Perception & Psychophysics
1996, 58 (3), 449-459

Structure from motion: A tolerance analysis

MAARTEN A. HOGERVORST, ASTRID M. L. KAPPERS, and JAN J. KOENDERINK
University of Utrecht, Utrecht, The Netherlands

We present a tolerance analysis that is applicable to a large group of stimuli used in structure-from-motion tasks. Human performance in structure-from-motion tasks reflects the fact that the visual system deals with projections of a 3-D world on the retina. A tolerance analysis reveals the relationship between the projections and the 3-D world. Any realistic model of the visual system should incorporate a tolerance analysis as a complete description of the stimulus. By way of example we apply the tolerance analysis to the stimuli used in two widely known experiments in which different properties of structure were tested—that is, perceived nonrigidity (Norman & Todd, 1993) and ordering in depth (Hildreth, Grzywacz, Adelson, & Inada, 1990). The analysis explains qualitatively the results of these experiments, illustrating that the results are to a large extent due to stimulus limitations rather than to mechanistic properties of the visual system. From our analysis it follows that far more sensitive measurements of the optic information are needed to obtain metric structure than affine structure.

For a proper evaluation of the performance of the visual system one needs to analyze the stimuli. From such an analysis it should become clear whether the information required for a certain task *is available* and *to what extent*. If we do not perform such an analysis we cannot decide whether the limits reached by the visual system are set by stimulus limitations or by limits of the visual system. Koenderink and van Doorn (1987) performed a tolerance analysis to measure the extent to which certain information is available in velocity flow fields. Here we will do the same for an important class of structure-from-motion stimuli, using multiple frames. As an example the tolerance analysis will be applied to two widely known experiments. In this introduction we will describe the structure-from-motion problem and the role a tolerance analysis can play in the investigation of it.

A tolerance analysis of a system consists of an analysis of the degree to which the output of the system is resistant against perturbations in the input. When the output fluctuates a lot with small fluctuations in the input, it can be said that the output is not well defined. To develop a model of the visual system one can make use of an ideal detector model. By *ideal detector* we mean a system that reaches the optimal performance with respect to some criterion. Note that what is an ideal detector heavily depends on the definition. By comparing the performance of the visual system with an ideal detector model one can determine the extent to which the visual system approaches ideal performance.

In this article we will show the use of a tolerance analysis in structure-from-motion (SfM) experiments. In SfM the input to the visual system consists of projections of the 3-D world on the retina (we discard input from other sensory systems). The fact that 3-D properties have to be inferred from projections highly limits performance in 3-D tasks. We will consider a world consisting of a number of landmarks. To obtain the 3-D structure of a number of landmarks from its projections, assumptions have to be made about the 3-D structure or the 3-D movement between the projections. Like many others, we assume that the 3-D movement is a rigid one. We assume that the projected locations of the landmarks will be available up to limited precision. More precisely, we assume that the locations available to the system are drawn from normal distributions around the actual ones. No other assumptions about the 3-D structure or the movement are being used. We will explain a method to develop an ideal detector model for rigidly moving objects given these assumptions. We will develop such a model for a situation in which the object rotates rigidly about an axis parallel to the image plane. By comparing performance of the visual system with that of the ideal detector model we can find out the degree to which it resembles the ideal detector model. Taking 2-D measures, which are more directly related to the stimulus, has several advantages over taking 3-D measures of performance in SfM tasks. As a measure of the performance of the visual system we propose the amount of noise in the input of the ideal detector model under which it reaches the same level of performance as the visual system. This supplies the means for comparing visual performance for different circumstances and different tasks.

The SfM task requires deriving the 3-D structure of the environment from the 2-D projections on the retina. When one moves around in the environment, the projections of objects at different distances from the eye move relative to each other on the retina. Aspects of the 3-D layout of the environment can be derived from these relative movements. Wallach and O'Connell (1953), among others, showed that the visual system is capable of deriving SfM from stimuli from which all cues about depth except motion parallax have been removed. Typically stimuli are composed of a

This research was supported by the Netherlands Organization for Scientific Research. Correspondence should be addressed to M. A. Hogervorst, Department of Experimental Psychology, University of Oxford, South Parks Rd., Oxford OX1 3UD, England (e-mail: m.hogervorst@psy.ox.ac.uk).

number of discrete markers. Here we will consider only such point configurations.

The SfM problem can be split up into subproblems. The first problem that arises is the so-called correspondence problem. This means that the markers should retain their individuality over time. In this paper we ignore this problem, although it is by no means a trivial one. Once the correspondence problem has been solved we can investigate the extent to which relative 2-D movement leads to a 3-D solution.

A widely held assumption in the literature is the "rigidity assumption": It is assumed that from all possible 3-D interpretations the visual system will choose a rigid one. However, when small changes are made in the stimulus, the available rigid interpretations can change considerably. Moreover, the changes can be such that no rigid interpretation remains. A reasonable assumption is that if the changes are within certain tolerances, the interpretation of the visual system will not change significantly—that is, whether a rigid object will be perceived is a matter of tolerances.

The environment is projected on the retina. The actual projection is often approximated by other projections, such as orthographic projection. In orthographic projection the markers are projected along straight, parallel lines that cross the projection plane orthogonally. A perspective projection of motion in depth can be approximated using orthographic projection by scaling each view according to the distance of the object from the eye. This maintains the dependence of overall size on overall distance across views, yet the individual views are still orthographic projections (Koenderinck & van Doorn, 1991). The nature of the information that is lost differs for different projections. A nice evaluation of the differences between the projections is supplied by Ullman (1983). A tolerance analysis should be applied to reveal the importance of the information lost by using a particular projection method.

A question recently raised in the literature on SfM addresses the nature of the perceived structure. This is closely related to the problem of temporal order. Does the visual system use only the velocity field, or can it use a higher order temporal description of the flow (e.g., accelerations)? Many authors have tried to address this question by relating the performance of the visual system to the theoretical minimal number of views necessary for a certain task. Theoretically, three orthographic projections of four points carry the information to recover the 3-D structure up to a uniform scaling factor (Ullman, 1979). This means that the relative lengths of 3-D line elements pointing in different directions can be compared. Two orthographic projections contain the information to recover the 3-D structure up to a uniform scaling factor and a combination of a shear and a stretch in depth (Bennett, Hoffman, Nicola, & Prakash, 1989; Huang & Lee, 1989; Koenderink & van Doorn, 1991). For many tasks a full description of the 3-D structure is not required. In fact, for many tasks it is sufficient to have an affine representation of an object. Koenderink and van Doorn (1991) showed that only affine methods are required on two parallel projections to obtain such a representation. When subjects are capable only of performing tasks that require a theoretical minimum of two projections and not tasks in which a minimum of three projections is required, this supports the view that subjects use only two projections.

