MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 706 March, 1983

Computational Studies in the Interpretation of Structure and Motion: Summary and Extension

Shimon Ullman

Abstract. Computational studies of the interpretation of structure from motion examine the conditions under which three-dimensional structure can be recovered from motion in the image. The first part of this paper summarizes the main results obtained to date in these studies. The second part examines two issues: the robustness of the 3-D interpretation of perspective velocity fields, and the 3-D information contained in orthographic velocity fields. The two are related because, under local analysis, limitations on the interpretation of orthographic velocity fields also apply to perspective projection. The following results are established:

- When the interpretation is applied locally, the 3-D interpretation of the perspective velocity field is unstable.
- The orthographic velocity field determines the structure of the inducing object exactly up to a depth-scaling.
- For planar objects, the orthographic velocity field always admits two distinct solutions up to depth-scaling.
- The 3-D structure is determined uniquely by a "view and a half" of the orthographic velocity field.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under the Office of Naval Research contract N00014-80-C-0505 and in part by the National Science Foundation Grant 79-23110MCS.

1 Introduction

When objects move in the environment, the images they cast upon our retinas undergo complex transformations. The human visual system can interpret these transformations to recover the three-dimensional (3-D) structure of the viewed objects and their motion in space.

This capacity to interpret structure from motion has been demonstrated in a number of studies[1,2]. Its earliest systematic investigation was carried out by Wallach and OConnell[3] in the study of what they have termed "the kinetic depth effect". In their experiments, an unfamiliar object was rotated behind a translucent screen, and the shadow cast on the screen by a distant light source was observed from the other side of the screne. In most cases, the viewers were able to describe correctly the hidden object and its motion, even when each static shadow projection of the object was unrecognizable, and contained no three-dimensional information.

The original kinetic depth experiments employed primarily wireframe objects which projected as sets of connected lines. Later studies[4,5,2] established that 3-D structure can be perceived from displays consisting of unconnected elements in motion, and under both continuous and apparent motion conditions.

Additional demonstrations of motion-based interpretation were provided by the remarkable experiments of Johansson[6,7]. These demonstrations were created by filming human actors moving in the dark with small light sources attached to their main joints. Each actor was thus represented by up to 13 moving light dots. The resulting dynamic dot patterns created a vivid, three-dimensional impression of the actors and their motion.

In this memo, I shall review the main results obtained to date in the computational study of the interpretation of structure from motion. These studies examine the problem from a theoretical standpoint in an attempt to attain two main goals. The first is what may be called the underlying computational theory of the task[8,9]. This theory tries to explain how the interpretation can be achieved in principle by any biological visual system or by a man-made device. The second computational goal is to develop and compare different schemes that can actually recover 3-D structure from motion. The study of these schemes, their relative merits and shortcomings, and the comparison of their performance with that of humans, can lead to a better understanding of the interpretation scheme embodied in the human visual system.

In the next section, I shall analyze the computational problem, and describe a scheme for recovering structure from motion. Only a brief outline will be given, since a detailed description can be found elsewhere.[10] Section 3 will review recent alternative schemes and additional computational results obtained to date. Finally, Section 4 will make some comparisons among the main different schemes. In particular, the recovery of structure from continuous velocity fields under orthographic and perspective projections will be examined.

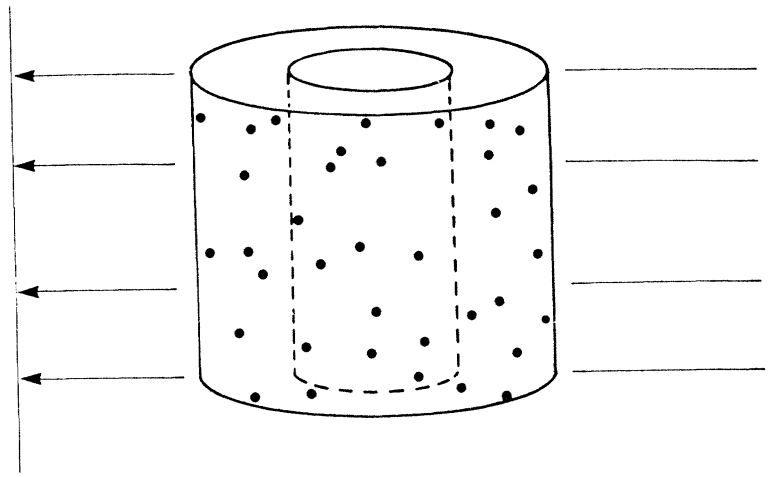

Figure 1 The interpretation of structure from motion. The dots comprising the two cylinders are projected on the screen (the outline of the cylinders in not shown in the actual presentation). The 3-D structure of the cylinders can be recovered from the motion of the dots across the screen (see Ullman, 1979).

2 The rigidity-based interpretation of structure from motion

In this section, we shall assume that the motion is given as a sequence of discrete frames, each one depicting a collection of unconnected elements. Figure 1 shows an example in which the elements are lying on the surface of two coaxial, invisible cylinders. The 3-D coordinates of all the dots are stored in a computer's memory, and their projection on the frontal plane is computed and presented on a CRT screen. The imaginary cylinders are then rotated (up to about 10 degrees between frames), and their new projection is computed and displayed on the screen. Each single static view of the cylinders appears as a random collection of dots. However, when the changing projection is viewed in a movie-like fashion, the elements in motion across the screen are perceived as two cylinders whose shapes and angles of rotation are easily determined.

How can this interpretation be achieved? The fundamental underlying problem is the ambiguity of the interpretation: there are many different motion patterns in space that could produce the same two-dimensional motion of the elements on the screen. To resolve this ambiguity, the interpretation scheme must incorporate some additional constraints that would rule out most of the possible interpretations and force a unique solution, which in most cases is also the correct one.

A possible constraint, suggested originally by Wallach and O'Connell[3] is a rigidity constraint. That is, the preferred interpretation is the one in which the elements move together as a rigid object rather than a collection of elements moving

independently in space. The suggested rigidity constraint raises, however, a number of problems. The first is the question of uniqueness: if the same 2D transformation of the elements can be produced by different 3-D objects, participating in different 3-D movements, then rigidity must be rendered an insufficient constraint for the 3-D interpretation task. A second problem is the possibility of false targets: if the elements participate in fact in a non-rigid motion in space, and if their 2D projoction happens to have a rigid interpretation, then this false rigid solution will be forced upon the elements. Finally, there is the multiple object problem: if the observed elements belong not to a single rigid object, but to distinct objects moving independently, then the collection as a whole will fail to have a rigid interpretation. A rigidity-based interpretation must, therefore, be applied somehow to relevant subcollections of the elements, and not to the entire scene at once.

A detailed analysis of these problems can be found elsewhere[2,10], together with a discussion of possible implications to human motion perception. Here, I shall only sketch briefly the main computational results. It has been shown by Ullman and Fremlin (in the "structure from motion" theorem[2]), that, given three distinct views of a moving object, it can be determined unambiguously whether they represent a single rigid object, and if they do, then the 3-D structure can be recovered uniquely. The object is defined here as a collection of identifiable points, and to obtain uniqueness, the object must contain at least four non-coplanar points.

Rigid objects in motion are thus determined uniquely on the basis of information that is local in both space and time. This result provides answers to the problems raised above. Uniqueness of the solution is guaranteed, provided that the moving object contains at least four non-coplanar points. The possibility of false targets is eliminated: it can be shown that the probability of four points that, in fact, do not belong to a single rigid object will happen to have a rigid interpretation is negligible. Finally, the locality of the interpretation suggests a solution to the multiple object problem. If the interpretation is restricted to local groups of about four elements each, then, due to the contiguity of objects, many of these groups will lie within a single object, and therefore their interpretation will not be affected by the additional objects. Given a scene containing several rigid objects in motion, a correct interpretation can therefore be obtained using the following scheme: divide the image into local groups of about four elements each, test each group for a unique rigid interpretation, and combine the results obtained for the different groups.

