INTERACTION OF DIFFERENT MODULES IN DEPTH PERCEPTION

Heinrich H. Bülthoff * and Hanspeter A. Mallot* **

* Massachusetts Institute of Technology, Center for Biological Information Processing
MIT E25-201, Cambridge, MA 02139, USA
** Joh. Gutenberg-Universität, Institut für Zoologie III, Saarstr. 21, D-6500 Mainz, FRG

ABSTRACT

The interaction of modules for depth perception was studied psychophysically by measuring the perceived depth of computer generated images showing simple solid objects with different combinations of depth cues. Accumulation of information from shading and stereo and vetoing of depth from shading by edge information have been found. Cooperativity and other types of interactions are discussed. If intensity edges are missing, as in smooth-shaded surfaces, the image intensities themselves could be used for stereo matching. Matching primitives other than edges were studied in additional experiments. The results are compared with computer vision algorithms for both single modules and their integration for 3D-vision.

1. Introduction

The problem of deriving a description of a three-dimensional scene from its two-dimensional images on the retina is the inverse of classical optics, wherein one has to find the two-dimensional image (brightness distribution) of a three-dimensional object. While the optics problem can be solved straightforwardly, the inverse problem is much harder to attack because a unique solution does not always exist. Furthermore the solution has to be stable, i.e., depend continuously on the image intensities. Computational studies have provided in recent years promising, although far from complete theories of the processes necessary to solve the ill-posed problem of deriving a three-dimensional scene description from two-dimensional images. It has become clear that a single module is not sufficient to solve this problem. Stereo and motion algorithms, for example, can work well under laboratory-controlled conditions (random dot stereograms and moving sinewave patterns), but quite often make severe errors under more natural conditions where specularity, inhomogeneous illuminations, and occlusion are common. We therefore argue that the analysis of the information processing involved should rely on complex natural images rather than non-complex synthetic images.

1.1. Complex vs Non-Complex Images

The human visual system works much more reliable for complex natural images than for non-complex synthetic images. For example it can analyze complex shapes in a natural scene under quite different viewing conditions but produces often ambiguous solutions for simple line drawings like the Necker cube. Similar observations can be made also for other vision modules like color, stereo and motion. Many illusions occur mainly when only single or few cues are available but are rare in complex natural situations because interaction of different cues can avoid false interpretations. In psychophysics, the study of this interaction can be facilitated by the use of computer graphic systems, which allows convenient control of different cues in complex synthetic images. Shading, for example, can be computed for arbitrary objects and ray-tracing and texture mapping techniques allow to compute synthetic images of three-dimensional scenes which can not be distinguished anymore from natural images (photographs).

Most studies of depth cues, both in psychophysics and in computer vision, deal with the reconstruction of a three-dimensional scene from one isolated cue, the most intensively studied one being stereo[13,19,22]. From the computational point of view, there also exist a number of studies on how to evaluate texture information[1,26], shading[12,15,25], and motion[3,11,33]. There is, however, little knowledge of how the information from these cues can be integrated by the human visual system.

1.2. Classification of Depth Cues

From the large number of cues, from which depth information may be inferred (for review see[3]) three types of cues may be distinguished:

- Primary depth cues that provide "direct" depth information, such as convergence of the optical axes of the two eyes, accommodation and unequivocal disparity cues.
- Secondary depth cues that may also be present in monocularly viewed images. These include shading, shadows, texture gradients, motion parallax, kinetic depth effect, occlusion, 3D-interpretation of line drawings, structure and size of familiar objects.
- Cues to flatness, inhibiting the perception of depth. Examples are frames surrounding pictures, or the uniform texture of a poorly resolving CRT-monitor.

CH2465-3/87/0000/0295$01.00 © 1987 IEEE

In the scope of computational vision, an alternative approach to a classification of depth cues could rely on the observation that different cues require a different amount of preprocessing. For example, convergence and accommodation can be evaluated straightforwardly, whereas stereo disparity requires the previous extraction of some matching primitives from the image. To evaluate occlusion or the apparent size of familiar objects, even more preprocessing is required and the objects may have to be recognized first. Since the recognition of objects might employ other, more easily accessible depth cues, it is clear that a complicated interaction of the different depth cues is to be expected. Only recently, attempts have been made to find general strategies for the integration of all this information in computer vision, e.g. by Poggio and Gamble[29].

1.3. Interaction of Depth Cues

In principle, there are several types of possible interactions between different depth cues, which are not mutually exclusive:

- **Accumulation:** Information from the different modules could be accumulated in a way similar to the (non-linear) summation known from spatial frequency channels (probability summation).
- **Veto:** There can be unequivocal information from one cue that should not be challenged by others. In general primary depth cues should override secondary depth cues.
- **Cooperation:** Especially in the case of poor or noisy cues, the modules might work synergistically.
- **Disambiguation:** Information from one module can be used to locally disambiguate a representation derived from another module. Also, a global ambiguity of depth-order (convex-concave) can occur from cues like shadows or kinetic depth[4].
- **Hierarchy:** Information derived from one cue may be used as raw data for another one.

1.4. Representation of Depth

In principle, there are many different ways to represent depth information. The most straightforward way is to produce a depth-map of all the points in the field of view. Another way is to segment the scene into distinguishable objects and describe the shape of the objects in more abstract terms. For the latter way, different approaches have been tried in the last decade. For example, Marr[17,18] proposed the $2\frac{1}{2}$D-sketch which includes rough distances to surface patches as well as their orientations and Koenderink & van Doorn[14,15] used the tools of differential geometry and related their ideas to Gestalt theories of perception.

For a psychophysical approach to these questions, we studied the depth perceived from computer generated images containing different combinations of depth cues. The shading and stereo cues could be either consistent or contradictory. In contrast to other studies of shape perception[23,32] we did not try to describe the shape by measuring the surface orientation of the displayed objects but rather tried to infer the shape from direct depth measurements of the surface of the objects. This was done by interactively adjusting a depth probe to the surface of an ellipsoidal object as described in the next chapter.

2. Methods

2.1. Computer Graphic Psychophysics

Images of smooth-shaded ellipsoids and flat-shaded polygonal ellipsoidal objects were generated by ray-tracing techniques or with a solid modeling software package (S-Geometry, Symbolics Inc.). The smooth objects were ellipsoids of revolution, the axis of revolution being perpendicular to the display screen. Textures and simple figures could be mapped onto the surface. The polygonal objects were derived from quadrangular tesselations of the sphere along meridian and latitude circles. These were elongated along an axis in the equatorial plane, the axis of elongation again being perpendicular to the display screen. Thus, the two types of objects differed mainly in the absence or presence of edges. As compared to spheres, the objects were elongated by the factors 0.5, 1.0, 2.0, or 4.0. With an original radius of 5 cm, this corresponds to depth values between 2.5 and 20 cm. In the following, all semi-diameters will be given as multiples of 5 cm. Examples for and smooth- and flat-shaded ellipsoids with elongation 1 (sphere) and 4 are shown in Plate 1.

