either the reference film or the signal film can be placed in the optical filter object plane. In some applications, prior knowledge of undesired component frequencies (noise) may be available. By placing the signal film in the optical filter object plane, noise components can be eliminated (along with the dc term) by using an appropriate optical filter. This application of optical filtering in the correlator system provides a means for improving the signal-to-noise ratio of the signal inputs.

In either case considered above, the amplitudes of the zero-order terms in the final slit represent the correlation functions. These correlation functions can be monitored by separate photocells or recorded on photographic film. The photographic record would consist of intensity-modulated bands, where each band corresponds to a channel. Whatever means is used to measure or record the correlation function, it must be kept in mind that the amplitude of the light in the zero-order term corresponds to the correlation function. Since the detection must be accomplished by an intensity-sensitive device, the square root of the detected value must be taken if the value of the correlation function (amplitude) is desired.

At this point some of the practical techniques used in implementing an optical correlator will be discussed. Comparing Figures 40 and 42, it is seen that a point source is implemented by an arrangement which includes a lamp and reflector, condenser and filter, and a slit. A suitable basic light source for an optical correlator is a 100-watt high-pressure mercury-arc lamp. A spherical reflector is used behind the lamp to increase the amount of light radiating in the direction of the optical-correlator axis. The light from the lamp is passed through a condenser and filter assembly. The optical filter (an interference-type filter) allows only monochromatic light (5,461 angstroms in this case) to be transmitted through it. A condenser assembly of high-quality lenses is used to image the light from the mercury-arc lamp into a small slit (primary slit). The fact that a mercury-arc source is small to begin with—and that its image is to be even smaller—requires the use of high-quality condenser lenses. The narrow primary slit produces circular wavefronts in the horizontal direction. In effect the light waves in a horizontal plane appear to be those radiated by a point source. The first collimator lens (located a focal length from the primary slit) changes the horizontal circular wavefronts into straight-line wavefronts (in a horizontal plane). Thus the light incident upon the object plane (reference plane in Figure 42) has uniform amplitude along a line perpendicular to the effective (vertical) slits inserted in this plane. The uniform horizontal illumination (horizontal-plane waves) of the film plane (reference plane) meets the requirements for the Fourier transform of the vertical lines to be produced. It is to be noted that the long dimension of the slit in the vertical direction does not act as a point source, and light along a vertical direction cannot be used for optical data processing. The lack of a point source characteristic in the vertical direction blurs the spectral points into vertical lines. This effect is shown in the filtering slit of Figure 42, where the vertical lines represent spectral lines.

An adjustable slit (the filtering slit of Figure 42) located in the transform plane can be used to block high-order spectral components. This adjustable filtering slit is effectively an adjustable low-pass filter. A wire is mounted vertically in this slit to block the zero-order spectral line. Thus the adjustable slit and wire eliminate the zero-order and all high-order spectral components above a selected value.

The final slit shown in Figure 42 blocks all output spectral orders except the zero order. This slit allows the selection of the zero-order terms, which by themselves represent the correlation function.

The resolution limits of optical correlators are dependent upon the lenses and films used, as discussed in an earlier subsection. At this point it is important to note that the resolution between spectral points is dependent also upon the width of the primary slit. Because of light spreading, a narrow slit results in better resolution than a wider slit.

Convolution

The convolution integral is given by

$$\rho_{12}(\tau) = \int_{-\infty}^{\infty} f_1(t) f_2(\tau - t) dt .$$

This integral resembles a correlation integral in that they both involve a displacement, multiplication, and integration. However, there is a difference in the expression of the displaced function. In correlation, the displaced function is expressed as $f_2(t+\tau)$, whereas in convolution it is expressed as $f_2(\tau - t)$. The difference is a folding or reflection in time. In other words, $f_2(\tau - t)$ is obtained by simply reflecting $f(t+\tau)$ about $t = 0$.

Figure 44 shows a graphical comparison between convolution and correlation. In this example it is assumed that a function $f_1(t)$ is to be convolved and correlated with the function $f_2(t)$. It is to be noted that in both convolution and correlation the function $f_2(t)$ is displaced to the left by τ, producing $f_2(t+\tau)$ as shown in the displacement diagrams. The next step in the convolution process is the folding or reflection (about $t = 0$) of $f_2(t+\tau)$ to produce $f_2(\tau - t)$. There is no corresponding step in the correlation process. This fact accounts for the difference between the two operations. The function $f_1(t)$ is then multiplied by $f_2(\tau - t)$ for convolution and by $f_2(t+\tau)$ for

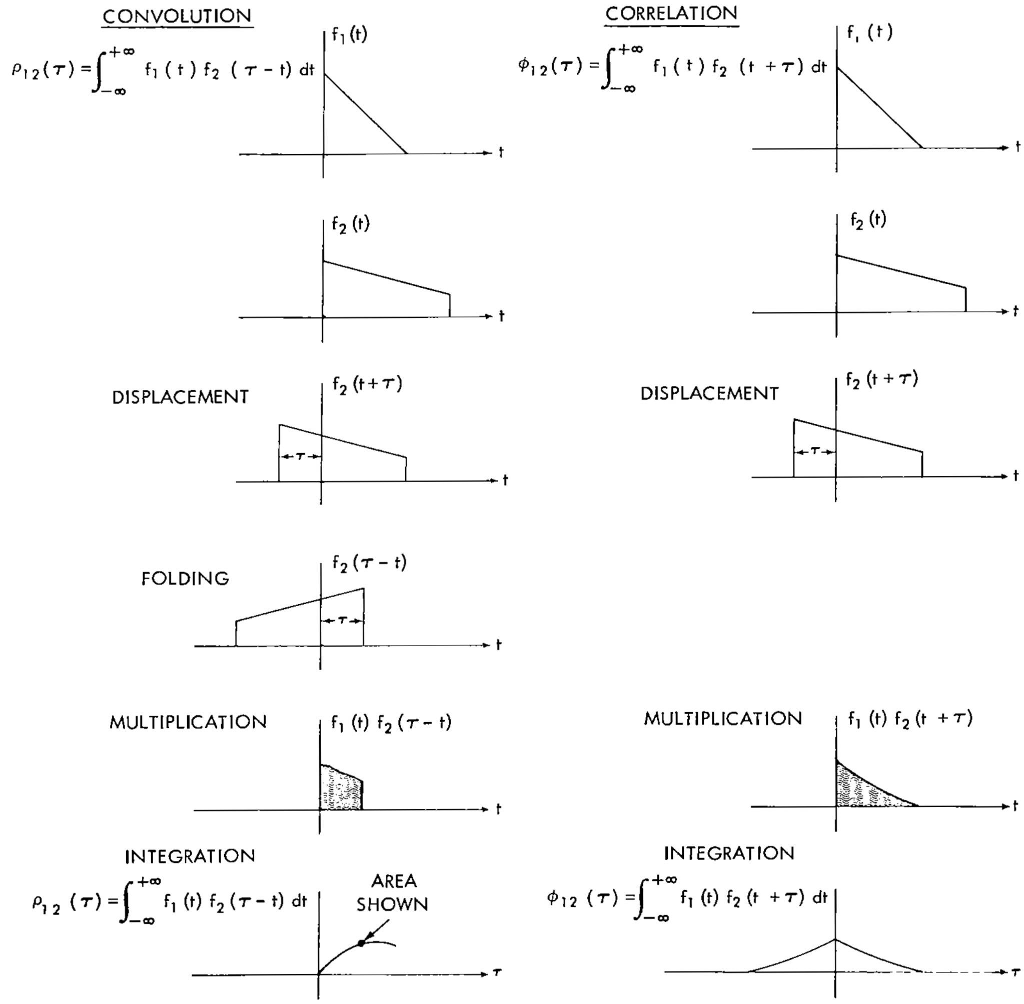

CONVOLUTION
$\rho_{12}(\tau) = \int_{-\infty}^{+\infty} f_1(t) f_2(\tau - t)\, dt$
$f_1(t)$
t
$f_2(t)$
t
DISPLACEMENT
$f_2(t+\tau)$
τ
t
FOLDING
$f_2(\tau - t)$
τ
t
MULTIPLICATION
$f_1(t) f_2(\tau - t)$
t
INTEGRATION
$\rho_{12}(\tau) = \int_{-\infty}^{+\infty} f_1(t) f_2(\tau - t)\, dt$
AREA SHOWN
τ
CORRELATION
$\phi_{12}(\tau) = \int_{-\infty}^{+\infty} f_1(t) f_2(t+\tau)\, dt$
$f_1(t)$
t
$f_2(t)$
t
DISPLACEMENT
$f_2(t+\tau)$
τ
t
MULTIPLICATION
$f_1(t) f_2(t+\tau)$
t
INTEGRATION
$\phi_{12}(\tau) = \int_{-\infty}^{+\infty} f_1(t) f_2(t+\tau)\, dt$
τ

Figure 44—Comparison between convolution and correlation.

correlation. The areas under the respective multiplication curves show the respective value of the convolution and correlation function for a given τ. In general, the area under this curve is given by an integral of the products from $-\infty$ to $+\infty$. The result of this integration is a function of τ. As shown in the curves labeled "Integration," a plot of the integral (area under multiplication curve) for all values of τ is a plot of the convolution and correlation functions.

An additional difference resulting from the above is that the convolution of $f_1(t)$ and $f_2(t)$ is the same regardless of which function is folded, provided the folded function is also the one displaced. The correlation function, in general, is different depending upon which function is displaced.

Optically we can see that the convolution function can be obtained in a way similar to that in which the correlation function was determined. The additional operation of folding required in convolution involves a 180° rotation of the film. Thus, if a 180° film rotation is used to represent the folding operation, the convolution function can be obtained using the same basic operation and configuration as for cross-correlation.

Convolution is an extremely important operation in the analysis of linear systems. The output of a linear system can be determined by convolving the given input with the impulse response of the system. For example, the output pattern of a two-dimensional optical spectrum analyzer can be predicted in advance by using convolution principles. The output of the spectrum analyzer is proportional to the convolution of the Fourier transform of the signal written on film with the image of the input light source. Normally when the input light is from a point source, the spectral image will be a set of points corresponding to the Fourier transform of the signal. If the input light source is diamond-shaped, the output function will be the convolution of the diamond shape with the Fourier transform of the signal film. In this case, the spectral image will contain diamond-shaped points instead of dots as in the point source case.

Pattern Recognition

Pattern recognition can be accomplished optically using several techniques previously discussed. For example, the Fourier transform of an input signal (representing a pattern) can be imaged onto a desired frequency pattern to determine whether the input signal spectra correspond to the desired spectra.

Cross-correlation techniques can be used to determine also whether a given input pattern corresponds to a desired pattern. Optical filtering can be performed during the cross-correlation process. For example, if a desired pattern consists of vertical lines, all other lines appearing in an input pattern can be filtered out and cross-correlation achieved for only the vertical lines present in the input pattern. Application of this principle allows for more flexible operation in optical pattern-recognition systems.

The optical techniques previously discussed not only lend themselves to pattern recognition, but will also be just as useful for pattern elimination. There are certain applications where the elimination of specific patterns in an image is desirable. For example, an undesirable variation in the density of a photograph due to grain can be eliminated. Since the variation in density due to photographic detail is usually of a much lower frequency than the variations due to grain, an appropriate low-pass filter will eliminate the (relatively) high-frequency grain components. Elimination of the components corresponding to grain can improve the quality of the photographic image. A similar improvement through filtering can be obtained by elimination of dot patterns in halftone photographs.

Another example of pattern elimination would be a televised picture containing 512 horizontal scan lines. It might be desirable to eliminate all horizontal scan lines corresponding to the frequency 512. It is possible to put an optical stop into the transform plane corresponding to the horizontal lines of the frequency of 512 Hz. It has been shown that the Fourier transform of the Fourier transform is an image of the original picture. Stopping the light in the Fourier transform plane corresponding to 512 Hz will produce an image of the original picture without the horizontal lines that correspond to a frequency of 512 Hz. At this point it might be noted that a television picture cannot be used directly as an input to an optical data processor. The main reason for this is that noncoherent light of low intensity, such as the light output from a television picture tube, cannot be used as an input to an optical data processor. It is possible, however, to photograph the image on the television picture tube and use the developed photographic image as an input signal to an optical data processor.

Optical processing techniques could have been used in the retrieval of the Mariner spacecraft signals representing pictures of Mars. The signals transmitted back to earth were processed by a digital computer to produce the final pictures. These pictures were found somewhat less than satisfactory. In order to emphasize certain areas of the pictures, the contrast ratios or scaling of the received bits were changed. Optical techniques could have been used to identify particular patterns. For example, if a particular frequency of lines at a particular inclination in the picture were to be accentuated, the corresponding points in the Fourier transform plane could be made more transparent than all others. This would result in accentuation of the selected lines in the final image. At the same time the horizontal scan lines could be eliminated by blocking the light in the transform plane corresponding to the scan line frequency.

The flexibility of optical data processing thus lends itself particularly to applications involving pattern recognition. It appears that, when these techniques have been fully developed, optical data processing will be an extremely powerful tool when applied to pattern recognition.

Analog Computation (Antenna Simulation)

An optical analog computer, like its electronic counterpart, can serve as a scale model to solve specific problems. One such application is in the simulation of antenna radiation patterns. In an optical system the lens acts on coherent light waves in a way similar to that in which an antenna acts on microwaves. Figure 45 shows the comparison between the optical and microwave systems. In the optical system, the wavefronts radiating out from a monochromatic light source are transmitted as parallel wavefronts (parallel rays) by the action of the transmitting lens. The

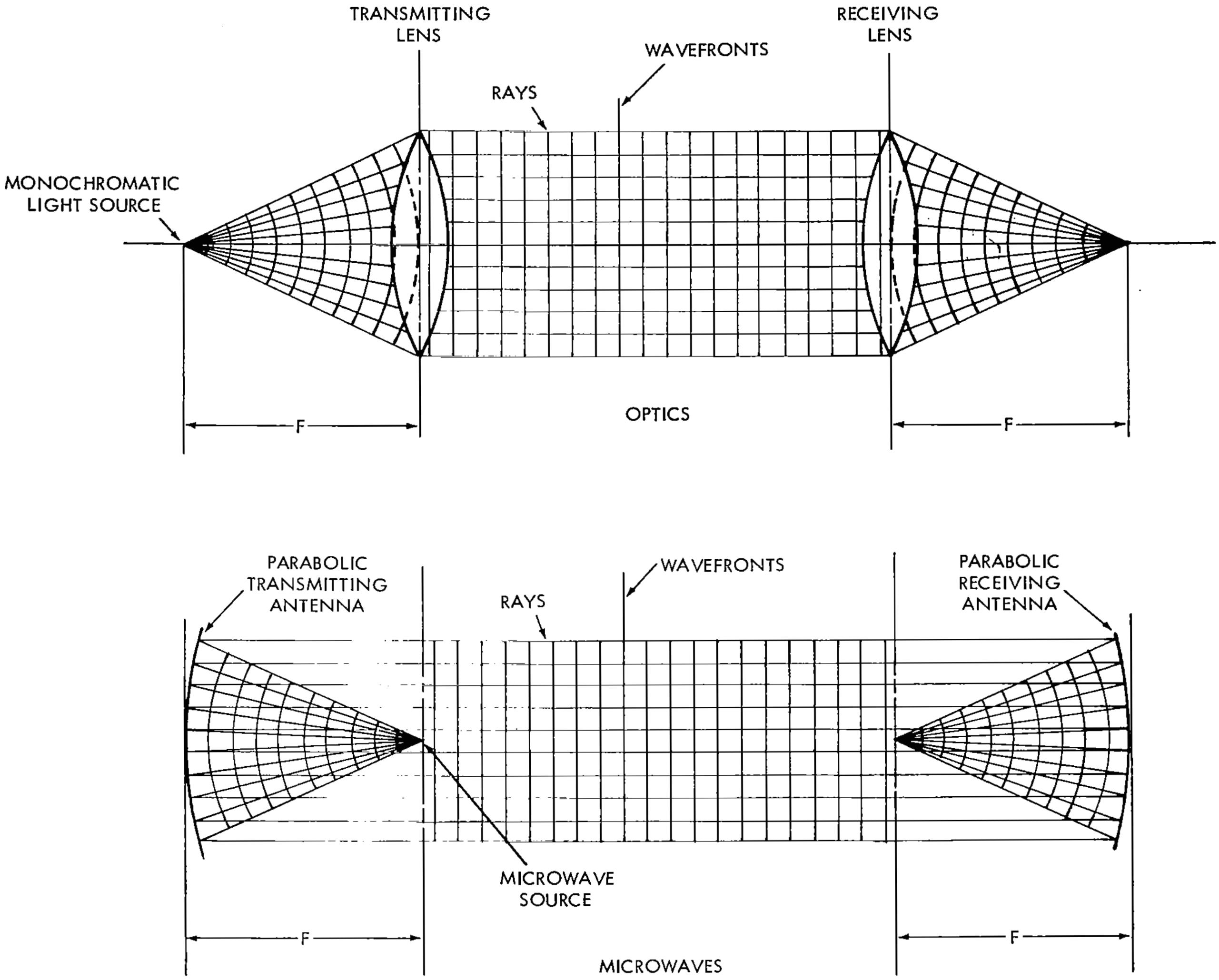


Figure 45—Comparison between optical and microwave transmitting and receiving systems.

receiving lens focuses these parallel wavefronts to a point. A similar action occurs in the microwave system. Microwaves radiating out from a microwave point source are made into parallel wavefronts by the action of the transmitting antenna. The receiving antenna effectively focuses these parallel wavefronts to a point.

The analog between an optical system and a microwave system can also be seen in the mathematics of the two systems. It has already been shown that the Fourier transform of a transmission function in the front focal plane of a lens is produced in the back focal plane of a lens. Now the approximate radiation pattern of an antenna is given by the Fourier transform of the aperture electric field distribution (source function). Thus it is apparent that an optical system with a transmission function representing the antenna electric field distribution function can be used to simulate the antenna system.

