Computer-generated Binary Holograms*

Abstract: Holograms synthesized by computer are used for constructing optical wavefronts from numerically specified objects. Elimination of the need for a physical object has made new applications possible, for example, three-dimensional computer output displays, synthetic prototypes for interferometric testing, and filters for various optical data processing operations. Our computer holograms differ from a normal hologram in that the transmittance is binary, yet they are able to construct general wavefronts and images efficiently and have several practical advantages over holograms with a continuous range of transmittance. Recent improvements that simplify the production of binary holograms and improve their performance are described and experimental work showing reconstruction of two- and three-dimensional images is presented.

Introduction

In 1965, encouraged by the spatial filtering work of Kozma and Kelly,1 we started to synthesize holograms with the aid of a computer.2 Since then we have studied the problem in more detail3 and have applied the computer-generated holograms to various spatial filtering experiments.2,4–6 Now several other groups, using a variety of techniques, are producing holograms by computer.7–19

There is considerable motivation for synthesizing holograms. Included is the fact that the object volume is no longer physically constrained by illumination coherence, vibration, or air turbulence considerations and it is possible to study certain holographic effects by simulation. For example, Leith and Upatnieks19 demonstrated that a ground glass in contact with an object may improve the light efficiency of a hologram, but it also creates a noisy image. Now with the help of a computer it might be possible to simulate an optimal ground glass that increases efficiency without introducing noise.

A more important reason for synthesizing holograms is to create optical wavefronts from objects that do not physically exist. A need to form such a wavefront from a numerically described object occurs whenever the results of a three-dimensional investigation, for example, the analysis of an x-ray diffractogram, must be displayed in three dimensions. Alternatively, a wavefront from a computer-generated hologram might serve as the interferometric prototype for testing a complex optical surface during its manufacture.

Another need for computer holograms arises from spatial filtering experiments. Sometimes, such as when making inverse filters, differential filters, or optical code translators, the filter function can be difficult to produce. Complex division and other mathematical operations performed by optical and photographic methods can be quite tedious compared with the ease with which a computer handles such tasks.4,6

There are other advantages of computer-generated holograms which are associated with the fact that a binary transmittance pattern can be used instead of the grey sinusoidal fringe pattern of ordinary holograms. Here we discuss some of the general features of computer-generated holograms and, in particular, the features of those types that have only a binary transmittance. Recent improvements in the theory of binary computer-generated holograms are presented with experimental demonstrations.

Hologram synthesis by computer

The process of synthesizing a hologram generally consists of four steps. First the propagation of the complex amplitude from the object to the hologram plane is computed. Actually, because of computer limitations, it is possible to compute the amplitude only at a finite number of
sampling points, but this constraint is quite tolerable since the amplitude is a band-limited function as demonstrated previously.

The second step, still within the computer, is to encode the complex amplitude as a real, non-negative function from which the hologram artwork can be generated on a graphic output device. An example of such an encoding scheme is the simulation of the interference fringes caused by interaction of the reference and object beams in conventional holography.

The final steps are to make the artwork and to reduce it to a reasonable size for diffracting light. Cathode-ray tubes, line printers, and mechanical plotters have been used to produce the artwork and the characteristics of these devices have determined to a great extent the differences among the various hologram generation techniques. The final photoreduction step could, of course, be eliminated by using special output devices to write the hologram directly in the desired size.

Basically the function of a computer-generated hologram is to create an optical wavefront from a set of computed data that are a proper sampling of the complex wavefront amplitude. If a very small antenna could be placed at each sampling point and driven at the correct amplitude and phase, the wavefront would be generated provided the sampling were sufficiently frequent. However, construction of such an antenna array, which radiates at optical frequencies using, for example, a proper distribution of thin dielectric films to provide the correct phase, is difficult.

Fortunately Lord Rayleigh was aware of a simpler method of making complex wavefronts. He knew that slight dislocations of some slits of a diffraction grating would give rise to "ghosts" in the diffraction spectra. While the path difference for wavelets from adjacent slits of a perfect grating in the first diffraction order is exactly one wavelength, the path length difference for wavelets from a dislocated slit and its neighbor will be greater or less than one wavelength. The deviation from an integral wavelength, also called a "detour phase," was only a nuisance to Lord Rayleigh, but has since been used by Hauk and Lohmann and by Taylor and his co-workers for representing phase functions. This detour phase also forms the basis for encoding the phase in computer-generated holograms.