The imaging geometry used in the computations is shown in Fig. 1. It differs from the usual camera geometry in that the image is constructed on a screen which is not perpendicular to the optical axis of the eyes. Note that the imaging geometry and therefore the image itself does not depend on the fixation point as long as the nodal points of the two eyes remain fixed at the positions E_l and E_r, respectively. Images were computed for a viewing distance of 120 cm and an interpupillary separation of 6.5 cm. When a point 10 cm in front of the center of the screen is fixated, Panum's fusional area of ±10 min of arc corresponds to an interval from 4.3 cm to 15.2 cm in front of the screen.

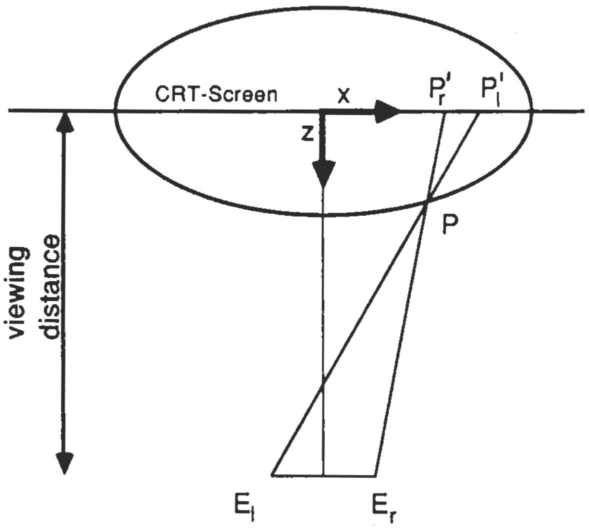

Figure 1 *Imaging geometry.* Projection onto the x-z-plane. Viewing distance is 120 cm. E_l, E_r: nodal points of the left and right eye, respectively. The distance between E_l and E_r is 6.5 cm. A point $P \in \mathbf{R}^3$ is imaged at P_l' for the view from the left eye and at P_r' for the view from the right eye.

For the computation of the smooth-shaded ellipsoids, a ray-tracing operation was performed. We write the equation of the ellipsoid as

$$\mathbf{x}^T \mathrm{A}\mathbf{x} = 1, \qquad \mathrm{A} = \begin{pmatrix} a^{-2} & 0 & 0 \\ 0 & b^{-2} & 0 \\ 0 & 0 & c^{-2} \end{pmatrix}, \qquad (1)$$

where a, b, c denote the semi-diameters. We usually have $a = b = 1$. For a ray from E to P',

$$\mathbf{x} = E + \mu(P' - E), \qquad \mu \in \mathbf{R}^+, \qquad (2)$$

this amounts to the solution for μ of the quadratic equation:

$$(E + \mu(P' - E))^T \; \mathrm{A} \; (E + \mu(P' - E)) = 1. \qquad (3)$$

The image intensity at point P' was computed from this solution according to an ideal Lambertian surface illuminated by parallel light from the z-direction. Note that for a point $\mathbf{x}$ on the surface of the ellipsoid $\mathbf{x}^T \mathrm{A}\mathbf{x} = 1$, the surface normal is simply $\mathrm{A}\mathbf{x}/\|\mathrm{A}\mathbf{x}\|$. The viewing direction and the axes of illumination and of revolution of the ellipsoid were aligned. Since our objects were convex, no cast shadows or repeated scattering had to be considered.

2.2. Experimental Procedure

We displayed either a pair of disparate images or one single (monocular) view of the object as seen from between the two eyes on a CRT-Color-Monitor (Mitsubishi UC-6912 High-Resolution Color-Display Monitor, Resolution (H x V) 1024 x 874 pixels; bandwidth ±3dB between 50 Hz and 50 MHz, short persistence phosphore). The disparate images were interlaced (even lines for left image and odd lines for right image) with a frame rate of 30 Hz. Both disparate and monocular images were viewed through shutter glasses (Stereo-Optic Systems, Inc.) which were triggered by the interlace signal to present the appropriate images only to the left and right eye. The objects were shown in black and white with a resolution of 254 gray-levels. The background was colored in half saturated blue.

Perceived depth was measured by adjusting a small red square-shaped (4 by 4 pixel) depth probe to the surface interactively (with the computer "mouse"). This probe was displayed in interlaced mode together with the disparate images. Thus, the accommodation was the same for viewing both the surface and the probe. Measurements were performed at 45 vertices of a cartesian grid in the image plane in random order. The initial disparity of the depth probe was randomized for each measurement to avoid hysteresis effects. Subjects were asked to move the cursor back and forth in depth until it finally seemed to lie directly on top of the displayed ellipsoidal surface. After some training, subjects felt comfortable with this procedure and achieved reproducible depth measurements. All stimuli were viewed binocularly. Subjects included the authors (corrected vision) and one naive observer.

2.3. Data Evaluation

For each set of 45 measurements we computed two characteristic numbers describing their shape by the following procedure. First, we performed a principle component analysis on all data sets. Variance of the perceived shapes was found mainly (0.95) along two directions in 45-space. The coefficient associated with these principle axes were used to describe the results. Since they are derived from all 45 measurements of a set, their scatter is very small. The results were confirmed by other methods of evaluation, such as computing a least square fit of an ellipsoid to the data.

3. Results

Four different image types were tested:

- Flat-shaded ellipsoid displayed stereoscopically; disparity and edge information (D^+E^+)
- Smooth-shaded ellipsoid displayed stereoscopically; disparity but no edge information (D^+E^-)
- Flat-shaded ellipsoid presented to both eyes identically; no disparity, but edge information (D^-E^+)
- Smooth-shaded ellipsoid presented to both eyes identically; neither disparity nor edge information (D^-E^-).

Each image type was tested for four different elongations (0.5, 1.0, 2.0, 4.0). The subjects did not know the elongation of the displayed objects. Altogether, 253 measurements were performed, each consisting of 45 adjustments of the depth probe to the perceived surface. Results were consistent in all three subjects with differences mainly in the standard deviation. The 16 plots of Fig. 2 show the averaged results of all subjects for the four types of experiments and the four different elongations.

3.1. Classification of the perceived objects

In Fig. 3, the first and the second principle component of the measured surfaces are shown, together with two analytical surfaces which allow an appropriate interpretation of these components. The first principle component is very close to an ideal ellipsoid (or sphere) which appears in Fig. 3c. A model of the second principle component is derived from the depth gradient of the sphere which, in cylindrical coordinates, is $z = r/\sqrt{1 - r^2}$. This 45-vector is orthogonalized (Gram-Schmidt) with respect to the sphere. The result is shown in Fig. 3d; it provides a reasonable fit of the second component. In what follows, we will use this theoretical frame derived from the ellipsoids depth and depth gradient rather than the actual principle components. The corresponding coefficients will be called *perceived elongation* and *deformation*, respectively.

