Pergamon

0042-6989(94)E0068-V

*Vision Res.* Vol. 35, No. 1, pp. 117–137, 1995

0042-6989/95 $7.00 + 0.00

# Recovering Three-dimensional Structure from Motion with Surface Reconstruction

ELLEN C. HILDRETH,*† HIROSHI ANDO,‡ RICHARD A. ANDERSEN,§‖ STEFAN TREUE§¶

*Received 9 April 1993; in revised form 3 March 1994*

**This paper addresses the computational role that the construction of a complete surface representation may play in the recovery of 3-D structure from motion. We first discuss the need to integrate surface reconstruction with the structure-from-motion process, both on computational and perceptual grounds. We then present a model that combines a feature-based structure-from-motion algorithm with a smooth surface interpolation mechanism. This model allows multiple surfaces to be represented in a given viewing direction, incorporates constraints on surface structure from object boundaries, and segregates image features onto multiple surfaces on the basis of their 2-D image motion. We present the results of computer simulations that relate the qualitative behavior of this model to psychophysical observations. In a companion paper, we discuss further perceptual observations regarding the possible role of surface reconstruction in the human recovery of 3-D structure from motion.**

Three-dimensional structure-from-motion  Surface reconstruction  Motion interpretation  Motion perception  Temporal integration

## INTRODUCTION

An important tool for the perceptual study of the recovery of three-dimensional (3-D) structure from motion has been the dynamic random-dot pattern, in which a random collection of points is moved across a two-dimensional (2-D) computer display in a way that is consistent with the projection of points from the surface of a 3-D object moving in space. From these displays, the human visual system derives a vivid impression of 3-D structure in the absence of other cues to 3-D shape. Of particular significance to the ideas presented here, human observers can derive the strong sense of a coherent surface, even when viewing only a sparse set of points in motion.

This paper addresses the computational role that the construction of a complete surface representation may play in the recovery of 3-D structure from motion. We first discuss the need to integrate surface reconstruction with the structure-from-motion (SFM) process, both on computational and perceptual grounds. We then present a model that combines a feature-based SFM recovery algorithm with smooth surface interpolation. This model allows multiple surfaces to be represented in a given viewing direction, incorporates constraints on surface structure from object boundaries, and segregates image features onto multiple surfaces on the basis of their 2-D image motion. Finally, we present the results of computer simulations that relate the qualitative behavior of this model to psychophysical observations. In a companion paper (Treue, Andersen, Ando & Hildreth, 1995), we discuss further perceptual observations that reinforce the important role that surfaced reconstruction plays in the human recovery of 3-D structure from motion.

The individual components of our model are based on methods that have been presented earlier in the computational literature of SFM recovery and surface reconstruction. Many of these methods have been tested on a range of synthetic and natural imagery, establishing their viability from a computational standpoint. The main contribution of the work presented here is the integration of these ideas in a framework that can provide one possible account for some seemingly complex phenomena in the perception of 3-D structure from motion. These phenomena have not before been related to the behavior of a computational model in the explicit way that we present in this paper. Our comparison between the performance of the model and human behavior is qualitative, and is intended to support fundamental aspects of the model. In order to conduct computer simulations, details of a model must be specified, but many of these details do

*Department of Computer Science, Wellesley College, Wellesley, MA 02181, U.S.A. [*Fax* 1 617 283 3642].
†To whom all correspondence should be addressed. ATR Human Information Processing Laboratories, Japan.
§Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.
‖Present address: Division of Biology, California Institute of Technology, Pasadena, CA 91125, U.S.A.
¶Present address: Division of Neuroscience, Baylor College of Medicine, Houston, Tex., U.S.A.

not critically effect the qualitative performance of the model. We therefore do not suggest that our model provides a detailed, quantitative account of human behavior.

*Computational motivations*

This section considers the computational motivations for combining SFM recovery with a surface reconstruction process. We first distinguish three terms that will be used in this discussion. Surface *interpolation* refers to a process that fills in unknown surface values between points with known surface data in a way that exactly fits the known data. Surface *approximation* refers to a filling-in process that only approximately fits through known surface points. Both processes implicitly assume that there is only one surface to be constructed. The term surface *reconstruction* refers to a more elaborate process that not only fills in surface values using known data (surface approximation), but also allows multiple surfaces to be represented in a given visual direction, and incorporates processes that detect and interpret surface boundaries, such as those associated with discontinuities in depth. The term "surface interpolation" is often used informally to refer to the component of the overall surface reconstruction process that fills in surface values on each individual surface, but the actual implementation of this process always involves an approximation rather than interpolation algorithm.

From a computational standpoint, there are several reasons to integrate SFM recovery with surface reconstruction. First, many SFM models are feature based, in that they first derive 3-D structure at the locations of image features such as intensity edges, corners and points (Ullman, 1983, 1984; Tsai & Huang, 1981; Barron, 1984; Aggarwal & Martin, 1988; Hildreth, 1988; Waxman & Wohn, 1988; Faugeras, 1993). If one goal of early visual processing is to produce a complete surface representation, in which depth or other surface shape information is known at every image location, then restricting the initial recovery of structure to the locations of features requires a subsequent stage in which a full surface is interpolated between the depths derived at sparse features.

Second, object boundaries play an important role in SFM recovery, and their detection and analysis can be considered a critical aspect of surface reconstruction. There is a need to segment the image into regions corresponding to distinct objects, because the constraints that are used to interpret the 3-D shape of a surface within a single object may differ from the constraints used to infer relative depth between objects undergoing different motions. For example, a single object surface will often obey the *rigidity* assumption, while the motion of multiple objects taken together usually will not. Thus it is important for the SFM recovery to consider whether a given set of features belongs to a single object or multiple objects.

With further regard to object boundaries, when an observer moves relative to a stationary environment, a depth discontinuity gives rise to a motion discontinuity in the image, with the change in projected image speed across the discontinuity proportional to the difference in depth between the viewed surfaces (Longuet-Higgins & Prazdny, 1980; Rieger & Lawton, 1985). In general, when object surfaces can undergo their own motion through space, only the order in depth of two surfaces meeting at a boundary can be inferred from relative image motion alone (Thompson, Mutch & Berzins, 1985). This relative 2-D motion can be used to infer whether the surfaces on either side of a boundary are curved and rotating in depth (Thompson, Kersten & Knecht, 1992). The surface reconstruction process should incorporate both quantitative information regarding the change in depth across an object boundary, and qualitative information, such as the order in depth of two adjacent surfaces, or whether a surface is curved along its boundary.

Third, a surface reconstruction process can facilitate the representation of multiple surfaces in a single visual direction, as in the case of transparency. The presence of multiple image velocities superimposed on a small image region can signal this transparency. If the different velocities of image features are caused by an observer moving relative to stationary transparent surfaces, then the relative image movement can be used to infer their relative depths. Thus it should be possible to segregate visual features onto multiple surfaces based on their image velocities, and to reconstruct multiple surface representations at each location in the image.

A further advantage of incorporating surface reconstruction into SFM recovery is that the reconstructed surface can serve as a later representation of 3-D structure, effectively replacing the depths of individual features on the surface. The construction of such a representation may facilitate tasks such as object recognition and manipulation. The reconstructed surface can also remain intact when individual features appear or disappear, for example, during occlusion by other objects, or during the self-occlusion that occurs along the boundary of an opaque, curved surface rotating in depth.

Finally, computational studies emphasize the difficulty of developing SFM algorithms that behave robustly in the presence of error in the 2-D motion measurements. An interpolation process may reduce the sensitivity of the 3-D recovery algorithm to error, by "smoothing out" small fluctuations in the computed structure due to this error.

The need to address all of the above issues through the integration of a 3-D recovery process with surface reconstruction has also been considered in other areas, such as in binocular stereo (e.g. Hoff & Ahuja, 1987).

*Perceptual motivations*

Perceptual observations suggest a role for surface reconstruction in SFM recovery and indicate how the structure-from-motion process contributes to the perception of complete 3-D surfaces. Consider our overall subjective experience when viewing dynamic random-dot displays. We perceive smooth, complete surfaces in

displays of sparse dots in motion, and discontinuities in the direction or speed of motion of the dots yield a strong impression of object boundaries with associated discontinuities in depth (Kaplan, 1969; Yonas, Craton & Thompson, 1987; Royden, Baker & Allman, 1988). SFM displays that depict points on a rotating surface yield a more compelling sense of 3-D structure than those that depict points distributed within a volume (Green, 1961; Todd, Akerstrom, Reichel & Hayes, 1988; Dosher, Landy & Sperling, 1989b).

Perceptual observations regarding our ability to interpret SFM displays of features with short lifetimes suggest a more direct role for surface interpolation (Husain, Treue & Andersen, 1989; Dosher, Landy & Sperling, 1989a; Landy, Dosher, Sperling & Perkins, 1991; Treue, Husain & Andersen, 1991; Treue *et al.*, 1995). Husain *et al.* (1989) constructed moving dot displays in which the lifetime of individual dots was systematically varied. The subjects' task was to distinguish between a "structured" stimulus, in which the moving points were projected from the surface of a transparent cylinder rotating around a central vertical axis, and an "unstructured" stimulus, in which the projected 2-D motion vectors derived from the structured stimulus were randomly shuffled. Each dot moved for a limited time and then disappeared and reappeared at another random location. Subjects require a total viewing time of several hundred msec to discriminate between the structured and unstructured stimuli, but the lifetime of individual points can be as little as 50–80 msec. Dosher *et al.* (1989a) showed that subjects can discriminate between different complex 3-D surfaces in moving dot displays in which each dot has a lifetime of only two frames, although performance may improve for a larger number of frames (Landy *et al.*, 1991). To account for these phenomena, a mechanism is required that allows the representation underlying the 3-D percept to be preserved when the moving points disappear, and allows new points appearing in different image locations to improve the representation of 3-D shape. One mechanism that satisfies this requirement is a spatial interpolation mechanism, where an "interpolated" representation is preserved when the points disappear, and the movement of newly appearing points improves the quality of the interpolated representation. The analysis presented in this paper shows that the incorporation of 3-D surface interpolation into the SFM recovery provides one possible account of the above phenomena.

Experiments by Treue *et al.* (1991) further support the existence of an interpolation mechanism. In displays with a small set of points in motion (12 points, with limited point lifetimes), if the points disappear and then reappear at the same initial image locations and repeat the same trajectories over time, rather than appearing at new random locations in the display, subjects are unable to distinguish between the structured and unstructured stimuli, even after extended viewing. Improvement of the 3-D percept occurs only when moving points cover a large number of spatial locations, which may be achieved either by presenting a small number of points at many different locations over different times, or by presenting a large number of points at each moment. Intuitively, one expects the result of an interpolation process to improve if data are given at a larger number of image locations. Thus, this experimental observation is qualitatively consistent with our intuition about how an interpolation mechanism would behave. SFM displays containing a larger number of points yield a more compelling, and sometimes more accurate 3-D percept (Green, 1961; Braunstein, 1962; Todd *et al.*, 1988; Dosher *et al.*, 1989b; Sperling, Landy, Dosher & Perkins, 1989).

A number of demonstrations by Ramachandran, Cobb and Rogers-Ramachandran (1988) show interesting interactions between multiple surfaces of moving points, and suggest an influence of the interpretation of object boundaries on perceived 3-D shape. In one demonstration, random dots on the surface of two coaxial, transparent cylinders are superimposed and rotated at different speeds. The two cylinders are the same size, so their surfaces occupy the same locations in 3-D space, but one cylinder is rotated at twice the speed of the other (see Fig. 5). Observers perceive two surfaces in each direction of motion that are separated in depth. This percept can also be obtained when the points have short lifetimes (Treue *et al.*, 1995). If interpolation is required to interpret SFM displays with short point lifetimes, then this observation suggests an ability to interpolate across multiple surfaces simultaneously.

In another demonstration, Ramachandran *et al.* (1988) present a display of two superimposed planes of random dots moving in opposite directions. Points reverse their direction of motion when reaching the edge of the display. Observers perceive the moving points as lying on the surface of a rotating cylinder, rather than two flat planes, suggesting that the interpretation of a boundary as being the edge of a curved surface, which might be inferred from the points "bouncing off" a virtual boundary in the image, can lead to the percept of a more highly curved surface. In a related demonstration, Ramachandran *et al.* found that if the edges of a display of moving points projected from a rotating cylinder are masked so that only a central triangular region is visible, or a narrower portion of the cylinder is visible, then observers perceive a rotating cone or a rotating cylinder with smaller radius, respectively (for related demonstrations, see Aloimonos & Huang, 1991; Thompson *et al.*, 1992: Treue *et al.*, 1995). Again, the visible edges of the display may be interpreted as the curved boundary of a rotating object, leading to the percept of a more highly curved surface.

Andersen (1989) conducted experiments with multiple planes of dots superimposed and translating under perspective projection, in which subjects were asked to evaluate the number of planes present and the relative depths between the planes. Subjects could accurately detect up to only three planes of dots at a time, and the perceived separation of the planes in depth increased with the simulated separation. These observations indicate the maximum number of surfaces that can be

represented simultaneously, and may suggest a role of grouping by speed in the SFM or surface reconstruction processes.

The experiments in our companion paper (Treue *et al.*, 1995) further support the general hypothesis that surface interpolation plays a role in SFM recovery by considering the following consequences of such an approach. First, surface interpolation allows the visual system to fill in surface information at locations that do not contain explicit image features, which may hinder our ability to specify regions on a surface that do or do not contain these features. Second, if the interpolated surface served as a later representation of 3-D structure, replacing that of the 3-D locations of individual features, then our final 3-D percept should follow the behavior of the surface rather than its features. Treue *et al.* (1995) provide evidence that the human recovery of structure from motion exhibits these two consequences.

