3-D Structure Perceived from Dynamic Information: A New Theory

Article in Trends in Cognitive Sciences • November 2003
DOI: 10.1016/j.tics.2003.08.007 • Source: PubMed

2 authors, including:

Corrado Caudek
University of Florence
90 PUBLICATIONS  655 CITATIONS

See profile
Image movement provides one of the most potent two-dimensional cues for depth. From motion cues alone, the brain is capable of deriving a three-dimensional representation of distant objects. For many decades, theoretical and empirical investigations into this ability have interpreted these percepts as faithful copies of the projected 3-D structures. Here we review empirical findings showing that perceived 3-D shape from motion is not veridical and cannot be accounted for by the current models. We present a probabilistic model based on a local analysis of optic flow. Although such a model does not guarantee a correct reconstruction of 3-D shape, it is shown to be consistent with human performance.

To perceive the 3-D shape of objects from two-dimensional (2-D) retinal images, our brain uses binocular as well as monocular cues such as motion, occlusion, texture and shading [1]. Although 3-D shape perception normally depends on the simultaneous presence of many sources of depth information, human observers show a remarkable ability to extract the 3-D structure of an object from motion cues alone. This ability is referred to as structure-from-motion (SfM) perception. The importance of motion for 3-D shape perception was demonstrated 50 years ago by Wallach and O’Connell [2]: in their classic demonstration, the shadow of a wire-frame figure appears flat when the wire frame is stationary, but pops out in depth as soon as the wire frame is rotated.

Currently, the generally accepted theory of SfM perception is formulated in terms of the inverse-optics problem: given a 2-D image, the observer needs to determine the 3-D object from which the image is a projection [3–5]. The inverse-optic problem does not have a unique solution: the same 2-D image, in fact, is consistent with infinite 3-D objects. Nevertheless, computer-vision algorithms have been devised that recover the ‘correct’ 3-D shape (if opportune assumptions are met), either up to scale factor [6–9], or up to a stretching in depth [10], and these algorithms have been proposed as models of human performance [6,11].

Although it is widely believed that human SfM perception is veridical (i.e. that observers’ perceptions correspond to faithful copies of the projected objects), we have recently shown that perceived 3-D shape from motion is distorted by more than a scale factor or a stretching in depth [12,13] and, therefore, cannot be accounted for by the computational algorithms that have been developed for machine vision. We propose a probabilistic model that, in general, does not guarantee a correct solution to the inverse-optic problem but, nevertheless, is consistent with both veridical performance and perceptual biases in many SfM tasks [14,15]. Before presenting its rationale, we will briefly describe the two classes of SfM models that have been proposed so far.

Euclidean and affine descriptions
The SfM models can be divided into two classes: Euclidean and affine models. Euclidean SfM algorithms recover an exact copy (up to a scale factor) of the projected objects [3–5]. If the projected object is a wedge, for example, then the recovered shape will also be a wedge with the same angle between its two planar surfaces as the wedge that generated the image. Euclidean SfM algorithms correctly recover the metric properties of the projected 3-D structures. For the simple example of Figure 1, these properties correspond to the slants of the two surfaces (i.e. the tangents of the angles α₀ and α₁ between the surfaces of the wedge and the frontal-parallel plane) and to the absolute distance between any pair of points (e.g. the distance between the points P₀ and P₁ in Figure 1).

From the same image projections, an affine SfM algorithm would only recover a copy of the projected wedge that is stretched in depth by an arbitrary amount [10]. In this case, the angle between the two planar surfaces, in general, would not be preserved. Affine geometry provides a more abstract representation of 3-D shape than Euclidean geometry, and affine SfM algorithms correctly recover only properties such as the depth order between pairs of points, the parallelism between lines defined by pairs of points on the surface, and the coplanarity among points.

Optic flow
To understand human SfM we need also to know what information observers use to derive 3-D shape from motion. An answer to this question will be attempted in the present and the next section.