3 Alternative schemes and additional results

The analysis outlined in the previous section has been formulated in terms of distinct views, distinct identifiable points, and a parallel projection of the moving objects. In this section, I shall examine similar results obtained under somewhat different formulations.

3.1 Perspective vs. parallel projection

The structure-from-motion theorem mentioned above assumed parallel or orthographic projection. Unlike perspective projection, orthographic projection is formed by parallel light rays that are perpendicular to the image plane. Under such projection, the interpretation is unique up to a possible reflection about the image plane. This is an inherent ambiguity, since the projections of a rotating object, and its mirror image rotating about the same axis in the opposite direction, coincide under parallel projection. It is not surprising, therefore, that under orthographic projection, and in the lack of any additional source of 3-D information, human observers experience spontaneous depth-reversals of the objects, accompanied by reversal in the observed direction of rotation.

The natural projection of 3-D objects onto the retina is a perspective rather than parallel projection (i.e., the projection rays are not parallel, but meet at a common point). Although it has been demonstrated (e.g., in the original kinetic depth experiments) that structure can be perceptually recovered from the parallel projection of moving objects, it is of interest to analyze the rigidity-based interpretation under perspective projection. Experiments with computer algorithms (some of which are described elsewhere[2]) have suggested that seven or eight points in two perspective views are sufficient for a unique interpretation. Six points were sometimes sufficient for a unique interpretation, but not always. Longuet-Higgins[11] has proposed an elegant algorithm for recovering 3-D structure from two perspective views of eight points. The computation required is particularly simple, involving primarily the solution of eight linear equations. This analysis did not provide, however, a uniqueness proof.

Recently, Tsai and Huang[12] provided a comprehensive analysis of the uniqueness problem under two perspective views. Their results established that seven points guarantee a unique interpretation, provided that (i) they do not lie on a pair of planes, one of which passes through the origin, and (ii) they do not lie on a single cone containing the origin. This means that, except for a few cases where the points happen to form some special configurations in space, the interpretation will be unique. Tsai and Huang provided, in addition, simple algorithms for the recovery of the motion and structure parameters.

3.2 On the meaning of "Computational Experiments"

Those who associate "experiments" primarily with the testing of human subjects may wonder what is meant here by experiments with computer algorithms. The answer lies in the fact that it is often easier to find a solution to a problem (at least an approximate one) than to prove its existence and uniqueness. Suppose, for example, that it is conjectured that the three-dimensional structure of an object can be recovered from its changing projection by solving a certain system of linear equations. The coefficients in these equations will be variables that assume different values for different objects in motion. It may be difficult to show that the system has, in general, one solution. It is a straightforward procedure, however, to solve the equations for particular examples and verify the solution and its uniqueness.

If we test the solution for a variety of examples, and consistently recover the correct solution, we have some reason to believe that the interpretation scheme is, in general, correct. It is still possible, however, that under certain conditions, the interpretation scheme will fail. Such conditions may be difficult to discover, and this is one reason why a comprehensive analytic analysis is more satisfaactory than computational experiments.

3.3 The use of velocity information

So far, the changing projection of the moving objects has been described in terms of discrete movie-like frames. It should be noted, however, that the formulation of the structure-from-motion theorem in terms of discrete views does not imply that the input image must be discrete rather than continuous. If a continuous motion extends long enough to contain three distinct views (and the qualification for "distinct" will depend on the accuracy of the imaging system), then it contains sufficient information for a unique interpretation. The theorem states this fact without excluding the possibility of implementing the computation using a continuous scheme.

A distinctive property of the scheme outlined in Section 2 is that the information used was expressed entirely in terms of the positions of the elements at different times. An alternative formulation[13,14] uses the velocities of the points as well as their positions.

The input to the computation consists then of a single perspective view in which the positions of the moving elements and their velocities are specified. This velocity-based formulation can be viewed as the limiting case of two frames, as the time interval between them approaches zero. The uniqueness problem then takes the following form: given the position and velocity of N points in the image, determine whether or not they belong to a single moving object, and find the 3-D structure of the object and its motion in space.

A preliminary theoretical problem is to determine the number N for which this recovery problem has a unique solution. Mathematically, this problem is still unresolved. A counting argument of equations and unknowns shows that at least five points would be necessary. A computer algorithm implemented by Prazdny[14] suggests that five points might also be sufficient. Since the computer algorithm proved sensitive to errors in the input, especially when the viewed object was small, it seems that a robust recovery algorithm would require more than five points. This problem of robustness is examined again in more detail in Section 4.3.

3.4 The use of a continuous velocity field

The scheme outline earlier relied on a small number of discrete elements. An alternative mathematical approach is to assume that the velocity of points in the image is known everywhere within a given region. The information is sometimes assumed to be known only locally; for example, the velocity field and its spatial

derivatives are known at a single point. This formulation can be thought of as a limiting case of the discrete formulation, as the distance between points approaches zero.

The most complete analysis to date of this problem has been provided by Longuet-Higgins and Prazdny[15]. Their analysis has established that the velocity field at a point has at most three different interpretations. More precisely, it showed that for nonplanar surfaces, given the velocity field and its first and second spatial derivatives at a point, there are at most three solutions to the surface orientation at that point. The analysis also provided a scheme for computing the solutions. The possible improvement of this result—in particular, a determination of whether the solution is in fact unique—poses an open question for future research.

In an analysis of the orthographic continuous field, Hoffman[16] has shown that 3-D structure can be recovered uniquely (up to the unavoidable reflection about the image plane) if both the velocity and the acceleration fields are known within a region.

3.5 Restricted motion and the interpretation of biological motion

The discussion so far has examined the general case of unrestricted motion. Additional results have been obtained for situations where certain limitations are imposed upon the motion of the viewed objects. In this section, I shall review the main results, together with their implication to problems in visual perception.

Results of considerable interest were obtained recently[17], which have shown that for a rigid rod moving in a plane, its length and orientation in space are determined uniquely by three views (under parallel projection). A similar result was obtained for two rigid rods hinged together end-to-tail. When such a pairwise-rigid configuration moves in a plane, its 3-D structure and motion are uniquely determined by only two parallel projections.

These results offer a powerful tool for the interpretation of biological motion (the kind of interpretation demonstrated by Johansson's experiments). Using 3-D measurements of motion in space, Hoffman has established that the arm and leg motion during locomotion often conforms to the planar motion constraint. Hoffman and Flinchbaugh[17] found that when the planarity-based scheme is applied to data obtained from Johannson-like experiments, a correct interpretation of the 3-D structure and motion of the moving light dots is often obtained. Their results suggest that the human visual system may incorporate processes that are capable of detecting pairs or triplets of feature points engaged in planar motion, and apply to them the planarity-based interpretation scheme. Figure 2 illustrates the application of the planarity-based interpretation scheme to six dots representing the human body in motion. The unconnected dots are shown in 2*a* and the rigid connection in 2*b*. These connections and their 3-D structure can be established by the planarity-based scheme for all the rigid links that obey the planarity constraint.

A similar scheme that can cope with somewhat less restricted motion was developed

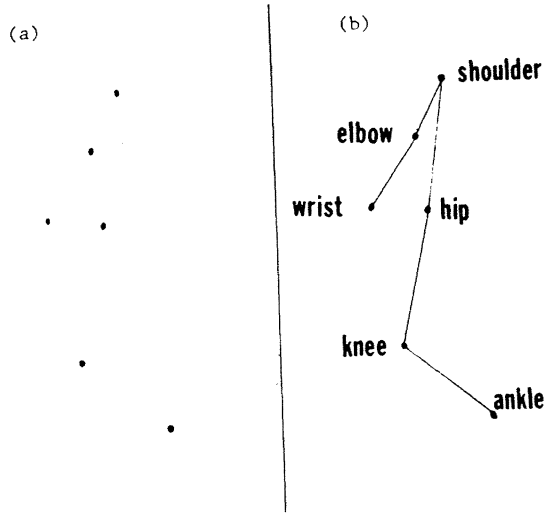

Figure 2 The interpretation of biological motion. (a) Six unconnected dots representing parts of the human body in motion. The motion of these dots can give rise to a vivid perception of a moving person. (b) Links made along the rigid connections. These links and their 3-D structure can be recovered by the planarity-based interpretation scheme (from Hoffman & Flinchbaugh, 1982).

by Webb and Aggarwal.[18] They have considered the interpretation problem for an object assumed to rotate continuously about a fixed axis. (In general, the axis of rotation may change with time.) They have applied this scheme successfully to instances of biological motion, but no mathematical results regarding the information required for a unique interpretation have yet been obtained.