## STRUCTURE-FROM-MOTION WITH SURFACE RECONSTRUCTION

Our analysis focuses on an approach that combines an independent surface reconstruction process with a feature-based SFM algorithm. This section provides an overview of the model and an initial example of its behavior from computer simulations. Later sections elaborate on the justifications of the model, and present the details of the implementation of this model that was used to conduct our computer simulations.

### *Summary of the model*

The overall structure of the model is illustrated in Fig. 1, and consists of a 2-D motion measurement stage, 3-D SFM recovery, 3-D surface reconstruction and temporal integration. The SFM recovery algorithm is motivated in part by Ullman's incremental rigidity scheme (Ullman, 1984), which builds up an accurate model of 3-D structure through incremental improvements over an extended time period. Ullman's original algorithm maintains an internal model of the structure

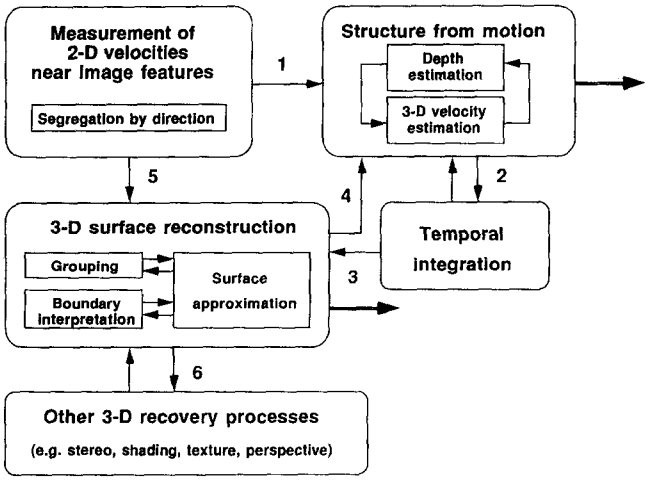


FIGURE 1. Diagram of a model that combines 2-D motion measurement, recovery of 3-D structure from motion, temporal integration and surface reconstruction. See text for details.

of a moving object, which is continually updated as new positions of image elements are considered. The initial 3-D model may be flat, if no other cues to 3-D structure are present, or it may be determined by other 3-D cues available. As each new view of the moving object appears, the algorithm computes new 3-D coordinates for points on the object, which maximize the rigidity in the transformation from the current model to the new positions. In particular, the algorithm minimizes the change in the 3-D distances between points in the model. The use of the rigidity constraint in this way allows the algorithm to interpret both rigid and nonrigid objects in motion. Ullman's original formulation assumed the input to consist of a sequence of discrete frames containing a set of feature points whose positions are obtained by orthographic projection. Extensions to this model use velocity information directly as input, and perspective projection (Grzywacz & Hildreth, 1987).

Limitations of existing SFM algorithms led Ando (1991, 1993) to develop the algorithm that is embodied in the model presented in this paper. The scheme is velocity and feature based, in that the inputs are the 2-D velocities of moving image features extracted continuously over time, and the outputs are the relative depths between these features and their relative 3-D velocities. (The explicit reliance on discrete image features can be relaxed.) The algorithm assumes perspective projection, although it can interpret images obtained under orthographic projection. At each moment, the algorithm alternates in an iterative fashion between computing 3-D velocities that maximize the rigidity of the moving configuration of points, and computing new depths of the features from a set of equations that relate image velocity, 3-D velocity and depth. The computed 3-D velocities are as consistent as possible with the image velocity measurements, while allowing some error in these measurements. Finally, there is an additional temporal integration process that effectively averages the depths computed over an extended time period, using an approach based on Kalman filtering (Gelb, 1974; Anderson & Moore, 1979). This temporal integration yields further improvement of the algorithm in the presence of error in the image motion measurements.

The surface reconstruction stage uses a surface interpolation algorithm that derives a complete surface from sparse depth information that simultaneously fits as closely as possible to the given depth data and is as smooth as possible. Our implementation of this stage is based on an algorithm proposed by Grimson (1981, 1983a), but there are many surface interpolation models that would be adequate for this stage, which we discuss later. The algorithm also incorporates constraints on 3-D shape imposed by object boundaries.

In organizing the overall SFM and surface reconstruction processes, we take into account the need to allow (1) grouping of the features by 2-D direction and speed of motion, (2) the simultaneous representation of multiple transparent surfaces, and (3) the influence of the interpretation of boundaries on the surface reconstruction process. Our model pieces together these mechanisms as

shown in Fig. 1. The various components are not all performed in a fully automatic way in the computer simulations presented in this paper, but the diagram in Fig. 1 indicates one hypothesis regarding where the information obtained by different modules of the system may be built into the overall computation.

Referring to Fig. 1, the measurement of 2-D image velocities in the vicinity of features forms the input to the SFM process (see the pathway labelled "1"), which then iterates between two computations that estimate relative 3-D depths and velocities. Temporal integration further improves upon the depth estimates derived from the SFM process, as indicated by the reciprocal pathways labelled "2". The SFM algorithm may be applied to all of the moving features together, regardless of their direction or speed of motion, or may follow a segregation of features that move in opposite directions of motion within limited image regions.

The new depths of the features derived from the SFM and temporal integration processes are fed into a separate surface reconstruction process that fits surfaces through the known depth points that are as "smooth" as possible (pathway labelled "3"). The result of this stage is a representation of complete surfaces, with explicit depths at each location on a fixed image grid that contains the moving features. In the case of transparency, we maintain a separate representation of each surface, and the surface approximation algorithm operates on each representation independently. Segregation of these surfaces may be derived from a segregation of image features by their 2-D direction of motion and input along the pathway labelled "5". Ultimately, this surface reconstruction process may represent a common integration point for information coming from other 3-D cues, such as stereo, texture and shading (pathways labelled "6").

The "boundary interpretation" component of the surface reconstruction process shown in Fig. 1 detects potential boundaries, for example, by detecting discontinuities in the 2-D direction or speed of motion, and infers whether the boundary is associated with a depth discontinuity and/or edge of a highly curved surface. Cues to the presence and type of boundary can come directly from 2-D motions or from other visual sources such as stereo (pathways labelled "5" and "6," respectively).

The results of the surface reconstruction process may be fed back into the SFM stage (pathway labelled "4"). In our simulations, this pathway is used when points disappear and reappear at different locations in the image. When points disappear, the global surface representations are preserved, and newly appearing points take on an initial depth given by the interpolated surfaces, allowing 3-D surface shape to improve over an extended time while the moving points persist for as little as two frames.

As we noted earlier, the SFM algorithm may be applied to all of the moving features regardless of their direction and speed of motion. However, the surface approximation algorithm is only applied to features undergoing similar or smoothly varying motions. The features are grouped by 2-D direction and speed prior to the surface approximation stage, and interpolation is performed independently on groups of features undergoing different motions in the same region of the image, as in the case of transparency.

*Example: the temporal buildup of 3-D shape*

To illustrate the temporal buildup of 3-D structure that results from the SFM and temporal integration stages, we present the results of a computer simulation using these two processes alone. In the simulation, 60 points were randomly positioned on the surface of a vertically oriented 3-D cylinder, and were rotated continuously around a central vertical axis. The image positions and velocities of the points were computed analytically, and noise was added in the form of Gaussian distributed perturbations of the image velocities. The added noise was scaled by the magnitudes of the velocity components. The initial 3-D structure considered by the algorithm was flat; that is, all points were initially assigned the same depth. At each moment, the depth and 3-D velocity computations were each performed once and new depths were derived using temporal integration.

Figure 2a shows a bird's eye view of the initial flat solution considered by the algorithm. Figure 2b–e shows the solution after 3°, 10°, 35° and 250° of total rotation, respectively. After a short time of only a few degrees of rotation, the computed depths of the moving points occupy a substantial volume that corresponds roughly to the overall extent of the cylinder. Over a more extended time, the 3-D structure improves further, eventually converging to the clear cylindrical shape. Figure 2F shows the result of the surface approximation algorithm applied to the final depths shown in Fig. 2e. The points were grouped by direction of motion prior to the interpolation stage, and surfaces were independently interpolated for the two groups. Separate pictures are shown in Fig. 2f for the front and back surfaces.

For comparison, Fig. 2g shows the results of the SFM and temporal integration algorithms applied to the unstructured stimulus used by Husain *et al.* (1989) and Treue *et al.* (1991). The points are distributed throughout a volume in the solution, and this general structure persists over more extended rotations. There is little difference between the results obtained for the unstructured stimulus and those obtained during the early stages of the analysis of the structured cylinder (Fig. 2b), but eventually the results of the two conditions clearly distinguish themselves, consistent with perceptual behavior.

## THE STRUCTURE-FROM-MOTION PROCESS

In this section, we further justify the choices made in the SFM and temporal integration components of the model, both on computational and perceptual grounds. In particular, we discuss the use of a feature-based algorithm, the use of a velocity-based versus

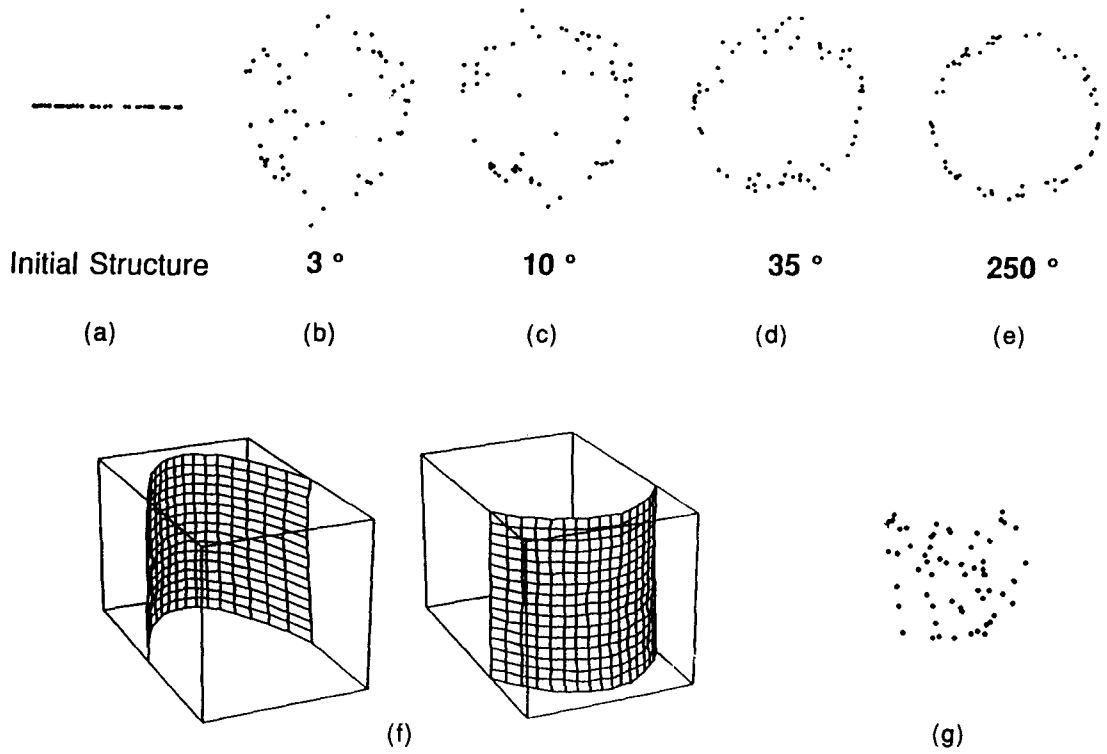


FIGURE 2. Example of the temporal buildup of 3-D structure from motion. 60 points were randomly positioned on the surface of a cylinder and rotated 1° per frame. Gaussian distributed noise was added to the image velocities of the points, with the space constant of the Gaussian, $\sigma = 0.5$, yielding an average relative error of 20% in the velocity components. After each rotation of the points, the depth and 3-D velocity computations were performed once, and roughly 60 iterations were performed within the 3-D velocity computation. (a) Bird's eye view of the initial flat solution. (b–e) The solution after 3°, 10°, 35° and 250° of total rotation, respectively. (f) The result of the surface approximation algorithm applied to the final depths shown in (e). Separate pictures are shown for the front and back surfaces. (g) The results of the structure-from-motion and temporal integration algorithms applied to the *unstructured* stimulus explored by Husain *et al.* (1989)

position-based scheme, and the underlying strategies for temporal integration.

*Using a feature-based structure-from-motion algorithm*

In the case of the moving dot displays, we first recover a "skeleton" 3-D structure at the locations of the points, and later interpolate a surface between the relative depths derived at these points. In the case of more general imagery, we assume that 3-D structure is first recovered in the vicinity of features, such as intensity edges, and is interpolated later to construct a full 3-D surface representation.

There are at least three reasons for concentrating the initial recovery of 3-D structure around the locations of image features. First, the motion measurements directly available in the vicinity of moving features are likely to be more reliable than those obtained in regions of weak or gradual variation of intensity, and subsequently should yield a more reliable recovery of 3-D structure. Second, if there does not exist well-defined features in some region of the image, it may be more appropriate to apply a different SFM strategy that is not based on the use of rigidity in the way we consider here. Suppose, for example, that a given region contains only smooth shading due to a smooth surface curving toward or away from the light source. The movement of the 2-D intensity pattern may not be correlated with the actual movement of the surface. If the light source moves, the pattern of shading will change, yielding a 2-D motion signal, even if the surface remains stationary. Shadows are often associated with slow variations of intensity whose motion is not coupled to the movement of fixed locations on a surface. Thus in general, the interpretation of the 2-D motion of shading and shadows to recover 3-D structure requires a different strategy than the interpretation of the motions of features that are rigidly attached to a moving surface, because the geometry of their motion and its relation to the 3-D shape of the surface is different. Features that are well-localized in the image, such as significant intensity edges, are more likely to correspond to fixed locations on the moving surface, and are therefore more appropriate locations for recovering 3-D structure initially, using the type of rigidity based SFM algorithm that we consider here. Finally, the ability to recover 3-D structure at isolated features, independent of the surface reconstruction process, allows the model to interpret displays of sparse features in motion that do not belong to a well-defined surface, for example, in the case of a random volume of points in motion or a natural image consisting of sparse texture. The SFM process should not rely critically on the presence of a coherent surface.

