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

A.I. Memo No. 1405 July, 1993
C.B.C.L. Paper No. 78

On Geometric and Algebraic Aspects of 3D Affine and Projective Structures from Perspective 2D Views

Amnon Shashua

Abstract

Part I of this paper investigates the differences — conceptually and algorithmically — between affine and projective frameworks for the tasks of visual recognition and reconstruction from perspective views. It is shown that an affine invariant exists between any view and a fixed view chosen as a reference view. This implies that for tasks for which a reference view can be chosen, such as in alignment schemes for visual recognition, projective invariants are not really necessary. The projective extension is then derived, showing that it is necessary only for tasks for which a reference view is not available — such as happens when updating scene structure from a moving stereo rig. The geometric difference between the two proposed invariants are that the affine invariant measures the relative deviation from a single reference plane, whereas the projective invariant measures the relative deviation from two reference planes. The affine invariant can be computed from three corresponding points and a fourth point for setting a scale; the projective invariant can be computed from four corresponding points and a fifth point for setting a scale. Both the affine and projective invariants are shown to be recovered by remarkably simple and linear methods.

In part II we use the affine invariant to derive new algebraic connections between perspective views. It is shown that three perspective views of an object are connected by certain algebraic functions of image coordinates alone (no structure or camera geometry needs to be involved). In the general case, three views satisfy a trilinear function of image coordinates. In case where two of the views are orthographic and the third is perspective the function reduces to a bilinear form. In case all three views are orthographic the function reduces further to a linear form (the "linear combination of views" of [31]). These functions are shown to be useful for recognition, among other applications.

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. Support for the A.I. Laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-91-J-4038. Support for the Center's research is provided in part by ONR contracts N00014-91-J-1270 and N00014-92-J-1879; by a grant from the National Science Foundation under contract ASC-9217041 (funds provided by this award include funds from ARPA provided under HPCC); and by a grant from the National Institutes of Health under contract NIH 2-S07-RR07047-26. Additional support is provided by the North Atlantic Treaty Organization, ATR Audio and Visual Perception Research Laboratories, Mitsubishi Electric Corporation, Siemens AG., and Sumitomo Metal Industries. A. Shashua is supported by a McDonnell-Pew postdoctoral fellowship from the department of Brain and Cognitive Sciences.

1 Introduction

The geometric relation between objects (or scenes) in the world and their images, taken from different viewing positions by a pin-hole camera, has many subtleties and nuances and has been the subject of research in computer vision since its early days. Two major areas in computer vision have been shown to benefit from an analytic treatment of the 3D to 2D geometry: visual recognition and reconstruction from multiple views (as a result of having motion sequences or from stereopsis).

A recent approach with growing interest in the past few years is based on the idea 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) [14, 26, 33, 34, 9, 20, 3, 4, 28, 29, 19, 31, 23, 5, 6, 18, 27, 13, 12] — and relevant to this paper are the work described in [14, 4, 26, 28, 29].

This paper has two parts. In Part I we investigate the intrinsic differences — conceptually and algorithmically — between an affine framework for recognition/reconstruction and a projective framework. Although the distinction between affine and projective spaces, and between affine and projective properties, is perfectly clear from classic studies in projective and algebraic geometries, as can be found in [8, 24, 25], it is less clear how these concepts relate to reconstruction from multiple views. In other words, given a set of views, under what conditions can we expect to recover affine invariants? what is the benefit from recovering projective invariants over affine? are there tasks, or methodologies, for which an affine framework is completely sufficient? what are the relations between the set of views generated by a pin-hole camera and the set of all possible projections $\mathcal{P}^3 \mapsto \mathcal{P}^2$ of a particular object? These are the kinds of questions for which the current literature does not provide satisfactory answers. For example, there is a tendency in some of the work listed above, following the influential work of [14], to associate the affine framework with reconstruction/recognition from orthographic views only. As will be shown later, the affine restriction need not be coupled with the orthographic restriction on the model of projection — provided we set one view fixed. In other words, an uncalibrated pin-hole camera undergoing general motion can indeed be modeled as an "affine engine" provided we introduce a "reference view", i.e., all other views are matched against the reference view for recovering invariants or for achieving recognition.

