Affine structure from motion

Jan J. Koenderink and Andrea J. van Doorn

Buys Ballot Laboratory, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands

Received November 1, 1989; accepted April 11, 1990

A mobile observer samples sequences of narrow-field projections of configurations in ambient space. The so-called structure-from-motion problem is to infer the structure of these spatial configurations from the sequence of projections. For rigid transformations, a unique metrical reconstruction is known to be possible from three orthographic views of four points. However, human observers seem able to obtain much shape information from a mere pair of views, as is evident in the case of binocular stereo. Moreover, human observers seem to find little use for the information provided by additional views, even though some improvement certainly occurs. The rigidity requirement in its strict form is also relaxed. We indicate how solutions of the structure-from-motion problem can be stratified in such a way that one explicitly knows at which stages various *a priori* assumptions enter and specific geometrical expertise is required. An affine stage is identified at which only smooth deformation is assumed (thus no rigidity constraint is involved) and no metrical concepts are required. This stage allows one to find the spatial configuration (modulo an affinity) from two views. The addition of metrical methods allows one to find shape from two views, modulo a relief transformation (depth scaling and shear). The addition of a third view then merely serves to settle the calibration. Results of a numerical experiment are discussed.

INTRODUCTION

When you walk around your chair, geometrical optics predicts how the image of the chair on your retinas is subject to continual severe deformation. The image is quite different from moment to moment. Yet, in your introspection, the chair remains the same trusty object of invariant shape. You would be worried if the chair were really deformed. The changing entity in your introspection is the relative position of you and your chair.

One interpretation is that your introspection is based on certain invariants of the transformations on your retinas. Such invariants are related to the shape of the chair. This idea has a venerable history, from Euclid's *Optics*[1] to Helmholtz[2] and Gibson.[3] The mathematics of the reconstruction of spatial structure from projections starts (in serious form) with Lambert's treatise on the free perspective (*Freye Perspektive*).[4] The key theorems were formulated by Pohlke[5] in 1853 and Hauck[6] in 1883. This led to lively discussions in the literature (with a decisive contribution by Kruppa[7]) that are largely obscure by now. The main application has been in photogrammetry. Only in recent times (largely the current decade) has the mathematical nature of the problem been taken up again and have major attempts to quantify perception of three-dimensional shapes been pursued.

The basic mathematical structure is simple. As the problem is usually framed, you have N views, say, of a spatial configuration consisting of M points in general position. (In fact, the problem—or problems—can be formulated in many different ways. We still stick to the present paradigm though.)

A "view" is a central projection, i.e., the points are mapped on a pencil of concurrent visual rays. Their common intersection is the center of projection or vantage point. The N vantage points are again assumed to be in general position.

"General position" means that slight perturbations of the configurations of N vantage points and M fiducial points will not lead to qualitative changes in a possible solution.

It is assumed that you can identify any given fiducial point in the different views and thus that a correspondence has been established.

It is also assumed that you have access to the full apparatus of spherical trigonometry: you may measure the angles between pairs of visual rays (apparent size) as well as the dihedral angles defined by triples of visual rays.

Finally, it is assumed that you know *a priori* that the spatial configuration of the fiducial points is rigid. That is, you may assume that the mutual distances between arbitrary pairs of fiducial points in three-dimensional space are equal in the case of all N views.

Then the problem is to find the spatial configuration of the M points (the shape) as well as the position of the vantage point relative to that configuration for the N views.

Needless to say, this problem has no solution: since you measure only angles, there can be no hope to recover distances. Thus people reformulate the problem as follows: Can you find a solution modulo a scaling factor? The answer is: Yes, sometimes, depending on the values of N and M.

In this paper we address only a slightly simplified form of the problem, namely, we restrict the discussion to orthographic projections. For this case, the so called structure-from-motion theorem states that "given three distinct orthographic views of four non-coplanar points in a rigid configuration, the structure and motion compatible with the three views are uniquely determined."[8] More information leads to an inconsistent problem, unless the motion is indeed rigid and the measurements of infinite precision (the case of redundant data); less information leads to continuous families of solutions. The equations are inconsistent even for three views of fo ur noncoplanar points if the movement fails to be rigid.

The structure-from-motion theorem is indeed almost intuitively obvious: technical drawings have shown at least three views for many years (the fact that these views are typically related in specific ways in technical drawings helps,

0740-3232/91/020377-09$05.00

of course), and any configuration of less than four points is planar anyway.

Psychophysical evidence has generally failed to show a satisfactory relationship between perception and the structure-from-motion theorem. The actual results are complicated, perhaps not quite consistent, and not easily summarized in a few sentences. Apparently, visual perception is not good with four points. It depends on the task, but typically performance increases with the number of points. On the other hand, the visual process often seems to base decisions on fewer than three views: you obtain vivid three-dimensional impressions with only two views (a good example is the case of binocular stereo), and additional views do not increase performance dramatically over the two-view case, although improvements certainly occur. Moreover, you may well doubt the ability of the visual system to handle the angular measurements with sufficient precision, and you have good reason to doubt the universal application of the rigidity hypothesis. Still, the rigidity assumption may well play a role in many cases.

For instance, you have no trouble perceiving the shape of the leaves of paper on which this paper has been printed when you bend them (bending is "rigid in the small," but not globally so[9]). Most people are of the opinion that they can gauge the shape of a cat pretty well when such an animal passes by (a deformation that is not even rigid in the small). It is only such troublesome objects as silk dresses that may look decidedly nonrigid, or even fluid. (According to romantic poetry this accounts for much of the charm.) In many of these cases static cues may be important; however, the fact that humans are able to deal with nonrigid shape from motion seems evident.

Introspection is not something to be trusted too much, of course. You would like to have some good psychophysics on spatial judgments in the absence of strict rigidity, with a variable number of views, for diverse spatial configurations, and so on. Some do exist, but not enough to constrain theoretical developments at this moment. One problem is that people do not know what to ask from the subject. You typically ask for expertise requiring metrical, Euclidean judgments (distances, angles, curvatures,...). However, it is clear that not all visual knowledge (i.e., that which promotes efficacious future action, for example, the ability to predict a future contour) needs to be of such a nature. In many cases it suffices to know aspects of mere spatial order, affine or projective structure, etc.

