MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
WHITAKER COLLEGE

A.I. Memo No. 1489
C.B.C.L. Paper No. 97

June, 1994

Relative Affine Structure: Canonical Model for 3D from 2D Geometry and Applications

Amnon Shashua and Nassir Navab

Abstract

We propose an affine framework for perspective views, captured by a single extremely simple equation based on a viewer-centered invariant we call *relative affine structure*. Via a number of corollaries of our main results we show that our framework unifies previous work — including Euclidean, projective and affine — in a natural and simple way, and introduces new, extremely simple, algorithms for the tasks of reconstruction from multiple views, recognition by alignment, and certain image coding applications.

Short version of this manuscript appears in the Proceedings of CVPR'94, Seattle, WA.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and at the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research is sponsored by grants from the Office of Naval Research under contracts N00014-92-J-1879 and N00014-93-1-0385; and by a grant from the National Science Foundation under contract ASC-9217041 (this award includes funds from ARPA provided under the HPCC program). Additional support is provided by the North Atlantic Treaty Organization, ATR Audio and Visual Perception Research Laboratories, Mitsubishi Electric Corporation, Sumitomo Metal Industries, and Siemens AG. Support for the A.I. Laboratory's artificial intelligence research is provided by ARPA contract N00014-91-J-4038. A. Shashua is supported by a McDonnell-Pew postdoctoral fellowship from the department of Brain and Cognitive Sciences. N. Navab is supported by a Nichidai Foundation fellowship provided by the MIT Media Laboratory.

1 Introduction

The geometric relation between 3D objects and their views is a key component for various applications in computer vision, image coding, and animation. For example, the change in the 2D projection of a moving 3D object is a source of information for 3D reconstruction, and for visual recognition applications — in the former case the retinal changes produce the cues for 3D recovery, and in the latter case the retinal changes provide the cues for factoring out the effects of changing viewing positions on the recognition process.

The introduction of affine and projective tools into the field of computer vision have brought increased activity in the fields of structure from motion and recognition in the recent few years. The emerging realization is that non-metric information, although weaker than the information provided by depth maps and rigid camera geometries, is nonetheless useful in the sense that the framework may provide simpler algorithms, camera calibration is not required, more freedom in picture-taking is allowed — such as taking pictures of pictures of objects — and there is no need to make a distinction between orthographic and perspective projections. The list of contributions to this framework include (though not intended to be complete) [17, 2, 30, 12, 46, 47, 13, 26, 7, 32, 34, 36, 25, 45, 29, 8, 10, 23, 31, 16, 15, 48] — and relevant to this paper are the work described in [17, 7, 13, 34, 36].

The material introduced so far in the literature, concerning 3D geometry from multiple views, focuses on the projective framework [7, 13, 36], or the affine framework. The latter requires either assuming parallel projection (cf. [17, 46, 45, 30]), or certain apriori assumptions on object structure (for determining the location of the plane at infinity [7, 28]), or assuming purely translational camera motion [24] (see also later in the text).

In this paper, we propose a unified framework that includes by generalization and specialization the Euclidean, projective and affine frameworks. The framework, we call "relative affine", gives rise to an equation that captures most of the spectrum of previous results related to 3D-from-2D geometry, and introduces new, extremely simple, algorithms for the tasks of reconstruction from multiple views, recognition by alignment, and certain image coding applications. For example, previous results in these areas — such as affine structure from orthographic views, projective structure from perspective views, the use of the plane at infinity for reconstruction (obtaining affine structure from perspective views), epipolar-geometry related results, reconstruction under restricted camera motion (the case of pure translation) — are often reduced to a single-line proof under the new framework (see Corollaries 1 to 6).

The basic idea is to choose a representation of projective space in which an arbitrarily chosen reference plane becomes the plane at infinity. We then show that under general, uncalibrated, camera motion, the resulting new representations can be described by an element of the affine group applied to the initial representation. As a result, we obtain an affine invariant, we call *relative affine structure*, relative to the initial representation. Via several corollaries of this basic result we show, among other things, that the invariant is a generalization of the affine structure under parallel projection [17] and is a specialization of the projective structure (projective structure can be described as a ratio of two relative affine structures). Furthermore, in computational terms the relative affine result requires fewer corresponding points and fewer calculations than the projective framework, and is the only next general framework after projective when working with perspective views. Parts of this work, as it evolved, have been presented in the meetings found in [33, 38], and in [27].

2 Notation

We consider object space to be the three-dimensional projective space $\mathcal{P}^3$, and image space to be the two-dimensional projective space $\mathcal{P}^2$. An object (or scene) is modeled by a set of points and let $\psi_i \subset \mathcal{P}^2$ denote views (arbitrary), indexed by i, of the object. Given two views with projection centers $O, O' \in \mathcal{P}^3$, respectively, the epipoles are defined as the intersection of the line $\overline{OO'}$ with both image planes. A set of numbers defined up to scale are enclosed by brackets, a set of numbers enclosed by parentheses define a vector in the usual way. Because the image plane is finite, we can assign, without loss of generality, the value 1 as the third homogeneous coordinate to every *observed* image point. That is, if (x, y) are the observed image coordinates of some point (with respect to some arbitrary origin — say the geometric center of the image), then $p = [x, y, 1]$ denotes the homogeneous coordinates of the image plane. When only two views ψ_o, ψ_1 are discussed, then points in ψ_o are denoted by p, their corresponding points in ψ_1 are denoted by p', and the epipoles are $v \in \psi_o$ and $v' \in \psi_1$. When multiple views are considered, then appropriate indecis are added as explained later in the text. The symbol $\cong$ denotes equality up to a scale, GL_n stands for the group of $n \times n$ matrices, and PGL_n is the group defined up to a scale.

A camera coordinate system is an Euclidean frame describing the actual internal geometry of the camera (position of the image plane relative to the camera center). If $p = (x, y, 1)^\top$ is a point in the observed coordinate representation, then $M^{-1}p$ represents the camera coordinates, where M is an upper-diagonal matrix containing the internal parameters of the camera. When M is known, the camera is said to be internally calibrated, and when $M = I$ the camera is in "standard" calibration mode. The material presented in this paper does not require further details of internal calibration — such as its decomposition into the components of principle point, image plane aspect ratios and skew — only the mere existence of M is required for the remaining of this paper.

3 Relative Affine Structure

The following theorems and corollaries introduce our main results which are then followed by explanatory text.

Theorem 1 (Relative Affine Structure [33]) *Let π be some arbitrary plane and let $P_j \in \pi$, $j = 1, 2, 3$ projecting onto p_j, p'_j in views ψ_o, ψ_1, respectively. Let*

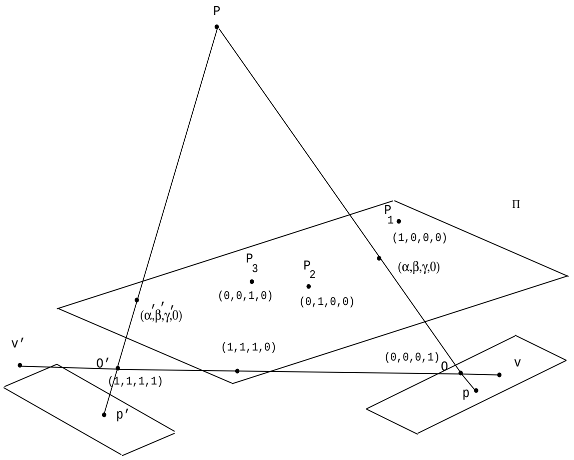


Figure 1: See proof of Theorem 1.