In the course of addressing these issues we derive two new, extremely simple, schemes for recovering geometric invariants — one affine and the other projective — which can be used for recognition and for reconstruction.

Some of the ideas presented in this part of the paper follow the work of [14, 4, 26, 28, 29]. Section 3 on affine reconstruction from two perspective views, follows and expands upon the work of [26, 14, 4]. Section 4 on projective reconstruction, follows and refines the results presented in [28, 29].

In Part II of this paper we use the results established in Part I (specifically those in Section 3) to address certain algebraic aspects of the connections between multiple views. Inspired by the work of [31], we address the problem of establishing a direct connection between views, expressed as functions of image coordinates alone — which we call "algebraic functions of views". In addition to linear functions of views, discovered by [31], applicable to orthographic views only, we show that three perspective views are related by trilinear functions of their coordinates, and by bilinear functions if two of the three views are assumed orthographic — a case that will be argued is relevant for purposes of recognition without constraining the generality of the recognition process. Part II ends with a discussion of possible applications for algebraic functions, other than visual recognition.

2 Mathematical Notations and Preliminaries

We consider object space to be the three-dimensional projective space $\mathcal{P}^3$, and image space to be the two-dimensional projective space $\mathcal{P}^2$. Within $\mathcal{P}^3$ we will be considering the projective group of transformations and the affine group. Below we describe basic definitions and formalism related to projective and affine geometries — more details can be found in [8, 24, 25].

2.1 Affine and Projective Spaces

Affine space over the field K is simply the vector space K^n, and is usually denoted as $\mathcal{A}^n$. Projective space $\mathcal{P}^n$ is the set of equivalence classes over the vector space K^{n+1}. A point in $\mathcal{P}^n$ is usually written as a homogeneous vector $(x_0, ..., x_n)$, which is an ordered set of $n+1$ real or complex numbers, not all zero, whose ratios only are to be regarded as significant. Two points $\boldsymbol{x}$ and $\boldsymbol{y}$ are equivalent, denoted by $\boldsymbol{x} \cong \boldsymbol{y}$, if $\boldsymbol{x} = \lambda \boldsymbol{y}$ for some scalar λ. Likewise, two points are distinct if there is no such scalar.

2.2 Representations

The points in $\mathcal{P}^n$ admit a class of coordinate representations $\mathcal{R}$ such that if $\mathcal{R}_0$ is any one allowable representation, the whole class $\mathcal{R}$ consists of all those representations that can be obtained from $\mathcal{R}_0$ by the action of the group GL_{n+1} of $(n+1) \times (n+1)$ non-singular matrices. It follows, that any one coordinate representation is completely specified by its standard simplex and its unit point. The standard simplex is the set of $n+1$ points which have the standard coordinates $(1, 0, ..., 0), (0, 1, 0, ..., 0), ..., (0, 0, ..., 0, 1)$ and the unit point is the point whose coordinates are $(1, 1, ..., 1)$. It also follows that the coordinate transformation between any two representations is completely determined from $n+1$ corresponding points in the two representations, which give rise to a linear system of $(n+1)^2 - 1$ or $(n+1)^2$ equations (depending on whether we set an arbitrary element of the matrix transform, or set one of the scale factors of the corresponding points).

2.3 Subspaces and Cross Ratios

A linear subspace $\Lambda \cong \mathcal{P}^k \subset \mathcal{P}^n$ is a hyperplane if $k = n-1$, is a line when $k=1$, and otherwise is a k-plane. There is a unique line in $\mathcal{P}^n$ through any two distinct points. Any point $\boldsymbol{z}$ on a line can be described as a linear combination of two fixed points $\boldsymbol{x}, \boldsymbol{y}$ on the line, i.e., $\boldsymbol{z} \cong \boldsymbol{x} + k\boldsymbol{y}$. Let $\boldsymbol{v} \cong \boldsymbol{x} + k'\boldsymbol{y}$ be another point on the line spanned by $\boldsymbol{x}, \boldsymbol{y}$, then the cross ratio of the four points is simply $\alpha = k/k'$ which is invariant in all representations $\mathcal{R}$. By permuting the four points on the line the 24 possible cross ratios fall into six sets of four with values $\alpha, 1/\alpha, 1-\alpha, (\alpha-1)/\alpha, \alpha/(\alpha-1)$ and $1/(1-\alpha)$.