With the proper light source and lens system a microwave antenna or antenna system can be simulated, and the field intensity of the antenna system can be determined from the optical model. Visual examination in a laboratory of the radiation pattern with side lobe structure is then possible. Assume light of wavelength 6×10^{-5} meters is used to simulate a microwave wavelength of 6 meters. One inch in the optical model will then represent 10^5 inches (approximately 8,333 feet) in the actual system. By the use of this optical scale model it becomes an easy matter to determine the parameters of the actual system and to change conditions. This is the simplest and most direct method to study and observe antenna systems in the laboratory.

In the optical simulation of a microwave antenna, the light-amplitude distribution of the simulator corresponds to the microwave electric field intensity distribution. As shown before, the light-amplitude distribution is given by the Fourier transform of the input light *amplitude*. Intensity is given by the square of the light amplitude without phase (i.e. the square of the absolute value of the complex light amplitude). The intensity distribution in the optical simulator is therefore given by the square of the absolute value of the Fourier transform of the input light amplitude (i.e. $I = |F(A)|^2$).

The near and far field patterns of a microwave antenna system can be determined from this optical scale model. A plane wavefront of light is used to illuminate an optical transparency which represents the distribution function of the antenna system being simulated. A lens placed a focal length away from the transparency will produce the Fourier transform of the transparency in its back focal plane. The far field pattern will be found in this back focal plane. By placing photographic film in this plane, it is possible to obtain a picture of the far field intensity distribution. Pictures of the entire antenna field pattern can be obtained by simply repositioning the recording film to the particular portion of the field pattern which is of interest.

In previous discussions we have considered the photographic input to optical data processors only as a recording of amplitude variations in light. It is possible to produce a photographic film which records phase variations as well as amplitude variations in light. Such films are more difficult to produce than the usual type, which records only the amplitude variations. The introduction of phase variations can, for example, be accomplished by varying the thickness of a negative. It was shown earlier how the usual photographic negative records the amplitude of light as a variable optical density. The thickness of the negative considered was uniform. Phase variations would be introduced if the thickness of the negative were allowed to vary. Another method for recording amplitude and phase on photographic film will be discussed later in the section on holography. Although techniques for producing variation of phase are difficult to achieve, they are possible. It is therefore possible to represent both amplitude and phase characteristics of antennas photographically and to study phase delays through optical simulation. In addition, the effect of the atmosphere on the antenna radiation pattern can be determined optically. The introduction of transparencies which introduce phase and amplitude variations corresponding to atmospheric effects would distort the antenna pattern just as the actual pattern is distorted in the atmosphere.

Pulse Expansion and Compression

In addition to data processing systems, special purpose systems can be realized using optical techniques. As an example, we can consider the optical pulse-expansion and -compression techniques which have been developed for use in radar systems. In radar systems, high-energy pulses are required for long range and narrow pulse widths are required for high resolution. Short-duration pulses (high-resolution) are limited in energy by practical considerations. The maximum energy (proportional to the square of the amplitude) of a pulse is limited by the ratings of the radar components. The energy of a pulse can be increased by increasing the pulse width; however, increasing the pulse width reduces the resolution capabilities of the radar. In conventional radar systems, a compromise is generally made between maximum range (energy) and resolution (narrow pulse width). It has been possible through optical techniques to transmit long-duration pulses (high-energy) and retain high resolution. This technique, which involves the expansion and compression of pulses, has been used in radar systems to obtain greater range and higher resolution for a given power capability.

The technique of pulse expansion and compression will be explained using Figure 46. Diagram A shows the pulse required to achieve a given range and resolution. The pulse width τ is determined by the resolution requirement, and the power P is determined by the range (energy) requirement. Now let us assume that the component ratings of the radar system limit the maximum pulse power to 1/5 (one-fifth) the desired power P. In order to obtain the energy necessary to meet the range requirement, it is necessary to increase the pulse width by a factor of 5, as shown in diagram B. Since the energy of a pulse is determined by the area under the curve representing the pulse, it is apparent that the pulses shown in diagram A and B have equal energy (therefore equal range). However, the pulse shown in diagram B no longer meets the resolution requirement, since the pulse duration has been increased (resolution decreased). This is the problem encountered in conventional radar systems.

Now we will consider the technique of expanding and compressing a pulse by optical methods, and later show that this method of expansion retains the desired resolution. Consider diagram C, which shows the spectrum for the pulse of width τ shown in diagram A. If the sinusoidal components of the spectrum curve (diagram C) were added together (with proper phase), the resultant would be a pulse of duration τ and amplitude A, as shown in diagram E. This implies that the infinite number of sine waves cancel one another completely outside the time interval τ. It is apparent that the same result would be obtained if each sine wave component were only of duration τ and superimposed (added) in the time interval τ as shown in diagram E. If we assume that these component waves are transmitted sequentially, as shown in diagram F, the pulse width would effectively be multiplied by a factor equal to the number of components used. In the example of Figure 46, only five components are used (diagram D), giving a pulse width of 5τ, as shown in diagram F. If the sequential signal of diagram F were divided into the five component waves and these waves were superimposed in a time interval τ, an approximation of the pulse of diagram A would be reconstructed, as shown in diagram E. It is to be noted that, since the waveform of diagram F has an effective pulse width of 5τ, the peak power required could be correspondingly less, just as in the case shown in diagram B. As mentioned above, the superpositioning of the component waves will result in an approximation of the pulse shown in diagram E. Since this pulse has an infinite number of components, any finite number of these components will produce only an approximation of the pulse. A fairly good approximation will be obtained by proper choice of the

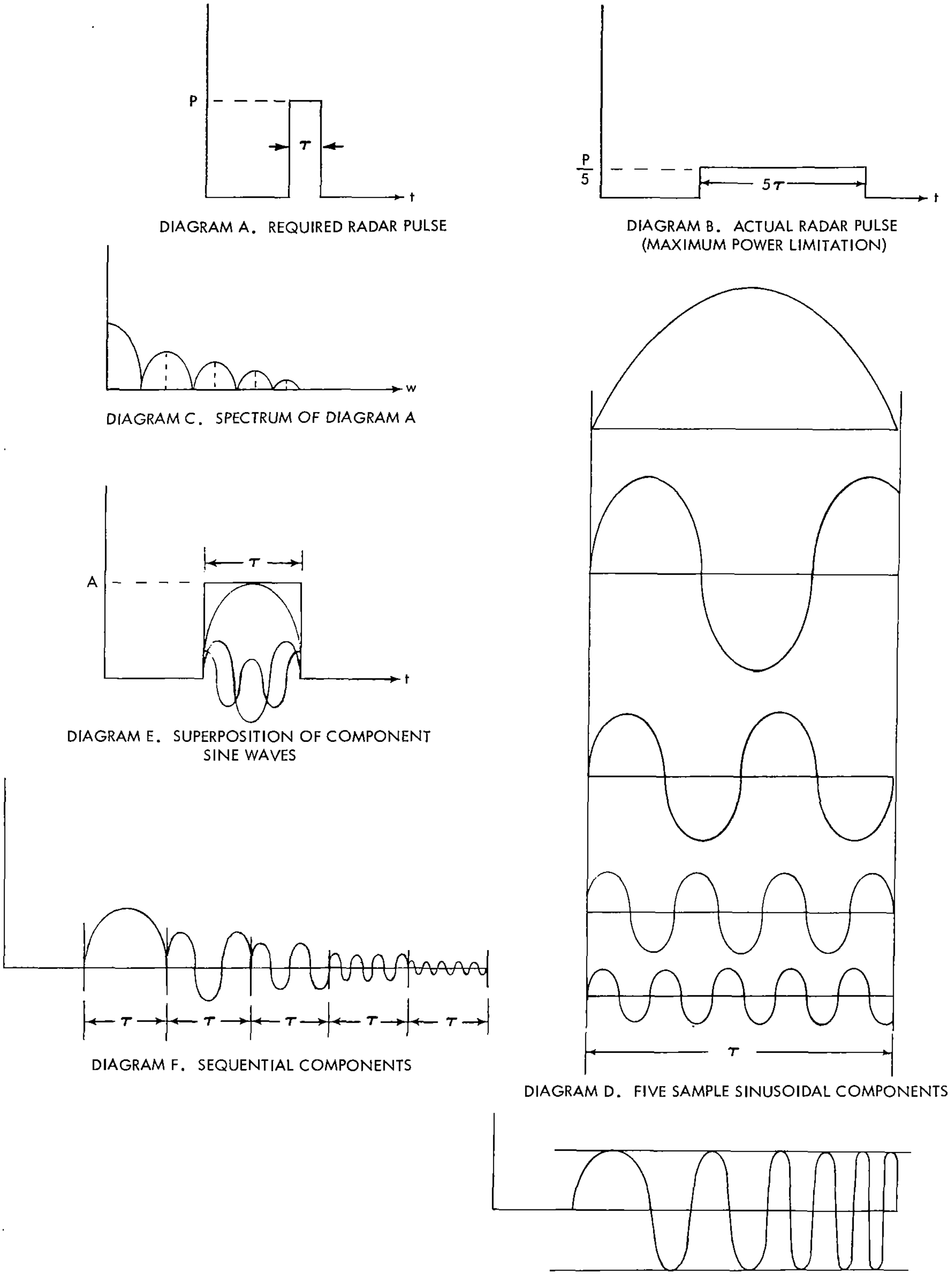


DIAGRAM A. REQUIRED RADAR PULSE

DIAGRAM B. ACTUAL RADAR PULSE (MAXIMUM POWER LIMITATION)

DIAGRAM C. SPECTRUM OF DIAGRAM A

DIAGRAM E. SUPERPOSITION OF COMPONENT SINE WAVES

DIAGRAM F. SEQUENTIAL COMPONENTS

DIAGRAM D. FIVE SAMPLE SINUSOIDAL COMPONENTS

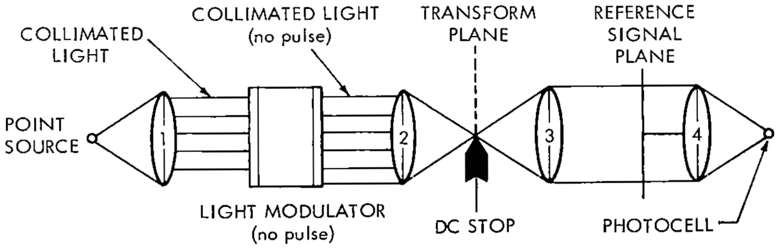
DIAGRAM G. EQUAL-AMPLITUDE COMPONENTS

Figure 46—Pulse expansion and compression.

component frequencies used in the sequential signal of diagram F. It is apparent that the approximation improves as the number of components is increased; however, this number is limited by the expansion-compression system. To take full advantage of the power capabilities, the sequential signal of diagram F is modified to produce equal-amplitude components as shown in diagram G.

The method for compressing the waveform of diagram F, Figure 46, to produce the pulse shown in diagram E, Figure 46, is shown in Figure 47. Figure 47 demonstrates the use of an ultrasonic light modulator in a pulse compression system. Any pulse introduced into the light modulator will diffract the light waves passing through in accordance with the variations in the index of refraction caused by compression. Lens 1 in Figure 47 collimates the light from a point source to provide plane wavefronts incident upon the light modulator. With no signal in the light modulator, the incident light is passed undiffracted, and lens 2 focuses the collimated light onto the dc stop in the transform plane. The dc stop allows light to reach lens 3 *only when* there is some signal in the light modulator. The combination of lenses 2 and 3 forms an image of the signal pattern (as represented by compressional waves in the modulator) onto the reference signal plane. Any light passing the reference signal plane is focused onto a photocell by lens 4. Consider a signal similar to that shown in diagram G of Figure 46, and assume the input aperture has a length equal to the length of the signal. As the signal passes through the light modulator, there is only one instant of time at which the complete signal will be present in the input aperture. Since the signal is known, a replica of it can be introduced in the reference signal plane. The photocell will detect maximum light only when the signal in the aperture of the light modulator corresponds to the reference signal. At the instant of time when the complete input signal (Figure 46, diagram G) appears in the aperture, a pulse indicating a correlation will appear at the photocell output. This output pulse is extremely short and in effect represents a pulse compression.

Figure 47—Pulse expansion and compression system.



The light modulator is also used to generate a signal similar to that shown in Figure 46, diagram G. In this case, the reference signal generates the expanded pulse. A narrow pulse is applied to the light modulator, and as it travels through the modulator its image moves along the reference signal film. Now the photocell will have an output whenever the imaged pulse falls on a transparent section of the reference film. Since the transparent sections of this film are spaced to correspond with the desired signal frequencies, as the imaged pulse moves across the reference signal film the desired signal frequencies will be produced by the photocell.

Examination of the radar pulses shown in Figure 48 will show that the resolution of the expanded pulse produced by the optical system is the same as the original desired pulse of diagram

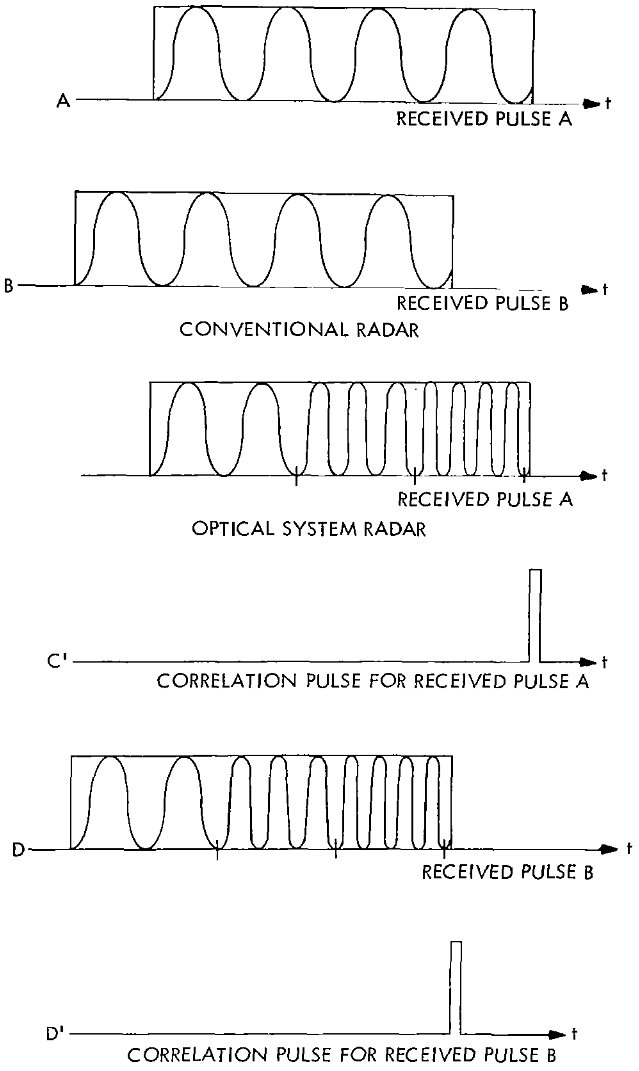


Figure 48—Comparison of radar systems.

A, Figure 46. First consider the conventional radar pulses A and B. These reflected pulses are displaced in time, since they are assumed to be reflected from two closely spaced objects. For the displacement shown, the conventional receiving system will be unable to detect the presence of two objects and will indicate the presence of a single target. Although the receiver is capable of detecting the presence of the received pulses, it is unable to detect whether two pulses overlap or only one pulse is being received. This limits the resolution of the system to target spacings which produce delays long enough so that the delayed pulse returns after the earlier pulse has been completely received. Now consider the expanded pulses C and D. These pulses are composed of different frequencies, as described for the pulse expansion process above. When the entire pulse is received and all the frequencies are correlated, a short correlation pulse is obtained. The pulse duration of these correlated pulses is extremely short, and therefore the resolution of the radar is high enough though the actual received pulse duration may be the same as for the conventional radar. In operation, when pulse C is in the light modulator aperture, an output pulse is produced. However, at this time D may already be partially in the aperture and a short time later, corresponding to the delay, another correlated output pulse (D') will be produced. Thus, although the effective pulse duration has been expanded, resolution does not decrease and may even be increased by using many frequency components in the optical pulse expansion-compression system.

PRINCIPLES OF HOLOGRAPHY

Introduction to Holograms

Certain emulsions, such as the Lippmann type, can be used for reproducing color pictures by recording the interference pattern produced by an incident and reflected wave. This process was discussed in the subsection "Photographic Films for Optical Processing," but its significance merits further discussion here. The argument used in this color process is shown in Figure 49. A very high-resolution emulsion is placed in contact with a reflecting surface such as mercury. Parallel incident-light waves will pass through the glass plate and emulsion to the mercury, from which they will be reflected back upon themselves. The incident waves would normally expose portions of the emulsion. However, in this case the portion of the waves reflected back by the mercury will interfere with the incident waves in the emulsion. Now only the areas of the emulsion where the incident and reflected waves reinforce will be exposed. No exposure will take place where the incident and reflected waves cancel each other. If the incident light is monochromatic, the location of exposed areas will be related to the wavelength of the incident light.

In effect these exposed areas form a diffraction grating. When white light is incident upon the developed emulsion at an angle, the color of the reflected light will correspond to the original color exposing the emulsion (refer to the subsection "Bragg's Effect").

The term "hologram" refers to the interference pattern recorded in a film emulsion by the interference between two (or more) wavefronts. Figure 50 shows an arrangement which can be used to record a hologram. In the

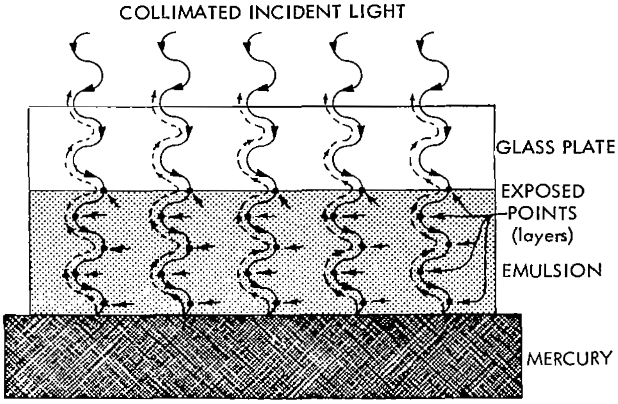


Figure 49—Lippmann color process.