There have been many attempts to obtain partial information from fewer than three views. All methods based on the measurement of instantaneous angular speed are of this type (they exploit two infinitesimally close views), and binocular stereo provides another well-known example.

There has been no serious attempt to stratify the structure-from-motion problem until now.

Such a stratification should explicitly identify the stages in the algorithm where the various *a priori* assumptions are introduced and where topological, affine, or metrical expertise is required.

Such a stratification is desirable for several reasons, one being the likelihood that the visual process will be similarly structured because the various strata will naturally apply to different classes of tasks and require different levels of expertise. Another reason is plain scientific method: You understand a theory only when you explicitly see where your assumptions enter; otherwise it is little better than a (possibly useful) hat trick, even if you know it to be correct. It is the difference between scientific understanding and an engineering solution. Felix Klein's famous stratification of geometry is an example of how such a stratification increases understanding.

STRATIFICATION OF THE STRUCTURE-FROM-MOTION PROBLEM

In this paper we skip the front end of the stratification, e.g., the solution of the correspondence problem, even if this is by no means a trivial problem. We assume N views of a configuration of M points, with the correspondences established. We do not assume rigidity or the ability to perform metrical operations in the pencil of visual rays, however. We do assume the ability to perform the elementary projective operations, though (in the visual field to draw a line through two points, find the intersection of two lines, find a point not on a given line, etc.). Below we also require the ability to perform affine operations (to bisect a line segment or to draw a line through a point parallel to a given line).

Clearly, the stratification could be pushed much further than wo do in the present paper. We merely indicate some layers that appear to be especially useful in view of the given problems.

Moreover, we limit the discussion to the case of parallel projection, or the case of a restricted field of view and simultaneously restricted depth range. (This case occurs when the largest diameter of the configuration is small with respect to the distance of the vantage point to the nearest fiducial point.) This limitation makes practical sense for two reasons:

1. Parallel projection is a good approximation to central projection if the field of view is small and the depth range restricted.
2. Arbitrary smooth deformations are locally equivalent to affine transformations.

More on this below.

Please notice that a restriction to parallel projection does not remove the dependence of apparent size on distance: We will consider pairs of views of objects in which the change in overall size indicates an (overall) depth difference that may far exceed the depth of relief of the object. Such views, however, are both considered parallel projections.

Basically, we identify the following logically distinct strata:

- From two views of four fiducial points we construct a unique affine frame. If a configuration of M points ($M > 4$) is given in two views, we can assign unique affine coordinates (α, β, γ) to the points. For example, the first quadruple of points is arbitrarily assigned the coordinate representation (0, 0, 0), (1, 0, 0), (0, 1, 0), and (0, 0, 1). Then we show how the fifth and all other points can be assigned unique affine coordinates (α, β, γ). This is the affine structure-from-motion theorem. The construction requires nothing but affine methods, i.e., it requires either the ability to bisect any given line segment or the (equivalent) ability to draw a line parallel to a given line through a given point. A Euclidean metric is not required.

- The introduction of the rigidity hypothesis and the ability to measure angles (distances in the visual field) permits a further stage of computation: you obtain a one-parameter family of Euclidean solutions. You know the axis of rotation but not the turn or the magnitude of the rotation. If you assume a turn, you fix the slants and tilts and vice versa. Thus this solution from two views yields the shape modulo a depth scaling and a shear. Apart from the true depth of relief and orientation with respect to the frontoparallel plane, you know the shape.
- The introduction of a third view restricts the solution to a unique one. (Or at least to a finite number—either two or four, pairwise related through a depth inversion—of solutions.) A pair of two-view affine solutions determines the relief scaling and orientation with respect to the frontoparallel plane or, equivalently, determines the turns.

Notice that the most essential part of the solution is already obtained in the first step. For instance, given the affine representation you can predict other views (e.g., a profile from two slightly different frontal views) or the equiluminance curves for a given direction of the light. (This is intuitively evident if you remember that shadow boundaries are outlines as seen from the light source.) The second step is in effect a fairly trivial affair, and the third step merely serves to fix the turn (a single degree of freedom).

Many well-known phenomena suggest that the human visual system may stop after the first or second stage. For instance, it is often hard to distinguish a relief representation from a sculpture in the round, except when you are permitted to take extreme side views (by walking around the object; *vide* Ref. 11). In many cases human observers appear to assume tacitly that the objects are predominantly spread out in the frontoparallel plane, except for a relief effect (*vide* Ref. 12). Below we show that you may indeed determine unique minimum slant solutions from two views.

AFFINE STRUCTURE-FROM-MOTION THEOREM

Suppose that you are handed two views of four points. You can find an affine representation of the configuration in the following trivial way: pick a point (any point) and call it $\mathcal{O}$. Assign it arbitrarily the coordinates (0, 0, 0). It will be the origin of the affine frame. Pick two other points (merely take care that the triple is not collinear), and call them $\mathcal{X}$ and $\mathcal{Y}$. Assign to these points the coordinates (1, 0, 0) and (0, 1, 0). We regard $\mathcal{O}\mathcal{X}$ and $\mathcal{O}\mathcal{Y}$ as the basis vectors in the X and Y directions. Of course these vectors need not be orthogonal, nor of unit modulus, either in projection or in space. The points $\mathcal{O}$, $\mathcal{X}$, $\mathcal{Y}$ define a fiducial triangle to which all the geometrical structure can be related. You have representations of these points in the two projections; these representations are affinely equivalent. Any other point $\mathcal{P}$ in the projection can be assigned unique XY coordinates. Just write $\mathcal{O}\mathcal{P}$ as a linear combination of $\mathcal{O}\mathcal{X}$ and $\mathcal{O}\mathcal{Y}$; then the coefficients are the desired affine coordinates.

The crucial point to notice is that if the point $\mathcal{P}$ is in the plane $\mathcal{O}\mathcal{X}\mathcal{Y}$, then its affine coordinates have to be the same in both views if the spatial configuration was subjected to an arbitrary linear transformation.

Thus we do not have to assume rigidity at all, but merely assume that the transformation between views is due to a three-dimensional linear transformation. This is a much weaker assumption than rigidity, because, as mentioned above, arbitrary smooth transformations are linear in the small. (This merely states that they are well approximated by the first derivative in any sufficiently small region. The insight that arbitrary smooth transformations are affine in the small is due to Tissot.[13]) Thus the present discussion is quite general for suitably restricted fields of view and in fact merely assumes coherency, or that materials usually stick together so that their transformations tend to be smooth.