3.6 Vertical rotation and horizontal translation: The computation of depth from stereoscopic disparity

A recent result by Longuet-Higgins[19] established uniqueness for the case of an object that rotates about the vertical axis and translate in the horizontal plane. When the motion is restricted in this manner, three points are, in general, sufficient for a unique interpretation. The significance of this result lies in its possible applicability to the computation of depth from stereoscopic disparity. In the stereoscopic case, the object is in fact, fixed, and the different views are obtained by the different viewing positions of the two eyes. This problem is formally equivalent to the interpretation of structure from motion, given only two views. The computation of depth from stereoscopic disparities requires, in principle, knowledge of the direction of gaze of the two eyes. Longuet-Higgins' result suggests that this information can be obtained without reliance on non-visual information. Assuming that the horizontal meridians

of the two eyes coincide accurately, three points (non-meridional, and not all lying in the vertical plane) are sufficient for the recovery of the two directions of gaze.

The main results regarding the unique interpretation of structure-from-motion discussed so far are summarized in Table 1 for general (upper half) and restricted (lower half) motion.

TABLE 1

MAIN RESULTS OF STUDIES ON THE RECOVERY OF STRUCTURE FROM MOTION

Unrestricted Motion		
Discrete Points & Views	Discrete Points, Velocities	Velocity Field & its Derivatives
4 points in 3 orthographic views (Ullman & Fremlin, 1979)	5 points and their velocities in a single perspective view (Prazdny, 1980)	Up to 3 solutions for general motion (Longuet-Higgins & Prazdny, 1980)
7 points in 2 perspective views (Tsai & Huang, 1981)		Unique solution from velocity and acceleration under orthographic projection (Hoffman, 1980)
Application: The recovery of 3-D structure from unrestricted motion.		
Restricted Motion		
3 orthographic views of two points in planar motion 2 orthographic views of 3 points in a "hinged" configuration, planar motion (Hoffman & Flinchbaugh, 1982)		
Application: Biological motion		
3 non-meridional points, vertical axis and horizontal translation (Longuet-Higgins, 1983)		
Application: The recovery of depth from stereo disparities		

Table 1: Uniqueness of the interpretation of structure from motion. The main results obtained to date are summarized for general motion (upper half) and restricted motion (lower half).

4 Remarks concerning the different schemes and their relevance to perception

The previous section examined the interpretation problem in a number of different formulations. Since a primary goal of these studies is to provide a computational basis for the study of perceptual phenomena, it is worthwhile to compare the relevance of the different schemes to human perception. Three questions will be examined briefly in this section. The first has to do with the applicability of mathematical results and algorithms to biological visual systems; the second has

to do with the use of positions versus velocity fields; the third with parallel and perspective projections. On the first two of these questions, I shall limit myself to a brief discussion of a few selected points.

4.1 Mathematical algorithms and biological visual systems

The results outlined in the preceding two sections were formulated in terms of mathematical propositions and algorithms. Two difficulties are sometimes raised regarding the applicability of such results to biological visual systems. The first is that, unlike an electronic computer, a biological system cannot be expected to solve the equations used in deriving the mathematical results. The second is that a biological system does not have access to the perfectly accurate data used in the mathematical abstraction.

A comprehensive examination of the first objection would be beyond the immediate goals of this review. The main answer lies, however, in the distinction between different levels of analysis: competence versus performance[20] or computational vs. algorithmic.[8]

The computational studies aim primarily at establishing principles such as rigidity or planarity that apply to any visual system facing the problem of interpreting structure from motion. Certain equations may be used in the derivation of such principles, but it does not follow, of course, that a system utilizing these principles would have to solve these equations in the process of interpreting structure from motion.

The problem of accuracy in the measurements and computation is an important one. To be of practical value, the interpretation scheme must be robust: small errors in the measurement of position and velocity, for example, should not lead to a complete breakdown of the interpretation scheme. This means that computational studies should not only explore what is possible under idealized conditions, but also examine the effects of small perturbations and errors. An example of such an analysis will be given below.

4.2 The use of positions vs. velocity fields

All of the structure-from-motion schemes examined about used certain measurements as their "inputs", and recovered the 3-D structure and motion in space as "output". Different schemes used different inputs; some used the positions (in the image) of the moving elements at different times, while others used their retinal velocities.

As noted earlier, the difference between the two formulations is not that velocity-based schemes are continuous and position-based ones discrete. A position-based computation, for example, can use the continually changing positions, without necessarily requiring discrete "snapshots", but also without using velocity measurements.

It is not clear at present which formulation, the position-based or the velocity-based, is more directly relevant to human motion perception, since the measurements employed by the human visual system are not fully known. It may prove valuable, therefore, to explore in more detail different schemes that are based on different types of inputs. The comparison of such schemes with the performance of the human visual system could provide some clues regarding the types of measurements employed by the visual system in the interpretation of structure from motion.

4.3 Parallel and perspective velocity fields

The projection of the external environment available to our eyes is perspective rather than parallel. What, then, is the relevance of the parallel projection studies? There are two answers to this question. The first is that humans can recover structure from motion under orthographic projection. The second answer lies in the fact that under local analysis, perspective and parallel projections are nearly identical. (Local analysis means here that the surface patch under analysis is restricted to a small part of the visual field, so that the dimensions of the patch are small compared to its overall distance from the viewer.) If the interpretation scheme is to be robust and insensitive to small errors, it must also be capable of coping with the minor differences between the two projections, and either projection can therefore be assumed. One can, in fact, use the two kinds of projection as a test for stability. If a given interpretation scheme can operate under perspective projoection but fails under orthographic projection, it cannot be a stable local interpretation method.

In the next section, I shall argue that (i) the recovery of 3-D structure from the instantaneous velocity field is impossible under orthographic projection, and (ii) for perspective projection, the recovery is unstable under local interpretation.

5 The orthographic velocity field.

In this section, we shall derive a complete characterization of the 3-D information that can be recovered from the instantaneous velocity field under orthographic projection.

5.1 The depth scaling proposition

Consider two surface patches such as S_1 and S_2 in Figure 3. The figure shows a cross-section of the surfaces from a side view. They are assumed to be rotationally symmetric with respect to rotations around the observer's line of sight so that S_1, for example, is part of the surface of a sphere. The observer is assumed to view the objects along the Y axis, which is his line of sight, or depth axis. X is the observer's horizontal axis, Z the vertical, and the $X - Z$ plane is called the image plane. The objects are assumed to be fixed at one point, which is taken as the origin of the coordinate system.