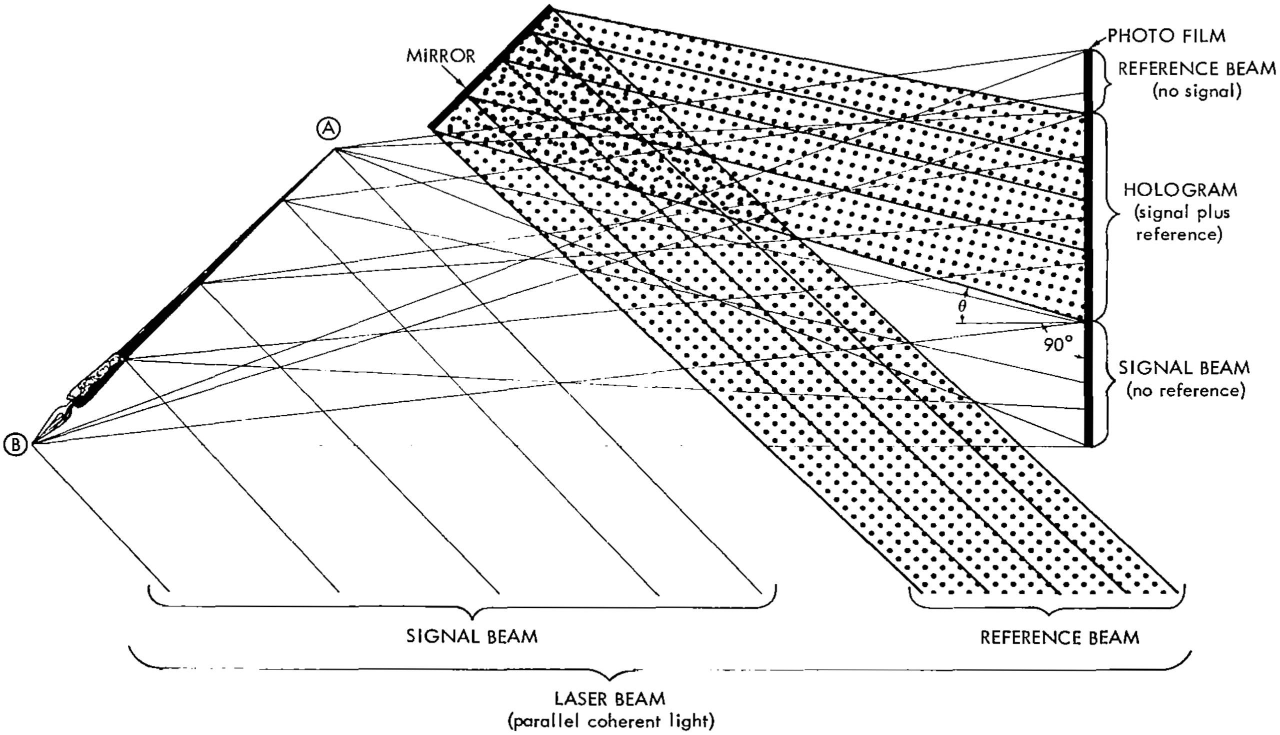

MIRROR
Ⓐ
Ⓑ
PHOTO FILM
REFERENCE BEAM (no signal)
HOLOGRAM (signal plus reference)
θ
90°
SIGNAL BEAM (no reference)
SIGNAL BEAM
REFERENCE BEAM
LASER BEAM (parallel coherent light)

Figure 50—Exposing a hologram.

process shown, the monochromatic parallel coherent light from a laser is divided into two beams. One beam is used as a reference beam and is reflected from a mirror to illuminate the photographic film. The second beam (signal beam) is reflected from an object (subject for which the hologram is made) to illuminate the photographic film. The signal and reference beams combine in the photographic emulsion and produce an interference pattern. As in the case above, the emulsion will be exposed only in areas where the beams reinforce and not where they cancel. The areas where only the signal beam, or only the reference beam, exposes the emulsion will not contain any useful information. Only the areas exposed by the interference pattern produced by the two beams will contain information regarding the object. These areas constitute the hologram of the object.

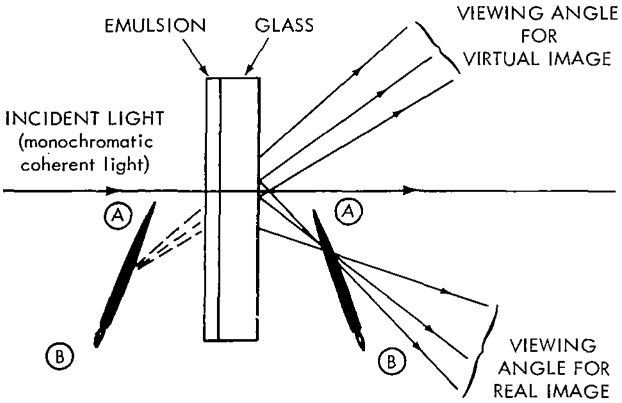
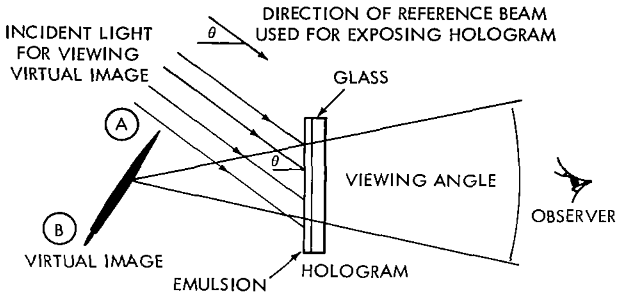

EMULSION
GLASS
VIEWING ANGLE FOR VIRTUAL IMAGE
INCIDENT LIGHT (monochromatic coherent light)
A
A
B
B
VIEWING ANGLE FOR REAL IMAGE

Figure 51—Viewing the images of a hologram.

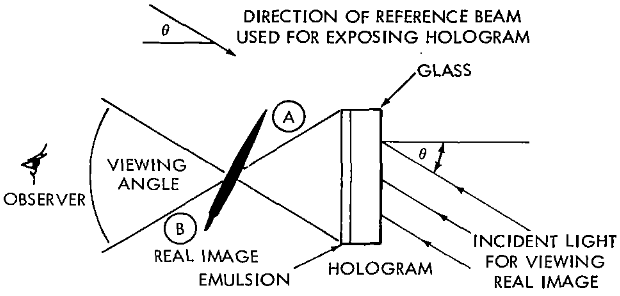
The general arrangement for viewing the images of a developed hologram is shown in Figure 51. Monochromatic coherent light is used to illuminate the hologram, and two wavefronts similar to that reflected from the original object are reconstructed. These wavefronts can be viewed separately (they travel in different directions), as shown by the two viewing angles in Figure 51. At one viewing angle the wavefronts appear to originate from an image behind the hologram plate. Since the light waves viewed do not actually pass through the image points, this image is called the *virtual* image. At the second viewing angle the wavefronts appear to originate from an image in front of the hologram plate. Since the light waves actually pass through the image points, this image is called the *real* image. The definitions used for the real and virtual images are the same as those used in conventional optical imaging theory. When the same reference beam is used for exposing as well as viewing the hologram, an exact three-dimensional image of the original object will be seen as illustrated in Figure 51. This three-dimensional characteristic of holograms is important. Just as in viewing an actual scene, the eye must refocus when looking at a near or a far portion of the reconstructed images. When the observer's view is obstructed by an object in the foreground, it is possible to look behind the obstruction by changing the angle of view (in those cases where it is possible to do so within the limits of the viewing angle). As the observer changes his position within the viewing angle, a distinct change in the perspective of the image is perceived.

Variations of the viewing method discussed above are possible. In fact, both theory and experiment indicate that the reconstruction process is sensitive to the angle of incidence used for the viewing light. The reconstructed images appear to be best when the viewing light is incident at the same angle as the original reference beam. In addition, it is found that one of the images is usually clearer than the other. That is, when the virtual image is very clear, the real image is difficult to see, and, when the real image is clear, the virtual image is difficult to see. It has also been found that, if the virtual image is the clearer of the two when the hologram is illuminated from one side, the real image is the clearer when the hologram is illuminated from the opposite side. This effect can be explained if the arrangement shown in Figure 51 is examined. If the direction of the incident light is reversed (i.e. imagine the light to travel from right to left), all the arrowheads indicating direction would be reversed in Figure 51. The images would then be viewed by looking toward the hologram from the left. Viewed from this new location, what had originally been a virtual image (when viewed from the right) now will appear as a real image (when viewed from the left). It is reasonable to assume, therefore, that if the virtual image was clear when viewing from the right (as illustrated in Figure 51) the real image will be clear when viewing from the left (reverse of Figure 51). Figures 52 and 53 illustrate in greater detail the formation of the real and virtual images.



INCIDENT LIGHT FOR VIEWING VIRTUAL IMAGE
θ
DIRECTION OF REFERENCE BEAM USED FOR EXPOSING HOLOGRAM
GLASS
A
θ
VIEWING ANGLE
OBSERVER
B
VIRTUAL IMAGE
EMULSION
HOLOGRAM

Figure 52—Viewing the virtual image.



θ
DIRECTION OF REFERENCE BEAM USED FOR EXPOSING HOLOGRAM
GLASS
A
VIEWING ANGLE
OBSERVER
θ
B
REAL IMAGE
EMULSION
HOLOGRAM
INCIDENT LIGHT FOR VIEWING REAL IMAGE

Figure 53—Viewing the real image.

Figure 52 shows the arrangement for viewing the virtual image (assumed to be the clear image). The virtual image appears behind the hologram plane. In effect, the hologram acts as a window through which the virtual image is viewed. Any small section of the hologram contains information similar to that of the entire hologram. When the virtual image is viewed through a small section of the hologram, the entire image may not be visible at one viewing angle. It is possible, however, to see all portions of the image by moving through the viewing-angle range. This movement truly gives the effect of looking through a window as if to see into a room, since, as the window is made smaller, one is required to move through a larger angle to see all parts of the room. Some holograms may be quite dark, giving the appearance of a dirty window. This effect presents difficulties in viewing the virtual image of the hologram. The previously discussed ("Photographic Film Basics") technique of bleaching to produce a positive eliminates this dirty-window effect. A bleached "positive" hologram seems to produce better imaging than does a "negative" hologram. It should be noted that, whether the hologram is developed as a negative or positive, the resulting image is the same. That is, the virtual image will always appear as a positive regardless of whether the hologram is a negative or a positive. This is easily understood if a negative is considered as the development of points of reinforcement and a positive is considered as the development of points of cancellation. Since the only difference will be a half-wavelength shift in the position of the interference patterns, either hologram will contain the same pattern and therefore the same information. Thus the same positive virtual image (and real image)is produced in either case.

We have considered the virtual image reconstructed from a hologram. As mentioned above, if we illuminate the hologram from the opposite side the real image can be viewed clearly. Figure 53 shows this arrangement, and comparison with Figure 52 will emphasize the reversal of direction specified in the discussion above. When illuminated by a reference beam as described, the hologram produces a wavefront pattern similar to the wavefront pattern reflected by the object used to make the hologram. The reconstructed real image appears suspended in mid-air between the observer and the hologram, as shown in Figure 53. Although the real image is a precise replica of the illuminated surface of the object used to make the hologram, it appears to the viewer as if he were looking through the object from the rear. This point can be shown by considering the configuration of Figures 50 and 51. In Figure 50 the original object is positioned so that point A is closer to the photographic film than is point B, and point A is on the top of the object. In Figure 51 it can be seen that the position of the virtual image is the same as that of the original object. Therefore, in viewing the virtual image, it appears that the original object is still behind the hologram. When one is viewing the real image, the point B is still farther away from the hologram plate than point A. The real image, however, is now between the observer and the hologram plate, therefore point B is now closer to the observer than point A. This gives the effect of looking through the object from the rear. Point A is still on the top of the object, a position which means that the real image is erect. Points farther away from the observer in the virtual image are closer to the observer in the real image.

If a sheet of film were placed at point A of the real image, it would photograph a focused image of point A and a defocused image of point B. It is important to note that a focused image of point A is produced without any lenses. If points A and B are close together, a true focused picture of the object can be obtained without lenses. If A and B are separated, individual pictures of the points A and B can be obtained by exposing separate films placed at A and B respectively. Therefore, the real image which is formed in front of the hologram can be photographed only piecewise, since the reconstructed image is three-dimensional and is not in a single plane (the usual form of a photographic emulsion). Generally there is some initial difficulty in focusing one's eyes on the real image suspended in space between the hologram and the eye. This fact makes the real image more difficult to see than the virtual image.

Using an original hologram, it is possible to produce a second hologram which will have a real image with true depth perception. Referring to Figure 53, it is apparent that the original hologram does not have a real image with true depth perception, because the observer is effectively looking through the object as if from the rear. True depth perception in a real image can be accomplished by using the original hologram to produce another hologram as shown in Figure 54. The original hologram is illuminated by a reference beam which, upon emerging from the hologram, is reflected by a mirror onto a new photographic emulsion. The real image formed by the original hologram also illuminates the new photographic emulsion. The reference beam reflected from the mirror and the light from the real image interfere to form a second hologram. After development, the second hologram is viewed as shown in the lower diagram of Figure 54. The real image of this second hologram will have true depth perception when viewed within the visual angle.

A hologram developed with any γ will generally be acceptable. A hologram with γ other than 2, however, will produce higher-order images (both real and virtual) of the object. The higher-order images decrease rapidly in intensity as the order becomes higher; however, when a hologram is made using an extremely bright and highly illuminated object, it may be possible to see a second-order image displaced from the first-order image. When one is viewing such a hologram, the first- and second-order images interfere with each other. That is, two distinct "pictures" displaced from each other but superimposed are seen in the viewing angle.

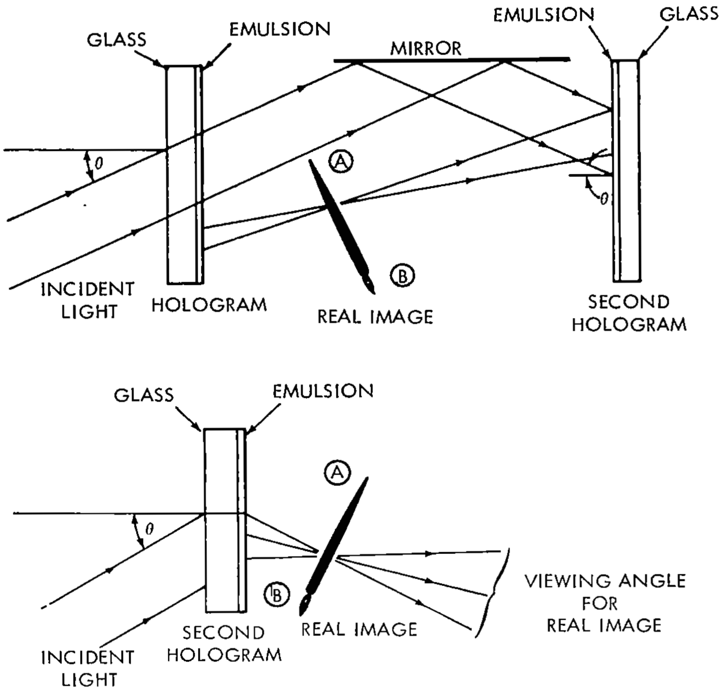


Figure 54—Hologram reproduced from another hologram.

The usual optical distortion problems do not occur in viewing holograms, since no lenses are used. In addition, large magnifications can be achieved without lenses by making the hologram with light of short wavelength and reconstructing images with light of long wavelength. This technique allows the viewed image to be magnified by a factor equal to the ratio of the viewing wavelength to the exposing wavelength. The maximum ratio of visible light wavelengths is less than 2, which limits magnification to values much below the normal lens magnification. If the wavelength ratio between infrared and X-rays is considered, magnifications of 10^8 are possible. However, magnifications of the order of 10^8 require coherent sources of infrared and X-rays which have yet to be developed.

In the making of a hologram, each point on the object reflects light which illuminates the entire hologram. As pointed out earlier, this effect allows any portion of the hologram to reconstruct the entire object. This principle also increases the dynamic tonal range of the reconstructed image of a hologram as compared to conventional photographic methods. This increase in dynamic tonal range is due to the fact that bright areas on the object have their energy spread over the entire recording film. In holograms only secondary effects, such as grain irregularities which produce random scattering, tend to reduce the image contrast. In conventional photography the dynamic range of film is quite limited, and very bright areas in an object are difficult to reproduce faithfully while still maintaining the overall tonal range.

Another interesting property of holograms is that it is possible to superimpose several holographic images on a single hologram. There are several techniques which can be used to achieve this result:

1. Using different incident angles of the reference beam, it is possible to produce on a single film many holographic images which can be reconstructed individually without appreciable crosstalk (one interfering with the viewing of another).

2. After exposing a photographic plate as described in 1 above, reorient the photographic plate (e.g. turn it upside down) and re-expose it by the same method as if it were a new plate. The images of either orientation can be viewed without interference from the images of the other.

3. A hologram of several transparencies (positives such as slides) can be made by placing the transparencies one behind another. Any particular transparency can then be reconstructed by taking a picture through the hologram with a camera which has a limited depth of field. This can be compared to taking a picture through a screen or fence. If the subject behind the screen is in focus, the screen will be out of focus. Under these conditions the subject behind the screen is easily discerned when the photograph is viewed. By changing the focus and limiting the depth of the field, it is therefore possible to take a picture of each transparency with essentially no interference from any of the images of the other transparencies.

