Current Developments in Optics and Vision (1968)

Pages
133

Size
5 x 9

ISBN
0309345812

William Benson and Milton A. Whitcomb, Editors; Armed Forces-NRC Vision Committee; National Research Council

CURRENT DEVELOPMENTS IN OPTICS AND VISION

Meeting of Committee on Vision, 1967

Edited by WILLIAM BENSON
and MILTON A. WHITCOMB
ARMED FORCES — NRC COMMITTEE ON VISION

NATIONAL ACADEMY OF SCIENCES — NATIONAL RESEARCH COUNCIL
Washington, D. C., 1968

Reproduction is sponsored by contract NONR 2300(05)
between Armed Forces–NRC Committee on Vision and ONR

Inquiries concerning this publication should be addressed to:
Armed Forces-NRC Committee on Vision
National Academy of Sciences—National Research Council
2101 Constitution Avenue, N.W.
Washington, D.C. 20418

Library of Congress Catalog Card Number 68-61735

CONTENTS

DYNAMIC PARALLAX
Stanley Ballard, Chairman

Dynamic Visual Cues in Flying
Aaron Hyman and Theodore Gold 3

Vision and Driving 22
Albert Burg

HOLOGRAPHY
Stanley Ballard, Chairman

Principles and Properties of Hologram Photography
Juris Upatnieks 35

Present and Future Applications of Holography
J. W. Goodman 43

Scientific Applications of Holography
Robert E. Brooks 57

IMAGE RESTORATION AND ENHANCEMENT
John L. Brown, Chairman

The Elastic Surface Transformation
Robert J. Hall and William F. Dossett 69

Image Processing as it Relates to the Human System 78
James L. Harris, Sr.

DYNAMIC PARALLAX

Stanley Ballard, Chairman

DYNAMIC VISUAL CUES IN FLYING

Aaron Hyman and Theodore Gold

Two work efforts conducted for the Navy by the Sperry Gyroscope Division of the Sperry Rand Corporation will be described in this paper. Both are experimental studies in which dynamic visual cues are presented to the subjects. The first study will discuss the characteristic findings obtained in the first phase of a study of visual perception in carrier landing sponsored by the Office of Naval Research.* The second study will briefly describe the program planned in a study of visual requirements for optical projection displays sponsored by the JANAIR Committee.**

CARRIER LANDING STUDY

Problem

The approach and landing on a conventional airfield is a demanding task for pilots. Special circumstances make landing on the deck of an aircraft carrier even more difficult. The limited size of the touchdown region on the deck, the pitch, roll, and heave motion of the deck, and the absence of visual aids in the approach zone aft of the carrier, are aggravating factors. Recognizing the hazardous nature of a military mission, landing a modern high-performance aircraft on the deck of a carrier can often be the most critical phase of the entire flight operation. The problem is made more acute by the fact that the landing maneuver always

*Contract Number Nonr 4801(000)
**Contract Number N00014-66-CO114

occurs at the end of the mission, when the pilot is most likely to feel fatigued. Also contributing to the critical nature of the carrier landing maneuver are: the structural limitations on the aircraft, which determine the permissible deck impact loads and the arresting hook loads, the high approach speeds, and the demanding control tasks in the approach configuration for high-performance jet aircraft.

The importance of the visual aspects of carrier landing led to a research program sponsored by the Physiological Psychology Branch of the Office of Naval Research. The aim of the program is the evaluation of the effects of some of the key visual factors on the pilot's ability to make critical visual judgments during carrier landings. Day and night operations were considered in these studies. The experimental program was initially concerned with the accuracy with which the pilot can estimate his position on the glide slope and his flight path or aim point during dusk and night carrier landings. The ability to perceive vertical position and flight path are primary factors in the visual control of the aircraft in the vertical plane. These studies were based on the pilot's using only natural cues in his visual environment for landings during a relatively quiet sea. Since the data from these studies provided base-line levels of visual performance, changes in performance caused by specific differences in the visual environment could then be evaluated. For example, how much does the ship motion due to state of the sea deteriorate performance? To what extent does a stabilized optical landing system enhance performance? The effects of heavier seas and of visual aids such as the optical landing system are two key factors that will be evaluated in later studies.

Apparatus

Visual perception was the primary concern in this program. A simulator was designed, therefore, to provide the pilot with an extremely realistic dynamic visual presentation. It was decided to eliminate the control aspect of the carrier landing problem in deference to the visual aspect. However, the equipment would be flexible enough to incorporate a pilot-operated control system when desired.

A perspective drawing of the apparatus designed for the study is shown in Figure 1, and a side view of the arrangement of the key elements in the system is shown in Figure 2. The pilot sat

in a moving cab and viewed a three-dimensional scale model of the carrier on an ocean extending from the approach zone to the horizon. The scale of the real world to the model is 600:1. A periscope with unit magnification was used by the pilot to obtain a viewing point at the required location in the model world. The cab moved longitudinally and laterally to simulate X and Y motion of the aircraft in relation to the carrier. The model ocean bed with carrier moved vertically to simulate altitude changes of the aircraft. Rotating optical elements in the periscope simulated attitude and heading changes of the aircraft. This combination of controls over translation and rotation provided a visual simulation of the motion of the aircraft with all six degrees of freedom. The rendition of the real world was in full color, with an instantaneous visual field of 42°. The apparatus was 12 feet wide, 23 feet long, and 9 feet high.

As shown in Figure 1, the four-wheeled cockpit gantry rode on two elevated tracks to permit lateral (Y) motion. The cross tracks, in turn, were supported by four wheels that ride on tracks to permit longitudinal (X) motion. Independent, but similar, drive

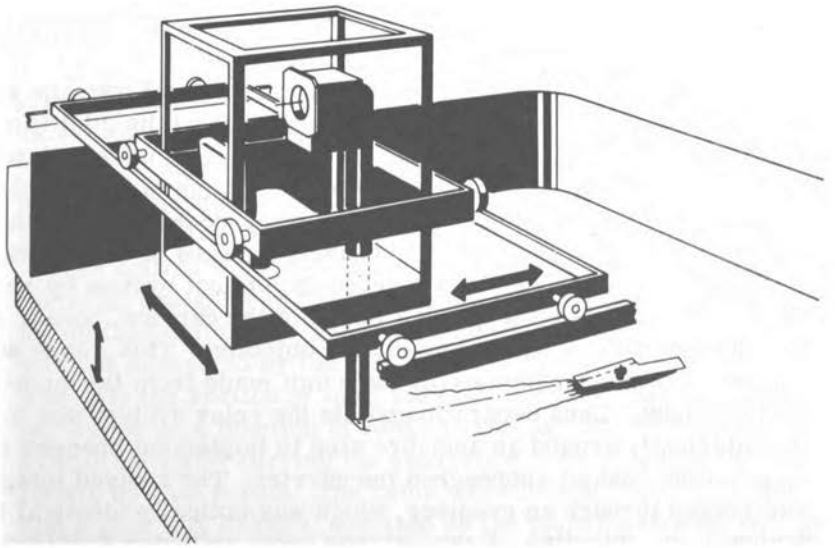

FIG. 1. Arrangement of the cab and ocean bed in the carrier landing simulator. The cab in which the pilot sat moves in the longitudinal (X) and the lateral (Y) directions, and the ocean bed with the carrier moves in the vertical (Z) direction. The pilot viewed the three-dimensional simulated external visual field through a periscope mounted in the cab.

systems were provided for the gantry along each set of tracks. Servomotors, which drive drums mounted on each of the chassis, were the prime movers. A continuous cable, secured to the ends of each track, was wrapped around each drum. The cable was wrapped so that rotation of the drum would wind one end of the cable onto the drum, while unwinding the other end. This action caused a translation of the drums, and, therefore, a cab movement along the tracks in the commanded direction.

The ocean bed was mounted on four vertical lead screws, secured to the main frame of the simulator. The ocean bed moved up and down as the lead screws were rotated simultaneously through bevel gears driven by a set of horizontal jack shafts. The master jack shaft was driven by a servomotor with a belt drive. Most of the weight of the ocean bed was supported by negator springs, to reduce axial loads on the lead screws.

The control of the drive system made possible linear aircraft trajectories between any two points. The flight paths were commanded by inserting an aim point at the carrier deck altitude, a flight path angle in elevation, and a ground track in azimuth. These inputs positioned the ocean bed vertically and the cockpit laterally as a function of range. The cockpit then moved toward the aim point. This scheme permits aircraft trajectories that were either straight lines or combinations of straight lines.

The periscope (Figure 2) served two purposes. First, the entrance pupil located at the scan mirror permitted the pilot to view the carrier from proper positions in model space, notwithstanding the small scale involved (600:1). Second, rotation of optical elements in the system reoriented the external visual field for the pilot to simulate pitch, bank, and heading changes of the aircraft.

Light from the model area entered the optical system by reflection from a scan mirror. An image of the carrier, ocean, and sky was formed by a symmetrical objective. This image was relayed by another symmetrical lens unit made from two high-quality lenses. Lens separation within the relay system was made symmetrically around an aperture stop to implement changes in focus as the cockpit approached the carrier. The relayed image was viewed through an eyepiece, which was optically identical in design to the objective. A two-mirror penta reflector provided a 90° bend in the optical path, and reverted the image so that the pilot's view was consistent with the view of the model space from the entrance pupil position. The final image of the carrier as seen by the pilot was located at approximately optical infinity. A silhouette of the structural members of the cockpit windscreen

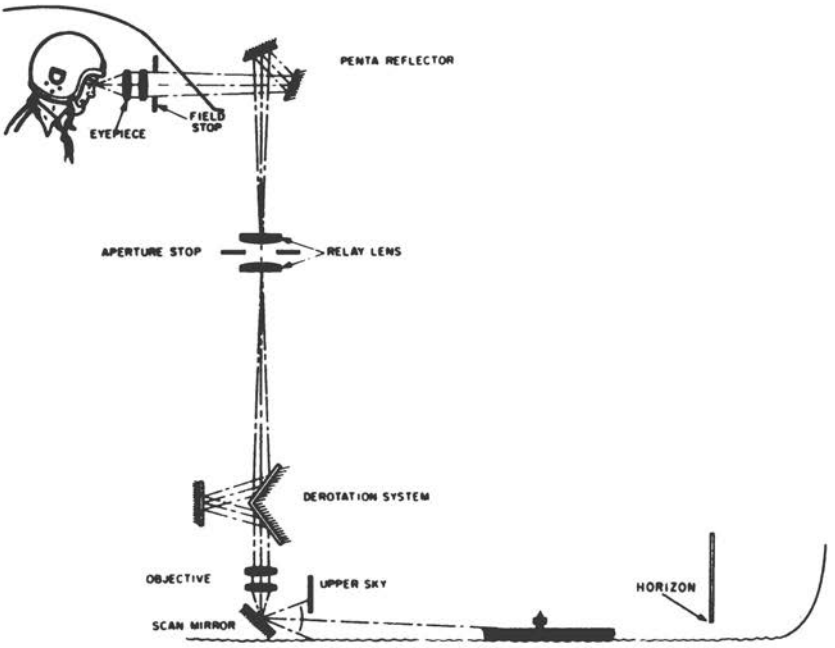

FIG. 2. Periscope through which the pilot viewed ocean, carrier, and horizon. Entrance pupil of the periscope was at the center of rotation of the scan mirror. Rotation of this mirror and of the optical de-rotation system permitted the simultaneous introduction of pitch, bank, and heading changes of the aircraft.

was introduced at a field stop in front of the eyepiece and located optically about 4 feet in front of the pilot.

The optical system comprised achromatic lenses. The symmetrical design used for the system reduced aberrations and minimized distortion. Blurring caused by curvature of the field was not perceived by the pilot because his major area of interest was only the portion of the field near the carrier. Focusing the periscope as a function of range to the carrier was accomplished by mechanically varying the separation between the two relay lenses by cam action. The cams were driven by a chain, sprocket, and a gear box linkage, actuated by the gantry drive system in the range (X) direction.

Changes in pitch were simulated by rotating the scan mirror about a horizontal (Y) axis. Changes in heading were affected by rotating the same mirror about a vertical (Z) axis. Since the mirror was at about 45°, a heading change also induced an apparent

change in the bank angle. The induced bank was compensated by rotating the de-rotation optics. Bank changes were accomplished with signal inputs to the de-rotation system. The attitude and heading optics were all driven by servomechanisms in response to command inputs.

Stereoscopic vision is not a significant factor for the aircraft-to-carrier distances with which these studies are concerned. Therefore, stereoscopic viewing was not duplicated in the simulator. Thus, although the periscope had a monocular eyepiece, provisions were made in the design to incorporate a set of binocular eyepieces if desired.

The exit pupil was made small to minimize position error in the simulation. With a scale of 600:1, any change in eye position in the simulator produces a translation of the aircraft in the simulated carrier approach zone amplified by a factor of 600.

In the visual field, the horizon was fixed vertically at the height of the entrance pupil of the periscope. Therefore, the elevation angle of the horizon was maintained at the same angle (zero degrees) during the entire approach to the carrier. The ocean bed curved upward behind the horizon generator so that the ocean could move vertically without interfering with the horizon line. The combination simulated an endless visual ocean. A translucent diffusing material placed close to the scan mirror simulated the upper sky.

The carrier was made from a plastic scale model of the USS Enterprise. The deck was modified to simulate standard deck markings and deck lighting. Deck lighting was provided from a single fluorescent fixture located under an opening in the hull beneath the ocean surface. Small crimped pinholes in the metal surface of the deck were illuminated by the central source to simulate the individual deck lights. Where necessary, individual pinholes were backed by a reflecting prism to provide a uniform luminance for the lights.

The ocean and the sky near the horizon were illuminated indirectly by four controllable flood lamps on each side of the apparatus. These lamps were directed against a white diffusing surface suspended above the ocean. The arrangement simulated sky lighting of the ocean surface. Small lamps, with the diffusing screens in front of each lamp, were mounted at intervals behind the horizon surface. These lamps provided controlled ocean illumination near the horizon. Horizon backlighting and the ocean and sky lighting were controlled independently to permit any desired contrast at the sky-ocean boundary. A wide range of

ambient conditions could thereby be simulated from daylight through dusk to night, with and without the visual presence of a horizon.

Pilots who made simulated carrier landings with the apparatus reported that it provides a high resolution, distortion free, full color, dynamic simulation of the external visual environment over the range of ambient conditions for which it was designed. In fact, the fidelity of the simulation was limited only by the accuracy and realism with which the visual field can be simulated to the small scale involved.

Procedure for Glide Slope Tests

The carrier landing research program has been planned as a continuing effort, encompassing a range of variables from the relatively simple to the complex elements involved in actual carrier landing operations. The first phase described here covers the pilots' ability to perceive both position on the glide slope and flight path angle (aim point). The tests were conducted under simulated calm sea conditions, i.e., without pitch, roll, yaw, or heave of the carrier. Aircraft trajectories were varied in elevation only. Two ambient lighting conditions were considered, dusk and night, each with and without the presence of a visual horizon. Ancillary visual landing aids were not included.

The glide slope test conditions are illustrated in Figure 3 (not to scale). All linear flight paths terminate at the correct aim point on the carrier deck, permitting the arresting hook to engage the center (no. 3) arresting cable. There are six test flight paths symmetrically placed above and below the ideal glide slope, which is at a 3-1/2° elevation angle. The test paths are displaced ±1/8°, ±1/4°, and ±1/2° from the glide slope. Pilots were requested to make judgments first in two categories, high and low, i.e., above and below the glide slope. Then to augment the data, the pilots were also requested to indicate at which of the three positions in the high and low categories they estimated the aircraft to be, i.e., H-3, H-2, H-1, L-1, L-2, and L-3, corresponding to the six possible positions. The data provided a 6 x 6 stimulus-response matrix for analysis. The glide slope itself was not used as one of the stimuli to avoid a forced choice situation when the aircraft was actually on course. The psychophysical technique used is essentially a modification of the classical method of constant stimuli, with the zero stimulus point omitted.

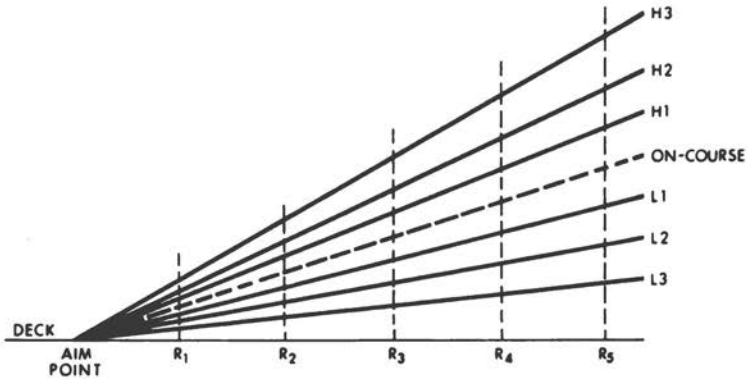

Fig. 3. Shows relationship between positions used as test points in the glide slope study. The on-course line terminates at the aim point on the carrier's deck. The test points are displaced by ±1/8° (H-1 and L-1), ±1/4° (H-2 and L-2), and ±1/2° (H-3 and L-3) from the on-course line. These same angular deviations are used at five ranges from the aim point.

Five ranges were used as variables, with initial values of 7000, 5600, 4200, 2800, and 1400 feet aft of the aim point. The pilot was exposed to the visual situation for 800 feet of travel to the carrier from each of these initial ranges. A single aircraft-to-carrier closing rate of 100 knots was used in this series of tests. The pilots were tested at each of the five ranges in succession, from far to near, in a single test run. The flight paths were varied at random among the six possible values, for each of the range intervals. A single run, therefore, provided five data points. A test session consisted of thirty test runs, yielding 150 data points (6 positions x 5 ranges x 5 replications). This scheme provides five replications of each flight path at each range. Each test session was devoted to a single ambient condition selected from four situations; dusk with horizon (DH), dusk without horizon (DH′), night with horizon (NH), and night without horizon (NH′). Each ambient condition was repeated once (8 test sessions). There were, therefore, a grand total of 1200 data points per pilot (150 test points per session with 8 test sessions).

Procedure for Aim Point Tests

From a stimulus standpoint, the perception of the flight path vector or aim point is interrelated with position. Therefore, the test

conditions shown in Figure 4 were selected for these studies. The same three positions on the glide slope (high, on, low), were used as starting points at each of three ranges (far, medium, near). Five possible aim points were simulated from each of these nine starting points in the approach zone. These aim points were: the correct aim point on the deck, two levels of overshoot, and two levels of undershoot. The overshoot and undershoot aim points were varied in position on the deck as a function of range to provide the same angular deviation from the correct aim point at the three ranges. Angular subtense was considered more appropriate than linear extent as an independent variable in these tests.

Two ambient conditions, dusk with horizon and night without horizon were used, along with two closing rates, 100 knots and 120 knots. Two sets of visual exposure intervals were also employed. This provided successive intervals of 8, 4, and 4 sec or 4, 4, and 2 sec for the three (far, medium, and near) ranges. The combination of two ambient conditions, two closing rates, and two exposure intervals yielded a total of 8 basic test conditions. Each test condition covered 225 data points (5 aim points x 3 glide slope positions x 3 ranges x 5 replications). These data points were distributed among 4 test sessions. Each pilot provided 1800 test points (8 x 225) in the aim point studies.

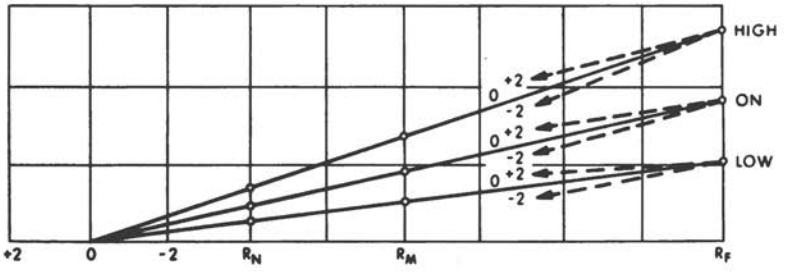

Fig. 4. Relationship of the positions and flight paths used in aim-point study. Three vertical positions (high, on, low) at each of three ranges far (R_F), medium (R_M), near (R_N) were used as starting points. Five flight paths (three shown) were generated from each of the nine original positions. Flight paths include direct intercept of aim point (0), two levels of overshoot (+1, +2), and two levels of undershoot (-1, 12). Deck intercepts for each level are selected so that they will subtend same visual angle at all three ranges.

Subjects

It was determined in preliminary glide slope tests that there were significant differences in performance between carrier qualified pilots and other military pilots. Therefore, six carrier qualified pilots were made available by the U.S. Navy for these studies, each for a two-week period. The pilots were selected from VF (fighter), VA (attack), and VS (antisubmarine warfare) squadrons.

Some Typical Findings

With the method of constant stimuli, the classical procedure is to have a subject judge a stumulus as greater than or less than a designated criterion. Repeated presentations of discrete values of the stimulus are made, and the data obtained are plotted in terms of the probability that responses will be judged to be greater than the reference criterion for the selected stimulus values presented. The functional relation thus obtained between stimulus and response is called the psychometric function.

It is possible, however, to obtain the same psychometric function for two subjects whose response probability density distributions for each of the selected stimulus values differ, provided their subjective scaling of the stimulus also differs appropriately. This is simply illustrated in Table 1 for two normal response

TABLE 1 Hypothetical data illustrating how two subjects generating probability density distributions having different standard deviations can provide identical psychometric functions because of differences in stimulus scaling.

	SUBJECT A			SUBJECT B		
Stimulus	Mean Response	Standard Deviation	P(R>0)[1]	Mean Response	Standard Deviation	P(R>0)[1]
-2	-2	1	.02	-1	1/2	.02
-1	-1	1	.16	-1/2	1/2	.16
0	0	1	.50	0	1/2	.50
1	1	1	.84	1/2	1/2	.84
2	2	1	.98	1	1/2	.98

1. Probability for responses greater than zero.

distributions (hypothetical data). For subject A, the mean response value for a given stimulus is equal to the magnitude of that stimulus, while for subject B, it is one half that amount. However, for subject A the standard deviation for the responses made for a given stimulus is twice that for subject B. This results in identical psychometric functions for the two subjects (see last column for each), yet the implications for their further training would be quite different. It was for this reason the subjects in the carrier landing study were additionally required to estimate the magnitude of the deviation from the reference criterion for each stimulus presented.

Typical data, obtained for one of the pilots, are presented in Table 2 for the glide slope study and in Table 3 for the aim point study. Using the method of least squares, a normal probability curve was fitted to the response distribution obtained with each stimulus. The mean response values and standard deviations listed in the tables were computed from these fitted curves.

A comparison of the data in Tables 2 and 3 shows that this pilot's behavior with regard to glide slope judgments differs radically from his behavior with regard to aim point judgments. In the former case, the standard deviation determined from the psychometric function (difference in stimulus values between the 50% level of probability and either the 84% or 16% level) was generally comparable to the standard deviations obtained for the individual response probability density distributions. This was not at all true for the aim point judgments, since the same pilot underestimated aim point errors by approximately a factor of three.

The classical psychometric functions for glide slope obtained with a second pilot are shown in Figure 5. Each plotted point, of course, represents an independent determination of probability level for a particular position on the glide slope. The plots represent the four ambient conditions considered, dusk and night each with and without a horizon.

The results for the six pilots indicated that in estimating position on the glide slope, the pilots' mean judgment was always close to the 3.5° on-course position under all conditions, with a maximum departure (constant error) of about 0.10°. This indicated absence of any pronounced bias in either the high or the low direction. Standard deviations were all about 0.25° or less, indicating good discrimination for changes in position. Based on these data, pilots could judge correctly that they are high or low 95% of the time, for departures from the on-course position of about 0.50° or more.

TABLE 2. Probability of a "high" response as obtained from curve fitting[1] and from the psychometric function for glide slope data for one subject.[2]

Condition	Performance Measure[1]	Position on Glide Slope (degrees) 3.00	3.25	3.38	3.63	3.75	4.00
Dusk with Horizon	R_M	-3.79	-1.52	-0.73	0.75	1.34	3.32
	R_σ	1.11	1.22	1.78	1.23	1.17	1.04
	P(R>3.5)	0.00	0.11	0.34	0.73	0.87	1.00
	F (Ψ)	0.00	0.10	0.32	0.80	0.94	1.00
Dusk without Horizon	R_M	-3.22	-1.72	-0.93	0.84	1.66	3.37
	R_σ	1.11	1.44	1.26	1.23	1.03	1.46
	P(R>3.5)	0.00	0.12	0.23	0.75	0.95	0.99
	F (Ψ)	0.00	0.12	0.28	0.76	0.96	1.00
Night with Horizon	R_M	-3.87	-1.98	-1.22	0.27	1.37	3.16
	R_σ	1.35	0.96	1.58	1.61	1.06	1.54
	P(R>3.5)	0.00	0.02	0.22	0.57	0.90	0.98
	F (Ψ)	0.00	0.02	0.22	0.54	0.90	0.98
Night without Horizon	R_M	-4.26	-1.60	-0.85	0.81	1.45	3.41
	R_σ	1.97	1.51	1.68	1.52	1.31	1.63
	P(R>3.5)	0.02	0.14	0.30	0.70	0.87	0.98
	F(Ψ)	0.00	0.16	0.28	0.70	0.92	1.00

1. A normal probability curve was fitted (by method of least squares) to the response probability density distribution for each stimulus value. R_M is the mean and R_σ is the standard deviation for the fitted curve; and P(R>3.5) is that area (under the fitted curve) representing the fraction of responses greater than 3.5 degrees, i.e., responses higher than reference response criterion. The function F(Ψ) is the fraction of judgments greater than 3.5 degrees, i.e., "high" responses, computed directly from the stimulus response matrix obtained experimentally. F(Ψ) is the psychometric function.

2. Data in Table 3 are for the same subject.

A graphic presentation of typical psychometric data relating to pilot's ability to estimate flight path, or aim point, are presented in Figure 6. From data of this type obtained from three pilots, the following conclusions can be drawn regarding aim point performance. In general, pilots' sensitivity to aim point changed as the standard deviation of the probability distribution varied between 0.50° and 1°. Comparing dusk and night perform-

TABLE 3. Probability of an "overshoot" response as obtained from curve fitting[1] and from the psychometric function, for the aim point data for one subject.[2]

Condition	Performance Measure[1]	Ratio of Aim Point Displacement over Range				
		-0.102	-0.051	0	0.051	0.102
Dusk with Horizon	R_M	-0.52	-0.24	0.37	0.59	1.05
	R_σ	1.39	1.34	1.23	1.22	1.47
	P(R>0)	0.36	0.43	0.62	0.68	0.76
	F(Ψ)	0.40	0.46	0.64	0.68	0.76
Night without Horizon	R_M	0.17	0.25	0.45	0.75	1.15
	R_σ	1.33	1.36	1.27	1.19	1.37
	P(R>0)	0.55	0.57	0.64	0.74	0.80
	F(Ψ)	0.43	0.59	0.64	0.72	0.81

1. A normal probability curve was fitted (by method of least squares) to the response probability density distribution for each stimulus value. R_M is the mean and R_σ is the standard deviation for the fitted curve; and P(R>0) is the area under the curve representing the fraction of "overshoot" responses. The function F(Ψ) is the fraction of judgments which were called "overshoots" plus one half the "on" judgments as obtained directly from the stimulus response matrix. F(Ψ) is the psychometric function.

2. Data in Table 2 are for the same subject.

ance, the sensitivity was poorer at night for each of the three pilots. These results pertain to data combined for the three ranges. When ranges are separated in the analysis, the pilot's mean judgments of aim point were found to be more aft of the correct aim point at the longer ranges. This indicated that pilots tend to overestimate their aim point at these ranges, i.e., the pilots feel they are overshooting 50% of the time with a stimulus which is an undershoot aim point. This effect was consistently reduced for the short (near) range.

Thus the first phase of the carrier landing study has been concerned with obtaining quantitative data covering the visual judgments of position and aim point made by pilots during carrier landing when there is a calm sea condition and no ancillary visual aids on the carrier. The effects of ship motion and the presence of the Fresnel lens optical landing system will be investigated in forthcoming studies. The composite data will

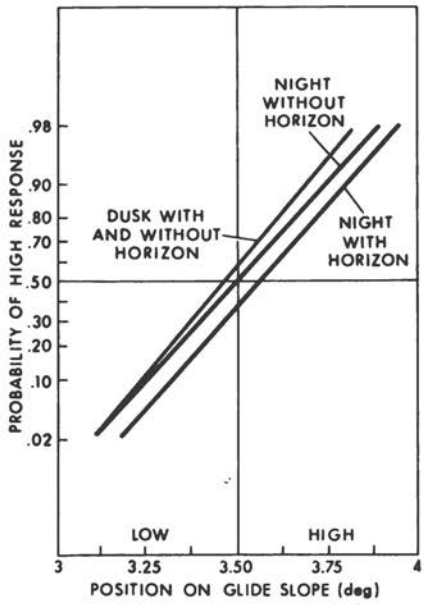

FIG. 5. Psychometric functions describing ability of one pilot to perceive position on glide slope under different ambient conditions. Probability that pilot will perceive he is high is plotted on probit paper as function of vertical position. Straight line indicates that random components of pilot response are normally distributed. Steepness of slope of line is measure of how sharply the pilot discriminates differences in position. Horizontal displacement from on-course line of position at which the probability is 50% is a measure of the pilot's bias in judging position. High 50% intercept indicates that pilot underestimates his position and a low intercept that he overestimates it.

permit evaluation of the individual influences of each of the parameters affecting the pilot's judgment.

VISUAL REQUIREMENTS STUDY FOR OPTICAL PROJECT DISPLAYS

Problem

The experiments to be conducted in this study have been designed primarily to measure the effects of binocular disparity on operator performance with virtual image optical projection displays (e.g.,

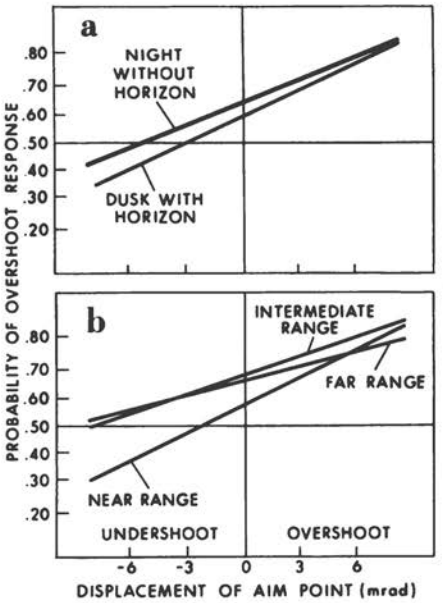

FIG. 6. Typical psychometric functions for one pilot's perception of aim point. In (a), pilot's performance at dusk is shown to be superior (steeper slope) than at night. In (b), pilot bias is shown to be less at near range than at farther ranges (smaller negative intercept). He is also more sensitive to changes at near ranges. Performance at medium and far ranges is about the same.

aircraft head-up windscreen displays). The effects of changing exit pupil size and axial head position will also be investigated.