2.4 Projections

Let $\mathcal{P}^{n-1} \subset \mathcal{P}^n$ be some hyperplane, and a point $O \in \mathcal{P}^n$ not lying on $\mathcal{P}^{n-1}$. If we like, we can choose the representation such that $\mathcal{P}^{n-1}$ is given by $x_n = 0$ and the point $O = (0, 0, ..., 0, 1)$. We can define a map

$$\sigma_o : \mathcal{P}^n - \{O\} \rightarrow \mathcal{P}^{n-1}$$

by

$$\sigma_o : P \mapsto \overline{OP} \cap \mathcal{P}^{n-1};$$

that is, sending a point $P \in \mathcal{P}^n$ other than O to the point of intersection of the line $\overline{OP}$ with the hyperplane $\mathcal{P}^{n-1}$. σ_o is the projection from the point O to the hyperplane $\mathcal{P}^{n-1}$, and the point O is called the center of projection (COP). In terms of coordinates $\boldsymbol{x}$, this amounts to

$$\sigma_o : (x_0, ..., x_n) \mapsto (x_0, ..., x_{n-1}).$$

As an example, the projection of 3D objects onto an image plane is modeled by $\boldsymbol{x} \mapsto T\boldsymbol{x}$, where T is a 3×4 matrix, often called the camera transformation. The set $\mathcal{S}$ of all views of an object (ignoring problems of self occlusion, i.e., assuming that all points are visible from all viewpoints) is obtained by the group GL_4 of 4×4 non-singular matrices applied to some arbitrary representation of $\mathcal{P}^3$, and then dropping the coordinate x_3.

2.5 The Affine Subgroup

Let $A_i \subset \mathcal{P}^n$ be the subset of points $(x_0, ..., x_n)$ with $x_i \neq 0$. Then the ratios $\bar{x}_j = x_j/x_i$ are well defined and are called affine or Euclidean coordinates on the projective space, and A_i is bijective to the affine space $\mathcal{A}^n$, i.e. $A_i \cong \mathcal{A}^n$. The affine subgroup of GL_{n+1} leaves the hyperplane $x_i = 0$ invariant under all affine representations. Any subgroup of GL_{n+1} that leaves some hyperplane invariant is an affine subgroup, and the invariant hyperplane is called the ideal hyperplane. As an example, a subgroup of GL_4 that leaves some plane invariant is affine. It could be any plane, but if it is the plane at infinity ($x_2 = 0$) then the mapping $\mathcal{P}^3 \mapsto \mathcal{P}^2$ is created by parallel projection, i.e., the COP is at infinity. Since two lines are parallel if they meet on the ideal hyperplane, then when the ideal hyperplane is at infinity, affine geometry takes its "intuitive" form of preserving parallelism of lines and planes and preserving ratios. The importance of the affine subgroups is that there exist affine invariants that are not projective invariants. Parallelism, the concept of a midpoint, area of triangles, classification of conics are examples of affine properties that are not projective.

2.6 Epipoles

Given two cameras with positions of their COP at $O, O' \in \mathcal{P}^3$, respectively, the epipoles are at the intersection of the line $\overline{OO'}$ with both image planes. Recovering the epipoles from point correspondences across two views is remarkably simple but is notoriously sensitive to noise in image measurements. For more details on recovering epipoles see [4, 29, 28, 5], and for comparative and error analysis see [17, 22]. In Part I of this paper we assume the epipoles are given; in Part II, where we make further use of derivations made in Section 3, we show that for purposes discussed there one can eliminate the epipoles altogether.

2.7 Image Coordinates

Image space is $\mathcal{P}^2$. Since the image plane is finite, we can assign, without loss of generality, the value 1 as the third homogeneous coordinate to every 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. Note that by this notation we are not assuming that an observed point in one image is always mapped onto an observed (i.e., not at infinity) point in another view (that would constitute an affine plane) — all what we are relying upon is that points at infinity are not observed anyway, so we are allowed to assign the value 1 to all observed points.

