A Geometric Invariant For Visual Recognition and 3D Reconstruction From Two Perspective/Orthographic Views

Amnon Shashua
Artificial Intelligence Laboratory
Dept. of Brain & Cognitive Sciences
M.I.T.
Cambridge, MA 02139

Abstract

We address the problem of reconstructing 3D space in a projective framework from two views, and the problem of artificially generating novel views of the scene from two given views. We show that with the correspondences coming from four non-coplanar points in the scene and the corresponding epipoles, one can define and reconstruct (using simple linear methods) a projective invariant, referred to as *projective depth*, that can be used later to reconstruct the projective or affine structure of the scene, or directly to generate novel views of the scene. The derivation has the advantage that the viewing transformation matrix need not be recovered in the course of computations (i.e., we compute structure without motion).

1 Introduction

This paper presents a study on the geometric relation between objects and their views (perspective and orthographic) geared towards developing tools with applications to 3D reconstruction and visual recognition. For this purpose we define a new projective invariant that can be computed from image measurements across two views (four corresponding points and the epipoles) using simple linear methods. The invariant is then used for reconstructing the 3D scene in projective or affine space, and for generating novel views of the scene/object directly — without going through projective coordinates and camera transformation.

We adopt the projective framework for representing 3D space as was also done recently by [6, 13, 9]. In a projective framework the scene is represented with respect to a frame of reference of five points whose location in space are unknown and can assume arbitrary general configurations in 3D projective space [22]. This allows us to work in a framework that does not make a distinction between orthographic and perspective views and does not require internal camera calibration, i.e., the internal camera parameters are folded into the camera transformations.

Related to 3D reconstruction is the application to visual recognition. The alignment approach to recognition ([20], and references therein) is based on the notion that the geometric relation between objects and their images can be used to create an equivalence class of images of an object of interest. This approach can be realized by storing a few number of "model" views (two, for example) and with the help of corresponding points between the model views and any novel input view, the object is "re-projected" onto the novel viewing position. Recognition is achieved if the re-projected image is successfully matched against the input image. We refer to the problem of predicting a novel view from a set of model views using a limited number of corresponding points, as the problem of *re-projection*.

The problem of re-projection can in principal be dealt with via 3D reconstruction of shape and camera motion. For purposes of stability, however, it is worthwhile exploring more direct tools for achieving re-projection. Most of the current tools available for this purpose assume orthographic projection [21, 10, 16]. The method of epipolar line intersection is a possibility for achieving re-projection under perspective [3, 15, 17] but, however, is singular for certain viewing transformations. For example, numerical instabilities arise when the centers of projection of the three cameras are nearly collinear, or equivalently, when the object rotates around nearly the same axis for all views. The re-projection methods introduced in this paper is not based on an epipolar intersection, but rather is based directly on the relative structure of the object, and does not suffer from any singularities, a finding that implies greater stability in the presence of noise.

We derive a geometric invariant defined by a single cross ratio along a ray cutting through the frame of reference. We show that the invariant is equal to the third projective coordinate, and therefore refer to it as *projective depth*. The invariant can be used later to

0-8186-3692-0/93 $03.00

recover homogeneous coordinates if desired, or used directly to achieve re-projection onto a third view. The derivation has the advantage that the viewing transformation need not be recovered in the course of the computations — only the projections due to two faces of the tetrahedron of reference. The geometric construction we use requires the projections of four scene reference points onto two views, and as the fifth reference point we use the camera's center of projection via the epipoles. The epipoles are used both as a fifth corresponding pair and a means for determining correspondences due to projections of various faces of the tetrahedron of reference.

Part of this work originally appeared in [18] describing the geometric invariant and its application to re-projection, and was derived independently of [6, 13, 9]. The later stage of reconstructing homogeneous coordinates given the recovered invariant is inspired by the work of [6].

2 Projective Framework and Related Work

In a projective framework the location of an object point is measured relative to a frame of reference of five points (a tetrahedron and a unit point) whose positions in space are unknown and which are allowed to map onto any general configuration of five points in 3D projective space. It is not difficult to show [19] that the space of images we can get out of this framework are no more than perspective and orthographic images of the scene, and images of images of the scene, produced by a pin-hole camera in which the camera's coordinate frame is allowed to undergo arbitrary affine transformations in space.

The projective framework enlarges the equivalence class of images of an object compared to the metric framework, but in return does not require internal camera calibration and does not make a distinction between orthographic and perspective projections. The internal camera parameters (focal length, principal point and image coordinates scale factors) are folded into the affine transformation of the camera coordinate frame ([14], for example) and, therefore, can assume arbitrary values (which can also change from one view to another). Orthographic images are included in this framework because any of the reference points (including the COP) can be anywhere in 3D projective space. These features of the projective framework imply greater stability in the presence of noise compared to the metric framework (see [1, 5, 4, 17] for discussions on the performance of metric structure-from-motion in the presence of noise).

Projective space can be represented by homogeneous or non-homogeneous coordinates. In a non-homogeneous representation a point P is represented by three cross ratios along three axes of the tetrahedron of reference (see Figure 1). A homogeneous representation is a tetrad (x, y, z, t) of coordinates which is typically realized by assigning the standard coordinates $(0,0,0,1), (1,0,0,0), (0,1,0,0), (0,0,1,0)$ and $(1,1,1,1)$ to the vertices of the tetrahedron O, U, V, W and the unit point T, respectively (see Figure 2). For example, the points with $t = 0$ are on the plane UVW, and the projection of P via O is the point with coordinates $(x, y, z, 0)$ (i.e., orthographic projection in coordinate space). In general, any ordered set of four numbers, not all zero, determine uniquely a point in space.