The head-up display is a relatively new cockpit display technique which provides flight control information to the pilot visually as he looks through the windscreen in the forward direction. The pilot has the visual impression that the images generated by optical projection are located in the real world, in front of the aircraft. These images are called virtual images because the light rays which produce them do not emanate from their apparent locations.

For head-up displays, optical projection systems with large fields-of-view and with qualities which permit considerable freedom of head movement for the pilot are desirable. High accuracy in positioning the images, both in relation to each other and to the real world, is also required. Ideally, this precision of image positioning should be independent of the pilot's head position.

Also, the same image should be presented to each eye if the head-up display is viewed binocularly.

On the basis of existing literature in vision, the visual requirements established for projection systems have led to large, heavy, complex and costly optical systems for certain applications. In addition, there is evidence that some head-up displays have manifested undesirable visual characteristics.

Appropriate visual design criteria for head-up display must be generated. Therefore, the need to conduct experimental studies which will objectively define the visual requirements relating to the design of wide field, virtual image, optical projections system is two-fold. First, a design must be evolved whereby the optical projection system matches the visual characteristic of the pilot, and, second, a reduction in the size, weight, complexity, and subsequent cost of the system must be accomplished.

A well-designed optical projection system must provide a large exit pupil to permit some freedom of head movement of the pilot. When this exit pupil is larger than the interpupillary distance of the eyes, viewing becomes binocular, and image distortion presents a major problem. When viewing is binocular, differences in image distortion for the two eyes become a critical matter. Such differences can give rise to eye strain, display image doubling, and an apparent depth displacement of image elements. With monocular viewing, distortion would give rise to an overlay error in those cases where display symbology must superimpose on real world elements, and tolerance for distortion at various eye positions with respect to the exit pupil would have been determined by the permissible overlay error.

With regard to head movement, the main problems involved are the effect of off-axis eye positions which occur with binocular viewing and the effect of a field stop that provides an instantaneous field less than the total available field-of-view. (Total available field-of-view can be a function of forward and aft movement of the head.) In the parts of the field seen by one eye only, there can be perceptual "wash-out" of projected imagery should retinal rivalry occur. In the binocularly viewed regions, undesirable retinal disparities, a phenomenon which is not possible in monocularly viewed regions, can occur.

An organized research effort is essential to solve these interrelated visual problems.

Apparatus

Specialized research apparatus is required to perform this investigation. It is highly desirable that the simulation of the head-up display be dynamic and that manipulation of the experimental parameters be simply accomplished. Two display devices will, therefore, be employed. One is a binocular viewing system of telecentric optical design in which the visual presentation to each eye is independently controlled. The disparity studies will be conducted on this apparatus. The other display device is wide field optical projection system with a large exit pupil. Effects of size of exit pupil and of axial head position will be studied with this equipment.

The optical schematic for the telecentric viewing device is shown in Figure 7. Since a telecentric system is of symmetrical optical design, distortion is at a minimum. In addition, each eye is provided with an identical viewing system so that even an extremely small distortion, if it should occur, is identical for the

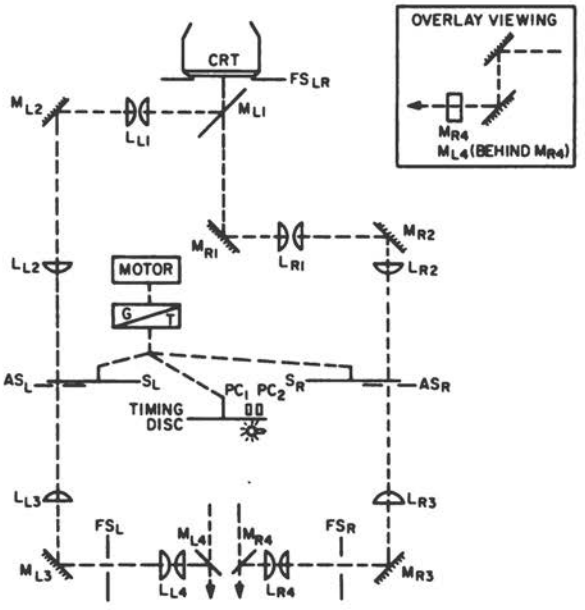

FIG. 7. Optical schematic of telecentric viewing system.

two eyes, and, hence, will generate no binocular disparity. Further precision is obtained by having both systems view a common object, for in this manner there will be no problem with regard to the accuracy of replication of the object for the two viewing systems.

This telecentric viewing system functions in the following manner. For right-eye viewing, the light from a display symbol presented on the CRT face is transmitted through a half-silvered mirror, M_{L1}, totally reflected by mirror M_{R1}, and collected by lens L_{R1}. An image is formed and reflected by a totally reflecting mirror M_{R2}. Then the light is converged by lens L_{R2}, passes through aperture stop AS_R, diverges and is collected by lens L_{R3} to form another image after reflection by mirror M_{R3}. The latter image is viewed by the right eye after it is collimated by lens L_{R4} and reflected by a half-silvered mirror, M_{R4}. In a comparable manner light from the CRT reaches the left eye via the left half of the system. Interpupillary distance is adjusted by moving mirrors M_{L4} and M_{R4} fore or aft in unison. The subject's head position is stabilized with an appropriate support mechanism.

The field-of-view is determined by field stops FS_{LR}, FS_R, and FS_L, Figure 7. FS_{LS} is common to both eyes, while FS_R and FS_L affect only the field for the right eye and left eye, respectively. Alternate presentations to the eyes is accomplished by proper phasing of shutters S_R and S_L. The two shutters will be driven by a common motor through a timing belt and geared pulley arrangement. Each shutter (S_R and S_L) will have two openings to halve the required speed of rotation. Luminance matching of the two systems will be obtained by inserting appropriate neutral density filters near lenses L_{R1} and L_{L1}. Aperture and field stops will be varied in discrete steps.

Binocular overlay viewing of the real world environment can be presented to a subject by having the viewer look through the half-silvered mirrors M_{L4} and M_{R4} and into a two-mirror periscope as shown in side view, in insert of Figure 7. A lens system may be necessary to image the real world overlay at infinity.

To obtain binocular disparity, disparate images will be presented alternatingly (above flicker fusion frequency) to the two eyes. This will be accomplished by the periodic switching in of programmed deviations in the x and y displacement fields of the CRT and synchronizing the shutters, so that one eye sees the symbols displayed in an undistorted field and the other eye sees it in a distorted field.

For exit pupil and axial head position studies, an existing

head-up display device will be used. This viewing device is a wide field, virtual image, optical projection system having a 25° field-of-view and a 3-inch by 5-inch exit pupil located 16 inches aft of the last optical component. This unit is mounted in a B-47 flight simulator and performs as a flight instrument. Dynamic symbology reflecting flight maneuvers and pilot control operations are presented.

Program Objectives

The overall aim of research program is the establishment of quantitative perception tolerances for optical design parameters. It is essential for the establishment of the design parameters to determine what disparity is permissible, how small can the exit pupil be, and whether or not the viewing must be binocular. A knowledge of the operational acceptability of binocularly disparity in a dynamic display of characteristic symbology superimposed on a non-disparate real world view and obtained for various amounts of overlap of the display fields seen by the left and right eye is required.

The specific areas of research to be studied during this program designed to aid in developing optimal designs for virtual image optical projection systems are as follows:

- Binocular disparity
- Tolerance for differing convergence requirements (i.e., focusing)
- Permissible disparity at boundaries of viewing region
- Alleviation of retinal rivalry
- Minimum desired exit pupil size
- Axial head position.

VISION AND DRIVING

Albert Burg

For those of us who have been actively engaged in traffic safety research for a number of years, it is indeed gratifying to see the tremendous upsurge of interest and activity in this area that has come about in the past 18 months. It is gratifying for a number of reasons. First, we are happy at the realization that, for the first time, what more closely approximates the proper degree of importance will be placed on this vital area, both by the federal government and by the American public. We also are justifiably pleased with the expectation that not only will this important area of research be more fully appreciated but that it will receive financial support more in proportion to the magnitude of the problem it hopes to solve.*

Finally, we are pleased with this new interest because it will help make people aware of the many difficult problems we have been and are faced with in traffic safety research, problems whose solutions are vital to improved safety and which, at the same time, have resisted all efforts at solution. The role of vision in driving is one of these problems and the one which is the subject of this paper.

*In this connection, it is interesting to compare the relative amounts of money spent to improve safety in various modes of transportation. Rough estimates of 1965 expenditures are as follows: In the U.S. space program, approximately $12,000 per astronaut per year was spent; in military aviation, some $250 per occupant per year; in civil aviation, $2 per occupant per year; and last (but least), for the automobile, approximately $.05 per occupant per year. Considering the relative accident rates, a serious disproportion is evident.

It has always been assumed, and logically so, that vision plays an important role in the task of driving. While this assumption has been traditionally and universally accepted and has been used by driver-licensing agencies as the basis for incorporating one or more vision tests in their procedures for evaluating driver's license applicants, there has been, in fact, no definitive experimental evidence relating visual ability to driving ability. As was pointed out in a detailed survey,[1] despite a reasonably large amount of published literature on vision in relation to driving, relatively little substantial research has been done and few, if any, basic relationships have been established.

Among the visual factors that frequently have been investigated with regard to driving performance are static acuity, field of vision, dominance, phoria, depth perception, dark adaptation, glare resistance, and color vision. In each research program, either no relationships were found between driving performance and the particular aspects of visual ability under investigation, or if some relationships were found, they were small and, more importantly, they were usually offset by contradictory findings by other investigators.

A close look at this research reveals that each study seems to have had one or more shortcomings. Prominent among these shortcomings are: (a) small sample size; (b) lack of representativeness of the sample; (c) questionable validity of the criterion variable (e.g., accidents, violations, performance of a simulated driving task, and so on); (d) incompleteness of accident records, when they are used as the criterion variable; and (e) inadequate description of the criteria used to classify "good" versus "poor" vision, or "good" versus "bad" drivers.

To summarize our current state of knowledge, it is obvious that at the present time there is no widely recognized evidence that vision is related to driving. Everyone would readily agree that there must be a relationship, but as yet there is no way of translating this"feeling" into valid and useful practice, such as, the development of practical vision test standards for driver-screening agencies.

The fact that previous attempts to relate vision and driving have been inconclusive may be due to any number of factors, including the following:

1. Accidents are rare and complex events, usually involving a number of factors. Vision is only one of these factors, and while it is undoubtedly important, vision's contribution may be small enough to make it difficult to show a significant relationship.

2. "Driving Ability," "Driving Performance," and "Driving Record" are three separate concepts. "Ability" is the one we would like to be able to measure but cannot with any degree of confidence. "Performance" is an approximation of "Ability," but cost, time and legal factors make it impossible to obtain any truly representative sample of driving performance of the large sample of drivers essential to research of this type. This leaves us with "Driving Record," which we can ascertain but without any assurance that we have obtained an accurate or complete estimate of the driver's on-the-road performance over the specified period of time because of the lack of consistency and completeness in reporting accidents and traffic citations.

Vision is very likely an important factor in both driving ability and driving performance. Driving record, however, is influenced strongly by the element of chance or luck and thus is, at best, an uncertain indicator of true driving ability. Unfortunately, since we do not as yet have the means to measure true driving ability, there is no alternative but to use driving record (which is two steps removed from driving ability) as the criterion.

3. The vision tests used may not be valid. That is, they may not be related to the visual functions used in driving. These functions, needless to say, have yet to be established.

4. The reliability of the vision test and/or of the measure of driving ability used (e.g., driving record) may be poor. It is well known, for example, that the correlation between driving records in adjacent periods of time is low.

5. What the person is capable of seeing and the use he makes of this capability in driving may differ substantially. In other words, there may be considerable disparity between an individual's physiological vision and his functional vision. (A complicating factor here is that the individual may, to a certain extent, be able to compensate for his visual shortcomings,[2] thus masking unsafe vision in driving.)

6. Using drivers as our subjects, as we must, we necessarily are dealing with a restricted range of visual capabilities. For example, a large sample of drivers will probably not contain any individuals whose static acuity is worse than about 20/100. Perhaps the truth of the matter is that when a person's acuity is no worse than 20/40 or 20/50, which is usually the case with drivers, his acuity is adequate to the driving task, and other factors, either visual or nonvisual, such at attitude, attention and "luck," play a much more important role.

It is unnecessary to delve further into the multiplicity of reasons that have been given to explain the lack of relationships found between vision and driving. Suffice it to say that the research that is currently underway at UCLA and that is about to be described was undertaken with full knowledge of the various pitfalls mentioned above, and a concerted effort has been made to avoid as many of them as possible.

AIMS OF CURRENT RESEARCH*

Early in 1961, the University, with the California Department of Motor Vehicles as a co-participant, received a grant from the U.S. Public Health Service to investigate the "Relationship Between Vision Test Scores and Driving Records." Utilizing a large sample of volunteer California driver's license applicants, the USPHS project (which represents the first phase of a long-term research program) seeks to determine whether a relationship exists between how well a person "sees," as measured by several standard and nonstandard vision tests and how well he "drives," as reflected in his driving record (accidents and convictions for traffic violations accumulated over a three-year period). The vision tests being used include: dynamic visual acuity, static (or standard) visual acuity, lateral visual field, lateral phoria, vision at low levels of illumination, glare recovery, and sighting dominance. Because of time considerations, all possible vision tests could not be included; however, those chosen were felt to have the greatest chance of success. Of these, the dynamic acuity test is the newest and most "nonstandard," and it is the keystone test of the battery.

Dynamic visual acuity (DVA) refers to the ability to discriminate an object when there is relative movement between the observer and the object. Previous research (summarized by Burg[1]) has indicated that it is difficult to predict an individual's DVA score from his static acuity score. In view of the fact that many

*The research herein described was partially supported by funds provided by the U.S. Public Health Service (Grant AC-00015) and by the California Transportation Agency and U.S. Bureau of Public Roads (California Standard Agreement 13600). The opinions, findings, and conclusions expressed are those of the author and not necessarily those of the supporting agencies.

activities, including driving, flying, ballplaying and the like, involve the discrimination of moving objects as a major aspect of the visual task, it is reasonable to suspect that performance on a dynamic test may be more closely correlated with task performance than is the score obtained on a test of static acuity. Put another way, it is logical to assume the DVA test to be a more valid indicator of performance of these types of tasks than is static acuity. It was on the basis of this assumption that the present study was conceived, although, needless to say, the research program ultimately put into effect far surpassed the original concept in scope and size.

For example, we are also currently involved in a long-term extension and modification of the original USPHS project. This second phase of the research program is jointly sponsored by the California Transportation Agency and U.S. Bureau of Public Roads, and while it has the same general purpose (and the same title, incidentally) as the USPHS study, this new project contains two different and important aspects:

1. Continued accumulation of driving record information for the drivers originally tested under the USPHS program in order that driving performance may be evaluated over a long period of time (up to 12 years). This will provide a broadly-based criterion variable to which visual performance (and the many other personal and driving variables being studied) can be related, thereby increasing significantly the validity and reliability of any obtained relationships.

2. Retesting of many of the original subjects to gain information relative to short-term (two to three years) deterioration of visual capabilities. The data thus obtained, in conjunction with the long-term driving record referred to above, will make possible the relating of visual deterioration to any changes in driving performance that may occur and will help provide a basis for a decision concerning the use of differential licensing requirements based on age.

The potential value of both studies lies in the expectation that from the results will come indications of what aspects of vision seem to be most important in driving and, consequently, what vision tests may be of greatest value in screening future driver's license applicants. A second, but equally important product of this research, is the accumulation of a very large amount of normative data describing the visual, personal and driving characteristics of a large sample of drivers. These data will prove

invaluable source material for future investigators working in the related areas of both traffic safety research and vision research.

METHOD

Details of the conduct and major findings of the U.S. Public Health Service study appear in another report.[3] A number of other reports[1, 4, 5, 6, 7, 8] have been published during the course of the study, and the reader is referred to them for detailed information not covered in this report.

To summarize briefly the conduct of the study, over a 32-month period some 17,500 volunteer driver's license applicants were interviewed and tested at 46 branch offices of the California Department of Motor Vehicles scattered throughout the State. The subjects, of both sexes and ranging in age from 16 to 92, constitute what appears from preliminary evaluation* to be a relatively representative sample of the California driving population. Each subject was first questioned regarding personal data and his driving characteristics and was then administered the previously listed vision tests.

Three-year driving records were obtained for approximately 14,400 of these tested subjects by utilizing the files of the Department of Motor Vehicles and California Highway Patrol and, when possible, insurance company records. (The remaining 3100 subjects are new or out-of-state drivers who have not as yet accumulated a three-year California driving record. They are being held until such a record is available for analysis.)

RESULTS

A number of major computer analyses of the data have been conducted thus far; these analyses, which are described in another report,[3] include tabular data presentations, correlational analyses, and multiple regression analyses. However, because of the massive amount of data accumulated in the USPHS study as well

*Personal and driving record characteristics of the test subject sample were compared with those of a two percent sample of California drivers[9] and with similar characteristics obtained for some 11,400 driver's license applicants who declined to participate in the study.

as the large number of additional data presently being accumulated on the second phase of the research program, it will be several years before the multitude of projected data analyses can be completed. A series of reports will be issued during this time that will describe not only the detailed visual performance characteristics of the subjects but also the relationship between certain types of visual performance (or other personal characteristics) and specific aspects of driving performance.

Based on the statistical analyses conducted to date, the major findings are as follows:

1. As is to be expected, the two driving record variables being considered (that is, accidents and convictions for traffic violations in a three-year period) correlate most highly with each other and next most highly with quantitative exposure (i.e., average annual mileage) and with age and age-related variables such as marital status and driving experience.

2. Of the vision tests investigated, only dynamic visual acuity and static visual acuity consistently appear to be related to driving record. Of these, DVA shows the strongest relationships. (Lateral visual field is a distant third.)

3. Considerable variation in the vision-driving relationship is found as a function of the sex and age of the group being studied.

It is important to point out that the primary purpose of obtaining such a large number of subjects was to permit breaking the total sample down into age-sex-mileage subgroups of sufficient size for proper statistical analysis. Every analysis conducted to date has supported the well-known fact that age, sex, and mileage are variables that greatly influence driving record; the analyses also reveal striking differences in visual performance on most of the tests as a function of age and sex. For example, the results show that:

1. Both accident and conviction frequencies increase markedly as mileage goes up. However, accident and conviction rates (per 100,000 vehicle miles) decrease slightly with increased mileage.

2. In general, convictions (whether frequency or rate) decrease with increasing age. Accidents (whether frequency or rate) behave similarly but to a lesser extent.

3. Males have poorer driving records than females, when relative frequencies of accidents or convictions are compared, but little or no difference in record exists between the two sexes when accident or conviction rates are compared.

4. For both sexes, visual performance on all tests but lateral phoria declines markedly and progressively with increasing age.

5. Average performance for males is superior to that for females on all static or dynamic acuity tests and on the lateral phoria test. Females are slightly superior in visual field, glare recovery, and low-illumination threshold.

DISCUSSION

Because of complex interactions that occur among many of the study variables, it becomes extremely difficult to derive general statements about the relationships between vision and driving that are applicable for both sexes, all ages, all measures of driving record, and so on. For example, when the effect of age is studied, poor vision is found to be positively related to poor driving for older drivers but negatively related to poor driving for young drivers. When all age groups are combined, however, the net result is a negative relationship between poor vision and poor driving (in terms of frequency of accidents and convictions) due to the fact that young drivers, who have the best vision and the poorest record, contribute a disproportionately larger percent of the total number of accidents and convictions than do the older drivers. To further complicate the issue, the above finding involving accident and conviction frequencies does not hold when accident rates are considered, in which case the relationships between poor vision and poor driving are all in the "expected" (i.e., positive) direction. Furthermore, when we analyze only those accidents in which vision cannot be ruled out as a possible contributing factor, stronger relationships with vision test scores are found than in the case in which all accidents, regardless of type, are considered.

The purpose of these examples is to emphasize the complexity of the problem under investigation and the need for many additional analyses of various permutations and combinations of the data to elicit specific vision-driving relationships that the more general analyses conducted thus far can only suggest. This does not mean, however, that the general findings obtained thus far should be considered equivocal; rather, the results of future analyses, which will consider in detail such things as specific accident and violation types and qualitative as well as quantitative exposure information, will supplement the present findings and provide an ever greater degree of understanding of the vision-driving problem than the present findings make possible.

As is to be expected, the interrelationships between visual test

performance and driving record are small; however, they are highly significant, due to the large sample sizes involved. For example, the simple correlation coefficients obtained range from 0.05 to 0.15, depending upon age, sex, the vision test involved and the driving record variable concerned. When the multiple regression analyses are considered, which indicate the relative contribution of a number of vision and nonvision variables to the prediction of accidents or convictions, the vision variables are shown to contribute far less than the major predictors of age, sex and mileage. The point is, however, that in the multiple prediction approach, which is the most promising for driver-licensing agencies, inclusion of vision variables significantly increases the accuracy and, therefore, the value of the prediction. This is a significant finding in itself and should act as a stimulus to even more detailed studies.

The fact that of all the vision variables studied, only DVA and static acuity show up as being consistently related to driver record, does not necessarily mean that no other vision test is of value in predicting accidents or convictions. As was pointed out earlier, because of time limitations, a number of vision tests that were adjudged least likely to be related to driving were not included in the present study. Until these tests are evaluated in an in-depth study such as the present one, the validity of this judgment cannot be ascertained. Furthermore, the planned analyses involving specific types of accidents and convictions may very well reveal relationships that exist with some of the older vision variables studied in the present research program.

Finally, the fact that DVA shows a stronger relationship with driving record than does static acuity is important (and, to a certain extent, gratifying) since it supports the assumption upon which the original study was based and thereby suggests the likelihood of a relationship between dynamic vision testing and other tasks as well.

At our present stage of analysis, it is too early to estimate the practical utility of the relationships uncovered thus far. Although the multiple regression analyses reveal that both static and dynamic visual acuity are of significant value in predicting accidents and convictions, the degree of their contribution varies markedly as a function of age and sex.

It is possible that the ultimate outcome of the analyses will be a set of relationships, each specific to a given set of personal and driving variables. While the total utility of such a set of relationships cannot as yet be estimated, one of the possible areas of

application that comes immediately to mind has to do with the institution of differential licensing requirements based on age (and possibly, also, on sex). For example, it is felt that the obtained data permit the generalization that vision and driving are much more consistently related to one another for the older age groups. Thus, it is reasonable to suspect that it would be practical to utilize a more comprehensive battery of vision tests for the older driver's license applicants than for the younger ones. Further, it might also be practical that once a driver reaches a certain (as yet undetermined) age, he be required to undergo driver's license renewal examinations, both visual and otherwise, at more frequent intervals than is the case for the younger driver. In California, for instance, the standard term of a license is four years, regardless of the age of the applicant. (Shorter-term licenses are given to drivers whose licenses have expired or who have accumulated too many convictions.) It might be advisable that after a driver reaches a certain age, he be given only a three-year (or two-year, or one-year) license from that point on, so that a closer check could be kept on his fairly rapidly deteriorating visual capabilities.

Incidentally, because of this relatively rapid change in visual capabilities that occurs after middle age, it is felt that a program of vision evaluation at regular intervals should be made a part of every state's driver's license renewal program. This, unfortunately, is not the case at the present time.

Finally, a word about drawing conclusions from the data analyses. The University's role in research is to add to the body of knowledge about a subject. Assuming the results we obtain are the consequence of properly conducted research, we have met our obligation. On the other hand, by far the more difficult task lies with operating agencies such as the Department of Motor Vehicles. Their task is to evaluate research findings in order to decide a course of action based not only on such findings, but on economic and time factors as well.

To give you an idea of the magnitude of the effects of such decisions in California, if the results of future planned analysis continue to show dynamic visual acuity to be the test most closely related to driving and if a reliable DVA test is developed that is also compact and not unreasonably expensive, then the Department is still faced with deciding whether the potential value of the DVA test justifies a capital equipment outlay that would certainly exceed a million dollars, not including future maintenance and replacement costs. To look at the problem in another way,

if the final results of our study were strongly to suggest the addition of five minutes of vision-testing in screening each applicant, this would necessitate the hiring of an additional 132 driver's license examiners, an increase of approximately 38% over the present manpower requirements. If this figure sounds too high, I ask you to remember that California issues nearly 13,000 new or renewal licenses each and every day. In a sense, then, the Department is directly responsible for the safety and welfare of a great many people but within restrictions of time and money, so the implications of each decision as regards the procedures for granting driver's licenses must be carefully studied before action is taken. It is our responsibility to help make this decision easier by providing them with as much information as possible.

REFERENCES

1. Burg, A. An Investigation of Some Relationships Between Dynamic Visual Acuity, and Driving Record. Los Angeles: University of California, Department of Engineering Report No. 64-18, 1964.
2. Fletcher, E. D. Visual acuity and safe driving. Journal of the American Optometric Association, 1949, 20, (7), 439-442.
3. Burg, A. The Relationship Between Vision Test Scores and Driving Record: General Findings. Los Angeles: University of California, Department of Engineering Report No. 67-24, June 1967.
4. Burg, A. & Coppin, R. S. Visual acuity and driving record. Highway Research Record, 1966, No. 122, 1-6.
5. Burg, A. "Some Preliminary Findings Concerning the Relation Between Vision and Driving Performance," Journal of the American Optometric Association, 38 (5), pp. 372-377, 1967.
6. Burg, A. Apparatus for measurement of dynamic visual acuity. Perceptual and Motor Skills, 1965, 20, 231-234.
7. Burg, A. Visual acuity as measured by dynamic and static tests: A comparative evaluation. Journal of Applied Psychology, 1966, 50 (6), 460-466.
8. Burg, A. "Light Sensitivity as Related to Age and Sex," Perceptual and Motor Skills, 24, pp. 1279-1288, 1967.
9. California Department of Motor Vehicles, Division of Administration, The 1964 California Driver Record Study, Sacramento: CDMV, Parts 1 through 8, December 1964 through April 1967.

HOLOGRAPHY

Stanley Ballard, Chairman

PRINCIPLES AND PROPERTIES OF HOLOGRAM PHOTOGRAPHY

Juris Upatnieks*

Holography is the name given to the new technique of photography which, in exact scientific language, is called photography by wavefront reconstruction. Holography differs from ordinary photography in that no lenses are used to form an image on the film. Instead, light waves themselves are recorded on the film in such a way that they can be recovered at a later time. This fundamental difference gives holography several exciting and unusual characteristics. This paper will briefly describe the development, basic principles, and properties of hologram photography.

Holography was invented in 1947 by Dennis Gabor at the Imperial College of London in England. He originated the basic idea of recording the light waves, theoretically proved that this should be possible, and experimentally demonstrated that the theory was correct. It did not gain widespread use at that time because of two difficulties with this process. For one, Gabor did not have efficient light sources like lasers and experiments were difficult to perform. The other obstacle was that with the original technique an out-of-focus image always formed on top of a focused image. Consequently, the resulting image quality was rather poor. Most of the effort following Gabor's invention was directed at eliminating the out-of-focus image, but interest eventually subsided.

Gabor also gave the name hologram to his type of a photograph. It comes from the Greek words "holo," meaning complete, and "gram" meaning recording. Thus the word hologram means

*Radar and Optics Laboratory, Willow Run Laboratories, Institute of Science and Technology, The University of Michigan, Ann Arbor, Michigan.

complete recording, which it is indeed. The image obtained from a hologram has all the optical properties of the original scene.

Work on Gabor's principles was resumed at the University of Michigan in the early 1960's. First of all, the out-of-focus image problem was solved theoretically and the principle was demonstrated with a simple experiment.[5] Shortly thereafter, the laser was invented and high resolution images of continuous-tone transparencies and three-dimensional objects were made.[6, 7] The demonstration of the capabilities of holography attracted widespread interest, and the field rapidly expanded into a major area of research in optics. Denisyuk,[8] in Russia, independently arrived at another technique of removing the out-of-focus image. The explosion in research effort in holography started in about 1964. In barely three years great advances in this field have been achieved.

The process of hologram recording has been previously described in a simplified manner.[9] We shall give here a brief review of the basic techniques and properties of holography.

Hologram recording depends on the property of interference of light. Only coherent light can interfere, and therefore we need coherent light to make a hologram type of photograph. Coherent light is monochromatic, and it has the property that many light waves arriving at a point can interfere: they can subtract or add depending on the relative phase of the waves. When waves from many light scatterers, such as surface irregularities of a diffusely reflecting surface, fall on a plane and interfere, the result is an extremely fine granular pattern that looks somewhat like film grain. Though seemingly meaningless, this pattern contains information about the direction and intensities of the component waves.

Coherent light has to be produced in some way since none of the natural sources of light are coherent. One way to get coherent light is to place a small pinhole in front of an ordinary light source, such as an arc lamp, and then pass it through an interference filter. The result is very weak partially coherent light which will interfere over limited path differences. This type of a light source was used by Gabor, Leith and Upatnieks in their early work. Lasers are much better sources since they are more coherent and more intense. Hologram type photographs of three-dimensional objects can be made only with lasers, while holograms of transparencies can be made either with laser or filtered arc lamps.

The basic optical system for making a hologram is shown in

Figure 1. Coherent light illuminates both the subject and a mirror nearby. The beam reflected by the mirror falls directly on the photographic plate. Also, light from the subject falls on the plate and combines with the other beam, called the reference beam. The reference beam interferes with the light reflected from the object, and this interference pattern is recorded on the film. This pattern is very fine for most three-dimensional scenes, and the photographic film must be capable of recording at least 1,000 lines/mm.

We may get some insight into the recording and reconstruction process of hologram images by considering the following spatial case. Consider that instead of the object shown in Figure 1 we have another mirror which reflects a second beam onto the photographic plate. The two beams interfere producing a pattern consisting of straight lines which is recorded on the film and this pattern is known as a diffraction grating. The properties of diffraction gratings have been known for a long time, and they are known to diffract light. If, after processing the film, we replace the recorded grating in its original position and illuminate it with only one beam of light, then the grating will diffract light. One of the first diffracted orders will be identical to the second beam originally falling on the film. Thus, a simple plain beam of light was recorded on film and recovered at a later time.