A similar argument has been made by Todd and Bressan (1990) and Todd and Norman (1991). They argued that performance in tasks in which three projections are required is relatively poor compared to performance in tasks in which only two projections are required. In their experiments, performance does not significantly improve with an increase in the number of projections from two to eight. By contrast, in Hildreth et al.'s (1990) experiments performance does increase when the number of frames is increased. This difference in behavior is probably related to the fact that in Hildreth et al.'s experiments the stimulus duration was coupled to the number of frames, whereas in the experiments of Todd et al. it was not. In the experiments of Todd et al., the objects rotated back and forth for some time. Todd and Bressan (1990) concluded that the human visual system does not use information about 3-D structure over more than two frames in such a case. A tolerance analysis can reveal the degree to which the information for a certain task is better specified if the number of frames is increased, or, more precisely, how the robustness against noise increases with an increasing number of frames.

In this article we develop an ideal detector model that can be applied to an important class of stimuli used in SfM. We demonstrate the value of such an analysis by applying it to the stimuli used in the experiments of Norman and Todd (1993) and Hildreth et al. (1990). The results of these experiments will be reevaluated.

Obtaining the Optimum Rigid Structure

The tolerance analysis we propose consists of an evaluation of the robustness against noise of an ideal detector model. We explain how an ideal detector model can be developed for a situation in which (1) the rigidity assumption holds, (2) the locations used as an input to the model are assumed to be drawn from Gaussian distributions around the actual projected positions the width of which is independent of the location, and (3) there is no information about the 3-D structure or 3-D movement. This means that there is no a priori probability distribution about the possible solutions; there is only a posteriori probability distribution. The solution we search for obtains the solution with the highest posteriori probability. We call this the *optimum rigid structure*. In a realistic setting some structures and movements are more likely to occur than other ones. Once the a priori probabilities are known, they can be incorporated in a Bayesian analysis. Freeman (1994) showed how this can be done for several examples. The solution obtained here resembles the solution that Freeman would call the solution with the highest "fidelity."

The method for obtaining this solution is to search for the rigid structure that minimizes the sum of the squared differences between the projected positions of this solution and the 2-D locations of the points in the projections. This means that we try a rigid solution and move it until the sum of the quadratic distances between the projections of its 3-D markers and the 2-D points in the frames is minimized. This requires a movement that differs from projection to projection. The method can be generally applied to all kinds of movement and projections.

Here we explain a method for obtaining this optimum solution for a restricted class of stimuli; specifically, stimuli that have parallel trajectories of the projected points due to orthographic projections. It is assumed that the object rotates rigidly about an axis parallel to the image plane.

Any rigid transformation of an object can be regarded as a translation and a rotation about any 3-D point. For objects under orthographic projection it is convenient to decompose the rotation into a rotation about an axis perpendicular to the image plane and a rotation about an axis in the image plane (Whittaker, 1924). The only component of the transformation that gives information about the 3-D structure under orthographic projection is the rotation about an axis in the image plane.

What characterizes such a rigid solution? Consider a number of points rotating about an axis in the frontoparallel plane, which is referred to as the y-axis. The z-axis is chosen parallel to the line of sight; the x-axis is parallel to the trajectories of the projected points. The only relevant variables are distances in the direction of the flow. This effectively reduces the dimensions to two. Because the information about the spatial 3-D structure is given by the *relative* movement of the projected points, one of the points can be taken as the origin of the coordinate system.

The n 3-D points rotate about the y-axis with radii ($r_1, \ldots, r_n$), initial angles with the z-axis ($\phi_1, \ldots, \phi_n$) and rotational velocity ω. The projected distances ($x_1, \ldots, x_n$) can be described as a function of time (t):

$$\begin{cases} x_1 = r_1 \cos(\omega t + \phi_1) \\ \vdots \\ x_n = r_n \cos(\omega t + \phi_n) \end{cases}.$$

One can construct a rigid object by fitting sinusoidal functions to the projected positions if constant rotational velocity is assumed.

When no assumption about the movement in time is made, a better fit can generally be obtained. Therefore, one should eliminate the time parameter. When no assumptions are made about the movement in time, the information about the structure is contained in the relative positions of the projected points in the projections. If we plot the positions in n dimensional space on the axes ($x_1, \ldots, x_n$) we will find an ellipse parametrized by the angle ωt. We will call this somewhat abstract space the *phase space*. The relation between such an ellipse and the 3-D solution is shown for $n = 2$ in Figure 1. The relative movement of every two points describes an ellipse—that is, every projection of the trajectory in n dimensional space onto a plane results in an ellipse. From this it follows that the trajectory in n dimensional space will be also an ellipse, and thus a planar curve. *A rigid solution is characterized by an ellipse in phase space and vice versa*. The fitted values of ($r_1, \ldots, r_n$) and ($\phi_1, \ldots, \phi_n$) describe the 3-D structure of the fitted 3-D configuration of markers. The ($x_1, \ldots, x_n$) values can be measured in the projections.

Given a set of 2-D points in a number of frames (with Index j) that move horizontally from frame to frame, every Frame j results in a set of projected distances $(x_1, \ldots, x_n)_j$, that is, a Point X_j in n-dimensional space. In case the frames can be considered as orthographic projections of a rigidly moving object rotating about a vertical axis, all points X_j lie on the same ellipse in phase space. In case the Points X_j describe some other trajectory in phase space, the frames cannot be regarded as projections of a rigidly moving configuration. One can obtain the optimum rigid solution by performing a least squares fit of an ellipse to the Points X_j in phase space. This is equivalent to minimizing the function Ψ^2 using the constraint that all $r_i \geq 0$:

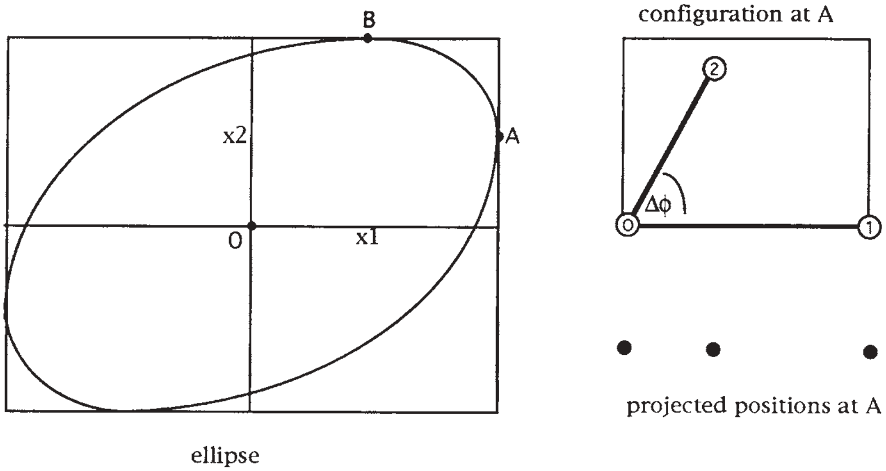

Figure 1. The projected length of Vector 1 versus projected length of Vector 2 with radii r_1 and r_2 and angle $\Delta\phi$ together with a top view of the configuration at Point A. At Point A: $(x_1, x_2) = (r_1, r_2 \cos \Delta\phi)$. At Point B: $(r_1 \cos - \Delta\phi, r_2)$.

$$\Psi^2 = \sum_{i,j} [x_{i,j} - r_i \cos(\theta_i + \alpha_j)]^2 .$$

The deviation of the actual trajectory from the fitted ellipse (Ψ^2) can be regarded as a measure of nonrigidity.