$p_o \in \psi_o$ and $p'_o \in \psi_1$ be projections of $P_o \notin \pi$. Let $A \in PGL_3$ be a homography of $\mathcal{P}^2$ determined by the equations $Ap_j \cong p'_j$, $j = 1,2,3$, and $Av \cong v'$, scaled to satisfy the equation $p'_o \cong Ap_o + v'$. Then, for any point $P \in \mathcal{P}^3$ projecting onto $p \in \psi_o$ and $p' \in \psi_1$, we have

$$\boxed{p' \cong Ap + k\boldsymbol{v}'} \tag{1}$$

The coefficient $k = k(p)$ is independent of ψ_1, i.e., is invariant to the choice of the second view, and the coordinates of P are $[x, y, 1, k]$.

Proof. We assign the coordinates $(1,0,0,0), (0,1,0,0), (0,0,1,0)$ to P_1, P_2, P_3, respectively. Let O and O' be the projection centers associated with the views ψ_o and ψ_1, respectively, and let their coordinates be $(0,0,0,1), (1,1,1,1)$, respectively (see Figure 1). This choice of representation is always possible because the two cameras are part of $\mathcal{P}^3$. By construction, the point of intersection of the line $\overline{OO'}$ with π has the coordinates $(1,1,1,0)$.

Let P be some object point projecting onto p, p'. The line $\overline{OP}$ intersects π at the point $(\alpha, \beta, \gamma, 0)$. The coordinates α, β, γ can be recovered by projecting the image plane onto π, as follows. Given the epipoles $v \in \psi_o$ and $v' \in \psi_1$, we have by our choice of coordinates that p_1, p_2, p_3 and v are projectively (in $\mathcal{P}^2$) mapped onto $e_1 = (1,0,0), e_2 = (0,1,0), e_3 = (0,0,1)$ and $e_4 = (1,1,1)$, respectively. Therefore, there exists a unique element $A_1 \in PGL_3$ that satisfies $A_1 p_j \cong e_j$, $j = 1,2,3$, and $A_1 v = e_4$. Note that we have made a choice of scale by setting $A_1 v$ to e_4, this is simply for convenience as will be clear later on. Let $A_1 p = (\alpha, \beta, \gamma)$.

Similarly, the line $\overline{O'P}$ intersects π at $(\alpha', \beta', \gamma', 0)$. Let $A_2 \in PGL_3$ be defined by $A_2 p'_j \cong e_j$, $j = 1,2,3$, and $A_2 v' = e_4$. Let $A_2 p' = (\alpha', \beta', \gamma')$. Since P can be described as a linear combination of two points along each of the lines $\overline{OP}$, and $\overline{O'P}$, we have the following equation:

$$P \cong \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ 0 \end{pmatrix} - k \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \mu \begin{pmatrix} \alpha' \\ \beta' \\ \gamma' \\ 0 \end{pmatrix} - s \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

from which it readily follows that $k = s$ (i.e., the transformation between the two representations of $\mathcal{P}^3$ is affine). Note that since only ratios of coordinates are significant in $\mathcal{P}^n$, k is determined up to a uniform scale, and any point $P_o \notin \pi$ can be used to set a mutual scale for all views — by setting an appropriate scale for A, for example. The value of k can easily be determined from image measurements as follows: we have

$$\mu \begin{pmatrix} \alpha' \\ \beta' \\ \gamma' \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} + k \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Multiply both sides by A_2^{-1} to obtain $\mu p' = Ap + k\boldsymbol{v}'$, where $A = A_2^{-1}A_1$. Note that $A \in PGL_3$ is a homography between the two image planes, due to π, determined by $p'_j \cong Ap_j$, $j = 1,2,3$, and $Av \cong v'$ (therefore, can be recovered directly without going through A_1, A_2). Similar proofs that a homography of a plane can be recovered from three points and the epipoles are found in [34, 29]. Since k is determined up to a uniform scale, we need a fourth correspondence p_o, p'_o, and let A, or v', be scaled such that $p'_o \cong Ap_o + v'$. Finally, $[x, y, 1, k]$ are the homogeneous coordinates representation of P, and the 3×4 matrix $[A, v']$ is a camera transformation matrix between the two views. □

Theorem 2 (Further Algebraic Aspects [27]) *Let the coordinate transform from $P = zp$ to $P' = z'p'$ be described by $P' = M'RM^{-1}P + M'T$, where R, T are the rotational and translational parameters of the relative camera displacement, and M, M' are the internal camera parameters. Given A, π, k defined in Theorem 1, let n be the unit normal to the plane π, and d_π the (perpendicular) distance of the origin to π, both in the first camera coordinate frame. Then,*

$$A \cong M'(R + \frac{Tn^\top}{d_\pi})M^{-1}, \tag{2}$$

and

$$k = \frac{z}{z_o}\alpha,$$

where z_o is the depth of $P_o \notin \pi$, and $\alpha = \alpha(p)$ is the affine structure of P in the case of parallel projection (the ratio of perpendicular distances of P and P_o from π).

Proof. Let $\tilde{P}$ be at the intersection of the ray $\overline{OP}$ with π. Then $\tilde{P}' = M'RM^{-1}\tilde{P} + M'T$. Since $n^\top M^{-1}\tilde{P} = d_\pi$, we have: $\tilde{P}' = M'(R + \frac{Tn^\top}{d_\pi})M^{-1}\tilde{P}$. Since the term in parentheses describes the homography due to π, we have $A \cong M'(R + \frac{Tn^\top}{d_\pi})M^{-1}$ — which is the generalization of the classical motion of planes in the calibrated case [9, 43]. For the point P we have:

$$\begin{aligned} \frac{z'}{z}p' &= M'RMp + \frac{1}{z}M'T \\ &= Ap + \left[\frac{1}{z} - \frac{n^\top(M^{-1}p)}{d_\pi}\right] M'T \\ &\cong Ap + \left[\frac{d_\pi - n^\top(M^{-1}P)}{zd_\pi}\right] v'. \end{aligned}$$

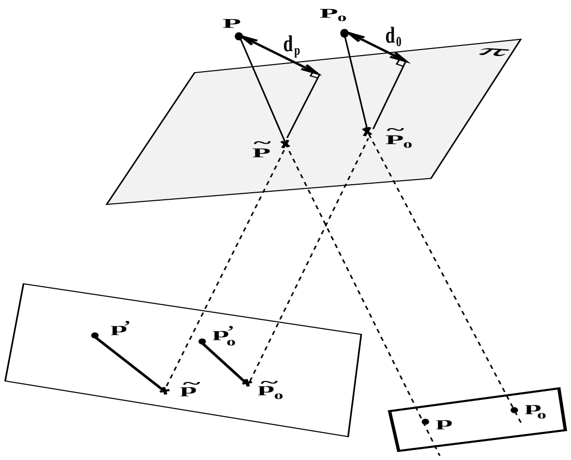


Figure 2: Affine structure under parallel projection is d_p/d_o. This can be seen from the similarity of trapezoids followed by the similarity of triangles: $\frac{p'-\tilde{p}}{p'_o-\tilde{p}'_o} = \frac{P'-\tilde{P}}{P'_o-\tilde{P}'_o} = \frac{d_p}{d_o}$.