Theory of Hologram Construction

Reflected light from an object is described by its amplitude and phase. Considering a point on the object, the reflected waves will travel outward from the point as if it were a source of light. Each point on the object, therefore, can be treated as a point source or origin for a light wave. The exact form of the wave pattern reflected from a three-dimensional object is extremely complex. In holography, the signal-bearing waves reflected from the objects are combined with reference waves to produce an interference pattern which is recorded photographically. Normally photographic film records the intensity of light incident upon it as an optical density pattern. A photographic emulsion therefore does not normally record the phase characteristics of the incident light. In order to record the phase characteristics of signal-bearing waves, it is necessary to represent the phase as intensity variations. When two coherent beams of light, with plane wavefronts, are incident upon a screen at different angles, a set of uniform parallel interference fringes will be produced. If an irregular wavefront coming from the object is substituted for one of the plane waves, a complex interference pattern will be produced. Variations in the amplitude of the signal-bearing wave (reflected from the object) will produce variations in the density of the interference fringes as photographically recorded. The variations in phase of the signal waves with respect to the reference waves produce variations in the spacings of the fringes recorded. Thus a hologram is an interference pattern produced by a reference beam of light and the light reflected from each point of the object. Both the phase and amplitude of the light reflected from each point of the object are recorded as intensity variations.

Essentially the hologram can be considered to be made up of many nonuniformly spaced and oriented slits. We have previously discussed how plane waves are diffracted at various angles from uniformly spaced slits. A hologram also has the various orders of diffracted wavefronts radiating from the effective slits (interference pattern) of the hologram. In the case of the hologram, however, the various orders of diffracted wavefronts are extremely complex, as are the orientation and spacing of the slits. That is, when the spacing of the slits is irregular (some regions have closer line spacings than others), there will be corresponding variations in the direction of propagation for the diffracted wavefronts. In addition, variations in density of the fringes will produce variations in the amplitudes of the diffracted wavefronts. In effect, therefore, slit spacing variations correspond to the phase of the signal beam, while the amplitude of the signal beam corresponds to the contrast of the exposed emulsion. The greater the contrast, the stronger the wavefront produced. Slits which have low contrast between a slit and a space will produce weak, or low-amplitude, diffracted wavefronts.

When a hologram is produced, it would normally be considered a negative. The fact that it produces positive pictures is due to the fact that the wavefronts emanating from it will be duplicates of the wavefronts emanating from the actual object. A contact print of the hologram would normally be considered a positive (i.e. opaque areas would now be transparent and vice versa). The image constructed from a copy of a hologram is also positive. As discussed earlier, the basic slit pattern of the hologram is not changed whether the hologram is a "negative" or a "positive" and therefore the same image will be produced in either case.

The contrast rendition of the original object is reproduced very closely by a hologram regardless of contrast properties of the photographic emulsion used to record it. Actually a hologram can be produced by a film emulsion which is capable of producing only two optical levels (transparent and opaque). The tonal rendition of the image will not suffer, because each individual slit diffracts some light to a point related to the object. Therefore, using only two levels of density in the hologram plate, the entire tonal range of the original object is reproduced by the combination of light from all the slits. The tonal range for a hologram can be much less than that required for normal photography. In normal photography the film must be capable of reproducing the relative intensity of a point on the original object at a corresponding point on the emulsion. In holograms the light radiated from each point on the object is distributed over the entire film surface. It is apparent from the difference between the two techniques that a higher-intensity point on an object which would saturate the film in normal photography can be recorded faithfully in holography

In effect, the hologram plate acts similarly as does a zone plate. The interference between the reference beam and the light reflected from one particular point on the object gives rise to a set of elliptical quasi-concentric interference fringes in the hologram. The two first-order diffraction wavefronts produced by these fringes produce the virtual- and real-image beams. The lens properties of these fringe patterns are similar to those of a modified group of zone plates. The spacing of the zones on a zone plate determines its focal length. The focal length will change if the frequency of the incident light changes. These effects are also present in a hologram. Any shrinkage in the hologram plate will change the spacing between zones of the effective zone plate and produce irregularities in the image. In addition, change of frequency or temporal noncoherence of the reference beam used to view the hologram will produce corresponding changes in the focal length of the effective hologram zone plates.

Coherent Light and Holograms

In order to view the holograms produced by the methods discussed above, it is necessary to use a laser beam to illuminate the hologram. It is possible, however, to produce a hologram-type interference pattern which does not require a laser for viewing. Figure 55 shows a method for exposing a standing-wave-type hologram. The laser beam is collimated and projected to illuminate the object. The reflected light from the object illuminates the hologram. A portion of the laser beam is diverted by a beam splitter to a mirror which reflects the light into the hologram. The light reflected by the object is again the signal beam. The reference beam is the diverted beam from the laser which is reflected by the mirror. It is to be noted that the reference beam and the signal beam from the object are traveling through the photographic plate in different directions. Standing waves are set up *in depth* in the emulsion. The emulsion will be exposed only at the points where the two waves add constructively. The similarity between this type of exposure and that used by Dr. Lippmann is evident. Both methods require standing waves to be set up in the emulsion (in depth). The previous type of hologram discussed required primarily a surface contrast record, whereas the standing-wave hologram forms exposed points within the emulsion. The exposed points within the emulsion can be considered as an interference filter.

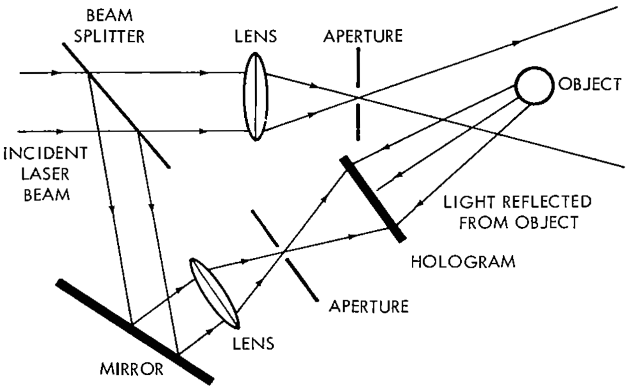


Figure 55—Method for producing a standing-wave-type hologram.

A hologram produced by the method shown in Figure 55 can be viewed using a point white light source. Figure 56 shows a point white light source illuminating a standing-wave-type hologram produced as shown in Figure 55. The three-dimensional hologram effect is evident when the standing-wave-type hologram is viewed by reflected light. The standing-wave-type hologram (or reflectance hologram) is much more difficult to produce than the "normal" transmission-type hologram. Some of the reasons for these difficulties are:

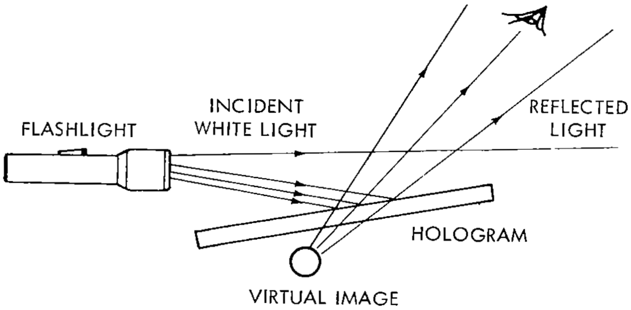


Figure 56—White-light viewing of a standing-wave-type hologram.

1. Exposing the standing-wave-type hologram is more difficult, since sound vibrations tend to vibrate the emulsion. Considering the emulsion as a thin membrane held at its outer periphery, the similarity to the membrane of a drum is evident. As in a drum, any sound vibrations will tend to make the membrane vibrate. Since the exposure is being made in the emulsion in depth, any vibrations will tend to blur it. In the normal hologram a slight

vibration (moving surface) will not cause distortions of as great a degree, since the interference pattern is recorded only on the surface.

2. When the image is formed in depth, any shrinkage of the film will change the spacing between the exposed points of a hologram. This distortion will in fact be present, since there is shrinkage in the normal processing of films.

3. The problems described for the Lippmann emulsion apply also to the standing-wave-type hologram. One such problem is the absorption of the signal beam before it can penetrate deep enough to form the standing waves in depth.

4. When white light is used to view the hologram, as the viewing angle changes the color of the hologram image will change. This color change is basically caused by the Bragg effect. The Bragg effect predicts a downward shift in the frequency of the reflected light when the viewing white light is incident at increased angles.

Although there are more problems involved in making a standing-wave-type (reflectance) hologram, it would seem that the ultimate future of holography lies in the standing-wave-type hologram. At GSFC we have succeeded in producing a standing-wave transmission hologram (viewed with transmitted light) which can be viewed with *white* light. To the best of our knowledge, this standing-wave transmission hologram is the first of its type to be produced and permanently fixed. The problems of final fixing of the surface-type hologram plate do not seem to be as great as for the standing-wave transmission type hologram. In the reflectance (or our transmittance) standing-wave-type hologram, the shrinkage due to processing is usually avoided by not pursuing the final fixing of the hologram plate. When the hologram is developed but not fixed, the developed image can be expected to fade eventually. Not fixing the hologram plate prevents much of the shrinkage, since the basic content of the emulsion after development will be the same as it was before (i.e. unexposed silver halide will not be removed from the emulsion).

As mentioned above, there is a change in the color of the subject matter recorded in a reflectance hologram when the viewing angle is changed. This color change occurs also in the case of a transmission-type standing-wave hologram. This color change, however, is not a significant disadvantage in comparison to the advantage of using a white light source for viewing. The possibility of using a white light source to view holograms provides the opportunity for significant practical application of holograms.

The hologram process requires temporal coherence because the light reflected from each point on the object must individually interfere with the reference beam. It is not necessary that the light from adjacent points on the object be able to produce an interference pattern. When viewing the hologram, the opposite requirements are necessary. For viewing, adjacent points on the hologram plate must be able to produce an interference pattern. This requirement implies that spatial coherence is necessary for viewing. The interference pattern produced by the light from adjacent points on a hologram plate gives rise to a reconstructed wavefront similar to that which emanated from the original object. Temporal coherence is not required for viewing, because each point in the light wave gives rise to a divergent wavefront. Some degree of temporal coherence is desired, however, since any variation in wavelength will tend to vary the focal length (change the divergence of the wavefronts). A point source of white light (as from a flashlight) has spatial coherence at large distances. The size of the point source as compared to the distance from the source determines the degree of spatial coherence. Temporal coherence is determined basically by the bandwidth of the light source. The use of a temporally noncoherent source will produce variations in depth of the object, while a lack of spatial coherence tends to produce longitudinal blurring.

The combination of the human eye and brain is capable of detecting partially obscured items. Basically the eye-brain combination integrates any variations out of a picture so that any slight blurring (due to spatial noncoherence) or slight variations in depth (temporal noncoherence) are eliminated. Figure 57 demonstrates the ability of the eye-brain combination to interpret and distinguish patterns although they are partially obscured. This basic process occurs unconsciously when one is viewing a hologram illuminated by partially (spatially or temporally) coherent light.

When producing a hologram, it is possible to illuminate the object through a ground-glass screen which will destroy the spatial coherence of the beam. Any point on the object then receives light from virtually the entire ground-glass screen and has some fixed value of amplitude. The frequency of the light at any particular point on an object will still vary in accordance with the frequency of the light source. If the light source is temporally coherent, the light at a point on the object will maintain temporal coherence even though the spatial coherence has been destroyed. The temporally coherent wave emanating from a point on the object will radiate out and interfere with the reference beam, producing an interference pattern on a photographic plate. This interference pattern is produced even though there is no spatial coherence.

A modification of the three-dimensional hologram described above has been produced—independently—at Bell Telephone Laboratories and at the University of Michigan. This modification permits three-dimensional multicolored images to be seen by reflected white light from a point source. Multicolored holograms are made by combining two laser beams of different colors to form a single beam. The single beam combination is then used to make a hologram. The angle between the reference and signal beams is made large—approximately 160°. (The usual range of angles is from 30° to 90°.) With this larger angle between the reference and signal beams, color information is recorded and ordinary white light can be reflected off the hologram to produce high-quality three-dimensional multicolored images.

Summary of Hologram Applications

The most immediate application for holograms is lensless photography which will produce true "three-dimensional" images from a flat film plate. Conventional three-dimensional techniques requiring stereo pairs and special viewing glasses do not permit the observer to move off center without causing distortion. When the observer (without special viewers) moves off center to view the hologram, he sees the imaged scene from a new angle. This allows viewing around, behind, under, and over objects. The observer must also focus his eyes on different parts of the scene just as if the objects were actually there. In conventional three-dimensional stereo pairs, the eyes

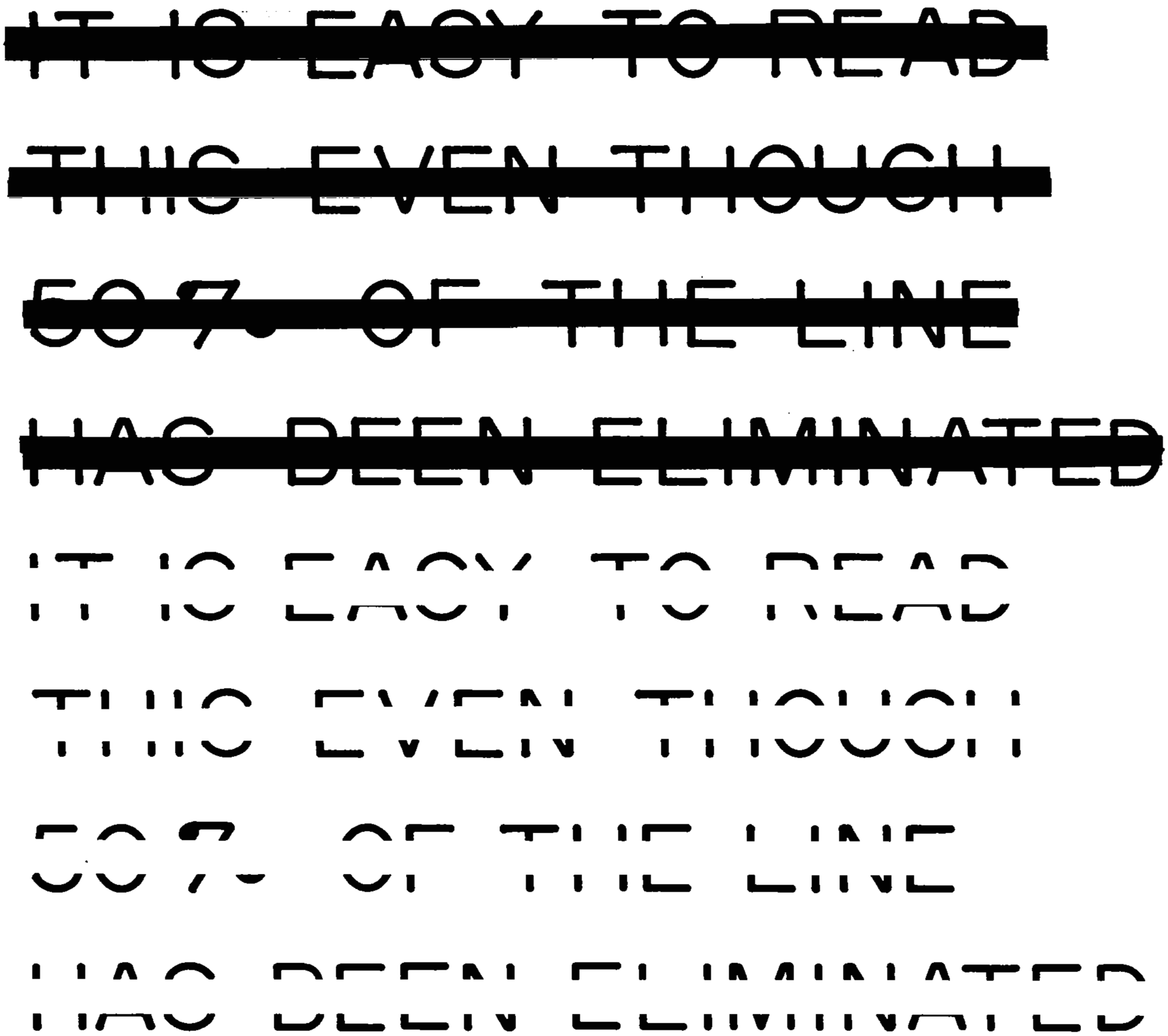


Figure 57—Eye-brain combination effects.

are focused to the picture plane and do not have to be refocused to see "depth." Holograms thus provide the means for obtaining true three-dimensional viewing which has previously been possible only with real scenes.

The fact that there is no one point on the hologram plate corresponding to a point on the object recorded will be very important in applications which require errorless information storage. Since each section of a hologram contains information concerning the entire recorded image, the destruction or loss of any section will not result in a loss of any significant amount of information. For example, a large scratch on the hologram will not be seen in the image.

As mentioned earlier, holograms can be used to obtain magnification greater than that possible in conventional lens systems. A hologram made with short wavelength and viewed with long wavelength will cause the object to appear magnified by a factor equal to the ratio of the two wavelengths. Additional magnification can be obtained by viewing the hologram in light which is more divergent than that which produced the hologram.

The display of information to man in a more effective manner through the use of holograms appears to be a significant possibility. The three-dimensional capabilities and even color effects will make the storage and presentation of more information possible. There are many other potential applications open to holograms. Initial efforts, however, will probably be directed along the lines indicated by the principles discussed above.

Transmission of Holograms

One of the basic problems in the applications of holography is the inherent difficulty in transmitting or reproducing a given hologram. The interference pattern recorded within the emulsion is such that it would be extremely difficult to reproduce by any means. Consider a contact print, which is the most accurate reproduction method used in normal photography. Reproduction of high-resolution holograms by contact printing is extremely difficult, since the hologram emulsion contains a three-dimensional interference pattern. In effect, contact printing will transfer only the shadow of the original hologram; it would not allow the reproduction of an exact duplicate in depth. Even assuming that the hologram was a surface pattern on the film, the minutest separation between emulsions would produce diffraction which would not allow an accurate transfer of the hologram pattern. The small spacings between information fringes on a hologram form effective apertures. The diffraction effects of these small apertures plus the refraction effects due to thickness of the plates will cause the light to be spread out. This spreading out of the light makes the accurate reproduction of holograms by contact printing impossible.

Unless a technique can be developed which will allow the information of a hologram to be transmitted, the full potential of holography as applied to information processing will not be realized. Enlarging a hologram and scanning it optically with a photocell are not an acceptable method. The electronic transmission of the detected information requires an excessively large bandwidth (or time). The normal projection-type copying with an enlarger has not been fully explored as a technique for reproducing holograms. However, it appears that the resolution capabilities of optical lenses will not be sufficient to allow copying by this means. A special type of projection copying which consists of the transmission of the hologram information on a light beam may offer a possible solution. In this method the original hologram information in the light beam would be reconstructed at the receiving end to produce a new hologram. Further investigation of this technique requires the prior development of light-communications techniques.

