MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo No. 1314 December 1991
C.B.I.P. Memo No. 70

RECOVERING THREE–DIMENSIONAL STRUCTURE FROM MOTION WITH SURFACE RECONSTRUCTION

Ellen C. Hildreth, Hiroshi Ando, Richard Andersen, Stefan Treue

ABSTRACT: 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 groups image features on the basis of their 2–D image motion to segregate features onto multiple surfaces. We present the results of computer simulations that relate the 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.

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.

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, such as stereo disparity, texture gradients or smooth shading. 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 process, both on computational and perceptual grounds. We then present a model that combines a feature–based structure–from–motion recovery 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 groups image features on the basis of their 2–D image motion to segregate features onto multiple surfaces. Finally, we present the results of computer simulations that relate the behavior of this model to psychophysical observations. In a companion paper (Treue, Andersen, Ando & Hildreth, 1992), we discuss further perceptual observations regarding the possible role of surface reconstruction in the human recovery of 3–D structure from motion.

Computational Motivations

We first provide a number of computational motivations for combining the recovery of 3–D structure from motion with a surface reconstruction process. We begin by distinguishing three terms that will be used throughout our discussion of the surface reconstruction process. Technically, 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 (in practice, computer vision systems usually perform surface approximation rather than interpolation). Both processes implicitly assume that there is only one surface to be constructed. The term surface *reconstruction* will be used to refer 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. Fig. 1 illustrates schematically, the distinction between these three terms. In Fig. 1a, the location in depth of a set of points is plotted as a function of horizontal position in the image. Points located a greater distance from the observer correspond to larger depth values. In Fig. 1b, the set of known depths are interpolated linearly between adjacent points. A smooth depth profile that only approximates the given depth data is shown in Fig. 1c. Fig. 1d shows a set of sparse points in depth that are sampled from a more elaborate set of surfaces. The reconstruction in Fig. 1e reveals a transparent cylinder (dashed contour) in front of a flat depth plane that is also occluded on the left by a slanted plane. The arrows highlight discontinuities in depth and the representation of the most distant plane is continued behind the transparent cylinder and the occluding slanted plane (see dotted line). The term "surface interpolation" will often be used informally to refer to the component of the overall surface reconstruction process that fills in the surface values on each individual surface, but the actual implementation of this process in our model will always involve an approximation rather than interpolation algorithm.

From a computational standpoint, there are several reasons to consider integrating the recovery of 3–D structure from motion with a surface reconstruction process. First, many structure–from–motion models are feature based, in that they first extract a set of moving image features such as intensity edges, corners, points, and so on, and then derive 3–D structure at the locations of these features (for review, see Ullman, 1983, 1984; Tsai & Huang, 1981; Barron, 1984; Aggarwal & Martin, 1988; Hildreth, 1988; Waxman & Wohn, 1988). If one of the goals of early visual processing is to produce a complete surface representation, in which depth or other surface shape information is known at every location in the image, 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 the recovery of 3–D structure from motion, and their detection and analysis can be considered a critical aspect of surface reconstruction. There is a need to segment the visual image into regions corresponding to distinct objects, because the nature of 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 structure–from–motion 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 in the scene generally 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 (for example, Longuet–Higgins & Prazdny, 1980; Rieger & Lawton, 1985). In general, when object surfaces are not all stationary, but can undergo their own motion through space, only the order in depth of two surfaces meeting at a boundary can be inferred from

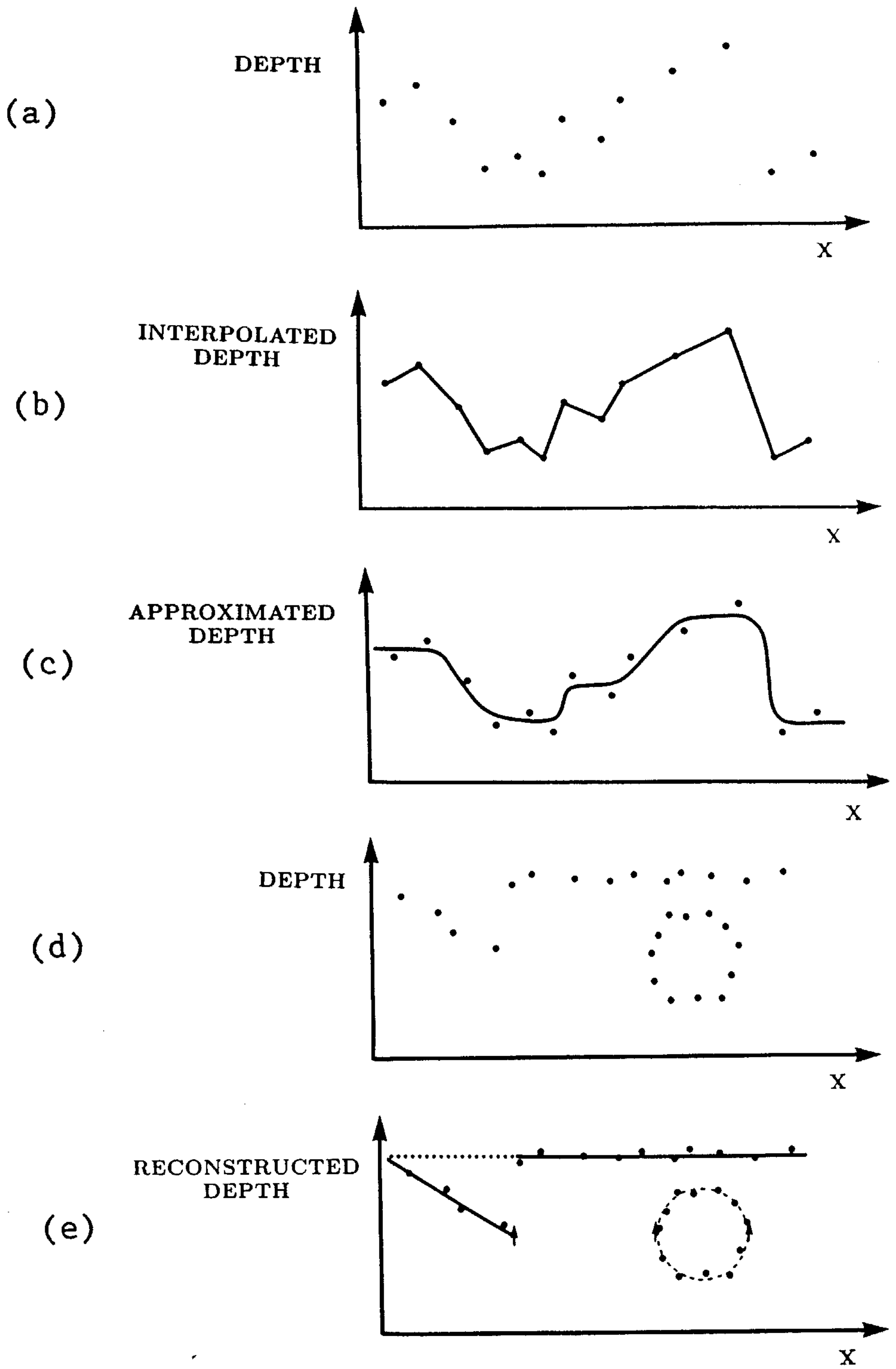

Figure 1. (a) One–dimensional depiction of a set of discrete depth samples, from which a complete surface representation is to be constructed. The depth of the points is plotted as a function of image location, X. Smooth surfaces are then constructed that (b) *interpolate* and (c) *approximate* the samples given in (a). (d) A new set of depth samples from three surfaces, one of which is transparent, that also contains depth discontinuities. (e) A *surface reconstruction* process builds a complete surface representation from the samples given in (d).

relative image motion alone. The relationship between the 2–D motion of a boundary and the motion of its surrounding image texture can be used to infer this relative order in depth of the two surfaces adjacent to the boundary (Thompson, Mutch & Berzins, 1985). Furthermore, 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, 1991). The surface reconstruction process should be able to 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 within 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 is 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. Even when an opaque surface occludes another, it may be desirable to continue the representation of the back surface for some distance behind the occluding edge of the front surface, as suggested by Nakayama, Shimojo and Silverman (1988) (an example of such a continuation is shown in Fig. 1e).

A further advantage of incorporating surface reconstruction into the structure–from–motion recovery process is that the reconstructed surface can serve as a later representation of 3–D structure, effectively replacing the depths (or relative depths) of individual features on the surface. The construction of such a representation of surface shape may facilitate subsequent tasks such as object recognition and manipulation. The reconstructed surface can also remain intact when individual features appear or disappear, for example, during momentary 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 structure–from–motion algorithms that behave in a robust manner in the presence of error in the 2–D motion measurements, or noise in the dynamic image. The incorporation of 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 error in the motion measurements.

Perceptual Motivations

There are a number of perceptual observations that either suggest a possible role of surface reconstruction in the recovery of 3–D structure from motion, or indicate how the structure–from–motion process contributes to the perception of complete 3–D surfaces. First, we can 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 (for example, Kaplan, 1969; Yonas, Craton & Thompson, 1987; Royden, Baker & Allman, 1988). Structure–from–motion displays that depict points on a rotating surface often yield a more compelling sense of 3–D structure than those that depict points distributed within a volume (for example, Green, 1961; Todd, Akerstrom, Reichel & Hayes, 1988; Dosher, Landy & Sperling, 1989b).

Perceptual observations regarding our ability to interpret structure–from–motion displays containing features with short lifetimes suggest a more direct role for surface interpolation in structure–from–motion recovery (Husain, Treue & Andersen, 1989; Dosher, Landy & Sperling, 1989a; Landy, Dosher, Sperling & Perkins, 1991; Treue, Husain & Andersen, 1991; Treue et al., 1991). 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 that was 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. In these displays, 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. In this case, observers cannot make the necessary 3–D judgement within a single point lifetime, and must integrate information across multiple point lifetimes. Dosher et al. (1989a) showed that subjects can discriminate between different complex 3–D surfaces in displays of moving random dots in which each dot has a lifetime of only two frames. Landy et al. (1991) later showed that two frames alone are sufficient for some determination of 3–D shape, although performance improves for a larger number of frames. 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. In this case, an "interpolated" representation (say, of the 2–D motion field or 3–D surface) 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 structure–from–motion recovery provides one possible account of this phenomenon.

A second observation by Treue et al. (1991) further supports the existence of an interpolation mechanism. It was found that 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. It appears that improvement of the 3–D percept in this case 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 might expect that the result of an interpolation process would improve if data were given at a larger number of image locations. Thus, this experimental observation is consistent with our intuition about how an interpolation mechanism might behave. Other perceptual studies also indicate that structure–from–motion displays containing a larger number of points can 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) using superimposed transparent cylinders and planes of dots 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 of these demonstrations, 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. 6). When presented with a display of the projected dots in motion, observers perceive two surfaces in each direction of motion that are separated in depth. The surface moving with slower speed is seen as being contained inside the faster surface. This percept can also be obtained when the points have short lifetimes, and the displays are constructed in a way that removes possible shading and texture cues due to a higher density of points along the curved boundaries of the cylinders (Treue et al., 1992). If interpolation is required to interpret structure–from–motion displays with short point lifetimes, then this observation suggests an ability to interpolate across a number of surfaces simultaneously.

