PERCEIVING ENVIRONMENTAL STRUCTURE FROM OPTICAL MOTION

Joseph S. Lappin
Vanderbilt University
Nashville, Tennessee

This technical report is for a Workshop on Visually Guided Control of Movement sponsored by NASA Ames Research Center, Rotorcraft Human Factors Branch, June 26 - July 14, 1989.

INTRODUCTION

Guiding a plane or helicopter over a natural terrain cluttered with objects of varying size, shape, position, and altitude requires extraordinary spatio-temporal coordination of the pilot's motor actions with optical information about the structure of the environment. Even though such visual-motor coordination is a commonplace achievement of human perception and action, the effort to understand this capability constitutes one of the main frontiers of contemporary science.

If the quantity of optical information utilized by a pilot in nap-of-the-earth flight, for example, is measured in terms of the number of bits per second per pixel that must be displayed by a high-fidelity, wide-field, realistic simulation by a computer graphics imaging system, then this quantity of information may approach roughly 10 billion bits per second - far beyond the capacity of state-of-the-art technology for acquiring or controlling optical image data. Human pilots, moreover, are able not only to visually acquire such optical information but to transform it in real time to coordinate the six-dimensional trajectory of an aircraft with the rapidly changing constraints of the surrounding environmental scene.

In fact, however, visually acquired optical information cannot yet be quantified. Despite extensive efforts and impressive progress in many relevant areas of science and technology over the past 25 years or so, we still lack a clear understanding of precisely what optical relationships constitute visual information. We cannot yet be certain exactly what properties can in principle or do in fact enable the real-time visual perception of 3D environmental structure.

Generally speaking, one of the most important sources of optical information about environmental structure is known to be the deforming optical patterns produced by the movements of the observer (pilot) or environmental objects. The visual salience and effectiveness of the information provided by such optical image motion has been amply documented by a large body of psychophysical research, by research on computer vision and robotics, and by a considerable body of experience in controlling flight in both real and simulated aircraft. As an observer moves through a rigid environment, the projected optical patterns of environmental objects are systematically transformed according to their orientations and positions in 3D space relative to those of the observer. The

detailed characteristics of these deforming optical patterns carry information about the 3D structure of the objects and about their locations and orientations relative to those of the observer.

The purpose of this paper is to examine specific geometrical properties of moving images that may constitute visually detected information about the shapes and locations of environmental objects. The basic theoretical ideas are the following:

(1) Optical information about environmental structure consists of two qualitatively different types of geometrical relationships which provide information about two different characteristics of environmental structure;

(2) First, information about the intrinsic geometric shape of environmental objects is primarily information about the differential structure of surfaces.

(3) Optical information about the differential structure of environmental surfaces is provided by local properties of the differential structure of moving images of the surfaces. In principle, this local image structure is sufficient to specify the metric structure of a local surface patch (up to a scalar), independent of other information or assumptions about the egocentric distance or orientation of the object relative to the observer.

(4) This information about local surface structure is based mainly on a rotation of the object relative to the observer (around some axis that does not pass through the observer's viewing position). The angular magnitude of this transformation provides a unique one-parameter transformation by which vision can represent the image transformations produced by the motions of environmental objects relative to the observer.

(5) Second, in contrast, the egocentric distance of an object, the distances between separated objects, the orientation of a given surface, and the observer's own location and motion within the environment all involve a qualitatively different aspect of the geometrical structure of the environment, specified by a different geometrical characteristic of the images. These geometrical properties reflect locations and orientations within an abstract 3-D Euclidean framework defined independently of the objects and motions within the space.

(6) Optical information about the structure of this abstract 3-D framework is defined by global properties of the images, specified by the (six) parameters of the perspective embedding of the 2D retinal image into Euclidean 3-space.

THE CONVENTIONAL CONCEPTION OF SPACE

"Space" is commonly regarded as an extrinsic framework in which Euclidean distances may be defined independently of the objects contained in the space. The sizes and shapes of objects, the distances between objects, and the velocities of objects moving in the space are usually considered as defined in relation to this framework, independently of the objects themselves. Thus, the spatial structure of optical data patterns on the retina has usually been represented in relation to the

anatomical arrangement of the photoreceptors - in a 2D space. Similarly, the structure of the 3D environmental space that contains both the observer's retinae and other environmental objects has been regarded as given a priori, independently of the observer and of other objects and events.

This conception of space has several important consequences: First, according to this geometrical representation, the problems of perceiving the structure of objects, their locations and orientation within 3D space, and the observer's own position and motion within this 3D space are all variations on the same general problem - namely, the problem of reconstructing or inferring a 3D environmental space from the 2D retinal optical patterns. As we shall see, however, two different aspects of environmental structure are probably described by two different geometrical characteristics of the retinal optical patterns.

Second, the 2D Euclidean distances on the retina cannot be isomorphic with 3D Euclidean distances in the environment. The mapping from the 2D Euclidean distances on the retina into the 3D Euclidean distances in the environment is necessarily ill-defined. Hence, computational solutions to the problem of recovering the 3D environmental framework require additional extra-retinal information, prior assumptions about natural constraints on the structure of environmental objects and events, and/or processes involving logical inductions and heuristics. Additional spatio-temporal information associated with movements of the observer and objects offers potentially important constraints on computational solutions of this problem, but this additional information is not sufficient to remove the fundamental limitation inherent in the mismatch in dimensionality of the retina and the environment.

Third, descriptions of the retinal optical data patterns are constrained by representing their spatial organization only in relation to the anatomical arrangements among the retinal photoreceptors. Thus, optical patterns are often represented as scalar fields - values of luminance at given spatial positions. Distances and other geometrical relations in the optical patterns are implicitly assumed to be defined only by reference to the retinal anatomy. Another common representation of the optical patterns is as a vector field - a binary relational structure where each vector is specified by two parameters, corresponding to its length and orientation or to the positions of its end-points. The optical velocity field is an example of such a binary relational structure, where the vectors correspond to successive space-time positions or directions and velocities of individual moving points. Because this geometric relational structure of the optical data patterns themselves is quite primitive, the visual computational processes required to recover the geometrical structure of environmental objects have necessarily been complex, time-consuming, and unreliable. For example, a difficult first computational step has been thought to involve solving the so-called "correspondence problem" - matching the spatial position of each point in one image with its corresponding position in a following image. The nature and difficulty of this problem, however, stems from representing the retinal spatial patterns as sets of points, without regard for the intrinsic geometric order of the optical patterns.

The justification for these geometrical representations of the optical patterns, however, has been based on presumption and convention rather than empirical evidence or theoretical analysis. We now examine an alternative representation of the geometric information in optical patterns which significantly simplifies the computational requirements for processing this information.

INTRINSIC GEOMETRY OF SURFACES AND THEIR IMAGES

The geometry of vision is simpler when described in terms of the intrinsic structure of surfaces. Surfaces are connected sets of points in 3D space, but they are just 2D manifolds - positions and changes in position on the surface can be described by just two independent parameters. (A "manifold" is a mathematical structure that is differentiable almost everywhere.) So long as one restricts attention only to points on the surface, geometrical relations in the abstract empty space outside the surface can be ignored as irrelevant to the geometry of the surface itself. Now the retinal images of surfaces are also 2D manifolds. Thus, the geometrical correspondence between surfaces and their retinal images is much closer than that between arbitrary collections of points in 3D Euclidean space and their perspective images on the 2D retinal surface.