2.8 General Notations

Vectors are always column vectors, unless mentioned otherwise. The transpose notation will be added only when otherwise there is a chance for confusion. Vectors will be in bold-face only in conjunction with a scalar, i.e., $\lambda\boldsymbol{x}$ stands for the scalar λ scaling the vector x. Scalar product will be noted by a center dot, i.e., $x \cdot y$, again avoiding the transpose notation except when necessary. Cross product will be denoted as usual, i.e., $x \times y$. The cross product, viewed as an operator, can be used between a vector x and a 3×3 matrix A as follows:

$$x \times A = \begin{bmatrix} x_2\boldsymbol{a}_3 - x_3\boldsymbol{a}_2 \\ x_3\boldsymbol{a}_1 - x_1\boldsymbol{a}_3 \\ x_1\boldsymbol{a}_2 - x_2\boldsymbol{a}_1 \end{bmatrix},$$

where $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ are the row vectors of A, and $x = (x_1, x_2, x_3)$.

Part I

3 Affine Structure and Invariant From Two Perspective Views

The key idea underlying the derivations in this section is to place the two camera centers as part of the reference frame (simplex and unit point) of $\mathcal{P}^3$. Let P_1, P_2, P_3 be three object points projecting onto corresponding points p_j, p'_j, $j = 1, 2, 3$, in the two views. We assign the coordinates $(1,0,0,0), (0,1,0,0), (0,0,1,0)$ to P_1, P_2, P_3, respectively. For later reference, the plane passing through

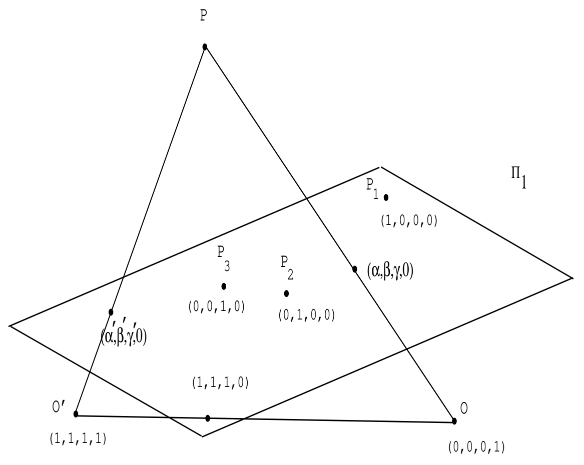

Figure 1:

P_1, P_2, P_3 will be denoted by π_1. Let O be the COP of the first camera, and O' the COP of the second camera. We assign the coordinates $(0,0,0,1),(1,1,1,1)$ to O, O', 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_1 has the coordinates $(1,1,1,0)$ (note that π_1 is the plane $x_3 = 0$, therefore the linear combination of O and O' with $x_3 = 0$ must be a multiple of $(1,1,1,0)$).

Let P be some object point projecting onto p, p'. The line $\overline{OP}$ intersects π_1 at the point $(\alpha, \beta, \gamma, 0)$. The coordinates α, β, γ can be recovered by projecting the image plane onto π_1, as follows. Let v, v' be the location of both epipoles in the first and second view, respectively (see Section 2.6). Given the epipoles v and v', 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 $A_1 \in PGL_3$ (3×3 matrix defined up to a scale) 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. It follows that $A_1 p = (\alpha, \beta, \gamma)$.

Similarly, the line $\overline{O'P}$ intersects π_1 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 $A_2 v' = e_4$. It follows that $A_2 p' = (\alpha', \beta', \gamma')$. 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 immediately follows that $k = s$. We have therefore, by the choice of putting both cameras on the frame of reference, that the transformation in $\mathcal{P}^3$ is affine (the plane π_1 is preserved). If we leave the first camera fixed and move the second camera to a new position (must be a general position, i.e., $O' \notin \pi_1$), then the transformation in $\mathcal{P}^3$ belongs to the same affine group. Note that since only ratios of coordinates are significant in $\mathcal{P}^n$, k is determined up to a uniform scale, and any point $P_o \notin \pi_1$ can be used to set a mutual scale for all views — by setting an appropriate scale for v', for example. The value of k can easily be determined 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} for which we get