It is somewhat more of a problem to establish the third affine frame vector. This entity cannot be defined in any single view but is well defined for a pair of views. We proceed as follows: Take the remaining point and call it $\mathcal{Z}$. It is assigned coordinates (0, 0, 1). So much for the trivial part. The real problem is to find the projection of the vector $\mathcal{O}\mathcal{Z}$.

The trick to arriving at the projection of the third affine frame vector is simple: just regard the point $\mathcal{Z}$ in the first view as the degenerated projection of a line segment $\boldsymbol{\mathcal{Z}}\mathcal{Z}$, where $\boldsymbol{\mathcal{Z}}$ belongs to the plane of the fiducial triangle.

Let us expand this idea a little. You may interpret the projection of $\mathcal{Z}$ to be the projection of not just one but actually two points, namely, the point $\mathcal{Z}$ itself and the imaginary point $\boldsymbol{\mathcal{Z}}$, which is to be the projection of $\mathcal{Z}$ on the plane $\mathcal{O}\mathcal{X}\mathcal{Y}$ for the first view. We call this point the trace of $\mathcal{Z}$ on the fiducial $\mathcal{O}\mathcal{X}\mathcal{Y}$ plane. Since $\boldsymbol{\mathcal{Z}}$ is in the $\mathcal{O}\mathcal{X}\mathcal{Y}$ plane, you may find its affine coordinates in the first view and construct its corresponding position in the second view. This position of the imaginary point $\boldsymbol{\mathcal{Z}}$ in the second view will in general be distinct from the projection of the real point $\mathcal{Z}$ (since the points are in general position, the point $\mathcal{Z}$ will not be in the plane $\mathcal{O}\mathcal{X}\mathcal{Y}$). This trick enables us to construct the third affine coordinate axis.

The directed line segment defined by the projections of $\boldsymbol{\mathcal{Z}}$ and $\mathcal{Z}$ in the second view will be taken as the projection of the third affine frame vector.

In order to appreciate the utility of these operations you may consider the arbitrary point $\mathcal{P}$ again. You have already seen how to obtain the XY coordinates of $\mathcal{P}$. In order to find the third coordinate, you merely perform the same trick all over again. Consider the projection of $\mathcal{P}$ in the first view to be the projection of two distinct points, namely, $\mathcal{P}$ itself and $\boldsymbol{\mathcal{P}}$, which is the trace of $\mathcal{P}$ on the fiducial plane $\mathcal{O}\mathcal{X}\mathcal{Y}$. You may construct the projection of $\boldsymbol{\mathcal{P}}$ in the second view and observe the projection of $\mathcal{P}$ in the second view. These two projected points define a directed line segment in the second view that has to be a multiple of the third frame vector (parallel to the third frame vector). If it is not, then the assumption of the affine transformation is falsified. (That is, the transformation is not smooth on the scale considered.) If it is, then the magnitude of the line segment relative to the third frame vector is the sought-for third coordinate (the sign is obviously relevant). In this way you may assign unique affine coordinates to every extra fiducial point.

You obtain a unique solution for configurations of more than four points up to an arbitrary affine transformation. Since rotations and homotheties (isotropic scalings) do not change the shape, you obtain the shape modulo an arbitrary shear.

This is sufficient to enable you to predict outlines for arbitrary viewing directions (viewing direction specified in terms of the frame, projection predicted modulo an affine transformation), the equiluminance contours for a given direction of light source (*idem*), and so forth. Although the affine solution does not permit predictions of a metrical nature, it is a true three-dimensional entity in the sense that it allows you to predict arbitrary views. We now present a numerical example.

NUMERICAL EXPERIMENT

It is a straightforward exercise to implement the affine stage numerically. You need a routine that enables you to find the image **b** of a vector **a**, say, under an affine transformation that carries the pair of vectors $\mathbf{f}_{1,2}$ into the pair $\mathbf{g}_{1,2}$. This is a problem in linear algebra.

First we write **a** as a linear combination of $\mathbf{f}_{1,2}$, say, $\mathbf{a} = \alpha\mathbf{f}_1 + \beta\mathbf{f}_2$. This leads to a set of simultaneous linear equations for the coefficients α and β with the solution

$$\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = D^{-1}\begin{bmatrix}\mathbf{f}_2\mathbf{f}_2 & -\mathbf{f}_1\mathbf{f}_2\\ -\mathbf{f}_2\mathbf{f}_1 & \mathbf{f}_1\mathbf{f}_1\end{bmatrix}\begin{bmatrix}\mathbf{f}_1\mathbf{a}\\ \mathbf{f}_2\mathbf{a}\end{bmatrix},$$

with

$$D = (\mathbf{f}_1\mathbf{f}_1)(\mathbf{f}_2\mathbf{f}_2) - (\mathbf{f}_1\mathbf{f}_2)^2.$$

This procedure succeeds whenever the vectors $\mathbf{f}_{1,2}$ are not collinear—that is, when the determinant $D \neq 0$. Then the required image of the vector **a** is $\mathbf{b} = \alpha\mathbf{g}_1 + \beta\mathbf{g}_2$.

To find the affine frame, take three points, $\mathcal{O}$, $\mathcal{X}$, and $\mathcal{Y}$, say. Define the vectors $\mathbf{f}_{1,2}$ as the projections of the directed line segments $\mathcal{OX}$ and $\mathcal{OY}$, respectively. The vectors $\mathbf{g}_{1,2}$ in the second view are similarly defined. The vectors $\mathbf{f}_{1,2}$ and $\mathbf{g}_{1,2}$ are the projections of the first two affine frame vectors in the two views.

Find the image of $\mathbf{f}_3$ in the way described above: In the first view, the projection of $\mathbf{f}_3$ is degenerated into a point. For instance, you may pick a fourth point $\mathcal{Z}$ (in an equally arbitrary manner as you did with the first triple) and consider the projections of $\mathcal{Z}$ and of the trace (relative to the first viewing direction) Z of $\mathcal{Z}$ on the $\mathcal{OXY}$ plane. In the first view, these projections (trivially) coincide, and the projection of $Z\mathcal{Z}$ degenerates into a point. In the second view, however, the projection of the line segment $Z\mathcal{Z}$ is nondegenerate. Regard the directed line segment $Z\mathcal{Z}$ as the projection of the third frame vector $\mathbf{g}_3$ in the second view. (You may shift the vector such that its tail is at the origin $\mathcal{O}$; however, this is not essential.)