Let $d_p = d_\pi - n^\top(M^{-1}P)$ the (perpendicular) distance from P to π. We thus have

$$k = \frac{z_o}{z}\frac{d_p}{d_o},$$

where d_o is the (perpendicular) distance of P_o from π (see Figure 3-a). Finally, note that the ratio $\alpha = d_p/d_o$ of the distances of P and P_o from π is the affine structure when the projection is parallel (see Figure 2). □

Corollary 1 *Relative affine structure k approaches affine structure under parallel projection when O goes to infinity, i.e., $k \longrightarrow \alpha$ when $O \longrightarrow \infty$.*

Proof. When $O \longrightarrow \infty$, then $z, z_o \longrightarrow \infty$. Thus $k = \frac{z_o}{z}\frac{d_p}{d_o} \longrightarrow \frac{d_p}{d_o}$ (see Figure 2). □

Corollary 2 *When the plane π is at infinity (with respect to the camera coordinate frame), then relative affine structure k is affine structure under perspective $k = z_o/z$, $A = M'RM^{-1}$, and, if in addition, the cameras are internally calibrated as $M = M' = I$, then $A = R$.*

Proof. When π is at infinity, then $d_p, d_o \longrightarrow \infty$. Thus $k = \frac{z_o}{z}\frac{d_p}{d_o} \longrightarrow \frac{z_o}{z}$. Also, $d_\pi \longrightarrow \infty$, thus $A \longrightarrow M'RM^{-1}$. (see Figure 3-b)□

Corollary 3 (Pure Translation) *In the case of pure translational motion of the camera, and when the internal camera parameters remain fixed, i.e., $M = M'$, then the selection of the identity homography $A = I$ (in Equation 1) leads to an affine reconstruction of the scene (i.e., the identity matrix is the homography due to the plane at infinity). In other words, the scalar k in*

$$p' \cong p + k\boldsymbol{v}'$$

is invariant under all subsequent camera motions that leave the internal parameters unchanged and consist of only translation of the camera center. The coordinates $[x, y, 1, k]$ are related to the camera coordinate frame by an element of the affine group.

Proof. Follows immediately from Corollary 2: the homography due to the plane at infinity is $A \cong M'RM^{-1}$. Hence, $A = I$ when $M = M'$ and $R = I$ (pure translational motion). □

Corollary 4 *The projective structure of the scene can be described as the ratio of two relative affine structures each with respect to a distinct reference plane $\pi, \hat{\pi}$, respectively, which in turn can be described as the ratio of affine structures under parallel projection with respect to the same two planes.*

Proof. Let k_π and $k_{\hat{\pi}}$ be the relative affine structures with respect to planes π and $\hat{\pi}$, respectively. From Theorem 2 we have that $k_\pi = \frac{z}{z_o}\frac{d_p}{d_o}$ and $k_{\hat{\pi}} = \frac{z}{z_o}\frac{\hat{d}_p}{\hat{d}_o}$. The ratio $k_\pi/k_{\hat{\pi}}$ removes the dependence on the projection center O (z/z_o cancels out) and is therefore a projective invariant (see Figure 4). This projective invariant is also the ratio of cross-ratios of the rays $\overline{OP}$ and $\overline{OP_o}$ with their intersections with the two planes π and $\hat{\pi}$, which was introduced in [34, 36] as "projective depth". It is also the ratio of two affine structures under parallel projection (recall that d_p/d_o is the affine structure; see Figure 2). □

Corollary 5 *The "essential" matrix $E = [v']R$ is a particular case of a generalized matrix $F = [v']A$. The matrix F, referred to as "fundamental" matrix in [7], is unique and does not depend on the plane π. Furthermore, $Fv = 0$ and $F^\top v' = 0$.*

Proof. Let $p \in \psi_o, p' \in \psi_1$ be two corresponding points, and let l, l' be their corresponding epipolar lines, i.e., $l \cong p \times v$ and $l' \cong p' \times v'$. Since lines are projective invariants, then any point along l is mapped by A to some point along l'. Thus, $l' \cong v' \times Ap$, and because p' is incident to l', we have ${p'}^\top(v' \times Ap) = 0$, or equivalently: ${p'}^\top[v']Ap = 0$, or ${p'}^\top Fp = 0$, where $F = [v']A$. From Corollary 2, $A = R$ in the special case where the plane π is at infinity and the cameras are internally calibrated as $M = M' = I$, thus $E = [v']R$ is a special case of F. The uniqueness of F follows from substitution of A with Equation 2 and noting that $[v']T = 0$, thus $F = [v']M'RM^{-1}$. Finally, since $Av \cong v'$, $[v']Av \cong [v']v' = 0$, thus $Fv = 0$, and $A^\top[v']^\top v' = -A^\top[v']v' = 0$, thus $F^\top v' = 0$. □

Corollary 6 (stream of views) *Given $m \geq 2$ views, let A_j and v'_j be the homographies of π and the epipoles, respectively, from view ψ_o to view ψ_j, and let the views of an object point P be p_j where the index j ranges over the m views. Then, the least squares solution for k is given by*

$$k = \frac{\sum_j (p_j \times v'_j)^T (A_j p_o \times p_j)}{\sum_j \parallel p_j \times v'_j \parallel^2}. \tag{3}$$

Proof. This is simply a calculation based on the observation that given a general equation of the type $\boldsymbol{a} \cong \boldsymbol{b} + k\boldsymbol{c}$, then by performing a cross product with $\boldsymbol{a}$ on both sides we get: $k(\boldsymbol{a} \times \boldsymbol{c}) = \boldsymbol{b} \times \boldsymbol{a}$. The value of k can be found using the normal equations (treating k as a vector of dimension 1):

$$k = \frac{(b \times a)^T (a \times c)}{\|a \times c\|^2}.$$

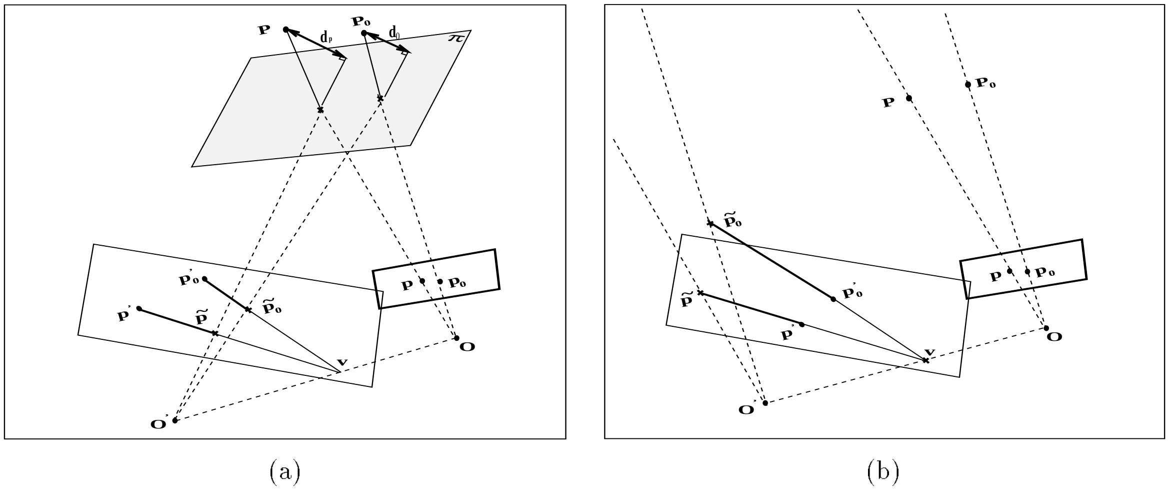


Figure 3: (a) Relative affine Structure: $k = \frac{z_o}{z}\frac{d_p}{d_o}$. (b) Affine structure under perspective (when π is at infinity). Note that the rays $\overline{OP}$ and $\overline{O'\tilde{p}}$ are parallel, thus the homography is the rotational component of motion.