As in the case of biological vision systems, it is not essential that motion measurements be restricted to instantaneous feature locations. The SFM process can initially use motion information within limited regions

around significant intensity changes in the image. This suggestion is consistent with the proposal of Dosher *et al.* (1989a; also see Landy *et al.*, 1991) that SFM recovery may be based on the outputs of first order motion energy filters, if we assume that these filters yield higher energy in the vicinity of strong intensity variations and that weak motion energy signals do not enter into this recovery.

There are at least three alternatives to the above approach. First, we could obtain initial constraints on 2-D or 3-D motions wherever there is any spatial and temporal variation of intensity (Horn & Schunck, 1981; Negahdaripour & Horn, 1987), but these measurements may be unreliable in regions where the spatial and temporal gradients of intensity are small. Second, we could obtain initial motion measurements in the vicinity of image features, but then immediately "fill in" a 2-D velocity field (Yuille & Grzywacz, 1988). A full 3-D surface could then be computed directly from the dense 2-D velocity or displacement field (e.g. Clocksin, 1980; Longuet-Higgins & Prazdny, 1980; Hoffman, 1982; Burss & Horn, 1983; Koenderink & van Doorn, 1986; Waxman & Wohn, 1988). However, the motion measurements obtained directly at image features will still be more reliable for recovering 3-D structure than those filled in through a 2-D interpolation process. Furthermore, in principle, no explicit surface interpolation is required in this case, but a surface reconstruction process may still be needed later to combine 3-D information from multiple cues. To cope with transparency, it may be necessary to represent multiple dense 2-D motion fields explicitly, in addition to multiple dense 3-D surfaces, which may be cumbersome.

As a third alternative, we could obtain motion measurements only at the locations of image features, and immediately compute a 3-D surface that is simultaneously consistent with the sparse motion measurements and is as smooth and as rigid as possible. The main disadvantage of this approach is its inability to interpret displays in which there is no well-defined surface. Also, our experience with considering this approach in more detail suggests that the two constraints of smoothness of the surface and rigidity of surface motion can compete against one another when applied simultaneously. A more integrated approach of this type has been suggested for the case of recovery of 3-D shape from stereo data by Hoff and Ahuja (1987).

*Positions versus velocities as input to the structure-from-motion recovery*

An issue that arises both for human SFM recovery and for the design of models is the use of the positions versus velocities of moving features as the input to this recovery. Computational methods that use velocities at one instant (e.g. Longuet-Higgins & Prazdny, 1980) are unstable in the presence of error. SFM algorithms exhibit better performance when applied to image sequences with larger spatial and temporal displacements between frames (Ullman, 1984; Yasumoto & Medioni, 1985; Bharwani, Riseman & Hanson, 1986; Grzywacz & Hildreth, 1987; Shariat & Price, 1990). An advantage to using positional information is the ability to relate directly the positions of moving features across longer distances in space and time, which can lead to a more robust recovery of 3-D structure. The SFM and temporal integration models proposed by Ando (1991, 1993) and presented here demonstrate that it is possible to integrate velocity information over time, continuously updating the computed 3-D structure, in a way that is robust in the presence of significant error.

With regard to human processing, studies that reveal our ability to recover structure from motion in displays with short point lifetimes suggest that this recovery may use motion information computed over a limited temporal window of 80–100 msec. (Husain *et al.*, 1989; Dosher *et al.*, 1989a; Landy *et al.*, 1991; Treue *et al.*, 1991). This minimal lifetime is similar to the minimal time required for accurate 2-D velocity estimation (McKee & Welch, 1985). The motion measurements that form the input to SFM recovery may encode image velocity or may capture information such as motion energy (Dosher *et al.*, 1989a; Landy *et al.*, 1991). Over a range of angular velocities of a rotating cylinder, it appears that points must be visible for a minimum time, rather than covering a minimum image displacement (Treue *et al.*, 1991), which may be more consistent with the use of velocities (or motion energy). A strictly position-based scheme may require a minimum displacement of the points to build up 3-D structure. Experiments showing that 3-D judgements can be made from two-frame motion sequences that are oscillated for an extended time period indicate that extended trajectories of moving points are not required for SFM recovery (Todd *et al.*, 1988; Braunstein *et al.*, 1990; Todd & Bressan, 1990). Finally we note that restricted lesions in area MT of monkey visual cortex, believed to play a significant role in the measurement of image motion, disrupts the ability to recovery 3-D structure from motion (Andersen & Siegel, 1990).

An alternative to using velocity information alone is to use both velocity and position information. If velocity measurements were used to guide the tracking of moving points over a more extended time, the limiting factor in this tracking process may still be the ability to measure velocities accurately. Clinical studies of patients with specific cortical lesions indicate that in rare cases, 3-D structure can be perceived in dynamic random dot patterns when the ability to analyze 2-D image velocity information is severely impaired, which may suggest some role for positional information in human SFM recovery (Vaina, Grzywacz & LeMay, 1990).

*Temporal integration through sequential updating of 3-D structure*

The combination of sequentially updating 3-D structure and temporal integration allows the SFM recovery process to interpret nonrigid motions and to cope with substantial error in the image motion measurements. Two factors contribute to the ability of the scheme to interpret nonrigid motions. First, there exists some

model of 3-D structure at each moment, which can change from one moment to the next, allowing the scheme to represent a changing 3-D structure. Second, the SFM algorithm only *maximizes* rigidity, rather than requiring objects to remain strictly rigid over time. If a viewed object changes nonrigidly, then the 3-D model computed by the SFM algorithm will be forced to change over time. From an initial flat structure, the computed 3-D model changes over time through incremental improvements, even for a rigid object in motion.

A number of factors contribute to the ability of the scheme to cope with large error in the image motion measurements. First, the temporal integration effectively averages computed 3-D structures over time, reducing the influence of errors that are temporally uncorrelated. Second, due to the relaxation of the rigidity constraint, large errors result in nonrigid distortions of the computed 3-D structure, rather than a complete breakdown of SFM recovery. Third, the algorithm computes 3-D velocities that only approximately satisfy the image motion measurements, allowing deviation of these measurements from the true projected motions.

The SFM and temporal integration algorithms are efficient in the way they use an extended sequence of continuously changing images. At each moment, the algorithms need only to consider a current 3-D model or average of past estimates, and the new image measurements. The full history of the motions of features over a long time period is implicitly captured in a single current model. This contrasts with an approach that achieves extension in time by storing a large number of images at each moment and processing them simultaneously.

Perceptual observations also support a model of this general type for human SFM recovery. First, many experiments indicate an incremental buildup of perceived 3-D structure over time (Wallach & O'Connell, 1953; White & Mueser, 1960; Braunstein & Andersen, 1984b; Doner *et al.*, 1984; Braunstein *et al.*, 1987; Siegel & Andersen, 1988; Husain *et al.*, 1989; Hildreth *et al.*, 1990; Treue *et al.*, 1991, 1995). The experiments by Hildreth, Grzywacz, Adelson and Inada (1990) suggest continued improvement in the accuracy of SFM judgements up to a second or so, which is significantly longer than the time required to judge image velocity accurately. This time frame is consistent with the observations of Husain *et al.* (1989). Some studies indicate that substantial 3-D information may be derived from only two frames (Todd *et al.*, 1988; Dosher *et al.*, 1989a; Braunstein, Hoffman & Pollick, 1990; Todd & Bressan, 1990), but these studies either oscillate a two-frame image sequence for an extended time period or use extended image sequences in which individual dots have a lifetime of only two frames, so they do not directly address how total viewing time *per se* influences the SFM recovery.

Second, we can cope with a broad range of nonrigid motions, including stretching, bending, transparency, random motions, and more complex types of deformation such as in biological motion displays (Johansson, 1973, 1978; Jansson & Johansson, 1973; Cutting, 1982; Todd, 1982, 1984, 1985; Loomis & Eby, 1988, 1989). Displays of rigid objects sometimes give rise to the perception of distorting objects (Wallach, Weisz & Adams, 1956; White & Mueser, 1960; Braunstein, 1976; Schwartz & Sperling, 1983; Braunstein & Andersen, 1984a; Adelson, 1985; Loomis & Eby, 1988, 1989). Thus the relaxation of the rigidity constraint is an essential component of any model proposed for the human recovery of structure from motion.

Perceptual experiments also suggest that human SFM recovery can cope with significant amounts of image noise (Petersik, 1979; Doner, Lappin & Perfetto, 1984; Todd, 1984, 1985; Husain *et al.*, 1989; Hildreth *et al.*, 1990), which sometimes leads to the perception of nonrigid distortions of the moving object.

Another property of our model is that the current estimate of 3-D structure constrains future estimates of structure. By measuring the consequence of manipulating observers' current perceived structure, Hildreth *et al.* (1990) found some evidence that human SFM recovery exhibits this behavior. This property also allows the algorithm to recover 3-D structure for as few as two or three points in motion, consistent with perceptual observations (Borjesson & von Hofsten, 1973; Lappin & Fuqua, 1983; Braunstein *et al.*, 1987; Petersik, 1987; Hildreth *et al.*, 1990). This is less information than theoretical studies suggest is needed for a unique interpretation of structure using the rigidity constraint alone (Ullman, 1979; Tsai & Huang, 1981).

### *Summary*

To summarize, the SFM and temporal integration components of our model incorporate four key aspects that we have attempted to justify on computational and perceptual grounds: (1) the initial recovery of 3-D structure in the vicinity of image features; (2) the use of 2-D velocity information in SFM recovery; (3) the incremental buildup of 3-D structure over extended time; and (4) relaxation of the rigidity constraint in order to interpret nonrigid objects in motion. These four aspects taken together provide a possible account of a wide range of SFM perceptual phenomena.

## THE SURFACE RECONSTRUCTION PROCESS

This section further justifies a number of aspects of the surface reconstruction process: (1) the separation of the SFM and surface interpolation components of the 3-D recovery process; (2) the grouping of points by their 2-D motion for surface reconstruction; (3) the interpolation of the "smoothest" surface consistent with the sparse 3-D data; and (4) the use of boundary constraints for surface reconstruction.

### *Separating structure-from-motion recovery and surface reconstruction*

Our motivation for separating the surface reconstruction process from the SFM recovery is primarily computational, as available experimental observations do not address this issue directly. The main reason is one of

parsimony. A number of visual cues contribute to the perception of 3-D structure, including binocular disparity, texture gradients, shading, contour shape, and perspective. Many computational models for these processes are feature based, requiring an interpolation stage to fill in 3-D information between image features. It may be more efficient to perform a common surface reconstruction that combines depth or surface orientation information derived from all of the 3-D cues together, rather than building an interpolation mechanism into each visual module [as suggested by Hoff and Ahuja (1987), for example, in the case of stereo vision]. Even if the processes analyzing different 3-D cues produce a dense representation of 3-D shape directly, it may be simpler to analyze the various cues somewhat independently and then integrate 3-D information derived from each cue at a level that constructs a common surface representation, rather than tightly linking the 3-D recovery processes themselves. [It has been suggested that the Bayesian approach provides a useful framework for reliable integration of various depth cues (see Durant-Whyte, 1988).]

A weaker argument for performing interpolation as part of a surface reconstruction process, rather than as part of the computation of the 2-D or 3-D motion fields, is that the additional constraints needed to interpolate a unique 3-D surface may be easier to justify on physical grounds than those required for interpolation of the motion fields. The algorithm proposed by Yuille and Grzywacz (1988), for example, minimizes variation in the velocity field. While minimal variation in 2-D velocity is loosely related to rigidity of 3-D structure (Ullman & Yuille, 1987), the smoothness constraint in 3-D surface reconstruction may be justified more directly on physical grounds (Grimson, 1982, 1983b).

Finally, with regard to perception, we repeat the argument that we can derive a strong sense of 3-D structure when there is only a weak sense of surface, and can recover 3-D shape from both synthetic and natural scenes containing only sparse texture. The separation of SFM recovery from surface reconstruction preserves the ability to cope with such visual patterns.

*Grouping points by 2-D motion*

At some stage, points may be segregated into different groups on the basis of their speed or direction of motion. One can easily justify using this segregation for surface reconstruction. Within a limited region of the image, points belonging to a single surface will tend to move with a roughly similar direction and speed of motion. When different motions are present, they are likely to be due either to the presence of multiple surfaces in the same visual direction, as in the case of transparency, or to two adjacent surfaces undergoing differing motions. This presence of multiple surfaces should be taken into account in building a complete surface representation.

On the other hand, the SFM process specifically uses *relative* motion to infer relative depth. The larger the relative motion between features, the stronger the SFM cue. If we segregate features first on the basis of direction or speed of motion, and recover the 3-D structure of the separate groups independently, the SFM recovery would be inherently less reliable, because it must depend on smaller relative motions between features. For objects such as the rotating transparent cylinder, which have significant variation in speed within each motion direction, points moving in each direction can be grouped together and the SFM algorithm can be applied separately to the two groups, without a significant loss of quality in the final solution. Physiological studies indicate that features moving in opposite directions do not interact during the measurement of motion in area V1, but there is a large degree of inhibitory interaction later in area MT (Snowden, Treue, Erickson & Andersen, 1991), which is critical to SFM recovery (Siegel & Andersen, 1988; Andersen & Siegel, 1990).