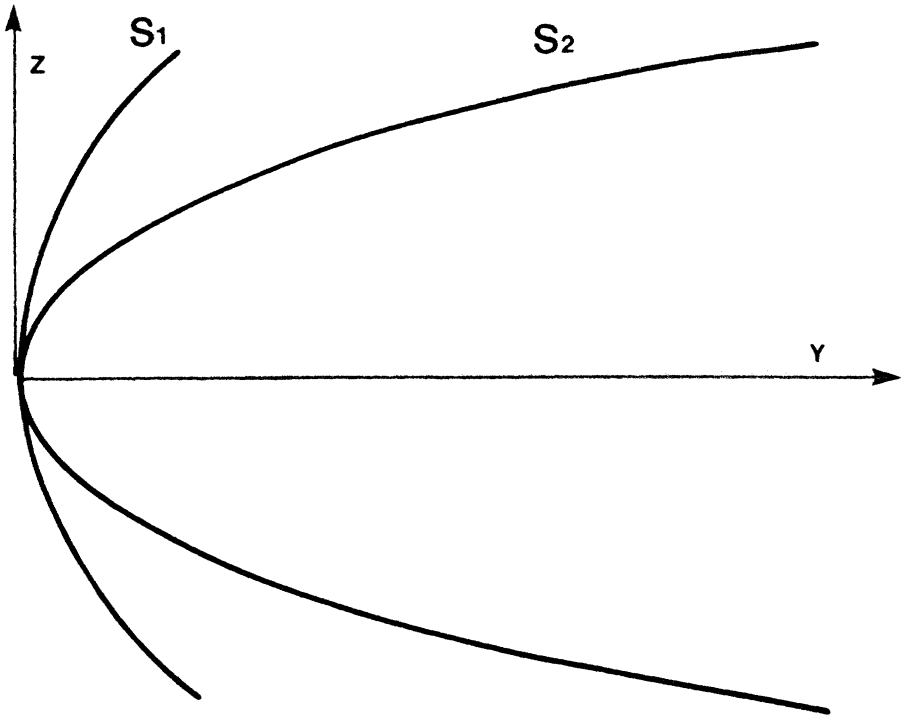

Figure 3 The ambiguity of the orthographic velocity field. The surface S_2 is obtained from S_1 via depth-scaling. Any motion of S_1 can be "mimicked" by S_2 in such a way that their velocity fields will coincide.

Let O_1 and O_2 denote the parallel projections of S_1 and S_2, respectively (on the $X - Y$ plane); P_1 and P_2 will denote the perspective projections of S_1 and S_2; O'_i and P'_i (for $i = 1, 2$) will denote the parallel and perspective velocity fields, respectively; $u(x, z)$ will denote the projected velocity in the horizontal direction and $v(x, z)$ in the vertical.

The surface S_2 in Figure 3 was obtained from S_1 by the following transformation: if (x, y, z) is a point on S_1, then (x, ky, z) is a point of S_2, for a fixed constant k ($k = 5$ in Figure 3). S_1 is assumed to rotate by some angular velocity ω_1 about the vertical axis, and S_2 rotates about the same axis with angular velocity $\omega_2 = \omega_1/k$. This transformation (including the rotations) will be called "depth-scaling", since the depth coordinate Y is scaled by a constant k.

Since the rotation is about the vertical Z axis, under orthographic projection, all the points will travel in the image plane parallel to the X axis. The projected velocity of a point on S_1 having spatial coordinates (x_1, y_1, z_1) will be $u_1 = -y_1\omega_1$. The corresponding point on S_2 (i.e., the point on S_2 whose projection coincides with that of (x_1, y_1, z_1)) has a projected velocity $u_2 = -y_2\omega_2$. Since by definition, $y_2 = ky_1, \omega_2 = \omega_1/k$, it follows that $u_1 = u_2$ for every image point, and therefore the velocity fields of S_1 and S_2 coincide.

In this example, the objects were assumed to rotate about the vertical Z axis, but S_2 can, in fact, "mimic" S_1 under arbitrary rotation. This claim can be established by noting that any rotation in space can be decomposed into two components: one is a rotation about an axis lying in the frontal $X - Z$ plane, the other is a rotation

about the line of sight which is perpendicular to the $X - Z$ plane. We shall call the first component XZ-rotation and the second Y —rotation. A general rotation of S_1 can be decomposed, therefore, into an XZ-rotation with angular velocity ω_1 and a Y-rotation with angular velocity θ_1. Let S_2 rotate by $\omega_2 = \omega_1/k$ about the same axis in the $X - Z$ plane, and assume $\theta_2 = \theta_1$. It is not difficult to see that with this choice of ω_2, the orthographic velocity fields of S_1 and S_2 will coincide.

Unlike the vertical rotation case, where the two objects had a common rotation axis, in this case (assuming $\omega_1 \neq 0, \theta_1 \neq 0$), the two objects would rotate about different spatial axes and still have identical velocity fields. We conclude that under orthographic projection, two objects can differ drastically in their shapes (e.g., in terms of their surface orientation and curvatures), axes of rotation and rotation speeds, and yet induce identical velocity fields.

Since the constant k used in the definition of S_2 can be chosen arbitrarily, the velocity fields admit not only two distinct interpretations, but an infinite family of surfaces for depth scaling by different factors k. (The definition of depth scaling includes the appropriate relation between the rotation components, $\omega_2 = \omega_1/k, \theta_2 = \theta_1$. It is also assumed that $\omega_1 \neq 0$.) It can be further shown that this family completely characterizes the set of confusable objects. We can summarize these claims in the following proposition:

- *The Depth Scaling Proposition.* If a non-planar surface S_1 is a possible rigid interpretation for a given orthographic velocity field, then S_2 is another possible interpretation if and only if it is obtained from S_1 via depth scaling.

The proof of this proposition is given in Appendix 1. It serves to give a complete characterization of all possible interpretations of the orthographic velocity field. Its first implication is that this interpretation is always non-unique. The second implication is that, although the interpretation is non-unique, properties that are invariant under depth-scaling can be recovered from the orthographic velocity field. For example, the depth ratio y_i/y_j for two points with image coordinates (x_i, z_i) and (x_j, z_j) can be recovered uniquely from the velocity field; extremal points in y and inflection points can also be recovered.

5.2 The orthographic velocity field of a planar object

The depth scaling proposition stated above holds for non-planar surfaces. For planar objects, the ambiguity is doubled: the orthographic velocity field admits exactly two distinct solutions, each determined up to a depth scaling. In addition, it would be possible to determine from the velocity field whether or not the inducing object is planar. These properties of the orthographic velocity fields of planar objects are established in Appendix 2.

5.3 Unique structure from a "view and a half"

As it turns out, the depth-scaling ambiguity of the solution can be resolved with the addition of a single view. More specifically, if the projected positions and velocities of five points in a general configuration (i.e., no four of which are coplanar) are

given at time t_1, and the positions of the same five points are given at a later time t_2, then the 3-D structure can be recovered uniquely (up to the unavoidable reflection ambiguity about the image plane). This "view-and-a-half" proposition is proven in Appendix 3.

In summary, under orthographic projection, the interpretation of the velocity field is non-unique. For non-planar objects, if S_1 is a possible interpretation, then S_2 is another possible interpretation, if and only if, it is related to S_1, via depth scaling. Only properties that are invariant under depth scaling can therefore be recovered from the orthographic velocity field. For planar objects, there are exactly two distinct interpretations up to depth scaling. The structure can be recovered uniquely for as few as five points in a general configuration if a single view is given in addition to the velocity field. A more detailed analysis, including the planar case, is given in the appendix.

6 The perspective velocity field

The perspective velocity field is, in a sense, richer in information than the orthographic field. While the orthographic field admits infinitely many interpretations, in the perspective case, the velocity field, even in an arbitrarily small neighborhood, can have no more than three interpretations.[15] For a sufficiently small patch of surface, perspective and orthographic projections are not very different, however, and one may suspect, therefore, that the local interpretation of perspective velocity fields in unstable. The argument is roughly as follows. Under orthographic projection, two widely different objects, S_1 and S_2, can have similar, and even identical velocity fields, O'_1 and O'_2, respectively. For a sufficiently large viewing distance, the perspective velocity fields, P'_1 and P'_2 become similar to O'_1 and O'_2, respectively. Consequently, P'_1 and P'_2 are also closely similar, and therefore, slight errors in the measured velocity fields can have large effects on the interpreted 3-D shape.