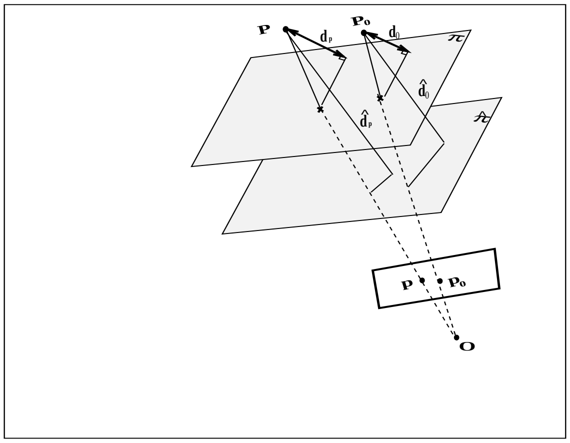


Figure 4: Projective-depth [34, 36] is the ratio of two relative affine structures, each with respect to a distinct reference plane, which is also the ratio of two affine structures (see Corollary 4 for more details).

Similarly, if in addition we have $\boldsymbol{a}' \cong \boldsymbol{b}' + k\boldsymbol{c}'$, then the overall least squares solution is given by

$$k = \frac{(b \times a)^T (a \times c) + (b' \times a')^T (a' \times c')}{\|a \times c\|^2 + \|a' \times c'\|^2}. \square$$

3.1 Explanatory Text

The key idea in Theorem 1 was to use both camera centers as part of the reference frame in order to show that the transformation between an arbitrary representation $\mathcal{R}_o$ of space as seen from the first camera and the representation $\mathcal{R}$ as seen from any other camera position, can be described by an element of the affine group. In other words, we have chosen an arbitrary plane π and made a choice of representation $\mathcal{R}_o$ in which π is the plane at infinity (i.e., π was mapped to infinity — not an unfamiliar trick, especially in computer graphics). The representation $\mathcal{R}_o$ is associated with $[x, y, 1, k]$ where k vanishes for all points coplanar with π, which means that π is the plane at infinity under the representation $\mathcal{R}_o$. What was left to show is that π remains the plane at infinity under all subsequent camera transformations, and therefore k is an affine invariant. Because k is invariant relative to the representation $\mathcal{R}_o$ we named it "relative affine structure"; this should not be confused with the term "relative invariants" used in classical invariant theory (invariants multiplied by a power of the transformation determinant, as opposed to "absolute invariants").

In practical terms, the difference between a full projective framework (like in [7, 13, 36]) and the relative affine framework can be described as follows. In a full projective framework, if we denote by f the invariance function acting on a pair of views indexed by a fixed set of five corresponding points, then $f(\psi_i, \psi_j)$ is fixed for all i, j. In a relative affine framework, if we denote f_o as the invariance function acting on a fixed view ψ_o and an arbitrary view ψ_i and indexed by a fixed set of four corresponding points, then $f_o(\psi_o, \psi_i)$ is fixed for all i.

The remaining theorem 2 and corollaries put the relative affine framework within the familiar context of affine structure under parallel and perspective projections, Euclidean structure and projective structure. The homography A due to the plane π was described as a product of the rigid camera motion parameters, the parameters of π, and the internal camera parameters of both cameras. This result is a natural extension of the classical motion of planes found in [9, 43], and also in [22]. The relative affine structure k was described as a product of the affine structure under parallel projection and a term that contains the location of the camera center of the reference view. Geometrically, k is the product of two ratios, the first being the ratio of the perpendicular[1] distance of a point P to the plane π and the depth z to the

[1] Note that the distance can be measured along any fixed direction. We use the perpendicular distance because it is the most natural way of describing the distance between a point and a plane.

reference camera, and the second ratio is of the same form but applied to a fixed point P_o which is used to set a uniform scale to the system. Therefore, when the depth goes to infinity (projection approaches orthography), then k approaches the ratio of the perpendicular distances of P from π and the perpendicular distance of P_o from π — which is precisely the affine structure under parallel projection [17]. Thus, relative affine structure is a generalization in the sense of including the center of projection of an arbitrary camera, and when the camera center goes to infinity we obtain an affine structure which becomes independent of the reference camera.

Another specialization of relative affine structure was shown in Corollary 2 by considering the case when π is at infinity with respect to our Euclidean frame (i.e., really at infinity). In that case k is simply inverse depth (up to a uniform scale factor), and the homography A is the familiar rotational component of camera motion (orthogonal matrix R) in the case of calibrated cameras, or a product of R with the internal calibration parameters. In other words, when π is at infinity also with respect to our camera coordinate frame, then relative affine becomes affine (the plane at infinity is preserved under all representations [7]). Notice that the rays towards the plane at infinity are parallel across the two cameras (see Figure 3-b). Thus, there exists a rotation matrix that aligns the two bundles of rays, and following this line of argument, the same rotation matrix aligns the epipolar lines (scaled appropriately) because orthogonal matrices commute with cross products. We have therefore the algorithm of [18] for determining the rotational component of standard calibrated camera motion, given the epipoles. In practice, of course, we cannot recover the homography due to the plane at infinity unless we are given prior information on the nature of the scene structure [28], or the camera motion is purely translational ([24] and Corollary 3). Thus in the general case, we can realize either the relative affine framework or the projective framework.

In Corollary 3 we address a particular case in which we *can* recover the homography due to the plane at infinity, hence recover the affine structure of the scene. This is the case where the camera motion is purely translational and the internal camera parameters remain fixed (i.e., we use the same camera for all views). This case was addressed in [24] by using clever and elaborate geometric constructions. The basic idea in [24] is that under pure translation of a calibrated camera, certain lines and points on the plane at infinity are easily constructed in the image plane. A line and a point from the plane at infinity are then used as auxiliaries for recovering the affine coordinates of the scene (with respect to a frame of four object points).