For every system that measures the relative projected positions with limited precision, there will be a range of undiscriminable stimuli representing different 3-D objects. In phase space this can be thought of as a band of indiscriminable trajectories. All ellipses within this band represent rigidly moving objects that are indiscriminable from each other. All nonelliptical trajectories represent nonrigidly moving objects that are indiscriminable from these rigidly moving objects. Under the rigidity assumption they will be perceived as rigid.

Equivalent Noise Level

Performance in SfM tasks is usually expressed in terms of 3-D properties of the simulated objects. Because these properties are available via the stimulus, it can be expected that performance is highly stimulus dependent. Therefore, it will be difficult to compare performance for different tasks. To compare performance in different situations, we therefore propose to compare performance of the visual system with that of the ideal detector model. As a measure of performance, we propose to take the equivalent noise level referred to the input. This is the amount of noise under which the ideal detector model reaches the same level of performance as that of the visual system. The "equivalent noise level" can be considered as a measure of the ability of the visual system to extract certain information from the stimulus. If the equivalent noise level is higher in a given situation than in another one, this implies that the visual system is better at extracting the information from the stimulus in the first case than in the second case. This is a stimulus-related measure of performance and therefore has advantages over expressing performance in 3-D measures. Consider, for instance, a situation in which three orthographic projections of four points are available. Ullman (1979) has shown that for this situation there is a unique 3-D interpretation (apart from a uniform scaling and an inversion in the image plane). Let us consider a situation in which the locations in the projections are subject to Gaussian noise around the actual location. Depending on the disturbance in the locations, the 3-D interpretation deviates from the original one in a different way. The transformation from 2-D locations to the 3-D interpretation is nonlinear. This means that averaging in 3-D introduces a bias that depends on the magnitude of the noise.

EXAMPLE 1
Stretching Experiments of Norman and Todd

Norman and Todd (1993) used objects consisting of 12 line elements inside a cube with edges of 10 cm. Under various conditions subjects indicated perceived rigidity on a "rigidity scale." From frame to frame the objects stretched in a certain direction before they rotated around an axis in the frontoparallel plane. In this way, the total amount of deformation built up over frames. In their first experiment, the objects stretched in the viewing direction. In their second experiment, the objects stretched in the direction perpendicular to the axis of rotation. Subjects perceived a rigid structure when the objects stretched in the viewing direction, whereas they perceived a nonrigid object if it stretched in the other direction. Here tolerance analysis is applied to the stimuli used in these experiments. This will reveal the extent to which the information for detecting nonrigidity is present in the stimuli. The parameters used are the same as those used in the experiments of Norman and Todd.

Affine Transformation

The nonrigid transformation used in these experiments is an affine transformation. If a 3-D object is subject to an affine transformation we do not have to find the optimum rigid solution for all 3-D points at the same time. In the case of such transformations it is convenient to describe the markers forming the object in terms of base vectors. All the other vectors can be described by linear combinations of the base vectors. An affine transformation involves a combination of rigid transformations and stretching and shearing transformations in all directions, that is, a general linear transformation in 3-D. Once the transformation of the base vectors is known, the transformation of the whole object is known. In general, this means that it is sufficient to consider the transformation of four points, which define three base vectors. All the other markers forming the object will transform in the same way. If this is a rigid transformation, all markers transform under the *same* rigid transformation (i.e., the object transforms rigidly). If the transformation is limited to two directions, as in the case of parallel trajectories, the transformation of two base vectors suffices to describe the transformation of the whole object.

We choose the same coordinate system as in the previous section: rotation about the y-axis, z-axis in the viewing direction, and the x-axis parallel to the trajectories of the projected points. As base vectors we take a vector connecting the origin $\mathcal{O}$ to $(x_1, z_1) = (0,10)$ and $\mathcal{O}$ to $(x_2, z_2) = (10,0)$. These vectors can be regarded as two edges of the cube used by Norman and Todd (1993), which had edges of 10 cm. The vectors associated with the elements in the cube are enclosed by the base vectors. Any vector within the cube can be thought of as a linear combination of the base vectors. In the analysis we neglect pixel noise. Therefore, deviations from the optimum rigid solution for the vectors inside the cube are also linear combinations of the deviations for the base vectors. They will be comparable to deviations of the base vectors (with a maximum deviation of the sum of the deviations of both base vectors).

Stretching Along the Line of Sight

While the object rotates, it stretches along the line of sight (the z-direction). From frame to frame this transformation consists of a stretch in z by a Factor f followed by a rotation over an angle $\delta\alpha$ about the y-axis:

$$\begin{pmatrix} x' \\ z' \end{pmatrix} = \begin{pmatrix} \cos\delta\alpha & -\sin\delta\alpha \\ \sin\delta\alpha & \cos\delta\alpha \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & f \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix},$$

with

$$f = 1/(2^{A*\sin(N\delta\alpha)}).$$

N is the frame number and the stretching is done with different amplitudes A.

Figure 2 shows how the lengths in 3-D of the two vectors vary with frame number for an amplitude A of 0.6 (this is the largest amplitude used by Norman & Todd, 1993).

Figure 3 shows the projected length of Vector 1 versus the projected length of Vector 2 (i.e., x_2 against x_1) in phase space. The best fit consists of the vectors with lengths of 5.2 cm (in the z-direction $= r_1$) and 10.2 cm (in the x-direction $= r_2$). This means that the 3-D structure of the best fit differs from the box used as initial input (rigid rotation of the box would give a circle with a radius of 10 cm, which is shown in Figure 3. The lengths in depth are compressed. The angle between the 3-D base vectors remains 90° (= $\Delta\phi$). This solution represents a rectangular box filled with line elements, width of 10.2 cm instead of 10 cm and a depth of 5.2 cm instead of 10 cm. The fit represents the data very well. The standard deviation of the points from the ellipse is 0.7 mm, which is equivalent to 3′ of arc at a distance of 76 cm. Considering these small deviations it is perhaps not surprising that Norman and Todd (1993) found that observers perceived the object as rigid.