In another demonstration, Ramachandran et al. (1988) present a display of two superimposed planes of random dots moving with opposite directions of motion. When points hit the edge of the display, they reverse their direction of motion so that points do not appear and disappear at the borders of the display. Ramachandran et al. report that in this case, observers perceive the moving points as lying on the surface of a rotating cylinder, rather than two flat planes. (A more extensive discussion of the perception of this display can be found in Treue et al. (1992).) This demonstration suggests that the interpretation of a boundary as being the edge of a curved surface, which might be inferred from the presence of points "bouncing off" a virtual boundary contour 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 in such a way 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 (also see Aloimonos & Huang, 1991; Treue et al., 1992). Again, the visible edges of the display may be interpreted as the curved boundary of a rotating object, leading in both cases to the percept of a more highly curved surface.

The possible influence of motion boundaries on perceived 3–D shape was also explored by Thompson et al. (1991). They constructed a dynamic random dot pattern in which a narrow vertical stripe containing dots moving with one direction and speed of motion was embedded in a surrounding pattern of dots moving with a different speed or opposite direction of motion. The borders of the slit remained stationary over time, so that dots continually appeared and disappeared along the borders of the slit. With appropriate combinations of slit size and dot velocities, the central vertical strip was perceived as a rotating cylinder. There is no variation in the 2–D motions of the dots near the borders of the strip to signal surface curvature in this case; Thompson et al. argue that qualitative information regarding the directions of motion on both sides of the boundary is used to infer the presence of surface curvature.

In a broader context, Nakayama et al. (1988) recently used the terms *intrinsic* and *extrinsic* to capture the relationship between a boundary that is observed in the image and the two surfaces on either side of the boundary. In particular, an image boundary is intrinsic to the surface to which it is physically connected. The characterization of whether a boundary is intrinsic or extrinsic to a given surface can influence the results of a number of perceptual processes, such as motion measurement, depth interpretation from stereo disparity, figure continuation and recognition (Nakayama et al., 1988). The demonstrations by Ramachandran et al. (1988; see also Aloimonos & Huang, 1991) and Thompson et al. (1991) hint that this characterization can influence perceived 3–D structure from motion. We elaborate on this point later and in our companion paper (Treue et al., 1992).

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 in the display and the relative depths between the planes. He found that subjects can accurately detect up to only three planes of dots at a time, and that the perceived separation of the planes in depth increases with the simulated separation. These observations also provide constraint on the surface reconstruction process, for example, by indicating the maximum number of surfaces that can be represented simultaneously, and may suggest a role of grouping by speed in the structure–from–motion or surface reconstruction processes.

The experiments in our companion paper (Treue et al., 1992) further support the hypothesis that surface interpolation plays a role in structure–from–motion recovery by considering the following possible consequences of such an approach. First, surface interpolation would allow 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. We provide evidence that the human recovery of structure from motion exhibits these two consequences. We also elaborate on a number of existing demonstrations by Ramachandran et al. (1988) and others, stressing aspects of these phenomena that are relevant to the work presented here.

STRUCTURE–FROM–MOTION WITH SURFACE RECONSTRUCTION

Our analysis focuses on one particular approach that combines an independent surface reconstruction process with a feature–based structure–from–motion algorithm. This section provides an overview of the model and an initial example of its behavior from computer simulations. Later sections, which elaborate on aspects of the model, further justify the choices that are made and consider alternatives.

Summary of the Model

The structure–from–motion 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 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 cues available, for example, from binocular stereo, shading, texture or perspective. 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. The original formulation presented by Ullman assumes the input to the recovery process to consist of a sequence of discrete frames, each containing a set of discrete feature points whose positions are obtained by orthographic projection of the scene onto the image plane. Extensions to the incremental rigidity scheme use velocity information directly as input and perspective projection (Grzywacz & Hildreth, 1987). Landy (1987) presents a parallel structure–from–motion model that implements a similar scheme in a cooperative network.

Limitations of existing structure–from–motion algorithms led Ando (1991) to develop the particular algorithm that is embodied in the model presented in this paper. The scheme is both velocity and feature based, in that the inputs to the algorithm 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. (We note later that the explicit reliance on discrete image features can be relaxed.) The algorithm assumes perspective projection of the scene onto the image plane, although it can interpret images obtained under orthographic projection as well. At each moment, the algorithm performs a number of iterations that alternate between a computation of 3–D velocities that maximize the rigidity of the moving configuration of points, as suggested by Ullman's incremental rigidity scheme, and a computation that derives new depths of the features from a set of equations that relate image velocity, 3–D velocity and depth. The computation of 3–D velocities also attempts to satisfy image velocity measurements as closely as possible, while allowing 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 process yields further improvement of the algorithm in the presence of significant error in the measurements of image motion. Details of this structure–from–motion algorithm are presented in a later section and in Ando (1991).

The surface reconstruction component of our model is based on a surface interpolation algorithm proposed by Grimson (1981, 1983a), which 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. Other approaches to smooth surface reconstruction have been explored, which also incorporate methods to detect and represent surface discontinuities (for example, Terzopoulos, 1988; Blake & Zisserman, 1987; Gamble & Poggio, 1987; Marroquin, Mitter & Poggio, 1987; Szeliski, 1988; for review, see Bolle & Vemuri, 1991). For simplicity, our simulations are based on Grimson's original algorithm, with minor modifications to incorporate constraints on 3–D shape imposed by object boundaries.

In organizing the overall structure–from–motion and surface reconstruction processes, we also take into account the need to incorporate additional mechanisms to allow for (1) grouping of the moving 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 current model pieces together these mechanisms as shown in Fig. 2. 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. 2 indicates one hypothesis regarding where the information obtained by different modules of the system may be built into the overall computation.

The measurement of 2–D image velocities in the vicinity of features forms the input to the structure–from–motion process (see the pathway labelled "1" in Fig. 2), which iterates between two computations that estimate relative 3–D depths and velocities. A

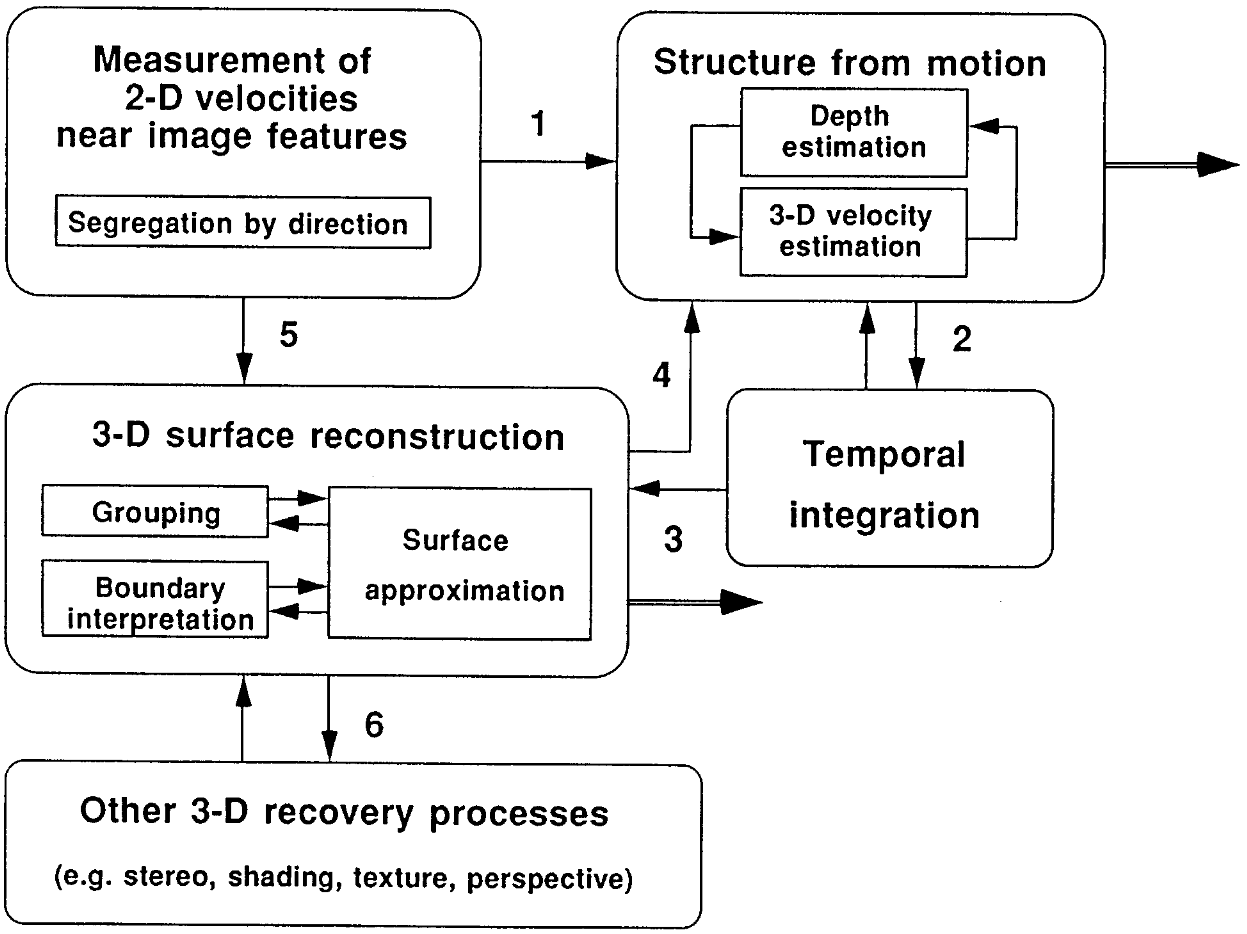

Figure 2. Diagram of a model that combines 2–D motion measurement, recovery of 3–D structure from motion, temporal integration and surface reconstruction. See text for details.

temporal integration process further improves upon the depth estimates derived from the structure–from–motion process, as indicated by the reciprocal pathways labelled "2". The structure–from–motion algorithm may be applied to all of the moving features together, regardless of their direction or speed of motion, or may follow a segregation process that at least separates features that move in opposite directions of motion within limited image regions (as in the case of images of points on the surface of a rotating transparent cylinder).

The new depths of the features derived from the structure–from–motion 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 surface depths given at each location on a fixed image grid that contains the moving features. In the case of transparency, we maintain separate representations 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 3–D information coming from other 3–D cues, such as stereo, texture, shading, and so on (pathways labelled "6").

The "boundary interpretation" component of the surface reconstruction process shown in Fig. 2 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 the 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 structure–from–motion stage (pathway labelled "4"). In the simulations described later, 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. This, for example, allows the 3–D shape of the surface to continue to improve over an extended time while the moving points persist for as little as only two frames.