The relative affine framework provides a single-line proof of the main result of [24], and Furthermore, provides an extremely obvious algorithm for reconstruction of affine structure from a purely translating camera with fixed internal parameters, as follows. The epipole v' is the focus of expansion and is determined from two corresponding points ($v' \cong (p_i \times p'_i) \times (p_j \times p'_j)$, for some i, j). Given corresponding points p, p' in the two views, the coordinates (x, y, k), where k satisfies $p' \cong p + k\mathbf{v}'$, are related to the Euclidean coordinates (with respect to a camera coordinate frame) by an element of the affine group. The scalar k is determined up to scale, thus one of the points, say p_o, should determine the scale by scaling v' to satisfy $p'_o \cong p_o + v'$ (note that p_o can coincide with one of the points, p_i or p_j, used for determining v'). In case we would like to determine the affine coordinates with respect to four object points $P_1, ..., P_4$, we simply assign the standard coordinates $(0,0,0), (1,0,0), (0,1,0)$ and $(0,0,1)$ to those points, and solve for the 3D affine transformation that maps (x_i, y_i, k_i), $i = 1, ..., 4$, onto the standard coordinates (the mapping contains 12 parameters, and each of the four points determines three linear equations).

To conclude the implications of Corollary 3, we observe that given the epipole v', we need only one more point match (for setting a mutual scale) in order to determine affine structure. This is obvious because the epipole is the translational component of camera motion, and since this is the only motion we assume to have, the structure of the scene should follow without additional information. This case is very similar to the classic paradigm of stereopsis: instead of assuming that epipolar lines are horizontal, we recover the epipole (two point matches are sufficient), and instead of assuming a calibrated camera we assume an uncalibrated camera whose internal parameters remain fixed, and in turn, instead of recovering depth we can recover at most the affine structure of the scene. Finally, the result that the homography due to the plane at infinity is the identity matrix can be derived by geometric grounds as well. Points and lines from the plane at infinity are fixed points of the homography; with an affine frame of four points we can observe four fixed points, and thus, a homography with four fixed points is necessarily the identity matrix.

The connection between the relative affine structure and projective structure was shown in Corollary 4. Projective invariants are necessarily described with reference to five scene points [7], or equivalently, with reference to two planes and a point laying outside of them both [36, 34]. Corollary 4 shows that by taking the ratio of two relative affine structures, each relative to a different reference plane, then the dependence on the camera center (the term z_o/z) drops and we are left with the projective invariant described in [36], which is the ratio of the perpendicular distance of a point to two planes (up to a uniform scale factor).

Corollary 5 unifies previous results on the nature of what is known by now as the "fundamental matrix" [7, 8]. It is shown, that for any plane π and its corresponding homography A we have $F = [v']A$. First, we see that given a homography, the epipole v' follows by having two corresponding points coming from scene points not coplanar with π — an observation that was originally made by [18]. Second, F is fixed, regardless of the choice of π, which was shown by using the result of Theorem 2. As a particular case, the product $[v']R$ remains fixed if we add to R a element that vanishes as a product with $[v']$ — an observation that was made previously by [13]. Thirdly, the "essential" matrix [19], $E = [v']R$, is shown to be a specialization of F in the case π is at infinity with respect to the world coordinate frame and the cameras are internally calibrated as $M = M' = I$.

Finally, Corollary 6 provides a practical formula for obtaining a least-squares estimation of relative affine structure which also applies for the case where a stream of views is available — in the spirit of [46, 42, 23, 41, 1, 5]. In the next section we apply these results to obtain a simple algorithm for relative affine reconstruction from multiple $m \geq 2$ views and multiple points.

3.2 Application I: Reconstruction from a Stream of Views

Taken together, the results above demonstrate the ability to compute relative affine structure using many points over many views in a least squares manner. At minimum we need two views and four corresponding points and the corresponding epipoles to recover k for all other points of the scene whose projections onto the two views are given. Let p_{ij}, $i = 0, ..., n$ and $j = 0, ..., m$ denote the i'th image point on frame j. Let A_j denote the homography from frame 0 to frame j, v_j, v'_j the corresponding epipoles such that $A_j v_j \cong v'_j$, and let k_i denote the relative affine structure of point i. We follow these steps:

1. Compute epipoles v_j, v'_j using the relation $p_{ij} F_j p_{io} = 0$, over all i. Eight corresponding points (frame 0 and frame j) are needed for a linear solution, and a least-squares solution is possible if more points are available. In practice the best results were obtained using the non-linear algorithm of [21]. The epipoles follow by $F_j v_j = 0$ and $F^\top v'_j = 0$ [7]. The latter readily follows from Corollary 5 as $[v'_j]A_j v_j \cong [v'_j]v'_j = 0$ and $A_j^\top [v'_j]^\top v'_j = -A_j^\top [v'_j]v'_j = 0$.
2. Compute A_j from the equations $A_j p_{io} \cong p_{ij}$, $i = 1, 2, 3$, and $A_j v_j \cong v'_j$. This leads to a linear set of eight equations for solving for A_j up to a scale. A least squares solution is available from the equation $p_{ij}[v'_j]A_j p_{io} = 0$ for all additional points (Corollary 5). Scale A_j to satisfy $p_{oj} \cong A_j p_{oo} + v'_j$.
3. Relative affine structure k_i is given by (3).

3.3 Application II: Recognition by Alignment

The relative affine invariance relation, captured by Theorem 1, can be used for visual recognition by alignment ([44, 14], and references therein). In other words, the invariance of k can be used to "re-project" the object onto any third view p'', as follows. Given two "model" views in full correspondence $p_i \longleftrightarrow p'_i$, $i = 1, ..., n$, we recover the epipoles and homography A from $Ap_i \cong p'_i$, $i = 1, 2, 3$, and $Av \cong v'$. Then the corresponding points p''_i in any third view satisfy $p'' \cong Bp + \boldsymbol{k}\boldsymbol{v}''$, for some matrix B and epipole v''. One can solve for B and v'' by observing six corresponding points between the first and third view. Once B, v'' are recovered, we can find the estimated location of p''_i for the remaining points p_i, $i = 7, ..., n$, by first solving for k_i from the equation $p'_i \cong Ap_i + k_i v'$, and then substituting the result in the equation $\hat{p}''_i \cong Bp_i + k_i \boldsymbol{v}''$. Recognition is achieved if the distance between p''_i and $\hat{p}''_i$, $i = 7, ..., n$, is sufficiently small. Other methods for achieving reprojection include the epipolar intersection method (cf. [26, 6, 11]), or by using projective structure instead of the relative affine structure [34, 36]. In all the above methods the epipolar geometry plays a key and preconditioned role. More direct methods, that do not require the epipolar geometry can be found in [35, 37].

3.4 Application III: Image Coding

The re-projection paradigm, described in the previous section, can serve as a principle for model-based image compression. In a sender/receiver mode, the sender computes the relative affine structure between two extreme views of a sequence, and sends the first view, the relative affine scalars, and the homographies and epipoles between the first frame and all the intermediate frames. The intermediate frames can be reconstructed by re-projection. Alternatively, the sender send the two extreme views and the homographies and epipoles between the first and all other intermediate views. The receiver recovers the correspondence field between the two extreme views, and then synthesizes the remaining views from the received parameters of homographies and epipoles. In case the distance between the two extreme views is "moderate", we found that optical flow techniques can be useful for the stage of obtaining the correspondence field between the views. Experiments can be found later in the text, and more detailed experiments concerning the use of optical flow in full registration of images for purposes of model-based image compression can be found in [4].