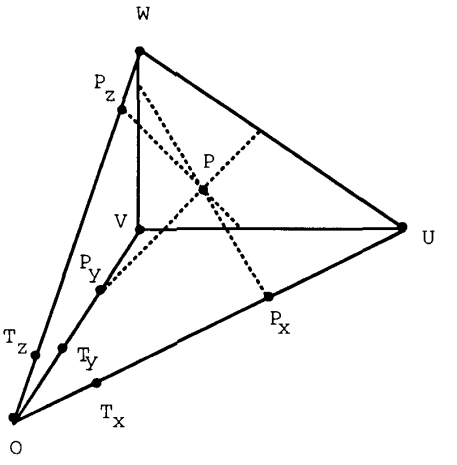

Figure 1: A non-homogeneous representation of space. The points O, U, V, W define the tetrahedron of reference. The point P_x is at the intersection of the plane PVW with the x-axis (the line OU). The point T_x is similarly constructed by replacing P with the unit point T (not shown in the drawing). The x coordinate of P is defined as the cross ratio of O, T_x, P_x, U (see [22], pp. 191).

A geometric reconstruction of non-homogeneous coordinates was recently proposed by Mohr *et al.* [13]. The authors use the projections of five scene reference points and the epipolar geometry (the "Essential" Matrix of [12] which is found by matching eight points) to determine the projections of the various stages of the construction needed to determine the three cross ratios for each point. The construction is elaborate and instead the authors propose and implement a direct non-linear algorithm for recovering the camera transformations between the scene and the two views.

Faugeras [6] proposes a linear algorithm for recover-

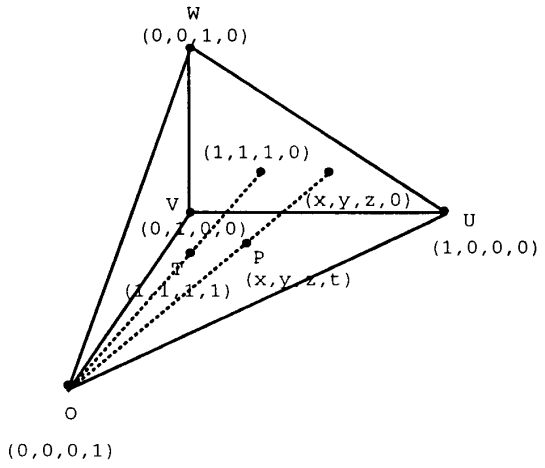

Figure 2: Homogeneous coordinates in space. If P is any point not on a face of the tetrahedron of reference, there exists four numbers x, y, z, t, all different from zero, such that the projections of P from the four vertices $(1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)$ respectively onto their opposite faces are $(0,y,z,t), (x,0,z,t), (x,y,0,t), (x,y,z,0)$ (see [22], pp. 194-195).

ing the camera transformations and the homogeneous coordinates. The projections of five scene reference points are used to determine each camera transformation matrix up to one unknown parameter (a camera transformation has 11 parameters and the correspondence between the reference points and their projections add five more unknowns, but produce 15 linear equations). The epipoles are then used as a sixth corresponding pair to fully determine (projectively speaking) the camera transformations. Once the camera transformations are recovered it becomes a simple matter to recover the homogeneous coordinates of any scene point whose projections in both views are known. Faugeras then considers the case of having four corresponding points instead of five. In that case the camera transformations are recovered up to four unknown parameters. Once these parameters are set (arbitrarily), then affine reconstruction becomes possible.

In our framework we do not recover the camera transformation matrices in order to achieve reconstruction. Instead we regard the camera's center as part of the projective reference frame making it necessary to use only four corresponding points coming from the scene. This still enables a projective reconstruction, and in addition to achieve an affine reconstruction in case the scene undergoes only affine transformations in space.

In the next section we derive the projective invariant and show how it can be computed given projections of four scene reference points (four corresponding points) and the corresponding epipoles. Section 4 describes the method by which 3D reconstruction is achieved given the recovered invariant, and shows that the invariant is equal to $1 - z$, where (x, y, z, t) are the projective coordinates of the scene. Section 5 describes two schemes for achieving re-projection, one using the invariant directly, and the other using the reconstructed structure. Section 6 briefly goes over two schemes for recovering the epipoles. Computer simulations for testing the stability of the scheme under noise were conducted and can be found in [18, 17].

3 The Projective Structure Invariant

Let the tetrahedron of reference consist of four scene points $P_1, ..., P_4$ and let the fifth reference point be the camera's COP denoted by O. Let P be an arbitrary point of interest, and consider the ray from O to P. As illustrated in Figure 3, the ray OP intersects the two faces $P_1P_2P_3$ and $P_2P_3P_4$ at $\tilde{P}$ and $\hat{P}$, respectively. We define our projective structure invariant as a cross ratio of $P, \tilde{P}, \hat{P}, O$, denoted by α_p:

$$\alpha_p =< P, \tilde{P}, \hat{P}, O >= \frac{\hat{P} - \tilde{P}}{P - \tilde{P}} \cdot \frac{P - O}{\hat{P} - O},$$

where distances are measured along the ray OP. We will use α_p for reconstructing the homogeneous coordinates (x, y, z, t) of P (and show that $\alpha_p = 1 - z$) and for re-projecting P onto novel views, but first we describe the way α_p can be computed from image measurements alone.