As we noted earlier, the structure–from–motion 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 structure–from–motion and temporal integration stages, we present the results of an initial 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 positions and velocities of the points were projected onto the image plane using perspective projection. These image measurements were computed analytically, and noise was added in the form of Gaussian distributed perturbations of the image velocities of the points. The added noise was relative, in that it 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 the temporal integration algorithm, which effectively averages the current depth estimates with an average of past depth estimates (see Ando (1991) for a discussion of the convergence properties of the algorithm).

Fig. 3a shows a bird's eye view of the initial flat solution considered by the algorithm. Figs. 3b–3e show the solution after 3°, 10°, 35° and 250° of total rotation, respectively. It can be seen that after a short time of only a few degrees of rotation, the computed depths of the moving points already 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. Fig. 3f shows the result of the surface approximation algorithm applied to the final depths shown in Fig. 3e. 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. 3f for the front and back surfaces.

For comparison, Fig. 3g shows the results of the structure–from–motion and temporal integration algorithms applied to the unstructured stimulus explored 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 extended rotations. It can be seen that there is little difference between the results obtained for the unstructured stimulus and those obtained during the early stages of recovery of the 3–D shape of the structured cylinder (e.g. Fig. 3b), 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 elaborate on the choices made in the structure–from–motion and temporal integration components of the model. In particular, we discuss three issues: (1) the use of a feature–based algorithm, (2) the use of a velocity–based versus position–based scheme, and (3) 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, corners, blobs, and so on, and is again 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 such as significant intensity edges. First, the motion measurements directly available in the vicinity of such moving features are likely

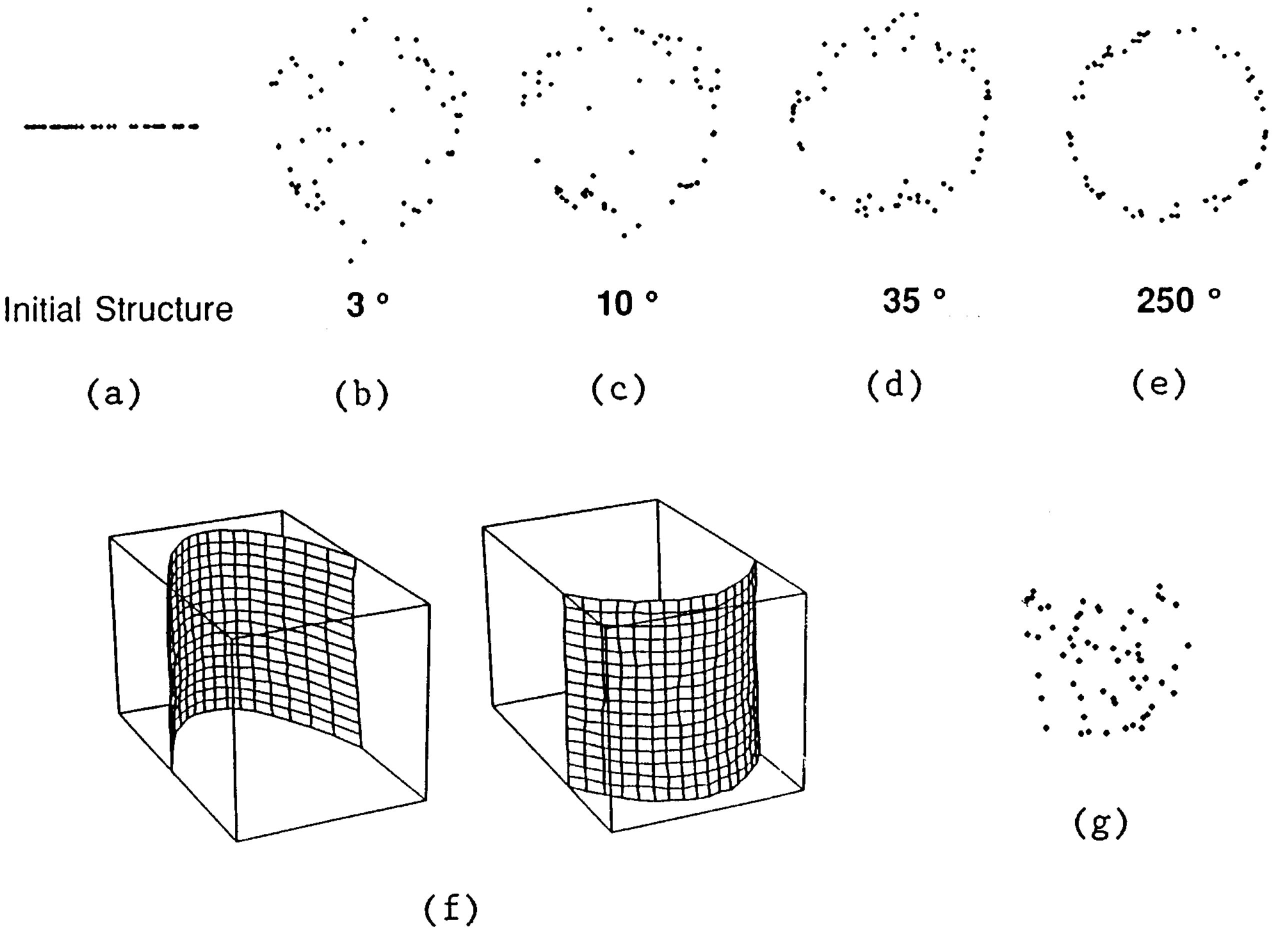

Figure 3. Example of the temporal buildup of 3–D structure from motion. 60 points were randomly positioned on the surface of a cylinder and rotated 1° per frame. Gaussian distributed noise was added to the image velocities of the points, with the space constant of the Gaussian, $\sigma = 0.5$, yielding 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).

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 structure–from–motion 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. There can be movement (or lack of movement) of the 2–D intensity pattern that is not correlated with the actual movement of the surface. For example, if the light source moves, the pattern of shading will change, yielding a motion signal, even if the surface remains stationary. In the extreme example of a rotating matte white cylinder or sphere, the surface can move without causing any change in the projected image intensities, thus yielding no 2–D motion signal. 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 interpretion of the 2–D motion of shading and shadows to recover 3–D structure requires a different strategy than the interpretion of the motions of fixed features on 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 structure–from–motion algorithm that we consider here.

Finally, the ability to recover 3–D structure at isolated features in motion allows the model to interpret displays of sparse features in motion that may not belong to a well–defined surface, for example, in the case of a random volume of points in motion. In general, the structure–from–motion process should not rely critically on the presence of a coherent surface. The separation of the structure–from–motion and surface interpolation stages allows 3–D structure to be recovered when a coherent surface does not exist.

As in the case of biological vision systems, it is not essential that the measurement of motion of a particular image feature be restricted to a single, instantaneous location, but we suggest that the structure–from–motion process initially makes use of motion information obtained within limited regions around the locations of significant intensity changes in the image. This suggestion can be made consistent with the proposal of Dosher et al. (1989a; also see Landy et al., 1991) that structure–from–motion 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. Further discussion of the biological implications of the approach presented here can be found in Treue et al. (1992).

There are at least three alternatives to the above approach. First, one could initially obtain constraints on 2–D or 3–D motions wherever there is any spatial and temporal variation of intensity (for example, Horn & Schunck, 1981; Negahdaripour & Horn, 1987), but as we noted earlier, such measurements may be unreliable in regions where the spatial and temporal gradients of intensity are small. As a second alternative, one could initially obtain motion measurements in the vicinity of image features, but then immediately "fill in" a 2–D velocity or displacement field, using a model such as that proposed by Yuille and Grzywacz (1988) (the 2–D velocity field could also be filled in implicitly through a polynomial approximation, as in Waxman and Wohn (1988)). A full 3–D surface could then be computed directly from the dense 2–D velocity or displacement field (for example, Clocksin, 1980; Longuet–Higgins & Prazdny, 1980; Hoffman, 1982; Bruss & 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 of some sort 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, one could obtain motion measurements only at the locations of image features, and then formulate a structure–from–motion recovery algorithm that computes a 3–D surface that is simultaneously consistent with the sparse motion measurements and is as smooth and rigid as possible. The main disadvantage of this approach would be 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 may compete against one another when applied simultaneously.

Positions versus Velocities as Input to the Structure–from–Motion Recovery

An issue that arises both for the human recovery of structure from motion 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 based on the use of velocities at one instant, such as the model proposed by Longuet–Higgins and Prazdny (1980), are very unstable in the presence of error in the image velocity measurements. It has been demonstrated that a number of structure–from–motion algorithms exhibit better performance when applied to image sequences with larger spatial and temporal displacements between frames (for example, Ullman, 1984; Yasumoto & Medioni, 1985; Bharwani, Riseman & Hanson, 1986; Grzywacz & Hildreth, 1987; Shariat & Price, 1990). Thus a potential computational 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 in the presence of error in the measurements of 2–D image positions or velocities, because the size of the relative motion between features is increased (Ullman, 1983). The structure–from–motion and temporal integration models proposed by Ando (1991) and presented here demonstrate, however, that it is possible to use only instantaneous velocity information and to integrate this information over time, continuously updating the computed 3–D structure, in a way that is robust in the presence of large amounts of noise.

With regard to human processing, studies that reveal our ability to recover structure from motion in displays with very short point lifetimes suggest that the human recovery of 3–D structure 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). The minimal lifetime required for successful recovery of 3–D shape is similar to the minimal time required for accurate 2–D velocity estimation (for example, McKee & Welch, 1985). The motion measurements that form the input to structure–from–motion recovery may encode image velocity or may capture information such as motion energy, as Dosher et al. (1989a; see also Landy et al., 1991) suggest. Over a range of angular velocities of a rotating cylinder, it appears that points must be visible for a minimum amount of 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), as one might expect that if a strictly position–based scheme were used, then a minimum displacement of the points may be necessary to build up 3–D structure. Other 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 to recover structure from motion (for example, Todd et al., 1988; Braunstein, Hoffman & Pollick, 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 recover 3–D structure from motion (Siegel & Andersen, 1988; Andersen & Siegel, 1990).

An alternative to the use of velocity (or other "instantaneous") information alone is that the visual system uses both velocity and position information. If, for example, velocity measurements were used to guide the tracking of the positions 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, may suggest some role for positional information in human structure–from–motion recovery (Vaina, Grzywacz & LeMay, 1990).

Temporal Integration through Sequential Updating of 3–D Structure

The combination of sequentially updating 3–D structure and using additional temporal integration allows the structure–from–motion 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 continuously changing 3–D structure. Second, the structure–from–motion component of the scheme relaxes the rigidity constraint by computing a set of 3–D velocities at each moment that only *maximize* rigidity, rather than requiring moving objects to remain strictly rigid over time. If a viewed object changes nonrigidly, then the 3–D model computed by the structure–from–motion algorithm will be forced to change over time. Furthermore, when the algorithm begins with an initial 3–D structure that is flat, the computed 3–D structure initially changes over time through incremental improvement, even for the case of a rigid object in motion.