There are a variety of forms, however, for encoding the modulus of the complex amplitude. In a manner similar to that of the conventional hologram formed by interference, the amplitude can be represented by the contrast of sinusoidal fringes. It is possible to avoid grey transmittance levels and yet to maintain complete control over the amplitude encoding with binary patterns by representing the fringe contrast by variable width slits, discontinuous slits, or dots of varying spatial density.

Originally we started synthesizing holograms with binary patterns because of the availability of computer-controlled mechanical plotters; however, we have since uncovered several features that favor the binary approach and that have warranted its further development.

Features of binary holograms
The ease and accuracy with which binary patterns can be photographically reduced and reproduced compared with grey patterns are well known and form the basis for the half-tone printing process widely used for pictures in newspapers and magazines. Like half-tone pictures the quality of the binary hologram is quite insensitive to nonlinear photographic effects; thus much less control over the exposure and development is needed during the reduction process. These holograms are, therefore, more suitable for mass production as, for example, read-only memory elements or master holograms in technical publications.

Another advantage of the binary hologram over the grey hologram is that it directs more light to the reconstructed image. In fact, by using emulsion relief effects or diffraction grating ruling techniques, it should be possible to synthesize blazed holograms. These could have very high efficiency as demonstrated recently by Sheridon who made blazed holograms by interference methods.

Using Kogelnik's definition of hologram light efficiency, we can compare the binary and grey hologram types by considering the diffraction efficiency of ideal regular gratings. As holograms these gratings reconstruct single-point images at \( x = x_0 \) and \( y = 0 \) in the output plane of the system shown in Fig. 1. The transmission functions of these gratings have a period \( d = \lambda f/x_0 \) and are illustrated in Fig. 2 for (a) conventional grey, (b) binary,
Figure 2: Amplitude and phase transmittance of (a) grey, (b) binary, (c) bleached grey, (d) bleached binary, and (e) blazed holograms. The parameter $A$ is proportional to the modulation level.

(c) bleached grey, (d) bleached binary, and (e) blazed holograms. The expressions for these functions are given in Table 1 with the term responsible for the point image listed separately. In all cases $A$ is a parameter describing the extent of modulation and it cannot exceed unity for the nonbleached, amplitude-modulated holograms (a) and (b).

The squared modulus of the point image term, which represents the fraction of incident light transmitted to the image, is plotted in Fig. 3 as a function of the modulation $A$. This figure shows that the efficiency of a grey hologram has a maximum of 6.2% whereas that of a binary hologram reaches 10%. The bleached versions of these holograms can attain efficiencies of 34% and 41%, respectively, while the blazed hologram value could reach 100%, at least for a single-point object, comparable to a volume phase hologram.

Actually the light efficiency advantage of binary holograms is even more significant in most practical situations. Typically, when making a grey hologram, the reference wave is two to three times larger in amplitude than the object wave, so that the fringe contrast does not extend beyond the linear region of the transmission-vs.-exposure curve of the photographic plate. The consequence of the strong reference beam is, therefore, to keep the modulation parameter $A$ of Fig. 3 less than about 0.5 for the grey holograms. Since this photographic linearity consideration does not apply to the binary holograms, the modulation parameter $A$ can remain near unity and yield maximum efficiency for these types.

For more complicated objects the efficiencies undoubtedly would be less and for bleached grey holograms the presence of image flare from the phase intermodulation noise must be considered. Nevertheless, the efficiency values obtained from the mathematically tractable point object case should be useful for comparison.

In addition to having higher efficiency, the reconstruction of a binary hologram yields less noise from light scattered by the photographic grain structure. At transmittance values near one there are very few grains to cause scattering, whereas near zero transmittance several layers of silver grains reliably make the emulsion opaque. On the other hand, the transmittance of a conventional grey hologram fluctuates near a value of 0.5 where grain scattering noise is most severe.

To realize the noise and efficiency advantages of binary holograms it is essential that the resolution of the photo-