The remarkable and important fact is that the differential structures of natural surfaces and their retinal images are isomorphic.[2] (In more technical jargon, we can say that the two structures are "diffeomorphic" - meaning that the mapping from the surface onto its image is one-to-one and differentiable and the same is true for the inverse mapping from the image onto the surface.) This isomorphism means that the differential structure of the retinal images of a surface provide rich and precise information about the differential structure of the environmental surface. This isomorphism holds for images defined by texture, motion parallax, and stereoscopic disparity; and even though it does not actually hold for images defined by illuminance, due in part to the conjoint influences of the directions of illumination and of gaze, the illuminance gradients do provide detailed information about the differential structure of the surface. It follows that the images defined by these various types of optical properties are also isomorphic with one another.

Particularly informative characteristics of the differential structure of a surface are given by its critical points; and a corresponding characterization in the image is provided by the critical values which are the images of the critical points. These critical points are isolated points and curves at which the differential map from the surface onto its image decreases in dimensionality from two to one dimension - at minima and maxima of the height of the surface (at peaks and valleys), at water sheds and water troughs, at parabolic lines or inflections where the curvature changes sign between regions of convexity and concavity, at saddle points, and at the discontinuities associated with sharp corners, where the derivatives of the surface vanish. (An additional set of discontinuities in the image corresponds to occluding and bounding contours where the surface is smoothly curved but the image is discontinuous. The surface positions corresponding to these image discontinuities do not remain constant as the object rotates relative to the observer.) The spatial pattern of these critical points provides a type of skeletal framework describing the topological structure of the surface independent of its orientation relative to the observer. Even in regions of the surface which appear and disappear from view due to changes in occlusion, the pattern of these catastrophic image changes is quite systematic and carries considerable qualitative information about the structure of the surface (see Koenderink & van Doorn, 1976).

IMAGES OF SURFACES DESCRIBED BY COORDINATE TRANSFORMATIONS

The mapping of the differential structure of a local surface patch onto that of its image can be very simply described as a linear coordinate transformation. This linear approximation of the perspective optical projection holds for "infinitely small" surface patches that may be locally approximated by a tangent plane at that location. Thus, the mapping of spatial relations on the surface onto those of its image can be locally described as a linear transformation of the two coordinates of the tangent plane at that location - from the intrinsic surface coordinates onto the retinal coordinates. Because the relative orientation of surface patches on the object and the image changes with the curvature of the surface and with the orientation of the object in the observer's visual field, the parameters of these coordinate transformations vary smoothly over the surface of the object.

Suppose that O2 represents the 2D manifold of the object surface, and that R2 represents the 2D manifold of the observer's retina. Then the linear map $v = O^2 \rightarrow R^2$ is locally specified by the following Jacobian matrix of partial derivatives:

$$V = \begin{bmatrix} \partial r^1 / \partial o^1 & \partial r^1 / \partial o^2 \\ \partial r^2 / o^1 & \partial r^2 / \partial o^2 \end{bmatrix}$$

Thus, suppose that $[dO] = [do^1, do^2]^t$ is a 2×1 column vector that describes an infinitesimal displacement on the surface in terms of two intrinsic coordinates on the object surface, and suppose that $[dR] = [dr^1, dr^2]^t$ is a corresponding description of the image of this vector in terms of the intrinsic coordinates of the retina. Then the transformation between these two coordinate systems produced by the optical projection from the object to its image on the retina is given by the linear equation

$$[dR] = V\,[dO] \tag{1}$$

and the inverse map is given by

$$[dO] = V^{-1}\,[dR] \tag{2}$$

where V is the Jacobian matrix given above. (The form of this equation is independent of the specific coordinate systems used to specify positions on the two manifolds. The coordinates need not intersect at right angles nor even be straight lines; they need only be differentiable and to provide a unique specification of each position on the manifold. The generality of this formulation seems especially relevant to vision, where no specific coordinate system can be assumed beforehand for any given environmental surface, and where the visually effective coordinates of the retina are not known.)

The coordinate transformation specified by the Jacobian matrix V may be understood as a local description of the retinal image. The parameters of this transformation need not be computed from more elementary data; these parameters constitute a representation of the image itself. The four

parameters simply quantify the local densities of the two retinal coordinates relative to each of the two object surface coordinates.

A principal characteristic of this representation is that it is a four-parameter description of the local relational structure of the image. Thus, these four parameters are necessary and sufficient for describing the local structure of the image.

Accordingly, the elementary optical image predicates for perceiving environmental structure from motion consist of the temporal deformation patterns of these four spatial parameters. Although the complexity of this relational structure exceeds that of the scalar or vector fields which are often used for describing such images, this greater complexity reduces the ambiguities associated with the so-called correspondence problem in matching component elements in successive images.

THE METRIC STRUCTURE OF A SURFACE FROM CONGRUENCE UNDER MOTION

The metric[3] structure of an arbitrary surface—providing quantitative measures of lengths, angles, and areas on the surface rather than in abstract empty space—involves an embedding of the surface into Euclidean 3-space. That is, perception of the intrinsic geometry of a surface involves three separate coordinate systems for describing any given infinitesimal displacement on the surface: intrinsic coordinates on the object's surface (O^2), retinal coordinates of the image of the surface (R^2), and Euclidean 3-space (E^3). Thus, we employ three differentiable mappings between three separate manifolds: $O^2 \rightarrow E^3$; $O^2 \rightarrow R^2$; $p: R^2 \rightarrow E^3$. A standard formula in differential geometry, the "first fundamental form," specifies the metric structure of a local surface patch based on its "natural" embedding, n, from O^2 into E^3. By using the chain rule for partial derivatives, we can express the natural embedding n as a composition of the functions v and p – i.e., $n = p \cdot v$ – and this leads easily to an expression for the metric structure of the retinal image of a local patch on the surface on an environmental object.

These relations are easily characterized by matrix equations. Let $[dX] = [dx^1, dx^2, dx^3]^t$ be a 3×1 column vector which specifies the lengths of a given displacement on each of the three orthogonal axes of E^3. Note first that the Pythagorean formula for distance can be written in matrix form as

$$ds^2 = [dX]^t[dX] = \sum_{k=1}^{3} \left(dx^k\right)^2 \qquad (3)$$

Next, we express this equation in terms of the intrinsic coordinates of the object surface. Let $N = \left[\partial x^k / \partial O^1\right]$, $k = 1,2,3$ and $i = 1,2$ be the 3×2 Jacobian matrix which transforms the description of a local surface patch from the O^2 into the E^3 coordinate system. Thus, $[dX] = N\,[dO]$. Now we substitute into the Pythagorean formula to express the quantity ds^2 in terms of the object surface coordinates:

$$
\begin{aligned}
ds^2 &= [N\,[dO]]^t\,[N\,[dO]] \\
&= [dO]^t\,N^t\;\;N\,[dO] \\
&= [dO]^t\,G\,[dO]
\end{aligned}
\tag{4}
$$

$G = N^tN$ is a symmetric 2×2 matrix whose entries are the metric tensor coefficients for the local surface patch,

$$
g_{i\;j} = \sum_k \left(\partial x^k / \partial O^i\right)\left(\partial x^k / \partial O^j\right)
$$

As may be seen there are three independent parameters in this matrix, g_{11}, $g_{12} = g_{21}$, and g_{22}. The values of these parameters remain invariant under rotational transformations of the E^3 coordinate system in which the object is described.