The fact that the perceived surfaces can be classified in terms of the two shapes shown in Fig. 3c,d reflects the occurrence of truncated and conical shapes in the original data, Fig. 2. When depth is inferred from shading or intensity-based stereo, the object appears more conical than an ellipsoid, the depth level of the periphery being pulled towards that of the occluding contour. In the case of intensity-based stereo, this may be due to a lack of matching primitives in the periphery of the smooth ellipsoids, or to problems related to the depth gradient.

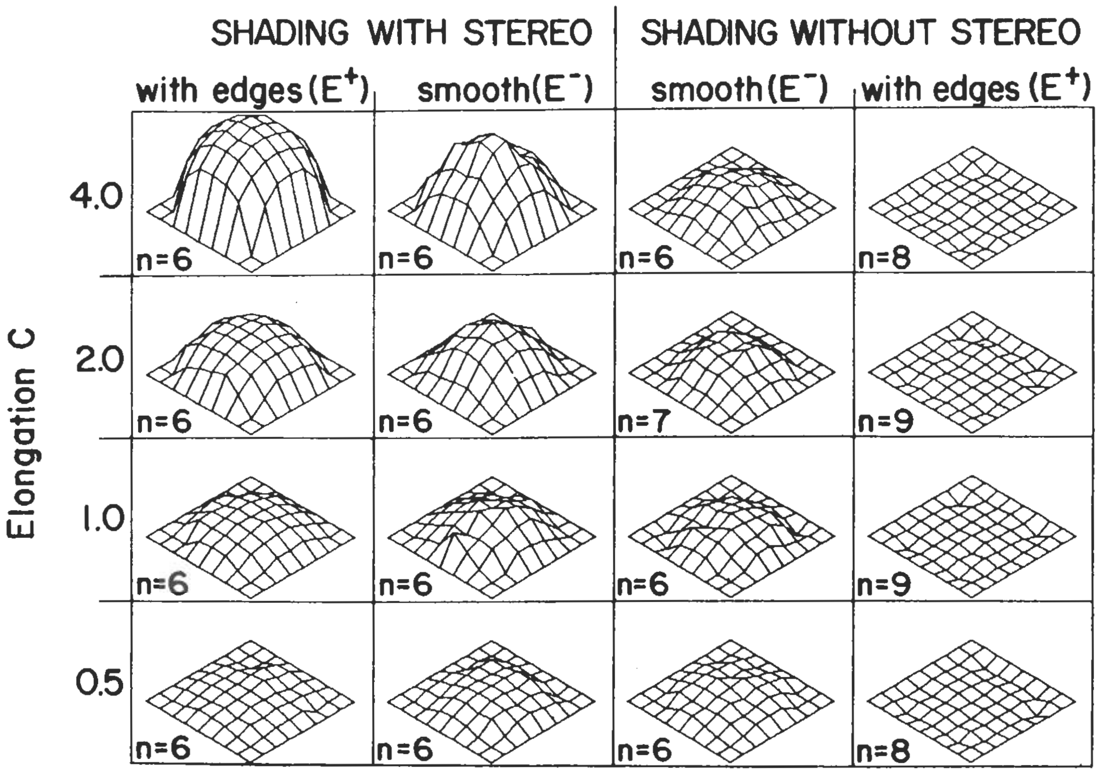

Figure 2 *Perceived surfaces* (depth not drawn to scale) For stimuli, cf. Plate 1a,b. Each plot shows the average of 6 - 9 sessions from three subjects. Perceived depth decreases with the following sequence of cue-combinations: shading, intensity-based and edge-based stereo ($D^+ E^+$); shading and intensity-based stereo ($D^+ E^-$); shading ($D^- E^-$); contradictory shading and edge information ($D^- E^+$). The elongation of the displayed objects is denoted by c.

3.2. Accumulation of Depth Information

The perceived elongation in the consistent images depends on the amount of information available. As can be seen from Fig. 4, the perceived elongation is almost correct when shading, intensity-based and edge-based disparity informations are available (D^+E^+). In the case of smooth-shaded disparate images (D^+E^-), the edges are missing and depth perception is reduced. When shading is the only cue (D^-E^-), perceived elongation is almost independent from the displayed elongation (but see Sect. 3.5). We could not test the case of intensity-based stereo without shading information for obvious reasons. From this, we can draw two conclusions:

- Perceived depth is the greater, the more information is available. That is to say, the information from different cues is accumulated.
- If no edges are present, intensity-based stereo can still provide important information. This might be used for surface interpolation for images with sparse edge information[8].

3.3. Edge-Based Stereo Vetoes Shading

In experiment D^-E^+, two identical images (no disparity) of flat-shaded ellipsoids (edges) were shown. Although shading alone provided some depth information as shown in experiment D^-E^-, the fact that edges occurred at zero disparity was decisive. The perceived depth did not vary with the elongation suggested by the shading (and perspective) information and took slightly negative values which, however, were not significantly different from zero. Since the perceived depth does not change with elongation, we may conclude that edge-based stereo matching overrides shading. This is an example of the veto-relationship mentioned in the introduction. This finding is confirmed by an additional experiment where a small stereo marker was attached to the smooth surface, cf. Sect. 5.1. Note, however, that this veto-relationship might occur only in the locally derived depth map. The global percept of the polygonal ellipsoid is not flat but convex.

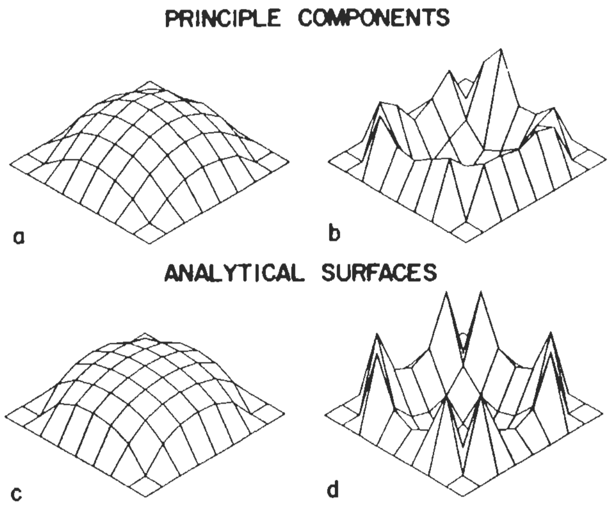

Figure 3 *Classification of the perceived surfaces.* **a,b.** Principle components. **a.** First component, $\lambda_1 = 94\%$. **b.** Second component, $\lambda_2 = 1.4\%$. **c,d.** Analytical surfaces that can be used to interpret the principle component data. **c.** An ideal ellipsoid is almost identical to the first component. The associated coefficient is used as a measure of the *perceived elongation*. **d.** The depth gradient of the ellipsoid leads to an analytical model of the second component. The associated coefficient describes the deviation of the perceived surface from an ellipsoid; it will be called *deformation*. Negative deformations correspond to a more cone-like percept, positive to a more cylindrical surface.