A technique for reproducing holograms by using Ronchi Rulings may be practical. A Ronchi Ruling is composed of evenly spaced parallel lines. The width of a line is exactly equal to the width of the space between lines. Ronchi Rulings have been used commercially for the preparation of half-tone negatives from continuous-tone photographs. A continuous-tone photograph is the normal photograph, which has continuous gradations from pure white to pure black. A half-tone negative has gradations represented by equally spaced dots of various sizes. That is, dot size variations produce a representation of the continuous-tone photograph. Two Ronchi Rulings are placed so that the lines are at right angles to each other, forming what is called a half-tone screen. This half-tone screen is placed in front of the negative so that the image formed by the camera lens must pass through it. By placing the screen a specified distance from the film emulsion during exposure, it is possible to break up the continuous image (formed by the lens) into a system of graded dots. The location of the screen for normal applications is usually given by the formula

$$\frac{\text{Screen-to-film distance}}{\text{Screen aperture}} = 64 .$$

A typical example would be a half-tone screen with 120 lines per inch (Ronchi Rulings). The screen aperture, which is the size of each open square,will be 1/240 of an inch. The screen-to-film distance would then be equal to 64 × 1/240 or 4/15 of an inch. Once the screen-to-film distance has been fixed, it is necessary to maintain the ratio of 64 to 1 by varying the lens aperture when the lens-to-film distance is changed. That is, the aperture stop is equal to 64 times the focal length of the camera divided by the lens-to-film distance. An adaptation of this technique might be applied to the reproduction of holograms. As mentioned earlier, a scratch or other imperfection in the hologram does not appreciably affect the viewed image. Therefore, a half-tone or some similar screen placed in front of the hologram for reproduction should not appreciably impair the reproduced image. The normal Ronchi Rulings of 65 or 130 lines per inch would not be suitable for hologram reproduction. However, diffraction gratings of 15,000 lines per inch and greater have been produced with no difficulty. This half-tone screen technique merits further investigation as a possible technique for hologram reproduction.

Still another technique that should be considered is the use of Moiré patterns. Moire patterns are produced when families of curves are superimposed, forming patterns very different from the original curves. As an example, these patterns can be observed by superimposing two Ronchi Rulings and rotating one with respect to the other. Applying the principle of Moiré patterns, it may be possible to change the information pattern of a hologram by using a reference ruling. The new pattern can be transmitted and the original hologram pattern reconstructed by using a reference ruling at the receiving end. Application of this technique may provide a means for reducing the bandwidth requirements and allow electronic transmission of hologram information.

ANALYSIS OF OPTICAL DATA PROCESSING SYSTEMS

Development of Block Diagram Approach

Now that we have completed a basic review of optics, optical data processing, and holograms, we will attempt a general analysis of optical transform systems. In this analysis a building-block approach is used to facilitate the mathematical development of some of the principles involved in optical systems.

The reversibility characteristics of Fourier transforms are shown by the flow diagram in Figure 58. The double path represented by oppositely directed arrows between the functions $f(x)$ and $F(\omega)$ indicates that each is given by the Fourier transform of the other. In other words, the Fourier transform of $f(x)$ is $F(\omega)$ and the Fourier transform (inverse transform) of $F(\omega)$ is $f(x)$. A single arrow from the amplitude spectrum $F(\omega)$ to the power spectrum ($P(\omega)$ indicates that this process is not reversible. Since the phase is lost in going from the complex amplitude spectrum to the power spectrum, it is impossible to determine the phase of the amplitude spectrum from the power spectrum. This irreversibility holds in any system that uses a square law detector to monitor the power of a signal with complex amplitude.

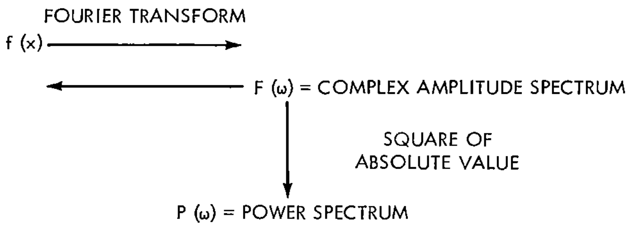


Figure 58—Fourier transform flow diagram.

Using the flow scheme discussed above and the optical terms discussed earlier, we can obtain an optical flow diagram as shown in Figure 59. It is noted that this optical flow diagram has the same reversibility characteristics as the Fourier transform diagram shown in Figure 58. The Fourier transform operation is performed by the focusing of a diffraction pattern, as shown earlier, and a photodetector performs the one-way squaring operation from the complex amplitude spectrum to the power spectrum (intensity spectrum).

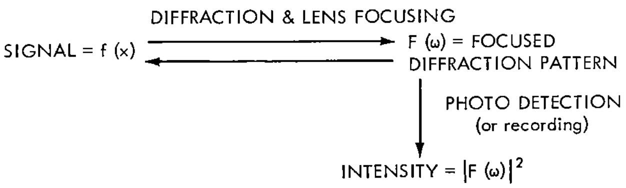
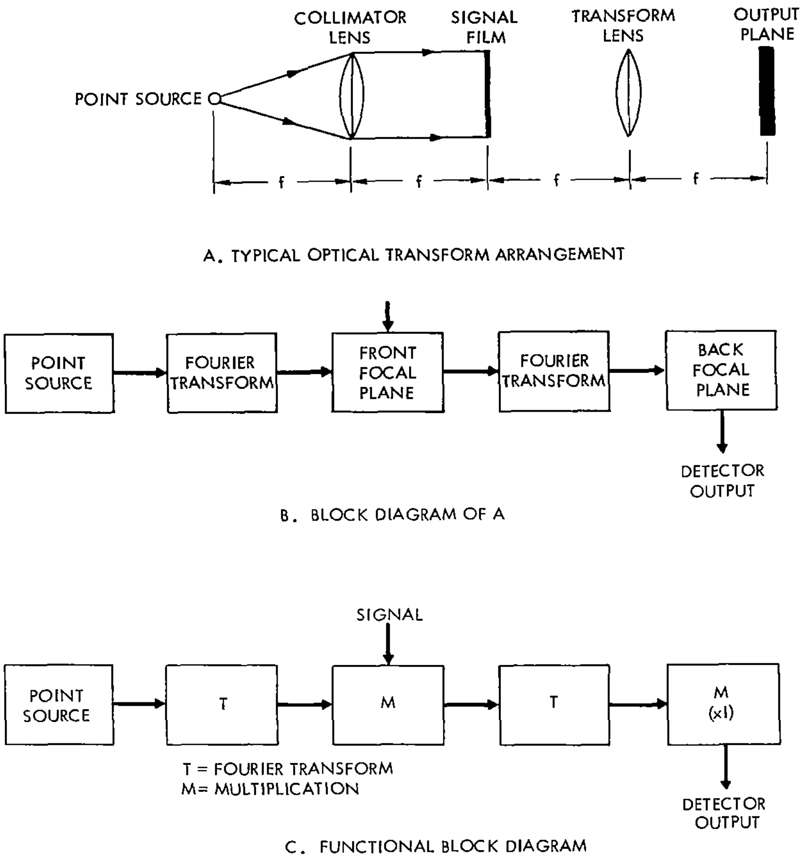

DIFFRACTION & LENS FOCUSING
SIGNAL = f (x)
F (ω) = FOCUSED DIFFRACTION PATTERN
PHOTO DETECTION (or recording)
INTENSITY = |F (ω)|²

Figure 59—Optical flow diagram.

From our previous discussions of diffraction theory, it is readily seen that the operations represented by the oppositely directed arrows in Figures 58 and 59 are performed when light passes from the front focal plane to the back focal plane of a lens. This immediately suggests that a lens can be represented by a transform function and a block diagram representation of an optical system can be developed, as shown in Figure 60. A typical optical transform arrangement is shown in Figure 60 (A). Figure 60 (B) shows a block diagram corresponding to this optical system. In this diagram, regular arrows represent the flow of information (as light amplitude) into or out of a function block, and diamond-head arrows represent information (e.g. films) inserted into, or extracted from, a focal plane. Examining the system illustrated in Figure 60 (A) it is to be noted that the signal film plane can be called the front focal plane of the transform lens, or the back focal plane of the collimator lens. This can lead to confusion, as in Figure 60 (B), where it is not clear to which lens the "Front Focal Plane" label refers. To eliminate this confusion and at the same time improve the functional significance of the block diagram approach, it is necessary only to recall that the light amplitude transmitted by a signal film is the *product* of the incident-light amplitude and the signal transmission function. It is immediately apparent that the focal plane



A. TYPICAL OPTICAL TRANSFORM ARRANGEMENT

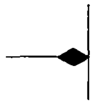
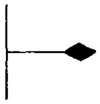
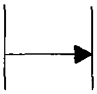
B. BLOCK DIAGRAM OF A

C. FUNCTIONAL BLOCK DIAGRAM

Figure 60—Development of an optical block diagram.

Table 5

Optical Block Diagram Symbols.

Symbol	Description	Symbol	Description
	Signal Input. Usually inserted as a film transmission function. Can be inserted only into a multiplier block.	( ) Source	Light Source. Label specifies type of source (e.g. point source). Note that there are no inputs.
	Signal Output. Photographically recorded or photodetected. Note that output is Intensity = Amplitude^2 and must be taken from a multiplier block.	T	Fourier Transform. Output-light amplitude is the Fourier transform of the input-light amplitude.
	Flow Path. Arrow indicates direction of information flow (direction of light propagation).	M	Multiplier. Output-light amplitude is the product of the input-light amplitude and the inserted transmission function. Note that, if no film is present, the transmission function is equal to 1.

notations used in Figure 60 (B) can be replaced by a "Multiply" notation. We can now represent our typical optical system as shown in Figure 60 (C). Note that if there is no signal film in a focal plane, the transmission function is equal to 1, as indicated by (x1) in the output plane represented in Figure 60 (C). To facilitate further applications of the block diagram symbols developed in Figure 60, Table 5 lists all symbols required in our following discussions.

Point Source and Collimated Light

It has been shown geometrically that a point source at the front focal point of a lens produces collimated light beyond the lens. This principle was used to obtain incident light of constant amplitude and phase in the optical systems discussed in this report. We will now investigate this principle, referring to the block diagram representation shown in Figure 61. It is seen from Figure 61 that collimated light is effectively the Fourier transform of the light amplitude of a point source. As mentioned above, the arrangement shown in Figure 61 is used to obtain light of constant amplitude and phase. In other words, the Fourier transform of the light amplitude of a point source was assumed to be a constant.

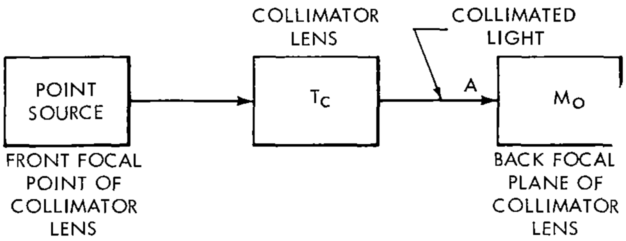

COLLIMATOR LENS
COLLIMATED LIGHT
POINT SOURCE
T_C
A
M_O
FRONT FOCAL POINT OF COLLIMATOR LENS
BACK FOCAL PLANE OF COLLIMATOR LENS

Figure 61—Point source and collimator lens.

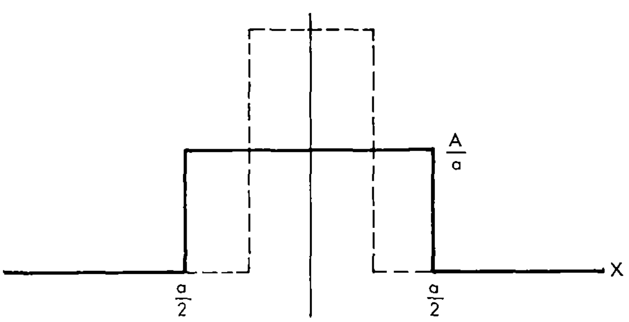


Figure 62—Square wave approaches impulse as $a \to 0$.

We will now consider an impulse function which can be used to mathematically describe an ideal point light source. An impulse is defined as the limit approached when the width of a square pulse is made to approach zero while the area of the pulse is held constant. If we consider a square pulse of width "a" and amplitude A/a, the area of the pulse will be A (constant). This type of square pulse is shown in Figure 62. As shown by the dashed square pulse in Figure 62, in order to maintain constant area, the amplitude (A/a) increases as the width "a" decreases. As the width "a" approaches zero, the amplitude becomes infinite. Mathematically the amplitude $S(x)$ of an impulse may be expressed as a function of x by the expression

$$S(x) = \lim_{a \to 0} \frac{A}{a} w(x) , \qquad (16)$$

where

A = constant area,

a = width of pulse,

$$w(x) = \begin{matrix} 1 \text{ where } |x| < a/2 \\ 0 \text{ where } |x| > a/2 \end{matrix} .$$

The term $w(x)$ defines the width of the pulse, since the amplitude $S(x)$ given by Equation 16 will be zero when $W(x)$ is zero.

Since an ideal point source has zero width, we can use the impulse function given by Equation 16 to describe the light amplitude of an ideal point source. The Fourier transform of the light amplitude can then be written

$$S(\omega) = \int_{-\infty}^{\infty} \lim_{a \to 0} \frac{A}{a} w(x) e^{-j\omega x} dx . \tag{17}$$

Since the integration is with respect to x, we can take the limit and amplitude terms outside the integral sign:

$$S(\omega) = \lim_{a \to 0} \frac{A}{a} \int_{-\infty}^{\infty} w(x) e^{-j\omega x} dx . \tag{18}$$

Since $w(x)$ is equal to zero for $|x| > a/2$, we need consider the integral only between the limits $-a/2$ and $a/2$, where $w(x)$ is equal to 1:

$$S(\omega) = \lim_{a \to 0} \frac{A}{a} \int_{-a/2}^{a/2} e^{-j\omega x} dx . \tag{19}$$

This integral is a standard form which may be found in any table of integrals. Performing the integration and evaluating at the limits, we obtain

$$S(\omega) = \lim_{a \to 0} \frac{A}{a} \left[\frac{-e^{-j(\omega a/2)} + e^{j(\omega a/2)}}{j\omega} \right] . \tag{20}$$

We can simplify Equation 20 by noting that the sine function can be expressed in terms of exponentials as

$$\sin\theta = \frac{e^{j\theta} - e^{-j\theta}}{2j} .$$

Substituting $\sin(\omega a/2)$ for the exponentials in Equation 20, we obtain

$$S(\omega) = \lim_{a \to 0} A \frac{\sin \frac{a\omega}{2}}{\frac{a\omega}{2}} . \tag{21}$$

If we apply the limit as "a" goes to zero we obtain

$$S(\omega) = A , \tag{22}$$

since $\left[\sin(a\omega/2)\right]/(a\omega/2)$ approaches 1 as "a" approaches zero. That $\left[\sin(a\omega/2)\right]/(a\omega/2)$ approaches 1 is immediately seen when we recall that $\sin\theta \cong \theta$ for small angles—i.e. $\left[\sin(a\omega/2)\right]/(a\omega/2) = (a\omega/2)/(a\omega/2) = 1$. Thus we have shown that an ideal point source described as an impulse

$$S(x) = \lim_{a \to 0} \frac{A}{a} w(x) \tag{23}$$

will produce light of constant amplitude A and constant phase when passed through a collimator T_c as shown in Figure 61.

It was shown that, for an ideal point source described by Equation 22, as the source width "a" decreases, the amplitude A/a increases. It should be obvious that in practice the amplitude of a source does not increase as the width of the source is decreased. The practical case for amplitude is more closely approximated by the expression $Aw(x)$, where it is assumed that the amplitude A is constant in magnitude and phase within the source width. This expression is more realistic than Equation 22, since it implies that the amplitude of light in the source remains constant as the width of the source is decreased. The Fourier transform of this light amplitude results in the light amplitude

$$S(\omega) = A \int_{-a/2}^{a/2} w(x)\, e^{-j\omega x}\, dx = \frac{aA}{2} \frac{\sin\frac{a\omega}{2}}{\frac{a\omega}{2}} . \tag{24}$$

For small "a,"

$$\frac{\sin\frac{a\omega}{2}}{\frac{a\omega}{2}} \cong \frac{\frac{a\omega}{2}}{\frac{a\omega}{2}} \equiv 1 ,$$

and the amplitude A can be written as

$$S(\omega) = \frac{aA}{2} . \tag{25}$$

Now as "a" is made small the amplitude $S(\omega)$ in the back focal plane decreases because of the factor "a" which appears in Equation 25.

It has been shown that constant light amplitude across the back focal plane is obtained only for an ideal source, as described by Equation 22 above. A practical source was shown by Equation 25 to produce an approximately constant light amplitude in the back focal plane of a collimator lens. This approximation was obtained by:

1. Neglecting amplitude and phase variations across the width of the source, and

2. Neglecting the variations due to the coordinate ω in the back focal plane (by assuming the approximation $\left[\sin(a\omega/2)\right]/(a\omega/2) = 1$. These approximations improve as the source width is decreased. In our following analysis we will assume that the collimator lens produces constant light amplitude and phase across the back focal plane of the collimator lens.

Optical Spectrum Analyzer

The simplest configuration for an optical data processor is that of a spectrum analyzer as shown in Figure 60, which is repeated in Figure 63 for convenience. The output of the collimator (T_c) is assumed to be a constant (A), as discussed in the preceding subsection. We will designate the signal input to the multiplier M_0 as $f(x)$. The output of M_0 will then be given by the product $Af(x)$. This product $Af(x)$ is operated on by T_1, resulting in the Fourier transform $AF(\omega)$. $AF(\omega)$ is then the input to multiplier M_1 and, since no other signal is inserted into M_1 (multiplication by 1), the output of M_1 is also $Af(\omega)$. The detected output will then be $|AF(\omega)|^2$ (see Figure 59), which is the power spectrum for the signal $f(x)$.