In the first view all points along the ray OP project onto a single point, denoted by p, in the image plane. Because internal camera parameters are folded into the affine component of camera motion, we can assign $p = (x, y, 1)$ where (x, y) are the observed image coordinates with respect to some image origin (say the geometric center of the image plane). Consider next a second view of the scene. The points $P, \tilde{P}, \hat{P}, O$ project onto generally distinct points denoted by $p', \tilde{p}', \hat{p}', v'$ which are also collinear. Because the two tetrads of points are projectively related, we have

$$\alpha_p =< P, \tilde{P}, \hat{P}, O >=< p', \tilde{p}', \hat{p}', v' >,$$

and therefore the structure invariant α_p can be computed from the projections onto the second view. The projection of O onto the second view is the epipole v', and similarly the projection of O' (the COP of the second camera position) onto the first view defines the other epipole v, and therefore v and v' are

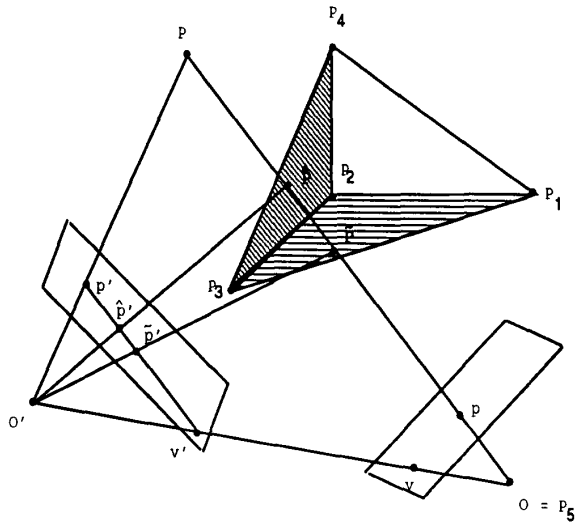

Figure 3: Projective structure of a scene point P is defined with respect to four reference points $P_1, ..., P_4$ and the center of projection O of the first camera position. The camera's center serves as the unit point in the projective frame of reference instead of a fifth scene point. The cross ratio, denoted by α_p, of the four points $P, \tilde{P}, \hat{P}, O$ uniquely fixes P with respect to the frame of reference. The cross ratio can be computed from the projections of $P, \tilde{P}, \hat{P}, O$ onto the second image plane. The projection of O is the epipole v' which can be computed from eight corresponding points [6]; the other projections $\tilde{p}', \hat{p}'$ can be recovered using the projections of the four reference points and the corresponding epipoles v, v'. Finally, since α_p is invariant it can be used for re-projection onto a third view and for reconstructing the projective structure of the scene.

corresponding points. We assume for now that the epipoles are known, and we will address the problem of finding them later (Section 6). The point p' is given to us (as we assume that correspondences between the two views has been established, as for example by [16, 17, 2]), and we can assign the coordinates $p' = (x', y', 1)$, where (x', y') are the observed image coordinates with respect to an arbitrary image origin. What is left is to recover the points $\tilde{p}'$ and $\hat{p}'$.

In order to determine $\tilde{p}'$ and $\hat{p}'$ we must recover the projective transformations due to the two faces $P_1P_2P_3$ and $P_2P_3P_4$, respectively. This can be done by identifying four coplanar points on each of the two faces, but instead we can make use of the epipoles again. For example, we can use the projections of P_1, P_2, P_3 onto both views and the corresponding epipoles to uniquely recover the 2D projective transformation A, that when applied to p will produce $\tilde{p}'$, up to a scale factor. This is expressed in the following proposition:

Proposition 1 *A projective transformation, A, which is determined from three arbitrary, non-collinear, corresponding points and the corresponding epipoles, is a projective transformation of the plane passing through the three object points which project onto the corresponding image points.*

Proof: Let $p_j \longleftrightarrow p'_j$, $j = 1, 2, 3$, be three arbitrary corresponding points, and let v and v' denote the two epipoles. First note that the four points p_j and v and the corresponding points p'_j, v' are the projections of four coplanar points in the scene. The reason is that the plane defined by the three object points P_1, P_2, P_3 intersects the line OO' connecting the two centers of projection, at a point — regular or ideal. That point projects onto both epipoles. The transformation A, therefore, is a projective transformation of the plane $P_1P_2P_3$. Note that A is uniquely determined provided that no three of the four points are collinear. □

Given the epipoles, therefore, we need just three points to determine the correspondences of all other points coplanar with the plane passing through the three corresponding object points. The transformation (collineation) A of the face $P_1P_2P_3$ is determined from the following equations:

$$Ap_j = \rho_j p'_j, \qquad j = 1, 2, 3$$
$$Av = \rho v',$$

where ρ, ρ_j are unknown scalars, and $A_{3,3} = 1$. One can eliminate ρ, ρ_j from the equations and solve for the matrix A from the three corresponding points and the corresponding epipoles. This leads to a linear system of eight equations (for more details see appendices in [14, 17]). Similarly, we can solve for the matrix E accounting for the projection of the face $P_2P_3P_4$ from the equations below:

$$Ep_j = \mu_j p'_j, \qquad j = 2, 3, 4$$
$$Ev = \mu v'.$$

If we set $\tilde{p}' = Ap$ and $\hat{p}' = Ep$ (note that $\tilde{p}'$ and $\hat{p}'$ are somewhere along the rays $O'\tilde{P}$ and $O'\hat{P}$, respectively), then the cross ratio α_p can be computed using the linear combination of rays result known in projective geometry ([8], for example) as follows: we represent p' and $\hat{p}'$ as linear combinations of v' and $\tilde{p}'$:

$$\rho p' = v' + k\tilde{p}'$$
$$\mu\hat{p}' = v' + k'\tilde{p}',$$

then $\alpha_p = \frac{k}{k'}$ (note that ρ and k are fully determined, and so are μ and k'). Note that we have made use of

the epipoles twice in our derivations. First, is because of having O as one of our reference points — this by definition brings the epipoles into the picture. Second, the epipoles were used in order to determine the image correspondences due to two faces of the tetrahedron of reference. Without the epipoles we would have needed an extra point on each face, hence loosing some generality because some of the reference points would have been coplanar. The computations for recovering α_p are simple and linear, and for convenience are summarized below:

1: Recover the transformation A that satisfies $\rho v' = Av$ and $\rho_j p'_j = Ap_j$, $j = 1,2,3$. Similarly, recover the transformation E that satisfies $\mu v' = Ev$ and $\mu_j p'_j = Ep_j$, $j = 2,3,4$.