A number of factors contribute to the ability of the scheme to cope with large amounts of error in the image motion measurements. First, the use of a temporal integration process that effectively averages computed 3–D structures over time reduces the influence of errors that are uncorrelated over time. Second, the relaxation of the rigidity constraint also allows the structure–from–motion recovery to cope with large errors, as these errors essentially result in nonrigid distortions of the computed 3–D structure. Third, the algorithm computes 3–D velocities that only approximately satisfy the image motion measurements, allowing some deviation of these measurements from the true motions projected onto the image plane.

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

Perceptual observations also support a model of this general type for the human recovery of 3–D structure from motion. First, many experiments indicate an incremental buildup of perceived 3–D structure over an extended time period (for example, Wallach & O'Connell, 1953; White & Mueser, 1960; Braunstein & Andersen, 1984b; Doner, Lappin & Perfetto, 1984; Braunstein et al., 1987; Siegel & Andersen, 1988; Husain et al., 1989; Hildreth et al., 1990; Treue et al., 1991, 1992). The experiments by Hildreth et al. (1990) suggest continued improvement in the accuracy of structure–from–motion judgements up to a second or so of viewing time, 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). A number of studies indicate that substantial 3–D information may be derived from only two frames (for example, Todd et al., 1988; Dosher et al., 1989a; Braunstein et al., 1990; Todd & Bressan, 1990), but these studies either oscillate a two–frame image sequence for an extended period of time, until the observer makes a judgement, 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 recovery of 3–D structure from motion. Landy et al. (1991) show that subjects can perform some discrimination of complex 3–D shapes from short–duration stimuli consisting of only two frames, although performance is better for more extended sequences.

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 (for example, Johansson, 1973, 1978; Jansson & Johansson, 1973; Cutting, 1982; Todd, 1982, 1984, 1985; Loomis & Eby, 1988, 1989). Displays of rigid objects can 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, allowing the interpretation of nonrigid motions, is an essential component of any model proposed for the human recovery of structure from motion.

Perceptual experiments also suggest that the human recovery of structure from motion can cope with significant amounts of noise in the moving image (for example, Petersik, 1979; Doner, Lappin & Perfetto, 1984; Todd, 1984, 1985; Husain et al., 1989; Hildreth et al., 1990). The addition of noise can sometimes lead to the perception of nonrigid distortions of the moving object.

Another property of our model is that, similar to the incremental rigidity scheme, the current estimate of 3–D structure can constrain future estimates of structure. By measuring the consequence of manipulating observers' current perceived structure, Hildreth et al. (1990) found some evidence that the human recovery of structure from motion exhibits this property. It also allows the algorithm to recover the 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 from motion using the rigidity constraint alone (Ullman, 1979; Tsai & Huang, 1981). Additional constraint comes from the fact that this particular model derives the most rigid interpretation, starting from an initial hypothesis that the object is flat.

THE SURFACE RECONSTRUCTION PROCESS

This section explores a number of aspects of the surface reconstruction process. In particular, we address the following issues: (1) the separation of the structure–from–motion and surface interpolation components of the 3–D recovery process, (2) the importance of the grouping of points by their 2–D motion for the surface reconstruction process, (3) the use of an interpolation scheme based on the computation of the "smoothest" surface consistent with the sparse 3–D data obtained by the structure–from–motion recovery, and (4) the importance of boundary constraints for surface reconstruction.

Separating Structure–from–Motion Recovery and Surface Reconstruction

Our motivation for separating the surface reconstruction process from the structure–from–motion 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, perspective, and so on. Many computational models for these visual processes are feature based, in that they derive 3–D information initially at the locations of image features such as intensity edges. Such models all require a subsequent interpolation stage to fill in 3–D information between image features, and 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 separate visual module. Even if the processes analyzing different 3–D cues could produce a dense representation of 3–D shape, 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.

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, 1988), the use of 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 here the argument that we can derive a strong sense of 3–D structure in situations where there is only a weak sense of surface, for example, in the case of volumes of isolated points in motion. It is therefore essential that the structure–from–motion process not rely critically on the existence of a well–defined surface. By postulating the existence of a structure–from–motion process that is separate from the surface reconstruction process, we preserve the ability to cope with such visual patterns.

Grouping Points by 2–D Motion

At some stage, moving points may be segregated into different groups on the basis of their speed or direction of motion. One can easily justify using this segregation of points for the surface reconstruction stage. 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 within a limited area of the image, 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 the presence of two surfaces undergoing different motions that are adjacent to one another with a boundary between them. Both possibilities signal the presence of multiple surfaces, which should be taken into account in building a complete representation of the structure of object surfaces. Thus it seems essential that grouping processes operating on the 2–D motion measurements precede the surface reconstruction stage.

On the other hand, the structure–from–motion process specifically uses *relative* motion to infer relative depth. The larger the relative motion between features, the stronger the structure–from–motion cue. If one were to segregate moving features first on the basis of direction or speed of motion, and then attempt to recover the 3–D structure of the separate groups independently, the structure–from–motion recovery would be inherently less reliable, because it now must depend on smaller relative motions between features. In the case of objects such as the rotating transparent cylinder, which contain oppositely directed motions everywhere in the image and significant variation in speed within each motion direction, points moving in each direction can be grouped together and the structure–from–motion algorithm can be applied separately to the two groups, without a significant loss of quality in the final solution. Alternatively, the structure–from–motion algorithm can be applied to all of the points together, although in general, the presence of motions in different directions superimposed in a limited region of the image is likely to arise from two independently moving transparent surfaces.

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 et al., 1991), which is critical to the recovery of structure from motion (Siegel & Andersen, 1988; Andersen & Siegel, 1990). These experiments indicate that in V1, two populations of neurons with different preferred directions of motion are activated by the two differently moving surfaces with little interaction between them, suggesting that a segregation of the two surfaces could be achieved early in the visual pathway. The strong inhibition between neurons of different preferred directions in area MT suggests that some form of integration of different motion directions is taking place at this level that could be relevant to the processes of structure–from–motion recovery or surface reconstruction.

Smooth Surface Interpolation

Our model uses a surface interpolation strategy that derives the "smoothest" surface consistent with the depth data given by the structure–from–motion recovery. Grimson (1980, 1982, 1983b) presented strong mathematical and physical motivations for this approach. From a mathematical perspective, this strategy can be formalized in a way that guarantees the computation of a unique surface. From a physical standpoint, one can show formally and rigorously using the physics of image formation, that the smoothest possible surface consistent with depth data derived from a feature–based 3–D recovery algorithm (i.e., from stereo, motion or other 3–D source) is most consistent with the structure of the image intensity function.

A second advantage of this strategy is that algorithms to perform smooth surface interpolation can be designed that 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 (for example, Grimson, 1983a; 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.

When viewing structure–from–motion displays that are constructed from smooth surfaces and contain a high density of points, we generally perceive a smoothly curved surface everywhere. If the density of points is low, however, the surface often appears to consist of planar facets surrounded by edges that connect local triplets of nearby points. Surface interpolation algorithms have been proposed that first perform a triangulation of a set of 3–D points, after which planar surface patches could be fit to local triplets of points (for example, Faugeras, LeBras–Mehlman & Boissonat, 1990), but current algorithms of this type use global operations over the entire set of points 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 a number of reasons, it is useful to incorporate explicit constraints regarding object or surface boundaries into the structure–from–motion and surface reconstruction processes. First, the explicit detection of boundaries allows segmentation into distinct objects or surfaces prior to the structure–from–motion recovery. As we noted earlier, the rigidity constraint that forms the basis of most structure–from–motion algorithms is more appropriately applied within the surfaces of single objects. The relaxation of the rigidity constraint in our model provides some ability to cope with multiple objects moving nonrigidly with respect to one another, but the algorithm would clearly 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, in the simulations that we have conducted, we found 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 tries to minimize the variation in depth with respect to distance in the image. At a curved boundary, depth is changing very rapidly with respect to distance in the image.

Finally, we note again that some of the demonstrations by Ramachandran et al. (1988; see also Aloimonos & Huang, 1991) and Thompson et al. (1991) summarized earlier suggest that object boundaries can play a significant role in the human perception of 3–D shape from motion. 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 in the image (Ramachandran et al., 1988; Treue et al., 1992). 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., 1991).

DETAILS OF THE MODEL

This section presents further details of the structure–from–motion, temporal integration and surface reconstruction components of the model presented earlier and illustrated in Fig. 2. To facilitate the development of these details, we first summarize previous formulations of Ullman's incremental rigidity scheme (Ullman, 1984) that assume orthographic projection of the scene onto the image plane.

Previous Formulations of the Incremental Rigidity Scheme

Ullman's original formulation of the incremental rigidity scheme (Ullman, 1984) assumes the visual input to consist of a sequence of frames, each containing a number of discrete points that correspond to identifiable features in the changing image. The scheme maintains and updates an internal model $M(t)$ of the viewed objects, which consists of a set of 3–D coordinates: $M(t) = (x_i(t), y_i(t), z_i(t))$. The scheme assumes orthographic projection onto the image plane, so that $(x_i(t), y_i(t))$ are the image coordinates of the i–th point, and $z_i(t)$ is the current estimate of depth at the i–th point. (The scheme can be modified to use perspective projection (Grzywacz & Hildreth, 1987).) When no other 3–D cues are present, the initial model $M(t)$ at $t = 0$ is taken to be flat; that is, $z_i(0) = 0$ (or some other constant value) for $i = 1, \ldots, n$, where n is the number of points in motion.

Given a current model $M(t)$ at time t and the image of the moving points in a new frame at a later time t', the problem is to compute a new model $M(t')$ such that the transformation from $M(t)$ to $M(t')$ is as rigid as possible. Since $x_i(t')$ and $y_i(t')$ are known, this requires the computation of the unknown depth values $z_i(t')$. (It is assumed that the correspondence between points in the two successive frames is known.) The new depth values are computed as follows. Let $l_{ij}(t)$ denote the distance between points i and j at time t. To make the transformation as rigid as possible, the values $z_i(t')$ for the new model are chosen so as to make $l_{ij}(t)$ and $l_{ij}(t')$ as similar as possible. For this purpose, Ullman defined a measure of the difference between $l_{ij}(t)$ and $l_{ij}(t')$ as:

$$d(l_{ij}(t), l_{ij}(t')) = \frac{(l_{ij}(t) - l_{ij}(t'))^2}{l_{ij}^3(t)}, \tag{1}$$

and formulated the recovery of structure as the computation of the $z_i(t')$ that minimize the following overall deviation from rigidity:

$$D_d(t,t') = \sum_{i,j} d(l_{ij}(t), l_{ij}(t')). \tag{2}$$

After the values $z_i(t')$ have been determined using this minimization process, the new model $M(t') = (x_i(t'), y_i(t'), z_i(t'))$ becomes the current model. A new frame is then registered and the process repeats itself. In this way, the scheme maintains rigidity by keeping the total distances between points in the model as constant as possible. The motivation for the cubic factor in the denominator of equation (1) is that the nearest neighbors to a given point are more likely to belong to the same object than distant neighbors, so that a point is more likely to move rigidly with its nearest neighbors. The $l^3_{ij}(t)$ factor diminishes the influence of distant points in the recovery of structure. Simulations with Ullman's scheme suggest that the use of this distance factor can yield improved results when the scene contains multiple objects undergoing different motions, but the cubic distance factor $l^3_{ij}(t)$ may decrease too rapidly with distance. A more flexible function to use is a Gaussian function such as $e^{\frac{-l^2_{ij}}{2\sigma^2}}$, whose space constant, σ, can be varied.