The image transformations produced by the relative motion between an observer and an object like that of Figure 1 can be described in terms of the 2-D motions of discrete texture elements belonging to the 3-D object. The
A 3-D rotating wedge produces an optic flow on the image plane. Such optic flow is represented by the blue and yellow vectors that correspond to the 2-D velocities of the projected features in two successive instants of time. In a coordinate system in which the $z$-axis corresponds to the line of sight, the $z$-depth distance between $P_0$ and $P_1$ is equal to the depth distance between $P_1$ and $P_2$. (b) View from above the wedge (left panel). $\Delta z$ represents the depth distance between $P_0$ and $P_1$ and between $P_1$ and $P_2$. $x_0$ is the size of the projection on the image plane of one of the two surfaces of the wedge and the frontal-parallel plane. Euclidean SfM models reconstruct a 3-D shape from the optic flow in which the angles $\alpha_0$ and $\alpha_1$ are derived in a veridical manner. In an affine algorithm (right panel) the wedge is stretched along the $z$-axis by four different amounts. In terms of an affine description, all these structures are equivalent. Consequently, in general, affine SfM algorithms do not derive the angles $\alpha_0$ and $\alpha_1$ in a veridical manner. (c) Motion of the projected texture elements can be described by the instantaneous velocities (blue arrows) and their change over time. This change over time can be measured by considering the values of each velocity vector in two successive instants of time (blue and yellow arrows). If we consider a small patch of the projected surface, its shape undergoes a non-rigid image transformation in two successive moments of time. For example, if this patch projects as a square in the first moment of time, successively the projected patch can take on the shape indicated in red. This 2-D shape change can be decomposed in four elementary transformations: a rigid image rotation ($\text{curl}$), an isotropic expansion or contraction ($\text{div}$), and two components of shear along vertical ($\text{def}_1$) and oblique axes ($\text{def}_2$). The two shear components quantify the amount of shape change and can be summarized by a unique value called $\text{def}$.

![Image 1](http://tics.trends.com)

**Figure 1.** (a) A 3-D rotating wedge produces an optic flow on the image plane. Such optic flow is represented by the blue and yellow vectors that correspond to the 2-D velocities of the projected features in two successive instants of time. In a coordinate system in which the $z$-axis corresponds to the line of sight, the $z$-depth distance between $P_0$ and $P_1$ is equal to the depth distance between $P_1$ and $P_2$. (b) View from above the wedge (left panel). $\Delta z$ represents the depth distance between $P_0$ and $P_1$ and between $P_1$ and $P_2$. $x_0$ is the size of the projection on the image plane of one of the two surfaces of the wedge and the frontal-parallel plane. Euclidean SfM models reconstruct a 3-D shape from the optic flow in which the angles $\alpha_0$ and $\alpha_1$ are derived in a veridical manner. In an affine algorithm (right panel) the wedge is stretched along the $z$-axis by four different amounts. In terms of an affine description, all these structures are equivalent. Consequently, in general, affine SfM algorithms do not derive the angles $\alpha_0$ and $\alpha_1$ in a veridical manner. (c) Motion of the projected texture elements can be described by the instantaneous velocities (blue arrows) and their change over time. This change over time can be measured by considering the values of each velocity vector in two successive instants of time (blue and yellow arrows). If we consider a small patch of the projected surface, its shape undergoes a non-rigid image transformation in two successive moments of time. For example, if this patch projects as a square in the first moment of time, successively the projected patch can take on the shape indicated in red. This 2-D shape change can be decomposed in four elementary transformations: a rigid image rotation ($\text{curl}$), an isotropic expansion or contraction ($\text{div}$), and two components of shear along vertical ($\text{def}_1$) and oblique axes ($\text{def}_2$). The two shear components quantify the amount of shape change and can be summarized by a unique value called $\text{def}$.

![Image 2](http://tics.trends.com)

**Figure 2.** (a) Image transformations produced by the vertical rotation of a planar surface. Initially, the 3-D surface projects a square on the image plane (pink). After rotation of the planar surface, the projected square is compressed (red). The rate of this compression represents the amount of $\text{def}$. Note that a large rotation of a surface having a small initial slant (top) could produce exactly the same amount of compression as a small rotation of a surface with an initial larger slant (bottom). (b) Family of slant and angular rotation magnitudes that produce the same image compression ($\text{def}$). Four instances of these $\omega$, $\sigma$ pairs are represented by the view from above the rotating surface.
overall motion of these feature elements is called optic flow [16]. For the present purposes, it is important to distinguish between a description of optic flow in terms of instantaneous velocities (the so-called first-order properties of the flow), and one in terms of velocities and accelerations (second-order properties) [17].