*Smooth surface interpolation*

Our model uses a surface interpolation strategy that derives the "smoothest" surface consistent with the depth data given by the SFM recovery. Grimson (1981, 1982, 1983b) presented strong mathematical and physical motivations for this general approach: from a mathematical perspective, this strategy guarantees the computation of a unique surface, and from a physical standpoint, one can show formally and rigorously using the physics of image formation, that the smoothest surface consistent with the sparse depth data derived at image features is most consistent with the image intensity function.

Many algorithms for smooth surface interpolation, including the one proposed by Grimson (1983a), use only local interactions between nearby locations in the image, but allow surface information to propagate over long distances, unless explicit evidence of boundaries is present (see also Terzopoulos, 1986, 1988; for review, see Bolle & Vemuri, 1991). Such algorithms also have no critical dependence on the spatial distribution of the data on the image grid. These factors are important in evaluating the biological feasibility of this approach.

The above models for surface reconstruction depend on the viewpoint of the observer. In principle, if the viewer moves relative to the surface, the computed 3-D shape can distort. Models have been proposed recently that minimize a measure of surface variation that allows a viewpoint invariant reconstruction of visible surfaces (Blake, 1991; Blake & Zisserman, 1991).

When viewing SFM displays that are constructed from smooth surfaces and contain a high density of points, we perceive a smoothly curved surface everywhere. If the density of points is low, the surface often appears to consist of planar facets surrounded by edges that connect local triplets of nearby points. Interpolation algorithms have been proposed that perform a triangulation of sparse 3-D points and fit planar surface patches (Faugeras, LeBras-Mehlman & Boissonat, 1990), but these algorithms use global operations to perform the triangulation. From a biological standpoint, it would be useful to explore extensions to such schemes that use local operations.

*Incorporating boundary constraints*

For two reasons, it is useful to incorporate explicit constraints regarding object or surface boundaries into the SFM and surface reconstruction processes. First, the explicit detection of boundaries allows segmentation into distinct objects or surfaces prior to the SFM recovery. The rigidity constraint that forms the basis of most SFM algorithms is more appropriately applied within the surfaces of single objects. Our relaxation of the rigidity constraint provides some ability to cope with multiple objects moving nonrigidly with respect to one another, but the algorithm would perform better if the locations of boundaries were identified and used to break the rigid links between image features on either side of the boundary. Second, our simulations suggest that it is important to incorporate explicit constraints on surface shape along the boundary of highly curved objects such as cylinders. Otherwise, the smoothness constraint has a tendency to "flatten out" the edges of the object, because this constraint minimizes the variation in depth with respect to distance in the image.

Demonstrations by Ramachandran *et al.* (1988; see also Aloimonos & Huang, 1991) and Thompson *et al.* (1992) suggest that object boundaries play a significant role in human SFM recovery. In particular, the presence of points bouncing off a virtual border in the image can lead to a percept of surface curvature, even when there exists no spatial variation in the speed of moving features near the border (see also, Treue *et al.*, 1995). The presence of a stationary boundary with points moving toward or away from the boundary and continually appearing and disappearing along the boundary can also suggest surface curvature (Thompson *et al.*, 1992).

## DETAILS OF THE MODEL

This section presents further details of the structure-from-motion, temporal integration and surface reconstruction components of the model.

*The structure-from-motion algorithm*

The SFM algorithm proposed by Ando (1991, 1993) and incorporated in our model is a velocity-based algorithm that builds upon earlier formulations of Ullman's incremental rigidity scheme (Ullman, 1984; Grzywacz & Hildreth, 1987). The formulation uses perspective projection, allows error in the image velocity measurements, allows the current estimates of depth to be modified, rather than assuming the current 3-D model to be fixed at each moment, and allows variable weighting of the strength of rigidity between pairs of image features. Ullman's incremental rigidity scheme has been tested on both synthetic and natural images (see also, Hildreth, 1988).

To describe the algorithm in more detail, let $(x_i, y_i)$ and $(\dot{x}_i, \dot{y}_i)$ denote the 2-D position and velocity of the $i$th point and let $(X_i, Y_i, Z_i)$ and $(\dot{X}_i, \dot{Y}_i, \dot{Z}_i)$ denote its 3-D position and velocity in space [for simplicity, we drop the argument $(t)$]. If we assume perspective projection with a focal length of one, then

$$(x_i, y_i) = \left(\frac{X_i}{Z_i}, \frac{Y_i}{Z_i}\right) \tag{2}$$

and

$$(\dot{x}_i, \dot{y}_i) = \left(\frac{\dot{X}_i - x_i\dot{Z}_i}{Z_i}, \frac{\dot{Y}_i - y_i\dot{Z}_i}{Z_i}\right). \tag{2}$$

At each moment, the algorithm estimates the depths $Z_i$ and 3-D velocities $(\dot{X}_i, \dot{Y}_i, \dot{Z}_i)$ that minimize a cost function consisting of two terms:

$$E_D + \lambda E_R \tag{3}$$

where $E_D$ describes the total error in the fit of the estimates to the 2-D velocity measurements, $E_R$ describes the total deviation from rigidity implied by the new estimates, and $\lambda$ is a constant that captures the trade-off between the two terms. The data term $E_D$ attempts to make the left and right sides of equation (2) above as similar as possible, and therefore minimizes the squared difference between the two. In addition, there may be variation in the confidence or reliability attributed to individual velocity measurements. The date term is then written as follows:

$$E_D = \sum_i [\beta_{xi}(\dot{x}_i Z_i + x_i\dot{Z}_i - \dot{X}_i)^2 + \beta_{yi}(\dot{y}_i Z_i + y_i\dot{Z}_i - \dot{Y}_i)^2] \tag{4}$$

where $\beta_{xi}$ and $\beta_{yi}$ are the weights associated with individual velocity measurements.

To derive the term that captures deviation from rigidity, let $l_{ij}$ denote the 3-D distance between two points $i$ and $j$. The change in this 3-D distance over time is given by $\dot{l}_{ij}$. We compute a set of 3-D velocities that minimizes the total change in 3-D distances between all pairs of points on the object. Weighting factors $w_{ij}$ capture the strength of the ridigity of the connection between point $i$ and point $j$. We therefore have:

$$E_R = \sum_{ij} w_{ij}(\dot{l}_{ij})^2. \tag{5}$$

The weighting factors $w_{ij}$ may depend inversely on the 3-D distance between two points $i$ and $j$ in the current 3-D model, possibly using a Gaussian function of distance. In our computer simulations, these weights were set to one for all pairs of features.

$E_R$ can be rewritten in terms of the 3-D velocities, depths and positions of features in the image as follows:

$$E_R = \Sigma w_{ij} \frac{[(x_iZ_i - x_jZ_j)(\dot{X}_i - \dot{X}_j) + (y_iZ_i - y_jZ_j)(\dot{Y}_i - \dot{Y}_j) + (Z_i - Z_j)(\dot{Z}_i - \dot{Z}_j)]^2}{(x_iZ_i - x_jZ_j)^2 + (y_iZ_i - y_jZ_j)^2 + (Z_i - Z_j)^2}. \tag{6}$$

The above cost functional is nonquadratic and its minimization normally requires the solution of a system of equations that are nonlinear in the parameters of depth and 3-D velocity. Standard optimization algorithms for solving nonlinear systems directly, such as gradient

descent methods, are slow and can become trapped in local minima of the solution space. To avoid the use of these nonlinear optimization methods, the algorithm uses a two-stage strategy to perform the minimization that alternates between computing new depths $Z_i$ and 3-D velocities $(\dot{X}_i, \dot{Y}_i, \dot{Z}_i)$. During the first stage of the computation, the depths are assumed to be fixed and only a new set of 3-D velocities is computed. In this case, the above cost function is now quadratic and its minimization can be performed by solving a system of linear equations. After a new estimate of 3-D velocities is obtained, a new set of depths is then computed using equation (2). [Note that equation (2) yields two independent estimates of the $Z_i$, which can be combined to obtain a single estimate using a weighted average of the two, with the weights depending, for example, on the reliability of the two image velocity measurements, $\dot{x}_i$ and $\dot{y}_i$.] The algorithm alternates between these two computations until some criterion is met, which may be a threshold on how much the solution is changing from one iteration to the next, or a fixed number of iterations [see Ando (1991, 1993) for further discussion].

*Temporal integration*

The depths computed from only two frames may not be accurate, as errors can occur for various reasons; for example, the retinal images can be blurred or distorted during the imaging process, the 2-D motion measurements may contain random noise, or the structure and motion of objects may violate the underlying assumptions of the motion measurement or SFM algorithms, such as the rigidity assumption. The goal of temporal integration is to estimate more reliable depths by combining information from multiple frames. The idea behind this process is that random errors may be smoothed out by effectively averaging the 3-D structures computed over time.

The integration algorithm presented here is formally related to a technique in optimal estimation theory called the Kalman filter (Kalman, 1960). The Kalman filter embodies a general framework for estimating dynamically changing random variables from noisy measurements (Gelb, 1974; Anderson & Moore, 1979). Recently, Kalman filtering has been applied to 3-D structure estimation problems (Matthies, Szeliski & Kanade, 1989; Heel, 1990a, b), and has been shown to improve significantly the quality of the estimated structure over time. Ando (1991, 1993) describes a scheme that uses the Kalman filter for a robust estimation of 3-D structure and velocities and relates this scheme to the human recovery of structure from motion. The algorithm presented in this section is based on this scheme, but here we explain only the basic concept behind the algorithm and summarize it briefly.

The temporal integration algorithm is designed to improve the accuracy of the estimated depths incrementally by maintaining and updating the current estimates as each new image is obtained. More specifically, as the velocities in the depth direction are computed at each moment, the depths at the next moment can be predicted by transforming the current estimates of depth using these velocities. Thus, at each moment, we have the predicted depths derived from past information and depths computed from the newly obtained motion information, which are integrated by taking an average of the two. The estimates of depth are then updated by replacing the previous estimates with these integrated depth estimates.

The averaging process can be performed by weighting the computed depths and the predicted depths by their reliability. The reliability of the newly computed depths depends on the properties of errors in the depths (or the variance of the noise in a statistical sense). Thus, it depends on how the errors are generated and conveyed in the earlier processes. Although it is difficult to model the sources of error precisely, some heuristics can be used; for example, when the velocity of a feature in the image is small, the computed depth is more sensitive to noise, so less weight can be given to this feature. The reliability of the predicted depths depends on the reliability of the depths computed in the past. The reliability of the estimates should increase as more reliable depths are integrated, so that the weights of the estimates can be updated sequentially by adding the previous weights with the weights of the newly computed depths. [For further discussion of the choice of these weights, see Ando (1991, 1993).]

The temporal integration algorithm can be summarized as follows. Let the depth and the 3-D velocities of a feature computed at time $t$ be denoted by $\tilde{Z}_t$ and $(\dot{X}_t, \dot{Y}_t, \dot{Z}_t)$, respectively. Let the estimates of depth at time $t$ be denoted by $\hat{Z}_t^-$ and $\hat{Z}_t^+$, where the symbol "^" denotes an estimate, and the superscripts "$-$" and "$+$" denote the estimates before and after the updating stage, respectively. The algorithm first predicts the estimate of depth for each feature at time $t$, $\hat{Z}_t^-$, from the previous estimate $\hat{Z}_{t-1}^+$, using the computed velocity in depth:

$$\hat{Z}_t^- = \hat{Z}_{t-1}^+ + \dot{Z}_{t-1}\Delta t \tag{7}$$

where $\Delta t$ denotes the interframe time interval. The algorithm then updates the current estimate of depth by taking a weighted average of the predicted depth $\hat{Z}_t^-$ and the newly computed depth $\tilde{Z}_t$ as follows:

$$\hat{Z}_t^+ = \frac{1}{\alpha_t^- + \alpha_t}\left(\alpha_t^- \hat{Z}_t^- + \alpha_t \tilde{Z}_t\right) \tag{7}$$

where $\alpha_t^-$ and $\alpha_t$ are the weights for the estimate of depth and the newly computed depth, respectively. The weight for the updated depth can be computed as the sum of these two weights, for example. This sequential averaging process integrates a number of past depth measurements without the need to store them all. As more images are obtained, the accuracy of the estimates of depth improves incrementally over time.

It was noted earlier that the surface interpolation algorithm that we used in our simulations is not viewpoint invariant. In principle, some viewpoint invariance can be achieved by using a known motion of the object surface or observer to predict the shape of the surface at

a later time, thereby taking the observer's viewpoint into account in the temporal integration stage.

*The surface interpolation algorithm*

A number of algorithms have been proposed for performing explicit smooth surface reconstruction from sparse depth data (Schumaker, 1976; Grimson, 1981, 1985; Boult & Kender, 1986; Terzopoulos, 1986, 1988; Blake, 1991; Blake & Zisserman, 1987, 1991; Gamble & Poggio, 1987; Marroquin, Mitter & Poggio, 1987; Szeliski, 1988; Geiger & Girosi, 1991; for review, see Bolle & Vemuri, 1991), any of which could be used for the interpolation component of our model. Most of these algorithms were developed in the context of interpolating sparse depth data derived from stereo (also see Hoff & Ahuja, 1987). Approaches based on fitting planar patches to local triplets of points may also be useful to consider (for example, Faugeras *et al.*, 1990). These methods differ mainly in the extent to which they address the detection of depth discontinuities, and in the particular algorithm used to compute the smoothest surface that fits the given depth data. In the earlier work of Grimson (1981), discontinuities were detected after the smooth surface interpolation and a gradient descent algorithm was used to compute the smooth surface. Algorithms proposed by Marroquin *et al.* (1987), Gamble and Poggio (1987) and Szeliski (1988) use a probabilistic optimization process in which the detection of discontinuities forms a more integral part of the surface reconstruction. Terzopoulos (1988), Blake and Zisserman (1987) and Geiger and Girosi (1991) present deterministic algorithms for computing piece-wise smooth surfaces that may contain discontinuities.