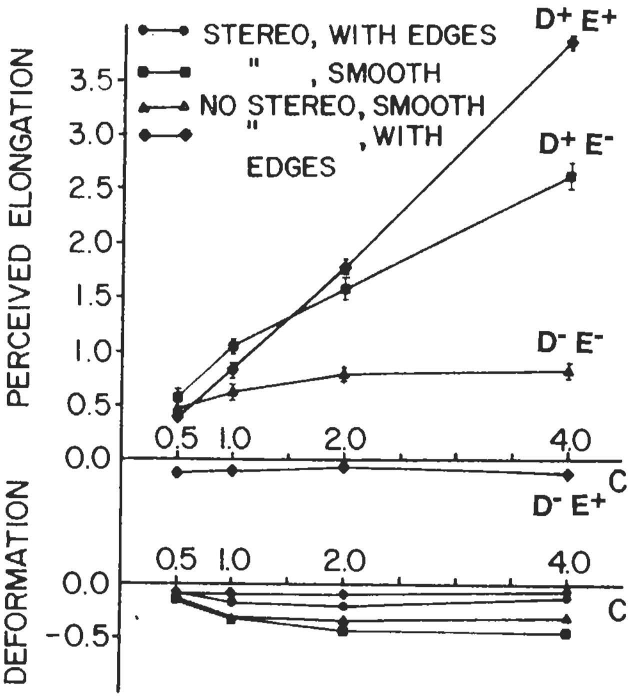

Figure 4 *Perceived elongation and deformation.* **top:** Depth perception improves as the number of available cues increases. The significant separation of the second and third curve (smooth shading with and without disparity) illustrates the influence of disparity information even in the absence of edges. **bottom:** Deformation (cf. Fig. 3b). In the experiments with disparate edges, the coefficients are negligible. In all other experiments, the coefficients are negative, i.e., a more conical surface is perceived.

3.4. Intensity-Based Stereo

If no edge information is available, depth can still be perceived. One could argue that this depth perception is due to shading information that is also present in the images D^+E^-. A comparison of the results (Fig. 4) for smooth-shaded images with and without disparity information, however, establishes a significant contribution of intensity-based disparity information. The curves for D^+E^- and D^-E^- are significantly separated for all elongations except 0.5. One could argue that even in the smoothly shaded images one salient edge is present, namely the occluding contour. However, this boundary was placed in the zero-disparity plane in all experiments. It therefore does not provide depth information. Note that the self-shadow boundary coincides with the occluding contour since illumination was from the front. A control experiment with oblique lighting directions confirmed the findings described here (cf. Sect. 5.1). For some general remarks on images without zero-crossings, see Sect. 4.4

3.5. Intensity-Based Stereo Does Not Veto Shading

If stereo matching can be performed without edge information, the depth cues in the experiment with smooth-shaded non-disparate images (D^-E^-) are contradictory in the sense that shading suggests some depth whereas stereo does not. A similar contradiction occurs in flat-shaded non-disparate images when edge-based stereo is considered. It appears that intensity-based stereo does not veto shading information, as did edge-based stereo in experiment D^-E^+. The contradiction, however, may be the reason for the saturation in the perceived depth from shading (Fig. 4).

4. Discussion

4.1. Images without Zero-Crossings

For the discussion of intensity-based stereo, the absence of zero-crossings in the Laplacians of images of smooth ellipsoids is crucial. First, we show that for an orthographically projected image of a sphere with Lambertian reflection function and parallel illumination, zero-crossings are missing.

Consider a hemisphere given in cylindrical coordinates by the parametric equation

$$z = \sqrt{1 - r^2}. \tag{4}$$

In the special case of a sphere, the surface normal simply equals the radius, i.e.,

$$\mathbf{n} = (r\cos\varphi, r\sin\varphi, \sqrt{1-r^2}). \tag{5}$$

Without loss of generality, we may assume the illuminant direction to be $\mathbf{l} = (0,0,1)$. For the Lambertian surface, we obtain the luminance profile

$$I(r) = I_0\,(\mathbf{l}\cdot\mathbf{n}) = I_0\,\sqrt{1-r^2}, \tag{6}$$

where I_0 is a suitable constant, i.e., the image luminance is again a hemisphere. For the Laplacian of I, we finally obtain

$$\nabla^2 I(r) = I''(r) - \frac{1}{r}I'(r) = -I_0\frac{r^2}{(1-r^2)^{\frac{3}{2}}}. \tag{7}$$

This is a non-positive function of r, with $\nabla^2 I(0) = 0$, i.e., the Laplacian of I has no zero-crossings.

Unfortunately, this result does not hold for ellipsoids with $c \neq 1$. A similar computation for an ellipsoid with elongation c yields

$$I_c(r) = I_0\frac{\sqrt{1-r^2}}{\sqrt{1-(1-c^2)r^2}}, \tag{8}$$

which reduces to Eq. 6 for $c = 1$. In Fig. 5a, where luminance-profiles are plotted for the elongations $c = 0.5, 1.0, 2.0,$ and 4.0, it can be seen that for $c \geq 2$ the curves are no longer convex. That is to say that the second derivatives of these profiles in fact have zero-crossings, and a similar result holds for the Laplacians. However, when filtering with the Laplacian of a Gaussian or with the difference of two Gaussians is considered, it turns out that these zero-crossings are insignificant for the elongations used here. Pixel-based convolutions failed to show the "edges" unequivocally, and even a Gaussian integration algorithm run on the complete function rather than on the sampled array produced no zero-crossings beyond the single-precision truncation error. We therefore conclude that the slight zero-crossings in the unfiltered Laplacian of our luminance profiles do not correspond to significant edges*.

4.2. Receptor Non-Linearities and Image Interpretation

Since the visual system does not work directly on image intensities but on spatially and temporally filtered and compressed (non-linear) signals, the effects of early visual processing in the retina have to be taken into account. Signal compression alone can significantly change image interpretation. Non-linearity in photoreceptors, for example can lead to an illusiary motion perception for time-varying signals that do not entail motion information[5]. In analogy, these non-linearities could induce edge information that is not present in smooth-shaded images. An additional source of zero-crossings not present in our image arrays is the non-linearity of the color monitor. If arbitrary non-linearities are considered, zero-crossings can be induced in every non-constant image, however smooth (e.g. by discretization). We therefore recalibrated the CRT to compensate either for the CRT non-linearity only, or for the non-linearities of both the CRT and the retina.