$$\mu p' = Ap - k\boldsymbol{v}', \tag{1}$$

where $A = A_2^{-1} A_1$. Note that $A \in PGL_3$ is a collineation between the two image planes, due to π_1, determined by $p'_j \cong A p_j$, $j = 1,2,3$, and $Av = v'$ (therefore, can be recovered directly without going through A_1, A_2). 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 $p'_o \cong Ap_o - v'$. Then k is an affine invariant, which we will refer to as "affine depth". Furthermore, $(x, y, 1, k)$ are the homogeneous coordinates representation of P, and the 3×4 matrix $[A, -v']$ is a camera transformation matrix between the two views. Note that k is invariant when computed against a reference view (the first view in this derivation), the camera transformation matrix does not only depend on the camera displacement but on the choice of three points, and the camera is an "affine engine" if a reference view is available. More details on theoretical aspects of this result are provided in Section 3.2, but first we discuss its algorithmic aspect.

3.1 Two Algorithms: Re-projection and Affine Reconstruction from Two Perspective Views

On the practical side, we have arrived to a remarkably simple algorithm for affine reconstruction from two perspective/orthographic views (with an uncalibrated camera), and an algorithm for generating novel views of a scene (re-projection). For reconstruction we follow these steps:

1. Compute epipoles v, v' (see Section 2.6).
2. Compute the matrix A that satisfies $Ap_j \cong p'_j$, $j = 1,2,3$, and $Av \cong v'$. This requires a solution of a linear system of eight equations (see Appendices in [19, 27, 28] for details).
3. Set the scale of v' by using a fourth corresponding pair p_o, p'_o such that $p'_o \cong Ap_o - v'$.
4. For every corresponding pair p, p' recover the affine depth k that satisfies $p' \cong Ap - k\boldsymbol{v}'$. As a technical note, k can be recovered in a least-squares fashion by using cross-products:

$$k = \frac{(p' \times v')^T (p' \times Ap)}{\| p' \times v' \|^2}.$$

Note that k is invariant as long as we use the first view as a reference view, i.e., compute k between a reference view p and any other view. The invariance of k can be used to "re-project" the object onto any third view p'', as follows. We observe:

$$p'' \cong Bp - k\boldsymbol{v}'',$$

for some (unique up to a scale) matrix B and epipole v''. One can solve for B and v'' by observing six corresponding points between the first and third view. Each pair of corresponding points p_j, p''_j contributes two equations:

$$b_{31}x_jx''_j + b_{32}y_jx''_j - k_jv''_3x''_j + x''_j = b_{11}x_j + b_{12}y_j + b_{13} - k_jv''_1,$$

$$b_{31}x_jy''_j + b_{32}y_jy''_j - k_jv''_3y''_j + y''_j = b_{21}x_j + b_{22}y_j + b_{23} - k_jv''_2,$$

where $b_{33} = 1$ (this for setting an arbitrary scale because the system of equations is homogeneous — of course this prevents the case where $b_{33} = 0$, but in practice this is not a problem; also one can use principal component analysis instead of setting the value of some chosen element of B or v''). The values of k_j are found from the correspondences p_j, p'_j, $j = 1, ..., 6$ (note that $k_1 = k_2 = k_3 = 0$). Once B, v'' are recovered, we can find the location of p''_i for any seventh point p_i, by first solving for k_i from the equation $p'_i \cong Ap_i - k_iv'$, and then substituting the result in the equation $p''_i \cong Bp_i - k_i\boldsymbol{v}''$.

3.2 Results of Theoretical Nature

Let $\psi_o \in \mathcal{S}$ be some view from the set of all possible views, and let $p_1, p_2, p_3 \in \psi_o$ be non-collinear points projected from some plane π. Also, let $\mathcal{S}_\pi \subset \mathcal{S}$ be the subset of views for which the corresponding pairs of p_j, $j = 1, 2, 3$, are non-collinear (A is full rank). Note that $\mathcal{S}_\pi$ contains all views for which the COP is not on π. We have the following result:

There exists an affine invariant between a reference view ψ_o and the set of views $\mathcal{S}_\pi$.