2: Compute α_p as the cross ratio of p', Ap, Ep, v', for all points p.

One can easily see how the projective invariant can be used to re-project the scene onto a third view. Simply perform Step 1 between the first and novel view (only four corresponding points and the corresponding epipoles are required). For any fifth point p, its corresponding point p'' in the third image can be found via α_p that has been recovered from the correspondence between p and p' (three points on the epipolar line and the cross ratio uniquely determine the fourth point p''). We will discuss re-projection and 3D reconstruction in more detail later, but before doing that it may be worthwhile to consider the situation of orthographic projection.

As mentioned previously, orthographic projection does not require special treatment because the reference frame can map onto any configuration including the case where O is at infinity. Within the proposed geometric construction there are two points worth mentioning regarding the case of orthographic projection. First, the invariant α_p remains fixed under any projective transformation of the second image plane (the view on which α_p is computed). In particular the projection onto the second view can be orthographic (cross ratios are well defined for parallel rays as well). Second, consider the case when the first view is orthographic, i.e., O is at infinity. In this case α_p turns into an affine structure invariant:

$$\alpha_p =< P, \tilde{P}, \hat{P}, \infty >= \frac{\tilde{P} - \hat{P}}{P - \tilde{P}}.$$

As a result, the projective invariant is defined and recovered under both orthographic and perspective projections. Therefore, in addition to enabling the use of uncalibrated cameras, we have the property (associated with the projective framework and not to the particular algorithm we proposed) that the size of field is no longer an issue as in a metric framework [1, 5].

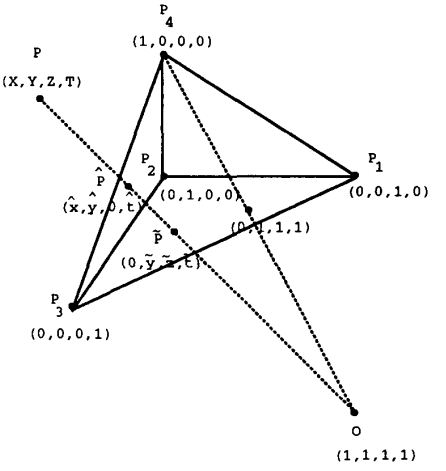

Figure 4: Reconstructing homogeneous coordinates of P (see text).

We next show how to reconstruct the homogeneous coordinate representation of the scene given that we have recovered α_p. Taken together, the central result is that we can recover projective structure without recovering the camera transforms using only four corresponding points and the corresponding epipoles.

4 Reconstructing Homogeneous Coordinates

Given the invariant structure α_p we can easily reconstruct the homogeneous coordinates (X, Y, Z, T) of any fifth object point P (its actually a sixth point overall, but its the fifth object point). We first assign the standard projective coordinates to our frame of reference as follows: the coordinates $(1,1,1,1)$ are assigned to O (the COP of the first camera position), then the coordinates $(0,0,1,0), (0,1,0,0), (0,0,0,1)$ and $(1,0,0,0)$ are assigned to the four reference points P_1, P_2, P_3 and P_4, respectively (see Figure 4).

In this choice of coordinate system we have that $\tilde{P} = (0, \tilde{y}, \tilde{z}, \tilde{t})$ and $\hat{P} = (\hat{x}, \hat{y}, 0, \hat{t})$. Note also that the projection of P_4 onto the plane $P_1P_2P_3$ is the point with coordinates $(0,1,1,1)$. In order to recover $\tilde{P}$ we map the image plane onto the plane $P_1P_2P_3$ by solving for the projective transformation B that is determined by the four following correspondences. Let $e_1, ..., e_4$ be the vectors $(0,1,0), (1,0,0), (0,0,1), (1,1,1)$. The correspondences $p_j \longleftrightarrow e_j$, $j = 1, ..., 4$, fully determine the projective transformation B, i.e., $Bp_j = \rho_j e_j$. We can therefore set the coordinates of $\tilde{P}$:

$$\tilde{P} = \begin{pmatrix} 0 \\ Bp \end{pmatrix}.$$

In a similar fashion we can recover $\hat{P}$ and with the knowledge of α_p we can determine the coordinates of P. We can also do that in a simpler way without recovering $\hat{P}$, as follows. We know that

$$\mu\hat{P} = O + s'\tilde{P},$$
$$\rho P = O + s\tilde{P},$$

and $\alpha_p = \frac{s}{s'}$. Because the third coordinate of $\hat{P}$ is always zero, we have $s' = -\frac{1}{\tilde{z}}$. Thus,

$$P = O - \frac{\alpha_p}{\tilde{z}}\tilde{P}.$$

Note that with this setup we have $\alpha_p = 1 - z$, which is the reason we referred to α_p as "projective depth". Also note that P is not determined uniquely when $\tilde{P}$ and $\hat{P}$ coincide (the ray OP intersects the line P_2P_3). In this case $\alpha_p = 0$ regardless of the position of P along the ray OP. This singularity can be easily detected ($Ap = \rho Ep$, for some scale factor ρ) and avoided by using the plane $P_1P_2P_4$ instead of $P_2P_3P_4$, for example. We have arrived at the following result:

Theorem 1 *In the case where the location of epipoles are known, then four corresponding points, coming from four non-coplanar points in space, are sufficient for computing the 3D homogeneous projective coordinates for all other points in space projecting onto corresponding points in both views. In the case the scene is undergoing an affine transformation in space, then the reconstructed scene is related to the true one by some unknown affine transformation.*

Note that the assignment of standard coordinates to the frame of reference is an arbitrary choice of representation and therefore, in the general case, the reconstructed structure is unique up to an unknown projective transformation of the scene. When the scene undergoes only affine transformations in space, then the COP can have fixed coordinates in space while allowing the remaining basis points $P_1, ..., P_4$ to have any arbitrary representation in projective space. Because the COP is part of the reference frame, it is always assigned the same coordinates regardless of the viewing position from which we choose to reconstruct the scene. Therefore, the reconstructed scene, using the algorithm described above, will be unique up to an unknown affine transformation in space, and not a general projective transformation. For convenience one can projectively transform the reconstructed coordinates (X, Y, Z, T) to $(X, Y, Z, X+Y+Z+T)$ which ensures that the fourth coordinate is non-zero.

In comparison with Faugeras' [6] results, the bottom line is the same, i.e., with four corresponding points and the corresponding epipoles we can achieve 3D reconstruction of projective or affine space. We approach the problem differently without first recovering the camera transformation matrices and instead recover first a geometric invariant α_p, which can then be used to reconstruct the homogeneous coordinates. Faugeras goes first through full reconstruction of the camera transformations using five corresponding points and the corresponding epipoles. In the case of four corresponding points (and the corresponding epipoles), Faugeras shows that the camera transformation can be recovered up to four unknown parameters. Once these parameters are set (arbitrarily) then reconstruction follows directly, and if one uses the same setting of the four parameters when reconstructing the scene from different view-points, then the reconstructions are only an affine transformation away from each other. In our case, instead of fixing four parameters in the camera transformation from the scene to the first view, we fix the coordinates of the COP by having it being part of the reference frame.

We next discuss the use of these results (the projective invariant or the reconstructed scene) for obtaining re-projection onto a third view.

5 Achieving Re-projection

Considering the two views we worked with so far as "model" views of an object of interest, we can use the projective invariant α_p or the homogeneous coordinates to re-project the object onto any novel view given a small number of corresponding points across the three views.

First, consider the use of α_p to achieve re-projection. Assume we have four corresponding points across the three views $p_j \longleftrightarrow p'_j \longleftrightarrow p''_j$, $j = 1, ..., 4$, and the epipoles v, v' between the two model views and u, u'' between the first model view and the novel view. From the correspondences $p_j \longleftrightarrow p''_j$, $j = 1, 2, 3$, and $u \longleftrightarrow u''$ we recover the collineation B, and similarly from the correspondences $p_j \longleftrightarrow p''_j$, $j = 2, 3, 4$, and $u \longleftrightarrow u''$ we recover the collineation D. Then, for any corresponding points $p \longleftrightarrow p'$, the third correspondence p'' can be recovered from the cross ratio α_p (computed from the two model views) and the three points Bp, Dp, u''.

An alternative method is to first reconstruct the homogeneous coordinates of all points of interest from the two model views (by using four corresponding points and the corresponding epipoles). We then need only six corresponding points between the first model view and the novel view in order to recover the camera transformation matrix T from the scene onto the novel view:

$$\rho_j p''_j = TP_j \qquad j = 1, ..., 6.$$

Note that we have 11 unknowns for T and 6 more unknowns for ρ_j, but we have 18 linear equations. Then,

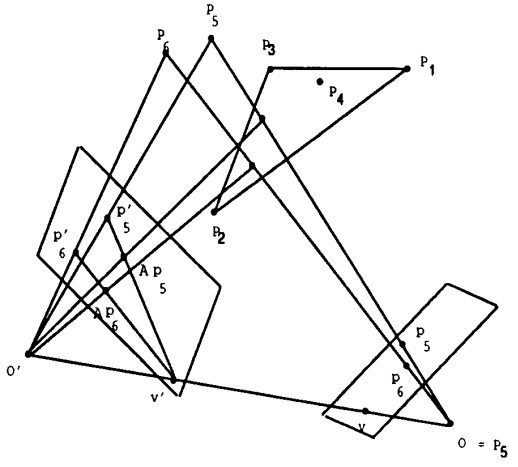

Figure 5: The geometry of locating the left epipole using two points out of the reference plane.

for any point p for which we have recovered homogeneous coordinates of the corresponding scene point P, we can recover the projection of P onto the novel view by,

$$\rho p'' = TP.$$

This method, although less direct than the previous one does not require the epipoles between the first model view and the novel view (which requires eight corresponding points), and therefore achieves re-projection with fewer corresponding points with the novel view.

For completeness we review next two methods for recovering epipoles from point correspondences between two views. Both methods are linear — one requires correspondences coming from six points, four of which are assumed to be coplanar, and the second method requires eight general correspondences.