Retinal non-linearities in both vertebrates[6,24] and invertebrates[16]have been modeled by saturation-type characteristics of the form

$$f(I) = \frac{I}{I + I_{0.5}} \tag{9}$$

where $I_{0.5}$ is a constant, given by the luminance which produces 50% of the maximal excitation. Among other things, $I_{0.5}$ depends on the adaptation of the eye. We repeated experiments D^+E^- and D^-E^-, i.e., those involving smooth-shaded images, with compensation for either monitor non-linearities or the combination of monitor and retina non-linearities with four different choices of the constant $I_{0.5}$. The results did not show significant differences from those obtained without corrections.

Figure 5b shows the luminance profile for an ellipsoid with elongation 4.0, and the effect of a non-linearity given in Eq. 9 for a number of choices of $I_{0.5}$. It turns out that in our experiments, the presumed receptor non-linearities tend to cancel the small zero-crossings rather than to create new ones. This is further support for our assumption that edges cannot be extracted from the smooth-shaded images. Mechanisms relying on zero-crossings either in the original image or in its first neural representation cannot account for the intensity-based stereo performance found in our experiments.

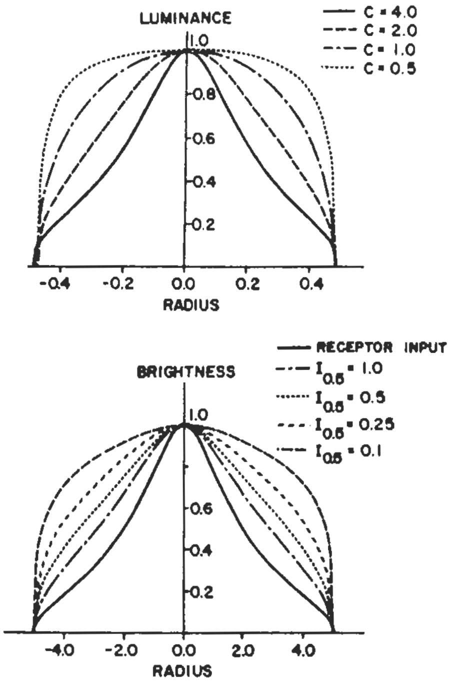

Figure 5 *Luminance profiles.* **a.** Luminance of ellipsoids with different elongations. The functions differ from those given analytically in Eq. 8 only in a slight distortion of the x-axis which is due to perspective rather than orthographic projection. Note that for elongations larger than 2.0, inflections occur. **b.** Simulated perceived brightness profiles for the ellipsoid with elongation 4.0 (the one with the most pronounced inflections in Fig. 5a). Receptor characteristics are accounted for by the non-linear compression described in Eq. 9. The non-linear compression tends to cancel the inflections (which might give rise to zero-crossings) rather than to enhance them.

*However, the absence of edges should be verified using other edge detection mechanisms as well.

5. Relation to Computational Studies

5.1. Edge-Based vs Intensity-Based Stereo

The major finding of this study, as far as single depth modules are concerned, is the strength of depth perception obtained from intensity-based stereo. In computational theory, most studies have focused on edge-based stereo algorithms (for review see[27]). This is due to the overall superiority of edge-based stereo which is confirmed by our finding that edge-based stereo gives a more reliable depth estimate than intensity matching. However, in the absence of edges and in surface interpolation, gray-level disparities appear to be more important than is usually appreciated.

A number of additional experiments was performed to confirm the involvement of intensity-based stereo and to study its relationship to edge-based stereo. First, we measured smooth-shaded ellipsoids (D^+E^-, D^-E^-) with oblique directions of illumination. Light sources were placed in the upper left and the lower right in front of the object (±14° azimuth and $\mp 13.6°$ elevation from the viewing direction). The results of these experiments are depicted in Fig. 6. Note that no depth values were determined in the dark (shadowed) parts of the images. The results confirm the original finding that intensity-based stereo is present and is much stronger than pure shape from shading. Furthermore, when illumination is from the lower right, stereo prevents depth inversions which occasionally occurred in the non-disparate images. One has to keep in mind, however, that in the case of oblique illumination, the self-shadow boundary provides some edge information which improves depth perception in the stereo images and inhibits it in the non-disparate cases. Nevertheless, these data show that our original findings were not critically dependent on the special lighting conditions used.

In a second series of control experiments, we studied the interaction of intensity-based and edge-based stereo. In contrast to the original measurements with flat-shaded ellipsoids where edge-information was distributed all over the surface, we placed a small dark ring (radius 7.5 mm, contrast 0.11) at the tip of the ellipsoid. The stereo disparity of this ring could be chosen independently from the disparity of the shaded surface. Three cases were tested: consistent disparities in ring and shading, no disparities in ring and shading, and a disparate ring in front of a non-disparate shaded image. The first two cases (left and right columns in Fig. 7) confirm the earlier findings of accumulation of depth information and vetoing. Although pure shape from shading yields some depth perception in the periphery, it is vetoed in the center by the non-disparate edge-information. The third case, a stereo ring in front of a non-disparate smooth image (middle columns in Fig. 7) provides information on the mechanisms involved in intensity-based stereo. For the elongations 1.0 and 2.0, the results are equal to those obtained with full stereo information, i.e. one salient stereo token in the center of the object (together with shape from shading) is sufficient to yield the same perception as a complete intensity stereo pair. Since our token coincides with the intensity peak, one could argue that intensity-based stereo is achieved by intensity peak matching. However, for the elongation 4.0, it seems that a single stereo match in the center of the object is not sufficient to produce the same percept as full intensity disparities. The difference between the results for the two subjects corresponds to an ambiguity which was experienced by both observers. For the large elongation, the object appears to consist of a solid base with about half the depth of the ring and a "glass dome" onto which the ring is drawn. While HAM adjusted the depth probe to this 'subjective surface', HHB measured the solid base. We conclude that at least for large disparities, one single token such as the intensity peak is not sufficient to yield the full depth percept.