This argument can be made more precise. Let S_1 and S_2 in Figure 4 be two surfaces rotating about the Z axis. As before, the direction of gaze is along the Y axis. S_2 is derived from S_1 in this case by extending the ray from the viewing point to each point on S_1 so that $y_2/y_1 = k$ for a given constant k. The surface S_2 depends in this case on the viewing point: when the viewing distance increases, the shape of S_2 in the perspective case approaches the orthographic depth scaling of S_1 by the constant k. (In Figure 4, it was assumed that the ratio of viewing distance to object size is such that the object occupies one degree of visual angle.) As before, we will choose $\omega_2 = \omega_1/k$. Finally, let H denote the distance of the origin Q from the observer.

The projected velocity of p_1 (a point on S_1 with $y \neq 0$) is (u_1, u_1). This velocity will be measured as an angular velocity, i.e., the number of deg/sec a point travels in the observer's field of view. The angular velocity of the corresponding point on S_2 is (u_2, v_2). The claim is that when H is sufficiently large, the vectors (u_1, v_1) and (u_2, v_2) become arbitrarily close. That is, as H grows, the ratio between their

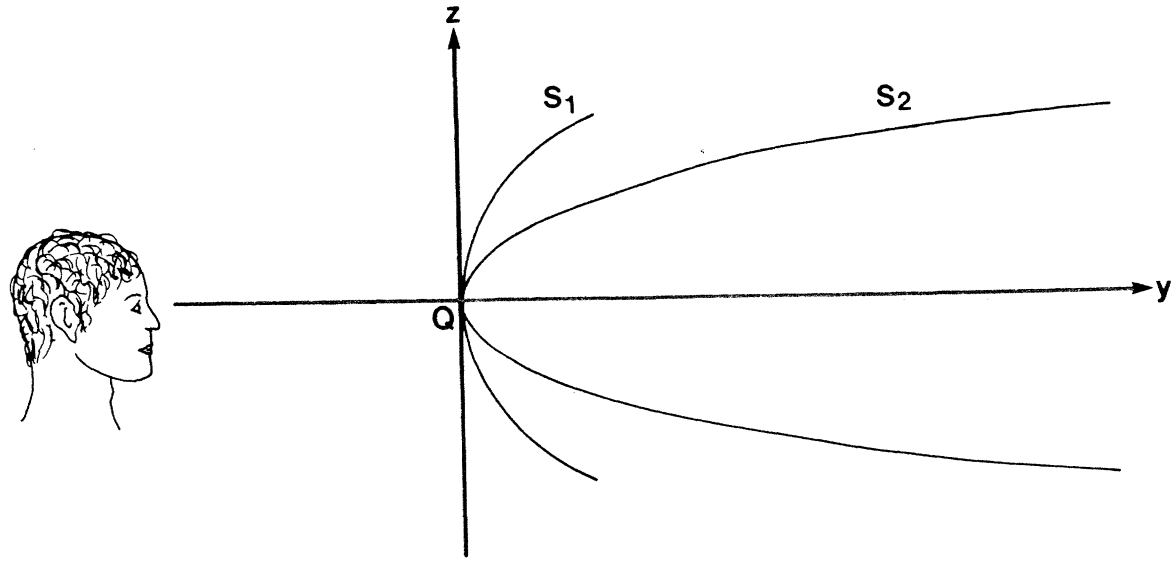

Figure 4 The instability of the local interpretation of perceptive velocity fields. Under local analysis, small errors in the measured velocity field can induce large errors in the interpreted 3-D structure. The difference in the velocity fields induced by S_1 and S_2 will be less than 3.2% when the surface patches occupy one degree of visual angle, and about 6% at twice this size.

magnitudes $\sqrt{u_1^2+v_1^2}/\sqrt{u_2^2+v_2^2}$ approaches 1, and their difference in direction, measured by $(v_1/u_1)-(v_2/u_2)$, approaches 0 (assuming $y \neq 0$ in the region, except at the origin).

The proof of this claim is straightforward, and will, therefore, not be detailed. It can be derived from the expressions given below for the speed ratio and the direction difference. These expressions follow directly from the definitions. The speed ratio (squared) is:

$$\frac{u_2^2+v_2^2}{u_1^2+v_1^2} = \frac{[(H+ky_1)y_1+x_2^2/k]^2+z_2^2x_2^2/k^2}{[(H+y_1)y_1+x_1^2]^2+z_1^2x_1^2} \quad \frac{(H+y_1)^4}{(H+ky_1)^4}$$

where

$$x_2 = x_1\frac{H+ky_1}{H+y_1} \quad z_2 = z_1\frac{H+ky_1}{H+y_1}.$$

The difference in direction is:

$$\frac{v_2}{u_2}-\frac{v_1}{u_1} = \frac{z_2x_2/k}{(H+ky_1)y_1+x_2^2/k} - \frac{z_1x_1}{(H+y_1)y_1+x_1^2}.$$

It can be seen from these expressions that for any point with $y_1 \neq 0$, the speed ratio approaches 1 and the difference in direction approaches 0 as H grows. Under certain conditions (which will not be elaborated), this will also happen uniformly within a region.

As a result, it is possible to construct drastically different objects whose velocity fields within a region are almost identical. That is, the differences in speed and direction at any given point within the region can be made arbitrarily small. As explained in the analysis of the orthographic case, this ambiguity is not restricted to rotation about the vertical axis, but can arise for arbitrary rotations as well.

The surfaces in Figure 4 illustrate that this problem can be quite severe. When the viewing distance is such that the surfaces in Figure 4*a* occupy one degree of visual angle, the differences in their perspective velocity fields within the entire one degree patch will not exceed 3.2%. At half the viewing distance, the maximum error about 6%.

Concluding Remarks

The comparisons discussed in the last sections can be combined with psychological experiments to gain further insight into the possible use of instantaneous velocity fields in the recovery of structure from motion by the human visual system. One can test, for example, whether shape parameters, such as surface orientation and curvature, can be recovered by the visual system for objects subtending one or two degrees of visual angle. Even a moderate success in this task would suggest that either our visual system measures velocities with high precision, or that the interpretation under these conditions does not rely on the instantaneous velocity field. The first of these alternatives does not seem attractive. Computational studies have indicated that the measurement of the velocity field is, in general, a difficult task,[21,22] and it is probably unrealistic to expect these measurements to reach the level of precision required for the interpretation task. The ability to interpret correctly the 3-D structure of small objects can therefore provide evidence against the use of the instantaneous velocity field in this task. One may expect instead to find in the visual system, the capacity to integrate information over time periods that allow sufficient excursion of the moving object, rather than to base the interpretation on instantaneous velocity measurements.

Acknowledgements: I thank E. Hildreth and E. Grimson for their help and comments. This work (with the exception of the "view and a half" proposition) appears in **Human & Machine Vision**, Beck & Rosenfeld, eds., Academic Press, in press.

References

1. Braunstein, M. L. **Depth Perception Through Motion**, New York: Academic Press, 1976.

2. Ullman, S. **The Interpretation of Visual Motion**, Cambridge, Ma: MIT Press, 1979.

3. Wallach, H., & O'Connell, D. N. "The kinetic depth effect". *J. Exp. Psych.*, **45**, 205-217, 1953.

4. Braunstein, M. L. "Depth perception in rotation dot patterns: effects of numerosity and perspective". *J. Exp. Psych.*, **64**, 415-420, 1962.

5. Green, B. F. "Figure coherence in the kinetic depth effect". *J. Exp. Psych.*, **62**, 272-282, 1961.

6. Johansson, G. "Visual perception of biological motion and a model for its analysis". *Percept. & Psycho.*, **14**, 201-211, 1973.

7. Johansson, G. "Visual motion perception". *Scientific American*, **232** (6), 76-88, 1975.

8. Marr, D. & Poggio, T. "From understanding computation to understanding neural circuitry". In: E. Poppel, et al (eds), **Neural Mechanisms in Visual Perception. Neuroscience Research Program Bulletin**, 15, 470-488, 1977.