Most of the above surface reconstruction algorithms have been applied to both synthetic and natural images, establishing their viability for computer vision systems. From a biological standpoint, most of these algorithms use simple, local operations that can be performed in parallel. Detailed psychophysical experiments that could possibly distinguish between different models have not yet been conducted. In order to test the general hypothesis that surface interpolation plays a critical role in human SFM recovery, we conducted simulations using a particular surface interpolation algorithm. In particular, we used Grimson's original surface approximation algorithm, primarily for its simplicity, with simple modifications to handle boundary information. We do not suggest this method as a specific quantitative model of surface interpolation in the human visual system.

In the case of Grimson's algorithm, a surface $S(x, y)$ is computed that fits through the known depth points $C(x, y)$ as closely as possible, and minimizes the total variation in depth, through minimization of the following expression:

$$\iint\left(\frac{\partial^2 S}{\partial x^2}+2\frac{\partial^2 S}{\partial x\partial y}+\frac{\partial^2 S}{\partial y^2}\right)\mathrm{d}x\,\mathrm{d}y + \lambda_s\sum_{\vartheta}S(x,y)-C(x,y)]^2 \quad (9)$$

where the discrete summation in the second term takes place over the set $\vartheta$ of points for which there is a known depth value. The first term expresses the variation in depth over the entire surface, and the second term measures how well the interpolated surface fits through the known depth data. $\lambda_s$ is a constant that captures the relative contribution of the smoothness and data in the surface reconstruction. Many standard optimization algorithms can be used to perform this interpolation (e.g. Luenberger, 1973).

A problem with the above algorithm as it stands is that it tends to flatten out the edges of highly curved objects such as the cylinders that we use in our simulations. We considered two modifications to handle depth constraints in the vicinity of object boundaries. The first is to "pin down" the depths at the edges of the object to the depth of the background plane. The second is to force the derivative of depth along the boundary to be high. [Alternatively, one could interpolate a representation based on surface orientation rather than depth and constrain the surface orientation along an object boundary to be perpendicular to the line of sight and to the 2-D projection of the boundary contour (Ikeuchi & Horn, 1981; Aloimonos & Huang, 1991).] Methods have been proposed to detect depth discontinuities, but we do not address this issue here; rather we only consider the consequence of placing boundary constraints into the surface reconstruction process.

When features from different surfaces move in opposite directions, they are first grouped by their direction of motion prior to surface interpolation. In the simulations presented in the next section, features with a horizontal component of motion to the right were segregated from those that moved to the left.

*Combining structure-from-motion with surface interpolation*

We directly combine the velocity based SFM algorithm, temporal integration, and Grimson's surface interpolation algorithm. We assume that the image sequence consists of a set of discrete features in motion, which may continually disappear and reappear at new locations. The initial surface is assumed to be at constant depth everywhere. The combined scheme consists of the following steps: (1) the set of discrete features undergoes small displacements in the image, and the SFM and temporal integration algorithms are used to compute a new 3-D structure for the features; (2) a smooth surface (or surfaces) is interpolated across the new depth values; and (3) some or all of the features may then disappear and reappear at other random locations in the image, and the newly appearing features are assigned an initial depth given by the interpolated surface(s) at the new locations. The process then repeats itself; the features undergo new displacements in the image, a new 3-D structure is computed, and so on. The surface reconstruction stage also uses the grouping of moving points by direction and speed of motion, and allows the independent interpolation of multiple surfaces.

## COMPUTER SIMULATIONS

This section presents the results of computer simulations conducted with the model described in the previous sections. We consider: (1) the ability of the model to cope with moving points having short lifetimes; (2) the degradation of the solution with fewer points in motion; (3) the performance of the model on the perceptual displays presented by Ramachandran *et al.* (1988); and (4) the influence of object boundaries on the surface reconstruction. Our simulations indicate that while surface reconstruction may play a key role in accounting for a number of these phenomena, there are other aspects of the motion measurement and SFM stages that may also contribute to our final 3-D percept. Some of these simulation results have appeared earlier in Ando (1991, 1993).

*Coping with short point lifetimes*

A primary perceptual motivation for considering the incorporation of surface interpolation into the SFM process is derived from our ability to perceive 3-D shape in displays with moving points that continually disappear and reappear, with short lifetimes (Hursain *et al.*, 1989; Dosher *et al.*, 1989a; Landy *et al.*, 1991; Treue *et al.*, 1991, 1995). The first simulation here demonstrates that the addition of a separate surface reconstruction process that embodies a surface interpolation algorithm successfully allows 3-D surface shape to build up incrementally in spite of a short persistence of moving points.

In this simulation, 60 points were randomly positioned on the surface of a vertically oriented 3-D cylinder and were rotated around a central vertical axis (the number of points was based in part on the psychophysical studies cited earlier). The radius of the cylinder was 15, its height was 30, and the cylinder was located a distance of 100 from the observer. The image positions and velocities of the points were computed analytically, using perspective projection. Relative noise was added in the form of Gaussian distributed perturbations of the velocities, giving an average error in the velocities of 20%. The initial 3-D structure considered by the algorithm was flat. Figure 3 shows the comparison between results obtained under two different conditions. In both cases, the points were rotated in increments of 2°, and after every 2° of rotation, half of the points disappeared and reappeared at different locations on the surface of the cylinder. As a consequence, each point persisted for only 4° of rotation. [In the experiments of Husain *et al.* (1989), the cylinder was rotated for 3.5° during the short point lifetimes of 100 msec, so the amount of rotation used here was comparable]. For the results shown in Fig. 3a, when the current points disappeared and new points appeared, the initial depths of the new points were placed back at the flat depth plane used as the initial solution. Thus motion information was only integrated over a rotation of 4° of the points. As shown in Fig. 3a, some structure is built up in this case, but there is no improvement of the solution after a few degrees of rotation. In contrast, for the results shown in Fig. 3b, a

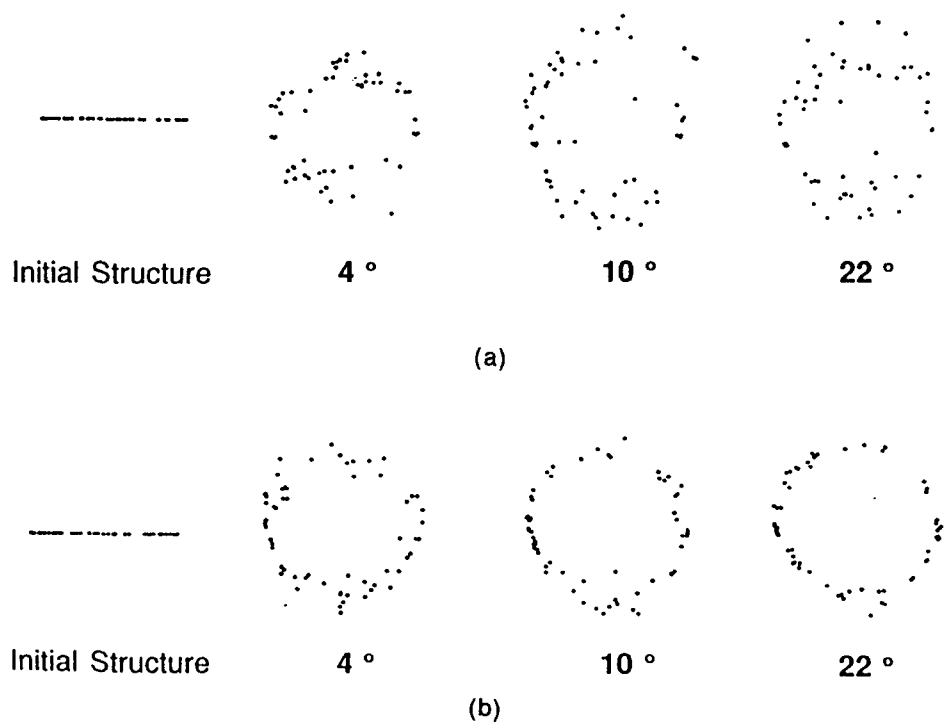


FIGURE 3. Comparison of the results of the model for displays containing points with short lifetimes (a) without surface interpolation and (b) with surface interpolation. The results are shown after total rotations of 4°, 10° and 22°. Relative Gaussian noise was added to the image velocities of the points, with $\sigma = 0.5$, yielding an average error of 20% in the velocity components. The depth and 3-D velocity computations were each performed once and roughly 60 iterations were performed within the 3-D velocity computation.

smooth surface was interpolated across the depth values obtained after every 2° of rotation, and this interpolated surface provided the initial depths for newly appearing points. With the added surface interpolation, there is a rapid convergence toward the cylindrical structure. The dense surface representation provided by the interpolation allows the temporal integration process to integrate motion information over a more extended time and helps to smooth out fluctuations in the 3-D structure derived from the SFM algorithm as a result of the added noise, by imposing a smoother surface interpolation on the sparse 3-D data. The interpolated surface also helps to speed up the iterative SFM computation by providing an initial 3-D solution that is closer to the true structure.

A basic assumption of the surface interpolation process is that only a single surface exists at each location in the image. For the case of this transparent cylinder, there are two surfaces at each location. To cope with this transparency, the moving points were segregated into two groups, depending on whether they were moving to the left or right in the image, and surface interpolation was performed separately on the two groups. [This grouping can also be performed simultaneously with surface interpolation, based on the reconstructed depths (Ando, 1993).]

Relating these results to perceptual demonstrations, in the case of the experiments of Husain *et al.* (1989) and Treue *et al.* (1991), subjects were asked to distinguish between an "unstructured" stimulus that can be seen as corresponding physically to a volume of randomly moving points, and a "structured" stimulus, in which points are placed on the surface of a cylinder. The results obtained without surface interpolation shown in Fig. 3a are essentially indistinguishable from a random volume of points. Therefore, without surface interpolation, our model would not be able to perform the discrimination task required in these experiments, similar to human observers. On the other hand, the results obtained with surface interpolation shown in Fig. 3b could clearly form the basis for a successful discrimination, with a relatively short total viewing time required.

The precise rate of buildup of 3-D structure over an extended image sequence depends on a number of parameters used in the SFM recovery process, including the factor $\lambda$ that captures the trade-off between the rigidity of the computed 3-D structure and the closeness of fit of the solution to the image velocity measurements (see equation 3), the distance metric used to weigh the rigidity of the connection between each pair of points (see the discussion around equation 5), the level of noise added to the input velocities, and the number of iterations of the depth and 3-D velocity computations at each time step. Some discussion of the influence of these parameters can be found in Ando (1991, 1993).

*Degradation with fewer points*

When viewing displays with fewer points in motion, Husain *et al.* (1989) and Treue *et al.* (1991) found that first, there is a general degradation in performance with fewer points, such that greater time is required to judge reliably whether a given stimulus is structured or unstructured. Second, if the points are repeated at the same initial locations after disappearing rather than jumped to new random locations, observers are unable to distinguish the structured and unstructured stimuli, even after long stimulus durations.

Aspects of the surface interpolation algorithm can lead to degradation in performance for fewer points. The particular algorithm used here is iterative and uses local operations at each iteration to propagate surface depth constraints from locations where depth information is known to locations at which there is no depth information given. In the case of a smaller number of points, the larger gaps that occur in the image require a larger number of iterations to fill in surface shape. This phenomenon is illustrated in Fig. 4. We show the results of the surface interpolation algorithm after the same number of iterations, for the case of 60 points and 6 points placed on a grid of size $17 \times 17$ elements, in Fig. 4a and b, respectively (again, the choice of the number of points used here is motivated in part by the psychophysical studies cited above). The initial data points were randomly sampled from the front surface of a cylinder, and no noise was added to their depths. Points with no initial depth information were placed on a background plane of constant depth that was roughly equal to the average depth within the front surface of the cylinder. After a fixed number of iterations, the solution obtained for 60 points is much closer to the true cylindrical shape. The solution obtained for fewer points eventually converges to a similar surface, but far more iterations are needed to obtain a comparable solution. When the depths of only a few points are given, the interpolated surface also depends critically on the spatial distribution of the points in the projected 2-D image. As an example, Fig. 4c was constructed from a set of 6 points whose positions were skewed towards one half of the image of the cylinder. A slanted plane emerges that does not curve backwards along both borders of the cylinder, due to a lack of explicit depth information on one side.

The biological vision system is not likely to use an algorithm that embodies discrete iterations as we consider here, but may use a process that requires time to converge to a solution, with the amount of time depending on the size of the gaps in the image. If a limited time is available at each moment, because the object is moving rapidly, the interpolation process may not have time to converge completely at each moment, requiring a larger number of views or longer total viewing time to yield a 3-D surface that is adequate for making the judgement of structured versus unstructured that is required in the Husain *et al.* (1989) and Treue *et al.* (1991) studies.