In the case of orthographic projection, only relative depth values, $z_i(t) - z_j(t)$, can be recovered, rather than absolute depth values, because under this form of projection, the image of a given object does not change with its absolute depth. In addition, 3–D structure is determined only up to a reflection about the image plane.

Grzywacz and Hildreth (1987) developed a continuous formulation of the incremental rigidity scheme, which uses velocity information at discrete feature points in a continuously changing image as input to the recovery of 3–D structure. In this formulation, it is again assumed that there always exists an internal model $M(t) = (x_i(t), y_i(t), z_i(t))$, and that the image velocities $\dot{x}_i(t)$ and $\dot{y}_i(t)$ are known. The problem is then formulated as the computation of the z components of velocity, $\dot{z}_i(t)$, that minimize the total continuous change in the distances between the points. The measure of overall deviation from rigidity is given by:

$$D_c(t) = \sum_{i,j} (\dot{l_{ij}}(t))^2 \tag{3}$$

where $\dot{l_{ij}}(t)$ denotes the time derivative of the distance $l_{ij}(t)$, which depends on the velocities $(\dot{x}_i(t), \dot{y}_i(t), \dot{z}_i(t))$. At each moment, the current model $(x_i(t), y_i(t), z_i(t))$ and image velocities $(\dot{x}_i(t), \dot{y}_i(t))$ are given, and new velocities in depth $\dot{z}_i(t)$ are computed. The final 3–D velocities are then used to derive a new model $(x_i(t'), y_i(t'), z_i(t'))$. The additional factor of l^3_{ij} that appears in the denominator of equation (1) (or a Gaussian function as noted above) could also be used in the measure shown here. In other respects, this continuous formulation is similar to Ullman's position based algorithm. A model of the structure of the moving points is built up by continually taking into account new velocity information over an extended time period.

Building Upon the Incremental Rigidity Scheme

The structure–from–motion algorithm proposed by Ando (1991) and incorporated in our model is a velocity–based algorithm that builds upon the continuous formulation of Ullman's incremental rigidity scheme described in the previous section in four ways: (1) it uses perspective projection, (2) it explicitly allows error in the image velocity measurements, (3) rather than assuming the current model of the depths of image features to be fixed at each moment and deriving only a new set of 3–D velocities, the algorithm allows the current estimates of depth to be modified, and (4) it allows variable weighting of the strength of rigidity between pairs of image features.

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

$$(x_i, y_i) = \left(\frac{X_i}{Z_i}, \frac{Y_i}{Z_i}\right) \tag{4}$$

and

$$(\dot{x}_i, \dot{y}_i) = \left(\frac{\dot{X}_i - x_i\dot{Z}_i}{Z_i}, \frac{\dot{Y}_i - y_i\dot{Z}_i}{Z_i}\right). \tag{5}$$

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:

$$E_D + \lambda E_R \tag{6}$$

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 λ 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 (5) 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 data term is then written as follows:

$$E_D = \sum_i [\beta_{xi}(\dot{x}_i Z_i + x_i\dot{Z}_i - \dot{X}_i)^2 + \beta_{yi}(\dot{y}_i Z_i + y_i\dot{Z}_i - \dot{Y}_i)^2] \tag{7}$$

where β_{xi} and β_{yi} are the weights associated with individual velocity measurements.

To derive the term that captures deviation from rigidity, the measure presented above in equation (3) is modified to incorporate weighting factors w_{ij} that capture the strength of the rigidity of the connection between point i and point j. We therefore have:

$$E_R = \sum_{ij} w_{ij}(\dot{l}_{ij})^2 \tag{8}$$

where $\dot{l}_{ij}$ denotes the time derivative of the 3–D distance between points i and j. 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, as suggested earlier.

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

$$E_R = \sum w_{ij} \frac{[(x_i Z_i - x_j Z_j)(\dot{X}_i - \dot{X}_j) + (y_i Z_i - y_j Z_j)(\dot{Y}_i - \dot{Y}_j) + (Z_i - Z_j)(\dot{Z}_i - \dot{Z}_j)]^2}{(x_i Z_i - x_j Z_j)^2 + (y_i Z_i - y_j Z_j)^2 + (Z_i - Z_j)^2}. \tag{9}$$

The above cost functional is non–quadratic 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 (5). (Note that equation (5) 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) 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 structure–from–motion 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 from multiple frames.

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 (Anderson & Moore, 1979; Gelb, 1974). Recently, Kalman filtering has been applied to 3–D structure estimation problems in computer vision (for example, Matthies, Szeliski & Kanade, 1989; Heel, 1990a,b), and has been shown to improve significantly the quality of the estimated structure over time. Ando (1991) 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. Further details of its derivation and underlying assumptions are discussed in Ando (1991).

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. The algorithm integrates the predicted depths and the newly computed depths 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 effectively 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).)

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 $\tilde{\mathrm{Z}}_t$ and $(\dot{X}_t, \dot{Y}_t, \dot{Z}_t)$, respectively. Let the estimates of depth at time t be denoted by $\hat{\mathrm{Z}}_t^-$ and $\hat{\mathrm{Z}}_t^+$, 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:

$$\hat{Z}_t^- = \hat{Z}_{t-1}^+ + \dot{Z}_{t-1}\Delta t \tag{10}$$

where Δt 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:

$$\hat{Z}_t^+ = \frac{1}{\alpha_t^- + \alpha_t}\left(\alpha_t^-\hat{Z}_t^- + \alpha_t\tilde{Z}_t\right) \tag{11}$$

where α_t^- and α_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 effectively achieves the integration of a number of past depth measurements without the need to store them all. Thus, as more images are obtained, the accuracy of the estimates of depth improves incrementally over time.

The Surface Interpolation Algorithm

A number of algorithms have been proposed for performing explicit smooth surface reconstruction from sparse depth data (for example, Schumaker, 1976; Grimson, 1981; Boult & Kender, 1986; Terzopoulos, 1986, 1988; Blake & Zisserman, 1987; Gamble & Poggio, 1987; Marroquin et al., 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. (We also noted earlier that approaches based on fitting planar patches to local triplets of points may be useful to consider (for example, Faugeras et al., 1990).) The various 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 was completed, and a standard 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. Our computer simulations used Grimson's original surface approximation algorithm, primarily for its simplicity, but we made some simple modifications to handle boundary information.

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:

$$\int\int(\frac{\partial^2 S}{\partial x^2} + 2\frac{\partial^2 S}{\partial x \partial y} + \frac{\partial^2 S}{\partial y^2})dxdy + \lambda_s \sum_{\vartheta}(S(x,y) - C(x,y))^2 \qquad (12)$$

where the discrete summation in the second term takes place over the set ϑ 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. λ_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 (for example, 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 have considered two simple modifications to handle constraints on the depth variations 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 both to the line of sight and to the 2–D projection of the boundary contour, as suggested by Ikeuchi and Horn (1981) and used by Aloimonos and 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.

Combining Structure–from–Motion with Surface Interpolation

We directly combine the velocity based structure–from–motion 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 then consists of the following steps: (1) the set of discrete features undergoes small displacements in the image, and the structure–from–motion 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 provides additional simulations that were conducted with the model described in the previous sections. In particular, we consider the following: (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 the interpretation of object boundaries on the overall surface reconstruction. Our simulations indicate that while the surface reconstruction process may play a key role in accounting for a number of these phenomena, there are other aspects of the motion measurement and structure–from–motion stages that can also contribute in a significant way to our final 3–D percept.

Coping with Short Point Lifetimes

A primary perceptual motivation for considering the incorporation of a surface interpolation mechanism into the structure–from–motion process is derived from our ability to perceive 3–D shape in displays with moving points that continually disappear and reappear, with short lifetimes (Husain et al., 1989; Dosher et al., 1989a; Landy et al., 1991; Treue et al., 1991, 1992). 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 used here was motivated in part by the psychophysical studies cited earlier). Using perspective projection, the positions and velocities of the points on the image plane were computed analytically. Relative noise was added in the form of Gaussian distributed perturbations of the velocities of the points, scaled by the magnitude of the velocity components. The initial 3–D structure considered by the algorithm was flat. Fig. 4 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 for each point lifetime used here was comparable.) For the results shown in Fig. 4a, 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

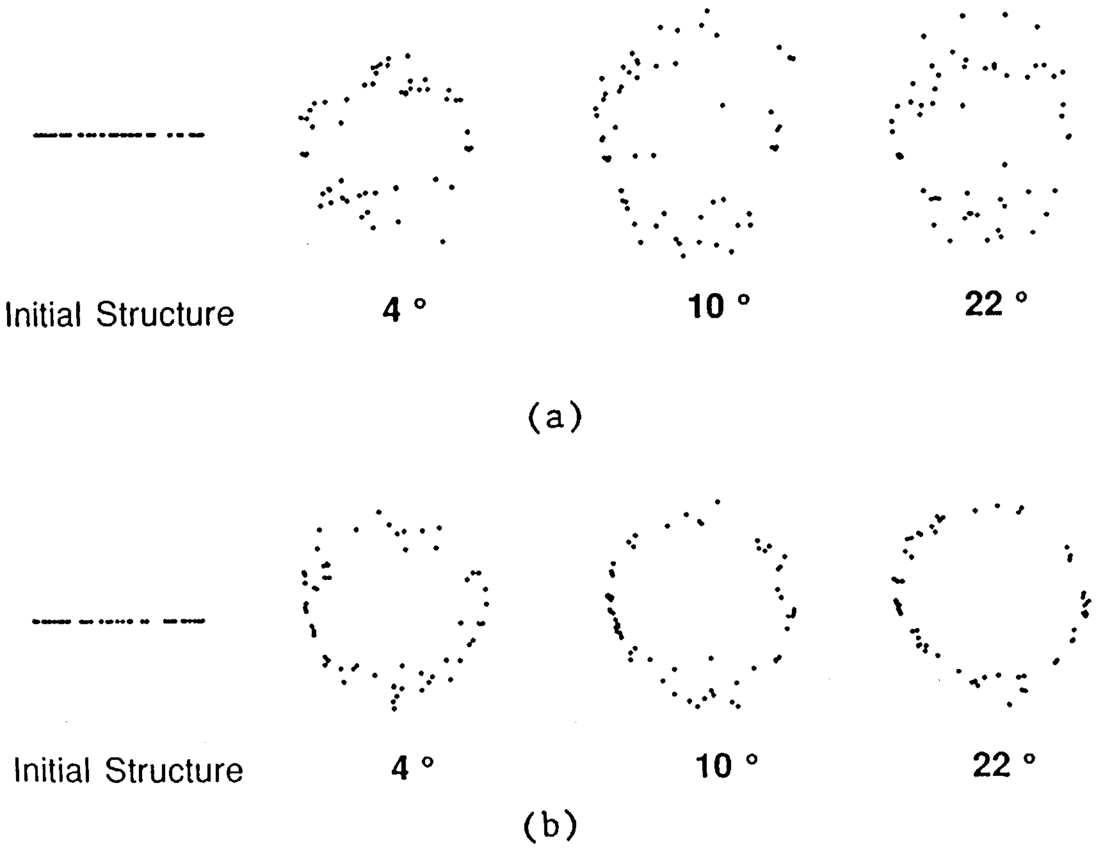

Figure 4. Comparison of the results of the model for displays containing points with short lifetimes (a) without surface interpolation and (b) with surface interpolation. The results are shown after total rotations of 4°, 10° and 22°. Relative Gaussian noise was added to the image velocities of the points, 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.

of 4° of the points. As shown in Fig. 4a, 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. 4b, a 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 stage, there is a rapid convergence toward the cylindrical structure. The surface interpolation allows the temporal integration process to integrate motion information over a more extended time and also helps to smooth out fluctuations in the 3–D structure derived from the structure–from–motion algorithm as a result of the added noise, by imposing a smoother surface interpretation on the sparse 3–D data.