9. Marr, D. **Vision**, San Francisco, Ca:W. H. Freeman, 1982.

10. Ullman, S. "The interpretation of structure from motion". *Proc. Roy. Soc. Lond., B*, **203**, 405-426, 1979.

11. Longuet-Higgins, H. C. "A computer algorithm for reconstructing a scene from two projections". *Nature*, **293**, 133-135, 1981.

12. Tsai, R. Y. & Huang, T. S. "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces". University of Illinois at Urbana-Champaign, Coordinated Science Laboratory Report R-921, 1981.

13. Clocksin, W. F. "Perception of surface slant and edge labels from optical flow: a computational approach". *Perception*, **9**, 253-269, 1980.

14. Prazdny, K. "Egomotion and relative depth map from optical flow". *Biol. Cyber.*, **36**, 87-102, 1980.

15. Longuet-Higgins, H. C. & Prazdny, K. "The interpretation of a moving retinal image". *Proc. Roy. Soc. Lond. B*, **208**, 385-397, 1980.

16. Hoffman, D. D. "Inferring shape from motion fields". MIT AI Memo 592, 1980.

17. Hoffman, D. D. & Flinchbaugh, B. E. "The interpretation of biological motion". *Biol. Cyber.*, **42**, 195-204, 1982.

18. Webb, J. A. & Aggarwal, J. K. "Visually interpreting the motion of objects in space". *Computer*, **14** (8), 40-46, 1981.

19. Longuet-Higgins, H. C. "The role of the vertical dimension in stereoscopic vision". *Perception*, in press.

20. Chomsky, N. **Aspects of the Theory of Syntax**, Cambridge, Mass:MIT Press, 1965.

21. Marr, D. & Ullman, S. "Directional selectivity and its use in early visual processing". *Proc. Roy. Soc. Lond. B*, **211**, 151-180, 1981.

22. Ullman, S. "Analysis of visual motion by biological and computer systems". *IEEE Computer*, **14 (8)**, 57-69, 1981.

APPENDICES

Appendix 1: The depth scaling proposition

This appendix uses the same notion as Section 4 above. As before, S_1 and S_2 are two rotating rigid surfaces. The rotation of S_1 can be decomposed into a component with angular velocity ω_1 (assumed to be non-zero) about an axis in the frontal $X - Z$ plane (XZ-rotation), and a second component (Y-rotation) with angular velocity θ_1 about the line of sight Y. The corresponding components for S_2 are ω_2 and θ_2. S_2 will be called a *depth scaling* of S_1 if:

- For every point (x, y, z) on S_1, (x, ky, z) is a point on S_2 for some constant $k \neq 0$.
- The rotations ω_1 and ω_2 are around the same axis and $\omega_2 = \omega_1/k$.
- $\theta_1 = \theta_2$.

The depth scaling proposition

- If a non-planar surface, S_1, is a possible rigid interpretation for a given orthographic velocity field, then S_2 is another possible interpretation if, and only if, it is obtained from S_1 via depth scaling.

Proof: We have seen in a previous sections that if S_2 is obtained from S_1 via depth scaling, then their orthographic velocity fields coincide. The converse statement that remains to be shown is that if S_1 and S_2 have identical velocity fields, then they are related via depth scaling.

This property clearly holds in the vertical rotation case where S_1 and S_2 both rotate about the Z axis. If the angular velocities of S_1 and S_2 about the vertical axis are ω_1 and ω_2 respectively, then the projected velocities at point (x, y) in the image are $-\omega_1 y_1$ and $-\omega_2 y_2$ respectively. This implies that for the velocity fields to coincide, the depth ratios y_2/y_1 at any given point in the image must equal ω_1/ω_2.

To show that the proposition holds under general rotation, we shall establish the following claim: if two rotating (non-planar) objects, S_1 and S_2, have the same orthographic velocity field, then $\theta_1 = \theta_2$ (where θ denotes, as before, the rotation component about the line of sight).

Let θ_1 be the Y-rotation of the first object. By rotating the velocity field induced by S_1 by $-\theta_1$ about the line of sight, the Y-rotation component of S_1 is cancelled, and the only component that remains is the XZ-rotation about some axis in the $X-Z$ plane. The resulting velocity field will have the property that all the velocity vectors will now be parallel to each other. We shall next show that $-\theta_1$ is the only Y-rotation which, when added to the velocity fields, results in a parallel velocity field. The implication will be that, given a velocity field, the Y-rotation component is uniquely determined, and hence, $\theta_1 = \theta_2$.

Without loss of generality, we can assume that in the resulting parallel velocity field, all the projected velocity vectors are in the direction of the X axis. This means that the image velocity component in the Z direction $v(x,z) = 0$ at every image point (x,z), and the velocity field $(u(x,z), v(x,z))$ can therefore be described as $(u(x,z), 0)$. Let us now add to this field a Y-rotation with some angular speed, θ. The combined velocity field will now be $(u(x,z) - z\theta, x\theta)$. This will again be a parallel velocity field only if either

$$u(x,z) - z\theta = 0 \tag{A1}$$

or

$$\frac{x\theta}{u(x,z) - z\theta} = c \tag{A2}$$

for some constant c. In either case, it can be readily verified that the velocity field $u(x,z)$ must have been induced by a planar surface. In the second case, for example, $u(x,z) = (\theta x/c) + \theta z$ and the original rotation of S_1 was assumed to be about the Z axis. For rotation about the Z axis with angular velocity ω, the velocity field is given by

$$u(x,z) = -\omega y(x,z). \tag{A3}$$

It follows that

$$y = -\frac{\theta}{\omega c}x - \frac{\theta}{\omega}z, \tag{A4}$$

which is the equation of a plane.

We conclude that for a non-planar object, starting from a parallel velocity field, a Y-rotation by any amount will destroy the parallelism of the field. It follows that, given the velocity field of S_1 (which is also the velocity field of S_2), there is one, and only one, Y-rotation, by $-\theta_1$, that will create a parallel velocity field. This rotation is defined by the velocity field itself, independent of the inducing object, and hence $\theta_1 = \theta_2$.

We can now add to both S_1 and S_2 a rotation component $-\theta_1$ that will cancel rotation about the line of sight. Their velocity fields will still concide, but now both

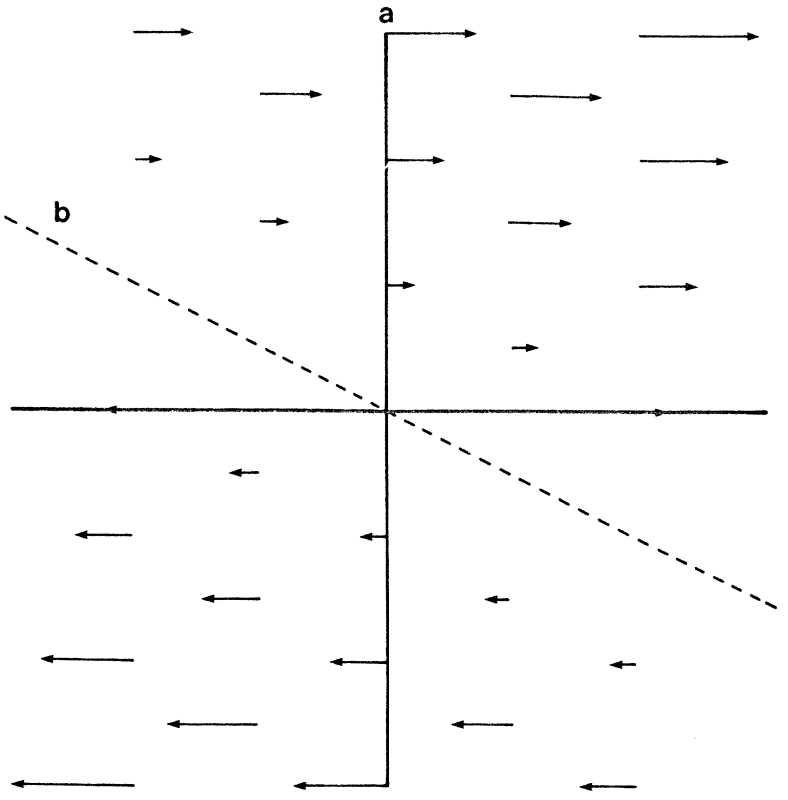

Figure 5 The velocity field of a planar surface. Line a is perpendicular to the field direction, and the field is nullified along b. This field has two different interpretations. In one, a is the rotation axis and b is the tilt line. In the second interpretation, roles are switched: a is the tile line, and b is the projection of the rotation axis.

objects rotate about a common axis in the image plane. Within loss of generality, this axis can be labeled the Z axis, and the general rotation can thereby be reduced to the case of vertical rotation.