This concludes the construction of the frame.

Now suppose that you have the two projections of any point $\mathcal{P}$, say. To find its affine coordinates in the frame, you first write $\mathcal{OP}$ as a linear combination of $\mathbf{f}_{1,2}$. This yields the first two coordinates $\alpha_{\mathcal{P}}$ and $\beta_{\mathcal{P}}$. The difference of the projection of $\mathcal{OP}$ in the second view and the image (found by the method outlined above) of that line segment in the first view has to be a vector that is collinear with the projection of the third frame vector. (If it is not, the assumption that the configuration suffered an affine transformation between the two views has been falsified.) The (signed) length ratio is the third coordinate, $\gamma_{\mathcal{P}}$.

Thus you end up with an affine model of the spatial configuration. This model has the coordinate representations

$$\mathcal{O} = (0, 0, 0),$$

$$\mathcal{X} = (1, 0, 0),$$

$$\mathcal{Y} = (0, 1, 0),$$

$$\mathcal{Z} = (0, 0, 1),$$

$$\mathcal{P} = (\alpha_{\mathcal{P}}, \beta_{\mathcal{P}}, \gamma_{\mathcal{P}}).$$

This may look rather trivial at first sight (it is a rather trivial affair!), but you may add an arbitrary number of points like $\mathcal{P}$, of course. The more points you add, the less trivial the solution appears: You really have constructed a three-dimensional model of the spatial configuration modulo an arbitrary affine transformation. This model suffices to predict possible contours or equiluminance curves, for instance, surely not a minor step toward shape calculation.

This procedure has been applied to the triple of projections illustrated in Fig. 1. (The triangulated head used in these examples is due to Rydfalk.[14]) As you see, the projections differ through a magnification, a cyclorotation (rotation about the axis of projection), a translation in the plane of projection, and a rotation about an axis orthogonal to the direction of view. These components are completely different for the transitions 0–1 and 1–2 (we denote the projections 0, 1, and 2). These projections are the input to the

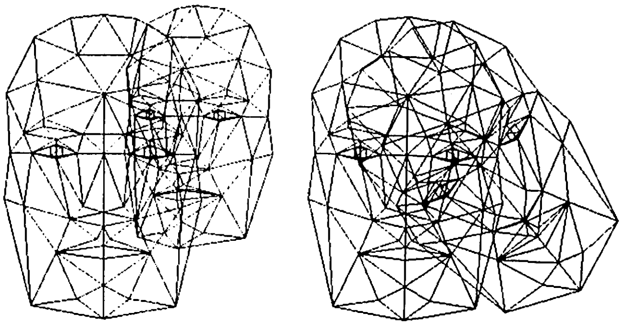

Fig. 1. Superposition of the 0th and 1st (left) and 1st and 2nd (right) views. Both figures contain one identical view (the 1st), a full frontal view of the triangulated face. The 0–1 transition is due to a rotation in space about the vertical (head shake) and a divergence; the 1–2 transition is due to a rotation about the horizontal (head nod) and a curl, or cyclorotation.

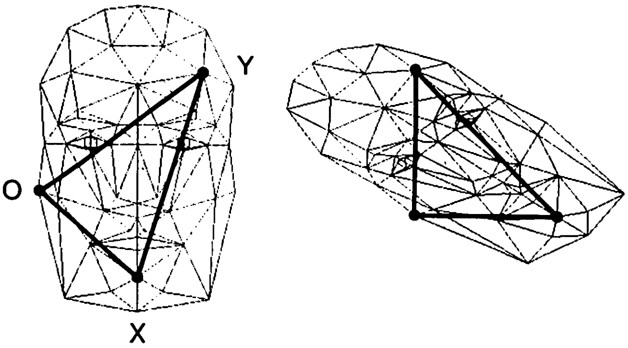

Fig. 2. The triangulated head as it appears in the 1st view with the fiducial triangle ($\mathcal{OXY}$) marked (left). The choice of fiducial points is essentially arbitrary. It is a good choice if the three points are not collinear in the projection. Notice that the fiducial triangle will be slanted and tilted with respect to the plane of projection, although its orientation cannot be calculated from any single view and is thus indeterminate. On the right an affinely equivalent view is presented. The algorithm runs on such representations in its first stage.

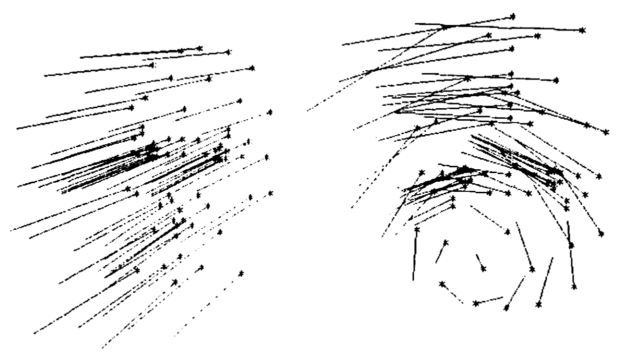

Fig. 3. Optical flow for the 0–1 (left) and 1–2 (right) transitions. Notice the strong divergence and curl.

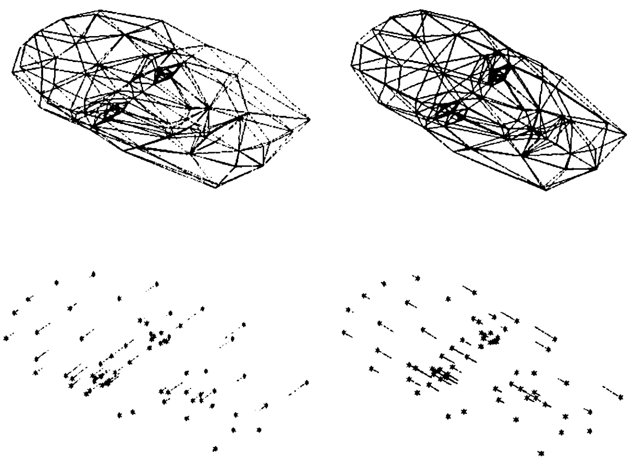

Fig. 4. Superpositions of the affine representations of the 0th and 1st and of the 1st and 2nd views. These representations differ merely through purely parallel shifts of the vertices, as is apparent from the affine flow depictions shown below.

algorithm. This is not the most general case that the algorithm can handle, which would include a general shear.