The fit is made using all 100 data points (i.e., $[x_1,x_2]$ values in 100 frames). The data are integrated over a large number of frames. If the integration takes place over a more limited number of frames the deviations will become even smaller. Note that the deviations are deviations in *relative* positions.

The best fit describes a structure that rotates at a speed that fluctuates over time. This is reflected by the fact that in Figure 3 points are more closely spaced in some regions than in others.

The increment angle between views is shown as a function of time in Figure 4. The increment angle of the fitted structure fluctuates over time. Norman and Todd (1993) reported that observers perceived a rigid object rotating with an angular velocity that fluctuated in time. This is consistent with the solution found here.

Figure 5 shows a view from above of the nonrigidly transforming box (used as input) and the fitted box in Frames 1, 3, 5, and so on. The projections of the edges on the image plane are very similar.

Stretching Perpendicular to the Line of Sight

While the object rotates, it stretches orthogonal to the line of sight (the x-direction). From frame to frame, the object stretches with a Factor f in the x-direction and then rotates over an angle $\delta\alpha$ about the y-axis:

$$\begin{pmatrix} x' \\ z' \end{pmatrix} = \begin{pmatrix} \cos\delta\alpha & -\sin\delta\alpha \\ \sin\delta\alpha & \cos\delta\alpha \end{pmatrix} \begin{pmatrix} f & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ z \end{pmatrix}.$$

Figure 6 shows x_2 against x_1 for an amplitude of 0.6 (the same as used above). Here the data fit an ellipse quite poorly. There is no reasonable rigid interpretation of the data. One should not be surprised therefore that Norman and Todd (1993) reported that observers perceived the object as nonrigid.

Perspective Projection

The algorithm for obtaining the optimum rigid solution applies only to objects under orthographic projection. Under perspective projection, distances in the direction of the axis of rotation have to be taken into account. Also, distances orthogonal to the axis of rotation no longer describe an ellipse. However, by using the graphical method of ellipse fitting, one can still get an idea of the deviations introduced by using orthographic projection instead of perspective projection.

In their experiments, Norman and Todd (1993) used a viewing distance of 76 cm. Figure 7 shows the deviations in the distances orthogonal to the axis of rotation that are in-

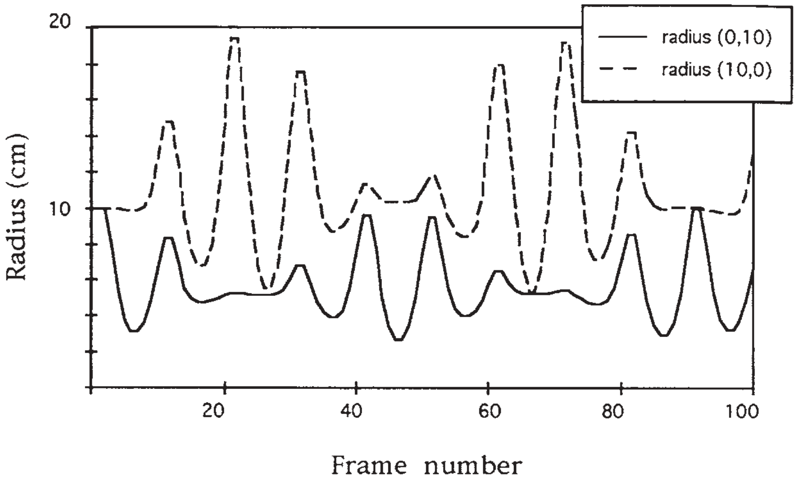

Figure 2. The 3-D radii of the vectors starting at (x,z) = (0,10) and (10,0) as a function of frame number if the object stretches in the viewing direction with an amplitude of 0.6 (see text).

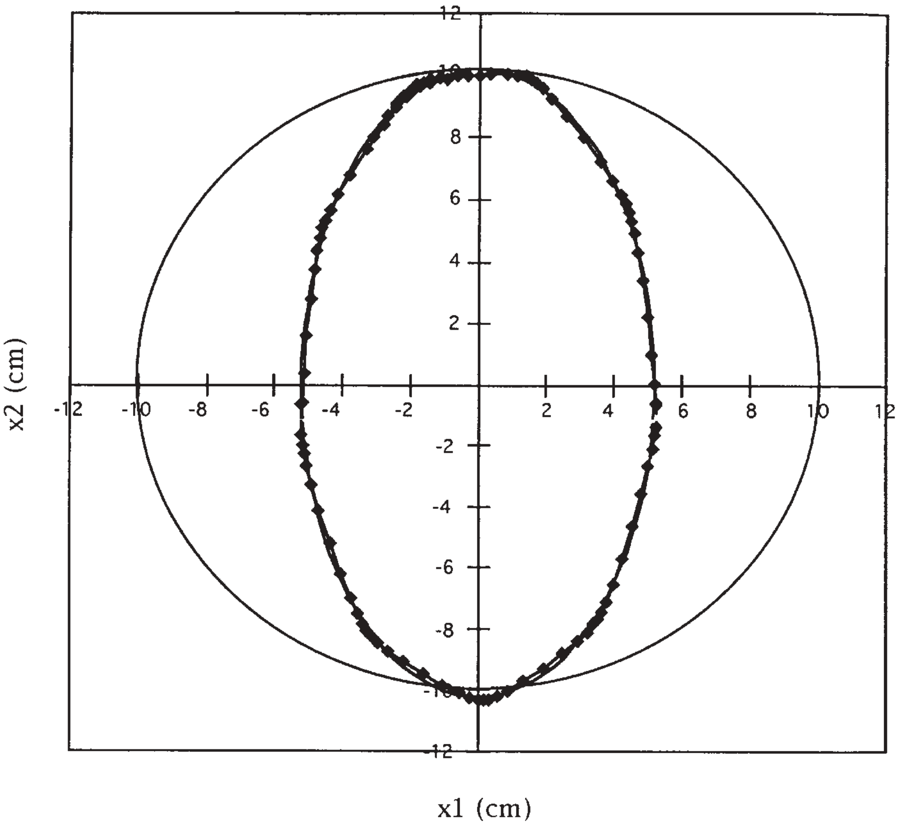

Figure 3. The projected length of Vector 1 versus projected length of Vector 2 if the object stretches in the viewing direction with an amplitude of 0.6. The circle with radius 10 describes the original input structure. The data points are fitted to an ellipse describing a configuration with vectors with lengths of 5.2 and 10.2 cm and an angle of 90º. Average root-mean square (RMS) deviation from the fit is 0.07 cm.

troduced by using orthographic instead of perspective projection. As expected, the data for the perspective projection do not describe an ellipse. The standard deviation between the differences in the projections of the perspectively projected and the orthographically projected points is 0.5 cm, equivalent to 21′ of arc. This means that the deviations introduced by using orthographic projection are larger than those introduced by stretching along the line of sight.