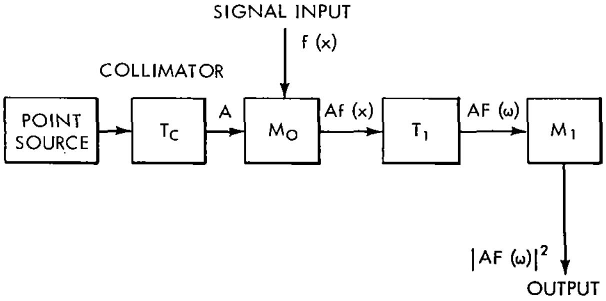


Figure 63—Optical spectrum analyzer.

We will now consider some of the details of the transform operation by developing the mathematics for a sample input signal. Let us consider a film with a transmission function of the form

$$f(x) = \frac{1}{2}\left[1 + \cos\omega_0\left(x + x_0\right)\right], \qquad (26)$$

where

ω_0 = spatial angular frequency,

x = spatial coordinate,

x_0 = displacement with respect to $x = 0$.

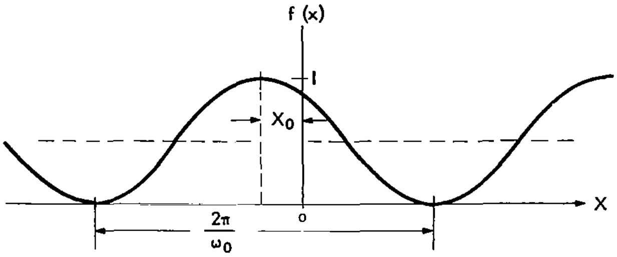


Figure 64—Graph of $f(x) = 1/2\left[1 + \cos\omega_0(x + x_0)\right]$.

The graph of the transmission function $f(x)$ given by Equation 26 is shown in Figure 64. A cosine function can be expressed in terms of exponentials as

$$\cos\theta = \frac{e^{j\theta} + e^{-j\theta}}{2}.$$

Expressing the cosine term in exponential form, the transmission function of Equation 26 can be rewritten

$$f(x) = \frac{1}{2}\left[1 + \frac{1}{2}e^{j\omega_0(x+x_0)} + \frac{1}{2}e^{-j\omega_0(x+x_0)}\right] . \tag{27}$$

In our previous discussion of the spectrum analyzer, we noted that a single-frequency transmission function, such as the $f(x)$ we are considering here, produced three spectral points. Intuitively we might expect each of the three terms in Equation 27 to correspond to one of the spectral points. We can prove that this is so by taking the Fourier transform of $Af(x)$. We will maintain the order of the three terms in Equation 27 so that we may examine our results for correspondence between each term and a spectral point. The Fourier transform operation on $Af(x)$ is represented by T_1 and is given by the expression

$$AF(\omega) = \int_{-\infty}^{\infty} Af(x)\, e^{-j\omega x}\, dx .$$

Letting A equal 1, and substituting the expression for $f(x)$ given by Equation 27, we obtain

$$AF(\omega) = \int_{-\infty}^{\infty} \frac{1}{2}\left(1 + \frac{1}{2}e^{j\omega_0(x+x_0)} + \frac{1}{2}e^{-j\omega_0(x+x_0)}\right)e^{-j\omega x}\, dx . \tag{28}$$

It was previously explained that the limits of integration can be replaced by the aperture limits, since $f(x)$ is zero beyond the aperture dimensions. Therefore, letting "$2a$" equal the length of the aperture, Equation 28 can be written as

$$AF(\omega) = \int_{-a}^{a} \frac{1}{2}e^{-j\omega x}\, dx + \int_{-a}^{a} \frac{1}{4}e^{j\omega_0(x+x_0)-j\omega x}\, dx + \int_{-a}^{a} \frac{1}{4}e^{-j\omega_0(x+x_0)-j\omega x}\, dx . \tag{29}$$

We can now factor the functions appearing under the integral and place constant factors outside the integral:

$$AF(\omega) = \frac{1}{2}\int_{-a}^{a} e^{-j\omega x}\, dx + \frac{e^{j\omega_0 x_0}}{4}\int_{-a}^{a} e^{j(\omega_0-\omega)x} + \frac{e^{-j\omega_0 x_0}}{4}\int_{-a}^{a} e^{-j(\omega_0+\omega)x}\, dx . \tag{30}$$

The integrals appearing in Equation 30 are standard forms which may be found in any table of integrals. Performing the integration and evaluating at the limit, we obtain

$$AF(\omega) = \frac{-e^{-j\omega a} + e^{j\omega a}}{2j\omega} + \frac{e^{j\omega_0 x_0}}{4}\left[\frac{e^{j(\omega_0-\omega)a} - e^{-j(\omega_0-\omega)a}}{j^{(\omega_0-\omega)}}\right] + \frac{e^{-j\omega_0 x_0}}{4}\left[\frac{-e^{-j(\omega_0+\omega)a} + e^{j(\omega_0+\omega)a}}{j^{(\omega_0+\omega)}}\right] . \tag{31}$$

We can simplify Equation 31 by noting that the sine function can be written in terms of exponentials as

$$\sin\theta = \frac{e^{j\theta} - e^{-j\theta}}{2j}$$

and using this exponential definition for the sine in rewriting Equation 31 as

$$\mathrm{AF}(\omega) = \frac{\sin a\omega}{\omega} + \frac{e^{j\omega_0 x_0}}{2}\,\frac{\sin a\left(\omega_0 - \omega\right)}{\left(\omega_0 - \omega\right)} + \frac{e^{-j\omega_0 x_0}}{2}\,\frac{\sin a\left(\omega_0 + \omega\right)}{\left(\omega_0 + \omega\right)} . \qquad (32)$$

We can change each term in Equation 32 into the familiar form $\sin\theta/\theta$ by multiplying $\mathrm{AF}(\omega)$ by a/a:

$$\mathrm{AF}(\omega) = a\left[\frac{\sin a\omega}{a\omega}\right] + \frac{a\,e^{j\omega_0 x_0}}{2}\left[\frac{\sin a\left(\omega_0 - \omega\right)}{a\left(\omega_0 - \omega\right)}\right]$$

$$+ \frac{a\,e^{-j\omega_0 x_0}}{2}\left[\frac{\sin a\left(\omega_0 + \omega\right)}{a\left(\omega_0 + \omega\right)}\right] . \qquad (33)$$

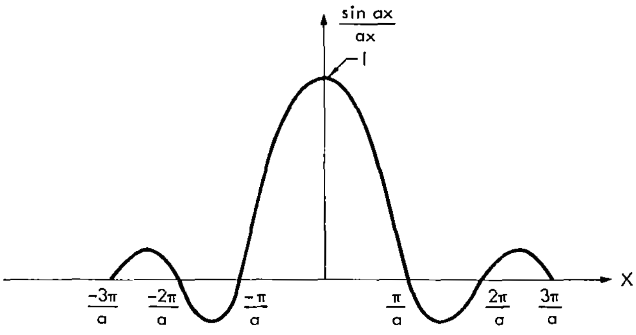
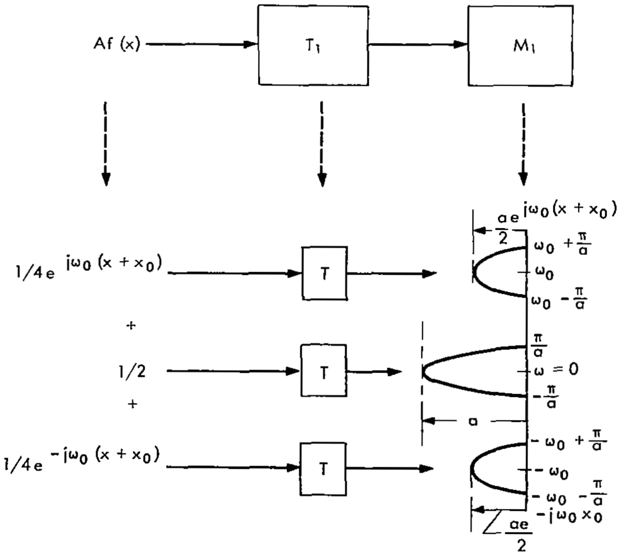


Figure 65—Sin ax/ax versus x.

The curve for $\sin ax/ax$ versus x is shown in Figure 65. It is seen that the function has its greatest peak for $x = 0$. The amplitudes of the higher-order peaks are seen to drop off rapidly. It can therefore be assumed that an approximation for the function $\sin ax/ax$ can be made by considering only the maximum peak centered at $x = 0$ with a total width of $2\pi/a$. Noting that we maintained the order of terms in our development of Equation 33, we can match corresponding terms between Equation 33 and Equation 27. The results of this match can be used to diagram the transform operation T_1 as shown in Figure 66. It is quite apparent in Figure 66 that each term in the exponential expression


Af (x)
T_1
M_1
$1/4e^{j\omega_0(x+x_0)}$
T
+
1/2
T
+
$1/4e^{-j\omega_0(x+x_0)}$
T
$ae^{j\omega_0(x+x_0)}$
$\frac{ae}{2}$
$\omega_0 + \frac{\pi}{a}$
ω_0
$\omega_0 - \frac{\pi}{a}$
$\frac{\pi}{a}$
$\omega = 0$
$-\frac{\pi}{a}$
a
$-\omega_0 + \frac{\pi}{a}$
$-\omega_0$
$-\omega_0 - \frac{\pi}{a}$
$-j\omega_0 x_0$
$\frac{ae}{2}$

$$f(x) = \frac{1}{2}\left[1 + \frac{1}{2}\,e^{j\omega_0(x+x_0)} + \frac{1}{2}\,e^{-j\omega_0(x+x_0)}\right]$$

Figure 66—Diagram of transform operation.

does correspond to a particular spectral point. If we note that the constant term +1 can be expressed as $e^{\pm j0}$, we can say that each exponential term $e^{j\omega_n x}$ in the expansion of $f(x)$ produces a spectral point at $\omega = \omega_n$. It should be apparent that this statement applies to the general case, since no restrictions were placed on ω_0 in the above example. Any signal consisting of component frequencies can be expanded similarly as for Equation 27. Each component frequency ω_n will result in exponentials of the form $e^{j\omega_n(x+x_n)}$ and $e^{-j\omega_n(x+x_n)}$, as in Equation 27. These exponential terms produce spectral points at ω ▫ ω_n and $\omega = -\omega_n$, as proved above. Thus each frequency component ω_n produces two spectral points, one at $\omega = \omega_n$ and another at $\omega = -\omega_n$.

The resolution of spectral points is dependent upon the aperture dimension $2a$. It is shown in Figure 65 that the width of a spectral point can be considered to be $2\pi/a$. This width specification means that two adjacent spectral points will begin to overlap when the difference between the corresponding frequencies becomes less than $2\pi/a$. In conclusion it should be noted that the detected output of a spectrum analyzer will be the square of the amplitude curves illustrated in Figure 66.

Optical Filtering

An optical filter system is diagrammed in Figure 67. This optical filter configuration is an extension of the spectrum analyzer shown in Figure 63. The addition of a transform (T_2) and a multiplier (M_2) to the spectrum analyzer forms the optical filter. The output of T_1 is $AF(\omega)$, as shown in Figure 67 (see the preceding subsection, "Optical Spectrum Analyzer"). Instead of detecting this signal as an output from M_1 as shown in Figure 63, a filter function $G(\omega)$ is inserted into M_1 and the output product $AF(\omega)\,G(\omega)$ is applied to the input of the transform T_2. To simplify our notation, we will adopt the shorthand notation $AF_f(\omega)$ to designate $AF(\omega)\,G(\omega)$—the subscript f indicates that the signal $F(\omega)$ has been modified, or filtered. Continuing this notation, we can designate the output of T_2 as $Af_f(x)$, where $f_f(x)$ is the Fourier transform of $F_f(\omega)$. The significance of the subscript f becomes clear if we consider the output amplitude $Af_f(x)$ as the filtered form of the input amplitude $Af(x)$. The detected output of M_2 will be $|Af_f(x)|^2$.

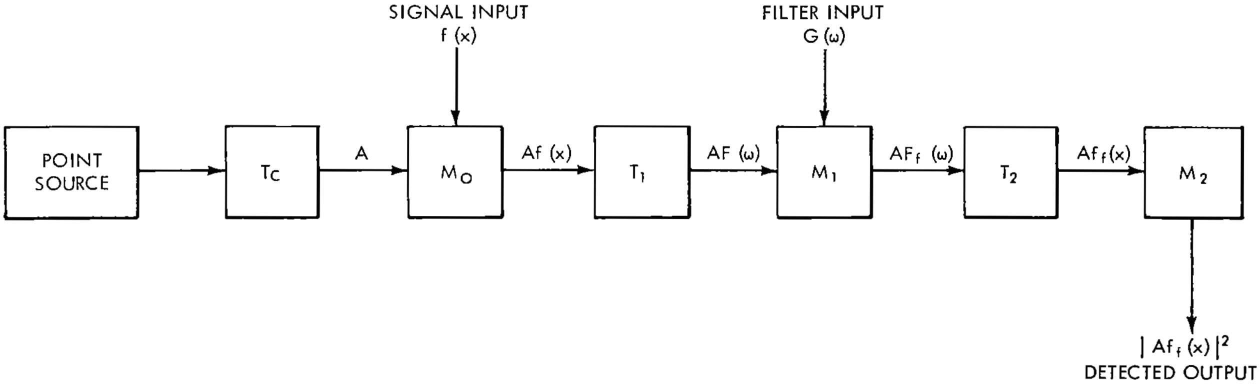


Figure 67—Optical filter system.

The basic operation of an optical filter can be clarified by examining the light amplitude $Af_f(x)$ present at the multiplier M_2. Letting "A" equal 1 (to simplify the discussion), and assuming that there is no filter signal present, i.e. $G(\omega) = 1$, the output of M_1 will be $F_f(\omega) = F(\omega)$, and the output of T_2 will be $f_f(x) = f(x)$. Thus, when there is no filter signal present, the light-amplitude variations at the input of M_2 will represent the signal input $f(x)$. Effectively, when no filter input signal is present, the signal input $f(x)$ is imaged onto the input of the multiplier M_2 (output plane). In this case the detected output of M_2 is $|f(x)|^2$. It is apparent that, if a filter signal $G(\omega)$ is inserted into M_1, the output of M_1 is effectively a modified (or filtered) spectrum of $f(x)$. The usual filter input signal $G(\omega)$ is inserted as a film with a transmission function which has values of 1 or zero depending on the coordinate ω. In effect, the filter input signal $G(\omega)$ blocks (multiplies by zero) any frequency component of $f(x)$ which corresponds to the values of ω for which $G(\omega)$ is equal to zero. Thus the light amplitude $f_f(x)$ at M_2 represents the filtered form of the input signal $f(x)$.

To examine some of the details of optical filter operation, we can again consider the input signal $f(x)$, given as

$$f(x) = \frac{1}{2}\left[1 + \cos\omega_0\left(x + x_0\right)\right] = \frac{1}{2}\left[1 + \frac{1}{2}e^{j\omega_0(x+x_0)} + \frac{1}{2}e^{-j\omega_0(x+x_0)}\right] . \qquad (34)$$

The light amplitude at the output of M_0 will be

$$Af(x) = \frac{A}{2}\left[1 + \cos\omega_0\left(x + x_0\right)\right] = \frac{A}{2}\left[1 + \frac{1}{2}e^{j\omega_0(x+x_0)} + \frac{1}{2}e^{-j\omega_0(x+x_0)}\right] . \qquad (35)$$

The transformed input to M_1 (as shown for Equation 31) will be

$$AF(\omega) = aA\left[\frac{\sin a\omega}{a\omega} + \frac{e^{j\omega_0 x_0}}{2}\,\frac{\sin a\left(\omega_0 - \omega\right)}{a\left(\omega_0 - \omega\right)} + \frac{e^{-j\omega_0 x_0}}{2}\,\frac{\sin a\left(\omega_0 + \omega\right)}{a\left(\omega_0 + \omega\right)}\right] \qquad (36)$$

If we use the approximation indicated by the amplitude curves shown in Figure 66 (neglect all peaks except the maximum), we need consider only the values of the filter function $G(\omega)$ in the neighborhood of each of the spectral points. To abbreviate our notation, we will designate the value of $G(\omega)$ in a region π/a wide as the value at the center point of the region. In other words, when we say that $G(\omega)$ is equal to zero at $\omega = b$, we will imply that $G(\omega)$ is equal to zero for ω in the region between $b - \pi/a$ and $b + \pi/a$. If $G(\omega)$ is equal to zero at $\omega = b$, a spectral point of $AF(\omega)$ that has its peak centered at $\omega = b$ will be blocked (multiplied by zero). Since there is a direct term-for-term correspondence between $AF(\omega)$ and $Af(x)$, blocking a point of $AF(\omega)$ centered at $\omega = b$ eliminates the corresponding term $\left(e^{jbx}\right)$ in $Af(x)$. As an example, consider the case where $G(\omega)$ is equal to zero at $\omega = 0$. The $(\sin a\omega)/a\omega$ term of $AF(\omega)$ (Equation 36) is centered at $\omega = 0$ and is therefore multiplied by zero. Since $(\sin a\omega)/a\omega$ of $AF(\omega)$ is blocked (multiplied by zero), the corresponding term $A/2$ of $Af(x)$ (Equation 35) is eliminated. The filtered transform $AF_f(\omega)$ for this example

(Equation 36 with the $(\sin a\omega)/a\omega$ term eliminated) can be written

$$AF_f(\omega) = \frac{aA}{2}\left[e^{j\omega_0 x_0}\,\frac{\sin a\left(\omega_0 - \omega\right)}{a\left(\omega_0 - \omega\right)} + e^{-j\omega_0 x_0}\,\frac{\sin a\left(\omega_0 + \omega\right)}{a\left(\omega_0 + \omega\right)}\right] . \tag{37}$$

The light amplitude $Af_f(x)$ at the input to M_2 for this example can be written

$$Af_f(x) = \frac{A}{2}\cos\omega_0\left(x + x_0\right) , \tag{38}$$

since the constant term $A/2$ was eliminated by blocking the corresponding term, $(\sin a\omega)/a\omega$, of $AF(\omega)$. The significance of the filter operation can be seen by comparing $AF_f(\omega)$ of Equation 37 with $AF(\omega)$ of Equation 36, and by comparing $Af_f(x)$ of Equation 38 with $Af(x)$ of Equation 35 (or $f(x)$ of Equation 34).