In Fig. 2 we illustrate the fiducial triangle $\mathcal{O}\mathcal{X}\mathcal{Y}$ in the 1st view (reference view). It is tilted and slanted with respect to the plane of projection, with tilt and slant being unknown to the algorithm, of course. We also illustrate an (affinely!) equivalent representation.

In Fig. 3 we illustrate the optical displacement fields for the 0–1 and 1–2 transitions. These fields show strong indications of divergences and curls.

In Fig. 4 we illustrate the affine displacement fields. These are fields of parallel displacement: in the affine coordinates the divergences and curls are automatically eliminated. This vividly illustrates a major asset of the affine method: the remapping effectively removes the effects of global curl and divergence that mess up the Euclidean flow field.

In Fig. 5 we illustrate the triad of affine frame vectors in the 0th and 2nd views. As you see, the 0–1 transition was a rotation about the vertical (a "head shake"); the 1–2 transition, about the horizontal (a "nod").

Finally, we show an affine view computed from the affine solution of the 0–1 transition. As you see, one may easily obtain a profile view from two nearly frontal views: the solution is truly three dimensional (Fig. 6). A Pearson rank-order correlation analysis[15] reveals that the affine depth order is essentially perfect. This is equally true for the 1–2 transition.

We have studied cases including nonrigid transformations. The results are equally good. Rigidity is absolutely irrelevant to the affine structure-from-motion problem.

Studies with randomly perturbed views reveal that this type of solution is rather robust.

RIGIDITY AND THE METRIC

Until now we have used only affine properties: bisection of line segments or the ability to draw a line parallel to a given line through a given point, and the ability to find the ratio of lengths of parallel line segments. Additional structure can be computed if you permit metrical concepts, e.g., the ability to bisect angles and to compare the lengths of nonparallel line segments.

In this paper we introduce metrical concepts and the rigidity hypothesis at the same stage. Note that the notion of rigidity itself depends on the metrical framework.

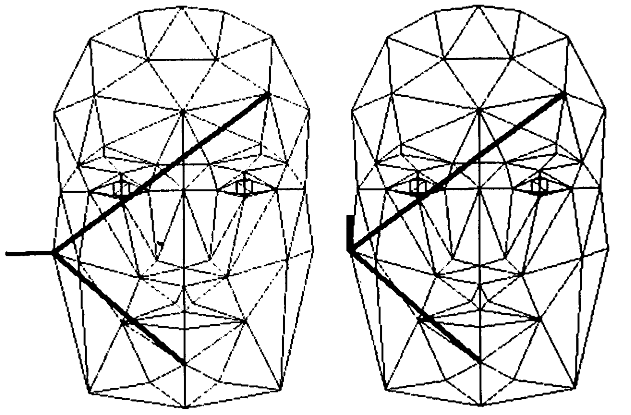

Fig. 5. Illustration of the 1st view, with the affine frame as defined in the 0th (left) and 2nd (right) views projected on it. The effects of the shake and nod of the head are immediately evident.

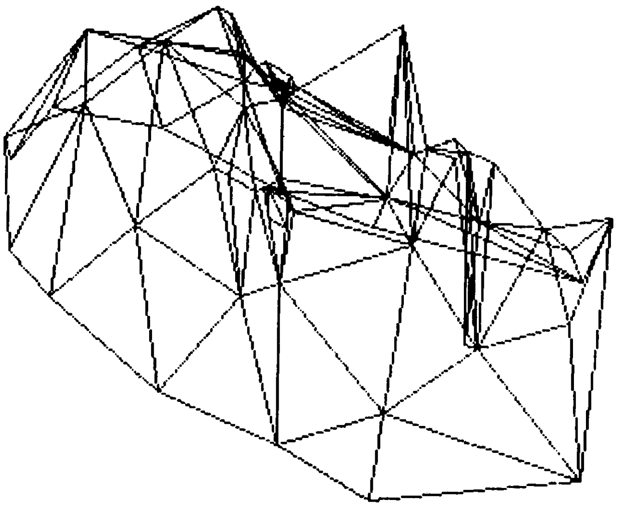

Fig. 6. Profile view computed from the affine solution using the 0–1 transition. This is affine shape-from-motion from two views, without invoking the rigidity hypothesis.

If the transformation between views is an isometry (note that the rigidity hypothesis is introduced at this point), then it must be the composition of a translation and a rotation. A translation in a frontoparallel plane merely produces a shift in the projection; you can get rid of it by putting the two projections of $\mathcal{O}$ into coincidence. This is easily done, and we will not consider this frontoparallel translational component in the sequel. Assume that it has been factored out. (This also yields an estimate of the frontoparallel translation, of course.)

The rotation can be decomposed into a rotation in the image plane and a rotation about an axis in a frontoparallel plane. It is easy enough to find the latter: you may consider the projection of the third affine frame vector to be the projection of a plane perpendicular to the axis of rotation in a frontoparallel plane. If you then construct the projection of that plane in the first view (only affine constructions are required), you may put the projections of planes of rotation in the two views into coincidence through a relative rotation of the two views. This factors out the rotation in the image plane. The procedure is easily implemented. We disregard this rotation about the viewing direction in the following paper[10] and assume it to have been factored out. As a side benefit you obtain a numerical estimate of the amount of cyclorotation. In our numerical studies (e.g., the case illustrated above) we find perfect agreement.

Because the axis of rotation in the plane of projection is known in both views, you may also correct for an overall scale difference that is due to a translation in depth. The point to notice is that the points on the axis of rotation are not changed by that rotation. Thus, if you project all points (in the projection) on this axis, you obtain a collinear sequence of points that has to be invariant. If it differs between the views, then it has to differ merely in scale, and the scaling factor can be determined through comparison of lengths in the projection. If the point sequences do not differ merely by scale, the rigidity assumption has been falsified. We find perfect estimates of magnification in our numerical experiments. Such magnifications can immediately be interpreted in terms of distance changes, of course. Below we assume correction for such overall magnification.