6 Recovering the Epipoles

In general, the epipoles can be recovered from six points [11] (four of which are assumed to be coplanar), seven points (non-linear algorithm, see [7]), or eight points [6]. The basic idea behind the six point method is that the ray connecting the COP of the first camera position O and any object point P projects onto an epipolar line in the second image, and therefore the epipole can be found by intersecting two epipolar lines (see Figure 5). Given six points $P_1, ..., P_6$ where $P_1, ..., P_4$ are coplanar and P_5, P_6 are out of that plane, first recover the projective transformation A that satisfies $\rho_j p_j = Ap_j$, $j = 1, ..., 4$, then the epipoles v' and v are obtained as follows:

$$v' = (p'_5 \times Ap_5) \times (p'_6 \times Ap_6),$$
$$v = (p_5 \times A^{-1}p'_5) \times (p_6 \times A^{-1}p'_6).$$

Note that the epipoles are represented as rays with respect to the camera centers, and therefore the case of parallel epipolar lines leads to a ray parallel to the image plane (third coordinate vanishes).

The basic idea behind the eight point method [6] is that since epipolar lines in both images are projectively related, then the epipolar geometry may be represented as a 2D correlation matrix. Let F be an epipolar transformation, i.e., $Fl = \mu l'$, where $l = v \times p$ and $l' = v' \times p'$ are corresponding epipolar lines. We can rewrite the projective relation of epipolar lines using the matrix form of cross-products:

$$F(v \times p) = F[v]p = \rho l',$$

where $[v]$ is a skew symmetric matrix (and hence has rank 2). From the point/line incidence property we have that $p' \cdot l' = 0$ and therefore, $p'^t F[v]p = 0$, or $p'^t Hp = 0$ where $H = F[v]$. The matrix H is a 2D correlation (i.e., maps points onto lines) and is also known as the "essential" matrix introduced by [12], and is of rank 2. One can recover H (up to a scale factor) directly from eight corresponding points, or by using a principle components approach if more than eight points are available. Finally, it is easy to see that

$$v' = Hp_i \times Hp_j,$$

where p_i, p_j are any two points that are not on the same epipolar line (by taking more than two points we can find a least-squares fit to v'). Alternatively, as proposed in [6], we can recover v by noting that $Hv = 0$ and therefore the epipole v can be uniquely recovered (up to a scale factor). Note that the determinant of the first principle minor of H vanishes in the case where v is an ideal point, i.e., $h_{11}h_{22} - h_{12}h_{21} = 0$. In that case, the x, y components of v can be recovered (up to a scale factor) from the third row of H.

7 Computer Simulation

We ran computer simulations to test the robustness of the re-projection method under various types of noise. Instead of measuring the error due to reconstruction we measured the errors due to re-projection onto a third view. The assumption being that the performance of the system (reconstruction and re-projection) largely depends on the quality of α_p, so we may as well observe noise effects on re-projection. We tested the system using both schemes for recovering the epipoles. In general, the 8-point scheme is significantly more sensitive to noise, and in practice additional corresponding points are required to achieve

reasonable recovery of the epipoles. The experiments we describe below use the 6-point scheme for recovering the epipoles. Because the 6-point scheme requires that four of the corresponding points be projected from four coplanar points in space, it is of special interest to see how the method behaves under conditions that violate this assumption, and under noise conditions in general.

The object we used for the experiment consists of 26 points in space arranged in the following manner: 14 points are on a plane (reference plane) ortho-parallel to the image plane, and 12 points are out of the reference plane. The reference plane is located two focal lengths away from the center of projection (focal length is set to 50 units). The depth of out-of-plane points varies randomly between 10 to 25 units away from the reference plane. The x, y coordinates of all points, except the points $P_1, ..., P_6$, vary randomly between 0 — 240. The points $P_1, ..., P_6$ have x, y coordinates that place these points all around the object (clustering these points together will inevitably contribute to instability).

We applied the following camera motion: The first view is simply a perspective projection of the object. The second view is a result of rotating the object around the point $(128, 128, 100)$ with an axis of rotation described by the unit vector $(0.14, 0.7, 0.7)$ by an angle of 29 degrees, followed by a perspective projection (note that rotation about a point in space is equivalent to rotation about the center of projection followed by translation). The third (novel) view is constructed in a similar manner with a rotation around the unit vector $(0.7, 0.7, 0.14)$ by an angle of 17 degrees.

We conducted three types of experiments. The first experiment tested the stability under the situation where $P_1, ..., P_4$ are non-coplanar object points. The second experiment tested stability under random noise added to all image points in all views, and the third experiment tested stability under the situation that less noise is added to the six points, than to other points.

7.1 Testing Deviation from Coplanarity

In this experiment we investigated the effect of translating P_1 along the optical axis (of the first camera position) from its initial position on the reference plane ($z = 100$) to the farthest depth position ($z = 125$), in increments of one unit at a time. The experiment was conducted using several objects of the type described above (the six points were fixed, the remaining points were assigned random positions in space in different trials), undergoing the same motion described above. The effect of depth translation to the level $z = 125$ on the location of p_1 is a shift of 0.93 pixels, on p_1' is 1.58 pixels, and on the location of p_1'' is 3.26 pixels. Depth translation is therefore equivalent to perturbing the location of the projections of P_1 by various degrees (depending on the 3D motion parameters).

Figure 6 shows the average pixel error in re-projection over the entire range of depth translation. The average pixel error was measured as the average of deviations from the re-projected point to the actual location of the corresponding point in the novel view, taken over all points. Figure 6 also displays the result of re-projection for the case where P_1 is at $z = 125$. The average error is 1.31, and the maximal error (the point with the most deviation) is 7.1 pixels. The alignment between the re-projected image and the novel image is, for the most part, fairly accurate.