Some of the possible filtering operations which can be performed on the function $f(x)$ considered here are given in Table 6. The results tabulated are found by the same method as discussed

Table 6

Example of Filtering.

$$Af(x) = \frac{A}{2}\left[1 + \cos\omega_0\left(x + x_0\right)\right] = \frac{A}{2}\left[1 + 1/2\,e^{j\omega_0(x+x_0)} + 1/2\,e^{-j\omega_0(x+x_0)}\right]$$

Filter Function $G(\omega)$			Spectral Terms Blocked			Output at M_2	
At $\omega = -\omega_0$	At $\omega = 0$	At $\omega = \omega_0$	Lower Sideband	dc	Upper Sideband	Amplitude	Intensity (Detected)
1	1	1				$\frac{A}{2}\left[1 + \cos\omega_0\left(x + x_0\right)\right]$	$\frac{A^2}{4}\left[1 + \cos\omega_0\left(x + x_0\right)\right]^2$
0	0	0	X	X	X	0	0
0	1	0	X		X	$\frac{A}{2}$	$\frac{A^2}{4}$
1	0	0		X	X	$\frac{A}{4}\,e^{-j\omega_0(x+x_0)}$	$\frac{A^2}{16}$
0	0	1	X	X		$\frac{A}{4}\,e^{j\omega_0(x+x_0)}$	$\frac{A^2}{16}$
1	0	1		X		$\frac{A}{2}\cos\omega_0\left(x + x_0\right)$	$\frac{A^2}{4}\cos^2\omega_0\left(x + x_0\right)$
0	1	1	X			$\frac{A}{2}\left[1 + \frac{1}{2}\,e^{j\omega_0(x+x_0)}\right]$	$\frac{A^2}{4}\left[\frac{5}{4} + \cos\omega_0\left(x + x_0\right)\right]$
1	1	0			X	$\frac{A}{2}\left[1 + \frac{1}{2}\,e^{-j\omega_0(x+x_0)}\right]$	$\frac{A^2}{4}\left[\frac{5}{4} + \cos\omega_0\left(x + x_0\right)\right]$

above. It should be noted that the results given in Table 6 are approximate, since only the maximum peaks have been considered in this development.

It might be enlightening to consider the determination of the detected output, which is given as intensity in Table 6. Consider the case where $G(\omega)$ is equal to zero only at $\omega = -\omega_0$. The output amplitude is given as

$$Af_f(x) = \frac{A}{2}\left[1 + \frac{1}{2}e^{j\omega_0(x+x_0)}\right] . \tag{39}$$

The output intensity is given by the square of the magnitude of $Af_f(x)$ and *not* by the square of $Af_f(x)$. The square of $Af_f(x)$ would be

$$\left[Af_f(x)\right]^2 = \frac{A^2}{4}\left[1 + e^{j\omega_0(x+x_0)} + \frac{1}{4}e^{j2\omega_0(x+x_0)}\right] \tag{40}$$

Since $\left[Af_f(x)\right]^2$ as given by Equation 40 has both magnitude and phase, intensity which we know to be real cannot be the square of $Af_f(x)$. Equation 39 can be rewritten as

$$Af_f(x) = \frac{A}{2}\left[1 + \frac{1}{2}\cos\omega_0\left(x + x_0\right) + j\,\frac{1}{2}\sin\omega_0\left(x + x_0\right)\right] , \tag{41}$$

where we have made use of the fact that $e^{jkx} = \cos kx + j\sin kx$. Now Equation 41 is an equation of the form

$$A(x) = r + jX , \tag{42}$$

where

$A(x)$ = amplitude dependent on x [$Af_f(x)$ in Equation 41],

R = real component [$A/2 + A/4\cos\omega_0\left(x + x_0\right)$ in Equation 41],

X = imaging component [$A/4\sin\omega_0\left(x + x_0\right)$ in Equation 41].

The magnitude of $A(x)$ as given by Equation 42 is given by

$$|A(x)| = \sqrt{R^2 + X^2} , \tag{43}$$

where $|A(x)|$ means "magnitude of $A(x)$."

Now, if we define intensity, or I, as the square of the magnitude of the complex light amplitude, we may write

$$I = |A(x)|^2 = R^2 + X^2 . \tag{44}$$

The intensity given by Equation 44 is real, since R and X are real, and therefore agrees with physical measurements. The right side of Equation 44 can be factored, giving

$$R^2 + X^2 = (R + j X)(R - j X) . \tag{45}$$

The first term $(R + jX)$ is recognized as our amplitude $A(x)$ as given by Equation 42. The second term differs only in the sign of the imaginary term, and is defined as the complex conjugate of $A(x)$. The complex conjugate is usually indicated by a star, i.e. $A^*(x)$ is the complex conjugate of $A(x)$. Equation 44 may now be given as

$$I = |A(x)|^2 = A(x) A^*(x) . \tag{46}$$

Returning to our example (Equation 39), the intensity is given as

$$I = |Af_f(x)|^2 = [Af_f(x)] [Af_f(x)]^* . \tag{47}$$

When a function is written as a complex exponential, the complex conjugate is formed by changing the sign of the imaginary exponent. Using the expression for $Af_f(x)$ given in Equation 39, Equation 47 can be written

$$I = |Af_f(x)|^2 = \frac{A}{2}\left[1 + \frac{1}{2} e^{j\omega_0(x+x_0)}\right] \frac{A}{2}\left[1 + \frac{1}{2} e^{-j\omega_0(x+x_0)}\right] . \tag{48}$$

Carrying out the multiplication, we obtain

$$|Af_f(x)|^2 = \frac{A^2}{4}\left[1 + \frac{1}{2} e^{-j\omega_0(x+x_0)} + \frac{1}{2} e^{j\omega_0(x+x_0)} + \frac{1}{4}\right] . \tag{49}$$

Recalling that

$$\cos \omega_0 (x + x_0) = \frac{1}{2} e^{-j\omega_0(x+x_0)} + \frac{1}{2} e^{j\omega_0(x+x_0)} ,$$

we can rewrite Equation 49 as

$$|Af_f(x)|^2 = \frac{A^2}{4}\left[\frac{5}{4} + \cos \omega_0 (x + x_0)\right] . \tag{50}$$

Equation 50 is the expression for the detected output (intensity) for this example. Note that the use of the complex conjugate to determine the square of the absolute value results in an intensity expression (Equation 44) which is real (no phase).

The usual application of filtering is to eliminate undesired frequencies. It is apparent from examination of the output amplitude column of Table 6 that all traces of a given frequency are eliminated only when *both* the upper and lower spectral points (sidebands) are blocked. That is, in order to eliminate a frequency ω_0, the filter function $G(\omega)$ must be equal to zero at both $\omega = \omega_0$ and $\omega = -\omega_0$. Additional filtering operations can be described by considering filter functions $G(\omega)$ with values other than just 1 or zero, and by considering the effects of blocking only a fraction of the region covered by a spectral point.

Optical Correlator

An optical correlator is shown in Figure 68. This optical correlator configuration is obtained by adding a transform (T_3) and multiplier (M_3) to the optical filter arrangement of Figure 67. The output of T_2 is $Af_{1f}(x)$, which corresponds to $Af_f(x)$ shown in Figure 67. Instead of detection of output from M_2, as in an optical filter, there is a reference signal $f_2(x)$ inserted into M_2, and the output product $Af_{1f}(x)f_2(x)$ is applied to the input of the transform T_3. The transform T_3 produces the spectrum of the product $Af_{1f}(x)f_2(x)$ in the plane represented by M_3. It was shown in the previous discussion of optical correlation ("Optical Signal Correlator") that the amplitude of this spectrum at $\omega = 0$ is the value of the correlation function of $f_2(x)$ and $f_{1f}(x)$. The selection of the portion of the spectrum at $\omega = 0$ is performed by a low-pass filter input $L(\omega)$ into M_3. This filter input signal $L(\omega)$ has a value equal to 1 in a very small region about $\omega = 0$, and a value equal to zero elsewhere. The light amplitude at M_3 is given by the product of $L(\omega)$ and the transform of $Af_{1f}(x)f_2(x)$. This product is the correlation function $\phi_{2\,1}$ when $f_1(x)$ is the displaced signal. The detected output at M_3 will be $|\phi_{2\,1}|^2$.

To examine some of the details of optical correlation, we will consider an input signal given as

$$f_1(x) = \frac{1}{4}\left[2 + \cos\omega_1\left(x + x_1\right) + \cos\omega_2\left(x + x_2\right)\right], \qquad (51)$$

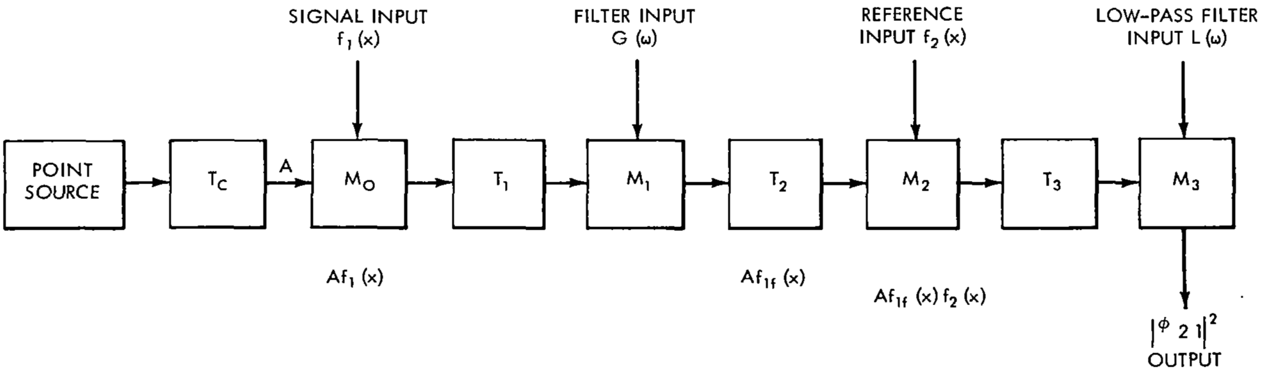


Figure 68—Optical correlator.

where ω_1 is not equal to ω_2 (i.e. $f_1(x)$ has two frequency components). Recalling that a cosine function can be expressed in terms of exponentials, we can write Equation 51 as

$$f_1(x) = \frac{1}{4}\left[2 + \frac{1}{2} e^{j\omega_1(x+x_1)} + \frac{1}{2} e^{-j\omega_1(x+x_1)} + \frac{1}{2} e^{j\omega_2(x+x_2)} + \frac{1}{2} e^{-j\omega_2(x+x_2)}\right] . \qquad (52)$$

Usually only the variable terms of $f_1(x)$ are of interest, and therefore a dc stop is inserted into M_1. A dc stop corresponds to a filter function $G(\omega)$ which is equal to zero at $\omega = 0$. As shown in the previous subsection "Optical Filtering," this type of filter function eliminates the constant term of the input signal. Multiplying Equation 52 by the light amplitude A and eliminating the constant term A/2, we can write Equation 52 as $Af_{1f}(x)$, which is the output of T_2, or

$$Af_{1f}(x) = \frac{A}{4}\left[\frac{1}{2} e^{j\omega_1(x+x_1)} + \frac{1}{2} e^{-j\omega_1(x+x_1)} + \frac{1}{2} e^{j\omega_2(x+x_2)} + \frac{1}{2} e^{-j\omega_2(x+x_2)}\right] . \qquad (53)$$

This signal $Af_{1f}(x)$ is applied to the input of M_2 and is multiplied by the reference signal $f_2(x)$ also inserted into M_2. Assuming $f_2(x)$ is given by

$$\begin{aligned} f_2(x) &= \frac{1}{4}\left(2 + \cos\omega_c x + \cos\omega_d x\right) \\ &= \frac{1}{4}\left[2 + \frac{1}{2} e^{j\omega_c x} + \frac{1}{2} e^{-j\omega_c x} + \frac{1}{2} e^{j\omega_d x} + \frac{1}{2} e^{-j\omega_d x}\right] , \qquad (54) \end{aligned}$$

where ω_c is not equal to ω_d, the product at the output of M_2 is given by the product of Equations 53 and 54 and can be written as

$$\begin{aligned} Af_{1f}(x) f_2(x) &= \frac{A}{64} e^{j\omega_1(x+x_1)}\left[4 + e^{j\omega_c x} + e^{-j\omega_c x} + e^{j\omega_d x} + e^{-j\omega_d x}\right] \\ &+ \frac{A}{64} e^{-j\omega_1(x+x_1)}\left[4 + e^{j\omega_c x} + e^{-j\omega_c x} + e^{j\omega_d x} + e^{-j\omega_d x}\right] \\ &+ \frac{A}{64} e^{j\omega_2(x+x_2)}\left[4 + e^{j\omega_c x} + e^{-j\omega_c x} + e^{j\omega_d x} + e^{-j\omega_d x}\right] , \\ &+ \frac{A}{64} e^{-j\omega_2(x+x_2)}\left[4 + e^{j\omega_c x} + e^{-j\omega_c x} + e^{j\omega_d x} + e^{-j\omega_d x}\right] . \qquad (55) \end{aligned}$$

We can let A equal 16 and multiply the brackets by the part of the exponential dependent on x,

$$
\begin{aligned}
A f_{1f}(x)\, f_2(x) = {} & \frac{e^{j\omega_1 x_1}}{4}\left[4e^{j\omega_1 x} + e^{j(\omega_1+\omega_c)x} + e^{j(\omega_1-\omega_c)x} + e^{j(\omega_1+\omega_d)x} + e^{j(\omega_1-\omega_d)x}\right] \\
& + \frac{e^{-j\omega_1 x_1}}{4}\left[4e^{-j\omega_1 x} + e^{-j(\omega_1-\omega_c)x} + e^{-j(\omega_1+\omega_c)x} + e^{-j(\omega_1+\omega_d)x} + e^{-j(\omega_1+\omega_d)x}\right] \\
& + \frac{e^{j\omega_2 x_2}}{4}\left[4e^{j\omega_2 x} + e^{j(\omega_2+\omega_c)x} + e^{j(\omega_1-\omega_c)x} + e^{j(\omega_2+\omega_d)x} + e^{j(\omega_2-\omega_d)x}\right] \\
& + \frac{e^{-j\omega_2 x_2}}{4}\left[4e^{-j\omega_2 x} + e^{-j(\omega_2-\omega_c)x} + e^{-j(\omega_2+\omega_c)x} + e^{-j(\omega_2-\omega_d)x} + e^{-j(\omega_2+\omega_d)x}\right] .
\end{aligned}
\tag{56}
$$

As shown in Figure 68, $Af_{1f}(x)\, f_2(x)$ given as Equation 56 is transformed by T_3 and then multiplied by $L(\omega)$ in M_3. Since Equation 56 is rather cumbersome to carry through completely, we will make use of the filter function $L(\omega)$ to simplify the equation. It was stated above that $L(\omega)$ was zero for all values of ω other than zero, i.e. the operation of filter $L(\omega)$ is to effectively block all spectral points not centered at $\omega = 0$. It was pointed out that blocking a spectral point will effectively eliminate the corresponding exponential in the original signal. Since Equation 56 is our signal, we can eliminate any exponential which does not produce a spectral point at $\omega = 0$. From our discussion of spectrum analyzers ("Optical Spectrum Analyzer") we found that each exponential term e^{jkx} produces a spectral point at

$$\left(\frac{\sin a(k-\omega)}{a(k-\omega)}\right) .$$

We can therefore eliminate all exponentials for which the term corresponding to k is not zero. For example, in the first bracket of Equation 56 the terms corresponding to k are ω_1, $(\omega_1+\omega_c)$, $(\omega_1-\omega_c)$, $(\omega_1+\omega_d)$, and $(\omega_1-\omega_d)$. The first, second, and fourth are definitely not equal to zero and can be eliminated. Since ω_c and ω_d are not equal, ω_1 cannot be equal to both and we can assume ω_1 is not equal to ω_d. This implies that $(\omega_1-\omega_d)$ is not zero and it can also be eliminated. Thus the only exponential remaining is $e^{j(\omega_1-\omega_c)x}$, which may or may not be eliminated depending on whether ω_1 is equal to ω_c. Applying the same elimination technique to the second bracket, we eliminate all terms except $e^{-j(\omega_1-\omega_c)x}$ by the same arguments. Since ω_2 cannot be equal to both ω_c and ω_d (ω_c and ω_d are unequal), we can assume ω_2 is not equal to ω_c. Proceeding thus with the same argument as for the first and second brackets, the third and fourth brackets can be reduced to $e^{j(\omega_2-\omega_d)x}$ and $e^{-j(\omega_2-\omega_d)x}$ respectively. Equation 56 can now be written

$$A f_{1f}(x)\, f_2(x) = \frac{e^{j\omega_1 x_1}}{4}\left[e^{j(\omega_1-\omega_c)x}\right] + \frac{e^{-j\omega_1 x_1}}{4}\left[e^{-j(\omega_1-\omega_c)x}\right] + \frac{e^{j\omega_2 x_2}}{4}\left[e^{j(\omega_2-\omega_d)x}\right] + \frac{e^{-j\omega_2 x_2}}{4}\left[e^{-j(\omega_2-\omega_d)x}\right] . \tag{57}$$

The location of the spectral points produced by the bracketed terms in Equation 57 depends on the values of ω_1 and ω_2. If ω_1 is not equal to ω_c, the first two terms produce spectral points at $\omega = \omega_1 - \omega_c$ and $\omega = -(\omega_1 - \omega_c)$ respectively. Since these points are not at $\omega = 0$, they are eliminated by the action of $L(\omega)$ and will contribute nothing to the correlator output. Likewise, if ω_2 is not equal to ω_d, the last two terms are eliminated by $L(\omega)$ and contribute nothing to the correlator output. Thus, if ω_1 is not equal to ω_c and ω_2 is not equal to ω_d, all terms in Equation 57 are eliminated and the output of the correlator will be zero (i.e. $\phi_{2\,1} = 0$).