If the object in Figure 1 moves rigidly relative to a stationary observer, then a veridical description of its full Euclidean structure can, in principle, be derived from second-order optic flow [6,7]. The first-order properties of optic flow are sufficient, conversely, to specify only the affine structure of this object (i.e. its 3-D shape up to a depth stretching) [10].

Many psychophysical findings have demonstrated that the judgments of Euclidean properties from optic flow are not veridical [11,18,19]. In particular, it has been shown, first, that judgments of Euclidean properties are far less accurate than judgments of affine properties [11], and second, that human observers have a very limited sensitivity for second-order temporal properties and thus rely mainly on velocities to recover 3-D information from optic flow [11,20].

Initially these two findings prompted researchers to hypothesize that the perceptual analysis of optic flow is restricted to a veridical recovery of the affine properties of the projected objects [11]. In a series of investigations, however, we have also shown that the affine properties are not recovered from optic flow in a veridical fashion [12,13,21]. These findings have thus led to the conclusion that the perceptual analysis of optic flow cannot be accounted for by either Euclidean or affine SfM algorithms [13]. To account for these findings, we have developed a model that derives the local orientation and motion of 3-D surface patches from an ambiguous property of first-order optic flow called deformation [14,15]. In the next section we will describe this property and its relation to the parameters of the projected 3-D shape.

The ambiguity of local optic flow

First-order optic flow can be described intuitively by representing the retinal projections with a sequence of frames, and by considering how a local patch of the projected surface is distorted from one frame to the next. Such distortions can be thought of as a sum of four elementary changes (see Figure 1c). Two of them modify only the orientation (\(\text{curl}\)) and the size (\(\text{div}\)) of the projected patch; the other two modify the shape of the 2-D projection and are therefore called shearing or deformation. The two shearing components are the only image transformations that are informative about 3-D shape [17].

For the sake of simplicity, let us consider the optic flow produced by the rotation about the vertical axis (\(\omega\)) of one of the two planar surfaces of the wedge represented in Figure 1. After the rotation, the initial projection of this surface on the image plane will be compressed. In this particular case, \(\text{def}\) coincides with the amount of compression (Figure 2a).

How is \(\text{def}\) related to the metric properties of the projected 3-D structures? It should be noted that the compression produced by, say, a small slant and a large angle of rotation produces the same deformation as a small rotation of a more slanted surface (see Figure 2). There are infinite combinations of slant and angular rotation, therefore, which give rise to the same \(\text{def}\) (Figure 2) [14]. In the instantaneous and local case (i.e. for two frames very close in time and for a very small patch), \(\text{def}\) represents the rate of compression (or shearing, in the general case), and it is related to the slant \(\sigma\) and the angular velocity \(\omega\) by:

\[
\text{def} = \sigma \omega
\]  

Functional MRI studies have identified in the middle temporal area (MT/V5) and in the fundus of the superior temporal sulcus (FST), the human brain areas that have a basic role in SfM processing [22–24]. Moreover, neural activity in the dorsal division of the medial superior temporal
Box 1. Distortions of perceived depth and shape

Another important characteristic of our model is that it derives perceived slant as a non-linear function of def (see Eqn 2 in main text). What does this imply for the judgments of relative depth? Consider the probe dots \( P_0 \) and \( P_1 \) in Figure 1a. From simple trigonometry, the depth difference \( \Delta z = P_0 - P_1 \) is equal to \( \Delta z = \Delta x \tan(\alpha) \), where \( \Delta x \) is the size of the planar patch projected onto the image plane and \( \sigma \) (slant) is the tangent of the angle \( \alpha \). Accordingly, the perceived depth separation \( \Delta z' \) is equal to \( \Delta z = \Delta x \sigma \). The relation between the perceived and the simulated depth difference, therefore, can be expressed in terms of the ratio between the two previous equations:

\[
\frac{\Delta z'}{\Delta z} = \frac{\sigma'}{\sigma}
\]  

(3)