EXAMPLE 2
Build-up Experiments of Hildreth et al.

Hildreth et al. (1990) provided evidence for the plausibility of the incremental rigidity scheme proposed by Ullman (1984) as a model for the visual system. In this scheme an "internal model" of the 3-D structure of the object is maintained. While the object moves around, the internal model is adjusted so that the change in the 3-D distances between the points (actually a function of these) from the previous model to the current model is minimized. The initial model consists of a flat frontoparallel object. To what extent do the experiments support the idea that the visual system uses such a model? Hildreth et al. investigated the following two predictions (called "critical predictions" by Hildreth et al.) of the model: (1) The accuracy of the 3-D structure will build up over an extended period of time, and (2) the use of a current 3-D model will influence the model at a later stage.

The stimuli consist of three points rotating about a vertical axis. The points rotate from frame to frame over an angle of 1.5º in 33 msec. The points are orthographically projected. In the final frame the points are evenly spaced in depth. This is shown in Figure 8. The subject has to indicate which of the three points is situated between the other two, in depth. Over a number of frames performance increases and reaches a plateau. The level of the plateau depends on the separation in depth (γ) between the points in the final frame. The build-up period varies roughly between 30º and 45º depending on the subject. In some of the experiments, Gaussian noise was added to the positions. The results showed a gradual degradation in performance with increasing noise level. Unfortunately, because the rotational speed was constant, the turn angle was coupled to time, so it is impossible to judge whether the percept built up over time or over turn angle.

Hildreth et al. (1990) modeled their results using Ullman's (1984) incremental rigidity scheme. This scheme can be implemented in different forms; the general behavior is

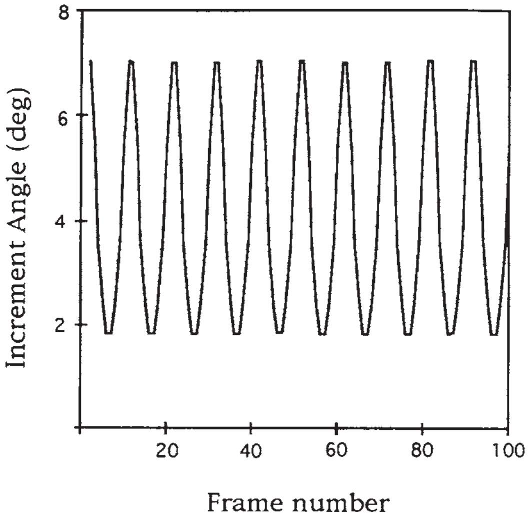

Figure 4. Increment angle between frames as a function of frame number as calculated from the best fit for stretching in the viewing direction with an amplitude of 0.6.

explained above. In this scheme no time parameter is incorporated explicitly. The internal model converges to the rigid interpretation with increasing number of frames. The convergence depends on the number of frames (if with the same turn angle more frames are used, performance will be better).

Can we explain the results on the basis of a tolerance analysis? The results can be fitted to any model that incorporates the following properties: (1) The positions are measured with limited precision, and (2) the information can be used over a limited, but extended amount of time or angular extent. The two "critical predictions" can be explained by these models as follows: (1) By integrating information over many views, performance improves, and (2) integrating information implies that a current fit will rely on former views.

For example, the performance of the ellipse-fitting algorithm is shown in Figure 9. The input to the algorithm consists of the perturbed projections of a configuration of three points rotating about the y-axis with an incremental angle between the frames of 5°. In the first frame, the three points form a configuration, as used by Hildreth et al. (1990). The depth of the first point is set to zero. One of the other points has a depth to the first point of $+\gamma$, the other one of $-\gamma$. An additional constraint is that the points have a minimum and a maximum distance of 0.5 and 2 times γ to the axis of rotation. The input to the algorithm consists of the perturbed x-values of the projected points in the frames. The perturbations are drawn from a uniform probability function around zero. Figure 9 shows the performance for different maximum deviations, indicated by the term *noise*. Every data point in the figure is an average of the performance over 1,000 trials. The performance is plotted as a function of turn angle for (1) different displacements (γ) with fixed noise of 1 and (2) for different levels of noise and fixed displacement of 10 units. The figure shows increasing performance with increasing turn angle and decreasing performance with increasing noise level. Performance increases to 100% for all levels of noise. This is evident from the fact that the effect of the noise of the output will gradually diminish with increasing number of views. To explain human performance, one should limit the integration over a limited amount of time or angular extent (corresponding to a limited amount of frames). In that case performance reaches a plateau at the point at which the full range is used for integration.

The experiments cannot be regarded as a critical test for investigating whether the visual system actually uses this particular scheme. The experiments show that we should exclude models of the visual system that use information only in the last few frames of the sequence. This leaves a large group of models that can explain the results satisfactorily.

DISCUSSION

The analysis of the stimuli used in the experiments of Norman and Todd (1993) shows that a stimulus representing a nonrigidly rotating object stretching in the viewing direction is very similar to a stimulus representing a

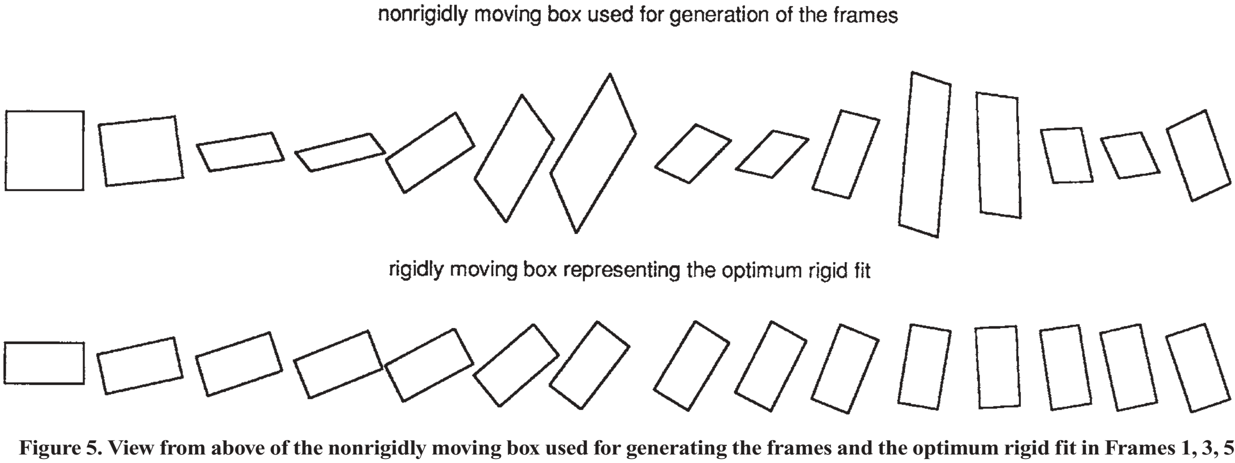
nonrigidly moving box used for generation of the frames
rigidly moving box representing the optimum rigid fit

Figure 5. View from above of the nonrigidly moving box used for generating the frames and the optimum rigid fit in Frames 1, 3, 5, and so on.