Now we employ the chain rule to find the metric tensor coefficients for the retinal image of the surface patch. By the chain rule we have $N = PV$, where P is the 3×2 Jacobian which embeds the retinal coordinates of the given surface patch into E^3, $P = \left[\partial x^k / \partial r^a\right]$, with a = 1,2. Thus, we have

$$
\begin{aligned}
G &= N^tN = [PV]^t\;[PV] \\
&= V^t\;\;P^t\;\;P\;\;V \\
&= V^t\;\;P^*\;\;V
\end{aligned}
$$

and

$$
ds^2 = [dO]^t\;\;V^t\;\;P^*\;\;V\;\;[dO] \tag{5}
$$

where $P^* = P^tP$ is a symmetric 2×2 matrix with entries

$$
p_{ab} = \sum_k \left(\partial x^k / \partial r^a\right)\left(\partial x^k / \partial r^b\right)
$$

In this construction of the metric tensor coefficients, $G = V^tP^*V$, the components of V are directly specified in the retinal image, and the three parameters of P^* are unknown free parameters which constitute the metric tensor coefficients for the retinal image of the surface patch. The values of P^* are not determined by a single image but can be estimated from optical information associated with movements of the object or observer.

A principal hypothesis in the present theory is that the perceived metric structure of retinally imaged surfaces is derived from invariance of the shape of the surface under rotational motions. This hypothesis contrasts with the more common approach of deriving the local surface structure from the global structure of the space it inhabits—e.g., from the relative depth values of neighboring regions

of the surface. Here, the surface structure is regarded as fundamental, and the structure of the space containing the object is derived from the isometries induced by the motion of the object.

If V and P are respectively the "visual" and "perspective" coordinate transformations for an initial image of a given surface patch, then suppose that U and Q are the corresponding coordinate transformations for a second image of the same surface patch seen from a different observational position. Because the metric structure of this surface patch remains invariant under motions in E^3 (rigid rotations as well as bendings; translations of course have no effect on the differential structure), then we have

$$G = V^t \; P^* \; V \; = \; U^t \; Q^* \; U \tag{6}$$

The parameters of the visual transformations V and U are given directly by the two successive images of the given surface patch, and the parameters of the metric embeddings P* and Q* are unknown parameters which must be found as solutions of these equations. These two sets of perspective parameters represent six unknown parameters, for each of the two images. Unfortunately, the values of these six parameters are not determined by this matrix equation since it involves only four independent equations.

If the values of Q* can be expressed as a one-parameter transformation of the values of P*, say $Q^* = f(P^*)$, where f( ) is the desired one-parameter transformation, then we would require only four independent parameter values as solutions for the four independent equations. The one-parameter transformation that yields these solvable equations is a rotation which moves the surface patch over a surface of revolution (see Guggenheimer, 1963, pp. 272-273). Thus, for example, if the 3D surface is a sphere rotated around an axis through its center, the perspective and metric embeddings, P and P*, of the image of any given surface patch on the sphere would be altered by such a rotation of the sphere, but the transformation of these perspective and metric parameters would be specified simply by the angle of this rotation.

Evidence that vision can indeed obtain sufficient information for perceiving the metric structure of a spherical surface from just two successive views which differ by such a rotational transformation was obtained by Lappin, Doner, and Kottas (1980) and Doner, Lappin, and Perfetto (1984). The displayed surfaces were seen as spherical even though the perspective projection used to display the surface seriously violated the normal perspective for 3D objects seen at that viewing distance, as if the object were seen from a position much closer than that at which it was actually presented. Related results were also reported by Lappin and Fuqua (1983), who found that observers exhibited "hyperacuity" for perceiving the center of the length of an imaginary line segment specified by three collinear dots rotated (about a nondisplayed collinear point) in a plane slanted by randomly varied amounts from the fronto-parallel plane. The surface of revolution in this case was a plane specified by the space-time trajectory of the rotating line segment. As before, this spatial discrimination performance was shown to be unaffected by the degree of polar perspective projection used to display the rotating line segment. Similar findings have also been obtained by Lappin and Love (in preparation) for discriminating the shapes of elliptical forms which were displayed stereoscopically on a plane slanted in depth by varying amounts. Large and variable magnifications of the stereoscopic disparities of the shapes were found to have little or no detrimental effect on discriminations of small differences in the relative shapes of these forms when they were rotated, although the shapes were

essentially indiscriminable when they were stationary. In general, then, the metric structure of these spherical and planar surfaces of revolution seems to have been accurately perceived, independently of the naturalness of the perspective with which they were displayed. Evidently, the perceived metric structure of these forms and spaces was enabled by the rotational transformations of the images of these forms.

Let us now examine the potential geometrical basis for this perceptual achievement. Suppose that PV and QU, respectively, are the perspective embeddings into E^3 of the first and second retinal images of a given surface patch, and suppose that these images are related by a rotation in E^3. Thus, we have

$$Q\ U = F\ P\ V \tag{7}$$

where F is a standard 3×3 rotation matrix. F is determined by three parameters, corresponding to the magnitudes of rotation around each of three previously given orthogonal axes. Any given momentary rotation occurs in only a single plane, however; if one of the orthogonal basis vectors happens to be perpendicular to this plane, then the rotation will have no effect on metric relations in that axis, and the metric relations in the two axes of the plane of rotation will trade off against each other.

Now if we wish to determine only the metric relations of the surface patch, ignoring the specific orientation of the surface relative to some other extrinsic reference system, then we can choose the basis vectors of E^3 so that one is perpendicular to the plane of rotation and the other two lie in the plane of rotation. Thus, the magnitude of rotation can be specified by a single parameter value, and the transformation can be described by a 2×2 matrix. We designate this restricted 2×2 matrix by F_2. Similarly, the changes in the perspective embedding parameters are also restricted to only two of the three axes of E^3, and the equations for the coordinate transformations produced by the rotation can also be described by 2×2 perspective matrices, say P_2 and Q_2. Thus, we can now rewrite Eq. (7) as the following equation involving only 2×2 matrices:

$$Q_2\ U = F_2\ \ P_2\ \ V \tag{8}$$

Because the matrices P_2 and Q_2 each now have an inverse, we can rearrange terms in Eq. (8) to represent the observed image deformation given by V and U in terms of an angular rotation in Euclidean coordinates between P_2 and Q_2:

$$U\ V^{-1} = Q_2^{-1}\ \ F_2\ \ P_2 \tag{9}$$

The left side of Eq. (9) specifies the observed image deformation defined by the two successive images V and U, and the right side is the representation of this deformation as a rotation in E^3. This matrix equation is composed of four independent quadratic equations in four independent parameters. Each of the terms on the left evaluates the relative magnitudes of the partial derivatives involving one of the two retinal coordinates for the second image of the surface patch relative to one of those for the first image of the same surface patch. The corresponding entries in the combined 2×2 matrix on the right side of Eq. (9) evaluate the relative magnitudes of partial derivatives that quantify the embedding of the same pair of retinal coordinates into the two Euclidean coordinates of the plane

of rotation. As the retinal coordinates for the image of a given surface patch expand (contract) in the second image relative to the first image, then the Euclidean embedding of the retinal coordinates contract (expand) in inverse proportion from the first to the second image, so that the metric embedding of the object coordinates of the given surface patch remain constant. (A more detailed presentation of the relevant equations is given by Lappin, in press.)