<table>
<thead>
<tr>
<th>Hologram type</th>
<th>Transmission function</th>
<th>Point-image term</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Grey</td>
<td>$\frac{1}{2} + \frac{A}{2 \pi} \sin \left( \frac{2\pi x'}{d} \right)$</td>
<td>$\frac{A}{4i}$</td>
</tr>
<tr>
<td>(b) Binary</td>
<td>$\frac{1}{2} + \frac{2A}{\pi} \sum_{n=1}^{\infty} \frac{1}{n} \sin \left( \frac{2\pi n x'}{d} \right)$</td>
<td>$\frac{A}{\pi i}$</td>
</tr>
<tr>
<td>(c) Bleached grey</td>
<td>$\exp \left[ i \frac{\pi A}{2} \sin \left( \frac{2\pi x'}{d} \right) \right]$</td>
<td>$J_1 \left( \frac{\pi A}{2} \right)$</td>
</tr>
<tr>
<td>(d) Bleached binary</td>
<td>$\frac{2i}{\pi} \sin \left( \frac{\pi A}{2} \right) \sum_{n=1}^{\infty} \frac{(-1)^n}{n} - \frac{1}{A} \sin \left( \frac{2\pi n x'}{d} \right)$</td>
<td>$\frac{2}{\pi} \sin \left( \frac{\pi A}{2} \right)$</td>
</tr>
<tr>
<td>(e) Blazed binary</td>
<td>$\frac{2}{\pi} \sin \left( \pi A \right) \sum_{n=1}^{\infty} \frac{(-1)^n}{n^2} - \frac{1}{A^2} \left[ A \sin \left( \frac{2\pi n x'}{d} \right) + i \cos \left( \frac{2\pi n x'}{d} \right) \right]$</td>
<td>$i \sin \left( \pi A \right)$</td>
</tr>
</tbody>
</table>

162

B. R. BROWN AND A. W. LOHMANN

IBM J. RES. DEVELOP.
reduction be sufficient to ensure a sharp-edged binary pattern. Specifically, the transition regions of the recorded edges should be an order of magnitude smaller than the typical aperture size. If this is not possible the rounded-off apertures will approximate the sinusoidal fringe structure of a grey hologram and the performance would, therefore, be reduced.

Although the binary hologram requires a higher spatial bandwidth than the grey type, this is usually not as significant for computer-generated holograms as is computational economy. In the sense of the Cauchy sampling theorem, the binary hologram is optimal because an image of \( N \) resolvable points can be reconstructed from a hologram of \( N \) computed apertures. The additional spatial bandwidth required in the recording step is quite analogous to the extra temporal bandwidth used in electronic digital telemetry systems. In both cases bandwidth utilization is traded for the use of binary signals to increase the signal-to-noise ratio over that of a system using analog signals.

With this analogy in mind it is tempting and perhaps instructive to view the binary hologram in the language of communication theory as Leith and Upatnieks have done for the conventional hologram. In this light the binary hologram can be considered to be a spatial equivalent of pulse modulation. The phase encoding method is analogous to pulse position modulation while the amplitude encoding method is similar to pulse width modulation. Figure 4a illustrates a complex function \( \tilde{u} = |\tilde{u}| \exp (i\varphi) \) that modulates the amplitude and phase of a sinusoidal carrier to form the real, non-negative transmittance \( H \) (Fig. 4b) of a conventional hologram. In Fig. 4c the corresponding pulse-modulated hologram is shown in which a binary transmittance \( B \) now represents the complex light amplitude. The Fourier spectra, Figs. 4d, e, and f, show that a binary hologram contains the same spectrum as the conventional hologram plus some higher harmonics. The shaded areas are high-pass and band-pass filters that may be used to eliminate the unwanted portions of the spectra during reconstruction.

If it were desired to transmit holographic information, a simple scan of a binary hologram would produce a pulse-modulated signal for transmission over a telemetry system and, as is well known from pulse modulation theory, considerable noise suppression could be obtained simply by clipping the received signal. Any remaining noise might be harmless because of the well-known insensitivity of holograms to local defects. Alternatively, the set of complex numbers that represent the object wave \( \tilde{u} \) at the sampling points in the hologram plane could be transmitted as a pulse-code-modulated binary signal with the actual hologram being synthesized at the receiver. The advantages of binary signals and holographic properties would be maintained, but now the transmission reliability could be improved further by incorporating an error detection and correction scheme into the pulse code.
Recent improvements in the theory

Previous work on binary holograms presented by this laboratory demonstrated an intuitive method for representing a wavefront by an array of binary apertures. The approximations involved in this method were discussed in a more recent paper along with a rigorous solution based on an iteration procedure.

Recently we have modified our method for making binary holograms for two reasons: (1) a certain phase ambiguity was not handled correctly in the previous work and this resulted in what we can refer to as the "gap and overlap problem" and (2) the computer time required for the iteration procedure was rather long relative to the value of the improvement obtained by it. We have now devised a new method to compensate for the approximations that, although not as rigorous as the iteration method, is simpler and is experimentally quite successful.

First, for a review of the binary hologram structure, a typical hologram is reproduced in Fig. 5. Each aperture is positioned a distance \( P \) from a sampling point according to the phase \( \varphi \) of the computed wavefront. The height \( W \) of the aperture is proportional to the modulus of the amplitude. Figure 6 shows the format and notation in detail for this hologram. Other rectangular aperture formats have been used and certainly non-rectangular ones are possible; however, the variable height rectangle is convenient for the mechanical plotter to draw and is the one we use most often.