7.2 Situation of Random Noise to all Image Locations

We next add random noise to all image points in all three views (P_1 is set back to the reference plane). This experiment was done repeatedly over various degrees of noise and over several objects. The results shown here have noise levels between 0–1 pixels randomly added to the x and y coordinates separately. The maximal perturbation is therefore $\sqrt{2}$, and because the direction of perturbation is random, the maximal error in relative location is double, i.e., 2.8 pixels. Figure 7 shows the average pixel errors over 10 trials (one particular object, the same camera motion as before). The average error fluctuates around 1.6 pixels. Also shown is the result of re-projection on a typical trial with average error of 1.05 pixels, and maximal error of 5.41 pixels. The match between the re-projected image and the novel image is relatively good considering the amount of noise added.

7.3 Random Noise Case 2

A more realistic situation occurs when the magnitude of noise associated with the six points used for setting the construction (epipoles and projections of the tetrahedron of reference) is much lower than the noise associated with other points, for the reason that we are interested in tracking points of interest that are often associated with distinct intensity structure (such as the tip of the eye in a picture of a face). Correlation methods, for instance, are known to perform much better on such locations, than on areas having smooth intensity change, or areas where the change in intensity is one-dimensional. We therefore applied a level of 0–0.3 perturbation to the x and y coordinates of the six points, and a level of 0–1 to all other points (as before). The results are shown in Figure 8. The average pixel error over 10 trials fluctuates around 0.5 pixels, and the re-projection shown for a typical trial

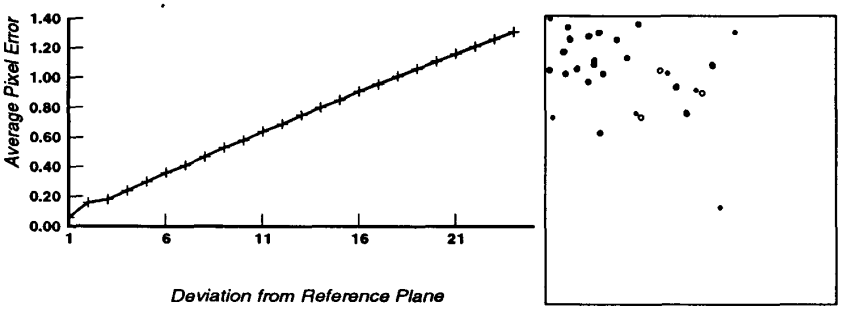

Figure 6: Deviation from coplanarity: average pixel error due to translation of P_1 along the optical axis from $z = 100$ to $z = 125$, by increments of one unit. The result of re-projection (overlay of re-projected image and novel image) for the case $z = 125$. The average error is 1.31 and the maximal error is 7.1.

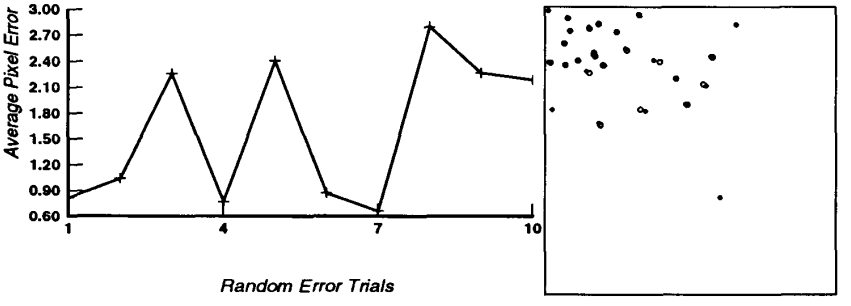

Figure 7: Random noise added to all image points, over all views, for 10 trials. Average pixel error fluctuates around 1.6 pixels. The result of re-projection on a typical trial with average error of 1.05 pixels, and maximal error of 5.41 pixels.

(average error 0.52, maximal error 1.61) is in relatively good correspondence with the novel view. With larger perturbations at a range of 0–2, the algorithm behaves proportionally well, i.e., the average error over 10 trials is 1.37.

8 Summary

We have described new techniques for two related problems: the problem of recovering structure from point matches, and the problem of visual recognition via alignment (the problem of re-projection). Our approach was based on recovering a geometric projective invariant that can then be used for both purposes: reconstruction and re-projection.

The key distinct features of our approach is, first, the definition of a new structural description (projective depth) which drives the applications of reconstruction and re-projection. Second, is the role played by the center of projection and the epipoles. Thirdly, shape reconstruction and re-projection are achieved without going through the computations of the camera transformation matrices (e.g., structure without motion). The overall features of the approach (shared with [6, 13, 9]) is that the system treats orthographic and perspective projections alike, and internal camera parameters are folded into the projection matrices, thereby allowing for views to be taken by uncalibrated cameras.

The projective depth invariant was recovered from four point matches arising from the projections of four non-coplanar object points, and the epipoles. The epipoles played a double role: first, the corresponding epipoles served as the projection of a fifth point in space, thereby allowing us to have a projective frame of reference while observing only four point matches from the scene. Second, with the epipoles we could determine the projections of various faces of the tetrahedron of reference — a task that otherwise would have required observing point matches coming from four coplanar points on each face. We then described two applications for which the invariant can be used for. First, we have shown that with the invariant we

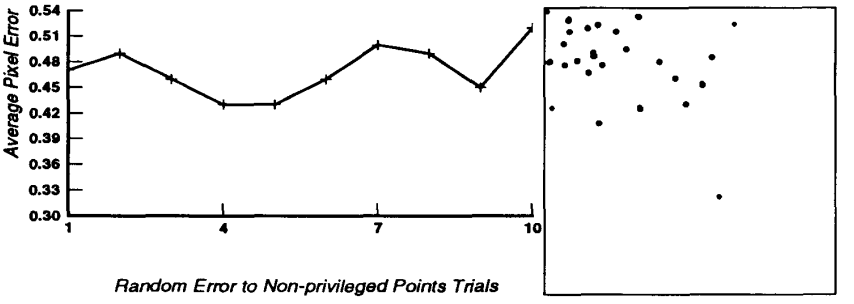

Figure 8: Random noise added to non-privileged image points, over all views, for 10 trials. Average pixel error fluctuates around 0.5 pixels. The result of re-projection on a typical trial with average error of 0.52 pixels, and maximal error of 1.61 pixels.

can achieve projective or affine reconstruction of the scene. Second, the invariant was shown to be equal to $1 - z$, where z is the third homogeneous coordinate of space. Thirdly, re-projection onto a third view was shown possible using the invariant directly without going through an explicit reconstruction of projective structure.