The interpolation scheme used here also computes a surface that is most smooth in a mathematical sense, through the local propagation of constraints. We could also consider a strategy more similar to a filtering operation that interpolates the surface by smoothing the given depth data with a function such as a Gaussian. If the data is sufficiently dense, such an operation may yield

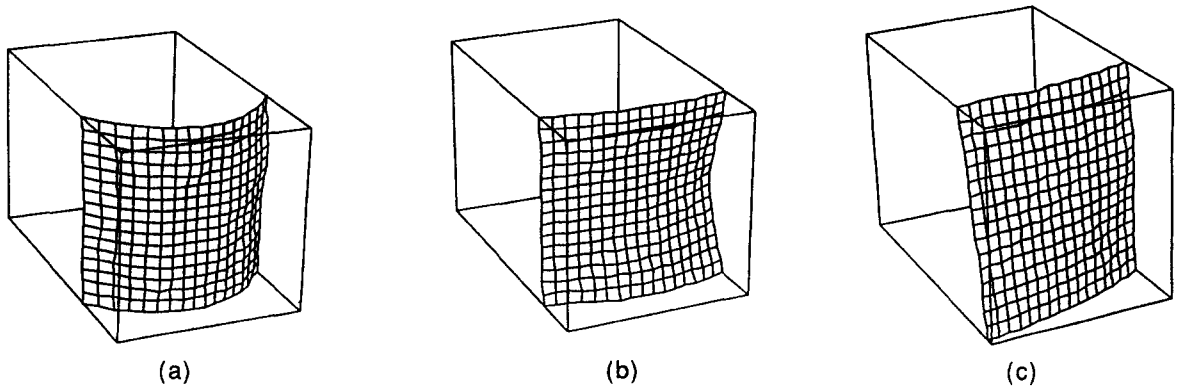


FIGURE 4. Degradation of the surface interpolation process for fewer points. A set of points in depth are first sampled from the surface of a cylinder and locations at which no explicit depth information is given are initially assigned a depth that is roughly the average of the known points. (a) The solution obtained after 50 iterations of the surface interpolation algorithm, for the case where 60 points are placed on a grid of size 17 × 17 elements. (b) The solution obtained after the same number of iterations, for the case of 6 points. (c) The solution obtained for a set of 6 points whose positions are skewed toward one side of the cylinder.

a reasonable approximation to the smoothest surface, but the results would degrade as the data become more sparse. Note that the detection of surface discontinuities can also become more difficult when image features are sparse.

Finally, it is also possible to associate a *confidence* with the surface depth information derived at each location in the image, and this confidence could depend on the distance to known depth points or the amount of time that has elapsed since an explicit data point appeared near a given location. As a consequence of these factors, the confidence associated with the surface information available at a particular location may decay over time, if no further evidence of surface shape is presented in this image region. For displays with only a few points that are oscillated at a small number of locations, the confidence in the derived surface may only be high in the vicinity of these few locations, and will always be low in regions of the image that are distant from the moving points. On the other hand, when the points are continually jumped to new locations, explicit depth data is obtained at a larger number of image locations, possibly leading to greater confidence in the surface information obtained over a larger portion of the image, which may facilitate the judgement of 3-D structure.

The above discussion focused on the degradation of the surface reconstruction process with sparse image features. The performance of the SFM recovery algorithm can also be affected by such sparse texture. For example, there are parameters used in the SFM recovery algorithm that can affect the relative performance of the overall model for sparse and dense patterns of points. Referring to equation (3), there is a factor $\lambda$ that controls the trade-off between the rigidity of the solution and the extent to which the solution is consistent with the image motion measurements. When the density of moving points is low, a larger value of $\lambda$, which places greater weight on the rigidity of 3-D structure, is needed to obtain a reasonable solution. If a fixed value of this parameter were used in all contexts, the solution would degrade when the image features are more sparse.

The use of a separate feature-based SFM recovery scheme such as the one presented here helps to reduce the degradation that could occur with sparse texture, by allowing 3-D structure to be recovered when there is no well-defined surface.

*Ramachandran et al.'s two-cylinders demonstrations*

The demonstrations by Ramachandran *et al.* (1988) suggest interesting interactions between multiple surfaces that are moving nonrigidly with respect to one another. This section shows that our model can account for a number of the experimental observations. We begin with a simulation of the demonstration in which two cylinders of the same size are superimposed in the same region of space, but rotated at different speeds. Human observers perceive two distinct surfaces in each direction of motion, with the faster surface bulging outward from the slower surface. Figure 5a shows a schematic illustration of the two cylinders that underlie the construction of the visual stimulus, and Fig. 5b shows a bird's eye view of the resulting percept. The SFM process embodied in our model derives a 3-D structure that is consistent with this percept, as illustrated in Fig. 6. Figure 6a shows a birds' eye view of the true 3-D structure and Fig. 6b shows the results of the SFM algorithm applied to two frames that were separated by 1° of rotation of the points. No noise was added to the image velocities. The points were subsequently grouped by their speed of motion, and separate 3-D surfaces were reconstructed for the two groups. (To group the points by speed, we divided the image into small regions, segregated the points within each region into two populations if there were two distinct peaks in a histogram of their speeds, and then grouped points from one region to the next that had similar speeds. This strategy clearly distinguished the two groups of points through the central part of the cylinder.) The results of this surface reconstruction stage are shown in Fig. 6c. The overall impression of one surface bulging out from the other is clearly conveyed in

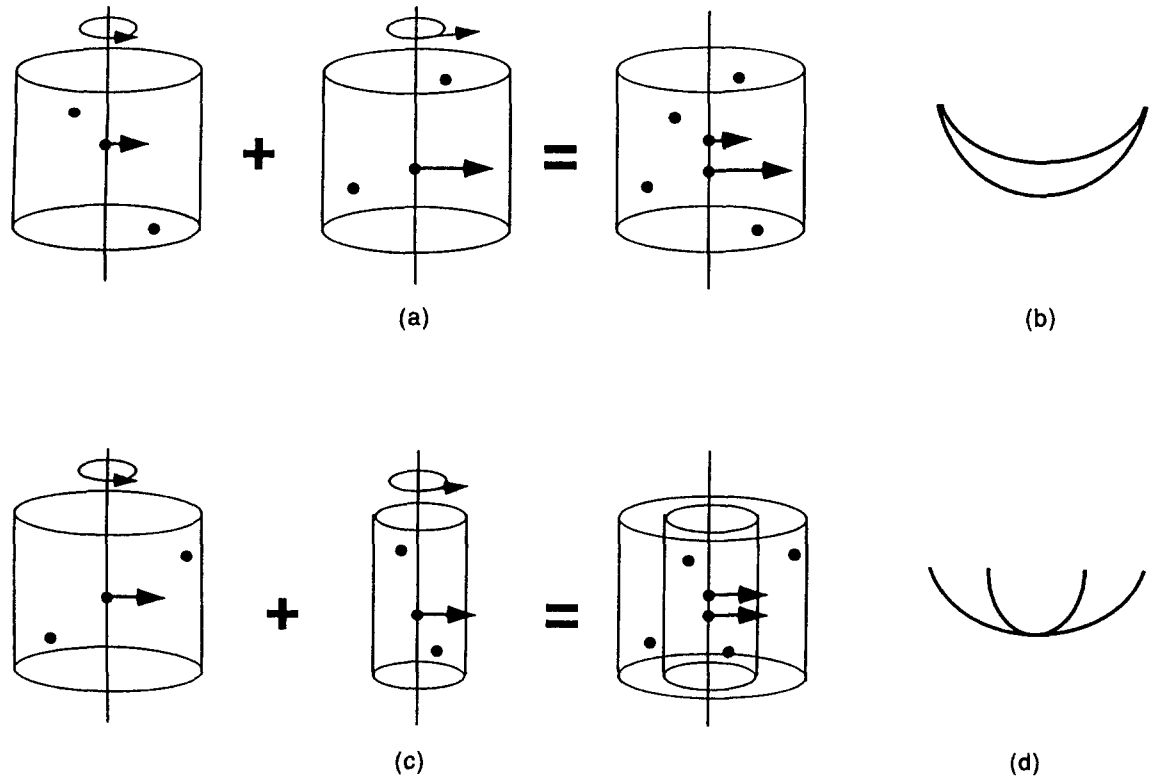


FIGURE 5. The perceptual demonstrations of Ramachandran *et al.* (1988). (a) Schematic drawing of the demonstration in which two cylinders of the same size are superimposed in the same region of space, but rotated at different speeds. (b) Bird's eye view of the percept obtained from a display created from (a). (c) Schematic drawing of the demonstration in which the cylinders are of different radius, but the relative speeds are adjusted so that the projected image speed is the same for points in the center of the two surfaces. (d) Birds' eye view of the percept obtained from a display created from (c).

these results. This phenomenon is also preserved with short point lifetimes (Treue *et al.*, 1995).

This result can be explained in terms of the goal of the SFM algorithm. The stimulus is nonrigid and the SFM algorithm effectively tries to interpret the stimulus as a single object that is deforming as little as possible over time. The total change in the 3-D distances between pairs of points in the computed 3-D structure is less than the total change in these 3-D distances in the true structure. The solution shown in Fig. 5b is, in fact, the *most rigid* structure consistent with the projected image velocities. The algorithm derives an interpretation similar to two embedded cylinders rotating rigidly with one another.

The next simulations address the demonstrations in which the cylinders are of different radius, but the relative speeds are adjusted so that projected image speed is the same for points in the center of the two surfaces. Figure 5c shows a schematic illustration of the two cylinders, and Fig. 5d shows a birds' eye view of the resulting percept. Figure 7a shows a birds' eye view of the true 3-D structure and Fig. 7b shows the results of the SFM algorithm applied to two frames separated by 1° of rotation of the points. No noise was added to the image velocities. The points were subsequently grouped by their speed of motion, with all of the points in the central region of the display participating in both groups, and separate surfaces were reconstructed for the two groups. The result of surface reconstruction is

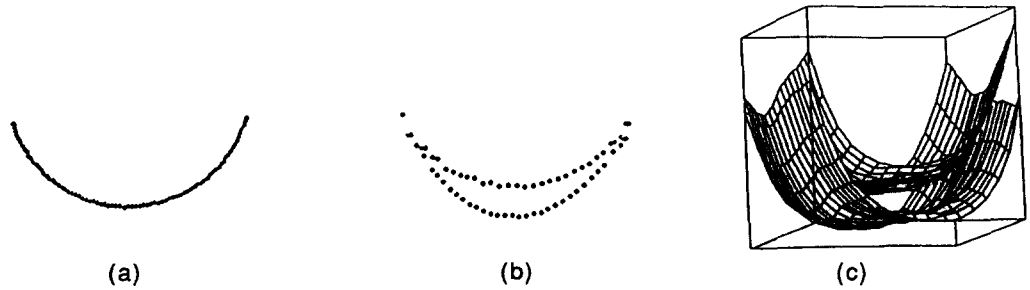


FIGURE 6. Simulations with the model applied to perceptual demonstrations of Ramachandran *et al.* (1988) shown in Fig. 5a and b. (a) Birds' eye view of the true 3-D structure. (b) The results of the structure-from-motion algorithm applied to two frames separated by 1° of rotation of the points. (c) The results of the surface reconstruction stage.

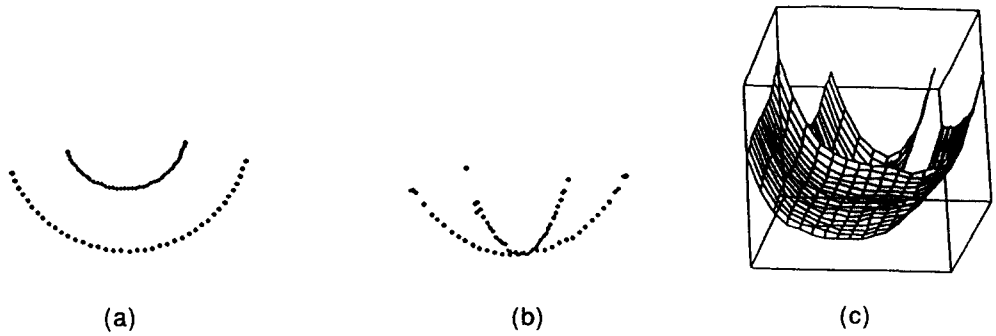


FIGURE 7. Simulations with the model applied to perceptual demonstrations of Ramachandran *et al.* (1988) shown in Fig. 5c and d. (a) Birds' eye view of the true 3-D structure. (b) The results of the structure-from-motion algorithm applied to two frames that are separated by 1° of rotation of the points. (c) The results of the surface reconstruction stage.

shown in Fig. 7c. The results capture the overall subjective impression of the two surfaces merging into a single surface in the center. This phenomenon is also preserved with short point lifetimes (Treue *et al.*, 1995). This result is again due in part to the fact that our model tries to interpret the stimulus as a single object whose 3-D structure is changing as little as possible over time.

For both of the above examples, the final surface structure is due primarily to the SFM recovery algorithm on its own, as indicated in Figs 6b and 7b. The subsequent interpolation is not critical here. Note, however, that in the case of our model, the additional surface interpolation stage is essential to account for the preservation of this final surface percept when short point lifetimes are used.

*The influence of the interpretation of boundaries*

The last issue that we address is the influence of constraints on 3-D shape provided by the interpretation of object boundaries. We again consider demonstrations by Ramachandran *et al.* (1988) that illustrate this influence in human SFM recovery.

The first demonstration simulates two superimposed planes of dots shearing along each other. In the perceptual display, there were stationary boundaries along the left and right edges of the display, and points changed their direction of motion when reaching the boundaries, giving the impression of points bouncing off the edges. The points otherwise underwent a pure translation across the display, either to the left or right. It was reported that human observers perceive a rotating cylinder when viewing these displays (Ramachandran *et al.*, 1988). In our companion paper, we note that observers perceive some curvature along the borders of the figure and some separation in depth between the two surfaces, but the overall percept is flatter than that of the true cylinder.