4 Experimental Results

The following experiments were conducted to illustrate the applications that arise from the relative affine framework (reconstruction, recognition by alignment, and image coding) and to test the algorithms on real data. The performance under real imaging situations is interesting, in particular, because of the presence of deviations from the pin-hole camera model (radial distortions, decentering, and other effects), and due to errors in obtaining image correspondences.

Fig. 5 shows four views, out of a sequence of ten views, of the object we selected for experiments. The object is a sneaker with added texture to facilitate the correspondence process. This object was chosen because of its complexity, i.e., it has a shape of a natural object and cannot easily be described parameterically (as a collection of planes or algebraic surfaces). A set of thirty-four points were manually selected on one of the frames, referred to as the first frame, and their correspondences were automatically obtained along all other frames used in this experiment (corresponding points are marked by overlapping squares in Fig. 5). The correspondence process is based on an implementation of a coarse-to-fine optical-flow algorithm based on [20] and described in [3].

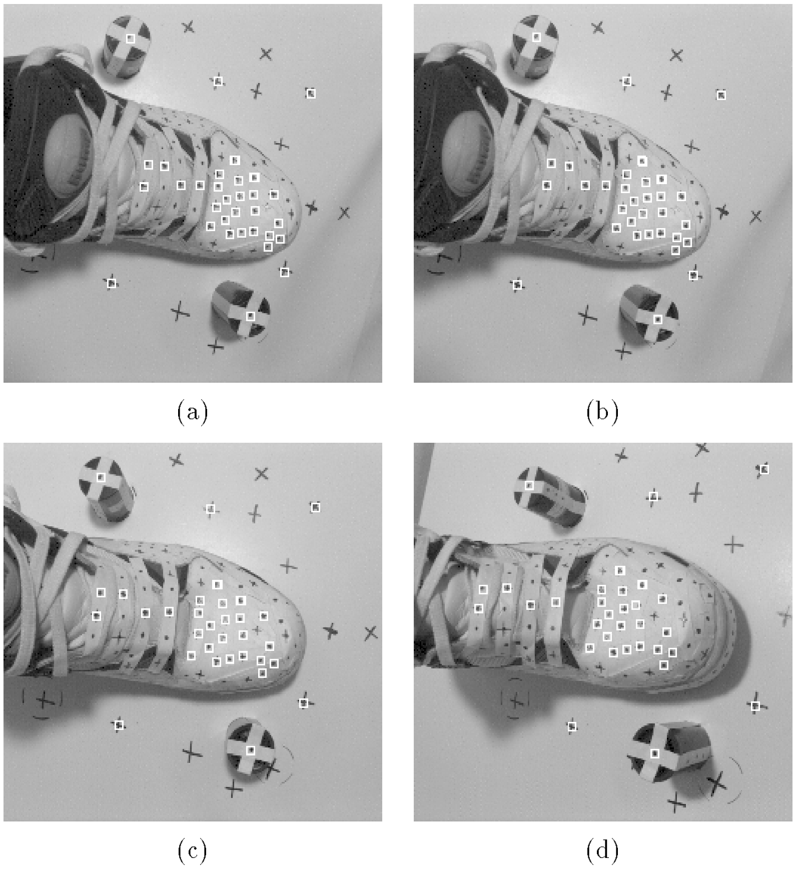
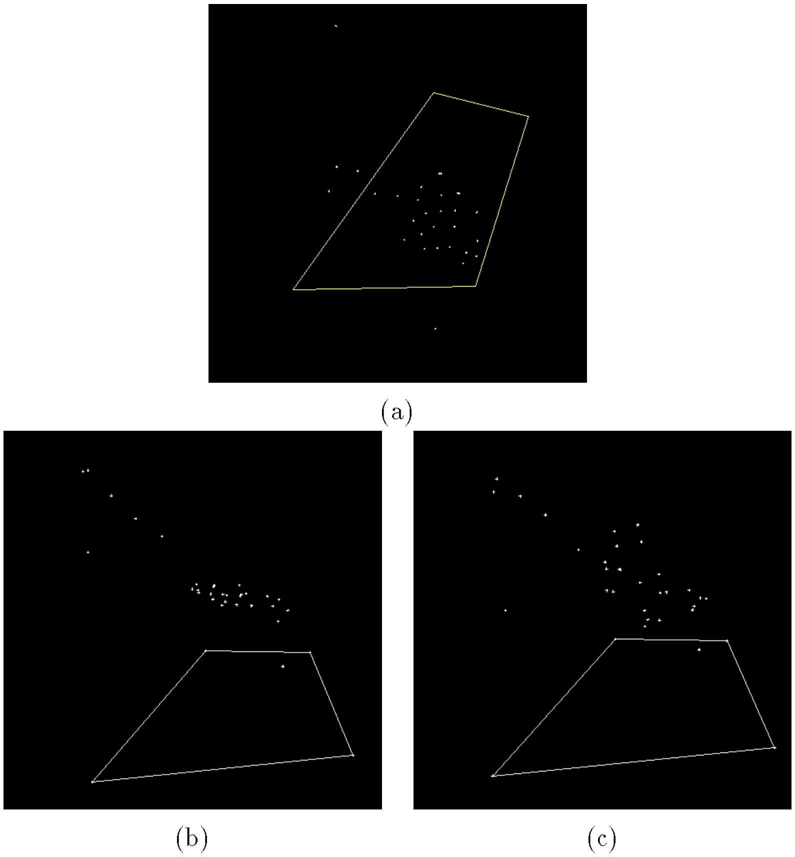


Figure 5: Four views, out of a sequence of ten views, of a sneaker. The frames shown here are the first, second, fifth and tenth of the sequence (top-bottom, left-to-right). The overlayed squares mark the corresponding points that were tracked and subsequently used for our experiments.


(a)

(b) (c)

Figure 6: Results of 3D reconstruction of the collection of sample points. (a) Frontal view (aligned with the first frame of the sneaker). The two bottom displays show a side view of the sample. (b) Result of recovering structure between the first and tenth frame (large base-line); (c) Result of recovery between the first and second frames (small base-line).

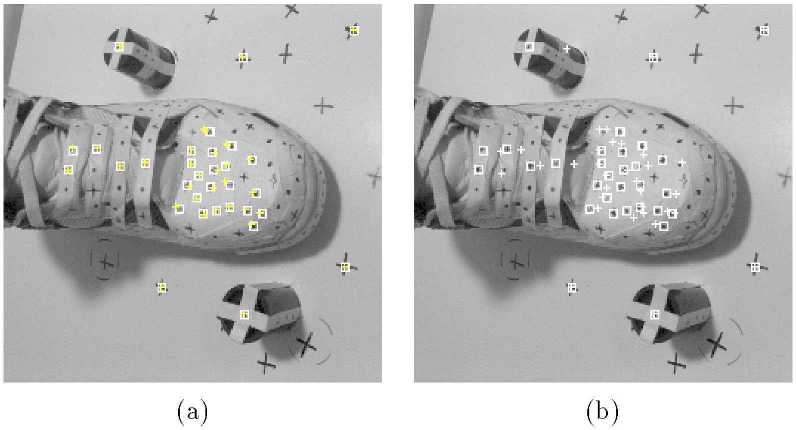

(a) (b)

Figure 7: Results of re-projection onto the tenth frame. Epipoles were recovered using the ground plane homography (see text). The re-projected points are marked by crosses, and should be in the center of their corresponding square for accurate re-projection. (a) Structure was recovered between the first and fifth frames, then re-projected onto the tenth frame (large base-line). Average error is 1.1 pixels with std of 0.98. (b) Structure was recovered between the first and second frames (small base-line situation) and then re-projected onto the tenth frame. Average error is 7.81 pixels with std of 6.5.