ROTATION AS THE BASIC TRANSFORMATION FOR PERCEIVING STRUCTURE FROM MOTION

One of the principal hypotheses implicit in these equations is that the angular magnitude of rotation in depth constitutes a fundamental relationship for visually perceiving the transformation between successive retinal images of a moving object. This representation is geometrically valid only in a restricted subset of cases, however - where the trajectory of the surface patch in Euclidean space-time is a surface of revolution. Most object surfaces and most trajectories do not really satisfy this condition. Even when an object rotates, the sequence of positions of most surfaces occupies a volume of space rather than a surface - e.g., consider a rotating cube or a sphere rotating around an axis that does not pass through its center. Accordingly, the perspective embedding of the images of the surface into E3 necessarily varies over time from one image to the next. Moreover, because this volume constitutes a three- dimensional rather than a two-dimensional manifold, the projective visual mapping of this manifold onto its retinal images is no longer diffeomorphic and no longer has a well- defined inverse.

Despite these apparent difficulties, the trajectory of an infinitesimal surface patch on a rotating object usually does approximate a section of a surface of revolution for at least a brief interval of time. The accuracy of the approximation improves as the area of the patch and the interval of time are reduced. For the "infinitesimally" small local patches on which the metric tensor is defined, the thickness of the volume is negligible in relation to the other two dimensions of the surface. Moreover, the neighboring patches on the object's surface have trajectories described by the same angular rotation, differing smoothly only in their radial distances from the axis of rotation. It is the differential structure of these radii of rotation that is the goal of these visual analyses, not the angular rotation parameters as such.

A second apparent limitation of this geometric approximation is that the group of motions in E^3 includes other motions besides rotations in depth. Translational movements of the observer as well as those of objects are common visual events and these transformations of the optic array are potentially important sources of visual information about 3D structure. Two different classes of such translational transformations are pertinent: (a) translations approximately parallel to the direction of gaze, which produce "looming" or divergence of the optical images, and (b) translations approximately perpendicular to the direction of gaze, yielding the classical "motion parallax" cue. Each of these cases is considered below. To anticipate, present evidence suggests that (a) the optical divergence patterns produced by approaching or receding objects are visually ineffective as information about surface structure (though potentially useful as a source of information about egocentric distance); and (b) the motion parallax patterns associated with translations roughly perpendicular to the direction of gaze are visually perceived as if produced by rotation rather than translation.

Translations that coincide with the direction of gaze produce optical flow fields characterized by divergence or "looming". The trajectory of any given point in the retinal image flows in a radial direction away from the so-called "focus of expansion" at a velocity which increases with its nearness to the observer and with its angular deviation from the observer's direction of gaze. Theoretically, the velocity field associated with such an optical flow pattern might provide visually effective information about both the egocentric distance of a given point - about its "time to contact", as Lee (1974) pointed out - and about the observer's direction of locomotion. Thus, the velocity fields associated with such optical divergence patterns might provide information about the relative depths of points on the surfaces of environmental objects.

When the direction of motion and the direction of gaze do not coincide or when the observer changes the direction of gaze during locomotion, the geometrical relations between the velocity field in the image and the distances of points form the observer become more complicated. The retinal image trajectories continue to point toward the vanishing point (the retinal image of the direction of gaze) despite changes in the relative direction of locomotion (see Regan & Beverley, 1982). The velocity field, however, is influenced by the direction of locomotion as well as by the distances of points from the observer. In principle, therefore, the velocity field might provide information about the orientation of a surface relative to the observer, as Prazdny (1983) and Perrone (1989) have indicated. This potential optical information about the structure and orientation of the surface is provided by the spatial derivatives of the velocities rather than by the directions of the image trajectories of the moving points.

So far as I am aware, however, human sensitivity to the differential structure of the velocity fields of these optical divergence patterns has not been shown to be sufficient for discriminating environmental surface structure. Indeed, experiments begun this summer at NASA-Ames by the working group on "Perceiving structure from motion" indicate that human sensitivities to this form of optical information are quite poor.

Observers were asked to discriminate the amount of slant of densely dotted planar surfaces away from the frontal parallel plane when these surfaces were displayed as if seen during translational motion in the direction of gaze - i.e., perpendicular to the display screen and moving toward the surface in question. As an additional visual reference, a ground plane was also visible, parallel to the simulated direction of motion and attached to the slanted surface along a horizontal line in the image.

All the observers of these displays were strikingly insensitive to the orientation of the simulated surface. The angle between the slanted plane and the frontal parallel plane was consistently and grossly underestimated, often by more than 45°. Even when the plane was nearly perpendicular to the direction of gaze, it often seemed slanted toward the observer. Moreover, the observers had little confidence in their judgments of the surface slant, and their judgments were inconsistent. Although we did not determine whether the observers might have been sensitive to the curvature of surfaces portrayed in this way by optical divergence patterns, the insensitivity to the angle between the ground plane and the slanted plane indicated that variations in curvature would not have been very visible. The perceived structure and orientation of the surface seemed to be influenced mainly by the 2D orientations of the image trajectories of the points in the optic flow pattern rather than by the velocity field as such, and these orientations are very poorly correlated with the relative depth or orientation of the surface.

In summary, presently available psychophysical evidence suggests that the divergence of optic flow fields is a poor source of information about environmental surface structure.

Next, we consider the perception of structure from motion parallax produced by translations in directions approximately perpendicular to the direction of gaze. Three aspects of the perception of these optical transformations are noteworthy: (1) These optical transformations produce much more accurate perception of environmental surface structure than the divergence patterns produced by translations parallel to the direction of gaze. (2) These transformations typically appear to have been produced by rotation rather than translation. (3) The tendency to perceive motion parallax patterns as rotation suggests how these perceptions are derived from local retinal information.

One of the questions examined in the experiment begun this summer at NASA-Ames was whether the perception of surface slant would be different when the simulated direction of translation was parallel to the slanted surface, so that the surface flowed horizontally over the display screen without changing the distance between the image and the surface. The motion parallax in these displays consisted of differential horizontal velocities which varied in inverse proportion to the distance of any given point from the observer's focal point-in contrast to the optical divergence patterns produced by moving in the direction of gaze toward the surface. The perceptual consequence of this change in the relative direction of motion was a dramatic improvement in the discriminability of the surface slant. This task was almost trivially easy in comparison with that when the direction of movement was toward the surface. Evidently, the optical information about surface structure was visually much more effective when the viewing position moved parallel to the surface.