Note that 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 of points.

Relating these results back to the perceptual demonstrations, in the case of the experiments of Husain et al. (1989; 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. 4a 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. 4b 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, in general, on a number of parameters used in the structure–from–motion recovery process, including the factor λ 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 eq. 6), the distance metric used to weigh the rigidity of the connection between each pair of points (see the discussion around eq. 8), 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).

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 in these sparse displays 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. This section presents the results of simulations with the surface interpolation algorithm on its own that may relate to our perception of displays with fewer points.

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. 5. 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 17×17 elements, in Figs. 5a and 5b, 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 of this dependence, Fig. 5c was constructed from a set of 6 points whose positions were skewed towards one half of the image of the cylinder. In this case, 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 it may use a process that requires time to converge to a solution, with the amount of time again 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 a reasonable approximation to the smoothest surface, but the results would degrade as the data become more 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 such factors as the distance to known depth points or the amount of time that has ellapsed since an explicit data point appeared near a given location. As a consequence of the second of these two 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 the case of 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

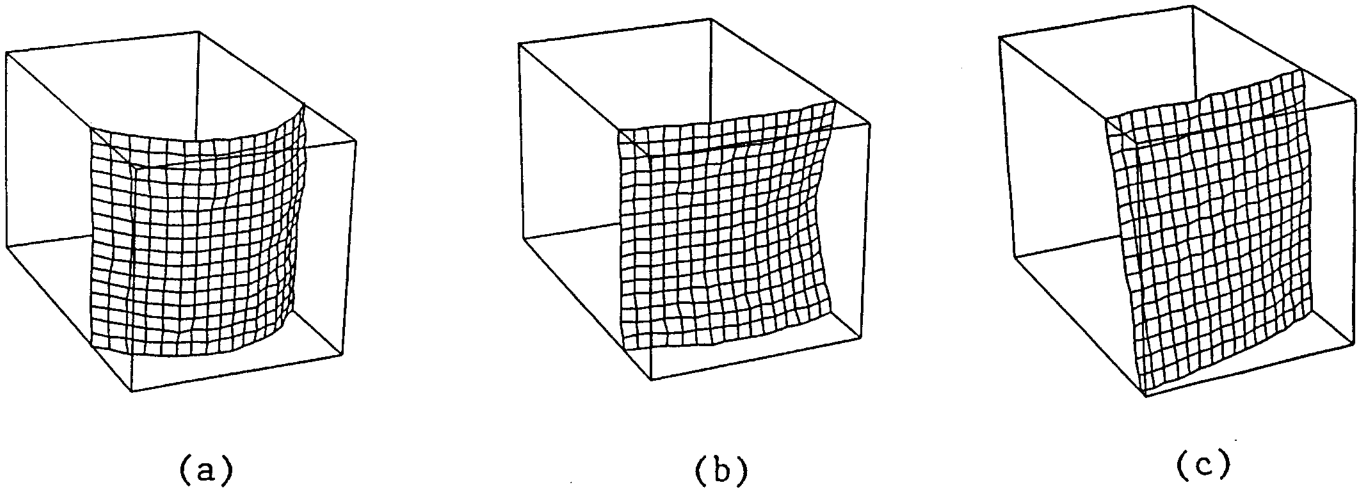

Figure 5. Degradation of the surface interpolation process for fewer points. A set of points in depth are first sampled from the surface of a cylinder and locations at which no explicit depth information is given are initially assigned a depth that is roughly the average of the known points. (a) The 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.

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 overall judgement of 3–D structure.

There are parameters used in the structure–from–motion recovery algorithm that also may affect the relative performance of the overall model for sparse and dense patterns of points. For example, referring to eq. 6, there is a factor λ 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 λ, 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.

Ramachandran et al.'s Two–Cylinders Demonstrations

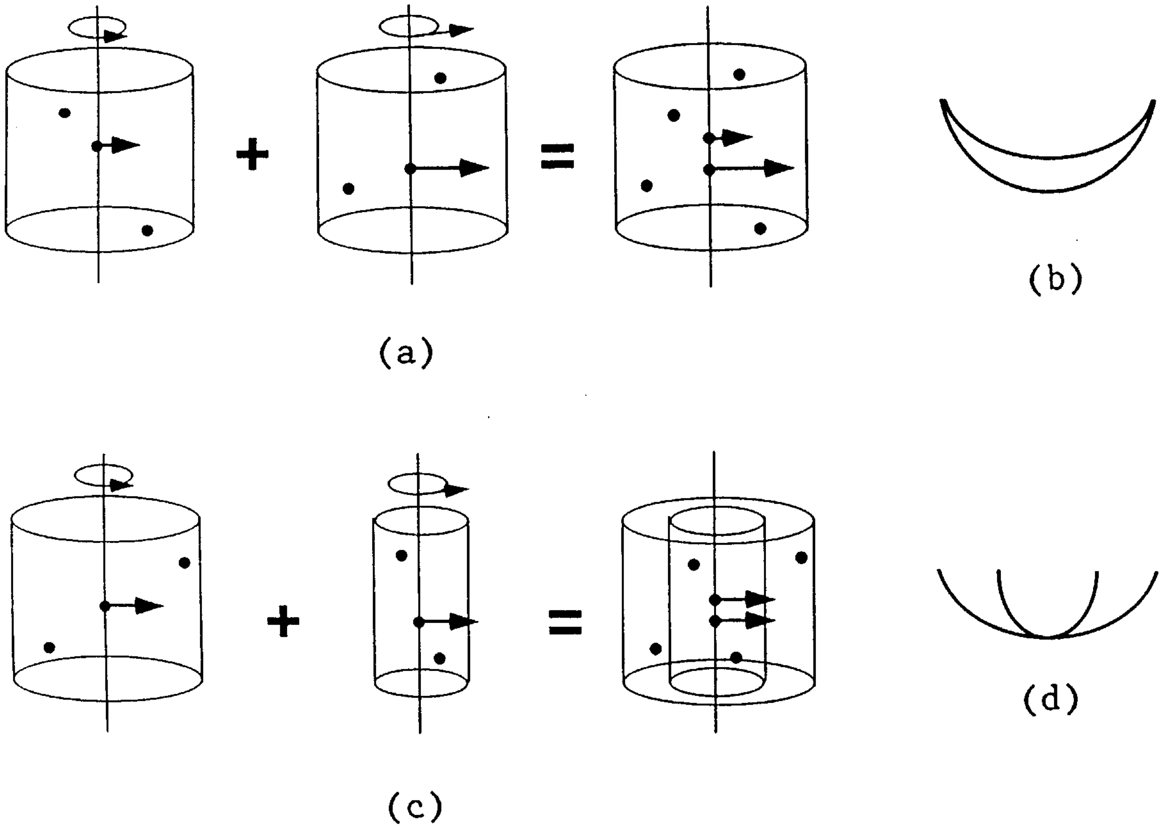

Figure 6. The perceptual demonstrations of Ramachandran et al. (1988). (a) Schematic drawing of the demonstration in which two cylinders of the same size are superimposed in the same region of space, but rotated at different speeds. (b) Birds' eye view of the percept obtained from a display created from (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).

Some of 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 the model we propose 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. Fig. 6a shows a schematic illustration of the two cylinders that underlie the construction of the visual stimulus, and Fig. 6b shows a birds' eye view of the resulting percept. The structure-from-motion process embodied in our model derives a 3-D structure that is consistent with this percept, as illustrated in Fig. 7. Fig. 7a shows a birds' eye view of the true 3-D structure and Fig. 7b shows the results of the structure–from–motion algorithm applied to two frames that were separated by 1° of rotation of the points. No noise was added to the image velocities in this case. The points were subsequently grouped by their speed of motion, and separate 3–D surfaces were reconstructed for the two populations undergoing different speeds. (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. 7c. The overall impression of one surface bulging out from the other is clearly conveyed in these results. This phenomenon is also preserved when short point lifetimes are used (Treue et al., 1992).

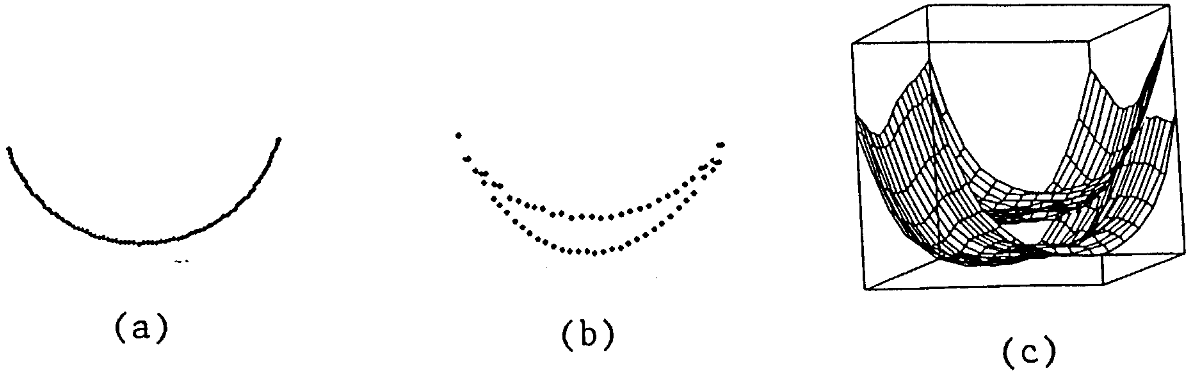

Figure 7. Simulations with the model applied to perceptual demonstrations of Ramachandran et al. (1988) shown in Figs. 6a and 6b. (a) Birds' eye view of the true 3–D structure. (b) The results of the structure–from–motion algorithm applied to two frames separated by 1° of rotation of the points. (c) The results of the surface reconstruction stage.