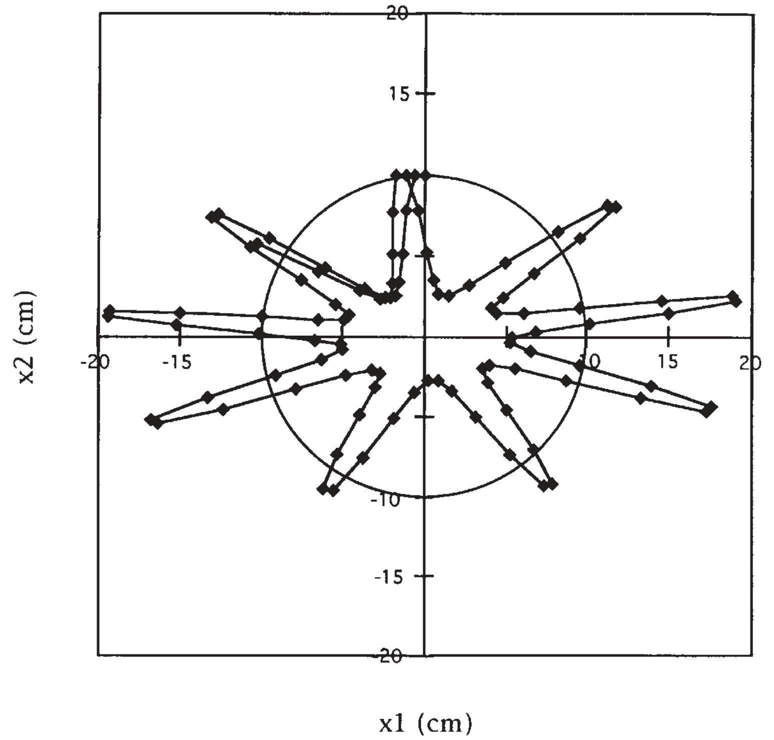

Figure 6. The projected length of Vector 1 versus projected length of Vector 2 if the object stretches perpendicular to the viewing direction with an amplitude of 0.6. The circle with radius 10 describes the original input structure. No fit is made to these data points.

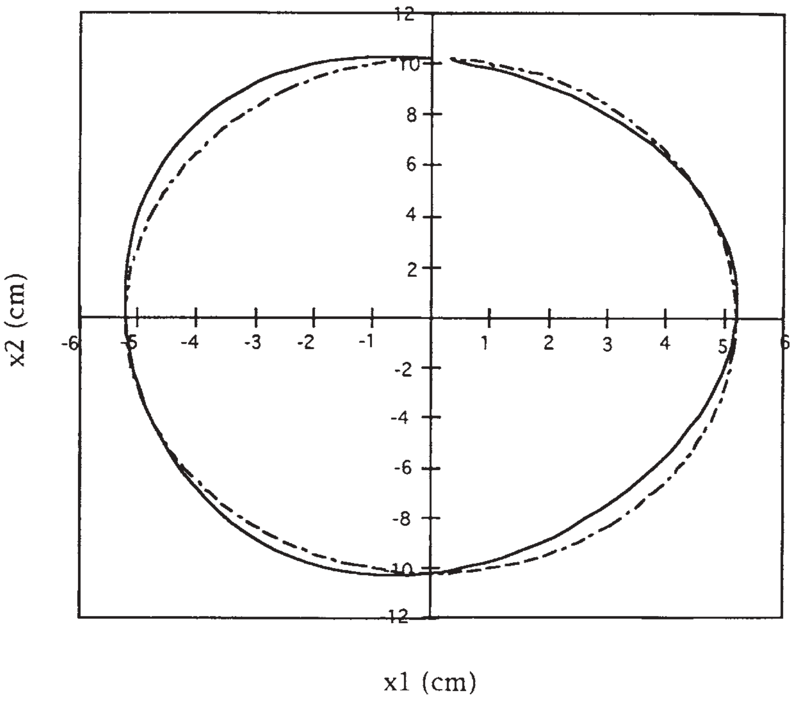

Figure 7. The projected length of Vector 1 versus projected length of Vector 2 for perspective (solid line) and orthographic (dashed line) projection. The rotating vectors have lengths of 5.2 and 10.2 cm and an angle of 90º (the fitted solution for stretching in the viewing direction); viewing distance is 76 cm (as used in the experiments of Norman & Todd, 1993).

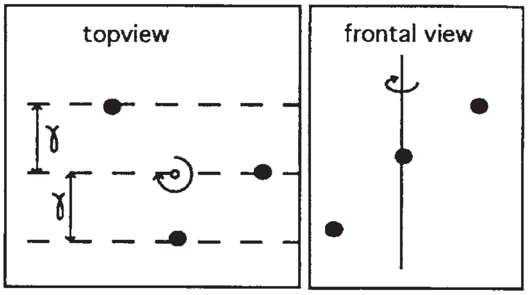

Figure 8. Top and frontal view of a configuration used by Hildreth, Grzywacz, Adelson, and Inada (1990) containing 3 points rotating about an axis in the frontoparallel plane. In the final frame the points are evenly spaced in depth with displacement γ.

rigid object rotating with varying angular velocity. In Norman and Todd's experiments, subjects indicated that they perceived a rigidly moving object in that case. This implies that the visual system prefers a rigid interpretation and does not deviate from such an interpretation if the differences between the projections of a rigid interpretation and the actual projections remain within certain limits. If the deviations become larger, a nonrigid object is perceived, as in the stimulus representing an object stretching in the image plane.

Moreover, from the analysis it follows that—for stretching in the viewing direction—the differences between the orthographic and the perspective projections of the rigid fit are larger than the differences between the orthographic projections of the rigid fit and those of the nonrigidly moving object. Although the differences are not of the same nature, it is not surprising that the nonrigidity remains unnoticed, since the deviations from perspective projection also remain unnoticed.

Norman and Todd (1993) applied an algorithm that uses three frames (Hoffman & Bennett, 1986) to prove that their stimulus contained the information needed to perform the task. Unfortunately, they did not include a tolerance analysis. If the projected positions are available with great accuracy, it is indeed possible to discriminate rigid from nonrigid. However, if some noise is added to the input, it is much harder to discriminate nonrigid from rigid structures on the basis of three successive frames. This becomes clear if pixel positions are used as an input.