Even though the optical transformation in the latter case was produced by a translation, the surfaces in these displays appeared to be rotating, as if the observer were moving around the arc of a large circle centered at some distant point beyond the field of view in the direction of gaze. This subjective impression is consistent with the theoretical idea that the perception of structure from motion is based mainly on optical transformations that are visually represented as rotations of surfaces in depth. Similar impressions of illusory rotation in motion parallax displays have also been observed by M. Braunstein and G. Andersen (personal communications, 1989).

Essentially the same phenomenon is involved in the "stereokinetic effect" (cf. Proffitt, Schmuckler, & Rock, 1989): In the standard demonstration, circular contours are arranged concentrically, centered about points that are laterally displaced from one another along a common invisible line. When these contours are rotated in the frontal parallel plane around a point at the center of the largest circle, the result is a strikingly compelling illusion of depth, with the contours whose centers are farthest from the center of the planar rotation appearing at the greatest depth from the plane of rotation and closest to the observer. (The Exploratorium in San Francisco has several fascinating demonstrations of this illusion.) This illusion results in part from local ambiguities about the direction of motion of the rotating circular contours, where the momentary velocities can be locally described by translations with a significant visible component perpendicular to the contour. Thus, differential local velocities are produced by the series of contours with varying curvatures and distances from the true center of rotation. The perceptual result is that the spatial pattern appears to be rigidly connected in depth and rotating around an axis which is tilted in depth in changing directions rotating around the line of sight. (More direct experimental evidence for this interpretation of the stereokinetic phenomenon will be reported by the author and his colleagues in the future).

The illusory rotation in depth that frequently occurs in these motion parallax patterns suggests some basic characteristics of the visual perception of spatial structure in moving optical patterns: First, the optical information that yields these perceptions seems to be local rather than global. Although rotations and translations may produce optical transformations that are globally quite different, even large global deformations that should in principle accompany the misperception of translations as rotations seem to go unnoticed; patterns which should appear plastic appear instead rigid. Local relations seem to govern the perceived global structure. These local relations are probably associated with spatial relations on connected surfaces.

Second, rotations seem to play a predominant role in visual representations of the optical transformation produced by moving objects and observers. The visual efficacy of these rotational representations may derive from the fact that translations may be locally approximated as rotations; the local first- and second-order derivatives are essentially the same in the two cases. The primacy of the rotational representations may be associated with preservation of local metric structure of a surface patch. Translations on the other hand, usually would not produce significant changes in the local differential structure of the image of a surface patch.

PERCEIVING THE 3-DIMENSIONAL FRAMEWORK OF ENVIRONMENTAL SPACE

So far, we have only examined the visual information about the surface structure of a single environmental object. This class of optical information does not specify the 3D structure of the space that contains that object; it does not specify the orientation of the object relative to either the observer or to some external reference; it does not specify the distance of the object from either the observer or from other separate objects; and it does not specify the location of the observer within this environmental space.

All of the latter properties involve the perspective optical projection from 3D Euclidean space (E^3) onto the 2D image surface (R^2). For any given local surface patch, this projective mapping is locally described by the six parameters of the 3×2 Jacobian matrix P, which embeds the retinal image of the surface patch into a specific coordinate system for E^3. As shown above, the values of these six parameters are not determined by the image transformations associated with rotation of the object; only the metric structure of the surface patch, described by the three metric tensor parameters of the matrix P*, can be derived from the invariance of object's structure under rotation.

The perspective projection from E^3 onto the retinal image is determined by the position of the observer's retina within the environment and by the direction of gaze. Thus, six parameters are needed to specify this perspective projection—three to specify the 3D position of the eye's focal point and three more to specify the fixation point or direction of gaze. These perspective parameters reflect global constraints among the local metric tensor parameters, P*, which embed images of the local surface patches into E^3. Similarly, the global perspective parameters also constrain the values of the local metric tensor parameters.

The perspective projective mapping onto the retinal image, from E^3 onto R^2, induces a version of hyperbolic geometry in the image: An infinite number of parallel lines may intersect at any given

point in the image. The lines which are regarded as parallel in the hyperbolic geometry of the 2D retinal image are the perspective images of lines that are parallel in E^3. The retinal images of lines that are parallel in E^3 converge at a point in the retinal image which is the image of an environmental point infinitely distant from the observer in that direction. Thus, if the observer were standing in a flat open field with no changes in elevation and looking "straight ahead" in a direction parallel to the ground plane, the images of parallel lines extending into the distance parallel to the direction of gaze would converge at a vanishing point on the horizon that is sometimes called the "center of vision." Other sets of parallel lines that extend in other directions parallel to this same ground plane will also converge at other image points that lie along the same horizon line.

This horizon line is important in the geometry of vision because it represents the observer's eye-height: The images of environmental objects above the observer's eye- height lie above the horizon line, and the images of objects below the observer's eye- height lie below the horizon line in the retinal image. Thus, the horizon line divides the retinal image into two regions, one region above and the other below the observer's eye. In most visual environments, however, the horizon line is not explicitly visible. But even when it is not explicitly visible in the retinal image it is implicitly specified by the convergence of image lines that are parallel to the ground plane. Thus, for example, if an observer is standing inside a rectangular room, the four lines defined by the intersections of the side walls with the floor and ceiling project onto the retinal image as four lines which if extended would cross at a single point corresponding to the observer's eye-height.[4]

Now it is useful to consider the retinal surface as a section of a sphere centered at the focal point of the observer's eye. This spherical set of potential visual images constitutes what is known as the optic array (cf. Gibson, 1966; Ch. 10; Cutting, 1986, Ch. 2; Johansson & Börjesson, 1989). Thus, the horizon line extended in all directions a full 360 degrees around the observer would define a great circle in the optic array—the intersection of the sphere with a plane passing through its center parallel to the ground plane, dividing the optic array into two equal hemispherical sections, one containing the images of objects above and the other containing the images of objects below the observer's eye. The shapes and locations of the images in the optic array are determined by the shapes and locations of environmental objects and by the location of the observer's station point within the environment. By definition, the optic array remains invariant under rotations of the eye, though of course the retinal positions of the images of objects are altered as the observer rotates his or her eye to look at various environmental objects.

The spatial relations associated with the position of the retina within the optic array constitute an important scourge of optical information about the orientation of the eye within the environment. Thus, pitch (rotation around the horizontal axis) is described by the elevation of the horizon line in the retinal image; roll (rotation around the "depth" axis parallel to the ground plane) is described by the angular orientation of the horizon line in the retinal images; and yaw (rotation around the vertical axis perpendicular to the ground plane) affects only lateral translation in the retinal image.

The importance of such spatial information for the observer's perceiving his or her orientation within the environment has been demonstrated in several recent psychophysical studies by Johansson and Börjesson (1989), Matin and Fox (1989), and Stoper and Cohen (1989). Changes in the orientation of a structured optical pattern were shown in these studies to exert a large influence on the perceived orientation of the (gravitational) ground plane. Although an explicit horizon line

was not visible in these studies, its location was implied by the convergence of straight lines which appeared to be parallel with each other and with the ground plane. Such optical information associated with the vanishing points of parallel lines is perceptually compelling, capable of dominating contradictory knowledge and proprioceptive information about the direction of gravity.