Appendix 2: The velocity fields of planar surfaces

For a rotating planar surface, when the velocity field is parallel to the X axis, it will have the following form:

$$u(x, z) = \alpha x + \beta z \qquad (5).$$

That is, the field depends linearly on x and z. This velocity field is illustrated in Figure 5. Two special lines are marked in this figure: line a (coinciding in this case with the z axis), which is perpendicular to the field direction, and line b (the dotted line in Figure 5) along which the field is nullified.

In describing the 3-D interpretation of this field, we shall make use of the following definition: a plane that is not parallel to $X - Z$ must intersect it along a straight line, which will be called the *"tilt line"* of the plane. We then have the following proposition:

- There are exactly two planar interpretations of the velocity field in Figure 5. In one interpretation, a is the axis of rotation, and b is the tilt line. In the second interpretation, the roles are switched: a is the tilt line, and b is the projection of the rotation axis (the axis itself lies on the planar surface). Each of these interpretations is determined as before up to depth scaling.

The proof of this proposition is rather straightforward, and will therefore be sketched briefly. Let $(\Omega_x, \Omega_y, \Omega_z)$ be the angular velocity vector of the rotating object. The velocity of a point (x, y, z) is given by the vector product $(\Omega_x, \Omega_y, \Omega_z) \times (x, y, z)$. The velocity field is required to have the form: $u(x, z) = \alpha x + \beta z$, $v(x, z) = 0$. Therefore,

$$\begin{aligned} \Omega_y z - \Omega_z y &= \alpha x + \beta z \\ \Omega_x y - \Omega_y x &= 0. \end{aligned} \tag{A6}$$

One solution to these equations arises when $\Omega_y = 0$. This implies that $\Omega_x = 0$ (if y is not identically 0) and $y = -(\alpha x + \beta y)/\Omega_z$. This solution corresponds to a plane whose tilt line is b, rotating about the vertical axis.

If $\Omega_y \neq 0$, then $\Omega_x \neq 0$ and $y = (\Omega_y/\Omega_x)x$. The surface must therefore be a plane with its tilt line coinciding with the Z axis.

In conclusion, the following statements summarize the analysis of the information content of the orthographic velocity field, for both planar and non-planar objects. Given the orthographic velocity field:

- It is possible to determine whether the inducing object is planar.

  If it is non-planar, then the interpretation is determined exactly up to depth scaling.

  If it is planar, then there are two distinct solutions up to depth scaling.

Appendix 3: Unique 3-D structure from a "view and a half"

As we have seen, a single view of the orthographic velocity field is insufficient to recover the 3-D structure of inducing objects. In this section, we shall see that with additional, "half-a-view", the 3-D structure of an object containing at least five points in a general configuration is uniquely determined.

Five points in space are arranged in a general configuration if no four of them are coplanar. It is assumed here that two views of the five points in motion are given.

The first view gives the projected positions (x_i, z_i), and the velocities (u_i, v_i), $i = 0..., 4$ of the five points at time t. The second view gives the position of the

points (x_i, z_i) at some later time t'. We shall assume that between t and t', the rotation of the object was less than 180 degrees.

The "view-and-a-half" proposition:

- Given two orthographic views of five points in general configuration, the first view specifying the position and velocities of the points, the second their position only, the 3-D structure of the five points is uniquely determined.

Proof: We have seen that the 3-D structure of non-planar objects can be recovered up to depth-scaling from the velocity field alone. The same result holds when the velocity field is known for five isolated points, rather than within a continuous region. What remains to be shown, therefore, is that the depth-scaling ambiguity can be removed with the addition of a single view at a later time t'.

The proof is divided into two parts. In the first part, the object is assumed to rotate about the vertical Z axis. The second part extends the results to general motion.

Rotation about the vertical axis

Without loss of generality, it can be assumed that the object is fixed at one point (x_o, y_o, z_o), which will serve as the coordinate system's origin. For a rotation about the vertical Z axis, the Z coordinates of all the points remain unaltered, and the x coordinates are transformed according to:

$$x'_i = x_i cos\alpha + y_i k sin\alpha \qquad i = 1...4 \qquad (A7)$$

where α is the unknown angle of rotation about the vertical axis between viewer V_1 and V_2. Both x_i and y_i are already known while k is the still unknown depth-scaling factor. We thus obtain four linear equations (for $1 \leq i \leq 4$) in two unknowns, $cos\alpha$ and $ksin\alpha$. Two independent equations are sufficient for a unique solution. Two equations of the form

$$x'_i = x_i cos\alpha + y_i k sin\alpha$$

$$x'_j = x_j cos\alpha + y_j k sin\alpha$$

will not be independent if $x_i y_j = x_j y_i$. We shall, therefore, fail to obtain a pair of independent equations only if $x_i y_j = x_j y_i$ for every $1 \leq i, j \leq 4$. This condition implies that all of the points are lying on a common plane passing through the origin in contradiction to the general configuration assumption. A unique solution for $cos\alpha$, $ksin\alpha$ is therefore guaranteed, and this solves α and k up to a sign (which is the inherent ambiguity with respect to reflection about the image plane).

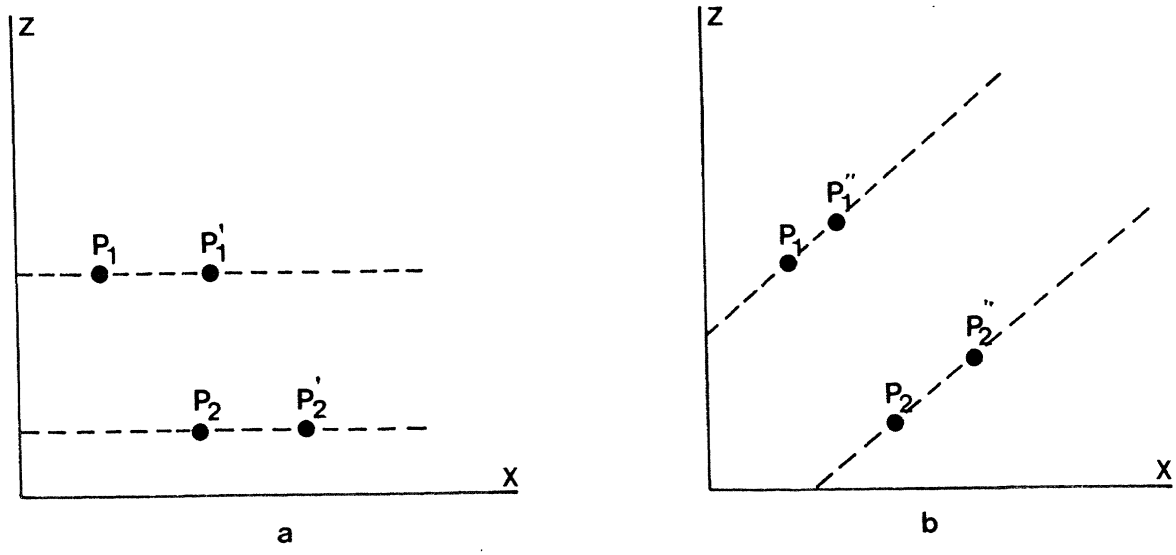

Figure 6 (a) The lines $P_1 - P'_1$ and $P_2 - P'_2$ are parallel to the X axis. (b) The points P'_1 and P'_2 in the figure have been rotated together about the origin by an angle α. Their new positions are P''_1 and P''_2 respectively. The lines $P_1 - P''_1$ and $P_2 - P''_2$ are parallel to each other.