Figure 10a shows the real radius and the radius calculated by the algorithm as a function of frame number for an object stretched in the viewing direction with an amplitude of 0.6. The calculated radius has a phase lag because the two former frames are used to calculate the structure in the current frame. The algorithm does not introduce any bias. Any algorithm using three frames in a correct way should give this result. *The algorithm is not unstable, but the information is not well defined.* Figure 10b shows the calculated radius of a rigidly rotating object formed by the vectors with initial (x,z)-values of (10.2,0) and (0,5.2) (the optimum fit to the object stretched in viewing direction) when the pixel positions are used (with same dimensions as in the experiment). In cases where the lines go to infinity, no solution is available. The deviations from the exact positions are comparable to the deviations introduced by stretching. Even the locations of regions with maximum deviation are comparable. The depth calculated from only three successive frames is very sensitive to noise in situations in which the vector is parallel to the viewing direction.

Pollick (1993) also analyzed the stimuli used by Norman and Todd (1993). He assumed that the total displacement of the projection of a point is twice the distance (r) to the axis of rotation. Given a projection of a point and its radius, the phase (α) of the point can be computed

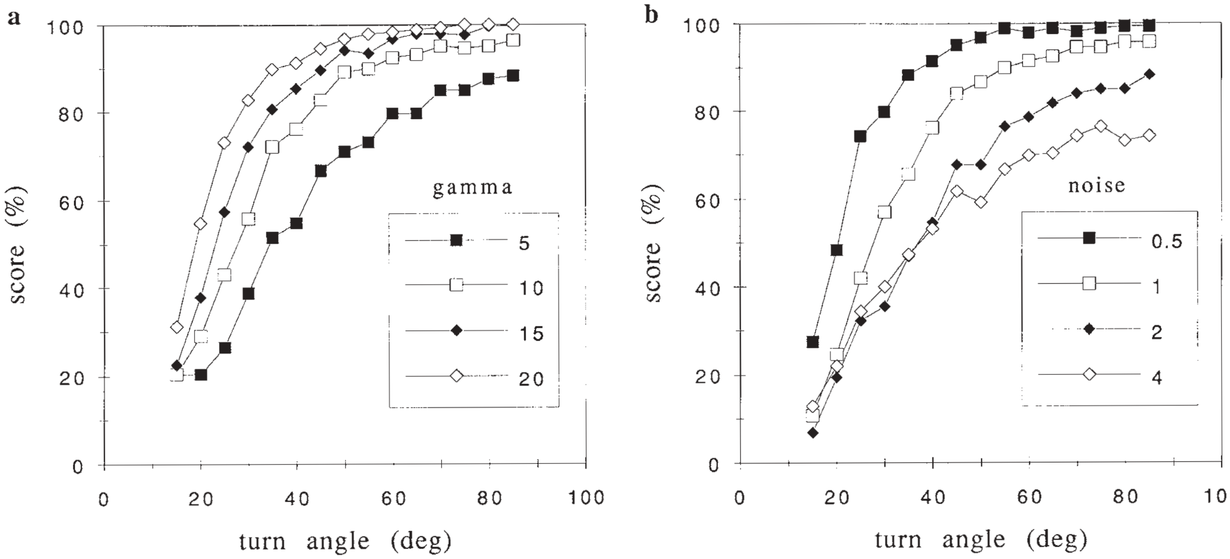

Figure 9. Performance of the ellipse-fitting algorithm in the experiments of Hildreth, Grzywacz, Adelson, and Inada (1990). Score as a function of turn angle with increment angle between frames of 5º (a) for different displacements (gamma) and fixed noise of 1, (b) for different levels of noise and fixed displacement of 10.

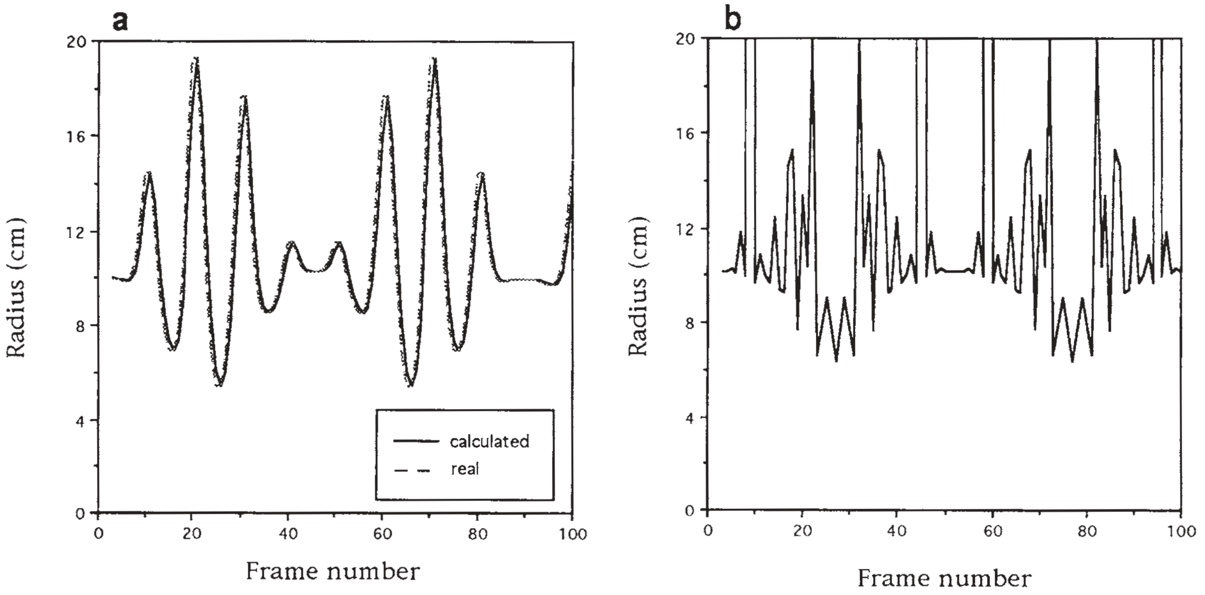

Figure 10. (a) Real and calculated radius of vector starting at $(x,z) = (10,0)$ as a function of frame number for stretching in the viewing direction with an amplitude of 0.6 as calculated by the Hoffman and Bennett (1986) algorithm using positions in three frames with infinite precision. (b) Calculated radius of vector starting at (10.2,0) as a function of frame number for rigid rotation as calculated by the Hoffman and Bennett (1986) algorithm using pixel positions in three successive frames (pixel dimensions as in the experiments of Norman & Todd, 1993) for an object consisting of vectors (10.2,0) and (0,5.2) (the optimum fit for stretching in the viewing direction).

$[r\cos(\alpha) = x]$. The sign of the displacement indicates whether the phase is negative or positive. If there is a rigid interpretation, all 3-D points should show the same incremental angle between frames.