The top and bottom rows of apertures in Fig. 6 depict the gap and overlap fault, which had been observed previously in more complicated holograms. To understand this problem we studied holograms of single-point objects because the correct holograms for this case are known to be regular gratings and any deviation could be traced easily. The mathematical oversight which was responsible for the gaps and overlaps was found to have occurred [in Eq. (25c) of Ref. 3] when we concluded from

\[ \exp(i\varphi_{nm}) = \exp(2\pi iP_{nm}) \]  

that

\[ P_{nm} = \varphi_{nm}/2\pi \]  

and ignored all the other possible solutions,

\[ P_{nm} = (\varphi_{nm} + 2\pi L_{nm})/2\pi, \]

where \( L_{nm} \) is any integer.

To eliminate this ambiguity we introduced a rule that the spacing between adjacent apertures should be as near as possible to one sampling interval. In other words \( L_{nm} \) was chosen so that

\[ |P_{n+1,m} - P_{nm}| \leq \frac{1}{2}. \]

Although this rule eliminated the gap and overlap problem, it did not compensate for a distortion produced by the simple algorithm. The distortion, which became evident in the observation of single-point holograms, compressed...
the image scale close to the zero order, but expanded it at the outer portions of the field. Fortunately, because the hologram binarization process is nonlinear, superposition of point-object results does not describe the general case; no distortion has been observed in our previous experimental work with extended objects.

Since image distortions in general are due to phase errors in the Fraunhofer plane, and because we implement the phase with the aperture positions \( P_{nm} \), it is evident that the original method, Eq. (2), needs to be modified. Although the iterative procedure\(^6\) did eliminate the distortion, a simpler method is to position the aperture according to the actual phase at the aperture itself rather than according to the phase at the nearby sampling point. Thus in one dimension the apertures are located so that

\[
2\pi P_n = \varphi((P_n + n) \delta x') + 2\pi L_n. \tag{5}
\]

This equation means that if the hologram is illuminated with a tilted plane wave,

\[
R(x') = \exp(2\pi i x' / \delta x') \equiv \exp(i\psi), \tag{6}
\]

the apertures are placed where the illuminating wave \( R = \exp(i\psi) \) equals the phase factor \( \exp(i\varphi) \) of the wavefront \( \tilde{u} = |\tilde{u}| \exp(i\varphi) \) to be reconstructed:

\[
\exp[i\psi(P_n \delta x')] = \exp[i\varphi(P_n \delta x')]. \tag{7}
\]

The essence of Eq. (7) is illustrated in Fig. 7b, while the old method according to Eq. (8) is demonstrated in Fig. 7a for comparison:

\[
\exp[i\psi(P_n \delta x')] = \exp(i\varphi_n). \tag{8}
\]

The drawback of the new method is that the phase \( \varphi(x') \) must now be known continuously over the hologram, not just at the sampling points. To compute the Fourier transform more often could become quite costly for large holograms, even using the Cooley-Tukey algorithm;\(^6\) however, a simple interpolation scheme has proved quite satisfactory.

In principle, the continuous Fourier transform of a finite object can be constructed exactly from an infinite number of discrete samples according to the Cauchy sampling theorem. It states for one dimension that

\[
\tilde{u}(x') = \sum_{n} u(n\lambda / \Delta x) \text{sinc}(x' \Delta x / \lambda) - n. \tag{9}
\]

Here the \( u(n\lambda / \Delta x) \) are the sampled values of the Fourier transform of an object distribution \( u(x) \) which is zero outside the interval \( |x| \leq \frac{1}{2} \Delta x \) and \( \text{sinc}(z) = \sin(\pi z) / \pi z \).

The evaluation of even a truncated sum of (9) can be very time-consuming; therefore we used a cubic polynomial to interpolate the phase \( \varphi(x') \) between the adjusted phases at the sampling points \( \varphi_n + 2\pi L_n \). The interpolation was based on a cubic form of Newton’s formula\(^8\) determined by four sampling points in the horizontal row surrounding the \( n \)th aperture location. The justification for using only horizontally neighboring points for interpolation is based on the former iteration corrections [Eqs. (23b') and (23c') of Ref. 3], the \( y' \) correction being an order of magnitude less significant than the \( x' \) correction. This is also reasonable intuitively because the corrections compensate for the apertures being shifted away from the sampling point only in the \( x' \) direction.