After these corrections the two normalized views merely differ through a rotation about an axis in a frontoparallel plane through the projection of $\mathcal{O}$. This axis is perpendicular to the projections of the plane of rotation constructed above. This type of transformation is the only component that generates depth information. As a side benefit, you have found the relative shift, the differential cyclorotation, and the size ratio that relate the two views.

Notice that this is not a trivial step: after all, the two views may differ quite a bit because of rotations about axes in frontoparallel planes. The estimation of the shift, curl, and magnitude differences is a well-recognized problem in optic flow analysis. Our algorithm yields a simple solution.

In order to proceed we need some formalism. Define a Euclidean frame $(\hat{e}_1, \hat{e}_2, \hat{e}_3)$, such that $\hat{e}_{1,2,3}$ are unit vectors, with $\hat{e}_1$ along the axis of rotation in the image plane and $\hat{e}_3$ along the line of sight.

Let $G_1\hat{e}_1 + G_2\hat{e}_2$ denote the depth gradient of the fiducial triangle $\mathcal{OXY}$; i.e., the depth of a point $\alpha\hat{e}_1 + \beta\hat{e}_2$ in the projection with respect to the frontoparallel plane through $\mathcal{O}$ is $\alpha G_1 + \beta G_2$. We also introduce the slant σ and the tilt τ of the fiducial plane through the equations $G_1 = \tan\sigma\cos\tau$ and $G_2 = \tan\sigma\sin\tau$.

Let the coordinates of the points $\mathcal{X}$ and $\mathcal{Y}$ in the projection (just disregard $\hat{e}_3$) be (X_1, X_2) and (Y_1, Y_2), respectively. Then the third coordinates must be $X_3 = G_1X_1 + G_2X_2$ and $Y_3 = G_1Y_1 + G_2Y_2$.

For a given turn ρ, say, the rotation can be represented by the matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\rho & -\sin\rho \\ 0 & \sin\rho & \cos\rho \end{bmatrix}.$$

The turn is the angle over which the object has (rigidly) turned about a frontoparallel axis from the first view to the second. Of the transformed coordinates, the first one is trivially unchanged, whereas the third one is not observable. The second coordinate is observable and yields the available information. The equations are

$$X_2{}^1 = X_2{}^0\cos\rho - \sin\rho(X_1{}^0G_1 + X_2{}^0G_2),$$

$$Y_2{}^1 = Y_2{}^0\cos\rho - \sin\rho(Y_1{}^0G_1 + Y_2{}^0G_2).$$

Here the upper indices label the views; the lower indices label the components. Because the turn ρ is unknown, we eliminate it from these equations in order to obtain a single equation in (G_1, G_2). This equation represents a one-parameter family of solutions for the two-view case. The parameter is the unknown turn ρ. The equation is quadratic in (G_1, G_2). The linear terms are absent. (Such one-parameter families of solutions consistent with rigid motion and two orthographic views have been discussed in the literature, e.g., in Refs. 16–18).

From the expression for the discriminant of this quadric, you may show that the locus of permissible points in gradient space is necessarily a hyperbola. As a consequence, you obtain a range of possible orientations of the tilt and a lower bound on the slant. There do exist two minimum-slant solutions, which are indeed unique because the hyperbola has exactly two points closest to the origin of gradient space. These minimum-slant solutions are mirror images with respect to the true frontoparallel plane. Such so-called solutions have perfect rank-order correlation, but the depth relief may be far wrong because the turn is typically way off.

The depth of a point equals

$$z = \frac{P_2{}^1\cos\rho - P_2{}^0}{\sin\rho};$$

thus the relief is indeed strongly dependent on the value of the turn. If the magnitude of the turn is small ($\rho \ll 1$), you have the simple relation

$$\rho z = P_2{}^1 - P_2{}^0,$$

i.e., the depth of relief (z) is inversely proportional to the estimated value of the turn. This simple relation is generally useful as a convenient rule of thumb.

In Fig. 7 we show the minimum-slant solution for the 0–1 transition of the example. In this case the true slant of the fiducial triangle was 27 deg, the tilt 224 deg, and the turn 20 deg. The minimum-slant solution picks a slant of 9 deg, a tilt of 168 deg, and a turn of 39 deg. (Tilt range is 93–243 deg, and slant range is 9–90 deg; thus the one-parameter family leaves the orientation of the fiducial triangle a considerable amount of leeway.) The relief is underestimated, but Pearson's rank-order correlation is perfect.

THREE VIEWS

When you obtain a third view, you are all set for a complete Euclidean solution.

From three views you can compute a pair of two-view solutions. (Actually you can compute three pairs of solutions. However, we consider the three views as part of a time series that has to be handled serially. Thus it makes sense to compare merely the 0–1 and 1–2 transitions and to disregard the 0–2 transition.) Each two-view solution represents a one-parameter family of solutions, as we showed in the previous section.

The one-parameter families of solutions for the 0–1 transition and the 1–2 transition are represented by their hyperbolic loci in gradient space. The pair of hyperbolas has either two or four intersections. (The case of no intersection occurs only in the case of nonrigid motions. If the motion is rigid, there has to be at least one solution and thus a pair of them.) These intersections represent either one or two pairs of solutions that are related through a reflection in the frontoparallel plane. In Fig. 8 we show the loci in gradient space for the example. There exists essentially a unique solution. It is indeed found to be numerically perfect, save for small rounding-error effects. In Fig. 9 we show a profile view computed from this solution; it is identical with the true profile, except for minor deviations caused by rounding errors.

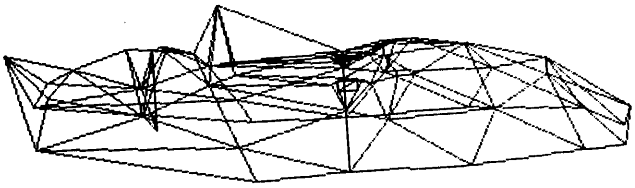

Fig. 7. Minimum-slant solution for the 0–1 transition (profile view). The relief is essentially perfectly recovered; it is merely the depth of the relief that has been misjudged (in this case underestimated; cf. Fig. 9 below).