Although the horizon line is an important aspect of the geometry of vision, its visibility should not be overemphasized: The position of the horizon line in the retinal image is no more visible than the orientation of the ground plane. The horizon line is simply the locus of vanishing points of lines parallel to the ground plane. If one does not know which lines are parallel to the ground plane, then neither does one know the location of the horizon line. Moreover, the horizon line is not a unique structural characteristic of the visual field; any point in the visual field can be the vanishing point of parallel lines in that direction.

Two lines parallel in E^3 but not parallel to the ground plane will converge in the image at a point that does not lie on the horizon line. Thus, for example, sets of parallel lines that are parallel with a vertical plane which is perpendicular to the horizon line would converge in the optic array at a point on a great circle which is perpendicular to the horizon line. If such a vertical arc passes through the center of vision, it divides the visual field into left and right halves, separating objects which lie to the left and right of the observer's direction of gaze. The full set of great circles passing through the center of vision forms a polar configuration radiating from the center of vision. Of course the polar configuration defined by this set of great circles is not unique to the center of vision. Similar sets of great circles can be described at any given point in the image. Every great circle in the optic array is the image of points that are infinitely distant from the observer in some direction parallel to a plane that contains the observer's station point.

Any set of parallel lines in E^3 is parallel to two orthogonal planes through the station point, and they would converge in the image at a point that is an intersection of the corresponding two orthogonal great circles in the optic array. Three parameters are needed to specify each of these vanishing points on the optic array—two parameters to specify any point on the sphere and another parameter to specify the orientation of the two orthogonal great circles at that position.

The physical significance of these vanishing points may be appreciated by considering the images of objects translating through the environment in a constant direction, with no rotation, as seen by an immobile eye of a stationary observer. The trajectories of all points on the object are parallel in E^3 and the images of these trajectories are straight lines which would converge in the image at a specific vanishing point. These converging image lines produced by an object's linear trajectory are also parallel in the hyperbolic geometry of the optic array.

Let us now consider a subset of these images of moving objects—those whose linear trajectories are normal to the spherical optic array, passing through its center at the observer's station point. The stationary images of such a moving object remain congruent with each other in the hyperbolic geometry of the image (even though their size changes in the Euclidean geometry of the image). Congruence is a property possessed by hyperbolic geometry as well as by Euclidean geometry. This congruence of the images of objects under this class of translational motions in E^3 serves to specify the structure of the 3D space in which these motions and objects occur. Figures 2 and 3 illustrate how such hyperbolic isometries in the image specify Euclidean isometries in an environmental 3-space. In

both of these illustrations the structure of the space is described by the congruence or symmetry of stationary objects repeated at varying locations throughout the space and its image. In natural visual environments such isometries are often revealed temporally, by the sequential images yielded by moving objects and moving observers.

When the trajectory of an object moving in relation to the observer is in a direction that does not coincide with the focal point of the eye, then its sequential images are not strictly congruent with one another in the hyperbolic geometry of the visual image sphere. In addition to the changes in size (in the Euclidean sense) produced by the changing distance between the object and the eye, the projected shape of any given surface patch also undergoes a deformation associated with a relative rotation in E^3, as described in the preceding sections of this paper. Such deformations of projected shape may be seen in vertical or diagonal sets of images of the square forms in Figure 2 or in corresponding directional sets of images of the flying-fish-like forms in Figure 3. Despite these projective deformations, the implied congruence of objects under motion in E^3 is immediately visible. That is, a fundamental characteristic of E^3 is its isometry under translations and rotations in any of the three orthogonal directions, and this isometry is displayed by the congruence of the sequential images of objects moving in relation to the observer.

The image transformations produced by the observer's translational motion through the environment are especially informative—about the scaling of the relative sizes of environmental objects, about the scaling of the relative distances of objects from the observer and from each other, and about the relative location and motion of the observer within the environment. These observer-produced image transformations are informative because they yield globally parallel trajectories in E^3 for the relative motions of points on environmental objects and surfaces - trajectories which diverge in the optic array from the horizon line and from the direction of locomotion, thereby describing and scaling the hyperbolic geometry of the optic array as an image of the environment.

Contrary to what might be supposed, these optical relations are more informative about the observer's direction of locomotion and involve disconnected points distributed over varying directions and distances from the observer. Cutting (1986) provides convincing experimental data on this effect, showing that the relative motions of contours lying directly ahead in a plane perpendicular to the path of locomotion yield much less accurate judgments about the direction of locomotion than do those laterally displaced from the path of locomotion. The parallel image trajectories of discrete texture elements distributed over an extended ground plane have been shown to provide sufficient optical information for accurate judgments of the direction of locomotion along linear (Warren, Morris, & Kalish, 1988; Warren & Hannon, 1988) and even curvilinear paths (Warren et al., in press). Recent results of G. J. Andersen (personal communications; Andersen & Dyre, 1989) indicate that similar performance can also be obtained from patterns of discrete points randomly distributed in a 3D cloud-like volume displayed as if the observer were translating through the cloud. Evidently, the global hyperbolic geometry of the optic array and hence the E^3 geometry of the environmental layout are best revealed by the divergence component common to the trajectories of spatially separate contours and edges distributed throughout the visual field.

The present geometric analysis of the optical information for perceiving one's own position and motion within the environment differs in two noteworthy respects from many other contemporary analyses: First, the hyperbolic geometry of the perspective projection of the environmental E^3 space

has been described in terms of the spherical optic array rather than the retinal images. By definition, the optic array remains invariant under rotational eye movements. In contrast, the velocities and directional trajectories of the retinal images of moving environmental objects are significantly affected by the eye movements involved in fixating or tracking various environmental objects. Accordingly, the changing optical patterns associated with the observer's motion through the environment constitute a much less direct source of information about the observer's position and motion in the environment when the spatial structure of these optical patterns is described by reference to the retinal coordinates. As emphasized in the earlier sections of this paper, however, there seems to be no compelling empirical or theoretical requirement for assuming that the spatial relations detected by vision must be referenced to the local retinal coordinates rather than to the neighboring optical pattern. In any case, the present analysis is based on the optic array, in which the spatial structure remains invariant under rotational eye movements.

Second, the present analysis is based on the spatial structure of the optic array and the transformations of this structure produced by moving objects and observers. In contrast, many contemporary theoretical analyses of optic flow have focussed on the velocity field (e.g., Cutting, 1986; Prazdny, 1983). In the present analysis, visible spatial information is provided by the spatial structure associated with parallelism of the image trajectories and with congruence of the successive images of moving objects. That is, if the moving optical patterns are represented as vector fields, where the velocity of a given point is represented by the length of an associated vector, then the visually detected spatial information is assumed to be associated with the directions rather than the lengths of these vectors.

Finally, we note again that the global nature of this optical information about the hyperbolic geometry of the spatial layout of the environment and the observer's position within it contrasts with the local nature of the optical information about the smooth surface structure of a single object. The former global information derives from parallelism and congruence associated with translations, whereas the latter information derives from local deformations produced by rotations. Presumably, the visual mechanisms for detecting these two functionally different classes of geometric information also differ from one another.

REFERENCES

Andersen, G. J. & Dyre, B. P. (1989). Spatial orientation from optic flow in the central visual field. Perception and Psychophysics, 45, 453-458.