If we have a complicated diffusely-reflecting three-dimensional object, then the interference pattern does not consist of simple

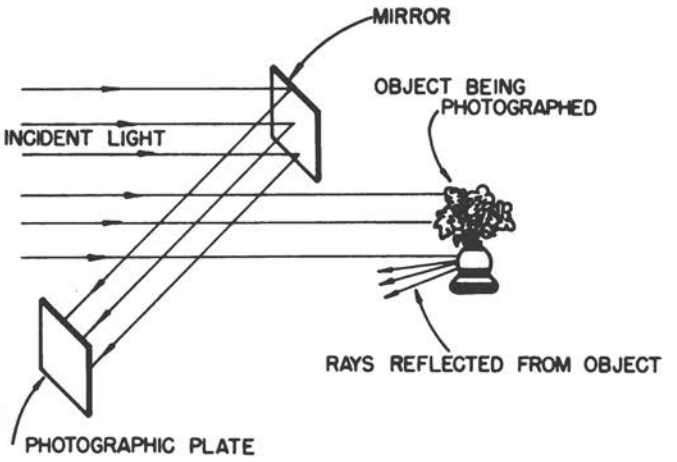

FIG. 1. The construction of a hologram. The incident light must be coherent but not necessarily collimated.

straight lines but of a complicated pattern that resembles film grain. If this pattern is recorded and then illuminated again with a duplicate of the reference beam, then the other beam will be recovered. This recovered beam is identical to the original light waves coming from the subject, and they are optically identical to the original light waves. A hologram can be considered as being a very complicated two-beam interference pattern which, when illuminated with one beam, will reproduce the other. The reference beam is chosen to be of a simple form, for example spherical, so that it can be easily reproduced in the reconstruction process. Figure 2 shows a typical hologram and Figure 3 illustrates the technique of reconstructing the images.

Similar to a grating, there are two diffracted orders in a hologram reconstruction. One of these, the virtual image, is an exact replica of the original light waves. When one looks at this image, the light rays appear to originate from an image situated some distance behind the hologram. Actually, the light rays come only from the plate where they are diffracted. The image is very realistic and has the effects of parallax and three-dimensionality:

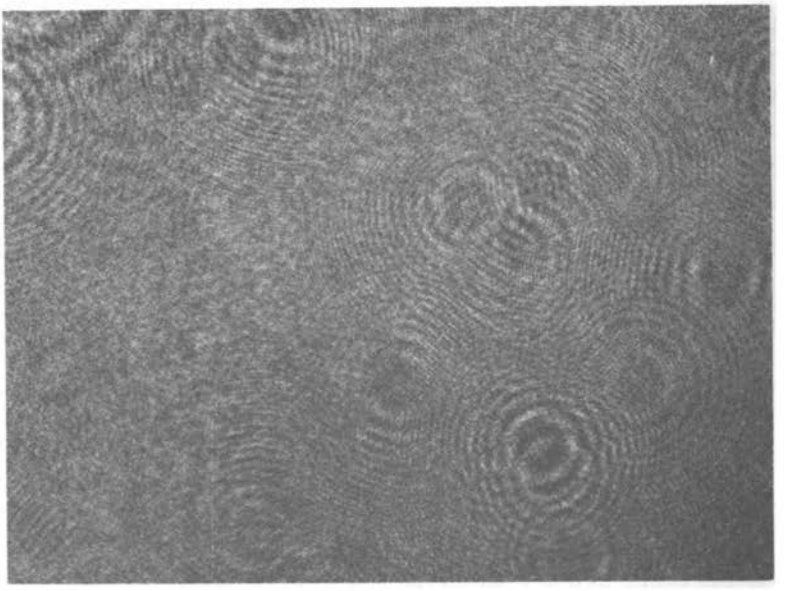

FIG. 2. A typical hologram. The visible pattern is caused by dust in the reference beam and it is not related to the image recorded on the hologram.

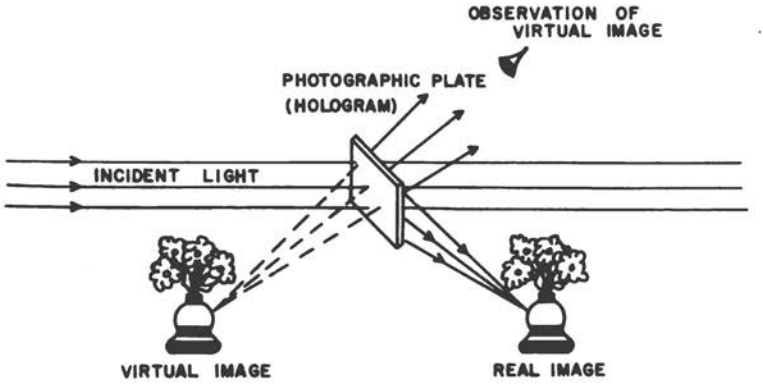

FIG. 3. Reconstruction of a hologram.

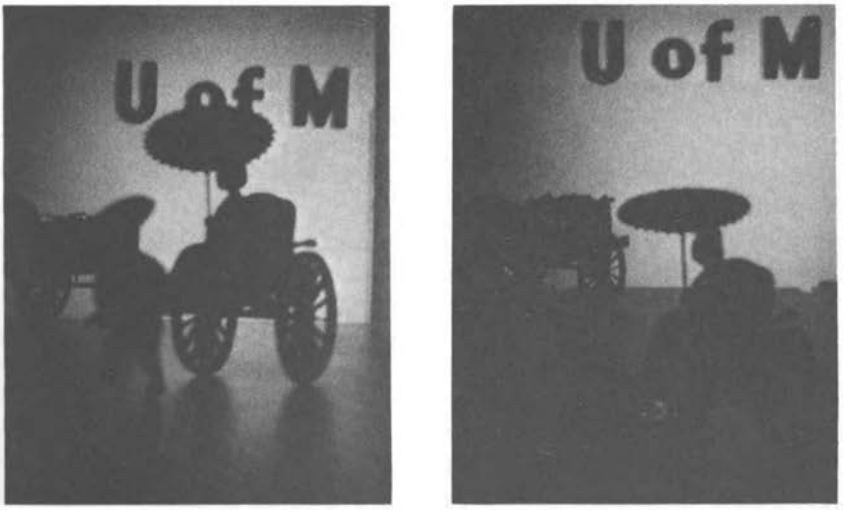

FIG. 4. Two-dimensional reconstructions from a hologram of a three-dimensional scene. Both images were made from a single hologram.

one can look around an object in front and see behind it. Figure 4 illustrates two different views obtained from one hologram.

The virtual image can be compared to a scene appearing behind a window. One can see the scene behind a window from different perspectives, limited only by the size of the glass. The scene is behind it and is completely three-dimensional.

Similarly, a scene appears behind a hologram, and the perspectives from which one can observe it are limited by the size of the hologram.

The other diffracted order from a hologram is different from the first one. These light waves have the same amplitude but conjugate phase: instead of diverging, they converge and form an image in space in front of the hologram. This image can be viewed with a magnifier or a microscope; or a photographic film can be placed at the position of focus, and an ordinary two-dimensional image can be obtained. The real image is not good for viewing for several reasons. Two of these reasons are that the real image is pseudoscopic and that one has difficulty in focusing the eyes on it.

Under certain circumstances, such as the reference beam being perpendicular to the hologram plane and emulsion being thin relative to the fringe spacing, the two images, real and virtual, appear simultaneously as shown in Figure 3. If the emulsion is sufficiently thick and fringe spacing is fine, then the reference beam must be duplicated in shape and position before a virtual image will be seen, and the real image may be completely suppressed. For most wide-angle holograms the conditions are such that only one of the images—the real or the virtual—will appear at a time, and the orientation of the plate is critical. This property can be used to store several holograms on one plate and to reconstruct them one at a time by changing the wavelength of light or the angle of the plate.[10] By changing wavelength of light, full color images can be made by recording three holograms with the primary wavelengths.[11, 12] By changing orientation of the plate in making the hologram, several images can be recorded in sequence and recovered at a later time.[10] The latter idea may lead to short hologram movies or cartoon-type presentations. The basic ideas for information storage in three-dimensional media were described by van Heerden[13] and Denisyuk,[14] and further theoretical and experimental work has been done by Leith et al.[15]

Holography possesses a number of unique and interesting properties which are listed here:

1. The image is always positive, both the first recording and photographic copies of it.

2. The image always appears granular, exactly like any diffusely reflecting surface in coherent light. Granularity, in general, can only be eliminated by making the image incoherent, which results in decreased resolution of the image. Coherence can be

decreased by using a partially coherent light source in the reconstruction process, by moving the position of the observer's eyes, or by other similar means.

3. Every part of the hologram contains the whole image but from a different perspective or direction. A hologram can be cut into many parts and each part will have the whole image.

4. The hologram can be made at one wavelength of radiation and reconstructed at another, say visible wavelength. This is a very useful property since one can essentially look at the world at a wavelength at which the eye does not respond.

5. The real hologram image can be magnified without using lenses by simply changing the divergence of the illuminating beam of light. Magnification without the use of lenses has applications in holographic microscopy.

6. Coherent light is required to make holograms which means that holograms can be made only of scenes illuminated by laser light. None of the natural light sources, such as sunlight or light from electric lamps, are suitable for this.

7. The scene must be motionless during exposure time. This requirement restricts it at the present time to rigid objects and laboratory types of setups. If pulsed lasers are used, then some motion is permissable. Pulsed lasers are not very coherent and are limited in their usefulness, but improvements are expected in the near future.

8. Coherence requirements are much lower for reconstructing than for making holograms. Lasers are the best sources, but filtered arc lamps also give good reconstructions. Some holograms are made in such a way that they also act as interference filters and select a narrow range of wavelength from a white-light source. These holograms can be viewed in sunlight or with flashlights.

Holography is a new field and still at its beginning. In the three years since it attracted attention great progress has been made, and we may expect it to grow further in the coming years. Holography already has found practical and useful applications in science and in the laboratory as an important tool of investigation and measurement.

REFERENCES

1. D. Gabor, Nature 161, 777 (1948).
2. D. Gabor, J. Appl. Phys. 19, 1191 (1948).
3. D. Gabor, Proc. Ray. Soc. (London) A197, 454 (1949).

4. D. Gabor, Proc. Phys. Soc. B64, 449 (1951).
5. E. N. Leith and J. Upatnieks, J. Opt. Soc. Am. 52, 1123 (1962).
6. E. N. Leith and J. Upatnieks, J. Opt. Soc. Am. 53, 1377 (1963).
7. E. N. Leith and J. Upatnieks, J. Opt. Soc. Am. 54, 1295 (1964).
8. Y. N. Denisyuk, Soviet Physics-Dakladay 7, 543 (1962); Doklady Akad. Nauk SSSR 144, 1275 (1962).
9. E. N. Leith and J. Upatnieks, Scientific American 212, No. 6, 24 (June 1965).
10. E. N. Leith, J. Upatnieks, A. Kozma and N. Massey, J. Soc. of Motion Pict. and Television Eng. 75, 323 (1966).
11. K. S. Pennington and L. H. Lin, Appl. Phys. Letters 7, 56 (1967).
12. A. A. Friesem and R. J. Fedorowicz, Appl. Optics 6, 529 (1967).
13. P. J. van Heerden, Appl. Optics 2, 393 (1963).
14. Y. N. Denisyuk, Opt. Spectry 15, 279 (1963).
15. E. N. Leith, A. Kozma, J. Upatnieks, J. Marks and N. Massey, Appl. Optics 5, 1303 (1966).

PRESENT AND FUTURE APPLICATIONS OF HOLOGRAPHY

J. W. Goodman

The current explosion of interest in Gabor's wavefront-reconstruction process[1, 2, 3] can be attributed to two independent developments: First, the introduction of the so-called "offset-reference" hologram by E. N. Leith and J. Upatnieks at the University of Michigan,[4, 5, 6] and, second, the invention of the laser. The mating of these two advances, one in technique and the second in technology, has led to a rapid growth of proposed and demonstrated application of what we now call holography. Our purpose here will be to review some of the more interesting and promising of these applications. Consideration is limited to that portion of holography which might be more accurately called wavefront-reconstruction imaging and will not consider those aspects of optical data processing that are sometimes also included in the term holography.

MICROSCOPY

Historically, microscopy was the potential application which motivated most of the early work on wavefront-reconstruction imaging. Gabor conceived of holography as a potential technique in electron microscopy. The prospect of constructing a lensless electron microscope was of considerable significance in view of the severe aberrations associated with electron lenses. While the envisioned application to electron microscopy has, for practical reasons, not yet been successful; nonetheless, microscopy has remained an important motivating force in holography.

In the present state of development, a sufficient number of

new applications of holography have been found so that microscopy can no longer be said to be the prime motivating force that it once was. In addition, while the use of holograms for optical microscopy has been amply demonstrated, it is clear that the new techniques are not serious competitors with the conventional microscope in ordinary, run-of-the-mill microscopy.

Nonetheless, there do exist two areas in which holography offers a unique potential for microscopy. First is the area of "high-resolution volume imagery." In conventional microscopy, high transverse resolution is achieved only at the price of a very limited depth of focus. Thus only a limited volume of the object space can be brought into focus at one time. It is, of course, possible to explore a large volume in sequence by continuously refocusing to examine new regions of the object volume, but such an approach is often unsatisfactory, particularly if the object is a dynamic one, continuously in motion. A solution to these problems may be obtained by recording a hologram of the object using a pulsed laser. The dynamic object is then "frozen" in time, but the recording retains all of the information necessary to explore the full object volume. If the hologram is illuminated, the real or virtual image may be explored with an auxiliary optical system. Sequential observation of the object volume is now acceptable because the object (i.e., the holographic image) is no longer dynamic. This approach has been fruitfully applied in the microscopy of three-dimensional volumes of biological specimens[7, 8] and in the measurement of particle-size distributions in aerosols.[9]

A second application in the field of microscopy, which must be said to be at too early a stage for full evaluation, is the potential use of holograms to perform X-ray microscopy.[10] The basic concept here is the use of X-ray illumination for the recording of a hologram and the use of optical illumination for the reconstruction process. Enormous magnifications can be achieved by this change of wavelength, but, more important, the use of X-rays for illumination could yield resolutions of a few angstroms. While such resolutions do not greatly exceed those achievable with electron microscopes, the use of X-rays might yield less sample heating than is produced by an electron beam and might also allow operation without the vacuum required in electron microscopy. The success or failure of holographic X-ray microscopy undoubtedly rests on future developments in the X-ray source and sensor technologies.

INTERFEROMETRY

Holography offers the capability of performing several rather unique kinds of interferometry. The ability to perform interferometry by wavefront-reconstruction rests on the fact that the images formed are coherent, with well defined amplitude and phase distributions. Any use of holography to achieve the superimposition of two coherent images will result in a potential method of interferometry.

The most powerful holographic interferometry techniques are based on the property that coherent addition of complex wavefronts can be achieved by multiple exposures of holograms.[11] Thus if a photographic emulsion is exposed sequentially to form n superimposed holograms, upon reconstruction the n corresponding coherent virtual images will be formed simultaneously and will mutually interfere. Similarly the n real images will mutually interfere.

The most dramatic demonstrations of this type of interferometry have been presented by Brooks et al.[12] and are discussed elsewhere in these Proceedings. It suffices to mention here that holographic techniques are particularly well suited for performing interferometry through imperfect optical elements (e.g., windows) that may exist between the measurement plane and the object, and this property makes possible certain kinds of interferometry which could not have been achieved by any classical means.

VIBRATION ANALYSIS

A holographic technique for determining the modes and depth of vibration of certain vibrating structures has been proposed and demonstrated[13] by Powell and Stetson of the University of Michigan. This technique may be regarded as a generalization of the multiple-exposure interferometry method to the case of a continuous time-exposure of a vibrating object.

With reference to Figure 1, consider a planar object, such as a drum-head, which is in sinusoidal vibration toward and away from the plane where a hologram is recorded. The exposure time is assumed to be much longer than a vibration period. Any one point on the structure spends most of its time at the peaks

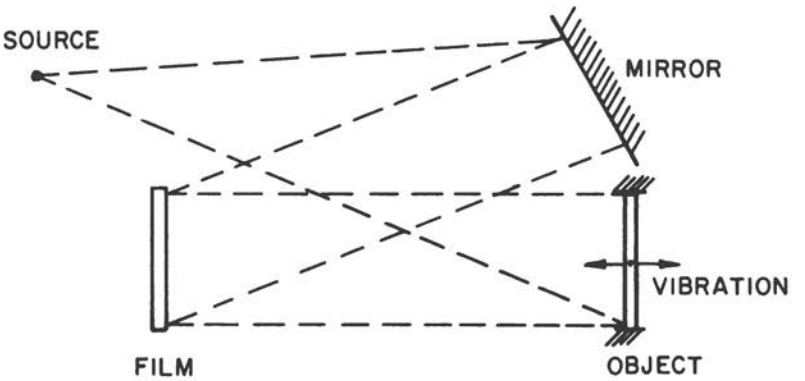

FIG. 1. Vibration analysis by holography.

of the vibration excursion, where the rate-of-change of displacement is minimal. Thus to a first approximation the intensity of the image at any one point may be regarded as resulting from interference of two objects, each existing at one of the peaks of the vibration cycle. At points where the peak-to-peak vibration depth is an even multiple of $\lambda/4$, the two effective objects produce images that are in phase and, therefore, add constructively; when the vibration depth is an odd multiple of $\lambda/4$, the two images are out of phase and add destructively. A more careful analysis shows that, for small subtended angles, the intensity of the image at any point is proportional to $J_0^2 (2 \pi m/\lambda)$, where J_0 is a zero-order Bessel function of the first kind, and m is the peak-to-peak vibration amplitude at the point in question.

Thus a time-exposed hologram of a vibrating object produces an image with superimposed light and dark fringes. From the characteristic structure of the fringe pattern, it is possible to identify the mode of vibration, and the vibration depth can be deduced by counting the number of fringes. The advantage of the holographic technique is, of course, that this information can be determined from a distance with no close contact between measurement apparatus and object required.

CONTOUR GENERATION ON THREE-DIMENSIONAL OBJECTS

The interference properties of multiple coherent images have also led to the development of techniques for obtaining three-dimensional images with superimposed constant-range contours.

Such techniques are applicable to the problems of cross-section tracing and contour mapping.

The most widely applicable of these contour-generation techniques is the so-called "two-frequency" method. The object in question is illuminated by coherent light containing two distinct frequency components. A hologram of the object is recorded, using a reference with the same two frequency components. When the resulting hologram is illuminated with single frequency light, two images with slightly different positions and magnifications are produced. These two coherent images will interfere, and for certain geometries the resulting light and dark fringes will be accurate indications of depth.

A dramatic demonstration of the potential of this technique has recently been published.[14] In this particular case the two-frequency illumination was obtained from two different lines of an argon laser. The object was a quarter, with the variations in depth being simply the reliefing on the coin. The two laser lines were separated by 6.5 nm, and resulting contours on the image were spaced at 0.02 mm increments of depth.

IMAGING THROUGH ABERRATING MEDIA

In many cases of practical interest, an optical system may be required to form images in the presence of uncontrollable aberrations. These aberrations may result from imperfections of the image-forming components themselves, or they may be introduced by an external phase-perturbing medium, such as the atmosphere. We discuss here three distinctly different holographic techniques for obtaining high resolution in the presence of severe aberrations.

The first technique of interest[15, 16] is applicable when the aberrating medium is constant in time and movable in space. As illustrated in Figure 2, a hologram of the aberrated object waves is recorded with an unaberrated reference wave. If the hologram is illuminated in such a way as to generate a real image of the object, a real image of the aberrating medium will likewise form between the hologram and the image plane. But it is a fundamental property of the real images formed by holography that they are the complex conjugate of the original object. Thus where the aberrating medium delayed the object wavefront, now at the image of the aberrating medium the corresponding portion of the wavefront is advanced. If the original aberrating medium is inserted

during the reconstruction process such that it exactly coincides with its conjugate image, then the aberrations are exactly cancelled, and an aberration-free real image forms.

This technique may find application in cryptography or the secure encoding of messages. The message to be encoded serves as the object. A diffuser is inserted between the object and the film plane. Illumination of the hologram in the ordinary manner will produce an image which is too badly aberrated to be read. However, the one person in possession of the very same diffuser used during the recording process can, by properly inserting his "decoding plate," obtain an unaberrated image.

A second technique of interest is illustrated in Figure 3. In this case the aberrating medium may be immobile but must be unchanging in time (or at best, only very slowly changing). To obtain high-resolution images in the presence of aberrations, a hologram of the aberrated wavefronts produced by a point-source is first recorded by means of an unaberrated reference wave. This hologram may now be used as a "compensating plate" to enable a more conventional optical system to form an aberration-free image. The operation of this technique rests on the fact that the "real image" term of transmittance of the hologram is proportional to the conjugate of the original aberrated wavefront incident upon it. Thus, if the hologram is reinserted in precisely the same position in which it was originally recorded, it will "straighten out" the wavefronts incident from each point on a more general object, allowing the conventional optical system which follows to form an unaberrated image.

This technique will work well over only a restricted field of view, for if an object point is too far from the position of the original point-source used in recording the hologram, the aberrations imparted to its wave may differ from the aberrations recorded on the hologram. This restriction is least severe when the aberrations are very close to the collecting aperture of the optical system. Upatnieks et al.[17] have successfully applied this technique to the compensation of lens aberrations, an application to which it is particularly well suited.

Finally, we consider a third technique[18] which may be applied to imaging through media which are movable or immovable and time-varying or time-independent. This technique is accomplished by passing both the reference wave and the object wave through the same aberrating medium. As indicated in Figure 4, the so-called "lensless Fourier transform" recording geometry is used. The reference emanates from a source that is coplanar with the

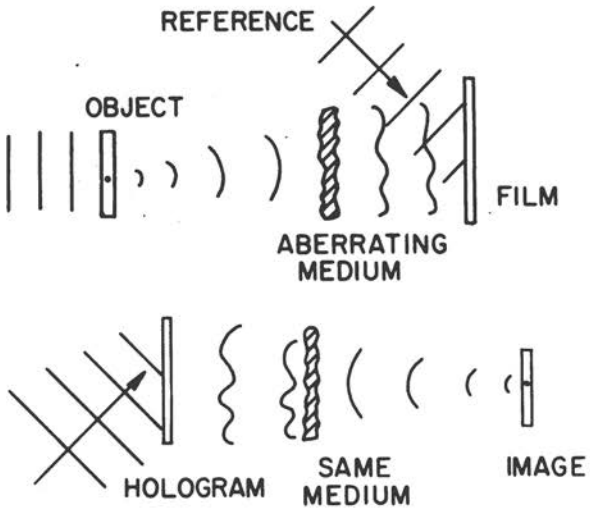

FIG. 2. The use of the original medium for compensating aberrations.

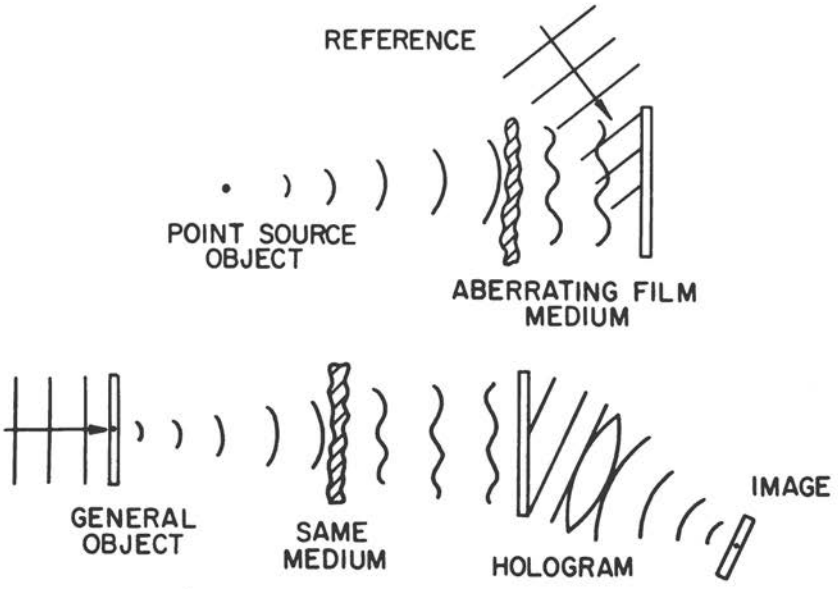

FIG. 3. The use of a hologram compensating plate.

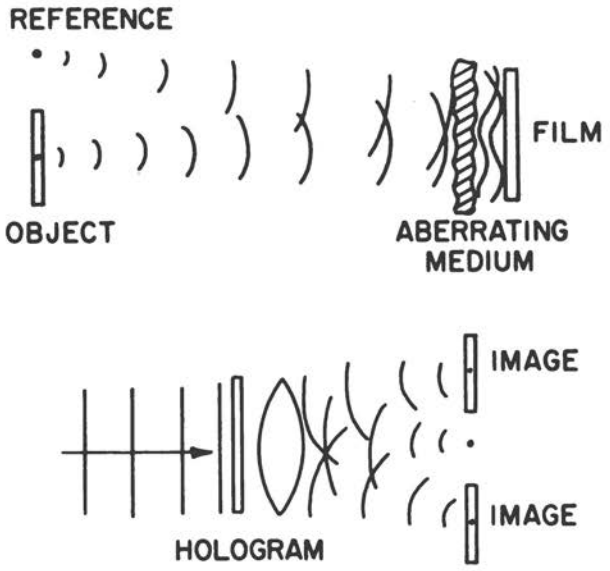

FIG. 4. Reference and objects waves both pass through the aberrating medium.

object. If the object is sufficiently close to the reference and if the aberrating medium is close to the recording plane, then the rays to any one point on the film from both reference and object undergo identical phase delays. As a consequence, the pattern of interference of the two waves is unaffected by the presence of the aberrating medium, and an unaberrated hologram is obtained. From this hologram an aberration-free image can be formed.

Note that this technique is most effective when the aberrations are close to the collecting aperture of the optical system. This is precisely the same condition for which the performance of a conventional imaging system is worst! It has been suggested that this technique may find application in ground-based imaging of coherently illuminated space objects orbiting well above the earth's atmosphere.[19]

Figures 5 through 9 illustrate experimental results obtained using this latter technique. The object is in this case a sign adjacent to a bright reference point. Figure 5 shows a conventionally formed image of the object and reference, no aberrating medium present. Figure 6 shows the corresponding wavefront-reconstruction images, again obtained in the absence of any aberrating medium. The aberrating medium (common shower glass) shown in Figure 7 was then introduced immediately in

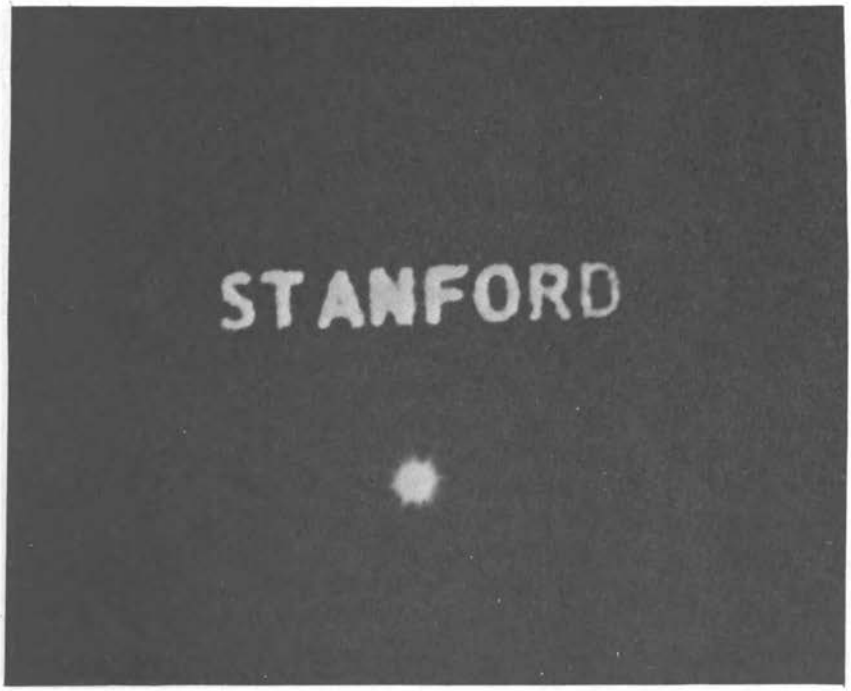

FIG. 5. Conventional image—no aberrations present.

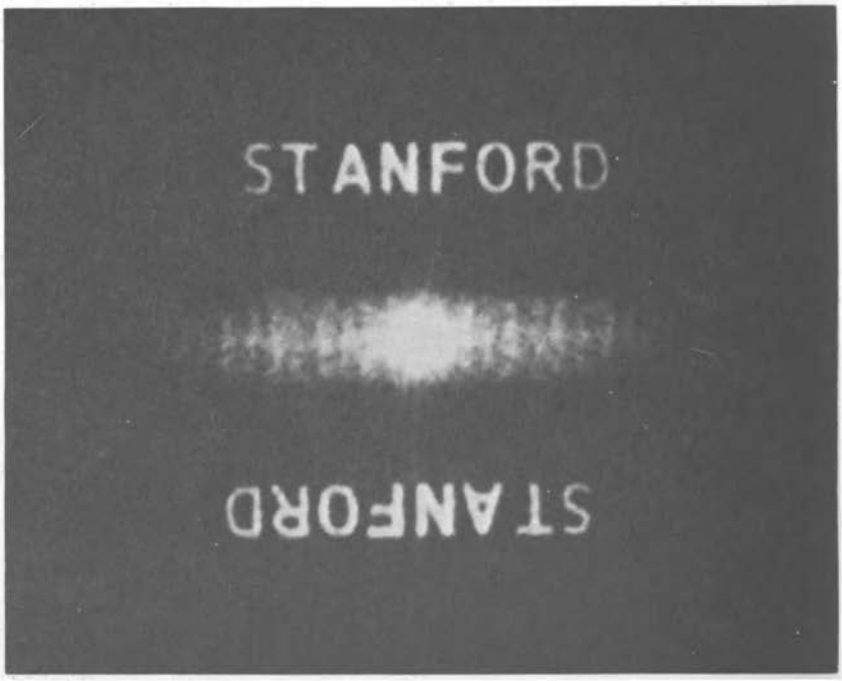

FIG. 6. Wavefront-reconstruction image—no aberrations present.