Three factors may contribute to the perception of curved surfaces in this demonstration. First, the initial motion measurements may be incorrect near the borders of the figure. The true velocities are constant up to the borders and suddenly change direction, but the spatial and temporal integration embodied in the motion measurement mechanisms is likely to distort this pattern of motion, yielding variation in the speed of motion of points near the borders that the SFM process then interprets as being due to surface curvature. A second factor is the "bouncing off" of points at the edges, which may provide a direct cue to the presence of a transparent, curved surface in rotation. The perception of curvature is weaker if the points disappear at the edges and reappear at some other location on the edge. Finally, if the SFM process combines all of the points and interprets their motion as due to a single object undergoing minimal distortion over time, a curved structure may emerge at this stage. In the simulations, we explore these three possible sources of the perception of curvature.

In the first simulation, we show the result of applying the SFM algorithm on its own to all of the moving points together, with no systematic error in the velocity measurements along the borders. The result is shown in Fig. 8a. The true structure consisting of two flat planes of points superimposed at the same depth is shown on the left. The computed 3-D structure, shown on the right, consists of two planes with only slight curvature, separated in depth. The magnitude of the depth separation increases with increased speed of motion of the points.

The next simulation adds systematic error to the input velocity measurements. We varied the image speed of points near the border so that speed drops off linearly with decreasing distance from the border. The SFM algorithm was applied to the resulting velocity pattern. The result is shown in Fig. 8b. Toward the center of the figure, the two surfaces are still fairly flat and separated in depth, but there is now more curvature near the edges. The precise shape of the surface near the edges depends on the particular velocity profile used.

In the final simulation, we show that if an explicit constraint is introduced along the two edges of the figure, forcing the gradient of the surface to be high along the edges, then surface interpolation on its own can yield a curved surface from an initial set of depths corresponding to points on two flat planes that are separated in depth. The results of this simulation for a single surface are shown in Fig. 8c. The added boundary constraint yields a curved surface near the edges of the

display. Our conclusion from these and earlier results is that the perception of curvature in these displays can be due to any or all of the above factors.

In a second demonstration, Ramachandran *et al.* (1988) masked off vertical sections on the left and right edge of a rotating random-dot cylinder. It was reported that the truncated cylinder is perceived as a cylinder with a smaller radius. We note in our companion paper that we do perceive a single, narrower object in rotation with higher curvature at its borders, but the true percept is flatter than that derived from the narrow cylinder. Furthermore, if the truncated cylinder is masked in such a way that the subject perceives a window in front of the cylinder, it no longer appears as a narrower, more highly curved object. In the terms suggested by Nakayama, Shimojo and Silverman (1988), when the image of the cylinder is simply truncated, the new virtual borders of the figure are interpreted as being *intrinsic* to the object surface, and are perceived as boundaries of an object rotating in depth. When the figure is surrounded by a mask that is perceived as an aperture, the borders of the moving pattern are now *extrinsic* to the inner surface and are no longer interpreted as the curved boundaries of an object rotating in depth.

One hypothesis consistent with the above observations is that subjects perceive the points as "bouncing off" the new edge, which leads to the inference that this is the edge of a curved surface, introducing higher curvature at this edge for the surface interpolation process (i.e. higher curvature than what is actually conveyed by the relative movement of the points at this edge). Viewers do report that the points appear to bounce off the edges of the display when they are not scrutinized. The SFM algorithm on its own, applied to all of the points together, yields the correct 3-D structure in this case, as shown in Fig. 9a. This is similar to the percept derived when the truncated cylinder appears to be viewed through an aperture. The reconstructed front surface is shown in Fig. 9b. If systematic error in the velocity measurements is introduced near the borders of the truncated cylinder, so that the speeds of motion of the points drop off near the borders, then the solution looks like a narrower object in rotation with higher curvature at its borders, although the separation in depth between the front and back surfaces is still high (see Fig. 9c). Finally, if in the surface reconstruction stage, constraints are placed along the two borders of the truncated cylinder that force the derivative of depth to be high, we also obtain a narrower, more curved object, similar to the result shown in Fig. 8c (see also, Aloimonos & Huang, 1991). When a mask is placed around the truncated cylinder that is perceived as an aperture, these boundary constraints may not be imposed (see also, Thompson *et al.*, 1992). Again, we conclude that a number of factors can lead to the perception of higher curvature in this display.

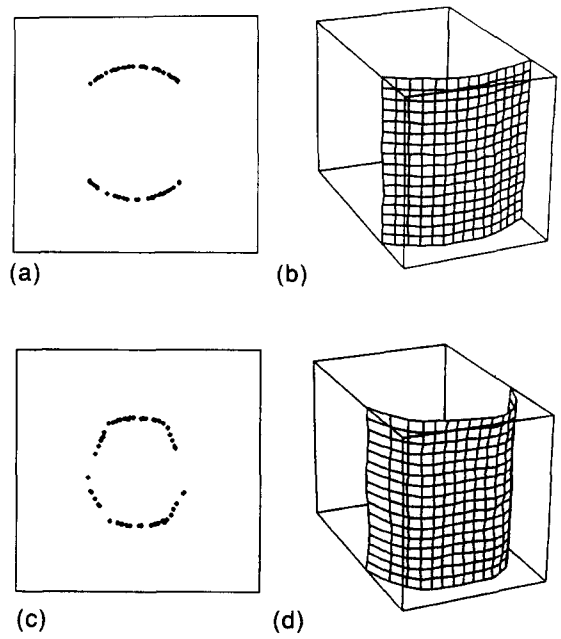


FIGURE 9. The influence of constraints on surface shape from object boundaries. The image of points on the surface of a vertical rotating cylinder is truncated along the left and right borders. (a) Birds' eye view of the 3-D structure obtained when the structure-from-motion algorithm is applied to all of the moving points together. The solution is essentially identical to the true structure of the points. (b) The reconstructed front surface without imposing the occluding boundary constraint. (c) The 3-D structure obtained when systematic error is added to the velocities of the moving points at the borders of the display. (d) The result of the surface interpolation algorithm applied to the depth information derived in (a), with added constraints that force the gradient of depth to be high along the borders of the display.

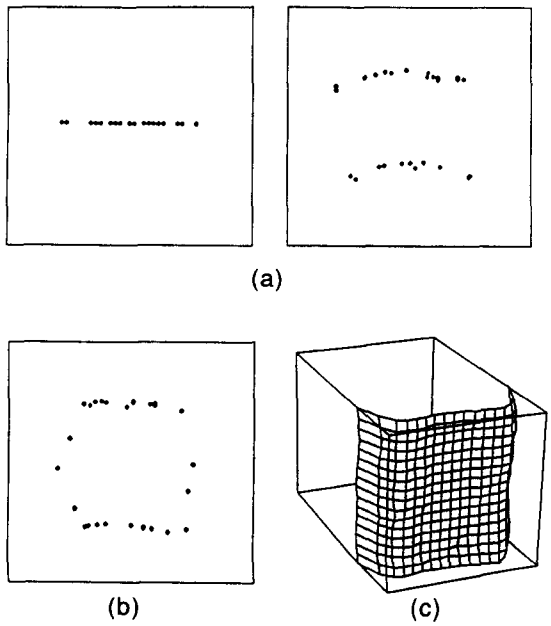


FIGURE 8. Simulations with the two-planes demonstration of Ramachandran *et al.* (1988). (a) The true structure (left) consists of two sets of points at the same depth, which translate to the left and right, respectively. On the right are shown the results obtained when the structure-from-motion algorithm on its own is applied to all of the points together. (b) The results obtained when systematic error is added to the image motions derived for points near the borders of the display. (c) The result of the surface interpolation algorithm applied to the depth information derived in (a) for one of the two surfaces, with added constraints that force the gradient of depth to be high along the borders of the display.

## SUMMARY AND CONCLUSIONS

This paper addressed the computational role that the construction of a complete surface representation plays in the recovery of 3-D structure from motion. We first discussed the need to integrate surface reconstruction with the SFM process on computational grounds, and then reviewed perceptual observations that support this need and place constraints on the nature of the underlying mechanisms. The experimental observations presented in our companion paper (Treue *et al.*, 1995) further strengthen our hypothesis regarding the important interaction of these two processes. We then presented a model that combines a feature-based SFM recovery algorithm, temporal integration and surface reconstruction. The latter component of this model allows multiple surfaces to be represented in a given viewing direction, incorporates constraints on surface structure from object boundaries, and segregates image features onto multiple surfaces on the basis of their 2-D image motion.

The results of computer simulations suggest that our model can provide a qualitative account for a number of perceptual phenomena regarding the possible role of surface reconstruction in SFM recovery. We also showed that aspects of the motion measurement and SFM recovery algorithms can contribute to some of these phenomena, in addition to surface reconstruction. The model is able to build up 3-D shape over an extended time period when the lifetimes of moving points are very short. In our model, surface interpolation is critical to achieving this particular capability. A number of factors in the overall process of 3-D shape recovery can yield degradation with fewer points in motion, consistent with human perception. Finally, our model can account for many of the demonstrations presented by Ramachandran *et al.* (1988) illustrating interesting interactions between multiple surfaces in motion and the influence of object boundaries on perceived shape. These latter conclusions are also based on extensions and clarifications of these demonstrations presented in our companion paper (Treue *et al.*, 1995).

Our work raises a number of questions for further investigation. One issue regards the quantitative aspects of the surface that humans perceive as being interpolated through explicit depth information, whether the data is derived from the SFM cue or other sources such as stereo. As we noted earlier, a number of algorithms have been proposed for performing smooth surface approximation, which may yield different behavior. Subjectively, we perceive a smooth surface when the data is dense, but may derive a more "faceted" surface when presented with sparser patterns. More sensitive psychophysical experiments are needed to distinguish between possible models for surface reconstruction.

A second question that arises is the precise nature of the grouping processes used to segregate points into different groups based on their image direction and speed, for the purpose of the SFM motion or surface reconstruction processes. This grouping task becomes more difficult, for example, when two or more curved transparent surfaces move with different speeds, yielding points moving with multiple speeds in small regions of the image that are also varying from one region to the next. It may also be possible to group features on the basis of depth itself, after some initial 3-D structure has been derived.

A third question is how to determine the appropriate surface boundary constraints automatically, from the observed pattern of 2-D image motion. This issue becomes especially important for the analysis of natural images that contain large regions of very sparse texture.

Finally, the hypothesis that there exists a separate surface reconstruction process that integrates 3-D information from multiple cues naturally raises the question of how this information is combined, particularly in situations where inconsistencies arise between these different cues.

## REFERENCES


Adelson, E. H. (1985). Rigid objects that appear highly non-rigid. *Investigative Ophthalmology and Visual Science, (Suppl.), 26,* 56.

Aggarwal, J. K. & Martin, W. (Eds) (1988). *Motion understanding.* Hingham, Mass.: Kluwer.

Aloimonos, J. & Huang, L. (1991). Motion-boundary illusions and their regularization. *Proc. IEEE Workshop on Visual Motion,* Princeton, N.J., October, 88–94.

Andersen, G. (1989). Perception of three-dimensional structure from optic flow without locally smooth velocity. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 363–371.

Andersen, R. A. & Siegel, R. M. (1990) Motion processing in the primate cortex. In Edelman, G. M., Gall, W. L. & Cowan, W. M. (Eds) *Signal and sense: Local and global order in perceptual maps* (pp. 163–184). New York: Wiley.

Anderson, B. D. O. & Moore, J. B. (1979). *Optimal filtering.* Englewood Cliffs, N.J.: Prentice-Hall.

Ando, H. (1991). Dynamic reconstruction of 3D structure and 3D motion. *Proc. IEEE Workshop on Visual Motion* (pp. 101–110). Princeton, N.J., October.

Ando, H. (1993). Dynamic reconstruction and integration of 3d structure information. Ph.D. thesis, MIT Dept & Brain & Cognitive Sciences, February.

Barron, J. (1984). A survey of approaches for determining optic flow, environmental layout and egomotion. *University of Toronto Technical Report on Research in Biological and Computer Vision, RBCV-TR-84-5.*

Bharwani, S., Riseman, E. & Hanson, A. (1986). Refinement of environmental depth maps over multiple frames, *Proc. IEEE Workshop on Motion: Representation and Analysis* (pp. 73–80). Charleston, S.C.

Blake A. (1991). Viewpoint-invariant reconstruction of visible surfaces. In Mayhew, J.E. W. & Frisby, J. P. (Eds). *3D model recognition from stereoscopic cues* (pp. 103–110). Cambridge, Mass.: MIT Press.

Blake, A. & Zisserman, A. (1987). *Visual reconstruction.* Cambridge, Mass.: MIT Press.

Blake, A. & Zisserman, A. (1991). Invariant surface reconstruction using weak continuity constraints. In Mayhew J. E. W. & Frisby, J. P. (Eds) *3D model recognition from stereoscopic cues* (pp. 111–117). Cambridge: MIT Press.

Bolle, R. M. & Vemuri, B. C. (1991). On three-dimensional surface reconstruction methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13,* 1–13.

Borjesson, E. & von Hofsten, C. (1973). Visual perception of motion in depth: application of a vector model to three-dot motion patterns. *Perception & Psychophysics, 13,* 169–179.

Boult, T. E. & Kender, J. R. (1986). Visual surface reconstruction using sparse depth data. In *Proc. IEEE Conference on Computer Vision Pattern Recognition* (pp. 68–76). June.

Braunstein, M. L. (1962). The perception of depth through motion. *Psychological Bulletin, 59*, 422–433.

Braunstein, M. L. (1976. *Depth perception through motion*. New York: Academic Press.

Braunstein, M. L. & Andersen, G. J. (1984a). A counterexample to the rigidity assumption in the visual perception of structure from motion. *Perception, 13*, 213–217.

Braunstein, M. L. & Andersen, G. J. (1984b). Shape and depth perception from parallel projections of three-dimensional motion. *Journal of Experimental Psychology: Perception and Performance, 10*, 749–760.