If ω_1 is equal to ω_c, the first two brackets in Equation 57 produce a spectral point at $\omega = 0$, since $\omega_1 - \omega_c = -(\omega_1 - \omega_c) = 0$. Likewise, if ω_2 is equal to ω_d, the last two brackets in Equation 57 produce a spectral point at $\omega = 0$, since $\omega_2 - \omega_d = -(\omega_2 - \omega_d) = 0$. Thus, for the case where $\omega_1 = \omega_0$ and $\omega_2 = \omega_d$, Equation 57 can be written

$$Af_{1f}(x)\, f_2(x) = \frac{e^{j\omega_c x_1}}{4} + \frac{e^{-j\omega_c x_1}}{4} + \frac{e^{j\omega_d x_2}}{4} + \frac{e^{-j\omega_d x_2}}{4} \,. \tag{58}$$

Since the coordinate x does not appear in any of the terms on the right side of Equation 58, the transform T produced at the output of T_3 will be the transform of a constant. Since the transform of a constant C is $2a\,c\left[(\sin a\omega)/a\omega\right]$, where $2a$ is the length of the aperture, the transform T of $Af_{1f}(x)\, f_2(x)$ can be written directly from Equation 58 as

$$T = \left[\frac{e^{j\omega_c x_1}}{2} + \frac{e^{-j\omega_c x_1}}{2} + \frac{e^{j\omega_d x_2}}{2} + \frac{e^{-j\omega_d x_2}}{2}\right] \frac{a \sin a\omega}{a\omega} \,, \tag{59}$$

where the terms inside the bracket can be considered constants. This transform T is applied to the input of M_3 and is multiplied by $L(\omega)$, which has the value 1 in a very small region about $\omega = 0$. The product at M_3 can be written

$$TL(\omega) = \left(\cos \omega_c x_1 + \cos \omega_d x_2\right) \frac{a \sin a\omega}{a\omega} \,, \tag{60}$$

where the identity

$$\left[\cos \theta = \frac{e^{j\theta}}{2} + \frac{e^{-j\theta}}{2}\right]$$

is used to substitute cosines for the exponentials in Equation 59. Since $L(\omega)$ equals 1 only in a very small region about $\omega = 0$, the ω appearing in Equation 60 is restricted to very small values and $(\sin a\omega)/a\omega$ can be approximated as being equal to 1 ($\sin a\omega \approx a\omega$ for small $a\omega$). Applying this approximation to Equation 60 and noting that the product $TL(\omega)$ is the correlation function $\phi_{2\,1}$, we can write

$$\phi_{2\,1} = TL(\omega) = a\left(\cos \omega_c x_1 + \cos \omega_d x_2\right) \,. \tag{61}$$

The detected output $|\phi_{2\,1}|^2$ for this case will be

$$|\phi_{2\,1}|^2 = a^2 \left(\cos \omega_c x_1 + \cos \omega_d x_2\right)^2 . \tag{62}$$

From the results obtained above, we can develop a general rule for determining correlator outputs. It was shown that a frequency component of the signal function which *does not match* a frequency component of the reference function *contributes nothing* to the correlation function (i.e. contribution to $\phi_{2\,1}$ is zero). In addition, it was shown that a frequency component of the signal which *does match* a frequency component of the reference *contributes a cosine term* (signal components originally expressed as cosines) to the correlation function. In other words, the correlation contribution of a signal component ω_n is zero if the frequency is *not* present in the reference and the contribution is $\cos \omega_n x_n$ if the frequency (ω_n) *is* present in the reference. Table 7 lists correlation functions for various combinations of ω_1, ω_2, ω_c, and ω_d. Direct application of the rules just described produces the results given in Table 7. A dash in the "Signal Frequency" column indicates that the particular signal frequency is not equal to a frequency in the reference frequency. The dash was used in such cases because any frequency not in the reference would produce the same (zero) contribution. Examination of the correlation functions listed in Table 7 shows that, when the reference contains more than one frequency, the correlation function can be used to determine which particular frequency or combination of frequencies is present in the signal function. As a final note, it should be pointed out that in our treatment we have assumed all frequency components to be of the same amplitude. In practice, this is not the case, and an amplitude factor (not necessarily the same) will appear in front of each frequency term.

Table 7

Sample Correlation Functions.

Reference Frequency	Signal Frequency		Correlation Functions	
	ω_1	ω_2	Amplitude - $\phi_{2\,1}$	Intensity - $\|\phi_{2\,1}\|^2$ (Detected)
ω_c and ω_d	-	-	0	0
ω_c and ω_d	ω_c	-	$a \cos \omega_c x_1$	$a^2 \cos^2 \omega_c x_1$
ω_c and ω_d	-	ω_d	$a \cos \omega_d x_2$	$a^2 \cos^2 \omega_d x_2$
ω_c and ω_d	ω_c	ω_d	$a \left(\cos \omega_c x_1 + \cos \omega_d x_2\right)$	$a^2 \left(\cos \omega_c x_1 + \cos \omega_d x_2\right)^2$
ω_c	-	-	0	0
ω_c	ω_c	-	$a \cos \omega_c x_1$	$a^2 \cos^2 \omega_c x_1$
ω_d	-	-	0	0
ω_d	-	ω_d	$a \cos \omega_d x_2$	$a^2 \cos^2 \omega_d x_2$

Single-Sideband Correlation

Interesting correlation results are obtained when we consider blocking (with an appropriate $G(\omega)$ in the multiplier M_1) one of the two spectral points corresponding to a signal frequency (in addition to the dc point). If we call the up and down spectral points the upper and lower sidebands, the case we are about to consider can be called single-sideband correlation.

We will consider only the development of the case for upper-sideband correlation. The lower-sideband case would follow the same development except for a change in the sign of the exponents. Single-sideband correlation is produced by inserting a special filter function $G(\omega)$ into M_1 (shown in Figure 68). For upper-sideband correlation, $G(\omega)$ would have a value of 1 for all values of ω greater than zero and a value of zero for ω equal to or less than zero. In effect, this filter function $G(\omega)$ blocks the dc spectral component plus all spectral terms on one side of the dc point. We will use the same signal function as in our development for (double-sideband) correlation in the preceding subsection. That is, signal function $f_1(x)$ is given as

$$f_1(x) = \frac{1}{4}\left[2 + \cos\omega_1\left(x + x_1\right) + \cos\omega_2\left(x + x_2\right)\right] . \tag{63}$$

In the previous discussion we used a $G(\omega)$ which was zero only at $\omega = 0$ (dc stop) and obtained a filtered signal $Af_{1f}(x)$, the output of T_2, which was given by Equation 53 as

$$Af_{1f}(x) = \frac{A}{4}\left[\frac{1}{2}e^{j\omega_1(x+x_1)} + \frac{1}{2}e^{-j\omega_1(x+x_1)} + \frac{1}{2}e^{j\omega_2(x+x_2)} + \frac{1}{2}e^{-j\omega_2(x+x_2)}\right] . \tag{53}$$

In our present example, $G(\omega)$ is zero at ω □ 0 and also at all negative values of ω. The second and fourth terms in the bracket of Equation 53 produce spectral points at $\omega = -\omega_1$ and $\omega = -\omega_2$. These points are blocked by the filter $G(\omega)$, and therefore the second and fourth terms in Equation 53 can be eliminated and $Af_{1f}(x)$ for single-sideband operation can be written

$$Af_{1f}(x) = \frac{A}{4}\left[\frac{1}{2}e^{j\omega_1(x+x_1)} + \frac{1}{2}e^{j\omega_2(x+x_2)}\right] . \tag{64}$$

We will use the reference signal $f_2(x)$, which was given in Equation 54 as

$$f_2(x) = \frac{1}{4}\left[2 + \frac{1}{2}e^{j\omega_c x} + \frac{1}{2}e^{-j\omega_c x} + \frac{1}{2}e^{j\omega_d x} + \frac{1}{2}e^{-j\omega_d x}\right] . \tag{54}$$

The product at the output of M_2 for single-sideband correlation is given by the product of Equations 54 and 64 as

$$Af_{1f}(x)\,f_2(x) = \frac{A}{64}e^{j\omega_1(x+x_1)}\left[4 + e^{j\omega_c x} + e^{-j\omega_c x} + e^{j\omega_d x} + e^{-j\omega_d x}\right] + \frac{A}{64}e^{j\omega_2(x+x_2)}\left[4 + e^{j\omega_c x} + e^{-j\omega_c x} + e^{j\omega_d x} + e^{-j\omega_d x}\right] . \tag{65}$$

If we compare Equation 65 to Equation 55, we can note that blocking a spectral term eliminates all terms dependent upon the exponential corresponding to the particular spectral term. For example, the second and fourth brackets in Equation 55 depend on $e^{-j\omega_1(x+x_1)}$ and $e^{-j\omega_2(x+x_2)}$ respectively. Since these exponentials are eliminated by $G(\omega)$ (corresponding spectral terms are blocked), the second and fourth brackets of Equation 55 are not present in Equation 65. To simplify our development from this point, we will make use of this elimination principle directly on the complete expression derived for the normal (double-sideband) correlation. If we apply the identity $\cos\theta = 1/2\, e^{j\theta} + 1/2\, e^{-j\theta}$, we can write the correlation function $\theta_{2\,1}$ of Equation 61 as

$$\phi_{2\,1} = a\left[\frac{1}{2}\, e^{j\omega_c x_1} + \frac{1}{2}\, e^{-j\omega_c x_1} + \frac{1}{2}\, e^{j\omega_d x_2} + e^{-j\omega_d x_2}\right] \qquad (66)$$

when $\omega_1 = \omega_c$ and $\omega_2 = \omega_d$. In the development of Equation 61, each of the terms in Equation 66 corresponds to a spectral term of $f_1(x)$. As pointed out in the comparison between Equations 55 and 65, $G(\omega)$ equal to zero for negative values of ω eliminates all terms dependent on negative exponential terms in $f_1(x)$. Thus the second and fourth terms in the bracket of Equation 66 can be eliminated, and $\phi_{2\,1}$ is given as

$$\phi_{2\,1} = \frac{a}{2}\left[e^{j\omega_c x_1} + e^{j\omega_d x_2}\right]$$

$$\text{when } \omega_1 = \omega_c \text{ and } \omega_2 = \omega_d\ ;$$

$$\phi_{2\,1} = 0 \text{ when } \omega_1 \neq \omega_c \text{ and } \omega_2 \neq \omega_d\ . \qquad (67)$$

The two cases given in Equation 67 are represented by the light amplitude at the correlator output (output of M_3). The first case follows from the elimination of terms in Equation 66 by the action of the filter $G(\omega)$. The second case corresponds directly to the results obtained for normal (double-sideband) correlation (see first line of entries in Table 7). The zero correlation results apply to both single-sideband and double-sideband correlation, since the development considered the elimination of individual exponential terms.

The basic difference between single-sideband correlation and the full double-sideband correlation of the preceding subsection can be seen by comparing Equations 66 and 67. In the single-sideband correlation given by Equation 67, we see that each frequency which appears in both the signal and the reference contributes a single exponential to the correlation function. In the double-sideband correlation given by Equation 66, a frequency which appears in both the signal and the reference contributes two exponentials (which can be combined into a cosine term) to the correlation function (see Table 7).

For a more complete comparison the (double-sideband) correlation functions which were listed in Table 7 are given for single-sideband correlation in Table 8 (upper sideband) and Table 9 (lower sideband). A comparison of Tables 8 and 9 shows that the single-sideband correlation function has the same form for upper or lower sideband operation. Opposite signs appear in the exponents of

Table 8

Single (Upper) Sideband Correlation.

Reference Frequencies	Signal Frequencies		Correlation Functions	
	ω_1	ω_2	Amplitude - $\phi_{2\,1}$	Intensity - $\lvert\phi_{2\,1}\rvert^2$ (Detected)
ω_c and ω_d	-	-	0	0
ω_c and ω_d	ω_c	-	$\frac{a}{2}\, e^{j\omega_c x_1}$	$\frac{a^2}{4}$
ω_c and ω_d	-	ω_d	$\frac{a}{2}\, e^{j\omega_d x_2}$	$\frac{a^2}{4}$
ω_c and ω_d	ω_c	ω_d	$\frac{a}{2}\left(e^{j\omega_c x_1} + e^{j\omega_d x_2}\right)$	$\frac{a^2}{2}\left[1 + \cos\left(\omega_c x_1 - \omega_d x_2\right)\right]$
ω_c	-	-	0	0
ω_c	ω_c	-	$\frac{a}{2}\, e^{j\omega_c x_1}$	$\frac{a^2}{4}$
ω_d	-	-	0	0
ω_d	-	ω_d	$\frac{a}{2}\, e^{j\omega_d x_2}$	$\frac{a^2}{4}$

the correlation functions $\phi_{2\,1}$ listed in the amplitude column; a "+" sign appears for upper sideband operation and a "-" sign appears for lower sideband operation. Since this sign difference is eliminated in determining the intensity (square of the absolute value of the amplitude), the detected output is exactly the same for either upper or lower sideband correlation.

The differences between double-sideband correlation (Table 7) and single-sideband correlation (Tables 8 and 9) can be seen by direct comparison. The most significant difference appears in the detected output for the case of one signal frequency matching one of two reference frequencies. In single-sideband correlation, a detected value of $a^2/4$ is obtained for a match of either frequency, That is, regardless of which frequency matches, the correlator output will be the constant $a^2/4$. In double-sideband correlation, the detected value depends on the matching frequency. That is, if ω_c is the matching frequency, $a^2 \cos^2 \omega_c x_1$ will be the output; and if ω_d is the matching frequency, $a^2 \cos^2 \omega_d x_2$ will be the output. Since the spacing between peaks of a squared cosine function depends upon the ω of the function, it is possible to distinguish frequencies by the peak spacing of the $a^2 \cos^2 \omega x$ outputs. Thus *single-sideband correlation* indicates that *a match* exists; *double-sideband correlation* not only indicates that a match exists, but also indicates *which match* exists. When several (more than one) signal frequencies match reference frequencies, both correlation methods produce usable results. The results depend on the matching frequencies, but the form of

Table 9

Single (Lower) Sideband Correlation.

Reference Frequencies	Signal Frequencies		Correlation Functions	
	ω_1	ω_2	Amplitude - $\phi_{2\,1}$	Intensity - $\lvert\phi_{2\,1}\rvert^2$ (Detected)
ω_c and ω_d	-	-	0	0
ω_c and ω_d	ω_c	-	$\frac{a}{2} e^{-j\omega_c x_1}$	$\frac{a^2}{4}$
ω_c and ω_d	-	ω_d	$\frac{a}{2} e^{-j\omega_d x_2}$	$\frac{a^2}{4}$
ω_c and ω_d	ω_c	ω_d	$\frac{a}{2}\left(e^{-j\omega_c x_1} + e^{-j\omega_d x_2}\right)$	$\frac{a^2}{2}\left[1 + \cos\left(\omega_c x_1 - \omega_d x_2\right)\right]$
ω_c	-	-	0	0
ω_c	ω_c	-	$\frac{a}{2} e^{-j\omega_c x_1}$	$\frac{a^2}{4}$
ω_d	-	-	0	0
ω_d	-	ω_d	$\frac{a}{2} e^{-j\omega_d x_2}$	$\frac{a^2}{4}$

the detected correlation function $\left(\lvert\phi_{2\,1}\rvert^2\right)$ is different for each method. This can be seen by comparing the results given in Tables 7 and 8 for ω_c and ω_d present in both reference and signal.

Concluding Statement

Holograms can be used to enhance the optical data processing techniques discussed in this document. The systems and techniques described in this document involved only amplitude (real) variable signals. Using holograms, data processing techniques can be devised to operate on complex signals which vary in both amplitude and phase. A thorough discussion of these systems would require mathematical descriptions a little beyond the intended scope of this document. For the same reason, the mathematical treatment of hologram construction and viewing was not included. The basic principles set forth in this document provide a good basis for the study of these more complex techniques. For those interested in further study, there are references and texts which begin with mathematical descriptions of hologram construction and proceed to develop the more complex aspects of optical data processing techniques.

It has probably occurred to some readers that the section on fundamentals of color was not referred to at any point in the discussion of optical data processing and holography. Color was included to more or less complete the review of basic optics and to provide some introduction to the problems involved in detection of the wavelengths of light with the human eye. These principles are important in the development and examination of photographic film storage systems described elsewhere in this document. The applications of color in optical data processing were not included, since the mathematical descriptions are beyond the intended scope of this document.

GSFC has acquired optical data processing equipment. When operational, this equipment will have the capability of performing the following functions:

Analog Computations

Spectrum Analysis

Cross-Correlation

Filtering

Matched Filtering

Convolution

Pattern or Signal (with noise) Recognition

Autocorrelation

Initial efforts have been directed towards performing these functions in non-real time with emphasis on standardizing the operation and determining the limitations of the equipment. Further work will be directed towards semi-automatic operations (in non-real time) with a final goal of real-time automatic operation.

Goddard Space Flight Center
National Aeronautics and Space Administration
Greenbelt, Maryland, March 28, 1969
125-23-02-05-51

NASA-Langley, 1970 —— 23

"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."

—NATIONAL AERONAUTICS AND SPACE ACT OF 1958

NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Scientific and technical information generated under a NASA contract or grant and considered an important contribution to existing knowledge.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

TECHNOLOGY UTILIZATION PUBLICATIONS: Information on technology used by NASA that may be of particular interest in commercial and other non-aerospace applications. Publications include Tech Briefs, Technology Utilization Reports and Technology Surveys.

Details on the availability of these publications may be obtained from:

SCIENTIFIC AND TECHNICAL INFORMATION DIVISION
NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Washington, D.C. 20546