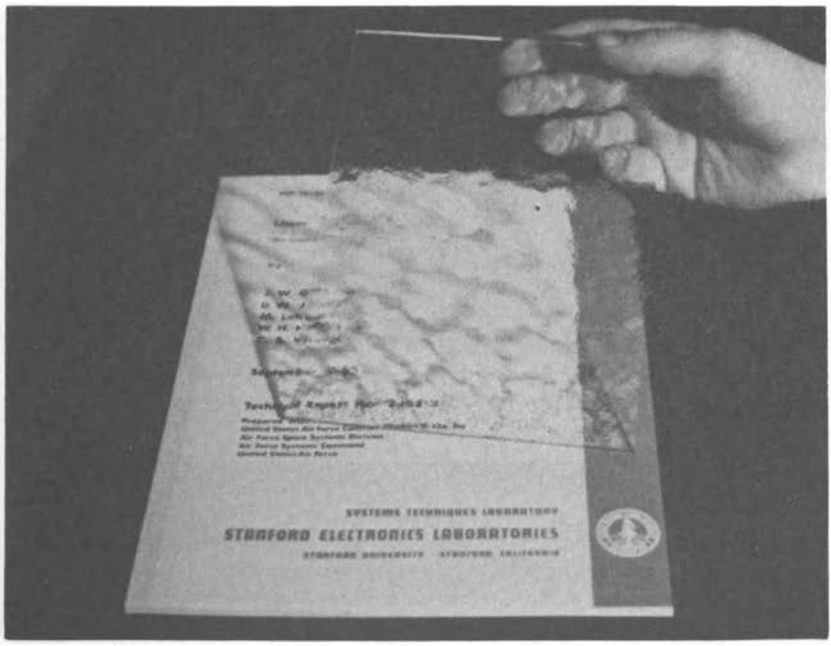

FIG. 7. The aberrating medium.

front of the collecting aperture of both systems. Figure 8 shows the resulting conventional image—resolution has been completely destroyed. Finally, Figure 9 shows the wavefront-reconstruction images obtained under identical conditions—the images have suffered only minor degradations.

PROGNOSIS FOR THE FUTURE

The source technology has represented a major constraint on applications of holography in the past, and, in spite of the revolution brought about by the laser, the same remains true today. While the readily available HeNe gas laser is, to a large extent, responsible for the widespread contemporary experimentation in holography, nonetheless this source in many cases merely "whets the appetite," spawning ideas of new applications for which its limited output power (1 to 100 milliwatts) is quite inadequate.

Developments are now taking place, however, which will undoubtedly broaden the class of problems to which holography may be applied. For example, the so-called "supermode" technique has been applied to the argon laser,[20] producing a single-frequency CW source at the 1/2 watt level with, in all probability, kilometers of coherence length.

FIG. 8. Conventional image—aberrations present.

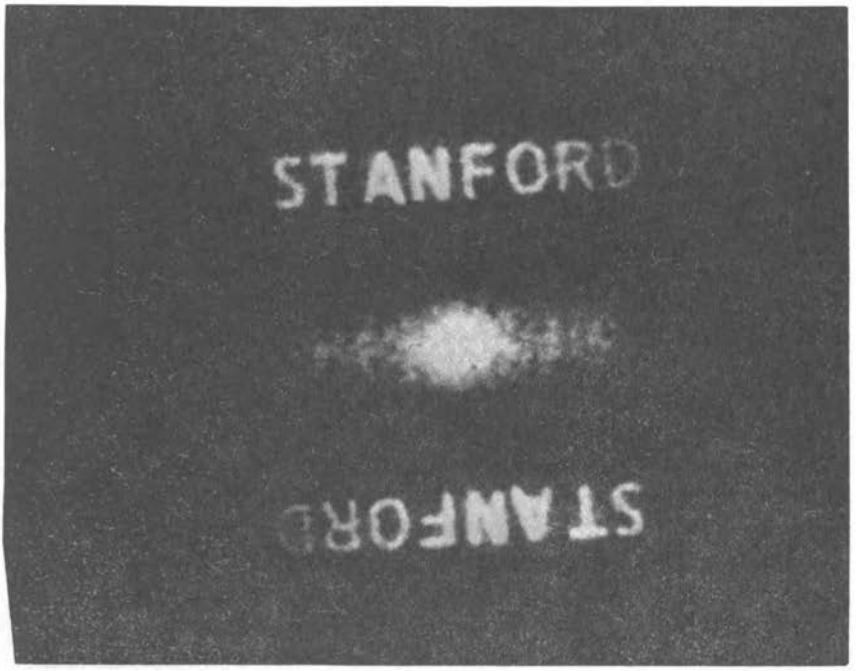

FIG. 9. Wavefront-reconstruction image—aberrations present.

Pulsed lasers are also improving to the point where they can find widespread application in holography. Modification of a commercially available pulsed ruby laser by my co-workers at Stanford has yielded what appears to be a single-mode source with one joule of output energy in a 20 nanosecond pulse. The coherence length of this source has been measured (by holography) as about four meters, while the pulse itself is only six meters long. Holograms have been recorded over a 100 foot path using this source in a monostatic radar configuration, and plans are now underway for a similar experiment over a 12 kilometer path.

The detector technology also poses significant constraints on holography. For scenes of wide angular subtense, enormous numbers of resolution-cells must exist on the recording medium, and high-resolution films with their attendant low sensitivity are unavoidable. For objects of small angular subtense, the requirements are more modest. If the object is weak, sensitivity is at a premium, and in some cases a highly sensitive electronic detector, such as an image orthicon, is both feasible and desirable.

Workers at Stanford have successfully detected a hologram directly on the photosensitive surface of a vidicon and formed the image on a digital computer, with a computation time shorter than is required for the processing of most photographic films.

Thus with the rapidly developing source and sensor technologies, holography too can be expected to develop, finding application to new problems that previously were unfeasible and generating new applications that are impossible to foresee at this time.

REFERENCES

1. D. Gabor, Nature 161, 77 (1948).
2. D. Gabor, Proc. Roy. Soc. (London) A197, 454 (1949).
3. D. Gabor, Proc. Phys. Soc. B64, 449 (1951).
4. E. N. Leith and J. Upatnieks, J. Opt. Soc. Am. 52, 1123 (1962).
5. E. N. Leith and J. Upatnieks, J. Opt. Soc. Am. 53, 1377 (1963).
6. E. N. Leith and J. Upatnieks, J. Opt. Soc. Am. 54, 1295 (1964).
7. R. F. Van Ligten and H. Osterberg, Nature 211, 282 (1966).
8. C. Knox, Science 154, 1195 (1966).
9. B. J. Thompson, J. H. Ward and W. R. Zinky, Appl. Optics 6, 519 (1967).
10. G. W. Stroke, An Introduction to Coherent Optics and Holography, Academic Press, New York, N.Y. (1966).
11. D. Gabor et al., Phys. Letters 18, 116 (1965).
12. R. E. Brooks, L. O. Heflinger and R. F. Wuerker, IEEE J. Quantum Elect. QE-2, 275 (1966).
13. R. Powell and K. Stetson, J. Opt. Soc. Am. 55, 1593 (1965).
14. B. P. Hilderbrand and K. A. Haines, J. Opt. Soc. Am. 57, 155 (1967).
15. E. N. Leith and J. Upatnieks, J. Soc. Photo. Instr. Engrs. 4, 3 (1965).
16. H. Kogelnik, Bell System Tech. J. XLIV, 2451 (1965).
17. J. Upatnieks, A. Vander Lugt and E. Leith, Appl. Optics 5, 589 (1966).
18. J. W. Goodman et al., Applied Phys. Letters 8, 311 (1966).
19. Report of the Woods Hole Summer Study, "Restoration of Atmospherically Degraded Images," National Academy of Sciences, Washington, D.C. (1966).
20. L. M. Osterink and R. Targ, "Single Frequency Light from an Argon FM Laser," Proceedings of the Symposium on Modern Optics, Polytechnic Press, Brooklyn, N.Y. (in press).

SCIENTIFIC APPLICATIONS OF HOLOGRAPHY

Robert E. Brooks

Although the spectacular ability of the hologram to create a true three-dimensional image is readily appreciated and perhaps is chiefly responsible for its current popularity, this is but one manifestation of the more basic property of the hologram to record and at a later time to reconstruct a complex optical wave.[1-4] The scientific implications of this basic property are enormous and are responsible for a number of important applications of the hologram.

MICROSCOPY

The use of holography for high resolution photography (microscopy) of subjects distributed through regions very large compared to those which can be recorded by conventional imaging systems of comparable resolution is of great interest.[5-7] Because the hologram records and reconstructs waves, focusing can be postponed until after the subject has been recorded. The need to focus occurs only when viewing the reconstructed image, but at this point the limited depth-of-field of the observer does not matter, for the observer can focus at will throughout the subject volume without compromising image resolution. In fact, the need to focus during viewing can be an advantage for it permits one to determine the longitudinal position of the subject.

The hologram contains an immense amount of information about the subject and becomes extremely useful when it is used to record transient or changing subjects. A hologram made with a pulsed laser can record the subject in a fraction of a micro-

second and the reconstructed image, frozen in time and containing all the subtleties of the original subject, can be leisurely examined in great detail.[5, 7-9]

Theoretically, the hologram can reconstruct the recorded wave without introducing aberrations, with image resolution limited only by the wave nature of light. In practice the resolution often is much less. In part this is due to physical distortions, aberrations, and insufficient resolution of the plate (usually photographic) upon which the hologram is recorded as well as limited coherence and aberrations in the laser beam.[9] Of great significance is the fact that there is additional image degradation in a perfect system (holographic or conventional) when monochromatic laser light is used. This results from the fact that the highly monochromatic light from the laser readily interferes with itself giving rise to spurious patterns which compete with the subject. The seriousness of these effects depends greatly on the nature of the subject and the type of illumination. When illuminated with non-diffuse (e.g., collimated) light, the subject does not blur and fade away when it goes out of focus. Rather, it gives rise to diffraction fringes which grow as it becomes more defocused. Although least objectionable with isolated particle or hair-like subjects, the fringes are distracting and reduce the accuracy with which the longitudinal subject position can be determined. If the subject is diffusely illuminated, the random (but stationary) phase differences of the waves forming the image cause the subject to have a granular appearance. The size scale of the granularity is of the order of the theoretical resolving power of the viewing system.[10, 11]

The most straightforward use of the hologram for imaging is to directly view the virtual image through the hologram with a long working-distance microscope. Because of the imposition of the hologram between the observer and the image, a long working-distance objective is generally needed. The arrangement is most suitable for applications requiring only moderate image resolution (typically 30 microns).

By reversing the direction of propagation of the reference beam with respect to the hologram during reconstruction, but otherwise leaving the geometry unchanged, an aberration-free real image can be obtained which is readily accessible to viewing with a short working-distance microscope, even though the

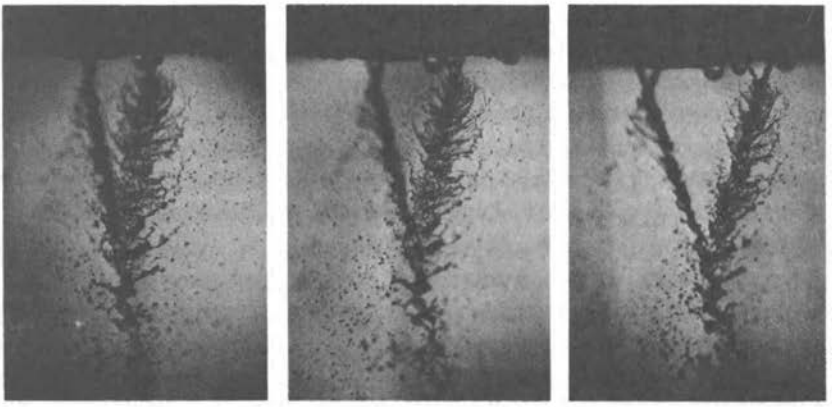

FIG. 1. Three views, spanning an angular range of 25°, of the intersection of two jets of water from a model rocket nozzle. All three photographs were taken of the reconstructed virtual image of a single hologram. A resolution of better than 1/10 millimeter can be obtained throughout the entire scene.

subject may be relatively large.* The image is aberration-free even though imperfect optical elements are placed between the subject and the hologram; during reconstruction the waves pass back through the same elements and any aberrations are "unraveled."[15-17] This latter property is useful for cancelling aberrations introduced when a relay lens is used to magnify the subject before it is recorded on the hologram. Although restricting the subject volume, such premagnification reduces the demands on the hologram and makes the attainment of high image resolution more practical.

Since the real image is viewed from the side opposite the hologram, it is accompanied by a depth inversion. That is, near parts appear far and far parts near. Although in most cases this is not a problem, depth inversion can be eliminated by making a hologram of the hologram and thereby causing a double inversion.[17]

In the one-beam (Gabor) holographic arrangement, a single

*Although it is possible to provide a magnified image without the use of lenses by changing the wavelength of geometry of the reference beam during reconstruction, such magnification is accompanied by aberrations.[12-14]

beam serves both as the reference and to back illuminate the subject.[18-20] Historically, the first holograms were produced in this way, and it continues to find application in holographic microscopy where the subject does not obscure a significant portion of the beam and the hologram is in the far field of the light diffracted by the subject.[5,7] Its advantage lies in its simplicity and its minimum demands on laser coherence and film resolution.[9,21]

INTERFEROMETRY

The ability of the hologram to record a complex optical wave and to synthetically generate this "stored" wave, faithful in both amplitude and phase, upon demand at a later time, has ideally suited the holograms to interferometry. The hologram can be used to generate either or both of two beams which are to be interferometrically compared. These can be generated by means of separate holograms or from the same hologram by means of a double exposure.

The unique wave storage property of the hologram makes it possible for the first time to cause interference between waves which exist at different times. This opens up an entirely new class of interferometric techniques and makes possible measurements which were hitherto not possible.

By separating the interfering beams in time rather than in space, one can use a common optical path for both beams. Such an interferometer is sensitive only to changes in the system and therefore permits the employment of imperfect optical elements without sacrificing accuracy of the measurements.

If the subject is diffusely illuminated, the resulting interferogram is "three-dimensional" in the sense that as one changes his viewing perspective he sees the interference pattern resulting from optical rays traversing the subject in that direction. This provides a great deal more information about asymmetric phenomena than can be obtained conventionally.

A very important feature of the common path interferometer is that it permits differential interferometric measurements to be made of subtle changes in an otherwise complex subject. Such applications might include viewing changes in a subject imbedded in an optically imperfect media, checking parts against a master, measuring erosion, or analyzing surface deformations due to mechanical or thermal expansion forces.

The simplest form of the holographic interferometer results from sequentially recording in time, on a single hologram, the two beams which are to be compared.[9, 22, 23] Upon reconstruction the two recorded waves are generated simultaneously and their interference can be viewed. If nothing is changed between the two exposures except to introduce the subject, a "perfect" infinite fringe interferogram of the subject results. If the subject is changed between the two exposures, then the interferogram documents the difference. There is no alignment to worry about (the system is self-aligning), and because the two holograms are locked together on the same plate, the reconstruction step is not critical. For ease of interpretation a finite fringe pattern is often preferred over the infinite fringe pattern.[24] This can be easily obtained by slightly changing the angle of the reference beam between exposures, the equivalent of inserting an additional thin wedge in the beam.

Real time interferometry, in which changes in the subject can

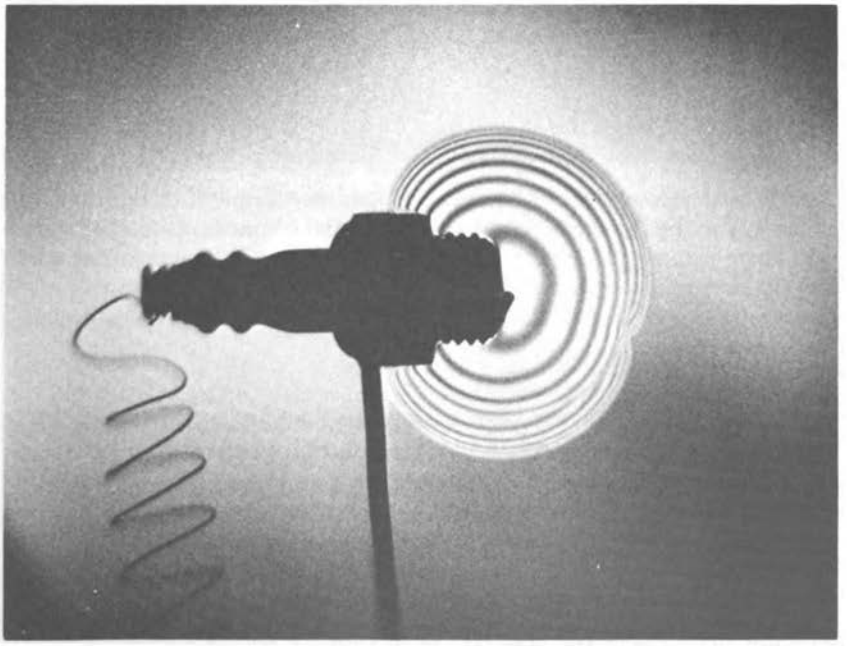

FIG. 2. A double-exposure holographic interferogram of the combustion front in an acetylene-air mixture as it propagates outward from a spark plug ignitor. The scene was illuminated with diffuse-light permitting the phenomena to be studied from many angles.

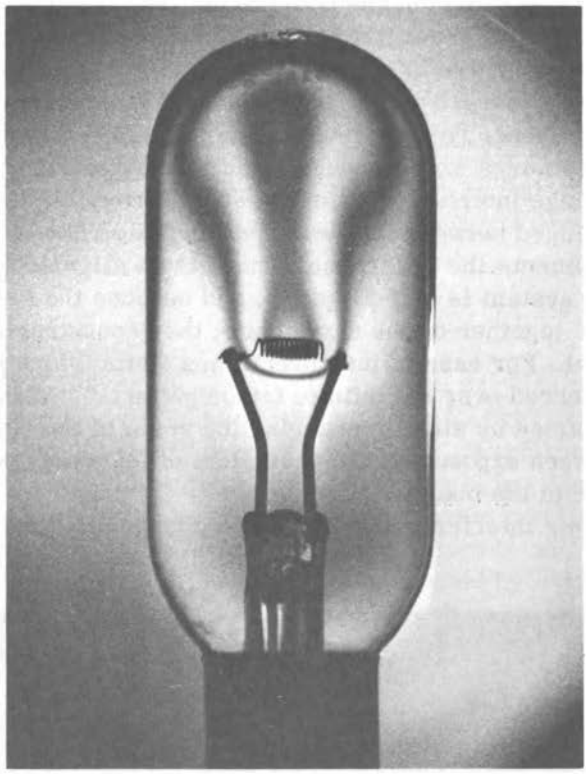

FIG. 3. A double-exposure holographic interferogram of density changes of the gas within the bulb due to heating by the filament. One exposure was made with the filament cold and the second with it heated. Because the interferometer is sensitive only to differences between the two exposures, the presence of the glass bulb does not affect the accuracy of the measurements.

be viewed as they occur, can be carried out by recording only one beam on the hologram.[24, 27] After the hologram is made, it is replaced into the recording apparatus exactly as during the recording step. Upon reillumination with the laser, one views simultaneously the waves from the actual subject (the hologram is transparent) and the synthetically generated waves stored on the hologram. If the subject is perturbed, the two sets of waves will not match and the interference fringes which appear are a quantitative measure of the perturbation. A finite-fringe pattern results if the angle of the reference beam is changed slightly during reconstruction.

Holography provides an extremely sensitive means for examin-

ing the mode structure of a periodically vibrating surface.[29] If a holographic recording (time exposure) is made of the surface while it is vibrating, the reconstructed image will appear dark or bright depending on the amplitude of the vibration. Thus, the image will be covered with a fringe pattern representing the loci of equal amplitudes of vibration, permitting a quantitative analysis of the vibrations.

SPECTROSCOPY

A simple hologram, made by recording the interference pattern produced by two collimated beams of coherent light, is a plane diffraction grating. The manufacture of transmission gratings by this method has been known for some time although their use for spectroscopy is limited by their low diffraction efficiency and scattered light.[30, 31] Their advantage lies in the fact that they can be produced in large sizes at relatively low cost and that ghost images, arising from systematic errors in ruled gratings, do not occur. Advances in the use of newer materials and techniques for the production of holograms may renew interest in the use of holograms for spectroscopy.[32, 33]

A novel spectrograph consisting only of an entrance slit, focusing hologram grating, and film plane (or exit slit) has been constructed. The hologram was produced by photographically recording the interference pattern produced from converging and diverging waves from a HeNe gas laser (6328 Å). The spectrograph clearly separated the Sodium D-lines with a resolution exceeding 1 Å.

ADDITIONAL APPLICATIONS

Although only a few of the more prominent applications of the hologram have been discussed, there are many other applications. A comprehensive bibliography with abstracts of papers on holography published through early 1966 has been prepared and should be of great value to the reader interesting in pursuing the subject further.[34]

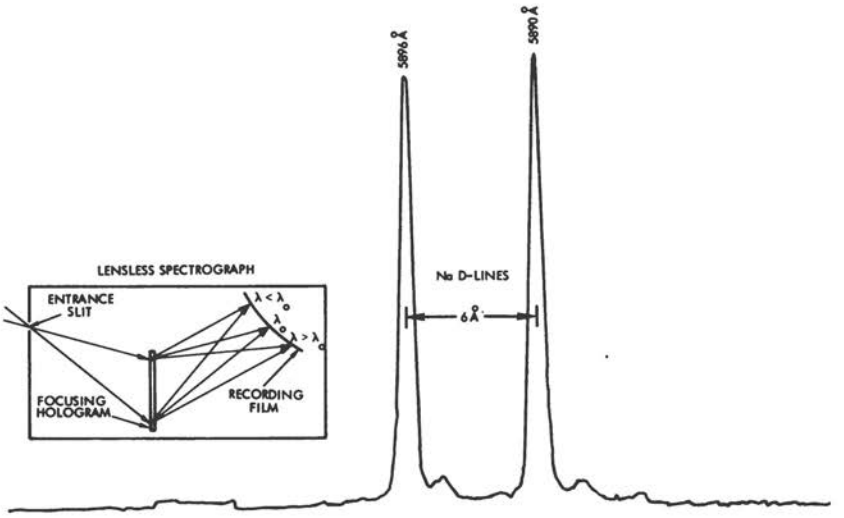

FIG. 4. A microdensitometer trace of a spectrogram produced using a focusing hologram grating as the dispersing element. The hologram was produced by photographically recording the interference between coherent converging and diverging beams from a HeNe gas laser (6328 Å).

REFERENCES

1. E. N. Leith and J. Upatnieks, "Reconstructed Wavefronts and Communication Theory," J. Opt. Soc. Am. 52, pp. 1130-1132 (1962).
2. E. N. Leith and J. Upatnieks, "Wavefront Reconstruction with Continuous Tone Objects," J. Opt. Soc. Am. 53, pp. 1377-1381 (1963).
3. E. N. Leith and J. Upatnieks, "Wavefront Reconstruction with Diffused Illumination and Three-Dimensional Objects," J. Opt. Soc. Am. 54, pp. 1295-1301 (1964).
4. E. G. Ramberg, "The Hologram—Properties and Application," RCA Review XXVII, pp. 469-499 (1966).
5. C. Knox, "Holographic Microscopy as a Technique for Recording Dynamic Microscopic Subjects," Science 153, pp. 989-990 (1966).
6. G. W. Ellis, "Holomicrography: Transformation of Image During Reconstruction a posteriori," Science 154, pp. 1195-1197 (1966).
7. "In-Line Hologram System for Bubble-Chamber Recording," J. Opt. Soc. Am. 57, pp. 275-276 (1967).
8. R. E. Brooks, L. O. Heflinger, R. F. Wuerker, and R. A. Briones, "Holographic Photography of High-Speed Phenomena with Conventional and Q-Switched Ruby Lasers, Appl. Phys. Ltrs. 7, pp. 92-94 (1965).
9. R. E. Brooks, L. O. Heflinger, and R. F. Wuerker, "Pulsed Laser Holograms," IEEE Jour. Quantum Electronics QE-2, pp. 275-279 (1966).

10. C. C. Cutler, "Coherent Light," International Science and Technology, Sept. 1963, pp. 54-63.
11. L. I. Goldfischer, "Autocorrelation Function and Power Spectral Density of Laser-produced Speckle Patterns," J. Opt. Soc. Am. 55, pp. 247-253 (1964).
12. E. N. Leith and J. Upatnieks, "Microscopy by Wavefront Reconstruction," J. Opt. Soc. Am. 55, pp. 569-570; and J. Opt. Soc. Am. 55, pp. 981-986 (1965).
13. R. W. Meier, "Magnification and Third-Order Aberrations in Holography," J. Opt. Soc. Am. 55, pp. 987-992 (1965).
14. E. N. Leith and J. Upatnieks, "Holograms: Their Properties and Uses," S.P.I.E. Journal 4, pp. 3-6 (1965).
15. H. Kogelnik, "Holographic Image Projection Through Inhomogeneous Media," Bell Syst. Tech. Jour., pp. 2451-2455 (Dec. 1965).
16. J. Upatnieks, A. Vander Lugt, and E. Leith, "Correction of Lens Aberrations by Means of Holograms," Applied Optics 5, pp. 589-593.
17. F. B. Rotz and A. A. Friesem, "Holograms with Nonpseudoscopic Real Images," Appl. Phys. Ltrs. 8, pp. 146-148 (1966).
18. D. Gabor, "A New Microscope Principle," Nature 161, p. 777.
19. D. Gabor, "Microscopy by Reconstructed Wavefronts," Proc. Roy. Soc. A197, pp. 454-487 (1949).
20. D. Gabor, "Microscopy by Reconstructed Wavefronts: II," Proc. Phys. Soc. B64, pp. 244-255 (1951).
21. G. B. Parrent, Jr., and G. O. Reynolds, "Resolution Limitations of Lensless Photography," S.P.I.E. Jour. 3, pp. 219-220 (1965).
22. L. O. Heflinger, R. F. Wuerker, and R. E. Brooks, "Holographic Interferometry," J. Appl. Phys. 37, pp. 642-649, Feb. 1966.
23. L. H. Tanner, "Some Applications of Holography in Fluid Mechanics," J. Sci. Instr. 43, pp. 81-83 (1966).
24. J. M. Burch, A. E. Ennos, and R. J. Wilton, "Dual- and Multiple-Beam Interferometry by Wavefront Reconstruction," Nature 209 pp. 1015-1016, March 1966.
25. J. Winckler, "The Mach Interferometer Applied to Studying an Axially Symmetric Supersonic Air Jet," Rev. Sci. Instr. 19, 307-322, May 1948.
26. R. E. Brooks, L. O. Heflinger, and R. F. Wuerker, "Interferometry with a Holographically Reconstructed Comparison Beam," Appl. Phys. Ltrs. 7, pp. 248-249, Nov. 1965.
27. K. A. Haines and B. P. Hildebrand, "Surface Deformation Measurements Using the Wavefront Reconstruction Technique," Appl. Optics 5, pp. 595-602, April 1966.
28. K. A. Haines and B. P. Hildebrand, "Interferometric Measurements on Diffuse Surfaces by Holographic Techniques," IEEE Trans. on Instrumentation IM-15, pp. 118, Dec. 1966.
29. R. L. Powell and K. A. Stetson, "Interferometric Vibration Analysis by Wavefront Reconstruction," J. Opt. Soc. Am. 55, pp. 1593-1598, Dec. 1965.

30. J. M. Burch and D. A. Palmer, "Interferometric Methods for the Production of Large Gratings," Optica Acta 8, pp. 73-80 (1961).
31. N. George and J. W. Matthews, "Holographic Diffraction Gratings," Appl. Phys. Ltrs. 9, pp. 212-215 (1966).
32. A. K. Rigler, "Wavefront Reconstruction by Reflection," J. Opt. Soc. Am. 55, pp. 1963 (1965).
33. H. Kogelnik, "Response and Efficiency of Five Hologram Types," XVII Modern Optics, Polytechnic Press (1967), pp. 605-617.
34. R. P. Chambers and J. S. Courtney-Pratt, "Bibliography on Holograms," Jour. S.M.P.T.E. 75, pp. 373-435 (1966); pp. 759-809 (1966); 76, pp. 393-394 (1967).

IMAGE RESTORATION AND ENHANCEMENT

John L. Brown, Chairman

THE ELASTIC SURFACE TRANSFORMATION*

Robert J. Hall and William F. Dossett

Concept of Display Processing

The principal restriction on the development of better displays for sonar, radar, and photographic systems is often attributed to the sensing and signal processing equipment. However, it seems more likely that our limited understanding and use of the human's perceptual processes is the major restriction. For example, studies in interference processing and display of active sonar signals[3] have shown that we are losing useful information because we have not translated it into a format that the human observer can perceive.

Psychophysical research has solved many display difficulties when detection of presence or absence of the signal has been the problem. But at present very little is known about the best procedures for manipulating and displaying the complex sets of stimuli produced by many of our current sensor systems. These deficiencies are even more apparent in systems where recognition operations are necessary to recover structural information about the target.

Recent studies dealing with the perception and selection of stimuli in complex tasks[4, 5] have emphasized exploration, search, and selection aspects in perceiving, rather than the classical

*This concept was first developed by R. J. Hall and James W. Miller[1] in a study of visual display enhancement and to our knowledge the only precursor is a related implementation by Menzel[2] who imprinted a microscopic image on an elastomer and stretched the imprint to amplify the microstructure of the image.

measures of receptor sensitivity based on the physical energy variation of the stimulus. They have suggested that information from the physical stimulation derives crucially from on-going patterns of change in the observer's natural environment. The observer uses many sensing operations, such as fixating first and then shifting his vantage point, thereby introducing changes that aid in perceiving the complex stimulus situation. These observer-controlled manipulations of change can be appropriately described as a type of perceptual processing transformation on the signal.

These studies coupled with our recent investigations suggest that it is desirable to develop signal processing and display procedures that will permit the observer to manipulate and transform the sensor inputs at the display. It is for this reason that we emphasize the term display processing as an entity operation, rather regarding these transformations in terms of separable functions indicated by the phrase "processing and display." Our point of emphasis here is that we are dealing with an inseparable, highly integrated process.

Elastic Surface Transformation

The elastic surface transformation (EST) is an attempt to develop a display processing technique that will permit the observer to manipulate and enhance marginal signals. In much of our work the signals are imagery of low contrast that has been degraded by serious edge-spreading problems.

The EST process, as we have developed it to date, consists of photographically transferring the information from some sensor to a uniformly prestretched surface. The surface is then relaxed in a continuous fashion, producing a uniform spatial compression. During compression the element size remains constant. The shrinking changes the spatial dispersion, which increases the density of elements per unit area. This increases the image contrast and enhances the dim edges and gradients. Optical magnification maintains the enhanced image at its original size.