If perceived slant is expressed in terms of Eqn 2, we obtain

\[
\sigma' = \frac{\sigma_{\text{max}}}{\omega_{\text{max}}} \sqrt{\sigma_0} = k \sqrt{\sigma}
\]  

(4)

From Eqn 3 and Eqn 4, moreover:

\[
\Delta z' = k \sqrt{\Delta z}
\]  

(5)

Eqn 5 reveals that, according to our model, the perceived depth difference \( \Delta z' \) between \( P_0 \) and \( P_1 \) should be proportional to the simulated depth separation between the two dots \( \Delta z \), and inversely related to the slant \( \sigma \) of the patch on which \( P_0 \) and \( P_1 \) are located. This prediction was confirmed by Domin and Braunstein [13].

Implications for perceived 3-D shape

Our model predicts that local judgments of perceived 3-D shape also violate another important property of real 3-D structures: internal consistency. Let us consider a 3-D structure made up of a circular arrangement of wedges, as shown in Figure 1b (this circular crown is composed of elements similar in shape to the wedge represented in Figure 1 in the main text). Suppose one moves along a closed path on the image-plane projection of this surface; each 2-D step along this path corresponds to a depth difference on the 3-D surface. In order for this 3-D shape to be internally consistent, the sum of the depth differences associated with the successive steps along this closed path, from point \( P_0 \) back to itself, must be equal to zero. According to our model, however, the perceived depth difference between the dots \( P_0 \) and \( P_1 \) should be smaller than the perceived depth difference \( P_2 - P_1 \), because the slant \( \sigma \) of the patch on which \( P_0 \) and \( P_1 \) are located is larger than the slant \( \sigma' \) of the patch on which \( P_2 \) is located. According to our model, therefore, \( P_2 \) should be perceived to be closer to the observer than \( P_0 \).

To test this prediction, we asked observers to make judgments of depth order along a closed path for the probe dots \( P_0 \) and \( P_1 \) located at the joints of the planar surfaces making up a circular-crown surface [12], similar to that shown in Figure 1b. As predicted by Eqn 5, the point \( P_2 \) was judged to be closer than \( P_0 \). The depth judgments along this closed path, therefore, can be thought of as an ‘always ascending’ staircase, an obvious violation of the property of internal consistency which would result in the 3-D shape represented in Figure 1d. This result, together with other similar findings [13], indicate that neither Euclidean nor affine geometries provide an adequate description of perceived 3-D shape from motion.

Figure 1. (a) Top view of a planar surface slanted away from the image plane by an amount \( \sigma = \tan(\alpha) \). \( \Delta x \) is the horizontal component of the image-plane projection of the surface. \( \Delta z \) is the depth distance between the points \( P_0 \) and \( P_1 \). (b) Side view of the circular crown surface used by Domin et al. [12] to investigate internal consistency of depth judgments in structure-from-motion. The \( z \)-axis represents the line of sight, and \( \Delta z \) is the depth distance \( P_0 - P_1 = P_0 - P_1 \) (c) Schematic representation of the plane defined by the points \( P_0 \) and \( P_1 \) of each wedge element of the circular crown. In successive trials, the probe dots \( P_0 \) and \( P_2 \) were located in different positions along the circular path indicated by dashed white line. Even if the probe dots \( P_0 \) and \( P_2 \) were always simulated at the same distance from the observer, because \( P_0 \) was perceived to be closer than \( P_2 \) for each wedge element composing the circular crown, the integration of the observers’ judgments produced the internal inconsistency schematically represented in (d).

area (MSTd) in the macaque has been found to be specifically sensitive to the def component of optic flow [25,26].

Probabilistic interpretation of a moving retinal image

Having identified in first-order optic flow the information that observers use to derive 3-D shape from motion, we need now to describe how perceived 3-D shape can be derived from def. Here we will summarize the model that we have proposed, which tries to account for both correct judgments and biases in human SfM [14]. This model is based on a local analysis of first-order optic flow and its aim is to find the most likely \( \sigma, \omega \) pair given def (Figure 3). Our model analyzes the visual motion that is generated by an object moving relative to a passive observer and, thus, it does not take into consideration the specific contributions of extra-retinal information to 3-D shape perception [27–29]. The rationale of the model can be exemplified in the following manner.