Uniqueness under general motion

As has been noted in earlier sections, general rotation can be decomposed into two components: a rotation about some axis (the XZ axis) lying in the image plane, following by a rotation (Y-rotation) about the line of sight. To extend the uniqueness result to general motion, we shall prove next that for five points in a general configuration, the Y-rotation and the XZ axis can be recovered uniquely from the two views. Consequently, it is always possible to "undo" the Y-rotation and re-label the XZ axis as the new Z axis, thereby reducing the motion between the two views to rotation about the vertical axis. The remainder of the proof is divided into two lemmas.

Lemma 1: Let P_1, P_2, P'_1 and P'_2 be four points in the $X - Y$ image plane, such that the lines $P_1 - P'_1$ and $P_2 - P'_2$ are both parallel to the X axis (figure 6a). Let P'_1 and P'_2 now rotate together in the plane about the origin by an angle α, and denote by P''_1, P''_2, the new position of P'_1, P'_2, respectively. Question: is there an angle α such that the new lines $P_1 - P''_1$ and $P_2 - P''_2$ will again be parallel to each other? (See fig 6.)

Claim: Under the following two conditions

(i) (P'_1, P'_2) is not a reflection of (P_1, P_2) about the Z axis;

(ii) at least one of the lines $P_1 - P_2$, $P'_1 - P'_2$ does not pass through the origin; there exists exactly one such angle α.

Proof: Let the coordinates of P_i be (x_i, z_i) of P'_i be $(x'_i z'_i)$ and of P''_i $(x''_i z''_i)$.

$$x''_i = x'_i cos\alpha - z'_i sin\alpha$$

$$z''_i = x'_i sin\alpha + z'_i cos\alpha \qquad (A8)$$

Following the rotation, the lines $(P_1 - P''_1)$ and $(P_2 - P''_2)$ are parallel, therefore their slopes coincide:

$$\frac{z_1 - z''_1}{x_1 - x''_1} = \frac{z_2 - z''_2}{x_2 - x''_2} \qquad (A9)$$

(There is also the possibility that $x_1 = x''_1$ and $x_2 = x''_2$ and therefore the denominators vanish. We shall see, however, that in this case, there is still a unique solution for α.)

Substituting for x''_i and z''_i from $(A8)$ yields the following. (Note that $z_1 = z'_1$, $z_2 = z'_2$.)

$$\frac{z_1 - x'_1 sin\alpha - z_1 cos\alpha}{x_1 - x'_1 cos\alpha + z_1 sin\alpha} = \frac{z_2 - x'_2 sin\alpha - z_2 cos\alpha}{x_2 - x'_2 cos\alpha + z_2 sin\alpha} \qquad (A10)$$

which reduces to the form

$$asin\alpha + bcos\alpha = b \qquad (A11)$$

where

$$a = x_1 x'_2 - x'_1 x_2$$

$$b = x_1 z_2 + x'_1 z_2 - z_1 x_2 - z_1 x'_2.$$

If a and b are not both zero, $(A11)$ has exactly one solution, given by:

$$sin\alpha = \frac{2ab}{a^2 + b^2}$$

$$cos\alpha = \frac{b^2 - a^2}{a^2 + b^2}. \qquad (A12)$$

If $a = 0$ and $b = 0$, then $(A11)$ provides two equations in x'_1 and x'_2. If these equations are independent, then their solution is $x'_1 = x_1$, $x'_2 = -x_2$, in violation

of condition (i) of the lemma. The equations are dependent when $x_1/x_2 = z_1/z_2 = x'_1/x'_2$ which violates assumption (ii).

Finally, if there is an angle α which makes the denominators in $(A9)$ equal to zero, this angle is still a solution, and the only solution to the equation is $(A11)$.

Let us consider next two frames of an object with five points (in general configuration) rotating about the Z axis. Let P_i and P'_i for $i = 0\ldots,4$ denote the first and second frames, respectively, without loss of generality $P_o = P'_o$. As in lemma 1, the lines $P_i - P'_i$, $i = 1\ldots,4$ are all parallel to the X axis. The points P'_i are now rotated about the origin, and their new positions are denoted by P''_i.

Lemma 2:

For every $0 < \alpha < 2\pi$, the lines $P_i - P''_i$ will no longer be all parallel to one another.

Proof: Suppose that an angle α exists such that $P_i - P''_i$, $i = 1\ldots4$ are all parallel to one another. Consider the triplet of points (P'_o, P'_1, P'_i) (where $P'_o = P_o$ is the origin, and $i = 2, 3, 4$). From lemma 1, such a triplet has a single angle α_i that would make the line $P_1 - P''_1$ parallel to $P_i - P''_i$. To make all the lines $P_i - P''_i$, $i = 2\ldots4$ parallel simultaneously, $\alpha_2 = \alpha_3 = \alpha_4 = \alpha$. From $(A11)$,

$$\frac{a_i}{b_i} = \frac{1 - cos\alpha}{sin\alpha} = c \qquad (A13)$$

where c is constant. ($sin\alpha \neq 0$ implies that $b_i \neq 0$, unless both a_i and b_i equal zero.) We therefore obtain that if $b_i \neq 0$,

$$\frac{a_i}{b_i} = \frac{x_1 x'_i - x'_1 x_i}{x_1 z_i + x'_1 z_i - z_1 x_i - z_1 x'_i} = c. \qquad (A14)$$

We now make use of the fact that the points (x'_i, y'_i) are obtained from $(x_i y_i)$ via rotation by some angle $\theta \neq 0$ about the Z axis, namely

$$x'_i = x_i cos\theta - y_i sin\theta. \qquad (A15)$$

Equation $(A14)$ now becomes

$$x_1(x_i cos\theta - y_i sin\theta) - x_i x'_1 =$$

$$c x_1 z_i + c x'_1 z_i - c z_1 x_i - c z_1 (x_i cos\theta - y_i sin\theta). \qquad (A16)$$

Equation (A16) holds for $i = 2, 3, 4$, including the points (if any) for which $a_i = b_i = 0$, in which case it is satisfied for every value of c. Note that x'_i, y'_i, z'_i do not appear in $(A16)$, which therefore reduces to the form:

$$Ax_i + By_i + Cz_i = 0 \qquad (A17)$$

where

$$A = x_1 cos\theta - x'_1 + cz_1 + cz_1 cos\theta$$
$$B = -x_1 sin\theta - cz_1 sin\theta$$
$$C = -cx_1 - cx'_1.$$

Equation $(A17)$ is an equation of a plane passing through the origin. All of the points (x_i, y_i, z_i) for $i = 2, 3, 4$ must obey $(A17)$, and are, therefore, coplanar, in contradition to the general configuration assumption. It can be assumed that A, B and C are not all zero. Examine, for instance, the coefficient B, which depends only on the "reference point" (x_1, z_1). If $B = 0$, we will choose a different point as the reference point (x, z_1). B cannot vanish for all points, since such a condition would imply

$$B = -x_i sin\theta - cz_i sin\theta = 0$$

for $i = 1...4$, which is again, an equation of a plane.

Conclusion: Given two orthographic projections of an object in a general configuration, the XZ axis and the Y-rotation are uniquely determined. If α denotes the Y-rotation between the two frames, then rotating the second frame by $-\alpha$ would make all the lines $(x_i, z_i) - (x'_i, z'_i)$ parallel to one another and the XY axis would then be perpendicular to these lines. Lemma 2 established that this angle α is unique, and hence, the Y-rotation and XY axis are determined uniquely by just two frames.