It should be emphasized, that, in the EST, the constancy of element size is maintained over a wide range of spatial compression. Changes in the dispersion change spatial relationships (signal pattern) by increasing the density and apparent contrast of the signal. This effect, which is that of a nonlinear transformation on the ratio of signal to noise, enhances marginal signals.

For example, when dim, unclear signals are placed on a prestretched surface and the surface is collapsed uniformly, the density and the sharpness of the images can be increased to the point where these images become clearly visible.

If we take a dim, symmetrical gradient on a prestretched surface, the mean square of the spatial distances separating the elements dispersed on the surface is equivalent to the variance of a statistical distribution. In the transformation process, the variance of this gradient is reduced about the mean value of the gradient (signal). Gains in signal-to-noise ratios are a result of such processing. It is particularly interesting to note that no additional data acquisition is involved, as in the case of multiple superpositioning of imagery.

The manner in which the EST affects edge-spread and contrast does not depend on some fixed relationship between these parameters, but rather on the initial flux density distribution of the energy which was photographically activated. The processing enables the observer to see those elements that constitute the image and to integrate them into patterns at various stages of shrinking. In other words, the elastic surface transformation does not add information but integrates and conveys more of it than conventional recording and display techniques.

IMPLEMENTATION

The major problem in the present research effort has been the development of a reliable, high-resolution, photographic transfer to the elastic surface. Losses that occur during the transfer of the original photo to the elastic surface must be minimal because marginal images and signals are particularly sensitive to any information losses.

A silver halide emulsion can be placed on the prestretched substrate and photographically exposed directly on the elastic surface. The procedures for developing and fixing are similar to those employed with conventional black and white photography. The present silver halide emulsions will withstand a shrink factor of approximately 15 to 1 without serious distortion.

The objective is to stretch and shrink the elastic surface without distortion and to control the rate and amount of shrinking so that it can be stopped and the image observed at any particular point in the transformation. At present the most satisfactory technique providing uniform surface stretching can be produced

by an inflatable spherical elastomer membrane. With the spherically inflatable elastomer the stretch is relatively uniform, and shrinking of the photographic information can be carefully controlled.

EXPERIMENTAL EXAMPLES

The objective of the current research was the enhancement of photographic information and sonar grams when the acoustic information was originally recorded on electro-sensitive paper.

Photographic Examples

The photographic information consisted of an outdoor scene. This scene contained both high-quality and marginal images. The marginal images consisted of distant objects, such as mountains, which were strongly attenuated by atmospheric haze. These marginal images were characterized by broad gradients and very low contrast.

Sonar Gram Examples

The sonar gram examples used in the present study contain dark traces on a gray background. The faint signals in the sonar gram are characteristic of weak signals embedded in a noisy background. For a given gram there may be a wide variation in signal strength, and some of the signals may be superimposed on one another. When viewing a sonar gram, the observer can often enhance the dim trace by tilting the gram and observing it at an inclined angle. This tilting of the gram suggested that it might be enhanced by a spatial compression similar to that produced by the elastic surface transformation.

RESULTS*

The present investigation has concentrated on developing a reliable research procedure for implementing the elastic surface

*For a more detailed report see Hall, Dossett, and Kelley.[6]

transformation. All examples exhibited were produced by means of the black and white silver halide implementation.

The method used to evaluate the image enhancement produced by the EST was concerned with the enhancement of the information that had been transferred to the prestretched surface. It compared the photographic information on the stretched surface with the information on the shrunken surface.

Enhancement of Marginal Photographic Images

The following example was produced by lowering the contrast and density of the information when it was transferred from a high-quality negative to the elastic surface. The first photo in Figure 1 is marginal or degraded imagery on the stretched surface. The second photo is the same image after the surface was allowed to shrink.

A comparison of photos 1 and 2 in Figure 1 shows that shrinking the surface has enhanced images that were invisible on the stretched surface. The exposed emulsion and the photographic information which it contained were the same. The only changes that occurred between the first and second photograph in Figure 1 are the spatial compression and optical magnification used to return the transformed picture to its original size. This example shows fairly dramatic enhancement, which suggests that the present technique can recover images well below the observer's threshold.

The Enhancement of Marginal Sonar Gram Signals

Sonar grams were transferred to the elastic surface by photographing the original gram and exposing a negative to an emulsion on a prestretched substrate. The first photo in Figure 2 shows the gram on the stretched substrate. The second shows the gram after the substrate has been allowed to contract. In this example, the photographic scales are slightly different. However, the numbered arrows indicate corresponding traces.

DISCUSSION

The greater part of the effort in this study has been devoted to obtaining suitable implementation. As a result, the number and

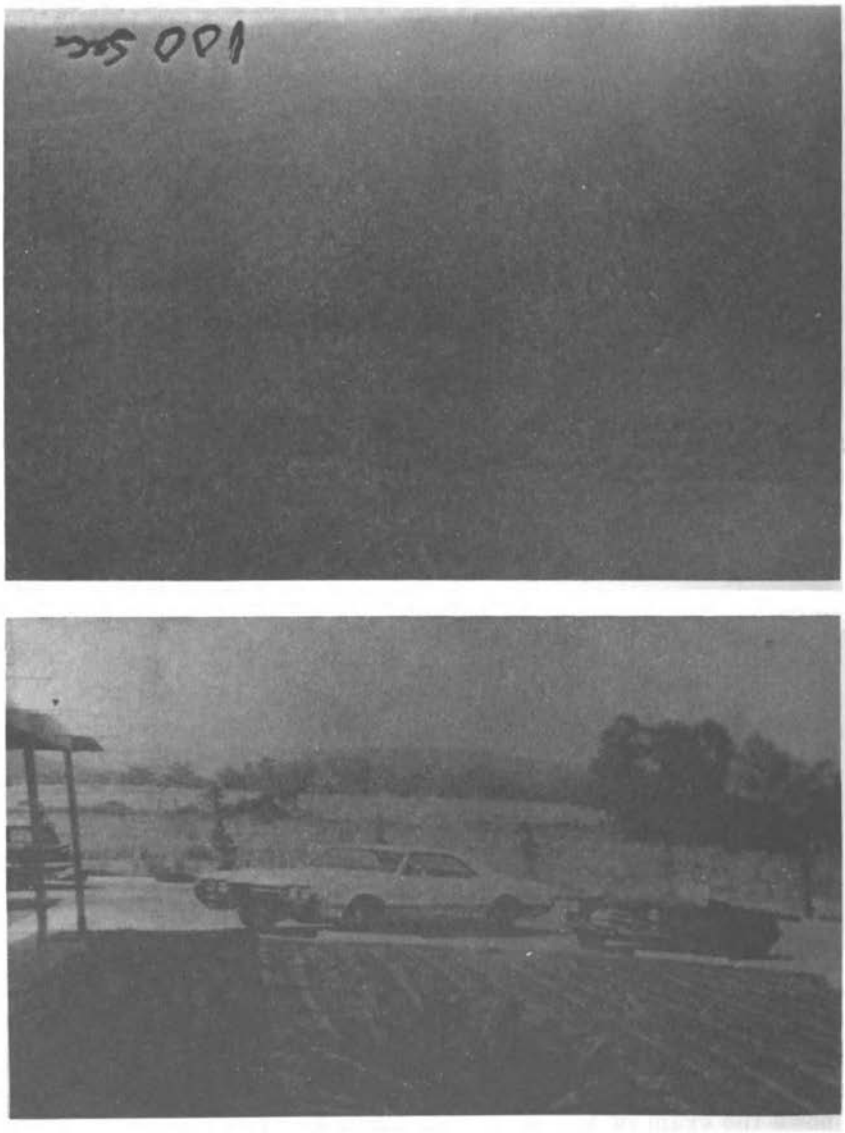

FIG. 1. At top is a marginal or degraded photograph on a prestretched surface. Below is the identical photograph after the surface has been allowed to shrink. The differences between the top and bottom photograph illustrate how the contrast and visibility of the marginal imagery can be enhanced by the EST.

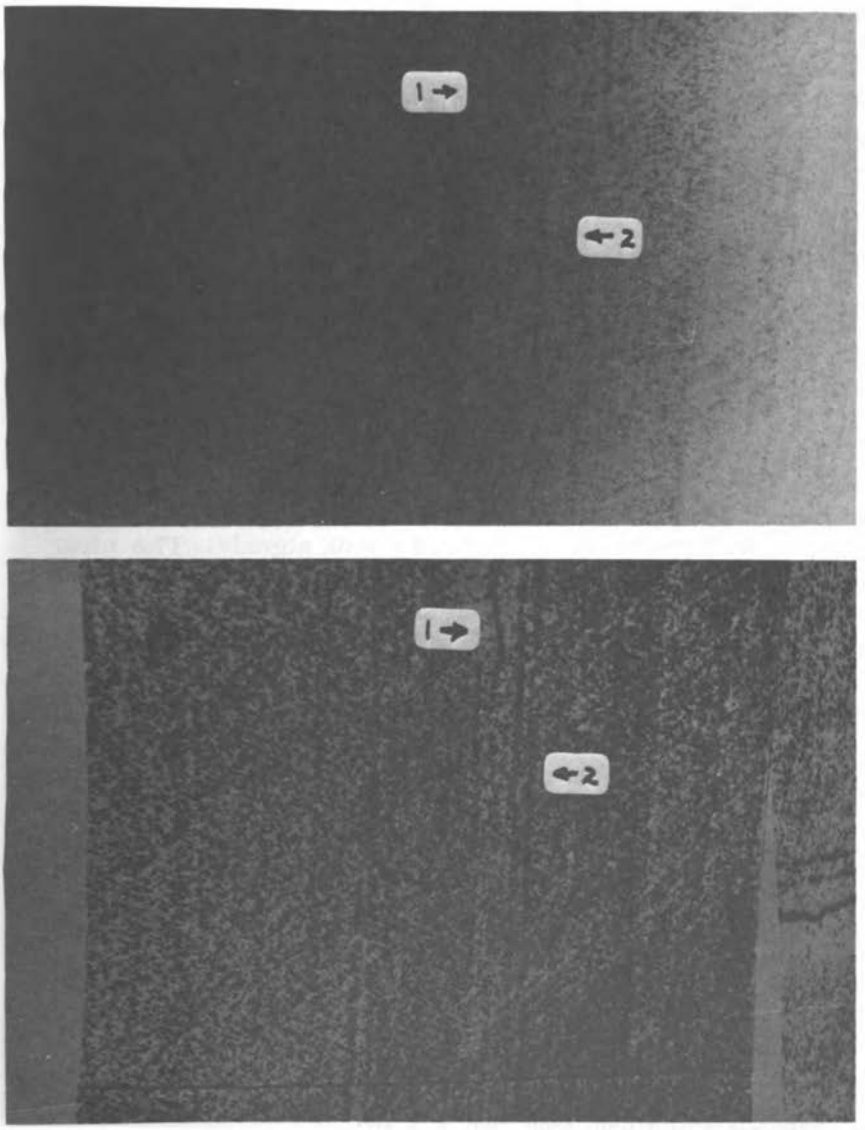

FIG. 2. Above is a sonar gram that was transferred to a prestretched elastic surface (top photograph). Then the gram was shrunk and optically magnified to its original size (bottom photograph). The arrows identify identical traces in each photograph.

variety of photographic images and sonar signals that we have processed with EST has been limited. However, the enhancement of marginal photographic images produced by placing a dim or underexposed image on the prestretched surface has shown that it is possible to enhance images that could not be seen prior to the transformation (see Figures 1 and 2). Second, attempts to enhance these dim images (prior to shrinking) by photographing and printing the imagery at different exposures were unsuccessful. Although our best results are from marginal images that were produced by degrading a high-quality photograph, the amount of the information recovered by the EST is convincing evidence that it can become a useful procedure for enhancing marginal images.

The sonar gram examples were originally recorded on an electro-sensitive paper. The limited dynamic range of this type of recording and the saturation produced by high amplitude restricts the EST's capacity to amplify weak signals. The ideal type of recording would be one where the density of uniform elements is modulated by the signal strength. Element modulation is different from the conventional amplitude recording techniques employed in sonar grams and many CRT displays because (a) the elements are of uniform size and (b) signal strength is a function of the density of these elements and their spatial distribution. Our results suggest that element modulation coupled with direct recording on a prestretched surface would be a powerful signal processing and display technique.

REFERENCES

1. Hall, R. J., Miller, J. W., Musselman, D., Earl, R., and Detambel, M. H. A study of visual display enhancement and techniques of color filtering. Tech. Doc. Report No. ESD-TDR-63-635, L. G. Hanscom Field, Bedford, Massachusetts, December 1963.
2. Menzel, E. Gesteigerte laterale Auflosung mit Lichtmikroskopen. Naturwissenschaften, 1959, Heft 3, S. 105.
3. Hall, R. J., Dossett, W. F., and Kiyan, N. R. Interference processing and display of active sonar signals (U). HFR tech. Rep. 771-1, 1967.
4. Dember, W. N., and Earl, R. W. Analysis of exploratory manipulatory and curiosity behavior. Psychological Review, 1957, 64, 91-96.
5. Gibson, James J. The Senses Considered as Perceptual Systems. Boston: Houghton Mifflin Company, 1966.
6. Hall, R. J., Dossett, W. F., and Kelley, G. R. A preliminary evaluation of the elastic surface transformation for processing and displaying passive sonar and photographic information. HFR tech. Rep. 775-1, 1967.

SUPPLEMENTARY REFERENCES

Davson, H. (Ed.) The Eye. Vol. 2. The Visual Process. New York: Academic Press, 1962.

Earl, R. W. Problem solving and motor skill behaviors under conditions of free choice. Unpublished doctor's dissertation, University of Michigan, 1957.

Gibson, James W. The usefulness of sensitivity. American Psychologist, 1963, 18, 1-15.

James, T. H. (Ed.) The Theory of the Photographic Process. (3rd ed.) New York: Macmillan Company, 1966.

IMAGE PROCESSING AS IT RELATES TO THE HUMAN SYSTEM

James L. Harris, Sr.

The term "image processing" as used in this paper refers to operations performed on a degraded image to correct or compensate the image for its defects. The result of such processing is an image which is called a restored image.

Figure 1 shows some examples of image processing. In the left-hand column are images degraded by various means as labeled. The discrete nature of these images results from a discrete-step photoelectric scanning of the original photographic films. The pictures in Figure 1 are photographs of a cathode-ray tube display of the data derived from the scanning operation.

Each of the degradations shown are isoplanatic degradations meaning that over the extent of the image every point in object space maps into the image plane as a characteristic point spread function located at the position in the image plane corresponding to the geometric projection of the object point and having a peak value directly proportional to the intensity of the object point. The degraded image is the linear sum of the contribution from all object space points. Mathematically stated, the degraded image is found by a linear, homogeneous convolution of the object and the point spread function. The point spread function, therefore, uniquely defines the degradation process.

The convolution integral equation contains three terms: the degraded image, the point spread function, and the ideal geometric image. If the degraded image is known (by measurement) and the point spread function is known a priori, then, under a certain class of conditions, the equation can be solved to determine the ideal geometric image, i.e., the restored image. Fourier transformations constitute a powerful tool for the solution of this type

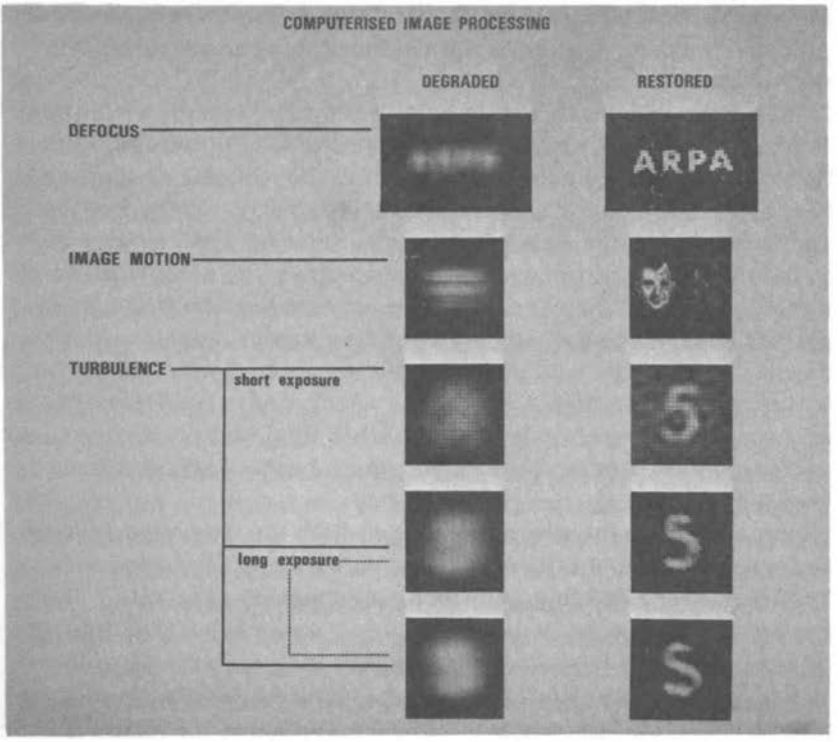

FIG. 1.

of convolution integral and were used to obtain the restored images shown in the right-hand column of Figure 1. The details of the mathematics of the operations are reported elsewhere in the literature[1, 2] and will be repeated in this paper. The actual operations were performed on a CDC-3600 computer at the UCSD Computer Center, and the restored images shown are photographs of a cathode-ray tube display of the computer output data.

Unfortunately, the image restoration operation is not as straightforward as the explanation given may imply. The difficulty arises from noise inherent in any measurement of an image; i.e., the degraded image can only be defined within certain precision limitations. In the case of photographic film, the noise is due to the film granularity. If, with a noisy measurement of the degraded image, an attempt is made to get an exact solution to the convolution integral equation, the disappointing result is a restored image which is almost totally noise. It becomes clear that image processing must settle for approximate solutions to

the convolution integral equations in order to achieve some compromise between quality of the restored image and noise level in the restored image.

Communications theory has dealt for many years with questions of optimum filtering for signals imbedded in noise. These theorems generally result from defining the signal, the noise, the characteristics of the system, and the objective of the system, and then solving the sets of equations relating these parameters to determine the filter which will maximize the achievement of the objective. In the case of image processing, the most important part of the system is the human visual system. Analytic descriptions of the human visual system adequate to determining optimum filtering do not exist at present, and a good part of image processing research today deals with a trial and error approach to finding the filtering operations which best satisfy the human visual system.

In comparing the restored images with the degraded images, it seems that more information is available to the human visual system. Where did this additional information originate? To answer this question, it is important to understand that information can be measured at various points in a system. Measured at the photographic film, there is a specific quantity of information in the degraded, noisy image. No amount of computer processing can increase that information level. The information perceived by the human visual system cannot exceed that which is on the film and will, in general, be less than that which is on the film. The ratio of the information perceived by the human visual system to that which existed on the display might be termed the informational efficiency of the display. Image processing deals with attempting to process and display information in such a way as to maximize the informational efficiency of the display. When the observer is able to extract all information from the display, 100% efficiency has been achieved and no further image processing is needed.

Since image processing deals with attempting to match the display of information to the information transmission characteristics of the human visual system, image processing research creates a new demand for obtaining more detailed knowledge of the workings of the human visual system in order that optimum image processing can be achieved. The modern high speed digital computer makes it possible to display information in almost any desired fashion. The question now becomes, What does the human visual system want to see in order to maximize performance? In finding

experimental answers to that question, it is possible that image processing research may provide a tool to add to those currently available for studying the characteristics of the human visual system.

It has previously been mentioned that, to obtain an image restoration, it is necessary to be satisfied with approximate solutions to the convolution integral. One important form of approximation amounts to limiting the number of terms in the two-dimensional Fourier series representation of an object. It becomes important to the image processor, therefore, to know how many Fourier series terms are required in order to convey a given picture to the human visual system. Figure 2 shows a series of pictures designed to answer that question for one specific recognition task. Elementary texts dealing with Fourier series analysis invariably show examples of the approximation of some elementary function with a Fourier series demonstrating that, as more and more terms in the series are included, the approximation of the function becomes better and better. Figure 2 is the two-dimensional equivalent of this type of example. The array of dots alongside the image identifies each of the two-dimensional Fourier series terms. For example, the dot, 1 (horizontal), 3 (vertical) is a two-dimensional luminance sinusoid which makes 1 cycle horizontally in the format of the picture while making 3 cycles vertically. The solid quarter of a circle having an origin at the (0, 0) dot (the average luminance or constant term) encloses those terms of the series which were used in the representation of the image. For example, the first pair of images used only three terms in the series, i.e., (0, 0), (1, 0), and (0, 1). By inspecting set of images in Figure 2, it is possible to obtain an estimate as to the number of Fourier terms which must be included in order to clearly distinguish between the two objects. These answers are essential to successful image processing.

Undoubtedly, the human visual system is capable of some image processing. It is interesting to speculate as to what properties of the human visual system limit the extent to which unaided image restoration is possible. As a guide to this speculation, the factors which limit the computer in performing restoration operations can be considered. The two primary limitations for the computer are the precision of the image measurements and the complexity of the operations required.

With respect to the human visual system and the question of precision of measurement, the author admits to an inadequate

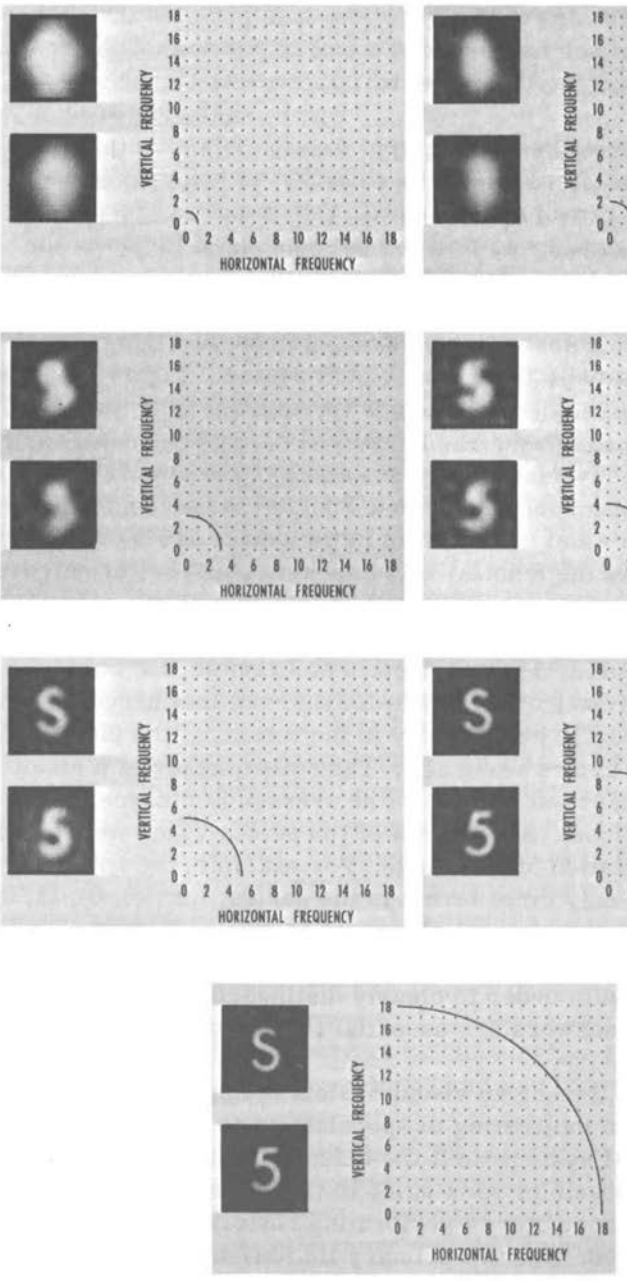
VERTICAL FREQUENCY
HORIZONTAL FREQUENCY
0 2 4 6 8 10 12 14 16 18
S
5

FIG. 2.

knowledge of the kinds and quantities of pertinent experimental visual data which have been acquired. Related to determining the direct limitations in image restoration capability due to precision of measurement, it would seem that the type of experimental data which is needed is that of a thorough study of the errors in making relative judgment of two luminance patches, with the distance between patches and the texture of the luminance map between the patches as important variables of the experiment. If such data exist or could be collected, it would be a matter of some interest to impose these same precision limitations on the digital computer and determine the extent to which this limits computer restoration capability.

A degraded image can be thought of as an encoded image; i.e., the degraded image is formed by operating on the undegraded image in a specified mathematical manner. Most image degradations are very poor encoding in that the information content is reduced in the process of encoding. However, there are some exceptions to this rule. It is interesting to consider these exceptions because extremely good restorations can be achieved with very modest precision of measurement so that they tend to become examples which stress the complexity of the operations rather than precision. Figure 3 is an example of such an encoding.

The image of Figure 3 is the image of an alpha-numeric character in which the phase of each of the spatial frequencies has been shifted in accordance with a table of phase shifts derived from a table of random digits. Since the phase of the noise is also random for each spatial frequency, the signal-to-noise ratio of the image is not altered by this operation, and the information content is not reduced, provided that the table of phase shifts exists as a priori information. In Figure 4 we see the table of phase shifts and the degraded image. With these two pieces of information available, the CDC-3600 can produce a perfect restoration in approximately four seconds time. It would seem that this might be a case in which the computational complexity exceeds the capability of the human visual system. Figure 5 shows the restored image. The discrete nature of the cathode-ray tube display system is again apparent and, in fact, somewhat detrimental in interpreting the image. Under some conditions, defocusing of the cathode-ray tube images, although an information destroying process, actually enhances the interpretability of the image as perceived by the human visual system. This fact in

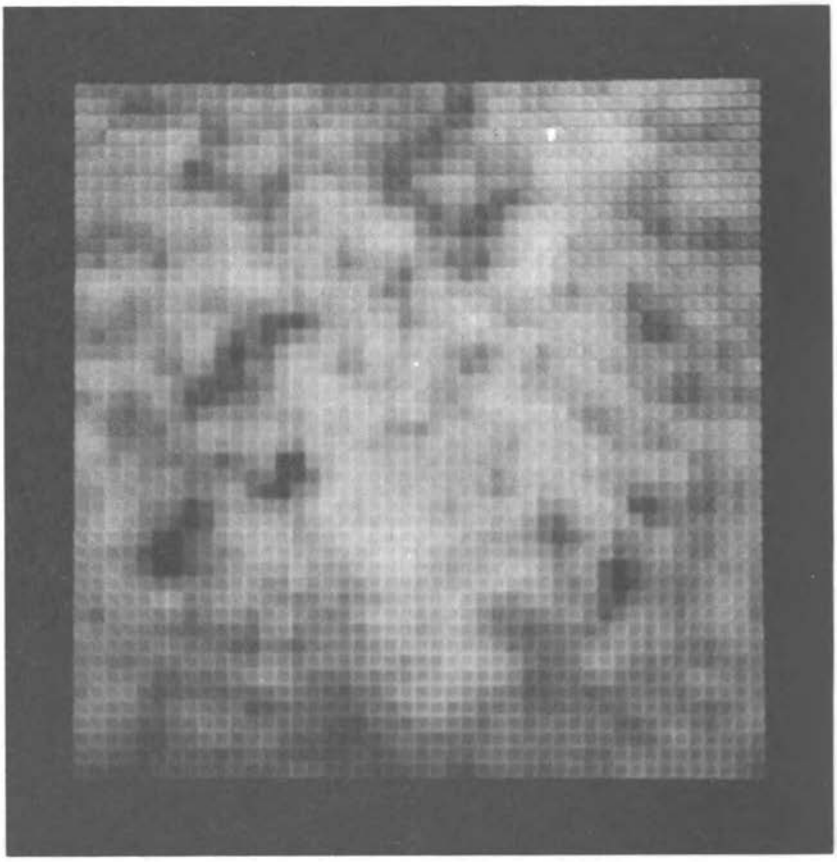

FIG. 3.

itself has interesting bearing on the properties of the eye-brain combination. A defocused version of Figure 5 is shown in Figure 6.

Another less exotic information preserving encoding is shown in Figure 7. Here the elements of Figure 5 have been redistributed according to a table of random digits. This very simple type of coding is easily unscrambled by the computer in a few milliseconds and precision requirements are quite modest, but it is an extremely difficult visual task.