This result can be explained in terms of the goal of the structure–from–motion algorithm. The stimulus itself is a nonrigid structure and the structure–from–motion 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. 6b is, in fact, the *most rigid* structure consistent with the projected image velocities. In some sense, 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. Fig. 6c shows a schematic illustration

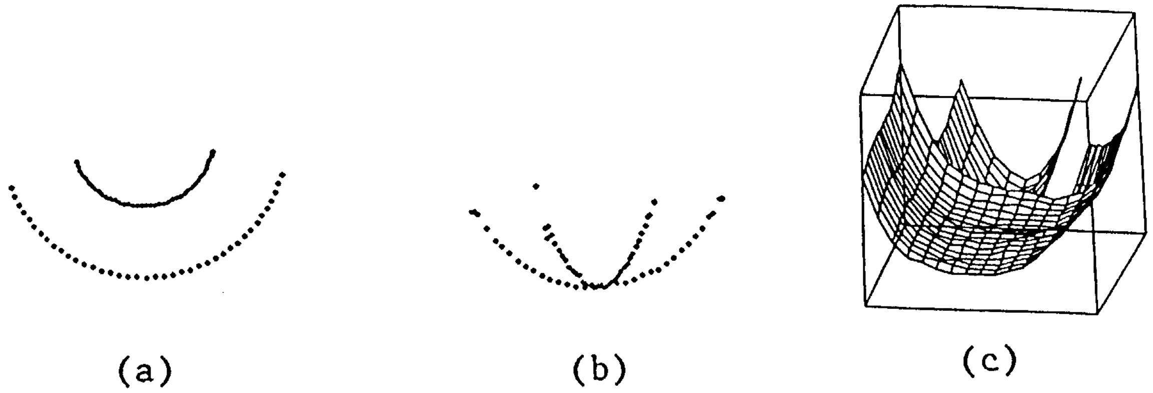

Figure 8. Simulations with the model applied to perceptual demonstrations of Ramachandran et al. (1988) shown in Figs. 6c and 6d. (a) Birds' eye view of the true 3–D structure. (b) The results of the structure–from–motion algorithm applied to two frames that are separated by 1° of rotation of the points. (c) The results of the surface reconstruction stage.

of the two cylinders in this case, and Fig. 6d shows a birds' eye view of the resulting percept. The results of the simulations with our model are shown in Fig. 8. Fig. 8a shows a birds' eye view of the true 3–D structure and Fig. 8b shows the results of the structure–from–motion algorithm applied to two frames separated by 1° of rotation of the points. Again, no noise was added to the image velocities in this case. 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 3–D surfaces were reconstructed for the two populations of points. The result of this surface reconstruction stage are shown in Fig. 8c. The results capture the overall subjective impression of the two surfaces merging into a single surface in the center. This phenomenon is also preserved when short point lifetimes are used (Treue et al., 1992). 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.

The Influence of the Interpretation of Boundaries

The last issue that we address is the possible influence of contraints on 3–D shape provided by the interpretation of object boundaries. We again consider demonstrations by Ramachandran et al. (1988) that illustrate this possible influence in the human recovery of structure from motion.

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). We elaborate on the perception of 3–D shape in this display in our companion paper (Treue et al., 1992); briefly, we 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 the two–planes demonstration. First, the initial motion measurement mechanisms may incorrectly represent the velocities of points 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 some variation in the speed of motion of points near the borders that the structure–from–motion 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 structure–from–motion process combines all of the points together and tries to interpret their motion as due to a single object undergoing minimal distortion over time, a curved structure may then emerge at this stage. In the simulations, we explore these different possible sources of the perception of curvature.

In the first simulation, we show the result of applying the structure–from–motion 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. 9a. The true structure consisting of two flat planes of points superimposed at the same depth is shown on the left and the computed 3–D structure is shown on the right. The computed structure 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. In particular, we varied the image speed of points near the border so that speed drops off linearly with decreasing distance from the border. The structure–from–motion algorithm was then applied to the resulting velocity pattern. The result is shown in Fig. 9b. 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 the surface interpolation algorithm 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. 9c. It can be seen that the added boundary constraint can yield a curved surface near the edges of the display.

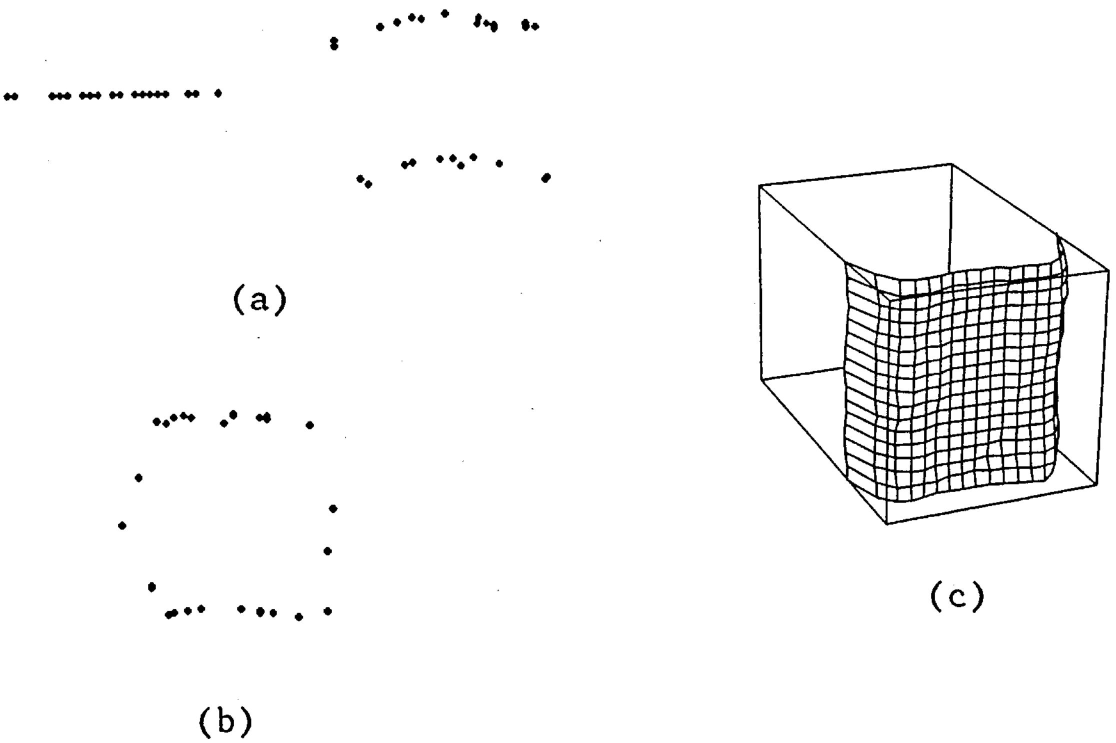

Figure 9. Simulations with the two–planes demonstration of Ramachandran et al. (1988). (a) The true structure (left) consists of two sets of points at the same depth, which translate to the left and right, respectively. On the right are shown the results obtained when the structure–from–motion algorithm 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.

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

In a second demonstration, Ramachandran et al. (1988) began with a random–dot cylinder and masked off vertical sections on the left and right edge of the cylinder. It was reported that the truncated cylinder is then perceived as a cylinder with a smaller radius. We note in our companion paper that we do perceive a single, narrower object in rotation in this case, 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 et al. (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 lead to the percept that the borders are the boundaries of an object rotating in depth. When the figure is surrounded by a mask that yields the percept of a surface viewed through 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. For illustration of these visual patterns and their physical interpretation, see our companion paper (Treue et al., 1992).

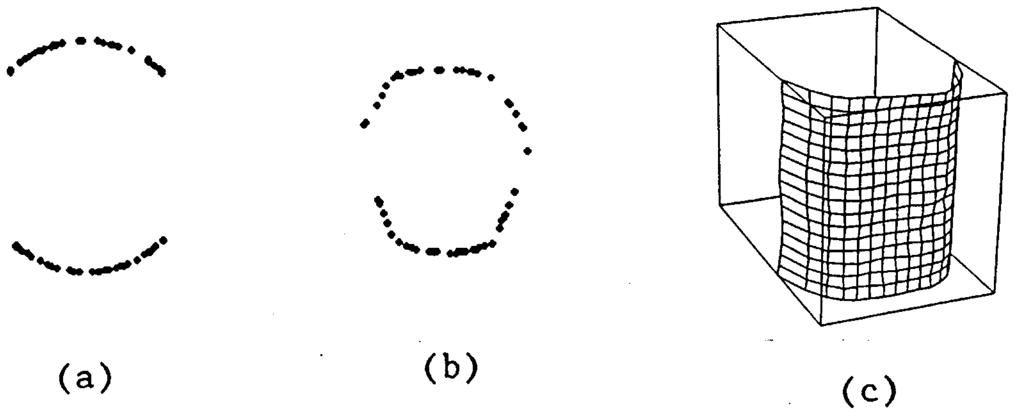

Figure 10. The influence of constraints on surface shape from object boundaries. The image of points on the surface of a vertical rotating cylinder is truncated along the left and right borders. (a) Birds' eye view of the 3–D structure obtained when the structure–from–motion algorithm is applied to all of the moving points together. The solution is essentially identical to the true structure of the points. (b) The 3–D structure obtained when systematic error is added to the velocities of the moving points at the borders of the display. (c) 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.

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 structure–from–motion algorithm on its own, applied to all of the points together, yields the correct 3–D structure in this case, as shown in Fig. 10a. This is similar to the percept derived from human observers when the truncated cylinder appears to be viewed through an aperture. 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. 10b). 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 along these borders, we also obtain a narrower, more curved object, similar to the result shown in Fig. 9c (a similar result is presented in Aloimonos and Huang (1991)). When a mask is placed around the truncated cylinder, giving rise to the perception of an aperture, these boundary constraints may not be imposed (see also, Thompson et al., 1991). Again, we conclude that a number of factors can lead to the perception of higher curvature in this display.

SUMMARY AND CONCLUSIONS

This paper has addressed the computational role that the construction of a complete surface representation may play in the recovery of 3–D structure from motion. We first discussed the need to integrate surface reconstruction with the structure–from–motion 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., 1992) further strengthen our hypothesis regarding the important interaction of these two processes. We then presented a model that combines a feature–based structure–from–motion recovery algorithm, temporal integration and a surface reconstruction process. 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 groups image features on the basis of their 2–D image motion for the purpose of segregating features onto multiple surfaces.

The results of our computer simulations suggest that the model presented here can account for a number of perceptual phenomena regarding the possible role of surface reconstruction in the recovery of 3–D structure from motion. We also showed that aspects of the motion measurement and structure–from–motion recovery algorithms can contribute to some of these perceptual phenomena, in addition to the use of a surface reconstruction process. 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, we showed that 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., 1992).

Our work also 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 structure–from–

motion 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 somewhat 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. A second question 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 stucture–from–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. 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.

Acknowledgements: We thank Eric Grimson and Shimon Ullman for useful comments on an earlier draft of this manuscript.

REFERENCES

Adelson, E. H. (1985). Rigid objects that appear highly non–rigid. *Invest. Ophthal. Visual Sci., Supplement, 26*, 56.