Epipoles were recovered by either one of the following two methods. First, by using the four ground points to recover the homography A, and then by Corollary 5 to compute the epipoles using all the remaining points in a least squares manner. Second, using the non-linear algorithm proposed by [21]. The two methods gave rise to very similar results for reconstruction, and slightly different results for re-projection (see later).

In the reconstruction paradigm, we recovered relative affine structure from two views and multiple views. In the two-view case we used either a small base-line (the first two views of the sequence) or a large base-line (the first and last views of the sequence). In the multiple view case, we used all ten views of the sequence (Corollary 6). The transformation to Euclidean coordinates was done for purposes of display by assuming that the ground plane is parallel to the image plane (it actually is not) and that the camera is calibrated (there was no calibration attempt made).

The 3D coordinates are shown in Fig. 6. Display (a) shows a frontal view (in order to visually align the display with the image of the sneaker). Other displays show a side view of the reconstructed sneaker under the following experimental situations. Display (b) is due to reconstruction under large base-line situation (the two methods for obtaining the epipoles produced very similar results; the multiple-view case produced very similar results as well). The side view illustrates the robustness of the reconstruction process, as it was obtained by rotating the object around a different axis than the one used for capturing the images. Display (c) is due to reconstruction under small base-line situation (both methods for obtaining the epipoles produced very similar results). The quality of reconstruction in the latter case is not as good as in the former, as should be expected. Nevertheless, the system does not totally brake-down under relatively small base-line situations and produces a reasonable result under these circumstances.

In the re-projection application (see Section 3.3), relative affine structure was recovered using the first and in-between views, and re-projected onto the last view of the sequence. Note that this is an extrapolation example, thereby performance is expected to be poorer than interpolation examples, i.e., when the re-projected view is in-between the model views. The interpolation case will be discussed in the next section, where relevance to image coding applications is argued for.

In general, the performance was better when the ground plane was used for recovering the epipoles. When the intermediate view was the fifth in the sequence (Fig. 5, display (c)), the average error in re-projection was 1.1 pixels (with standard deviation of 0.98 pixels). When the intermediate view was the second frame in the sequence (Fig. 5, display (b)), the results were poorer (due to small base-line and large extrapolation) with average error of 7.81 pixels (standard deviation of 6.5). These two cases are displayed in Fig. 7. The re-projected points are represented by crosses overlayed on the last frame (the re-projected view).

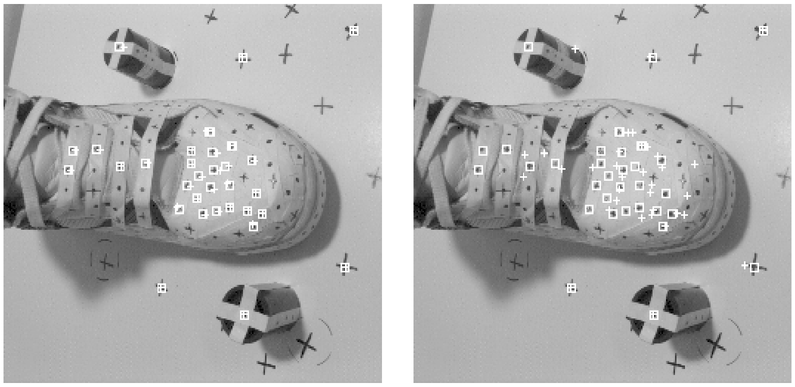
When the second method for computing the epipoles was used (more general, but generally less accurate), the results were as follows. With the fifth frame, the average error was 1.62 pixels (standard deviation of 1.2); and with the second frame (small base-line situation) the average error was 13.87 pixels (standard deviation of 9.47). These two cases are displayed in Fig. 8. Note that because all points were used for recovering the epipoles, the re-projection performance, only indicates the level of accuracy one can obtain when all the information is being used. In practice we would like to use much fewer points from the re-projected view, and therefore, re-projection methods that avoid the epipoles all together would be preferred — an example of such a method can be found in [35, 37].

For the image coding paradigm (see Section 3.4), relative affine structure of the 34 sample points were computed between the first and last frame of the ten frame sequence (displays (a) and (d) in Fig. 5). Display (a) in Fig. 9 shows a graph of the average re-projection error for all the intermediate frames (from second to ninth frames). Display (b) shows the relative error normalized by the distance between corresponding points across the sequence. We see that the relative error generally goes down as the re-projected frame is farther from the first frame (increase of base-line). In all frames, the average error is less than 1 pixel, indicating a relatively robust performance in practice.

5 Summary

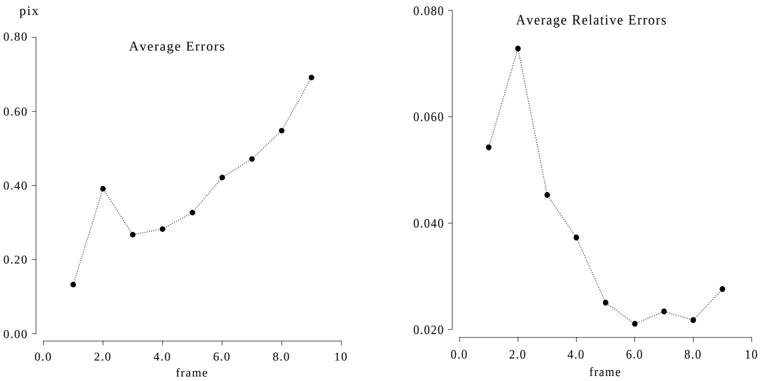
The framework of "relative affine" was introduced and shown to be general and sharper than the projective results for purposes of 3D reconstruction from multiple views and for the task of recognition by alignment. One of the key ideas in this work is to define and recover an invariant that stands in the middle ground between affine and projective. The middle ground is achieved by having the camera center of one arbitrary view as part of the projective reference frame (of five points), thus obtaining the first result described in Theorem 1 (originally in [33]). The result simply states that under general uncalibrated camera motion, the sharpest result we can obtain is that all the degrees of freedom are captured by four points (thus the scene may undergo at most 3D affine transformations) and a single unknown projective transformation (from the arbitrary viewer-centered representation $\mathcal{R}_o$ to the camera coordinate frame). The invariants that are obtained in this way are viewer-centered since the camera center is part of the reference frame and are called "relative affine structure". This statement, that all the available degrees of freedom are captured by four points and one projective transformation, was also recently presented in [40] using different notations and tools than those used here and in [33, 38].

This "middle ground" approach has several advantages. First, the results are sharper than a full projective reconstruction approach ([7, 13]) where five scene points are needed. The increased sharpness translates to a remarkably simple framework captured by a single equation (Equation 1). Second, the manner in which the results were derived provides the means for unifying a wide range of other previous results, thus obtaining a



(a) (b)

Figure 8: Re-projection onto the tenth frame. Epipoles are computed via fundamental matrix (see text) using the implementation of [21]. (a) Large base situation (structure computed between first and fifth frames): average error 1.62 with std of 1.2. (b) Small base-line situation (structure computed between first and second frames): average error 13.87 with std of 9.47.