Let us suppose that \( \sigma \) and \( \omega \) are normalized, so that they can vary in the range 0–1 (Figure 3). If def is measured without noise, then all the possible \( \sigma, \omega \) pairs whose product is equal to def lie on a hyperbola (Figure 2b). If the measurement of def is perturbed by Gaussian noise, then the 2-D probability density function of \( \sigma, \omega \) given def is not uniform and has a center of mass corresponding to the point of the hyperbola that is closest to the origin of the coordinate system. The center of mass of this probability density distribution, therefore, corresponds to the point \( \sigma = \omega = \sqrt{\text{def}} \). In general, if \( \sigma \) and \( \omega \) vary in the ranges \([0, \sigma_{\text{max}}]\) and \([0, \omega_{\text{max}}]\) respectively, then the posterior probability distribution of a \( \sigma, \omega \) pair given
\( \def^\) has a maximum value corresponding to:

\[
\def' = \frac{\sigma_{\text{max}}}{v_{\text{max}}} \quad \omega = \frac{\omega_{\text{max}}}{v_{\text{max}}} \quad \def^.
\]

According to our proposal, then, observers interpret first-order local optic flow by choosing the most likely 3-D solution – that is, the slant and angular velocity magnitudes that maximize the posterior probability of a \( \sigma, \omega \) pair given \( \def^\).

To test our model, we simulated the optic flow produced by the rotation of a planar random-dot surface and asked observers to judge perceived surface slant and angular rotation magnitudes [14]. Consistent with the predictions of the model, we found that perceived slant and angular rotation were an increasing function of \( \def^\) whereas the influence of simulated \( \sigma \) and \( \omega \) was negligible.

**The rigidity assumption is not biologically plausible**

One of the fundamental assumptions of both Euclidean and affine models is that 3-D objects undergo a rigid rotation unless otherwise specified by the projected image transformations [6,30,31]. We proved that this assumption (one of the fundamental constraints embedded in the algorithms deriving 3-D shape from motion) is not biologically plausible [21].

Let us consider again the wedge in Figure 1. The two planar surfaces of this wedge have different orientations (i.e. slants). If this 3-D structure is rotated rigidly with respect to the observer, then these two planar surfaces will project different deformations (Eqn 1, above). If, according to Eqn 2, the amount of perceived rotation for each surface is derived from \( \def^\) then, in general, a reliable discrimination between rigid and non-rigid motion would not be possible.

This prediction was confirmed by one of our investigations. We asked observers to discriminate rigid from non-rigid motion in stimulus conditions in which two surfaces were rotated rigidly or non-rigidly, and projected either the same or different deformations [21]. The results showed that perceived rigidity was determined by the distribution of \( \def^\) and not by simulated rigidity. By means of Eqn 2, moreover, we were also able to account for misperceptions in the orientation of axis of rotation [32], the discrimination between constant and non-constant 3-D angular velocity [33], and the segmentation of two overlapping velocity fields [34]. A further characteristic of our model, which results on distortions of perceived depth, and implications for perceived shape are discussed in Box 1.

**Concluding remarks**

Even in the absence of other depth information, optic flow is an effective depth cue from which 3-D information of the observed scene can, in principle, be derived. This article has reviewed two hypotheses regarding observers’ interpretation of optic flow. One hypothesis is that the perceptual analysis of optic flow is global and veridical according to the ‘inverse-optic’ approach, as suggested by most of the models of perceived S/M that have been proposed thus far. A second hypothesis is that observers analyze optic flow in a heuristic and patch-way fashion. According to this second hypothesis, the most likely 3-D interpretation is assigned to local (ambiguous) first-order properties of the flow and, consequently, the perceptual interpretation is, in general, not veridical, nor internally consistent. The empirical evidence indicates that both veridical judgments and biases in the local judgments of optic flow can be accounted for by this heuristic hypothesis. Rather than solving the ‘inverse-optic’ problem, therefore, the visual system seems to rely on probabilistic interpretations, perhaps derived from learning. Because the perceptual interpretation of optic flow is not veridical, the goal for future research is to understand how local information is integrated through space and time to achieve a global and coherent perceived 3-D shape [35–37].

**References**


http://tics.trends.com