Aggarwal, J. K. & Martin, W., eds. (1988). *Motion Understanding.* Hingham, MA: Kluwer Academic Pubs.

Aloimonos, J. & Huang, L. (1991). Motion–boundary illusions and their regularization. *Proc. IEEE Workshop on Visual Motion*, Princeton, NJ, October, in press.

Andersen, G. (1989). Perception of three–dimensional structure from optic flow without locally smooth velocity. *J. Exp. Psych.: Human Perc. Perf.*, *15*, 363–371.

Andersen, R. A. & Siegel, R. M. (1990). Motion processing in the primate cortex. In *Signal and Sense: Local and Global Order in Perceptual Maps*, eds. G. M. Edelman, W. L. Gall, W. M. Cowan, New York: Wiley & Sons, 163–184.

Anderson, B. D. O. & Moore, J. B. (1979). *Optimal Filtering.* Englewood Cliffs, NJ: Prentice–Hall.

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

Barron, J. (1984). A survey of approaches for determining optic flow, environmental layout and egomotion. *Univsity 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*, Charleston, SC, 73–80.

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

Bolle, R. M. & Vemuri, B. C. (1991). On three–dimensional surface reconstruction methods. *IEEE Trans. Patt. Anal. Machine Intell., 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. *Percept. Psychophys.*, *13*, 169–179.

Boult, T. E. & Kender, J. R. (1986). Visual surface reconstruction using sparse depth data. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, June, 68–76.

Braunstein, M. L. (1962). The perception of depth through motion. *Psychol. 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. *J. Exp. Psych.: Human Percept. Perf.*, *10*, 749–760.

Braunstein, M. L., Hoffman, D. D. & Pollick, F. E. (1990). Discriminating rigid from nonrigid motion: Minimum points and views. *Percept. Psychophys.*, *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. *J. Exp. Psych.: Human Percept. Perf.*, *13*, 335–343.

Bruss, A. & Horn, B. K. P. (1983). Passive navigation. *Comp. Vision, Graph. Image Proc.*, *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. *J. Exp. Psych.: Human Percept. Perf.*, *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 Res.*, *29*, 1789–1813.

Dosher, B. A., Landy, M. S. & Sperling, G. (1989b). Ratings of kinetic depth in multidot displays. *J. Exp. Psych.: Human Percept. Perf.*, *15*, 816–825.

Faugeras, O. D., LeBras–Mehlman, E. & Boissonat, J. D. (1990). Representing stereo data with the Delaunay triangulation. *Artif. Intell.*, *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 Trans. Patt. Anal. Machine Intell.*, *13(5)*, 401–412.

Gelb, A. (ed.) (1974). *Applied Optimal Estimation*, Cambridge: MIT Press.

Green, B. F. (1961). Figure coherence in the kinetic depth effect. *J. Exp. Psych.*, *62*, 272–282.

Grimson, W. E. L. (1981). *From Images To Surfaces: A Computational Study Of The Human Early Visual System*, Cambridge, MA: MIT Press.

Grimson, W. E. L. (1982). A computational theory of visual surface interpolation. *Phil. Trans. Royal Soc. London, Series B*, *298*, 395–427.

Grimson, W. E. L. (1983a). An implementation of a computational theory of visual surface interpolation, *Comp. Vision, Graph. Image Proc.* *22*, 39–69.

Grimson, W. E. L. (1983b). Surface consistency constraints in vision, *Comp. Vision, Graph. Image Proc.*, *24*, 28–51.

Grzywacz, N. M. & Hildreth, E. C. (1987). The incremental rigidity scheme for recovering structure from motion: position vs. velocity based formulations. *J. Opt. Soc. Amer. A*, *4*, 503–518.

Heel, J. (1990a). Dynamical motion vision. *Robotics and Autonomous Systems*, 6(3), 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, *Can. J. Physiol. Pharmacol.*, *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. *Percept. Psychophys.*, *48*, 19–36.

Hoffman, D. D. (1982). Inferring local surface orientation from motion fields. *J. Opt. Soc. Amer.*, *72*, 888–892.

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

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

Ikeuchi, K. & Horn, B. K. P. (1981). Numerical shape from shading and occluding boundaries. *Artif. Intell.*, *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. *Percept. Psychophys.*, *14*, 201–211.

Johansson, G. (1978). Visual event perception. In *Handbook of Sensory Physiology*, eds. R. Held, H. W. Leibowitz, H.–L. Teuber, Berlin: Springer–Verlag.

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

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

Landy, M. S. (1987). A parallel model of the kinetic depth effect using local computations. *J. Opt. Soc. Amer. A*, *4*, 864–877.

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 Res.*, *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. *Proc. Royal Soc. London Series B*, *208*, 385–397.

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

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

Luenberger, (1973). *Introduction to Linear and Nonlinear Programming.* Reading, MA: Addison–Wesley.

Marroquin, J, Mitter, S. & Poggio, T. (1987). Probabilistic solution of ill–posed problems in computational vision. *J. Amer. Statist. Assoc.*, *82*, 76–89.

Matthies, L. H., Szeliski, R. & Kanade, T. (1989). Kalman filter–based algorithms for estimating depth from image sequences. *Int. J. Comp. Vision*, 3, 209–236.

McKee, S. P. & Welch, L. (1985). Sequential recruitment in the discrimination of velocity. *J. Opt. Soc. Amer. 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(1)*, 55–68.

Negahdaripour, S. & Horn, B. K. P. (1987). Direct passive navigation. *IEEE Trans. Patt. Anal. Machine Intell.*, *PAMI–9*, 168–176.

Petersik, J. T. (1979). Three–dimensional object constancy: coherence of a simulated rotating sphere in noise. *Percept. Psychophys.*, *25*, 328–335.

Petersik, J. T. (1987). Recovery of structure from motion: Implications for a performance theory based on the structure–from–motion theorem. *Percept. Psychophys.*, *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. *Percept. Psychophys.* *44*, 390–393.

Rieger, J. H. & Lawton, D. T. (1985). Processing differential image motion. *J. Opt. Soc. Amer. 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 *Approximation Theory II*, eds. G. G. Lorentz, C. K. Chui, L. L. Schumaker, New York: Academic, 203–267.

Schwartz, B. J. & Sperling, G. (1983). Nonrigid 3–D percepts from 2–D representations of rigid objects. *Invest. Ophthal. Visual Sci.*, *24*, (3, Supplement), 239.

Shariat, H. & Price, K. E. (1990). Motion estimation with more than two frames. *IEEE Trans. Patt. Anal. Machine Intell.*, *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(9)*, 2768–2785.

Sperling, G., Landy, M. S., Dosher, B. A. & Perkins, M. E. (1989). Kinetic depth effect and identification of shape. *J. Exp. Psych.: Human Percept. Perf.*, *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: *From Pixels to Predicates: Recent Advances in Computational and Robotic Vision*, A. Pentland (ed.), Norwood, NJ: Ablex.

Terzopoulos, D. (1988). The computation of visible–surface representations. *IEEE Trans. Patt. Anal. Machine Intell., 10*, 417–438.

Thompson, W. B., Kersten, D. & Knecht, W. R. (1991). Structure–from–motion based on information at surface boundaries. *Submitted for publication.*

Thompson, W. B., Mutch, K. M. & Berzins, V. (1985). Dynamic occlusion analysis in optical flow fields. *IEEE Trans. Patt. Anal. Machine Intell.*, *PAMI–7*, 374–383.

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

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

Todd, J. T. (1985). The perception of structure from motion: Is projective correspondence of moving elements a necessary condition? *J. Exp. Psych.: Human Percept. Perf.*, *11*, 689–710.

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. *Percept. Psychophys.*, *43*, 179–188.

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

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

Treue, S., Andersen, R. A., Ando, H. & Hildreth, E. C. (1992). Structure from motion: Perceptual evidence for surface interpolation. *Manuscript in preparation.*

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: MIT Press.

Ullman, S. (1983). Computational studies in the interpretation of structure and motion: summary and extension. In *Human and Machine Vision*, ed. J. Beck, B. Hope, A. Rosenfeld. 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 Comp.*, *2*, 420–435.

Wallach, H. & O'Connell, D. N. (1953). The kinetic depth effect. *J. Exp. Psych.*, *45*, 205–217.

Wallach, H., Weisz, A. & Adams, P. A. (1956). Circles and derived figures in rotation. *Amer. J. Psych.*, *69*, 48–59.

Waxman, A. M. & Wohn, K. (1988). Image flow theory: A framework for 3–D inference from time–varying imagery. In: *Advances in Computer Vision*, vol. 1, ed. C. Brown, Hillsdale, NJ: Lawrence Erlbaum, 165–224.

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

Williams, D. & Philips, G. (1987). Cooperative phenomena in the perception of motion direction. *J. Opt. Soc. Amer. A*, *4*, 878–885.

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

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

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
December 1991

3. REPORT TYPE AND DATES COVERED
AIM 1314

4. TITLE AND SUBTITLE
Recovering Three-Dimensional Structure from Motion with Surface Reconstruction

5. FUNDING NUMBERS
N00014-85-K-0124
IRI-8719394
IRI-8657824
EY 07492

6. AUTHOR(S)
Ellen C. Hildreth, Horishi Ando, Richard Andersen, and Stefan Treue

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Artificial Intelligence Laboratory
545 Technology Square
Cambridge, Massachusetts 02139

8. PERFORMING ORGANIZATION REPORT NUMBER
AIM 1314
C.B.I.P. 70

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Information Systems
Arlington, Virginia 22217

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AD-A271844

11. SUPPLEMENTARY NOTES
None

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution of this document is unlimited

12b. DISTRIBUTION CODE
UNLIMITED

13. ABSTRACT *(Maximum 200 words)*

ABSTRACT: 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 groups image features on the basis of their 2–D image motion to segregate features onto multiple surfaces. We present the results of computer simulations that relate the 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.

14. SUBJECT TERMS (key words)
3D motion
surface perception
motion analysis
structure from motion
surface reconstruction

15. NUMBER OF PAGES
48

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
UNCLASSIFIED

CS-TR Scanning Project
Document Control Form

Date : 09/08/94

Report # 1314

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
□ Laboratory for Computer Science (LCS)

Document Type:

□ Technical Report (TR) ☒ Technical Memo (TM)
□ Other:

Document Information

Number of pages: 48
Not to include DOD forms, printer intstructions, etc... original pages only.

☒ Single-sided or □ Double-sided

Print type:

□ Typewriter □ Offset Press ☒ Laser Print
□ InkJet Printer □ Unknown □ Other:

Check each if included with document:

☒ DOD Form □ Funding Agent Form □ Cover Page
□ Spine □ Printers Notes □ Photo negatives
□ Other:

Page Data:

Blank Pages(by page number):

Photographs/Tonal Material (by page number):

Other (note description/page number):

Description : Page Number:
MARKS FROM XEROXING ACROSS CENTER OF SOME PAGES.

Scanning Agent Signoff:
Date Received: 09/08/94 Date Scanned: 09/14/94 Date Returned: 09/15/93

Scanning Agent Signature: Michael W. Cook