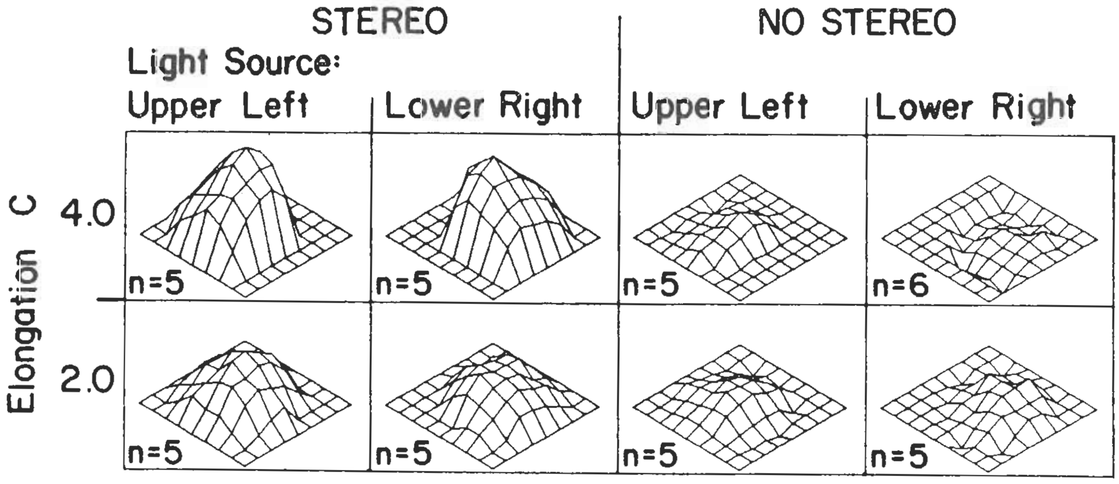

Figure 6 *Perceived surfaces for oblique illuminations* (Format as in Fig.3) For stimuli, cf. Plate 1c. Illumination from the upper left (first and third column) and from the lower right (second and fourth column). No depth was perceived in the self-shadow regions. The data confirm the relevance of intensity-based stereo and show the independence of our findings from the lighting conditions.

Mayhew & Frisby[22] propose a modification of the Marr-Poggio model[19] where matches in the two images may occur before edge-detection is complete. In particular, they discuss peaks in image irradiance as additional matching primitives. However, it appears that their experimental data can be explained with level- (rather than zero-) crossings in the Laplacian of the image irradiance, or with a shift of the zero-crossings due to some prior filtering as well[7]. On the other hand, our experimental data seem to support the notion of intensity peak matching if disparities are not too large. For large disparities other mechanisms have to be taken into consideration.

In another model, Grimson[10] makes explicitly use of binocular shading differences for the interpolation of surfaces between good matches (i.e., between edges). Unfortunately, his model is not directly comparable to our study for the following reasons: First, the information that Grimson's algorithm recovers from shading is the surface orientation along zero-crossings. In our experiments with smooth ellipsoids, the only zero-crossing is the occluding contour of the object where the surface orientation does not depend on the total elongation of the object; it is always perpendicular to the image plane. Second, Grimson's model requires a specular component in the reflectance function of the object. Until now, our experiments explored only purely Lambertian surfaces. We shall, however, include different reflectance functions and lighting conditions in future studies. At any rate, it is an interesting result that human observers are able to evaluate binocular shading information in the Lambertian case. From this we may conclude that a mechanism different from the one proposed by Grimson is involved.

There are different stereo algorithms that do not rely on matching primitives at all. For example, Gennert & Horn[8] have developed a new intensity-based stereo matching method that makes use of a spatially linear transformation to relate gray-levels in the two images. In future work, we are planning to relate these studies to psychophysical investigations as well.

5.2. Shape from Shading

The case of pure shape from shading is studied in our experiment D^-E^-. Ikeuchi & Horn[12] provide a computational theory of shape from shading. Their algorithm starts out from the occluding contour of a given object and successively computes first the surface-orientation and subsequently the depth within the surface. As an example, Ikeuchi & Horn discuss the image of a sphere with a Lambertian reflectance function, illuminated by parallel light from the viewing direction. This example can be directly compared to our experiment. As can be seen from their Fig. 15, the algorithm converges fastest in the vicinity of the occluding contour, i.e., in the periphery of the sphere, whereas errors persist for some iterations in the center. Eventually, however, the correct result is recovered. Interestingly, although the same dependence of the error on the position is found in our experiments, the result that the human observers eventually derive is different: in our experiments, depth from shading is significantly underestimated, although the observation time was not limited.

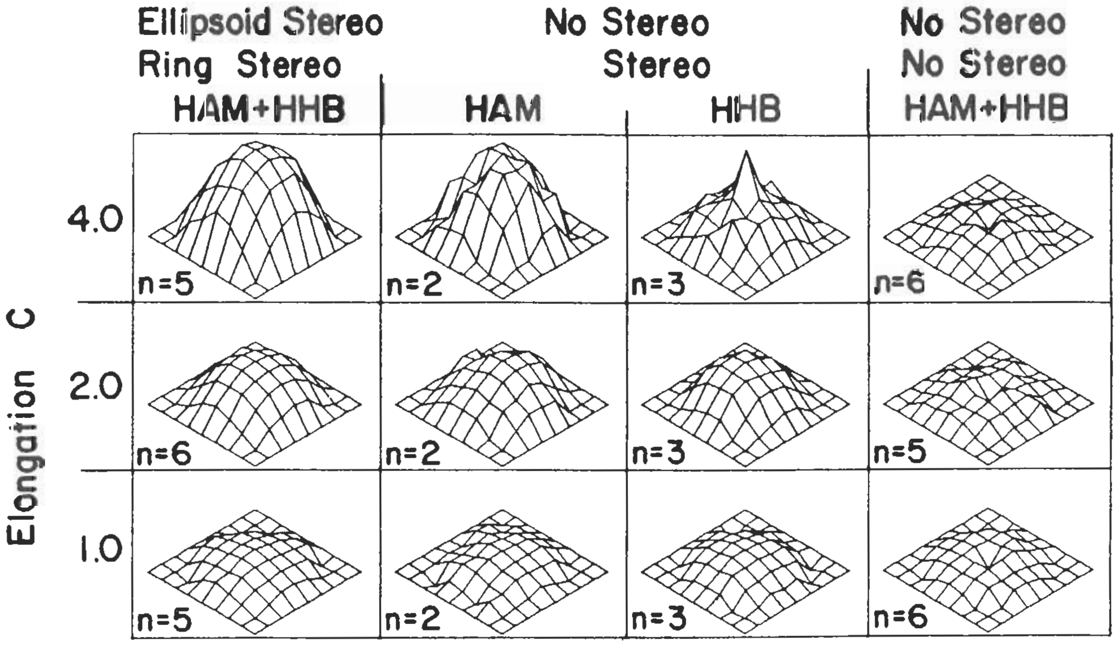

Figure 7 *Perceived surfaces for smooth shading combined with a small stereo marker* (Format as in Fig. 3) Edge-based stereo information cancels shape from shading (right column). When the marker has the correct disparity, intensity-based stereo does not further improve the percept, at least for small elongations. For the elongation 4, the data are ambiguous. (For further discussion, see text.)