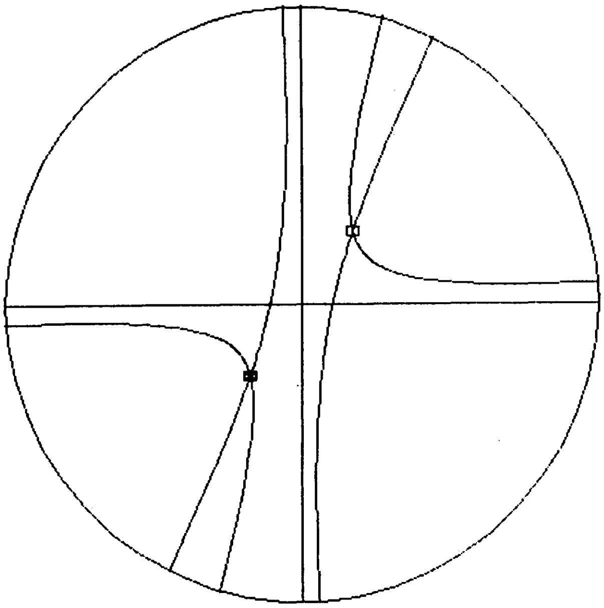

Fig. 8. Depiction of gradient space with the one-parameter families of solutions for the 0–1 and 1–2 transitions drawn in. Also marked are the true situation and the two Euclidean solutions. One solution coincides with the true shape; the other is a depth-inversed replica. Gradient space is depicted in polar coordinates. The radial coordinate is the slant (the angular value is zero at the origin and 90 deg at the boundary circle). The angular coordinate is the tilt.

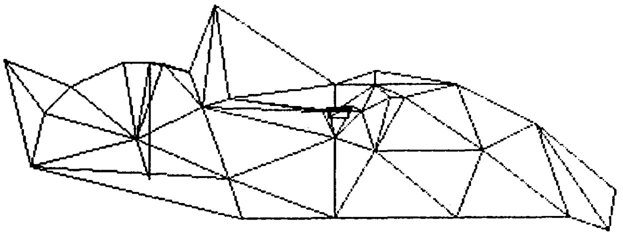

Fig. 9. Profile view computed from the Euclidean solution. It is identical to a true profile view.

This method of comparison of different pairs of views is reminiscent of taking a profile view to mensurate the orientation of the fiducial triangle and the depth of relief, a trivial affair. The essential structure-from-motion part of the solution has already been solved at the first stage. This is apparent from the perfect rank-order correlation of the initial affine solution.

CONCLUSIONS

We have shown how the structure-from-motion problem may be solved in a stratified, highly structured manner. Here we discuss some of the implications.

The first stage of the solution assumes

- A small field of view,
- A smooth transformation in three-space,
- Affine constructions in the visual field

and finds a three-dimensional model of the configuration modulo an arbitrary affine transformation.

The first two assumptions are in fact closely related: for a small enough field of view, you regard a small enough part of the configuration such that the smooth transformation is closely approximated with its first order differential. That is an affine transformation (the Cauchy–Green tensor of the kinematics of deformable media). The method itself yields a convenient check on the validity of the affine approximation: all line segments from arbitrary points to their traces on the fiducial $\mathcal{OXY}$ plane must be parallel. This is one of the interesting features of the affine representation: the affine flow for a rigid motion (or a three-dimensional affinity) is a parallel flow; thus any influences of divergences and curls that make the Cartesian flows so hard to interpret are automatically canceled. There is no need to search for the epipolar lines at all. This argument may be inverted, with the useful result that the affine flow enables you to find the epipolars in a simple manner.

If the transformation were really affine (or rigid) to begin with, the small field of view would appear unnecessarily restrictive. There is another reason for this restriction, however. The whole procedure hinges on the key observation that three points define the affine transformation of all the points coplanar with them. This permitted the crucial trace construction for the fourth point. Now consider the case of projective constructions. Given the fate of four points, you can find the fate of all points coplanar with them. At first sight this appears to open the way to generalizing our construction to the projective case (hence central projection and arbitrary fields of view). This does not work, though, for the simple reason that an arbitrary quartuple of points fails to be coplanar with probability one. On the other hand, any triple of points cannot fail to be coplanar. Thus there can be no projective generalization of this particular algorithm.

The small-field restriction means in practice that large fields have to be handled patchwise; you have to break up the scene into patches that are small enough for the first-order deformation kinematics to be applicable within the tolerances. This is necessary, anyway, if the transformation in three-space is an arbitrary smooth one.

The second stage uses two assumptions, namely, that

- The transformation in three-space is an isometry (rigidity) and
- Metrical constructions in the visual field are allowed.

Notice that, because the field of view is small, we require merely local rigidity. A bending, for instance, represents an infinitesimal isometry[9] and can be handled by the method. Then the axis of rotation in the frontoparallel plane will vary from patch to patch, of course. The differential rotation between patches reveals the nature of the bending itself. If the transformation is not even a bending, this step fails; you must content yourself with the affine part of the solution. The method itself allows a check on the rigidity hypothesis.

It is possible to reformulate the first or both the first and the second steps in terms of motion parallaxes; you need to consider only vantage points that are infinitesimally close and to introduce spatial derivatives. You then obtain the relations that have been described earlier in the literature (*vide* Ref. 16). For instance, the fact that a scaled depth solution can be obtained from two views under the assumption of rigidity has been noted in the past. The method described here can handle sets of arbitrary vantage points, however, so there is no need for the views to be taken from similar positions. The only problem that can arise is due to the fact that most real objects are opaque (there can be no correspondence between a frontal and a rear view of the object for obvious reasons).

If the viewer has additional prior information concerning the rotation (e.g., because the transformation was generated by an ego movement or, in the binocular case, if the eye-separation vector is known), then the complete solution has been reached at this second stage. Even a rough estimate of the turn suffices to fix the solution in many practical cases. The estimate merely sets the scale; it does not enter into the structure-from-motion calculation. Thus the (possibly appreciable) inaccuracy of the estimate does not affect the estimate of the shape; its effect is confined to the depth scaling and shear. This appears to be a likely strategy of the human visual system: the orientation of the fiducial triangle can often be roughly estimated on the basis of monocular cues. If such cues are weak, the system may fall back in part or completely on the minimum-slant solution. There is indeed a noticeable phenomenal regression to the real object in human depth judgments (*vide* Ref. 12).