(a) (b)

Figure 9: Error in re-projection onto the intermediate frames (2–9). Structure was computed between frames one and ten. (a) average error in pixels, (b) relative error normalized by the displacement between corresponding points.

canonical framework. Following Theorem 2, the corollaries show how this "middle ground" reduces back to full affine structure and extends into full projective structure (Corollaries 1 and 4). The corollaries also show how the "plane at infinity" is easily manipulated in this framework, thereby making further connections among projective affine and Euclidean results in general and less general situations (Corollaries 2 and 3). The corollaries also unify the various results related to the epipolar geometry of two views: the Essential matrix of [19], the Fundamental matrix of [7] and other related results of [13] (Corollary 5). All the above connections and results are often obtained as a single-line proof and follow naturally from the relative affine framework.

Finally, the relative affine result has proven useful for derivation of other results and applications, some of which can be found in [39, 37, 35]. The derivation of those results critically rely on the simplicity of the relative affine framework, and in some cases [37, 35] on the sharpness of the framework compared to the projective framework.

References

[1] Ali J. Azarbayejani, Bradley Horowitz, and Alex Pentland. Recursive estimation of structure and motion using relative orientation constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 294–299, New York, NY, June 1993.

[2] E.B. Barrett, 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.

[3] J.R. Bergen and R. Hingorani. Hierarchical motion-based frame rate conversion. Technical report, David Sarnoff Research Center, 1990.

[4] D. Beymer, A. Shashua, and T. Poggio. Example based image analysis and synthesis. A.I. Memo No. 1431, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, October 1993.

[5] T.J. Broida and R. Chellapa. Estimating the kinematics and structure of a rigid object from a sequence of monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13(6):497–513, 1991.

[6] S. Demey, A. Zisserman, and P. Beardsley. Affine and projective structure from motion. In *Proceedings of the British Machine Vision Conference*, October 1992.

[7] 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.

[8] O.D. Faugeras, Q.T. Luong, and S.J. Maybank. Camera self calibration: Theory and experiments. In *Proceedings of the European Conference on Computer Vision*, pages 321–334, Santa Margherita Ligure, Italy, June 1992.

[9] O.D. Faugeras and F. Lustman. Let us suppose that the world is piecewise planar. In O. D. Faugeras and Georges Giralt, editors, *International Symposium on Robotics Research*, pages 33–40. MIT Press, Cambridge, MA, 1986.

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

[11] O.D. Faugeras and L. Robert. What can two images tell us about a third one? Technical Report INRIA, France, 1993.

[12] D.A. Forsyth, J.L. Mundy, A.P. Zisserman, A.P. Coelho, C. Heller, and C.A. Rothwell. Invariant descriptors for 3-D object recognition and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-13(10):971–991, 1991.

[13] R. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–764, Champaign, IL., June 1992.

[14] D.P. Huttenlocher and S. Ullman. Recognizing solid objects by alignment with an image. *International Journal of Computer Vision*, 5(2):195–212, 1990.

[15] D.W. Jacobs. *Recognizing 3-D objects from 2-D images.* PhD thesis, M.I.T Artificial Intelligence Laboratory, September 1992.

[16] D.W. Jacobs. Space efficient 3D model indexing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 439–444, 1992.

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

[18] 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.

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

[20] B.D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings IJCAI*, pages 674–679, Vancouver, Canada, 1981.

[21] Q.T. Luong, R. Deriche, O.D. Faugeras, and T. Papadopoulo. On determining the fundamental matrix: Analysis of different methods and experimental results. Technical Report INRIA, France, 1993.

[22] Q.T. Luong and T. Vieville. Canonical representations for the geometries of multiple projective views. Technical Report INRIA, France, fall 1993.

[23] R. Mohr, F. Veillon, and L. Quan. Relative 3d reconstruction using multiple uncalibrated images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 543–548, New York, NY, June 1993.

[24] T. Moons, L. Van Gool, M. Van Diest, and E. Pauwels. Affine reconstruction from perspective image pairs. In *The 2nd European Workshop on Invariants*, Ponta Delagada, Azores, October 1993.

[25] 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.

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

[27] N. Navab and A. Shashua. Algebraic description of relative affine structure: Connections to euclidean, affine and projective structure. Technical Report 270, Media Laboratory, Massachusetts Institute of Technology, 1994.

[28] L. Quan. Affine stereo calibration for relative affine shape reconstruction. In *Proceedings of the British Machine Vision Conference*, pages 659–668, 1993.

[29] L. Robert and O.D. Faugeras. Relative 3D positioning and 3D convex hull computation from a weakly calibrated stereo pair. In *Proceedings of the International Conference on Computer Vision*, pages 540–544, Berlin, Germany, May 1993.

[30] 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.

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

[32] 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.

[33] A. Shashua. On geometric and algebraic aspects of 3D affine and projective structures from perspective 2D views. In *Proceedings of the 2nd European Workshop on Invariants*, Ponta Delagada, Azores, October 1993. Also MIT AI Memo No. 1405, July 1993.

[34] A. Shashua. Projective depth: A geometric invariant for 3D reconstruction from two perspective/orthographic views and for visual recognition. In *Proceedings of the International Conference on Computer Vision*, pages 583–590, Berlin, Germany, May 1993.

[35] A. Shashua. Algebraic functions for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994. in press.

[36] A. Shashua. Projective structure from uncalibrated images: structure from motion and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994. In press.

[37] A. Shashua. Trilinearity in visual recognition by alignment. In *Proceedings of the European Conference on Computer Vision*, Stockholm, Sweden, May 1994.

[38] A. Shashua and N. Navab. Relative affine structure: Theory and application to 3D reconstruction from perspective views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, Washington, 1994.

[39] A. Shashua and S. Toelg. The quadric reference surface: Applications in registering views of complex 3d objects. In *Proceedings of the European Conference on Computer Vision*, Stockholm, Sweden, May 1994.

[40] G. Sparr. A common framework for kinetic depth, reconstruction and motion for deformable objects. In *Proceedings of the European Conference on Computer Vision*, pages 471–482, Stockholm, Sweden, May 1994.

[41] R. Szeliski and S.B. Kang. Recovering 3D shape and motion from image streams using non-linear least squares. Technical Report D.E.C., December 1992.

[42] C. Tomasi and T. Kanade. Factoring image sequences into shape and motion. In *IEEE Workshop on Visual Motion*, pages 21–29, Princeton, NJ, September 1991.

[43] R. Tsai and T.S. Huang. Estimating three-dimensional motion parameters of a rigid planar patch, II: singular value decomposition. *IEEE Trans. on Acoustic, Speech and Signal Processing*, 30, 1982.

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

[45] 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.

[46] D. Weinshall. Model based invariants for 3-D vision. *International Journal of Computer Vision*, 10(1):27–42, 1993.

[47] D. Weinshall and C. Tomasi. Linear and incremental acquisition of invariant shape models from image sequences. In *Proceedings of the International Conference on Computer Vision*, pages 675–682, Berlin, Germany, May 1993.

[48] I. Weiss. Geometric invariants and object recognition. *International Journal of Computer Vision*, 10(3):201–231, 1993.