The algorithm of Ikeuchi & Horn does make some errors when the required knowledge on the light source position and the reflectance properties of the surface are not known exactly. The types of errors reported from numerical experiments are asymmetric distortions for false assumptions of the light source position and overestimation of depth when false reflectance functions are assumed. In our psychophysical studies, the main errors were of different types. As can be seen from Figs. 3,4, errors included underestimation of elongation and the deformation of the ellipsoidal shape to a more cone-like percept. Asymmetric deformations did not occur even for the obliquely illuminated objects.

5.3. How Useful is Shading as a Cue for Depth?

Todd & Mingolla[23,32] used psychophysical techniques to investigate how observes analyze shape by use of shading cues. According to their results, the human observer makes errors up to 50% in estimating shape from shading. A similar result has been reported in[2], showing that shading of a cylindrical surface could deviate substantially from natural shading before a change in the perceived shape can be detected. This is well in line with our psychophysical findings which suggest that non-disparate shading is a poor cue to shape. It is, however, in contrast to the intuition of artists who use shading as a primary tool to depict objects in depth.

Is it possible that we are not asking the right question when we try to analyze shape with psychophysical tools? Obviously everybody can describe the shape of a vase in a photograph even without any texture on it. In principle, shading can provide only information about surface orientation and not absolute depth measurements. But as Todd and Mingolla have shown, a long training phase is required for subjects to point out the normal surface on simply shaded rigid bodies. And even after the training phase subjects make a lot of errors. A precise measurement of surface-slant and tilt does not seem to be necessary for humans to describe shape. If we do not use slant of surfaces ($2\frac{1}{2}$D-sketch) it seems likely that we use other cues to construct a depth-map of an object.

In the study reported here, we tried to answer this question by measuring the perceived depth directly with a stereoscopically viewed depth probe. This seems to be a much simpler task for the subjects and indeed we did not need a long training phase to obtain consistent depth measurements. Surprisingly, this method worked for shading cues alone (no disparity). This is not obvious, since it involves a cross comparison of supposedly more or less independent modules and also comparison of local (depth probe) versus global (shading) information. On the other hand, our depth probe requires binocular viewing even for non-disparate images (pure shape from shading). The rivalry between shape from shading and intensity-based stereo (cf. Sect. 3.5) may be partly responsible for the poor shape from shading performance. To avoid this we are currently developing a paradigm to measure shape from shading monocularly. With this paradigm we can analyze also other cues, eg. texture gradients and occluding contours (see Plate 1d) which would show similar problems with a local stereo depth probe.

5.4. Interaction of Depth Modules

Concrete predictions as to what types of interactions should occur between different depth cues are still difficult to obtain from computational studies. Therefore, we hope that psychophysical studies will in turn provide useful hints for computational investigations as to how an integration of depth information could work. In this section, we try to relate our results to some of the emerging concepts of visual integration.

Accumulation is a simple type of interaction that can be implemented in a number of different ways. Consider for example Marr's $2\frac{1}{2}$D-sketch[17,18]. Information on surface orientation can be collected from different modules such as shading, texture (density- and deformation-gradient), or 3D interpretations of line drawings. It seems natural that performance improves when more information is available.

Similar results should be obtained with the approach of regularization theory[30]. Originally introduced as a unified theory of a number of different modules in early vision, it is equally suited to model the integration of different modules by joint optimization of different sets of data[31]. Depending on the choice of the particular loss-functions, the described interaction types of accumulation and cooperativity are likely to occur. In fact, it should be possible to infer the form of the minimized functional from the particular type of summation found psychophysically between the involved modules.

More 'asymmetric' types of interaction, such as veto or disambiguation, can be expected from models of surface interpolation[9] that start with reliable depth information typically obtained from disparate edges and employ other modules, especially shading, to improve the interpolation between the sites of the edges (R. Wildes, pers. communication). The combination of edge and shading information is thus similar to the combination of occluding contours and shading in[12]. A similar relationship has been assumed between edge-based stereo and binocular shading (intensity-based stereo)[10].

Recently, Poggio[28] proposed another formalism for the integration of different depth modules, based on a probabilistic approach to optimization by non-convex functionals[20,21]. The advantage of this coupled Markov Random Fields approach over regularization theory lies in the possibility of simultaneous segmentation and (piecewise) smoothing of the image. As far as the experiments discussed here are concerned, the results should not be significantly different from those of regularization. However, if other cues such as occlusion are considered, more complex types of interactions are to be expected from the coupled Markov Random Field approach.

Acknowledgments. We are grateful to E. Hildreth and T. Poggio who contributed to this work and read the draft of the paper. We also like to thank B. Saxberg, and R. Wildes for helpful discussions, I. Bülthoff for being a subject, and J. Little for assistance on the Lisp-Machine. The work was performed during a postdoctoral fellowship of HAM at MIT which was made possible by T. Poggio and a grant from the Deutsche Forschungsgemeinschaft.

References

[1] Bajcsy R. & Lieberman L. Texture gradient as a depth cue. *Computer Graphics and Image Processing* vol. 5, pp. 52-67, 1976

[2] Barrow H.G. & Tenenbaum J.M. Recovering intrinsic scene characteristics from images. In: *Computer Vision Systems*, A. Hanson & E. Riseman (eds.), Academic Press, New York, San Francisco, London, 1981

[3] Braunstein M.L. Depth Perception through Motion. Academic Press, New York, San Francisco, London, 1976

[4] Braunstein M.L., Andersen G.J., Rouse M.W., Tittle J.S. Recovering viewer-centered depth from disparity, occlusion, and velocity gradients. *Perception & Psychophysics*, vol. 40, pp. 216-224, 1986

[5] Bülthoff H.H. & Götz K.G. Analogous motion illusion in man and fly. *Nature*, vol. 278, pp. 636-638, 1979

[6] Dawis M. A model for light adaptation: producing Weber's law with bleaching-type kinetics. *Biol. Cybern.*, vol. 30, pp. 187-193, 1978

[7] Geiger D. & Poggio T. Level crossings and Panum area. *MIT Artif. Intelligence Lab. Memo.* in prep., 1987

[8] Gennert M. & Horn B.K.P. *MIT Artif. Intelligence Lab. Memo.* in prep., 1987

[9] Grimson W.E.L. A computational theory of visual surface interpolation. *Phil. Trans. Roy. Soc. London B*, vol. 298, pp. 395-427, 1982

[10] Grimson W.E.L. Binocular shading and visual surface reconstruction. *Computer vision, graphics and image proc.*, vol. 28, pp. 19-43, 1984

[11] Hildreth E.C. The measurement of visual motion. *The MIT Press*, Cambridge, MA., and London, U.K., 1983

[12] Ikeuchi K. & Horn B.K.P. Numerical shape from shading and occluding boundaries. *Artificial Intelligence*, vol. 17, pp. 141-184, 1981

[13] Julesz B. Foundations of cyclopean perception. *The University of Chicago Press* Chicago and London, 1971