In summary, image processing, as defined in this paper, exists for the sole purpose of attempting to increase the informational efficiency of the display system. Present optics and

PHASE COMING FOR FIRST QUADRANT

	0	1	2	3	4	5	6	7	8	9	10	11
0	-6.1636-01	-9.4481-02	-9.6115-01	6.9169-01	-6.4316-02	5.8239-01	-7.1748-01	-5.7116-01	9.1146-01	1.2238-01	-6.7586-02	-2.8476-01
1	6.1596-01	-1.1496-01	-3.9196-01	8.8726-01	-9.8336-01	1.4316-01	4.1268-01	7.9468-01	6.8656-01	-2.6795-02	-9.5356-01	8.8426-01
2	5.1766-01	-6.8436-01	3.9986-01	1.7456-01	-6.3636-01	-6.8126-02	-7.4746-01	-2.8336-01	5.3606-01	8.4176-01	5.7576-01	2.1766-01
3	9.5856-01	9.2556-01	7.4176-01	-1.5736-02	6.8676-01	6.5866-01	-1.8796-01	7.8868-02	-3.9856-01	-6.2466-01	7.6346-01	-3.1966-01
4	5.7176-01	-2.0996-01	-1.5576-01	6.9426-01	-6.8686-01	-5.8776-01	-9.5676-01	-2.8766-01	-6.4316-01	7.2626-01	1.9176-01	-6.9536-01
5	-7.7736-01	-6.1466-01	7.9916-01	2.3606-01	7.9676-01	-1.3686-01	-1.7846-01	-6.1976-01	9.2776-01	9.4796-01	5.1716-01	-1.6196-01
6	-1.5366-01	6.5466-01	-1.7796-01	6.8486-01	9.8466-01	1.6856-01	-9.7176-02	6.1336-01	-8.2846-01	9.7316-01	-9.7386-02	3.7186-01
7	-6.1536-01	-8.6486-01	9.1616-01	8.4796-01	-7.7466-01	9.5486-01	-6.7796-01	5.1996-01	6.6936-01	6.6986-01	5.9216-01	5.3416-01
8	7.7376-02	5.2696-01	-1.7846-01	-5.3726-01	-8.9936-01	-6.8526-01	-9.7376-01	-5.6136-01	9.7746-01	7.7996-01	2.5466-02	6.8486-01
9	-9.5476-01	3.3596-01	-8.8766-01	9.9486-01	9.9586-01	-6.6966-01	-1.5986-01	7.1566-02	-8.6156-01	9.5436-01	-4.7576-02	9.5596-01
10	-7.5616-01	-7.8426-01	8.6476-01	5.8636-01	-7.2946-01	3.1146-01	-4.9296-01	-4.8866-01	9.7156-01	-6.9716-01	6.7566-02	9.1186-01
11	-1.9356-01	-7.7946-01	6.9446-01	-1.1076-01	-7.7356-01	-4.1776-01	3.4256-01	-5.5446-01	-7.7426-01	5.9916-01	9.4496-01	1.6446-01
12	5.2866-01	-2.6456-01	-5.7716-01	-9.8746-01	6.8346-01	1.0796-01	6.1636-01	9.3786-02	-4.9886-01	-1.3706-01	-1.4976-01	8.1936-01
13	6.5796-01	2.1756-01	9.7936-01	7.2466-01	-8.8866-01	-4.7726-01	1.9876-01	-1.6846-01	6.7256-01	7.6866-01	5.7356-01	-6.8176-01
14	9.9616-01	-8.3706-01	7.5216-01	9.9866-02	6.7176-01	-1.8536-02	-1.1886-01	9.4816-02	-4.8836-01	-4.7876-01	5.3796-01	-9.8706-01
15	1.9276-01	-9.9486-01	-8.9546-01	7.0146-01	6.7956-01	1.3416-01	9.9296-01	6.3416-01	9.9826-01	7.8226-01	-6.7856-01	1.5936-01
16	-7.9376-02	8.8996-01	5.7446-02	8.8286-01	-6.6856-01	-8.7216-01	-1.5776-02	-9.9986-01	5.6486-02	-4.5316-01	1.6196-01	-1.0136-01
17	-9.5476-01	6.9716-01	6.7496-01	-9.8726-01	-9.5376-01	6.1856-01	1.7106-01	4.4586-01	9.8196-01	-3.7696-01	7.4886-01	-7.1426-01
18	6.7546-02	-2.0726-02	9.1766-01	6.8126-01	7.6196-01	7.4096-01	3.3756-01	-8.4176-02	3.7656-01	-8.6956-01	9.7366-01	-5.3676-01
19	-7.7366-01	-1.6416-02	-1.7596-01	8.5756-01	-6.7326-01	1.7536-01	7.4866-01	-9.3666-01	-8.1836-01	8.5966-01	-7.4166-01	8.9926-02
20	-1.1406-01	-3.3656-01	-8.5196-01	3.9156-01	9.0196-01	3.8716-02	1.7316-01	3.7816-02	-3.2086-01	-5.8806-01	1.4696-01	-6.8986-01
21	9.9176-01	8.6476-01	8.3246-02	-5.6736-01	1.8646-01	-4.3106-01	-6.8816-01	-5.1356-01	6.2936-01	-9.6486-01	6.7266-01	-2.2486-01
22	-7.1176-01	-8.9486-01	-7.4486-01	-5.6916-01	-8.8886-01	1.9396-01	6.6396-01	5.8196-01	-8.8216-01	-3.3956-01	5.7466-02	6.1486-01

	12	13	14	15	16	17	18	19	20	21	22
0	-6.4276-01	-6.2726-01	7.9416-01	5.9166-01	8.4676-01	-1.9386-02	8.9406-01	8.7976-01	-6.7466-02	3.9776-01	-8.1846-01
1	6.9276-01	6.2196-01	5.6946-01	-6.3366-02	-1.9986-02	5.3466-01	7.8126-03	-7.8986-01	-7.7166-01	-8.4176-01	6.2316-01
2	5.7116-01	-9.6396-01	-8.3946-01	6.8726-01	7.1996-01	-3.5866-01	-8.6466-01	-2.6976-01	1.7366-03	-3.5196-01	6.2696-01
3	5.9276-01	1.2636-01	-5.9866-03	-1.6126-01	5.1266-01	-6.7136-02	4.5106-01	-6.6176-01	-7.1396-01	-9.4136-02	-6.5116-01
4	7.4646-02	7.7866-01	3.6256-01	1.5926-01	1.7826-01	6.1606-01	9.9856-01	-7.1796-01	7.7446-01	-7.1976-01	-1.8746-01
5	-5.5766-01	6.6656-01	-7.9836-02	9.7966-02	6.8856-01	-7.1656-01	5.7496-01	-6.6626-01	-9.9836-01	-5.8516-01	-9.7836-01
6	-7.6996-01	-6.2696-02	3.9566-01	8.8496-01	3.6676-01	1.0756-01	2.8476-01	9.6176-01	-7.3226-01	5.1746-01	5.9996-01
7	6.0916-01	-7.9706-01	-9.9466-01	-7.0516-02	5.4496-01	-2.7656-01	-6.2236-01	-2.6416-01	6.9246-01	-8.6956-01	6.9276-01
8	-7.1156-01	-2.6996-01	7.1146-01	-3.7856-01	5.5796-01	-7.0366-01	-6.5656-01	-1.1496-01	6.1706-01	5.2186-01	-1.8216-01
9	-6.9246-01	-9.6416-01	6.9186-01	5.2866-01	5.5946-01	4.7826-01	6.5306-01	1.7726-02	-9.4016-02	2.9616-01	-5.5406-01
10	-4.9946-01	-6.4486-01	6.9676-01	-1.9376-01	-6.7466-01	-6.0486-01	-8.4286-01	-7.4726-01	9.7326-01	9.8696-01	9.5416-01
11	9.9176-01	-7.2946-01	-8.7416-01	-9.8196-01	6.4976-01	7.2696-01	-9.9866-01	-7.1136-01	-3.1976-02	-8.7726-02	9.7436-01
12	7.5146-01	7.7586-02	-5.7126-01	9.5476-01	-6.6786-02	-4.0316-01	8.9216-01	2.5926-01	-7.4646-01	-3.2946-01	9.9566-01
13	-1.5876-01	-8.4376-01	-1.4486-01	8.3856-01	-9.9756-01	-3.7726-01	-1.7726-01	8.1466-01	3.7956-02	-8.1916-01	-7.9296-01
14	5.6996-01	8.8596-01	9.9546-01	-9.3836-01	7.2936-01	-9.8176-01	-6.9476-01	1.8966-01	-6.9256-01	-7.8466-01	-9.9136-01
15	9.7136-02	-1.9376-01	6.7816-01	9.2796-01	9.7976-01	5.2526-01	3.8936-01	-2.2926-01	-6.4856-01	-1.9726-01	6.8176-01
16	4.2976-02	-8.1386-01	2.2336-01	9.4786-01	3.6966-02	-9.8716-01	6.1436-01	-1.9746-02	-9.1376-01	3.2766-01	9.8476-01
17	8.8596-01	5.9786-01	-4.8836-01	9.7466-01	9.8516-01	4.6936-01	-3.4886-01	-9.8136-01	9.9636-01	8.9466-01	1.4846-02
18	3.6686-01	1.8926-01	9.7996-01	1.7316-01	-9.7266-01	7.8996-01	-9.2396-01	-5.1986-01	-3.1986-01	2.9746-01	6.6496-01
19	5.3446-01	7.8686-01	-6.9946-01	9.6396-01	-1.9166-01	7.1856-01	-9.2776-01	7.0966-01	7.4816-01	3.6476-01	9.0346-01
20	7.3966-01	8.4696-01	9.2296-01	-8.6286-01	6.4696-01	-6.2956-01	2.3826-01	6.2726-01	4.3316-01	4.7976-02	3.8696-01
21	-9.6856-01	8.3446-01	-9.9496-01	-6.8496-01	4.3956-01	6.3796-01	-9.9656-01	-7.5446-01	3.7986-01	8.9726-01	-6.7466-01
22	5.9926-01	8.8826-01	-7.5176-01	8.9476-01	-1.6916-01	6.8696-02	-7.6236-01	-2.8886-01	9.9886-01	-9.4516-01	-9.9276-01

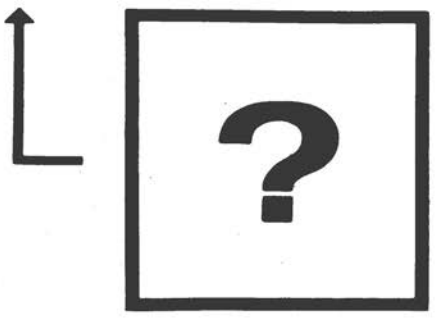

FIG. 4.

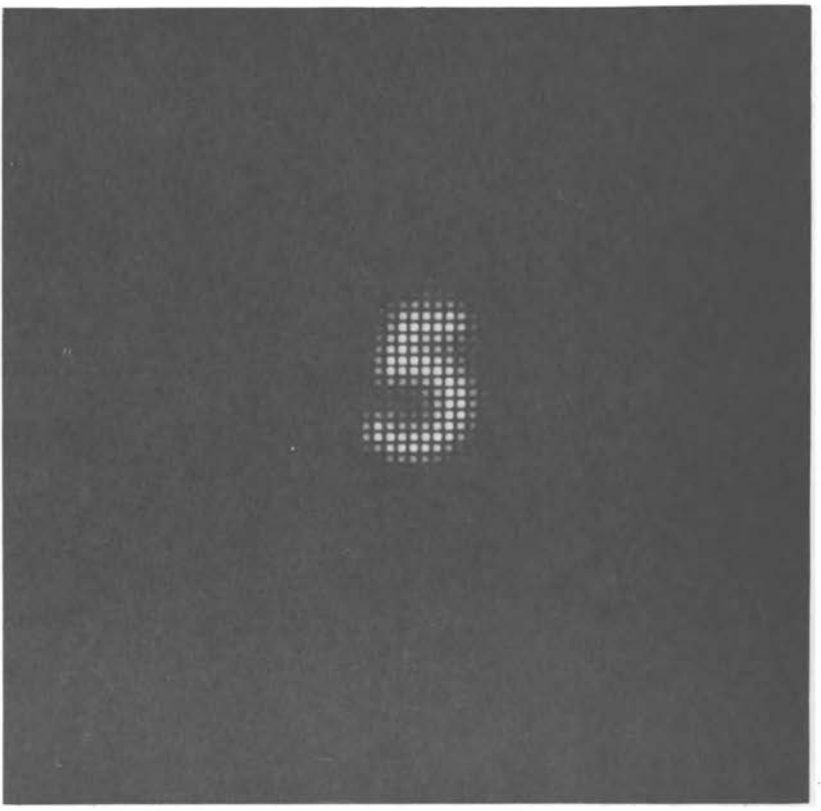

FIG. 5.

computer technology make it possible to display information in almost any fashion desired. The real question becomes, What does the human visual system want to see? Image processing, therefore, adds a new demand for better understanding of the human visual system, its characteristics, its capabilities, and its limitations. Because image processing is built around deficiencies of the human visual system, it also becomes a tool for research directed toward understanding the human visual system and should be added to the collection of such tools as already exists.

REFERENCES

1. J. L. Harris, Sr., J. Opt. Soc. Am. 56, 569 (1966).
2. B. L. McGlamery, J. Opt. Soc. Am. 57, 293 (1967).

FIG. 6.

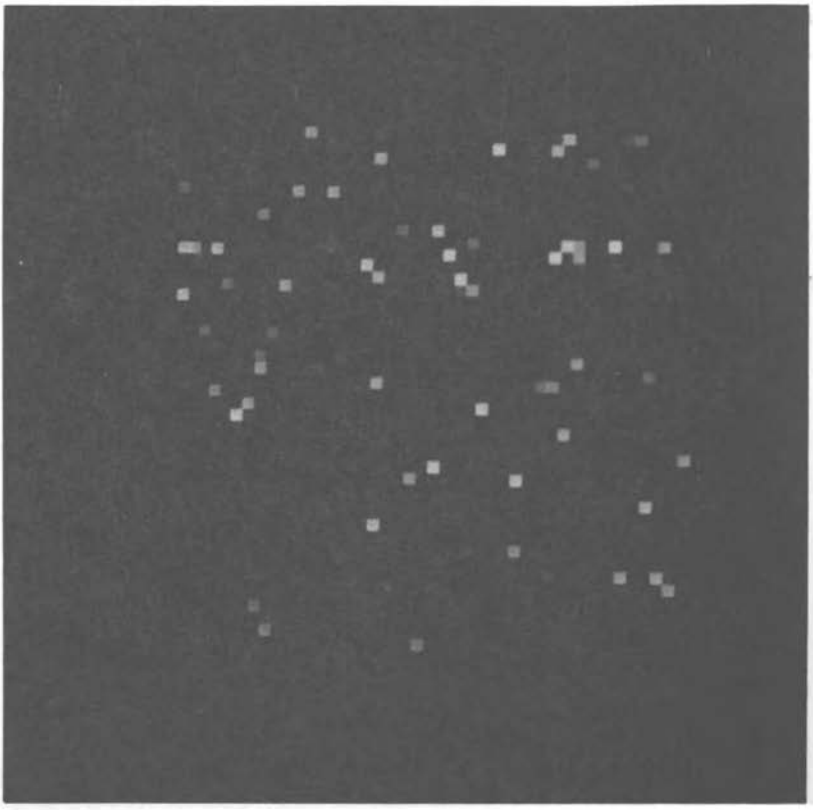

FIG. 7.

EXPERIMENTS IN SPATIAL FILTERING

A. W. Lohmann*

A coherent optical image forming system can be considered to be a linear filter. The complex amplitude, u(x, y), of the light in the object plane acts as input, and the complex amplitude, v(x, y), of the image is the output. Spatial filtering means that the spectrum of spatial frequencies, $\tilde{u}(\nu_x, \nu_y)$, of the object is multiplied by the filter function, $F(\nu_x, \nu_y)$, and the product is the spatial frequency spectrum, $\tilde{v}(\nu_x, \nu_y) = \tilde{u}(\nu_x, \nu_y)F(\nu_x, \nu_y)$, of the image. The spatial frequency spectrum can be calculated by means of a two-dimensional Fourier transformation:

$$\tilde{u}(\nu_x, \nu_y) = \int\int u(x,y)\, \exp[-2\pi i(x\nu_x + y\nu_y)]dxdy.$$

Spatial filtering experiments are fairly easy to implement if the object, u(x, y), is illuminated with a monochromatic plane wave which comes from a point source and is collimated by a lens. The object spectrum, $\tilde{u}(\nu_x, \nu_y)$, is then displayed in the so-called Fraunhofer plane, which is defined as the image plane of the point source and which is situated somewhere between the input plane and the output plane. If, for example, one wants to suppress a particular spatial frequency, (ν_x, ν_y), one has to insert an opaque spot at $(x = a\nu_x, y = a\nu_y)$. In other words, the filter function $F(\nu_x, \nu_y)$, is zero for this particular frequency. The constant, a, depends on the wavelength and on the geometry of the optical setup.

The fundamentals of spatial filtering were described 94 years ago by E. Abbe. In 1932, F. Zernike used a particular spatial filtering system in his phase contrast microscope for the detection

*I would like to thank Byron Brown, Dieter Paris and Harald Werlich for their excellent cooperation.

of phase objects. Twenty years later, A. Maréchal applied spatial filtering for image processing. His work has been continued and summarized by J. Tsujiuchi.[1] E. O'Neill[2] demonstrated the usefulness of spatial filtering for data processing. The same objectives were pursued by the optics group at the University of Michigan, particularly by L. Cutrona, E. Leith, A. Vander Lugt, and A. Kozma. Their work is summarized in a forthcoming article in Optica Acta by A. Vander Lugt.[3]

Our optics group has developed a new approach for making spatial filters. This activity is an outgrowth of our work on computer-generated holograms. After defining the spatial filter function, $F(\nu_x, \nu_y)$, which is supposed to achieve a certain effect, we calculate $F(n\delta\nu, m\delta\nu)$ at a mesh of sampling points. Next we binarize F which then is drawn in black and white by an automatic plotter. This drawing (Figure 1) is then photoreduced onto film or plate. Although this transparency has only two transmittance levels, zero and one, it nevertheless can perform the operation of any complex filter function, $F = A \exp(i\varphi)$. This has been explained previously[4, 5] by means of the diffraction theory for gratings with defects.

The applications of our spatial filtering experiments, which were presented at the Optical Society Meeting in the Fall of 1966, can be divided into three groups: repetition of classical methods with computer-generated spatial filters, differential operations for enhancing certain features such as edges and corners, and pattern recognition with matched filters[4] and other types of filters. The names and the functions of the filters realized so far are given in Table 1.

As an illustration, a reciprocal filter, $F(\nu_x, \nu_y) = 1/\tilde{u}(\nu_x, \nu_y)$, is shown in Figure 1. This filter is aimed at detecting the letter "E," among other letters. The input is shown on the left of Figure 2 and the output on the right. The strongest peak appears at a position corresponding to that of the "E."

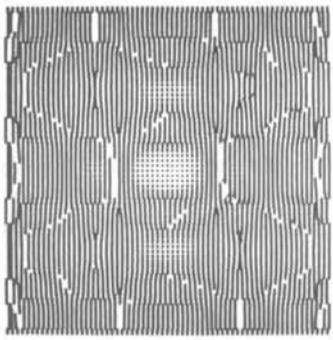

FIG. 1. Reciprocal filter for the detection of "E."

We feel that the prospects for computer-generated spatial filters are good because the advantages of the electronic digital computer and of the optical analog computer supplement each other very well. The digital computer is flexible enough

TABLE 1

Wolter's Schlieren method	$F(\nu_x, \nu_y) = \begin{cases} +1 ; \nu_x \geq 0 \\ -1 ; \nu_x < 0 \end{cases}$
Phase contrast	$F(\nu_x, \nu_y) = \begin{cases} iA ; \nu_x \approx 0 \approx \nu_y \\ 1 ; \nu_x \neq 0 \neq \nu_y \end{cases}$
$u(x, y) \rightarrow \frac{\partial u}{\partial x}$	$F(\nu_x, \nu_y) = 2\pi i \nu_x$
$u(x, y) \rightarrow \frac{\partial^2 u}{\partial x \partial y}$	$F(\nu_x, \nu_y) = -(2\pi)^2 \nu_x \nu_y$
Matched filter	$F(\nu_x, \nu_y) = \tilde{u}^*(\nu_x, \nu_y)$
Reciprocal filter	$F(\nu_x, \nu_y) = 1/\tilde{u}(\nu_x, \nu_y)$
Gradient correlation[6]	$F(\nu_x, \nu_y) = (\nu_x^2 + \nu_y^2)\, \tilde{u}^*(\nu_x, \nu_y)$
Code translation[7]	$F(\nu_x, \nu_y) = \tilde{v}(\nu_x, \nu_y)/\tilde{u}(\nu_x, \nu_y)$

FIG. 2. Input (left) to and output (right) of the filter in Fig. 1.

to produce almost any conceivable spatial filter, which can then be used in a relatively simple optical data processor, since it has the capability to handle many input data simultaneously.

REFERENCES

1. J. Tsujiuchi, Progr. in Opt. (E. Wolf, ed.), Vol. 2, p. 133 (1963).
2. E. L. O'Neill, IRE Trans. IT 2, 55 (1956).
3. A. Vander Lugt, Opt. Acta 14, ... (1967).
4. B. R. Brown, A. W. Lohmann, Appl. Opt. 5, 967 (1966).
5. A. W. Lohmann, D. P. Paris, Appl. Opt. 6, ... (1967).
6. S. Lowenthal, Y. Belvaux, C. R. Acad. Sc. Paris B262, 413 (1966).
7. A. W. Lohmann, D. P. Paris, H. W. Werlich, Appl. Opt. 6, ... (1967).

VISUAL FACTORS RELATED TO THE DESIGN AND USE OF DIRECT-VIEW ELECTRO-OPTICAL DEVICES

H. Richard Blackwell

Notable technical advances made during the past few years have ushered in the era of direct-view electro-optical devices which present the eye with exotic visual displays and the visual scientist with exciting new problems and opportunities. There is no doubt that these devices provide greatly improved visual capabilities when natural light levels are low. The problems presented to the vision scientist arise from the need that he provide data on the operating characteristics of the eye for the use of the designer who wishes to match the characteristics of the electro-optical device to those of the eye. The opportunities arise from the possibility that we may learn something of the processes underlying normal, unaided vision by studying the characteristics of the visual system when its optical input is altered in describable ways by electro-optical devices.

Direct-view electro-optical devices have ordinary optical components similar to those of visual telescopes and electronic components which amplify natural photon density to produce image-intensification. The images they produce are reasonably literal displays of natural objects with enhanced luminance and amplified noise properties. The degree to which scene luminance is enhanced depends upon the gain of the electronic components. The amplified noise is partially proportional to real object luminance, due to photon amplification, and partially unrelated to real object luminance, due to electronic noise.

Parameters available to the designer include the usual optical parameters of visual telescopes, magnification and field of view, and in addition, the electro-optical parameters, gain and noise. Furthermore, the spatial frequency response need not be uniform

over the field of view but may be graduated for some such purpose as to match the variable spatial frequency response of the visual field of the unaided eye.

Parameters of the visual system which are obviously related to the selection of an optimum electro-optical design include the trade-off of noise and gain, the spatial frequency response for different portions of the visual field, ocular scanning patterns for different target and background situations, and the interactions of these variables.

Theoretical questions which should be elucidated by study of vision with electro-optical devices include the basic decision processes for various target-background situations and the relevance of noise to these decision processes.

OLD AND NEW DESCRIPTIONS OF VISUAL TELESCOPES

It is instructive to contrast the methods developed for evaluating optical devices in the forties with methods required for evaluating electro-optical devices today. Classically, visual telescopes were described in terms of magnification and image quality, the latter defined by aberration indices. Toward the close of World War II, it became evident that whereas these descriptions evaluated telescopes quite well in terms of their usefulness at daytime football games, they failed miserably at describing their usefulness in detecting military targets at night. Howard Coleman first reported to this Committee the revolutionary notion that the usefulness of telescopes for this purpose depended upon their light transmittance and contrast rendition more than upon their image quality. Furthermore, Coleman showed that the overall utility of telescopes could be predicted quite well from these parameters together with magnification, by direct use of visual detection contours such as those I reported in 1946.[1] These contours represented the threshold contrast as a function of angular size for each of a number of different background luminances. Detection range for a given target varies with size, contrast, and luminance; and these could be related directly to the magnification, contrast rendition, and transmittance of first one then another visual telescope.

It must be emphasized that successful use of this approach in understanding the usefulness of visual telescopes depended upon the use of the telescopes under very restricted conditions. The telescopes were mounted and were vibration-free. The observers

were more-or-less correctly aligned with respect to the optical axes of the telescopes and did not have to search the environment for targets. These conditions obviously are rarely met in real-life use of visual telescopes but were used in special experiments in the hope that order could be brought out of chaos by restricting the number of variables at work.

Of course, even before the era of electro-optical telescopes, optical technology underwent the sine-wave revolution. It was found very useful to describe optical systems in terms of the contrast rendition they display to amplitude modulation of different spatial frequencies. The mathematical tools this transformation has brought to optics are powerful. The method has the tremendous advantage that the joint effects of systems of optical elements used in series may be easily computed. It is to be noted that this approach introduces a dimension omitted by Coleman's analysis, one related to the earlier descriptions of optical systems in terms of image quality, that is, spatial sharpness. Of course, there is no doubt that the extent to which amplitude modulations are rendered accurately for different spatial frequencies relates to the usefulness of a visual telescope in watching daytime football. But, to repeat history, we may well wonder whether the detection of military targets at low luminances is any more related to spatial frequency response than it was to image quality. Leaving aside the matters of gain and noise for the moment, we may say that these remarks apply equally well to electro-optical devices. The spatial frequency response of the optical and electronic components may be described in terms of modulation transfer functions (MTF).

Now the designer of the optical or electro-optical device wishes to optimize his design and to do so he must investigate the match of his device with the eye of the user. Of course, the designer thinks to use the powerful tools of sine-wave description for the machine-man system and inquires as to the sine-wave response characteristics of the human visual system. Early studies of the MTF of the human visual system were based upon the assessment of visual functions similar to those required in the daytime football game, such functions as the judgment of sharpness of photographs.

A much simpler method of defining the spatial frequency response of the visual system involves the use of single-frequency test objects in the form of sine-wave gratings. It is convenient to measure the spatial frequency response by the simple device of reducing the contrast of a sine-wave grating to a threshold and deducing the MTF for each single-frequency test object by the

contrast reduction required to bring the grating to threshold. When the threshold is that of mere detection of luminance ripple, the MTF is clearly being measured under conditions approximating threshold seeing and the data would seem directly relevant to the description of the visual detection of military targets. Values of MTF were obtained in this manner by A. S. Patel[2] at each of a variety of luminance levels, the lowest of which approaches the level encountered in the use of electro-optical devices. Thus, these data might well be expected to describe the effectiveness of electro-optical devices differing with respect to their values of MTF.

SPATIAL FOURIER ANALYSIS AND SPATIAL ISOMORPHIC MAPPING

We have a choice of how we think about the sine-wave description of the total human visual system. The MTF simply describes the effectiveness of the system at transmitting amplitude modulations (luminance contrasts) of different spatial frequencies (angular extents). We could imagine that the visual system actually performs a spatial Fourier analysis of the complex luminance distribution of a real scene and that the results of the analysis are transmitted along separate transmission lines with different efficiencies. Some further processing would have to occur before a decision could be made as to the information conveyed in the scene. As far as I know, no one (save a student of mine) has seriously considered that the visual system performs such an analysis. Rather, the sine-wave description of the visual system is usually thought of as a tool rather than a model. In this case, the visual system is presumably believed to have a spatial mapping property in which the real world is mapped isomorphically upon some CNS cell layer. Of course, other possibilities exist. Spatial information could be coded immediately into any number of terms, and there might not even be an isomorphic mapping.

For the moment, I prefer to think of the visual system as possessed of isomorphic mapping. The facts of spatial interaction have always made it clear that this mapping is a one-to-many, many-to-one mapping which may have both facilitatory and inhibitory properties. I have wondered what the spatial frequency description of the visual system implies for such a mapping system and have analyzed MTFs from this point of view.

TWO SPATIAL ISOMORPHIC MAPPING FUNCTIONS

Some years ago, Kincaid, Blackwell, and Kristofferson[3] first described a method for deriving a hypothetical spatial isomorphic mapping function from detection threshold data on circular targets consisting of luminance increments. The method assumed that each point in space mapped into a symmetrical area in a CNS cell layer with a graded amplitude of neural "excitation." The excitation was assumed to be linearly related to target luminance over a small range. The excitation produced by different points in space but mapped onto the same CNS cell was assumed to add algebraically. The map for a single point was assumed to have only a single maximum. Both positive (facilitatory) and negative (inhibitory) values were allowed. Under these conditions, targets such as circles will produce maps with single maxima. The detection criterion was assumed to be a fixed total amount of excitation, independent of the size of a circular target. Under these conditions, we may deduce the map of a single point, called the "element contribution function" from threshold data on circular targets of varying size.

Thanks to my student and colleague, Thomas G. Cleaver, we are now able to compute the element contribution function (ECF) by an analytic computer program rather than by the geometric approximations used in the earlier work. Figure 1 contains data Blackwell and Smith[4] have reported for the case of no measurable background luminance. Figure 2 presents the element contribution function computed from these data by the new program.* This is the presumed point spread function of the visual system derived on the basis of the element contribution notion. Note that there are only positive values, indicating an entirely facilitatory process.

We next developed a computer program to compute what might be called figure spread functions. The point spread function is involuted with the "luminance profile" of a target of interest, using a 50 by 50 matrix, each cell of which has 1 minute dimensions. The relative threshold luminance increment is taken as inversely proportional to the value of the figure spread function

*The analytic and experimental work on the new isomorphic mapping functions reported here has been supported by Contract No. DAAK02-67-C-0140 with the U.S. Army Engineer Research and Development Laboratories, Fort Belvoir, Virginia.

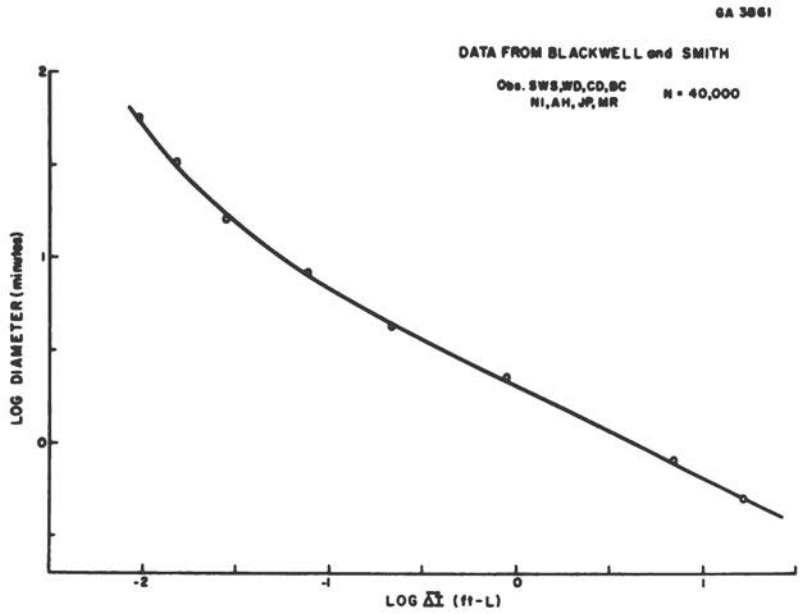

FIG. 1. Detection thresholds for circular targets without measurable background luminance.

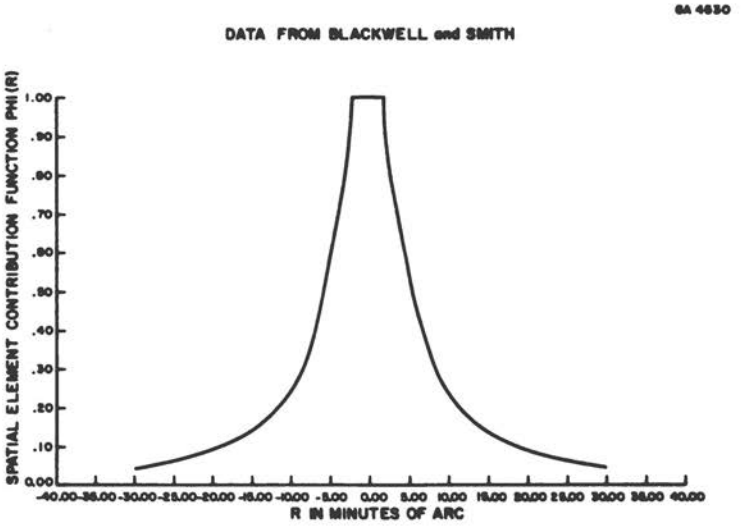

FIG. 2. Point spread function computed by element contribution analysis.

at its maximum. It is not necessary to know in advance at what point within the figure spread function the maximum occurs, since the program records values of the involution for all values of the matrix. The program also yields a graphical readout so that the figure spread function is represented by a "contourograph" representing slices through the three-dimensional figure.