The result implies that, within the framework of uncalibrated cameras, there are certain tasks which are inherently affine and, therefore, projective invariants are not necessary and instead affine invariants are sufficient (it is yet to be shown when exactly do we need to recover projective invariants — this is the subject of Section 4). Consider for example the task of recognition within the context of alignment [30, 11]. In the alignment approach, two or more reference views (also called model views), or a 3D model, are stored in memory — and referred to as a "model" of the object. During the recognition process, a small number of corresponding points between the reference views and the novel view are used for "re-projecting" the object onto the novel viewing position (as for example using the method described in the previous section). Recognition is achieved if the re-projected image is successfully matched against the input image. This entails a sequential search over all possible models until a match is found between the novel view and the re-projected view using a particular model. The implication of the result above is that since alignment uses

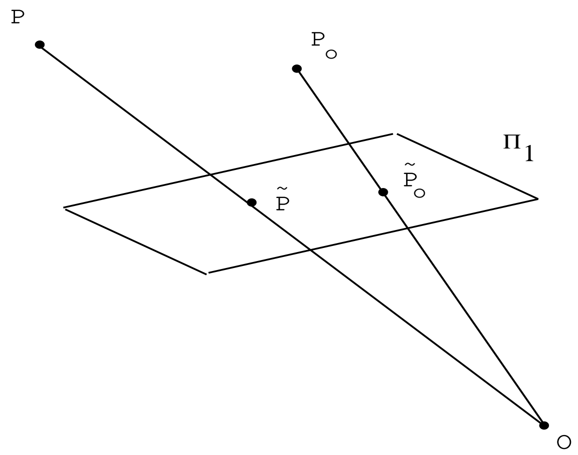

Figure 2:

a fixed set of reference views of an object to perform recognition, then only affine machinery is really necessary to perform re-projection. As will be shown in Section 4, projective machinery requires more points and slightly more computations (but see Section 9 for discussion about practical considerations).

The manner in which affine-depth was derived gives rise to a refinement on the general result that four corresponding points and the epipoles are required for affine reconstruction from two perspective views [4, 29]. Our derivation shows that in addition to the epipoles, we need only three points to recover affine structure up to a uniform scale, and therefore the fourth point is needed only for setting such a scale. To summarize,

In case where the location of epipoles are known, then three corresponding points are sufficient for computing the affine structure, up to a uniform but unknown scale, for all other points in space projecting onto corresponding points in both views.

We have also,

Affine shape can be described as the ratio of a point P from a plane and the COP, normalized by the ratio of a fixed point from the reference plane and the COP.

Therefore, affine-depth k depends only three points (setting up a reference plane), the COP (of the reference view) and a fourth point for setting a scale. This way of describing structure relative to a reference plane is very similar to what [14] suggested for reconstruction from two orthographic views. The difference is that there the fourth point played the role of both the COP and for setting a scale. We will show next that the affine-depth structure description derived here reduces exactly to what [14] described in the orthographic case.

There are two ways to look at the orthographic case. First, when both views are orthographic, the collineation A (in Equation 1) between the two images is an affine transformation in $\mathcal{P}^2$, i.e., third row of A is $(0, 0, 1)$. Therefore, A can be computed from only three corresponding points, $Ap_j \cong p'_j$, $j = 1, 2, 3$. Because both O and O' are at infinity, then the epipole v' is on the plane $x_2 = 0$, i.e., $v'_3 = 0$, and as a result all epipolar lines are parallel to each other. A fourth corresponding point p_o, p'_o can be used to determine both the direction of epipolar lines and to set the scale for the affine depth of all other points — as described in [14]. We see, therefore, that the orthographic case is simply a particular case of Equation 1. Alternatively, consider again the structure description entailed by our derivation of affine depth. If we denote the point of intersection of the line $\overline{OP}$ with π_1 by $\tilde{P}$, we have (see Figure 2)

$$k = \frac{\frac{P-\tilde{P}}{P-O}}{\frac{P_o-\tilde{P}_o}{P_o-O}}.$$

Let O (the COP of the first camera) go to infinity, in which case affine-depth approaches

$$k \longrightarrow \frac{P-\tilde{P}}{P_o-\tilde{P}_o},$$

which is precisely the way shape was described in [14] (see also [26, 27]). In the second view, if it is orthographic, then the two trapezoids $P, \tilde{P}, p', Ap$ and $P_o, \tilde{P}_o, p'_o, Ap_o$ are similar, and from similarity of trapezoids we obtain

$$\frac{P-\tilde{P}}{P_o-\tilde{P}_o} = \frac{p'-Ap}{p'_o-Ap_o},$$

which, again, is the expression described in [14, 26]. Note that affine-depth in the orthographic case does not depend any more on O, and therefore remains fixed regardless of what pair of views we choose, namely, a reference view is not necessary any more. This leads to the following result:

Let $\bar{\mathcal{S}} \subset \mathcal{S}$ be the subset of views created by means of parallel projection, i.e., the plane $x_2 = 0$ is preserved. Given four fixed reference points, affine-depth on $\mathcal{S}$ is reference-view-dependent, whereas affine-depth on $\bar{\mathcal{S}}$ is reference-view-independent.