[14] Koenderink J.J. & van Doorn A.J. The internal representation of solid shape with respect to vision. *Biol. Cybern.*, vol. 32, pp. 211-216, 1979

[15] Koenderink J.J. & van Doorn A.J. Photometric invariants related to solid shape. *Optica Acta*, vol. 27, pp. 981-996, 1980

[16] Kramer L. Interpretation of vertebrate photoreceptor potentials in terms of a quantitative model. *Biophys. Struct. Mechanism*, vol. 1, pp. 239-257, 1975

[17] Marr D. Vision. Freeman and Company, San Francisco, 1982

[18] Marr D., Nishihara H.K. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. Roy. Soc. Lond. B*, vol. 200, pp. 269-294, 1978

[19] Marr D., Poggio T. A computational theory of human stereo vision. *Proc. Roy. Soc. London B*, vol. 204, pp. 301-328, 1979

[20] Marroquin J.L. Surface reconstruction preserving discontinuities. *MIT Artif. Intelligence Lab. Memo.* 792, 1984

[21] Marroquin J.L., Mitter S. & Poggio T. Probabilistic solution of ill-posed problems in computational vision. *Proc. Image Understanding Workshop* Baumann (ed.) Scientific Applications International Corp., 1986

[22] Mayhew J.E.W. & Frisby J.P. (1981) Psychophysical and computational studies towards a theory of human stereopsis. *Artificial Intelligence*, vol. 17, pp. 349-385, 1985

[23] Mingolla E.& Todd J.T. Perception of solid shape from shading. *Biological Cybern.*, vol. 53, pp. 137-151, 1986

[24] Naka K.I. & Rushton W.A.H. S-potentials from color units in the retina of fish (*Cyprinidae*). *J. Physiol. (London)*, vol. 185, pp. 536-555, 1966

[25] Pentland A.P. Local shading analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 170-187, 1984

[26] Pentland A.P. Shading into texture. *Artificial Intelligence*, vol. 29, pp. 147-170, 1986

[27] Poggio G. & Poggio T. The analysis of stereopsis. *Ann. Rev. Neurosci.*, vol. 7, pp. 379-412, 1984

[28] Poggio T. Integrating vision modules with coupled MRFs. *MIT Artif. Intelligence Lab. Working Paper* 285, 1985

[29] Poggio T. MIT Progress in Understanding Images. *Proc. Image Understanding Workshop* Baumann (ed.), Scientific Applications International Corp., 1987

[30] Poggio T., Torre V., Koch C. Computational vision and regularization theory. *Nature*, vol. 317, pp. 314-319, 1985

[31] Terzopoulos D. Integrating visual information from multiple sources. In: Pentland, A.P. (ed.) From pixels to predicates, Ablex Publishing Corp., Norwood, N.J. 1986

[32] Todd J.T. & Mingolla E. Perception of surface curvature and direction of illumination from patterns of shading. *Journal of Experimental Psychology: Human Perception and Performance*, vol. 9, pp. 583-595, 1983

[33] Ullman S. The interpretation of visual motion. *The MIT Press*, Cambridge, MA., and London, U.K., 1979

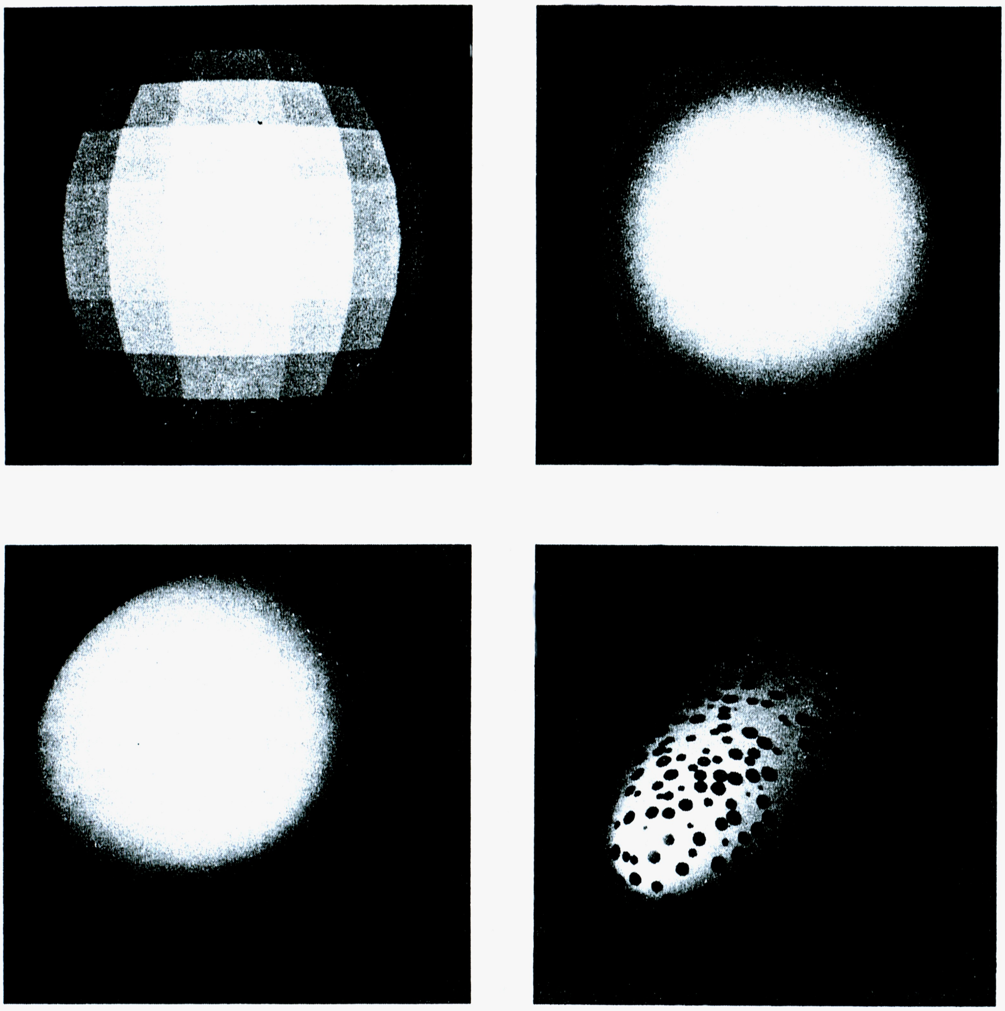

Plate 1 *Testpatterns (all examples for central view).* **a.** flat-shaded sphere ($c = 1$). **b.** smooth-shaded sphere ($c = 1$). **c.** smooth-shaded ellipsoid ($c = 4$) with oblique illumination. **d.** smooth-shaded textured ellipsoid ($a = 0.75$, $b = 0.75$, $c = 1.5$).