Figure 11 shows the incremental angles calculated by this algorithm for two vectors if the object stretches in the viewing direction with an amplitude of 0.6. The incremental angles are very similar for both vectors. This algorithm works quite well in this situation. Pollick (1993) investigated more points and found only 1% of nonrigidity (measured as the variances of the interpoint distances).

Pollick's (1993) algorithm appears to be rather ad hoc. A structure is deduced by assuming constant distance from the 3-D points to the axis of rotation. This distance or radius is calculated using only two frames: the frames in which a projected point reaches its maximum distance from the axis of rotation. This means that an error in measuring these extremes results in errors in the phase in all frames (these frames are treated as special frames but should not be). If these frames are not available, the radius should be calculated in a different way. A result of Pollick's analysis is that the errors in the phase depend on the phase. Given a constant error in the projected position, the error in its phase will be larger for larger projected radii. A more fundamental argument against such a method is that one assumes that if the nonrigidity is detected, the visual system reconstructs a 3-D interpretation that has changing phase differences of the 3-D points but that has no changing radii. Whether this is the case is not yet clear. In our method, no assumptions are made about the 3-D interpretation of nonrigidly transforming objects.

The experiments of Hildreth et al. (1990) show that the visual system combines information over an extended but limited amount of time or angle of rotation. Any reasonable model of the visual system should display the same characteristic. Our analysis of their experiments suggests that they do not offer a particularly critical test of the incremental rigidity scheme as a model of the visual system in SfM tasks. The evaluation shows that a tolerance analysis is needed to determine whether any model of the visual system is a reasonable one.

The tolerance analysis can be used to compare performances in affine and metric tasks. The affine structure can be deduced from the velocity field. To obtain metric structure, higher order temporal derivatives (e.g., acceleration) have to be taken into account. Therefore, it can be expected that deducing affine structure will be more robust than deducing metric structure. Let us consider a representation of a stimulus by a trajectory in phase space. The affine structure is determined when the tangent to the trajectory is determined (first-order temporal description of the flow). To deduce metric structure one needs to know the curvature of the trajectory (second-order temporal description of the flow). In general the determination of the tangent is far more robust than the determination of the curvature. This is in agreement with the fact that humans are better at doing tasks needing affine structure than at tasks needing metric structure (Todd & Bressan, 1990 and Todd & Norman, 1991).

In conclusion, an analysis of the extent to which information is specified in the stimulus is required for a proper evaluation of the performance of the visual system. The examples provided here show that the level of performance can be determined largely by stimulus limitations, rather than by mechanistic limits of the visual system. By comparing the performance of the visual system with that of an

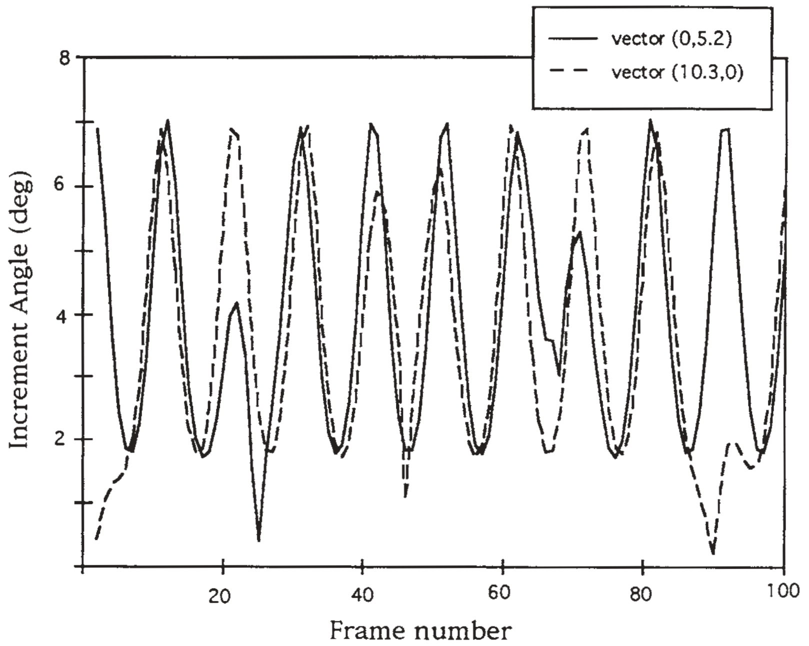

Figure 11. The increment angle between frames in the reconstruction of Pollick (1993) for two vectors of an object that stretches in the viewing direction with an amplitude of 0.6 and initial positions of (10,0) and (0,10).

ideal detector, we can find out the extent to which visual performance is determined by limits of the system itself.

REFERENCES

Bennett, B. M., Hoffman, D. D., Nicola, J. E., & Prakash, C. (1989). Structure from two orthographic views of rigid motion. *Journal of Optical Society of America A*, **6**, 1052-1069.

Faugeras, O. D., & Maybank, S. (1990). Motion from point matches: Multiplicity of solutions. *International Journal of Computer Vision*, **4**, 225-246.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, **368**, 542-545.

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

Hoffman, D. D., & Bennett, B. M. (1986). The computation of structure from fixed axis motion: Rigid structures. *Biological Cybernetics*, **54**, 71-83.

Huang, T., & Lee, C. (1989). Motion and structure from orthographic projections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**, 536-540.

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

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

Koenderink, J. J., & van Doorn, A. J. (1987). Facts on optic flow. *Biological Cybernetics*, **56**, 247-254.

Koenderink, J. J., & van Doorn, A. J. (1991). Affine structure from motion. *Journal of Optical Society of America A*, **2**, 377-385.

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

Norman, J. F., & Todd, J. T. (1993). The perceptual analysis of structure from motion for rotating objects undergoing affine stretching transformations. *Perception & Psychophysics*, **53**, 279-291.

Pollick, F. E. (1993). The tradeoff between motion and structure: Rigid interpretations of affine stretching displays. *Investigative Ophthalmology & Visual Science*, **34**, 1034.

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

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

Todd, J. T., & Norman, J. F. (1991). The visual perception of smoothly curved surfaces from minimal apparent motion sequences. *Perception & Psychophysics*, **50**, 509-523.

Treue, S., Husain, M., & Andersen, R. A. (1991). Human perception of structure from motion. *Vision Research*, **31**, 59-76.

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

Ullman, S. (1983). Recent computational studies in the interpretation of structure from motion. In A. Rosenfeld & J. Beck (Eds.), *Human and machine vision* (pp. 459-480). New York: Academic Press.

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

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

Whittaker, E. T. (1924). *Analytische Dynamik der Punkte und starren Körper* [Analytical dynamics of points and rigid bodies]. Berlin: Springer-Verlag.

(Manuscript received May 18, 1994;
revision accepted for publication August 21, 1995.)