The actual position \( P_{nm} \) in two dimensions of the \( n \)th aperture of each row was obtained by solving Eq. (5) numerically by Newton’s method using the interpolated values for the adjusted phase on the right hand side of the equation. After locating the aperture position \( P_{nm} \), the aperture height \( W_{nm} \) was then made proportional to the interpolated modulus \( |\tilde{u}| \) of the complex amplitude at the aperture position.

**Experimental results**

The binary holograms used to reconstruct the images of this section have over 16,000 apertures in a 128 \( \times \) 128 array and are similar to Fig. 5. They were computed with an IBM System/360 Model 50 computer in four to eight minutes depending on the object complexity. The 70-cm-square artwork was plotted in three strips to accommodate the 25-cm width of the CalComp 565 incremental plotter. Plotting time was typically 40 minutes. After assembling the strips, a 180 \( \times \) photoreduction onto Kodak 649F film was made in one step using a Leitz Summicron-R 50-mm camera objective. The resulting holograms are 4 mm square with an average aperture spacing of 0.03 mm.
An example of the complex structure that can be reconstructed from such a hologram is shown in Fig. 8. The object was designed on a grid $112 \times 112$ points that could be white, grey, or black. The finest bar patterns indicate the highest achievable resolution. Larger areas at the top of the image show the capability for reconstructing grey tones. Unfortunately the noise introduced by simulating a diffuser with a random phase over the object array is fairly severe in this case.

In the image of Fig. 9 the diffuser noise was reduced by applying a random phase only to every second object point in every other row. Phases of intermediate points were obtained by linear interpolation to achieve a smoother distribution. Although this method does reduce image noise, the light diffraction efficiency is also reduced.

Since a binary hologram can reconstruct a general wavefront, it can, of course, reconstruct wavefronts from three-dimensional objects. Like Waters$^{18}$ and Lesem, Hirsch, and Jordan,$^{13}$ we use a Fourier type hologram that forms two real images near the focal point of the reconstruction beam. Calculation of the light propagation from three-dimensional objects is performed according to the usual parabolic approximation to the Fresnel-Kirchhoff
The wavefront to be reconstructed by the hologram is, therefore, the superposition of the Fourier transforms of each object plane in depth, modified by a quadratic phase factor,
\[
\varphi_0 = \pi(x'^2 + y'^2)/\lambda z, \quad (11)
\]
where \( z \) is the hologram-to-image-plane distance.

If a lens of focal length \( f \) is placed near the hologram, it contributes a quadratic phase of
\[
\varphi_L = \pi(x'^2 + y'^2)/\lambda f; \quad (12)
\]
thus the hologram need have only the difference
\[
\varphi_0 - \varphi_L = \pi(x'^2 + y'^2)(f - z)/\lambda fz \quad (13)
\]
encoded in it. This phase factor can be quite small if \((f - z) \ll f\). When this is the case most of the focal power is provided by the lens and high spatial frequencies associated with the phase factor are avoided. Figures 10 and 11 show images reconstructed in three dimensions from such holograms. The planes in depth may be focused individually as shown in the figures.

The lateral extension of these images is about 2 mm; however, in Fig. 10 the planes are separated by 1 mm in depth whereas the separation in Fig. 11 is 10 mm. These separations correspond respectively to 1.3 and 13 times the depth of focus of this setup. The depth of focus is defined as
\[
\Delta z = \pm 2 \left( \frac{f}{D} \right)^2 \lambda, \quad (14)
\]
where \( D \) is the diameter of the hologram. The varying size of the undiffracted zero order on the right hand side of Fig. 11 indicates the relative distances of the image planes from the focal point of the reconstruction beam.

**Conclusions**

During our work we have increasingly found the binary holograms to be an efficient digital-to-optical information linkage. Not only are they efficient in terms of optical performance, but they are easy to produce and have favorable information transfer properties as well. Perhaps the most significant aspect of the computer-generated hologram is that it physically connects the digital technologies of numerical computation and pulse transmission with the capabilities of coherent optical systems for parallel processing and graphic display. As the computational ability to handle large arrays of \(10^5\) to \(10^6\) elements improves, it seems likely that such a link will gain importance. The computer-generated holograms could be useful for relating such quantities of data to the human and physical world. Conversely, as the processing of graphic information becomes more common, computer-generated filters in optical systems could aid in direct optical processing or the conversion of graphic information to digital form.

**Acknowledgment**

We are very grateful for the skillful help of H. Werlich in performing the experimental work and in preparing the figures for this paper.
References


31. Ibid., p. 441.

Received August 1, 1968