Cutting, J. E. (1986). Perception with an eye for motion. Cambridge, MA: MIT Press, Bradford Books.

Doner, J., Lappin, J. S., & Perfetto, G. (1984). Minimal conditions for the visual detection of structure and motion in three dimensions. Science, 209, 717-719.

Gibson, J. J. (1966). The senses considered as perceptual systems. Boston: Houghton-Mifflin.

Guggenheimer, H. W. (1963). Differential geometry. New York: McGraw-Hill. (Reprinted by Dover Publications, New York, 1977).

Johansson, G. & Bîrjesson, E. (1989). Toward a new theory of vision: Studies in wide- angle space perception. Ecological Psychology, 1, 301-331.

Koenderink, J. J. & van Doorn, A. J. (1976). The singularities of the visual mapping. Biological Cybernetics, 24, 51-59.

Lappin, J. S. (in press). Perceiving the metric structure of environmental objects from motion, self-motion and stereopsis. In R. Warren & A. H. Wertheim (Eds.), The perception and control of self-motion. Hillsdale, NJ: Erlbaum.

Lappin, J. S., Doner, J. F., & Kottas, B. L. (1980). Minimal conditions for the visual detection of structure and motion in three dimensions. Science, 209, 717-719.

Lappin, J. S. & Fuqua, M. A. (1983). Accurate visual measurement of three-dimensional moving patterns. Science, 221, 480-482.

Lappin, J. S. & Love, S. R. (in preparation). Visual measures of stereoscopic form from congruence under motion.

Lee, D. N. (1974). Visual information during locomotion. In R. B. MacLeod & H. L. Pick, (Eds.), Perception: Essays in honor of J. J. Gibson. Ithaca, NY: Cornell University Press.

Matin, L. & Fox, C. R. (1989). Visually perceived eye level and perceived elevation of objects: Linearly additive influences from visual field pitch and gravity. Vision Research, 29, 315-324.

Perrone, J. A. (1989). In search of the elusive flow field. Proceedings of the workshop on visual motion. Washington, DC: Computer Society of the IEEE.

Prazdny, K. (1983). On the information in optical flows. Computer vision, graphics, and image processing, 22, 239-259.

Proffitt, D. R., Rock, I., & Schmuckler, M. (1989). Visual bases of the stereokinetic effect. Paper presented at the Fifth International Conference on Event Perception and Action, Miami University. Oxford, OH, July, 1989). Manuscript n preparation.

Rogan, D. & Beverley, K. I. (1982). How do we avoid confounding the direction we are looking and the direction we are moving? Science, 215, 194-196.

Stoper, A. E. & Cohen, M. M. (1989). Effect of structured visual environments on apparent eye level. Perception and Psychophysics, 46, 469-475.

Warren, W. H., Jr. & Hannon, D. J. (1988). Direction of self-motion is perceived from optic flow. Nature, 336, 162-163.

Warren, W. H., Jr., Mestre, D. R., Blackwell, A.W., & Morris, M. W. (in press). Perception of curvilinear leading from optical flow. Journal of experimental psychology: Human perception and performance.

Warren, W. H., Jr., Morris, M.W., & Kalish, M. (1988). Perception of translational heading from optical flow. Journal of experimental psychology: Human perception and performance, 14, 646-660.

NOTES

1. Preparation of this report was supported in part by NIH Grants EY-05926 and P30- EY-08126. The mathematical ideas have been greatly influenced by John Ratcliffe (Dept. of Mathematics, Vanderbilt University) and by discussions with Alan Peters (Dept. of Electrical Engineering, Vanderbilt University).

2. Three technical qualifications bear mention: First, this isomorphism applies, of course, only to the visible regions of the surface. On any given curved surface, especially those surrounding opaque solid objects, some source regions of the surface will generally be occluded from view by other regions of the same surface or by other separate surfaces which are closer to the observer in the same visual direction. Second, this isomorphism also assumes that the scale of resolution with which the environmental surface is described corresponds with that of its image and that this scale of resolution falls within the range of resolution capabilities of the visual system. Third, some surfaces are transparent, resulting in the images of separate surfaces superimposed on the same retinal location. Nevertheless, none of these three technical qualifications should be considered to invalidate the essential correspondence between the two manifolds.

3. The term metric is used in the conventional mathematical sense: A relation m(a,b) between two elements a and b is said to be a metric relation if it satisfies the following axons for all a, b, and c: (i) non-negativity: m(a,b) Ú 0; (ii) symmetry: m(a,b) = m(b,a); (iii) reflexivity: m(a,a,) = 0; (iv) triangle inequality: m(a,c,) Û m (a,b) + m(b,c). Euclidean distances in abstract empty space constitute a special case of metric relations. Of particular interest in the present context are metric relations over curved surfaces, which remain invariant under bending of the surface.

4. I am grateful to Steven Tschantz for pointing this out to me.

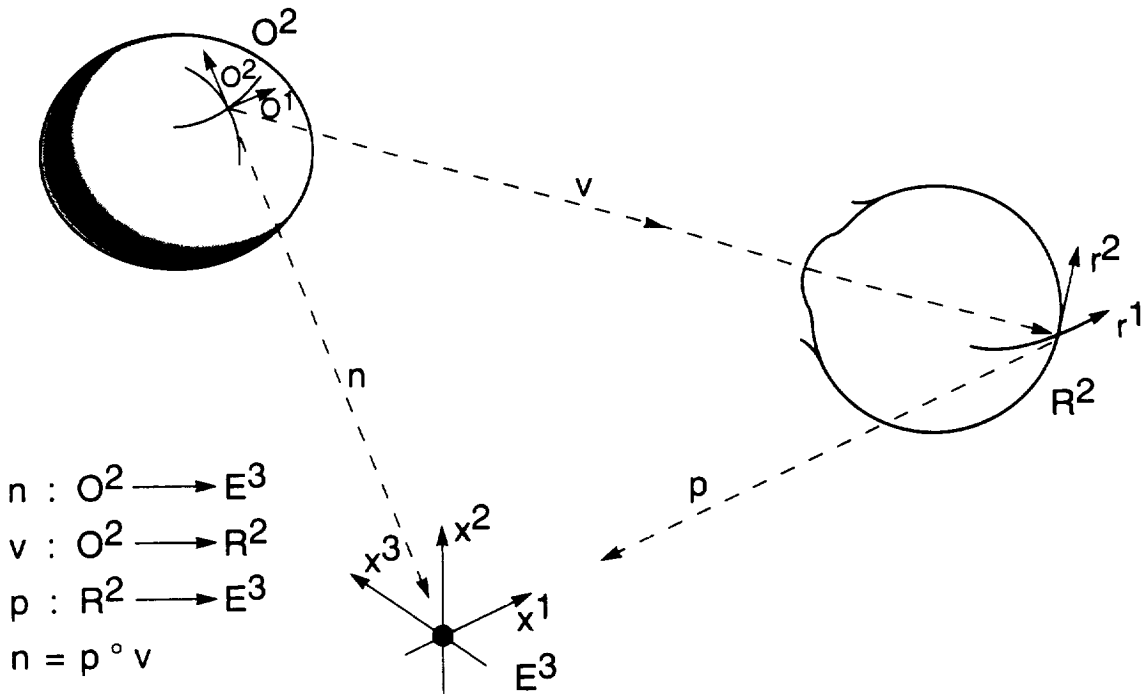

Figure 1. A schematic illustration of the relationship between three separate coordinate systems for describing the surface structure of an environmental object and its image, and the mappings between these coordinate systems.