The algorithms for reconstruction requires eight corresponding points, or six assuming four of them are coming from coplanar points in the scene. For re-projection, the result is that the more we recover about the scene and the camera transformation the less point matches are needed. We have seen that if projective structure is recovered, then only six point matches with the novel view are required for linear re-projection (via recovery of the camera transform matrix). If the projective invariant is used instead, then eight point matches are required.

We have shown that the invariant is complete in terms of shape description, that is, one can reconstruct the projective coordinates of space without further information from the images. Representing structure in terms of a single invariant number (like depth measurements in metric reconstructions) carries certain advantages over representations in terms of invariant coordinates (a tetrad of coordinates per point). For example, in a number of algorithms that manipulate depth measurements, such as the use of Kalman filters for reconstruction over sequences of views, one can simply replace "depth" by "projective depth" without changing much the basic structure of the algorithm. A compact shape descriptor is also likely to be less sensitive in unstable situations, like in the case of reconstruction or re-projection when the base-line is relatively small. For example, the numerator and denominator of the invariant become proportionally small for small motions of the camera, implying a relatively stable situation. These cases and other uses of the invariant are planned for future research.

Acknowledgments

Part of this work was supported by NSF grant IRI-8900267. The author is currently supported by a McDonnel-Pew postdoctoral fellowship from the department of Brain and Cognitive Sciences, MIT.

References

[1] G. Adiv. Inherent ambiguities in recovering 3D motion and structure from a noisy flow field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11(5):477–489, 1989.

[2] I.A. Bachelder and S. Ullman. Contour matching using local affine transformations. In *Proceedings Image Understanding Workshop*. Morgan Kaufmann, San Mateo, CA, 1992.

[3] E.B. Barret, M.H. Brill, N.N. Haag, and P.M. Pyton. General methods for determining projective invariants in imagery. *Computer Vision, Graphics, and Image Processing*, 53:46–65, 1991.

[4] T. Broida, S. Chandrashekhar, and R. Chellapa. recursive 3-d motion estimation from a monocular image sequence. *IEEE Transactions on Aerospace and Electronic Systems*, 26:639–656, 1990.

[5] R. Dutta and M.A. Synder. Robustness of correspondence based structure from motion. In *Proceedings of the International Conference on Computer Vision*, pages 106–110, Osaka, Japan, December 1990.

[6] O.D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *Proceedings of the European Conference on Computer Vision*, pages 563–578, Santa Margherita Ligure, Italy, June 1992.

[7] O.D. Faugeras and S. Maybank. Motion from point matches: Multiplicity of solutions. *International Journal of Computer Vision*, 4:225–246, 1990.

[8] D. Gans. *Transformations and Geometries.* Appleton-Century-Crofts, New York, 1969.

[9] R. Hartley, R. Gupta, and Tom Chang. Stereo from uncalibrated cameras. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 761–764, Champaign, IL., June 1992.

[10] J.J. Koenderink and A.J. Van Doorn. Affine structure from motion. *Journal of the Optical Society of America*, 8:377–385, 1991.

[11] C.H. Lee. Structure and motion from two perspective views via planar patch. In *Proceedings of the International Conference on Computer Vision*, pages 158–164, Tampa, FL, December 1988.

[12] H.C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.

[13] R. Mohr, L. Quan, F. Veillon, and B. Boufama. Relative 3D reconstruction using multiple uncalibrated images. Technical Report RT 84-IMAG, LIFIA — IRIMAG, France, June 1992.

[14] J. Mundy and A. Zisserman. Appendix — projective geometry for machine vision. In J. Mundy and A. Zisserman, editors, *Geometric invariances in computer vision.* MIT Press, Cambridge, 1992.

[15] J.L. Mundy, R.P. Welty, M.H. Brill, P.M. Payton, and E.B. Barret. 3-D model alignment without computing pose. In *Proceedings Image Understanding Workshop*, pages 727–735. Morgan Kaufmann, San Mateo, CA, January 1992.

[16] A. Shashua. Correspondence and affine shape from two orthographic views: Motion and Recognition. A.I. Memo No. 1327, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, December 1991.

[17] A. Shashua. *Geometry and Photometry in 3D visual recognition.* PhD thesis, M.I.T Artificial Intelligence Laboratory, AI-TR-1401, November 1992.

[18] A. Shashua. Projective structure from two uncalibrated images: structure from motion and recognition. A.I. Memo No. 1363, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, September 1992.

[19] A. Shashua. Algebraic functions of image coordinates across three perspective/orthographic views. A.I. Memo No. 1405, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1993.

[20] S. Ullman. Aligning pictorial descriptions: an approach to object recognition. *Cognition*, 32:193–254, 1989. Also: in MIT AI Memo 931, Dec. 1986.

[21] S. Ullman and R. Basri. Recognition by linear combination of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13:992—1006, 1991. Also in M.I.T AI Memo 1052, 1989.

[22] O. Veblen and J.W. Young. *Projective Geometry, Vol. 1.* Ginn and Company, 1910.