The first use to which this process was put involved "predicting" the detection thresholds for circular targets of varying diameter. The predicted data were in precise agreement with the smooth curve shown in Figure 1. (This exercise checks out both the program used to derive the element contribution function and the program used to calculate figure spread functions).

Next, the involution was used to predict the detection thresholds for rectangles, one of whose dimensions was maintained constant at a large value, with the other dimension varied. The figure spread functions had single maxima. The predicted form of the threshold data is illustrated by the curve in Figure 3. As we shall see, these predictions are exceptionally well supported by experimental data.

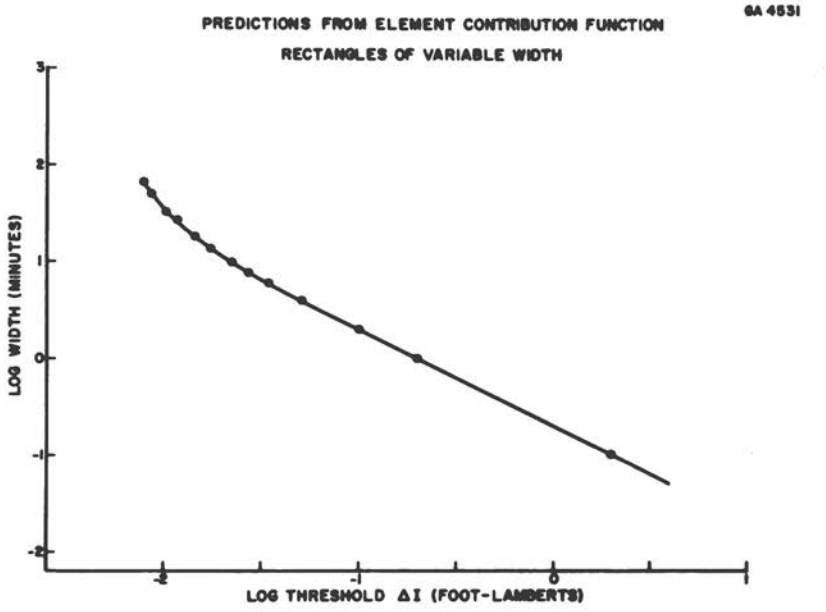

FIG. 3. Predicted detection threshold data based on element contribution analysis.

A point spread function has also been derived from the sine-wave response data obtained by Patel,[2] using the method described above in which the amplitude modulations (contrast) of sine-wave gratings were reduced to threshold recognition of the presence of a luminance ripple. Patel took the first step by computing a line spread function from his sine-wave response data. To do so, he had to assume that the system was linear over a small range and that effects were algebraically additive. The line spread function for an average grating illuminance of 3 trolands is presented in Figure 4. Note that all values are positive. To obtain a point spread function from this line spread function, we used a computer program which makes use of a matrix inversion sub-routine. The resulting point spread function is presented in Figure 5. Note that there are negative values at about 3 minutes from the center. The accuracy of the inversion method was checked by showing that a line spread function computed by involuting the point spread function agreed with the original line spread function.

Next, we computed figure spread functions for circles and one-dimensional rectangles by involuting this point spread function with appropriate target luminance profiles. Detection threshold

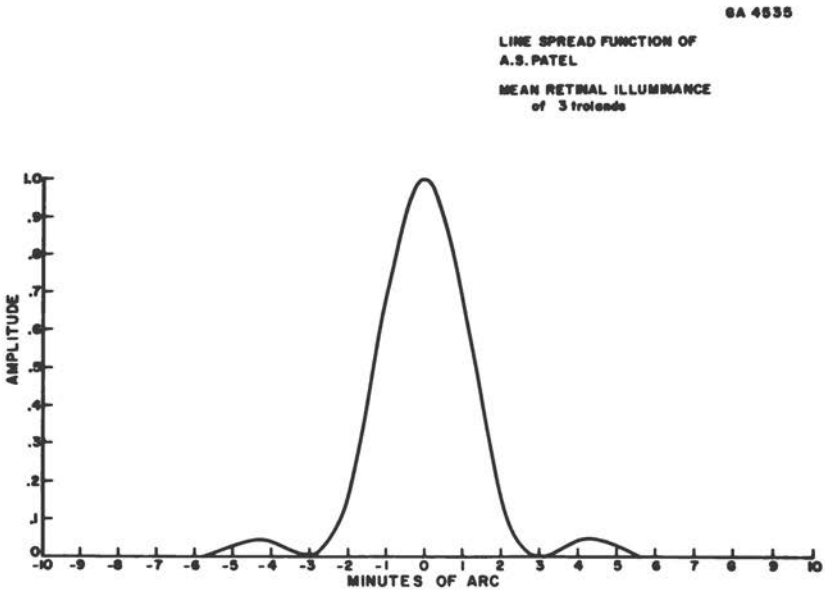

FIG. 4. Line spread function computed by sine-wave response analysis.

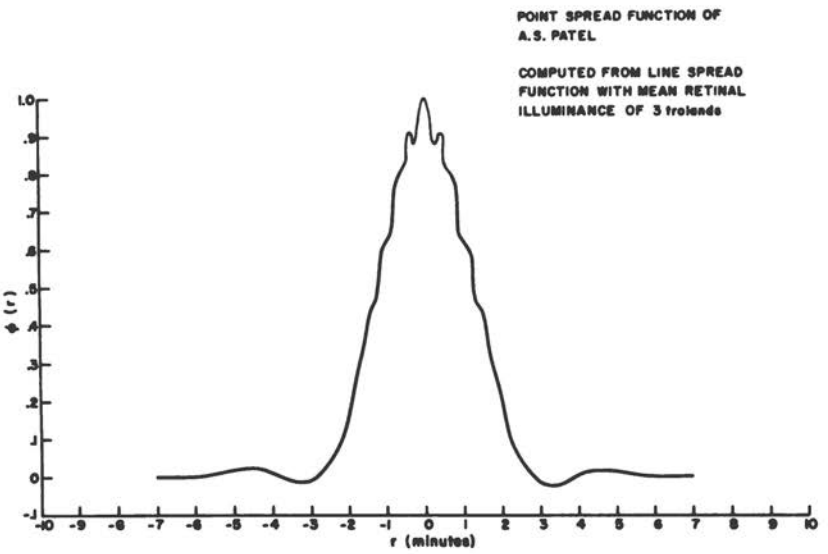

FIG. 5. Point spread function computed by sine-wave response analysis.

luminance increments were predicted as before, with the results shown in Figures 6 and 7.

We may compare the threshold relations for circles and one-dimensional rectangles predicted by the two isomorphic mapping functions by adjusting the values of threshold ΔI to agreement for small target dimensions. (This adjustment is necessary since no constraint was made on the area contained beneath the two point spread functions.) Such a comparison is made in Figure 8 for circles and in Figure 9 for the rectangles. In Figure 8, the data presented in Figure 1 are repeated. It is obvious that the threshold relations predicted for circular targets on the basis of the MTF point spread function do not agree with the experimental data. Of course, the "prediction" based upon the Element Contribution Function (ECF) point spread function agree with the data since these very data were used to derive the point spread function.

Experimental data obtained for one-dimensional rectangular targets with no measurable background luminance are presented in Figure 9. These data are seen to agree very well with predictions made on the basis of the ECF point spread function, and very badly with predictions based on the MTF point spread function.

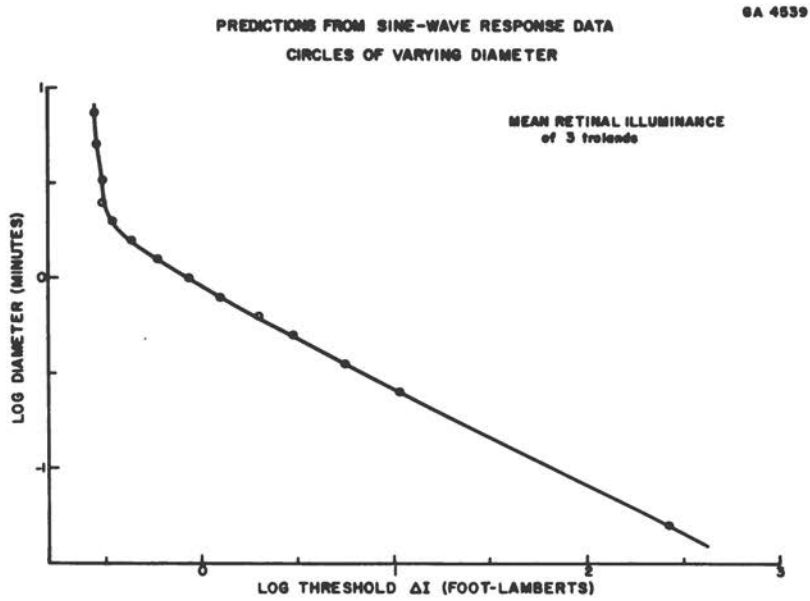

FIG. 6. Predicted detection threshold data based on sine-wave response analysis.

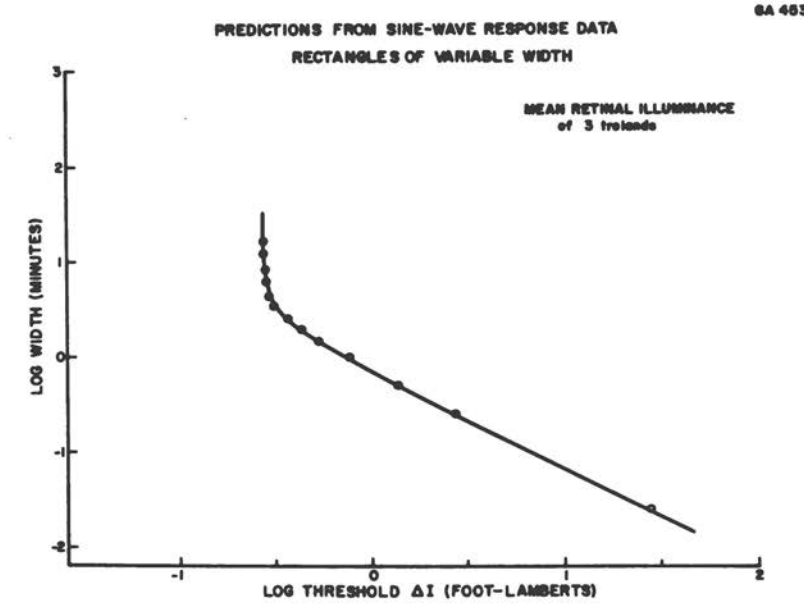

FIG. 7. Predicted detection threshold data based on sine-wave response analysis.

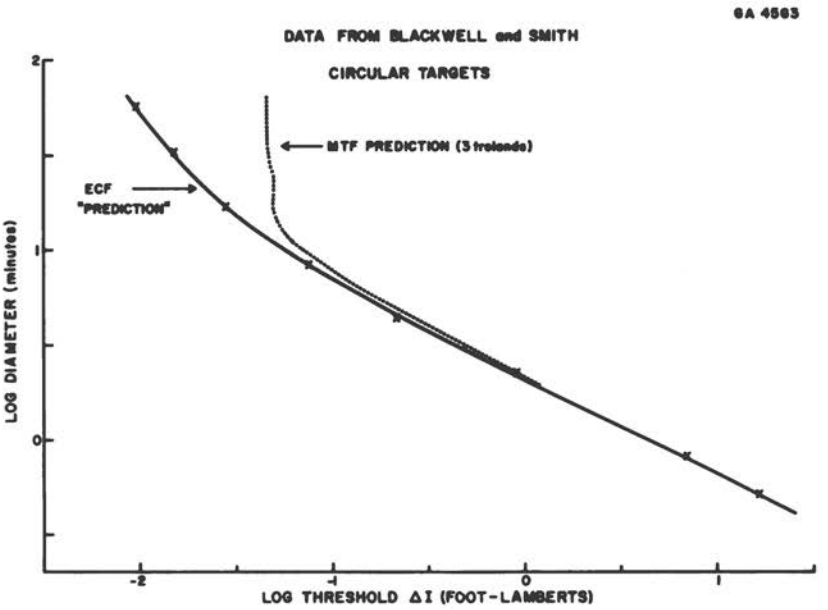

FIG. 8. Detection data and theoretical predictions based upon element contribution and sine-wave analyses.

These basic relationships between data and the predictions based upon point spread functions derived from sine-wave and element contribution analyses apply at different values of background luminance. As shown elsewhere,[5] the point spread functions derived from element contribution analyses narrow as luminance is increased but retain all positive values. The point spread functions derived from sine-wave analyses also narrow as luminance in increased and develop much more extended ranges of negative values. The differences between the predictions for circles and one-dimensional targets made by the two point spread functions increase as luminance is increased. Furthermore, the discrepancies between experimental data and the predictions made with the sine-wave analysis point spread function increase as luminance is increased. Thus, it seems that the two isomorphic mapping functions truly disagree and that the detection data we have available for analysis clearly favor the element contribution analysis.

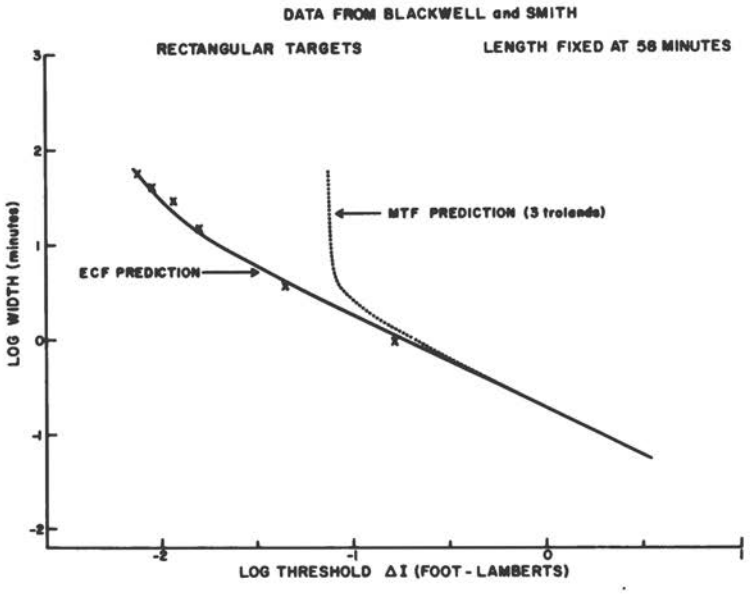

FIG. 9. Detection data and theoretical predictions based upon element contribution and sine-wave analyses.

SOME COMMENTS ON QUANTUM NOISE

Numerous experiments reported since 1939 have established the influence of quantum noise upon the detection threshold obtained with no measurable background luminance, at least under conditions in which the number of photons involved is quite small. In addition, there have been repeated efforts to relate the changes in detection threshold contrast which occur as a function of background luminance to the changes in the percentage of quantum noise involved. The original concept of deVries[6] was simply that the threshold number of additional photons needed for detection of a target is proportional to the standard deviation of photon variability produced by the background. Since the standard deviation of photon variability is proportional to the square root of the average number of photons, this leads to the prediction that threshold contrast is proportional to the inverse of the square root of background luminance.

If accurate, this simple prediction of the relationship between threshold contrast and luminance would be most useful in evaluating the relative importance of gain and noise in electro-optical

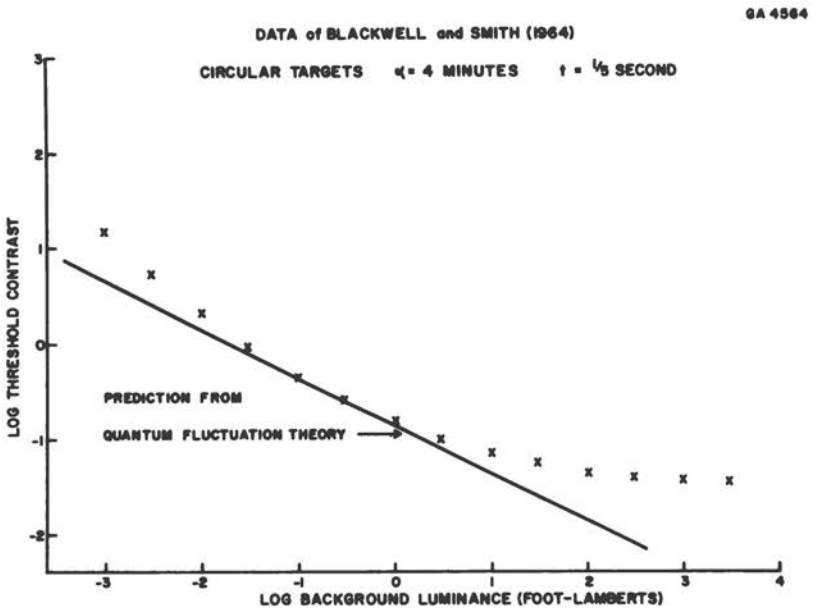

FIG. 10. Detection data at different luminances and the prediction of the quantum fluctuation theory.

devices. Unfortunately, however, the experimental evidence clearly does not support the "square-root law." Figure 10 illustrates some highly precise new threshold detection data[7] for circular targets viewed foveally and with natural pupils. The line illustrates the square-root law. It is apparent that the data really do not conform at all to the postulated relationship. Long the strongest supporter of the square-root law, Bouman[8] now concedes the inapplicability of the law to describe the experimental data over the range of luminances of interest. Bouman's "explanation" for the discrepancies takes the form of new theoretical models at both high and low luminances. The high luminance model leads to contrast constancy, and the low luminance model leads to contrast proportionally to luminance, the two asymptotic conditions for the data in Figure 10. Even if the expanded model is the theory of choice, the extensions at high and low luminance largely vitiate the value of the quantum fluctuation theory for the purpose of evaluating the relative significance of gain and noise in electro-optical systems.

The really interesting question, of course, concerns the relation between the properties of noise and the effect of noise upon

detection. Background noise will obviously influence target detection only if this noise can be confused with target signal. In the case of ordinary quantum noise, this would not seem to be the case. We may also wonder to what extent the postulated neural noise produced by background is sufficiently like the postulated target signal for confusion to exist. I am hoping that the papers by Russell Morgan and Wilson Tanner which follow will provide some useful information on this question.

Obviously, these comments about noise leave us very far from an understanding of the effects we might expect from the two classes of noise produced by electro-optical devices. Exploring the effects of manipulable noise of the two classes upon target detection should be most interesting.

THE COMPLEXITIES OF REAL USE OF ELECTRO-OPTICAL DEVICES

All of the foregoing discussion can be relevant to the use of electro-optical systems only under very restricted conditions. The devices must be vibration-free, and the users' eye must be fixed on-axis with respect to the devices and used to fixate only one point in the visual field. Then perhaps we can use our knowledge of the detection capability of the eye to evaluate displays differing in luminance, noise, magnification, and spatial frequency response.

In the real world, of course, the situation is remarkably different. The devices will often vibrate with respect to the eye of the user; the user will not be properly aligned with respect to the optical axis, and he must scan the environment before him in search of targets, both searching different portions of the field of the device and moving the device with respect to the world. During World War II, research on telescopes demonstrated the vast discrepancy between these two sets of conditions. Performance was markedly degraded under realistic conditions in comparison with simplified conditions, due presumably to factors such as the image blur produced by vibration and the loss of image quality due to de-centering of the device and the observers' eye. Of course, the same kind of effects can be expected in the new era of electro-optical telescopes.

In addition to all these complications, there is a special kind of problem introduced by the use of search and scanning which forces us into a different form of discourse. How do magnifica-

tion and field of view affect performance of telescopes? The effects of these variables depend strongly upon the precise conditions under which detection must occur. If the observer has a great deal of information about the world, he can use it to program his search tactics in such a way so that small field of view and high magnification will be optimal for detection. On the other hand, if the observer must actually examine the entire visual scene for possible targets, optimum performance will shift to lower magnifications and wider fields of view.

We have much to learn about the type of controlled behavior which governs visual search with restricted fields of view. This behavior is undoubtedly sequential and strongly dependent upon visual information obtained and hypotheses based upon this information. To what extent does detection begin with suspicion of an object viewed off-axis and end with on-axis confirmation? The answer to this question will have important implications for the spatial frequency characteristics designed into different portions of the field of view of electro-optical devices. And these design decisions will be related to our knowledge of the spatial frequency characteristics of different portions of the visual field of view.

We are now aware that different portions of the visual field possess up to four separate classes of receptor processes, differing intrinsically in spectral response, spatial frequency response, and integrating time. It also appears that neural networks from the same receptor populations may differ in their spatial frequency and temporal characteristics. Who can say how these different systems respond to the two classes of noise introduced by electro-optical devices.

Surely these vague remarks suggest the vast extent of our ignorance and the potential usefulness of such new knowledge as we may collect in the coming years for the design and use of electro-optical devices.

REFERENCES

1. Blackwell, H. R., Contrast Thresholds of the Human Eye. J. Opt. Soc. Amer. 36, 624 (1946).
2. Patel, A. S., Spatial Resolution by the Human Visual System. The Effect of Mean Retinal Illuminance. J. Opt. Soc. Amer. 56, 685 (1966).
3. Kincaid, W. M., Blackwell, H. R., and Kristofferson, A. B., Neural Formulation of the Effects of Target Size and Shape upon Visual Detection. J. Opt. Soc. Amer. 50, 143 (1960).

4. Blackwell, H. R. and Smith, S. W., The Effects of Target Size and Shape on Visual Detection II. Engng. Res. Inst. Report 2144-334-T, University of Michigan, p. 19 (1959).
5. Blackwell, H. R., Neural Theories of Simple Visual Discriminations. J. Opt. Soc. Amer. 53, 129 (1963).
6. deVries, H., The Quantum Character of Light and its Bearing upon Threshold of Vision, the Differential Sensitivity and Acuity of the Eye. Physica 10, 553 (1943).
7. Blackwell, H. R. and Smith, S. W., Status of Research on the Specification of Quantity of Illumination. Institute for Research in Vision Report, Ohio State University, p. 10 (1964).
8. Bouman, M. A., Vos, J. J., and Walraven, P. L., Fluctuation Theory of Luminance and Chromaticity Discrimination. J. Opt. Soc. Amer. 53, 121 (1963).

THE INFLUENCE OF SPATIAL AND TEMPORAL BANDWIDTH ON THRESHOLD CONTRAST SENSITIVITY OF THE EYE

Russell H. Morgan

Several years ago, Rose[1] in the United States and deVries[2] in the Netherlands independently proposed that visual performance as measured by the minimum perceptible contrast of an image (i.e., threshold contrast) is directly related to the random fluctuation in energy flux density or noise prevailing in the visual signal. This has been an important concept for it allows the development of mathematical models with which visual performance may be predicted under a wide range of conditions.

The noise in a visual signal may arise from a number of sources, depending upon the complexity of the visual system. For example, if the system includes television equipment, noise is introduced by the electron beam of the photoreceptor tube as well as by the several stages of electronic amplification. A source of noise common to all visual systems is the light composing image itself. Radiation comprises discrete photons whose number in any given situation fluctuates randomly about some average value in accordance with well-known statistical principles. Such fluctuations arise because of the probabilistic nature of photonic processes. Hence, when light of given intensity is projected on an observer's eye, a definite number of photons may be expected to reach and stimulate the retina in a given time interval. However, there is no certainty that precisely such a number will actually arrive. Consequently, the number of incident protons fluctuates from moment to moment and from region to region about some average value approaching the expected number.

Rose postulated that the minimum perceptible contrast of an image is proportional to the noise level prevailing in the visual

signal, that is,

$$C_t = k \cdot S_N, \tag{1}$$

where S_N is the root-mean-square value of the noise signal and k is a proportionality constant, the so-called threshold signal-to-noise ratio of the visual system. He also developed a method for the calculation of S_N for situations in which an observer views a series of round dots on a uniform background and where the noise is due entirely to random photon fluctuation.

The value of equation 1, in the prediction of visual performance under any given set of circumstances is clearly limited by one's ability to determine S_N. In this paper, a method[3] for calculating the noise levels of visual signals is proposed which differs from those suggested by others heretofore. The method is based upon statistical principles frequently used to calculate noise in physical systems (e.g., electronic amplifiers) which transmit or respond to modulated signals which are in a quantized energy state of flux. Because visual systems and their elements characteristically fall into such a class of systems (i.e., they transmit or respond to quantized signals that are modulated in time and space) the method should be generally useful. Indeed, the method appears to have considerable value in the development of models of visual perception in which predicted and measured data are closely similar.

DEFINITION OF TERMS

In the discussions which follow, a number of terms are used which, for the sake of clarity, are defined at the outset.

Photoreceptor. A surface composed of elements capable of reacting to an incident energy pattern (photons, electrons, etc.); e.g., the regina of the eye, photographic film, photoelectric surfaces.

Statistical unit. A quantity counted statistically as an independent event in a quantized energy state of flux; e.g., photons, electrons, neural impulses or aggregates of any one of these acting as a unit.

Spatial frequency (ν_x, ν_y, or ν_z). The number of cycles per unit length of a pattern whose quantized energy flux density varies periodically in one of the dimensions of space.

Temporal frequency (ν_t). The number of cycles per unit of time of a pattern whose quantized energy flux density varies periodically as a function of time.

Passband (bandwidth). The range of frequencies transmitted by or reacting with an element of an optical system; e.g., an eye which responds to spatial frequencies between ν_1 and ν_2 cycle/mm exhibits a passband of $\nu_2 - \nu_1$.

Modulation transfer function, $A(\nu)$. A relationship which expresses the normalized response of a photoreceptor when exposed to a sine wave test pattern, as a function of the spatial or temporal frequency of the pattern.

Contrast (C). The ratio of the root-mean-square value of the variable component of a pattern's energy flux density to the pattern's average flux density. In the case of a uniformly illuminated bar pattern, object contrast, C_O, is equal to $(B_1 - B_2)/(B_1 + B_2)$ where B_1 and B_2 are maximum and minimum test-object luminances respectively.

Noise contrast (C_N). The ratio of the root-mean-square value of the randomly fluctuating component of a pattern's energy flux density to the pattern's average flux density.

NOISE IN SINGLE PHOTORECEPTOR SYSTEMS

To determine the magnitude of the random fluctuation or noise prevailing in a visual system, consider first (Figure 1) a uniformly illuminated surface, x-y, composed of a large number of photoreceptive elements, which receive n photons in an incremental area, Δx, Δy, during a time interval, Δt. Because of the discontinuous nature of light, these photons may be expected to fluctuate about some average value, N, when sample measurements are made from time to time at various locations within the x-y plane; that is,

$$\langle n \rangle = N, \tag{2}$$

where the brackets signify an average over an equilibrium ensemble.

The mean-square value of the fluctuation may be shown statistically to be

$$\langle (n - N)^2 \rangle = \langle dn^2 \rangle = N. \tag{3}$$

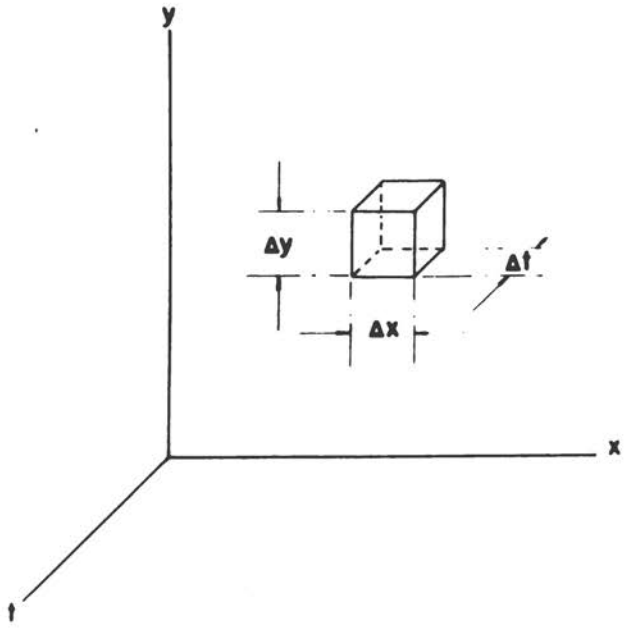

FIG. 1.

Now, if $\bar{n}$ is the average number of photons/mm^2/sec received by the x-y plane, then

$$\langle dn^2 \rangle = \bar{n} \cdot \Delta x \cdot \Delta y \cdot \Delta t. \tag{4}$$

From the definition of noise contrast given heretofore, equation 4 indicates that

$$C_N^2 = \frac{\langle dn^2 \rangle}{(\bar{n} \cdot \Delta x \cdot \Delta y \cdot \Delta t)^2} = \frac{1}{\bar{n} \cdot \Delta x \cdot \Delta y \cdot \Delta t}. \tag{5}$$

For equation 5 to be useful in the calculation of noise contrast, the photoreceptor surface must have sharply defined limits of resolution in both the space and time domains; i.e., the surface must have characteristics that such an element, $\Delta x \cdot \Delta y \cdot \Delta t$, is resolved with perfect clarity, whereas a just smaller element is not recorded at all. However, as is well known, optical systems do not exhibit these properties. Instead, their response to a series of progressively smaller elements shows gradual deterioration in accordance with their frequency response

characteristics. Hence, a more satisfactory expression for noise contrast must be sought.

Now optical systems are characteristically ergodic; that is, the fluctuation associated with an incremental element, $\Delta x \cdot \Delta y \cdot \Delta t$, is the same whether one samples in the x-, y-, or t-axis or in any combination of the three. Therefore, by Fourier analysis, each of the elemental dimensions in equation 5 may be replaced by an appropriate partial integral as follows:

$$1/\Delta x = 2 \int [A(\nu_x)]^2 \, \delta\nu_x \tag{6a}$$

$$1/\Delta y = 2 \int [A(\nu_y)]^2 \, \delta\nu_y \tag{6b}$$

$$1/\Delta t = 2 \int [A(\nu_t)]^2 \, \delta\nu_t, \tag{6c}$$

where $A(\nu_x)$, $A(\nu_y)$, and $A(\nu_t)$ are the modulation transfer functions of the photoreceptor in the x-, y-, and t-dimensions respectively. Consequently, equation 5 becomes

$$C_N^2 = \iiint \delta C_N^2$$

$$= \frac{2\int[A(\nu_x)]^2 \, \delta\nu_x \cdot 2\int[A(\nu_y)]^2 \, \delta\nu_y \cdot 2\int[A(\nu_t)]^2 \, \delta\nu_t}{\bar{n}}. \tag{7}$$

The partial integrals of equations 6a - 6c and of equation 7 represent sharply defined passbands, spatial and temporal, equivalent from the standpoint of noise generation to the broader and less sharply defined passbands to which the photoreceptor is responsive. For this reason, these integrals are often referred to as noise-equivalent-passbands or bandwidths and denoted respectively by the symbols, $\Delta\nu_x$, $\Delta\nu_y$, $\Delta\nu_t$. Hence, equation 7 may be written

$$C_N^2 = \frac{2\Delta\nu_x \cdot 2\Delta\nu_y \cdot 2\Delta\nu_t}{\bar{n}}. \tag{8a}$$

Many photoreceptors are isotropic in the x- and y-dimensions. Under these circumstances, $\Delta\nu_x = \Delta\nu_y$ and equation 8a may be simplified to

$$C_N = 2\Delta\nu_x \cdot (2\Delta\nu_t/\bar{n})^{1/2}. \tag{8b}$$

Equations 8a and 8b constitute the general noise equations of a photoreceptor system. They indicate that photonic noise is governed by the number of available photons and by certain photoreceptor characteristics, namely the spatial and temporal bandwidths of the system. This is in contrast to the noise equations

proposed heretofore in which it has been suggested that photonic noise is governed by the characteristics of the image to be recorded as well as by the number of available photons. The difference is a fundamental one. As we shall see, the new equations remove some of the inconsistancies that have prevailed heretofore in the predicted values and physiological observations of visual threshold signal-to-noise contrast ratios. It also appears that they will allow the development of mathematical models of visual systems with an ability to predict visual performance on more than a very limited basis.