The third stage assumes merely that the viewer has access to a third view (or at least to a sequence of pairs of views; it is likely that stable shape percepts build up over time). This stage merely serves to calibrate the depth scale. It does not enter into the structure-from-motion calculation proper. It can be regarded as a refinement of the method of taking a profile view to judge the (frontal) depth directly (in the image plane). There appears to be little doubt that human observers often use this method: it has many times been described explicitly in the technical literature on the process of sculpturing; (*vide* Ref. 11).

In this paper we have merely indicated a principle. We have also demonstrated that the principle works through numerical simulations. If you had to convert the principle into a practical (that is, robust) algorithm, much more would

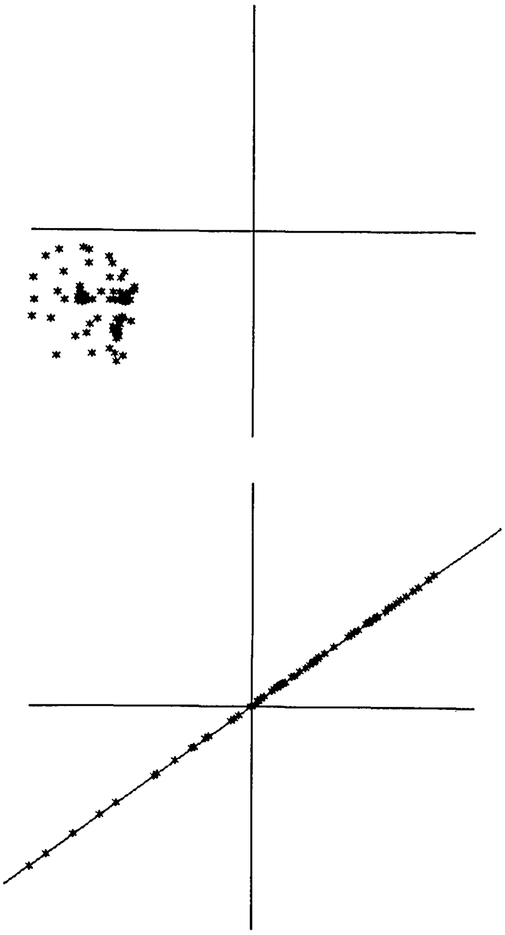

Fig. 10. Cartesian flow space (top) and affine flow space (bottom) for the 0–1 transition. Flow space is the space of flow vectors (with their tails at the origin, markers placed at their tips). The flow in Cartesian flow space is a dispersed cloud. Affine flow space, on the other hand, shows the structure: the cloud is restricted to a linear subspace. Linear regression analysis in affine flow space is indeed the core of the affine shape-from-motion algorithm.

have to be done, however. Obviously the choice of the triple $\mathcal{OXY}$ will be important (intuitively one expects a large support and a small slant to be desirable properties). You would probably not pick just any point as $\mathcal{Z}$ but rather would try to find the direction that most closely approximates the directions of $\mathcal{PP}$ for all the other points (here symbolized as $\mathcal{P}$). [This is in fact what has been done in our implementation. The flow-space representation of the flow—all flow vectors moved to the origin—for the Cartesian flow is a cloud, extended in two dimensions, but for the affine flow it is a linear subspace of flow space. In Fig. 10 we illustrate this for the 0–1 transition. We merely do a linear regression on the (highly elongated) cloud in affine flow space. This procedure is extremely stable in the presence of perturbations.] However, such considerations, important as they may be in practice, are of no importance for the basic principles involved.

ACKNOWLEDGMENTS

The research of Jan Koenderink was supported by the ESPRIT project InSight of the European Community; that of Andrea van Doorn, by the U.S. Air Force Office of Scientific Research under grant 87-0380.

REFERENCES AND NOTES

1. Euclid, *Optics* (ca. 300 B.C.); see also H. E. Burton, "The optics of Euclid," J. Opt. Soc. Am. **35**, 357–372 (1945).
2. H. von Helmholtz, *Handbuch der physiologischen Optik*, 1st ed., 1866 (3rd, posthumous, ed., Voss, Hamburg, 1909–1911).
3. J. J. Gibson, *The Perception of the Visual World* (Houghton Mifflin, Boston, Mass., 1950).
4. J. H. Lambert, *Freye Perspektive*, 2nd ed. (Heidegger, Zürich, 1774).
5. Pohlke's law (1853), as cited by E. Müller, *Vorlesungen über darstellende Geometrie* (Deuticke, Leipzig, 1923), Vol. 1.
6. G. Hauck, "Neue Konstructionen der Perspektive und Photogrammetrie," J. Reine Ang. Math. **95**, 9 (1883).
7. E. Kruppa, "Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung," Sitzungber. Akad. Wiss. Wien Math. Naturwiss. Kl. Abt. 2a **122**, 1939–1948 (1913).
8. S. Ullman, *The Interpretation of Visual Motion* (MIT Press, Cambridge, Mass., 1979).
9. J. J. Koenderink and A. J. van Doorn, "Depth and shape from differential perspective in the presence of bending deformations," J. Opt. Soc. Am. A **3**, 242–249 (1986).
10. W. A. de Grind, J. J. Koenderink, and A. J. van Doorn, "Viewing-distance and variance-of-movement detection," submitted to J. Opt. Soc. Am. A.
11. A. von Hildebrand, *Das Problem der Form in der bildenden Kunst* (Heitz and Mundel, Strassburg, 1913).
12. R. H. Thoules, "Phenomenal regression to the real object," Parts I and II, Brit. J. Psychol. **21**, 339–359; **22**, 1–30 (1931).
13. A. Tissot (1859), discussed in K. Strubecker, *Differentialgeometrie* (de Gruyter, Berlin, 1969).
14. M. Rydfalk, "Candide, a parametrized face," Rep. LiTH-ISY-I-0866 (Linköping University, Linköping, Sweden, 1987).
15. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C, The Art of Scientific Computing* (Cambridge U. Press, Cambridge, 1988).
16. J. J. Koenderink and A. J. van Doorn, "Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer," Opt. Acta **22**, 773–791 (1975).
17. B. M. Bennett, D. D. Hoffman, J. E. Nicola, and C. Prakrash, "Structure from two orthographic views of rigid motion," J. Opt. Soc. Am. A **6**, 1052–1069 (1989).
18. T. S. Huang and C. H. Lee, "Motion and structure from orthographic projections," IEEE Trans. Pattern Anal. Machine Intell. **11**, 536–540 (1989).