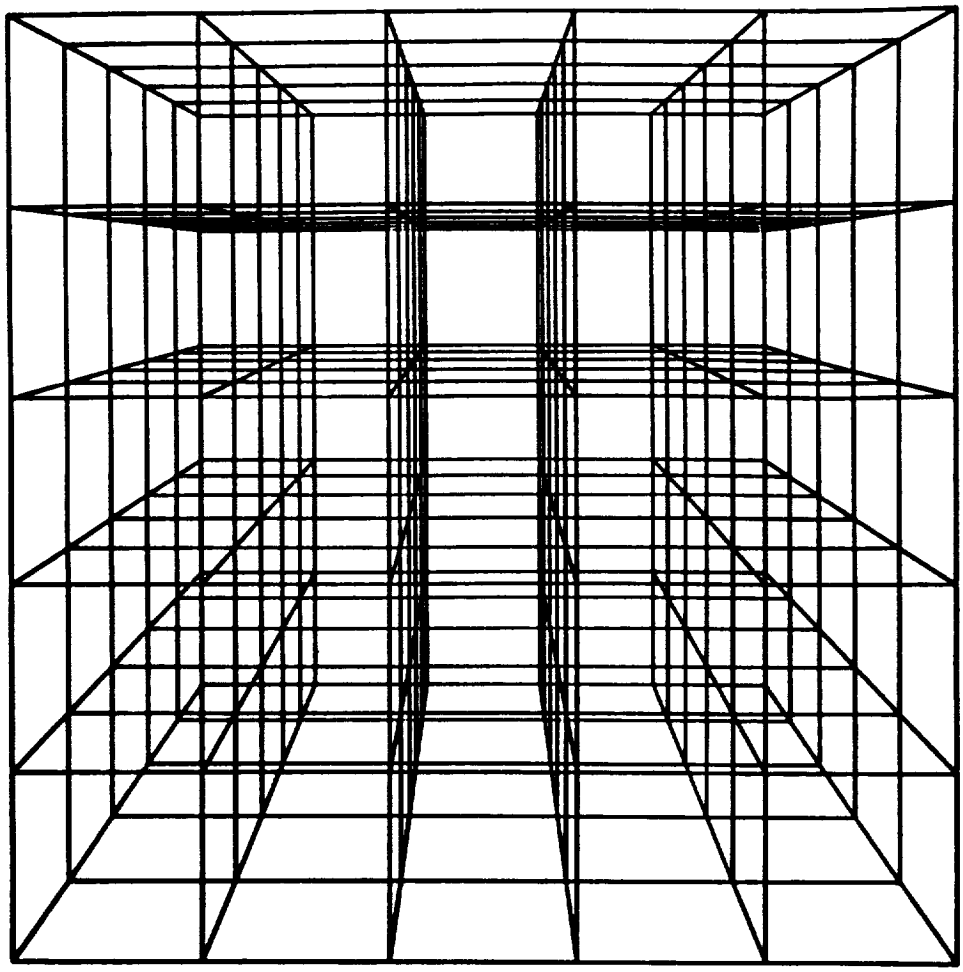

Figure 2. A perspective image of a 5×5 cube. (I am indebted to Steven Tschantz for providing this illustration.)

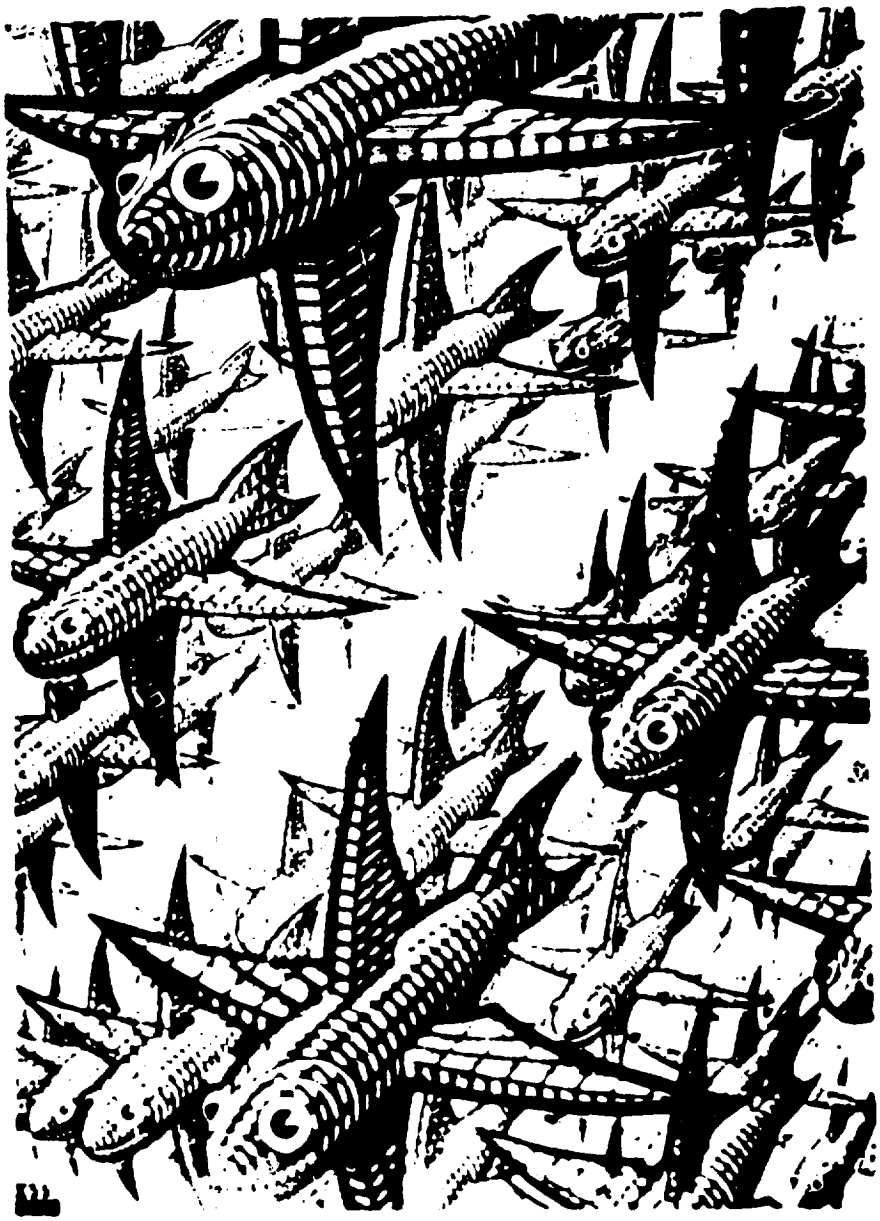

Figure 3. Depth - a wood engraving by M. C. Esher, 1955.