Consider next the resulting camera transformation matrix $[A, -v']$. The matrix A depends on the choice of three points and therefore does not only depend on the camera displacement. This additional degree of freedom is a direct result of our camera being uncalibrated, i.e., we are free to choose the internal camera parameters (focal length, principal point, and image coordinates scale factors) as we like. The matrix A is unique, i.e., depends only on camera displacement, if we know in advance that the internal camera parameters remain fixed for all views $\mathcal{S}_\pi$. For example, assume the camera is calibrated in the usual manner, i.e., focal length is 1, principle point is at $(0, 0, 1)$ in Euclidean coordinates, and image scale factors are 1 (image plane is parallel to xy plane of Euclidean coordinate system). In that case A is an orthogonal matrix and can be recovered from two corresponding points and the epipoles — by imposing the constraint that vector magnitudes remain unchanged (each point provides three equations). A third corresponding point can be used to determine the reflection component (i.e., making sure the determinant of A is 1 rather than -1). More details can be found in [27, 15]. Since in the uncalibrated case A is not unique, let A_π denote the fact that A is the collineation induced by a plane π, and let k_π denote the fact that the affine-depth also depends on the choice of π. We see, therefore, that there exists a family of solutions for the camera transformation matrix and the affine-depth as a function of π. This immediately implies that a naive solution for A, k, given v', from point correspondences leads to a singular system of equations (even if many points are used for a least-squares solution).

Given the epipole v', the linear system of equations for solving for A and k_j of the equation

$$\mu p'_j = Ap_j - k_j \boldsymbol{v}',$$

from point correspondences p_j, p'_j is singular, unless further constraints are introduced.

We see that equation counting alone is not sufficient for obtaining a unique solution, and therefore the knowledge that A is a homography of a plane is critical for this task. For example, one can solve for A and k_j from many correspondences in a least-squares approach by first setting $k_j = 0$, $j = 1, 2, 3$ and $k_4 = 1$, otherwise the solution may not be unique.

Finally, consider the "price" we are paying for an uncalibrated, affine framework. We can view this in two ways, somewhat orthogonal. First, if the scene is undergoing transformations, and the camera is fixed, then those transformations are affine in 3D, rather than rigid. For purposes of achieving visual recognition the price we are paying is that we might confuse two different objects that are affinely related. Second, because of the non-uniqueness of the camera transformation matrix it appears that the set of views $\mathcal{S}_\pi$ is a superset of the set of views that could be created by a calibrated camera taking pictures of the object. The natural question is whether this superset can, nevertheless, be realized by a calibrated camera. In other words, if we have a calibrated camera (or we know that the internal camera parameters remain fixed for all views), then can we generate $\mathcal{S}_\pi$, and if so how? This question was addressed first in [12] but assuming only orthographic views. A more general result is expressed in the following proposition:

Proposition 1 *Given an arbitrary view $\psi_o \in \mathcal{S}_\pi$ generated by a camera with COP at initial position O, then all other views $\psi \in \mathcal{S}_\pi$ can be generated by a rigid motion of the camera frame from its initial position, if in addition to taking pictures of the object we allow any finite sequence of pictures of pictures to be taken as well.*

The proof has a trivial and a less trivial component. The trivial part is to show that an affine motion of the camera frame can be decomposed into a rigid motion followed by some arbitrary collineation in $\mathcal{P}^2$. The less trivial component is to show that any collineation in $\mathcal{P}^2$