NOISE IN OTHER PHYSICAL SYSTEMS WHOSE PERFORMANCE IS LIMITED BY QUANTUM STATISTICS

The foregoing method of determining noise levels in simple visual systems will be recognized as identical to that employed in electronic systems to calculate the noise current created by a temperature-saturated diode feeding into an amplifier having a bandwidth of $\Delta\nu_t$ cycles/sec. To illustrate this, let us consider the conditions prevailing when such a diode emits n electrons during a time interval, Δt. Since electron emission is a random process, these electrons may be expected to fluctuate about some average value N, when sample measurements are made from time to time; that is,

$$n = N. \tag{9}$$

As before, the mean-square value of the fluctuation or noise is

$$(n - N)^2 = dn^2 = N. \tag{10}$$

Also, it will be apparent that the ratio of the mean-square fluctuation in the number of electrons during time, Δt, to the square of the average number of emitted electrons in the same period is equal to the ratio of the mean-square of the noise current of the diode, i_n, to the square of the diode's average current, i; that is,

$$\frac{dn^2}{N^2} = \frac{i_n^2}{i^2} . \tag{11}$$

Now if $\bar{n}$ is the average number of electrons emitted per second (i.e., $N = \bar{n} \cdot \Delta t$), then from equations 10 and 11,

$$\frac{i_n^2}{i^2} = \frac{1}{\bar{n}\Delta t}. \tag{12}$$

Because amplifiers do not characteristically have sharply defined limits in respect to the time intervals to which they will and will not respond, just as was the case with optical systems, it is necessary to replace the elemental dimension, $1/\Delta t$, of equation 12 by the integral

$$1/\Delta t = 2\,A(\nu_t)^2\,d\nu_t\,. \tag{13}$$

Hence,

$$\frac{i_n^2}{i^2} = \frac{2\,A(\nu_t)^2\,d\nu_t}{\bar{n}} = \frac{2\,\Delta\nu_t}{\bar{n}}\,. \tag{14}$$

But $\bar{n}$ equals i/e, where e is the charge on the electron. Therefore,

$$i_n^2 = 2\,ei\,\Delta\nu_t. \tag{15}$$

Equation 15 is the classical formula expressing the noise current of a temperature-saturated diode feeding into a system having a bandwidth of $\Delta\nu_t$. Its derivation is clearly identical to that used to obtain the noise levels generated by random photon fluctuation in optical systems. It is perhaps appropriate to point out that in both systems we are dealing with signals which are in a quantized energy state of flux which are impressed upon system elements having limited bandwidth characteristics in one or more dimensions of time and space.

CALCULATIONS OF RETINAL NOISE CONTRAST

If the surface illustrated in Figure 1 is the retina of the eye, the observed noise contrast is greater than that indicated by equation 8b because only a fraction of the incident photons admitted to the retina by the pupil produce neural impulses. Transmission losses within the ocular media and absorption losses at the retinal surface are present. Photon-to-neural-impulse conversion may also introduce a further diminution in the effective number of incident photons. Hence, the noise contrast due to photon fluctuation at the retina is

$$C_{Nr} = 2\,\Delta\nu_{xr}(2\,\Delta\nu_{tr}/\bar{n}\,T\,a\mu_s)^{1/2}, \tag{16a}$$

where μ_s is the number of statistical units produced in the form of neural impulses by the rods or cones per incident photon, T is the transmittance of the ocular media, a is the fractional absorption of the retinal photoreceptors and the subscripts, r, indicate that the noise-equivalent-passbands are specifically those of the retina.

It is usually convenient to express the illumination on the retina in trolands instead of $\bar{n}$. If a value of 3.5×10^7 photon mm^{-2} sec^{-1} is taken as equivalent to 1 troland, then

$$C_{Nr} = 4.8 \Delta\nu_{xr} (\Delta\nu_{tr} / B\,T\,a\mu_s)^{1/2} \times 10^{-4}, \qquad (16b)$$

where B is retinal illumination in trolands.

Values of the spatial noise-equivalent-passband, $\Delta\nu_{xr}$, of the retina have been calculated for various levels of retinal illumination by Schade[5] and are tabulated in column 2, Table 1.

The sensitivity of the eye, measured with sinusoidal time-dependent stimuli has been studied by Kelly.[6] From these data, the eye's temporal noise-equivalent-passband, $\Delta\nu_{tr}$, may be readily calculated if one assumes that the eye's temporal modulation transfer function is directly related to its temporal amplitude sensitivity. Values of $\Delta\nu_{tr}$, calculated on this assumption, are listed in column 3 of Table 1.

The statistical transfer ratio, μ_s, of the retina has been investigated by a number of workers and is now generally believed to have a value close to unity for both the rods and cones. The fractional absorption, a, of the rods seems to be near 0.35 when the retina is illuminated with white light; a somewhat higher value may be applicable to the cones. The transmittance of the eye's

TABLE 1. Evaluation of Retinal Noise Contrast

Retinal Illumination (trolands)	$\Delta\nu_{xr}$ (cy.mm^{-1})	$\Delta\nu_{tr}$ (cy.sec^{-1})	C_{Nr}
5×10^3	85	30	7.3×10^{-3}
9×10^2	70	26	1.4×10^{-2}
9×10^1	47	20	2.6×10^{-2}
9×10^0	28	15	4.1×10^{-2}
9×10^{-1}	18	10	6.7×10^{-2}
9×10^{-2}	12	10	1.4×10^{-1}
9×10^{-3}	7.5	(10)	2.9×10^{-1}
9×10^{-4}	5.5	(10)	6.4×10^{-1}

ocular media has been studied by Ludvigh and McCarthy[7] and has been shown to have a value close to 0.5 through most of the spectrum above a wavelength of 440 millimicrons. When these values are substituted in equation 16b, retinal noise contrast may be determined (see column 5, Table 1).

Many visual systems, (e.g., those using television or photographic equipment) include one or more photoreceptors in addition to the observer's retina. Because the response of a photoreceptor is probabilistic, noise is generated by each photoreceptor stage. The calculation of noise levels under these circumstances can become quite involved but nevertheless follows the same principles set forth above. Morgan[3] has recently given a detailed discussion of the methods used in the calculation of noise in complex visual systems and hence further attention to this subject need not be given here.

CONTRAST SENSITIVITY AND NOISE CONTRAST

It has been pointed out earlier in this paper that an image projected upon the retina may be expected to be seen only when its contrast, C_r, is equal to or exceeds some multiple of the noise contrast; that is,

$$C_r = k C_N \cdot \tag{17}$$

The validity of this concept has recently been tested[4] under conditions in which an observer was required to determine his threshold contrast perception when viewing a bar pattern of fixed spatial frequency displayed on a television monitor and on which were impressed noise signals of known amplitude. The relationship set forth in equation 17 seems to hold over an extensive range of noise levels (see Figure 2).

When the scene under observation is a test object bearing a sinusoidal pattern in the space domain, the relationship between retinal contrast and test object contrast is

$$C_r = C_o \cdot A_{cr}(\nu_x) \cdot S, \tag{18}$$

where $A_{cr}(\nu_x)$ is the composite spatial modulation transfer function of the complete visual system (in simple visual systems, this function is simply the combined functions of the eye lens and retina) and where S is a factor reflecting the loss of contrast due to scattering processes in the visual system.

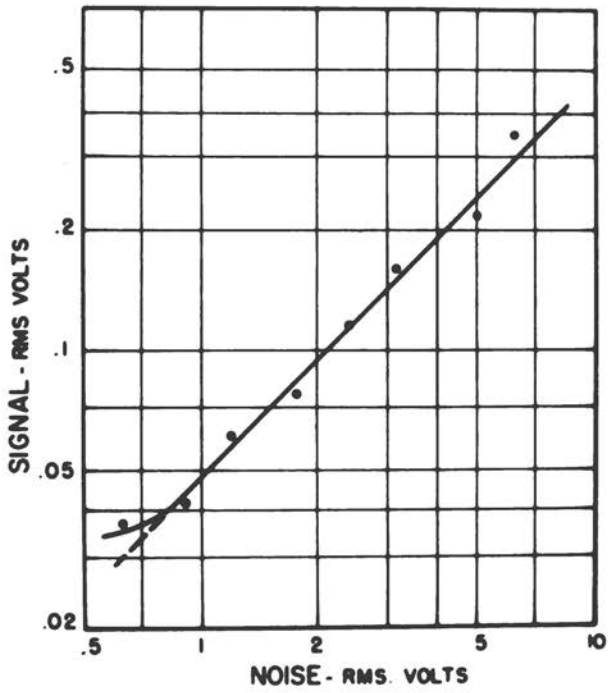

FIG. 2. Threshold contrast values, measured in terms of the rms. voltage of a square-wave pattern applied to a television kinescope, plotted as a function of the rms. voltage of a white noise signal also measured at the kinescope impact. Viewing distance: 1.25 meters; visual angle: 15°.

When equations 17 and 18 are combined, visual performance expressed in terms of threshold test object contrast, C_{To}, is given by the formula

$$C_{To} = \frac{k \cdot C_N \cdot S}{A_{cr}(\nu_x)} \quad . \tag{19a}$$

For simple visual systems, the value of C_N is that given by equation 16b. Under these conditions, threshold test object contrast becomes

$$C_{To} = 4.8\, kS \cdot \Delta\nu_{xr}/A_{cr}(\nu_x) \cdot \Delta\nu_{tr}/B\,T\,a\mu_s^{1/2} \times 10^{-4}. \tag{19b}$$

Equation 19a and 19b constitute a mathematical model with which the threshold contrast of a sinusoidal test object may be predicted in simple visual systems. It suggests that the modula-

tion transfer function of a system may be determined from reciprocal threshold contrast values plotted as a function of spatial frequency. It also indicates that these equations may be used to predict threshold contrast for other types of test object if the modulation transfer function for sinusoidal patterns is replaced by an appropriate transfer function which gives the composite response of the visual system in terms of the spatial characteristics of the test object in question.

The value of k in equations 19a and 19b may be determined for sinusoidal test patterns from measurements of threshold test object contrast made under conditions where C_N equals C_{Nr}, the value of $A_{cr}(\nu_x)$ equals 1.0 and S approaches unity; that is, where the noise contrast is governed entirely by retinal performance, the sine wave contrast sensitivity of the eye is maximal and scattering effects are sufficiently small that they can be neglected. Under these circumstances, k is simply equal to C_{To}/C_{Nr}.

Column 3 of Table 2 lists threshold test object contrast values for several levels of retinal illumination under conditions where $A_{cr}(\nu_x) = 1$. These data were obtained from contrast sensitivity measurements reported by de Palma and Lowry[8] for illumination levels above 9×10^1 trolands and by Van Nes and Bouman[9] for levels of 9×10^1 trolands and below. The Van Nes and Bouman data are averages of measurements made in the red, green and blue regions of the spectrum at each light level. Column 2 lists corresponding values of C_{Nr} taken from Table 1. Values of k are tabulated in column 4.

TABLE 2. Evaluation of Threshold Signal-to-Noise Ratio, k, for Sinusoidal Test Objects

Retinal Illumination (trolands)	C_{Nr}	C_{To} for $A_{cr}(\nu_x) = 1.0$	k
5×10^3	7.3×10^{-3}	1.1×10^{-3}	0.15
9×10^2	1.4×10^{-2}	1.4×10^{-3}	0.10
9×10^1	2.6×10^{-2}	2.3×10^{-3}	0.09
9×10^0	4.1×10^{-2}	4.1×10^{-3}	0.10
9×10^{-1}	6.7×10^{-2}	7.5×10^{-2}	0.11
9×10^{-2}	1.4×10^{-1}	2.1×10^{-2}	0.15
9×10^{-3}	2.9×10^{-1}	3.9×10^{-2}	0.14
9×10^{-4}	6.4×10^{-1}	1.2×10^{-1}	0.19

It will be observed that, in general, k remains quite constant throughout the range of retinal illumination studied. One should perhaps expect this since the computational processes performed by the brain on retinal signals are not likely to be different at one level of illumination from that at another. It will also be seen that values of k tend to rise at both high and low levels of illumination. At high light levels, this may be due to saturation of the visual system which causes perturbations that override photon fluctuation. At low levels, the observation may be due to the fact that noise contrast is sufficiently high that nonlinearity of the visual system may cause noise levels to appear higher than calculated.

It is perhaps of some interest to speculate upon the significance of the factor, k. It seems possible that its value reflects the various auto-correlation processes which the brain must perform to permit recognition of the ordered optical signals recorded in the presence of noise. Hence, values of k are likely to be influenced by the type of test object under observation and the conditions under which the optical signals are presented to the observer. For example, although it seems unlikely that the value of k would be substantially different for sinusoidal and dot patterns observed on a uniform field when long observation times are permitted, it possibly may change quite substantially when observation times are shortened, when test patterns are made more complex, or when these patterns are superimposed on nonuniform fields. Hence, investigations of the manner in which k changes when viewing conditions are varied may be expected to provide important information on the manner in which visual data are processed by the brain.

REFERENCES

1. Rose, A. Sensitivity performance of the human eye on an absolute scale. J. Optic. Soc. Amer. 38:196, 1948.
2. de Vries, H. The quantum character of light and its bearing upon threshold of vision, the differential sensitivity and visual acuity of the eye. Physics 10:553, 1943.
3. Morgan, R. H. Threshold visual perception and its relationship to photon fluctuation and sine wave response. Am. Jour. Roent., Rad. Ther. & Nuc. Med. 93:982, 1965.
4. Morgan, R. H. Data in process of publication.
5. Schade, O. H. Optical and photoelectrical analog of eye. J. Optic. Soc. Amer. 46:721, 1956.

6. Kelly, D. H. Visual responses to time-dependent stimuli. I. Amplitude sensitivity measurements. J. Optic. Soc. Amer. 51:422, 1961.
7. Ludvigh, E. and McCarthy, E. F. Absorption of visible light by refractive media of human eye. Arch. Opthalmol. 20:37, 1938.
8. de Palma, J. J. and Lowry, E. M. Sine wave response of visual systems. II. Sine wave and square wave contrast sensitivity. J. Optic. Soc. Amer. 52:328, 1962.
9. Van Nes, F. L. and Bouman, M. A. The effects of wavelength and luminances on visual modulation transfer. Proc. of Coll. on Performance of the Eye at Low Luminances; Excerpta Medica Found., Int. Cong. Series No. 125, Amsterdam, 1965.

VISUAL DETECTION OF OSCILLOSCOPIC TRACINGS*

Wilson P. Tanner, Jr., Danna B. Main, Theodore E. Cohn

In this paper we are dealing with a visual problem involving noise that is different in nature from the problems discussed by the previous authors. In particular, we are considering the problem of attempts to read physiological records for the purpose of describing the response in the record to some physical event. In the early days, the nature of these responses was determined largely by visual inspection. More recently, with the availability of computers, various averaging techniques have been employed. Presumably the assumption is that those parts of the records constituting aspects irrelevant to the response are random events which, over a large number of observations, will average to zero while the response itself will be additive. The result of the averaging, then, is thought to be a description of the response. In an attempt to illustrate visually the nature of the averaging procedure, we placed a polaroid camera on an oscilloscope face and ran six hundred traces on the same film. In doing this, we simulated the response with a well defined signal and added to this the random noise as illustrated in Figure 1. In the figure, we have two illustrations of the signal we employed (I-a), two illustrations of the random noise (I-b), and two illustrations of the signal-plus-noise (I-c). The noise sample in the latter two is not the same noise sample as in the first two.

Figure 1-II illustrates six hundred superimposed trials of signal-plus-noise. By drawing a line through the most dense portion of the figure, one can see what the average would be. In this

This work was supported by the Office of the Surgeon General, Contract No. DA-49-193-MD-2585.

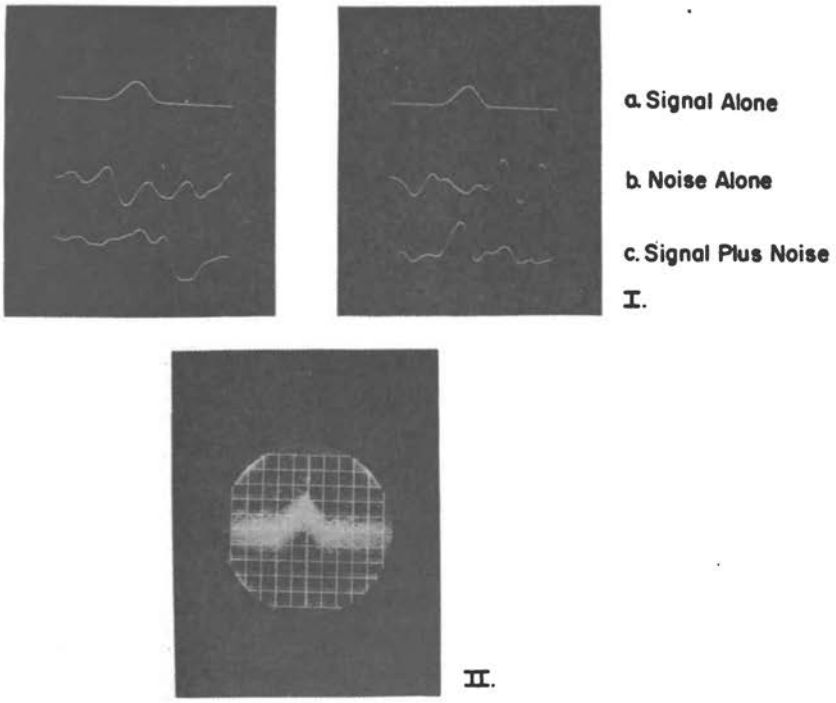

FIG. 1. Ia) Two Photographs of Signal Alone
Ib) Two Photographs of Noise Alone
Ic) Two Photographs of Signal-Plus-Noise
II) 600 Superimposed Trials of Signal-Plus-Noise

case, it should be noted that this average will yield a reasonably true description of the signal. It should be pointed out that, in this case, the signal waveform and starting time were precisely the same for each trace. The result is one that would be expected if a computer of average transients (CAT) were employed under these precise conditions.

Figure 2 shows the introduction of a small amount of starting-time uncertainty in the position on which the signal occurred on the sweep. The signal has the same waveform as shown previously. Figure 2-a illustrates the results of superimposing one hundred trials of signal alone when starting-time uncertainty has been introduced. In Figure 2-b, traces of signal-plus-noise for a signal with starting-time uncertainty are superimposed. Again,

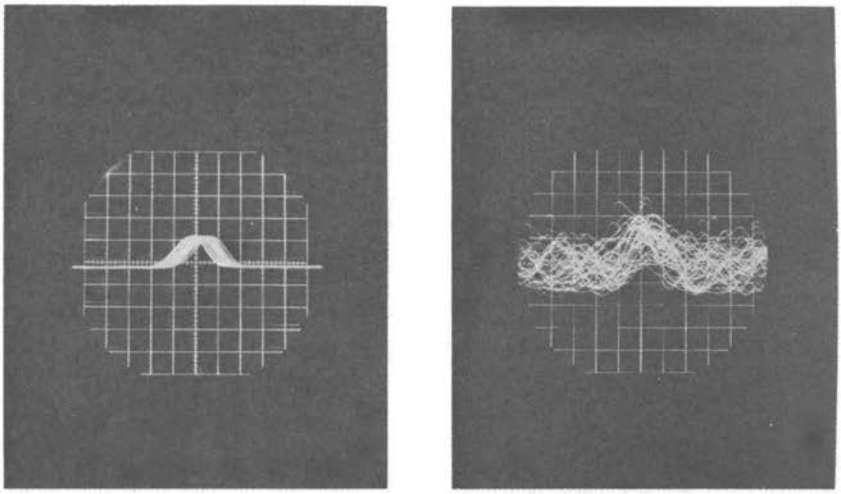

FIG. 2. a) Superimposed Trials of Signal with Starting Time Uncertainty
b) Superimposed Trials of Signal-Plus-Noise for Signal with Starting Time Uncertainty
100 μsec/cm 20 mv/cm

drawing a line through the maximum density, we see the result of averaging. In this case, the average would describe a signal of greater duration and lesser amplitude than what was actually employed. If the signal had both positive and negative components and sufficient starting-time uncertainty, one might expect that the signal would also average to zero, and no response would be interpreted as having occurred.

Psychophysical experiments have indicated that the observer's knowledge of signal parameters yield ROC curves on normal-normal paper with slopes that depend upon the amount of uncertainty. Information obtained from experiments employing a human observer may be combined with that obtained from computer-averaging techniques, such as the CAT. For example, we know that the human observer is capable of performing certain tasks such as concept formation and pattern recognition. We do not know yet how to program computers to perform these same tasks. After all, when you start looking for the response, you do not know what to look for. Let us put a human observer into an experiment and program him by monetary payoffs. We present him with simulated physiological records. In some of these records, the physical event which might lead to a response has occurred and, in others, it has not. We ask him each time to make a statement about whether or not he thinks the physical event occurred. By

varying the payoffs, we should influence both hit and false alarm probabilities, thus, collecting data sufficient to construct an ROC curve. The level of his performance should indicate a lower bound on the detectability of this response in the physiological noise. For example, if the human observer performs well in this task and the averaging technique suggests no response, perhaps because the response has both negative and positive aspects, then, obviously, there is something wrong with averaging. The slope of the ROC curve should give some indication as to the degree of uncertainty in the position and the form of the response.

PROCEDURE

The observer viewed an oscilloscope face in a dimly lighted room and was told that on each trial one of two events would occur: (a) noise alone might be present or (b) signal-plus-noise might be present. He was instructed to press one of two buttons corresponding to each of the two conditions. He was told nothing about the signal characteristics. After each response, he was informed by lights which of the two events had occurred. The actual event and the observer's response on each trial were recorded on counters and also punched onto IBM cards. The events, signal-plus-noise (S + N) and noise alone (N), occurred randomly with an a priori probability of 0.5. An experimental session consisted of eight hundred trials taken within a two hour period. The observer was given a one minute rest period after each run of one hundred trials and a five to ten minute rest period after four hundred trials. He was paid a base hourly wage plus a bonus dictated by one of the payoff matrices shown in Figure 3. The observer was tested for at least five consecutive days at each of these payoff conditions.

In Experiment 1 all signal parameters were fixed when the S + N event occurred. The signal consisted of a decaying exponential pulse filtered at 20 cps and 4 kc with a peak amplitude of 2 mV after filtering. The noise consisted of 15 mV (true RMS) filtered between 1 cps and 5 kc. In Experiment 2, the same pro-

		Observer Response					
		(1)		(2)		(3)	
		A	CA	A	CA	A	CA
TRANSMITTER	S+N	+0.1	-0.1	+0.3	-0.3	+0.1	-0.1
READING	N	-0.1	+0.1	-0.1	+0.1	-0.3	+0.3

FIG. 3. Experimental Payoff Matrices

cedures as those used in Experiment 1 were followed with the same observer employed. In this case, however, the signal was the same as that described above except for slight starting-time uncertainty. The signal could begin anywhere within a distance of 1.5 cm on the oscilloscope face, which corresponds to a starting-time uncertainty of approximately 150 microseconds, about one-half the duration of the signal.

RESULTS

Figure 4 is a plot of the hit probability [$P_{SN}(A)$] against the false alarm probability [$P_N(A)$] for the fixed starting-time condition. Each point is based on the data from eight hundred trials. Each group of five points was obtained under the three different payoff conditions. The ROC curve is of the form that would be expected if the signal has no uncertain parameters. The average

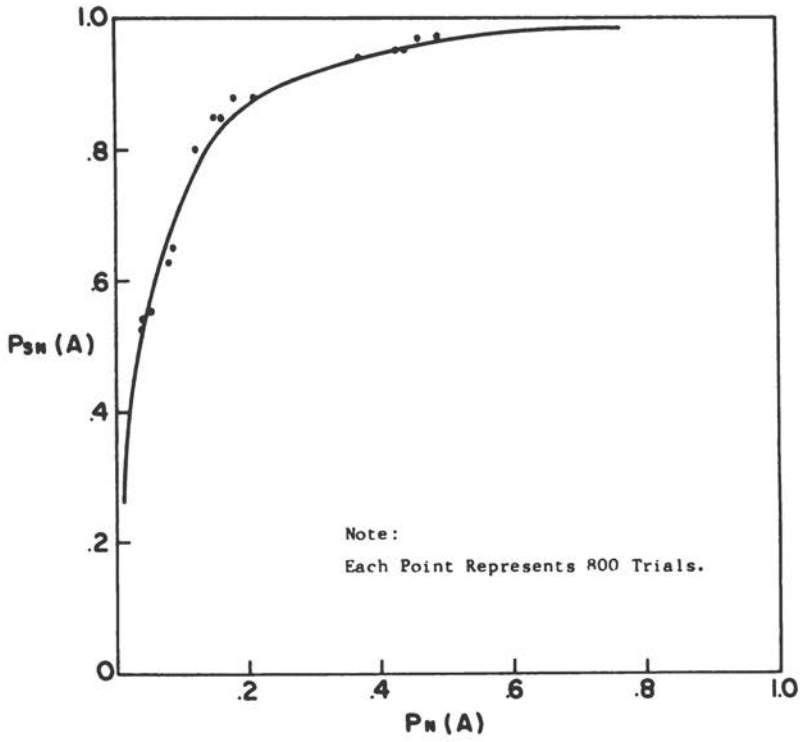

FIG. 4. Receiver Operating Characteristic for Fixed Signal.

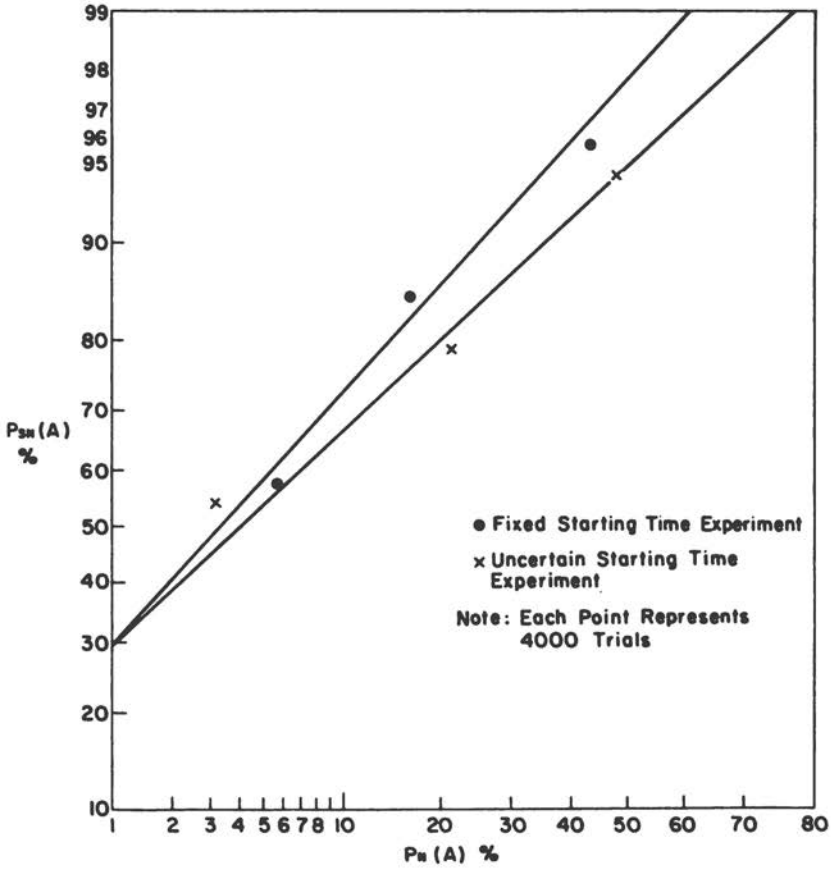

FIG. 5. Receiver Operating Characteristics for Both Experiments.

of each group of five points is replotted on double probability paper in the upper curve in Figure 5. It can be seen that the slope of this curve is approximately one. The lower curve, which represents the data from the uncertain starting-time experiment, has a slope less than one. The difference in the slopes reflects the starting-time uncertainty of the observed signal.

DISCUSSION

These experiments illustrate the human"computer's" usefulness in determining not only a lower bound of detectability of a

response in a record but also an indication of the uncertainty in the parameters of the response. We are not suggesting that all computers be retired in favor of human observers in the tedious task of record reading. But we do suggest that information from experiments employing human observers and computers might lead to more precise descriptions of physiological responses. For example, the experimenter may program a template for the physiological response he is studying. The template, based on the experimenter's previous knowledge, may be as detailed as specified waveform or, at the other extreme, a simple time interval. The experimenter may cross-correlate this template, trial by trial, with physiological records that result when input A occurs (S + N, for example) and again when input B occurs (N, perhaps). The detectability given the template may be determined. As the experimenter's template approaches the form of the physiological response he is studying, the detectability will increase. Given any template, the degree of uncertainty in the parameters of the response under study is indicated by the slope of the ROC curve.

In conclusion, we suggest that the advent of averaging computers has made possible the collection of information from biological systems that was a formidable task in the not so distant past. We must carefully examine, however, the assumptions made when these techniques are used. By working within the framework of the theory of signal detectability, it is possible to quantify at least a lower bound of detectability of a physiological response in a record, whether we use unaided human observers or a programmed template. More important, the framework enables us to state the uncertainty in the physiological response in terms of the slope of an ROC curve.