Braunstein, M. L., Hoffman, D. D. & Pollick, F. E. (1990). Discriminating rigid from nonrigid motion: Minimum points and views. *Perception & Psychophysics, 47*, 205–214.

Braunstein, M. L., Hoffman, D. D., Shapiro, L. R., Andersen, G. J. & Bennett, B. M. (1987). Minimum points and views for the recovery of three-dimensional structure. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 335–343.

Bruss, A. & Horn, B. K. P. (1983). Passive navigation. *Computer Vision, Graphics and Image Processing, 21*, 3–20.

Clocksin, W. F. (1980). Perception of surface slant and edge labels from optical flow: A computational approach. *Perception, 9*, 253–269.

Cutting, J. E. (1982). Blowing in the wind: Perceiving structure in trees and bushes. *Cognition, 12*, 25–44.

Doner, J., Lappin, J. S. & Perfetto, G. (1984). Detection of three-dimensional structure in moving optical patterns. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 1–11.

Dosher, B. A., Landy, M. S. & Sperling, G. (1989a) Kinetic depth effect and optic flow—I. 3D shape from fourier motion. *Vision Research, 29*, 1789–1813.

Dosher, B. A., Landy, M. S. & Sperling, G. (1989b). Ratings of kinetic depth in multidot displays. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 816–825.

Durrant-Whyte, H. F. (1988). *Consistent integration and propagation of disparate sensor observations*. New York: Kluwer.

Faugeras, O. (1993). *Three-dimensional computer vision: A geometric viewpoint*. Cambridge, Mass.: MIT Press.

Faugeras, O. D., LeBras-Mehlman, E. & Boissonat, J. D. (1990). Representing stereo data with the Delaunay triangulation. *Artificial Intelligence, 44*, 41–88.

Gamble, E. B. & Poggio, T. (1987). Visual integration and the detection of discontinuities: The key role of intensity edges. *MIT Artif. Intell. Lab. Memo, 970*.

Geiger, D. & Girosi, F. (1991). Parallel and deterministic algorithms from MRFs: surface reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*, 401–412.

Gelb, A. (Ed.) (1994). *Applied optimal estimation*. Cambridge, Mass.: MIT Press.

Green, B. F. (1961). Figure coherence in the kinetic depth effect. *Journal of Experimental Psychology, 62*, 272–282.

Grimson, W. E. L. (1981). *From images to surfaces: A computational study of the human early visual system*. Cambridge, Mass: MIT Press.

Grimson, W. E. L. (1982). A computational theory of visual surface interpolation. *Philosophical Transactions of the Royal Society of London B, 298*, 395–427.

Grimson, W. E. L. (1983a). An implementation of a computational theory of visual surface interpolation. *Computer Vision, Graphics and Image Processing, 22*, 39–69.

Grimson, W. E. L. (1983b) Surface consistency constraints in vision. *Computer Vision, Graphics and Image Processing, 24*, 28–51.

Grimson, W. E. L. (1985). Computational experiments with a feature based stereo algorithm. *IEEE Transactions of Pattern Analysis and Machine Intelligence, PAMI-7*, 17–34.

Grzywacz, N. M. & Hildreth, E. C. (1987). The incremental rigidity scheme for recovering structure from motion: Position vs. velocity based formulations. *Journal of the Optical Society of America A, 4*, 503–518.

Heel, J. (1990a). Dynamical motion vision. *Robotics and Autonomous Systems, 6*, 297–314.

Heel, J. (1990b). Direct estimation of structure and motion from multiple frames. *MIT Artif. Intell. Lab. Memo, 1190*.

Hildreth, E. C. (1988). Computational studies of the extraction of visual spatial information from binocular and motion cues. *Canadian Journal of Physiology and Pharmacology, 66*, 464–477.

Hildreth, E. C., Grzywacz, N. M., Adelson, E. H. & Inada, V. K. (1990). The perceptual buildup of three-dimensional structure from motion. *Perception & Psychophysics, 48*, 19–36.

Hoff, W. & Ahuja, N. (1987). Extracting surfaces from stereo images: An integrated approach. *Proc. 1st International Conference on Computer Vision* (pp. 284–294). London, June.

Hoffman, D. D. (1982). Inferring local surface orientation from motion fields. *Journal of the Optical Society of America, 72*, 888–892.

Horn, B. K. P. & Schunck, B. G. (1981). Determining optical flow. *Artifical Intelligence, 17*, 185–203.

Husain, M., Treue, S. & Andersen, R. (1989). Surface interpolation in three-dimensional structure-from-motion perception. *Neural Computation, 1*, 324–333.

Ikeuchi, K. & Horn, B. K. P. (1981). Numerical shape from shading and occluding boundaries. *Artificial Intelligence, 17*, 141–184.

Jansson, G. & Johansson, G. (1973). Visual perception of bending motion. *Perception, 2*, 321–326.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics, 14*, 201–211.

Johansson, G. (1978). Visual event perception. In Held, R., Leibowitz, H.W. & Teuber, H.-L. (Eds). *Handbook of sensory physiology*, Berlin: Springer.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, March*, 35–46.

Kaplan, G. (1969). Kinetic description of optical texture: The perception of depth at an edge. *Perception & Psychophysics, 6*, 193–198.

Koenderink, J. J. & van Doorn, A. J. (1986). Depth and shape from differential perspective in the presence of bending deformations. *Journal of the Optical Society of America A, 3*, 242–249.

Landy, M. S., Dosher, B. A., Sperling, G. & Perkins, M. K. (1991). The kinetic depth effect and optic flow—II. First- and second-order motion. *Vision Research, 31*, 859–876.

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

Longuet-Higgins, H. C. & Prazdny, K. (1980). The interpretation of moving retinal images. *Proceedings of the Royal Society of London B, 208*, 385–397.

Loomis, J. M. & Eby, D. M. (1988). Perceiving structure from motion: Failure of shape constancey. *Proc. 2nd International Conference on Computer Vision* (pp. 383–391). Tampa, Florida, December.

Loomis, J. M. & Eby, D. M. (1989). Relative motion parallax and the perception of structure from motion. *Proc. IEEE Workshop on Visual Motion* (pp. 204–211). Irvine, Calif.

Luenberger, (1973). *Introduction to linear and nonlinear programming*. Reading, Mass.: Addison-Wesley.

Marroquin, J., Mitter, S. & Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association, 82*, 76–89.

Matthies, L. H., Szeliski, R. & Kanade, T. (1989). Kalman filter-based algorithms for estimating depth from image sequences. *International Journal of Computer Vision, 3*, 209–236.

McKee, S. P. & Welch, L. (1985). Sequential recruitment in the discrimination of velocity. *Journal of the Optical Society of America A, 2*, 243–251.

Nakayama, K. Shimojo, S. & Silverman, G. (1989). Stereoscopic depth: Its relation to image segmentation, grouping, and the recognition of occluded objects. *Perception, 18*, 55–68.

Negahdaripour, S. & Horn, B. K. P. (1987). Direct passive navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-9*, 168–176.

Petersik, J. T. (1979). Three-dimensional object constancy: Coherence of a simulated rotating sphere in noise. *Perception & Psychophysics, 25*, 328–335.

Petersik, J. T. (1987). Recovery of structure from motion: Implications for a performance theory based on the structure-from-motion theorem. *Perception & Psychophysics, 42*, 355–364.

Ramachandran, V. S., Cobb, S. & Rogers-Ramachandran, D. (1988). Perception of 3-D structure from motion: The role of velocity

gradients and segmentation boundaries. *Perception & Psychophysics, 44*, 390–393.

Rieger, J. H. & Lawton, D. T. (1985). Processing differential image motion. *Journal of the Optical Society of America A, 2*, 354–360.

Royden, C., Baker, J. & Allman, J. (1988). Perception of depth elicited by occluded and shearing motions of random dots. *Perception, 17*, 289–296.

Schumaker, L. L. (1976). Fitting surfaces to scattered data. In Lorentz, G. G., Chui, C. K. & Schumaker, L. L. (Eds) *Approximation theory II* (pp. 203–267). New York: Academic Press.

Schwartz, B. J. & Sperling, G. (1983). Nonrigid 3-D percepts from 2-D representations of rigid objects. *Investigative Ophthalmology and Visual Science, (Suppl.), 34*, 239.

Shariat, H. & Price, K. E. (1990). Motion estimation with more than two frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*, 417–434.

Siegel, R. M. & Andersen, R. A. (1988). Perception of three-dimensional structure from two-dimensional visual motion in monkey and man. *Nature, 331*, 259–261.

Snowden, R. J., Treue, S., Erickson, R. G. & Andersen, R. A. (1991). The responses of area MT and V1 neurons to transparent motions. *J. Neuroscience, 11*, 2768–2785.

Sperling, G., Landy, M. S., Dosher, B. A. & Perkins, M. E. (1989). Kinetic depth effect and identification of shape. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 826–840.

Szeliski, R. S. (1988). Bayesian modeling of uncertainty in low-level vision. PhD thesis, Department of Computer Science, Carnegie Mellon Univ., Pittsburgh, August.

Terzopoulos, D. (1986). Integrating visual information from multiple sources for the cooperative computation of surface shape. In Pentland, A. (Ed.). *From pixels to predicates: Recent advances in computational and robotic vision*. Norwood, NJ: Ablex.

Terzopoulos, D. (1988). The computation of visible-surfacer representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 10*, 417–438.

Thompson, W. B., Mutch, K. M. & Berzins, V. (1985). Dynamic occlusion analysis in optical flow fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-7*, 374–383.

Thompson, W. B., Kersten, D. & Knecht, W. R. (1992). Structure-from-motion based on information at surface boundaries. *Biological Cybernetics, 66*, 327–333.

Todd, J. T. (1982). Visual information about rigid and nonrigid motion: A geometric analysis. *Journal of Experimental Psychology, 8*, 238–252.

Todd, J. T. (1984). The perception of three-dimensional structure from rigid and nonrigid motion. *Perception & Psychophysics, 36*, 97–103.

Todd, J. T. (1985). The perception of structure from motion: Is projective correspondence of moving elements a necessary condition? *Journal of Experimental Psychology: Human Perception and Performance, 11*, 689–710.

Todd, J. T. & Bressan, P. (1990). The perception of 3-dimensional affine structure from minimal apparent motion sequences. *Perception & Psychophysics, 48*, 419–430.

Todd, J. T., Akerstrom, R. A., Reichel, F. D. & Hayes, W. (1988). Apparent rotation in three-dimensional space: Effects of temporal, spatial, and structural factors. *Perception & Psychophysics, 43*, 179–188.

Treue, S., Husain, M. & Andersen, R. A. (1991). Human perception of structure from motion. *Vision Research. 31*, 59–75.

Treue, S., Andersen, R. A., Ando, H. & Hildreth, E. C. (1995). Structure-from-motion: Perceptual evidence for surface interpolation. *Vision Research, 35*, 139–148.

Tsai, R. Y. & Huang, T. S. (1981). Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. *Univ. Illinois Urbana-Champaign, Coordinated Science Laboratory report R-921.*

Ullman, S. (1979). *The interpretation of visual motion.* Cambridge, Mass.: MIT Press.

Ullman, S. (1983). Computational studies in the interpretation of structure and motion: Summary and extension. In Beck, J., Hope, B. & Rosenfeld, A. (Eds) *Human and machine vision.* New York: Academic Press.

Ullman, S. (1984). Maximizing rigidity: The incremental recovery of 3-D structure from rigid and rubbery motion. *Perception, 13*, 255–274.

Ullman, S. & Yuille, A. L. (1987). Rigidity and smoothness of motion. *MIT Artif. Intell. Lab. Memo 989.*

Vaina, L. M., Grzywacz, N. M. & LeMay, M. (1990). Structure from motion with impaired local-speed and global motion-field computations. *Neural Computation, 2*, 420–435.

Wallach, H. & O'Connell, D. N. (1953). The kinetic depth effect. *Journal of Experimental Psychology, 45*, 205–217.

Wallach, H., Weisz, A. & Adams, P. A. (1956). Circles and derived figureds in rotation. *American Journal of Psychology, 69*, 48–59.

Waxman, A. M. & Wohn, K. (1988). Image flow theory: A framework for 3-D inference from time-varying imagery. In Brown, C. (Ed.) *Advances in computer vision* (Vol. 1, pp. 165–224). Hillsdale, N.J.: Erlbaum.

White, B. W. & Mueser, G. E. (1960). Accuracy in reconstructing the arrangement of elements generating kinetic depth displays. *Journal of Experimental Psychology, 60*, 1–11.

Yasumoto, Y. & Medioni, G. (1985). Experiments in estimation of 3-D motion parameters from a sequence of image frames. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* (pp. 89–94). New York: IEEE.

Yonas, A., Craton, L. G. & Thompson, W. B. (1987). Relative motion: Kinetic information for the order of depth at an edge. *Perception & Psychophysics, 41*, 53–59.

Yuille, A. L. & Grzywacz, N. M. (1988). A computational theory for the perception of coherent visual motion. *Nature, 333*, 71–74.

---

*Acknowledgements*—This paper describes research done at the Artificial Intelligence Laboratory, the Center for Biological Information Processing, and Whitaker College at the Massachusetts Institute of Technology. Support for the A. I. Laboratory's research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-85-K-0124. The Center's support is provided in part by the Office of Naval Research, Cognitive and Neural Sciences Division, the National Science Foundation (IRI-8719394 and IRI-8657824) and the McDonnell Foundation. This work was also supported by a grant to E. Hildreth and R. Andersen from the Educational Foundation of America and a grant to R. Andersen from the National Institutes of Health (EY 07492). S. Treue was supported by the Poitras Foundation.