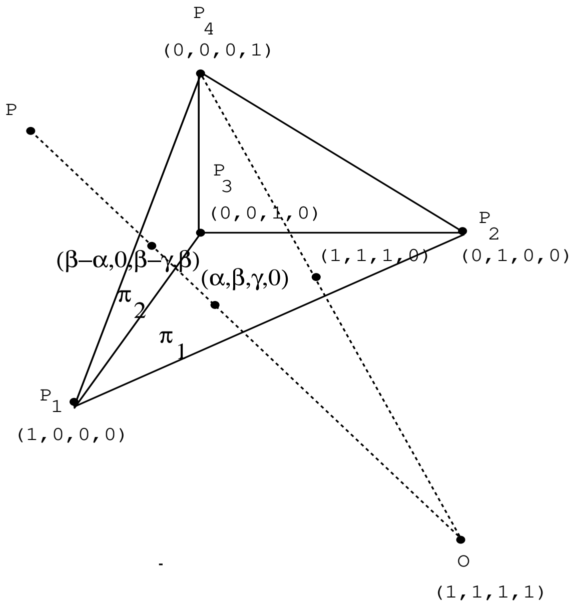

Figure 3:

can be created by a finite sequence of views of a view where only rigid motion of the camera frame is allowed. The details can be found in Appendix A.

The next section treats the projective case. It will be shown that this involves looking for invariants that remain fixed when any two views of $\mathcal{S}$ are chosen. The section may be skipped if the reader wishes to get to Part II of the paper — only results of affine-depth are used there.

4 Projective Structure and Invariant From Two Perspective Views

Affine depth required the construction of a single reference plane, and for that reason it was necessary to require that one view remained fixed to serve as a reference view. To permit an invariant from any pair of views of $\mathcal{S}$, we should, by inference, design the construction such that the invariant be defined relative to two planes. By analogy, we will call the invariant "projective depth" [29]. This is done as follows.

We assign the coordinates $(1,0,0,0),(0,1,0,0)$ and $(0,0,1,0)$ to P_1,P_2,P_3, respectively. The coordinates $(0,0,0,1)$ are assigned to a fourth point P_4, and the coordinates $(1,1,1,1)$ to the COP of the first camera O (see Figure 3). The plane passing through P_1,P_2,P_3 is denoted by π_1 (as before), and the plane passing through P_1,P_3,P_4 is denoted by π_2. Note that the line $\overline{OP_4}$ intersects π_1 at $(1,1,1,0)$, and the line $\overline{OP_2}$ intersects π_2 at $(1,0,1,1)$.

As before, let A_1 be the collineation from the image plane to π_1 by satisfying $A_1p_j \cong e_j$, $j=1,...,4$, where $e_1=(1,0,0), e_2=(0,1,0), e_3=(0,0,1)$ and $e_4=(1,1,1)$. Similarly, let E_1 be the collineation from the image plane to π_2 by satisfying $E_1p_1 \cong e_1, E_1p_2 \cong e_4, E_1p_3 \cong e_2$ and $E_1p_4 \cong e_3$. Note that if $A_1p=(\alpha,\beta,\gamma)$, then $E_1p=(\beta-\alpha,\beta-\gamma,\beta)$. We have therefore, that the intersection of the line $\overline{OP}$ with π_1 is the point $P_{\pi_1}=(\alpha,\beta,\gamma,0)$, and the intersection with π_2 is the point $P_{\pi_2}=(\beta-\alpha,0,\beta-\gamma,\beta)$. We can express P and O as a linear combination of those points:

$$P \cong \begin{pmatrix}\alpha\\ \beta\\ \gamma\\ 0\end{pmatrix} + \kappa\begin{pmatrix}\beta-\alpha\\ 0\\ \beta-\gamma\\ \beta\end{pmatrix},$$

$$\begin{pmatrix}1\\ 1\\ 1\\ 1\end{pmatrix} \cong \begin{pmatrix}\alpha\\ \beta\\ \gamma\\ 0\end{pmatrix} + \kappa'\begin{pmatrix}\beta-\alpha\\ 0\\ \beta-\gamma\\ \beta\end{pmatrix}$$

Consider the cross ratio κ/κ' of the four points O,P_{π_1},P_{π_2},P. Note that $\kappa'=1$ independently of P, therefore the cross ratio is simply κ. As in the affine case, κ is invariant up to a uniform scale, and any fifth object point P_o (not lying on any face of the tetrahedron P_1,P_2,P_3,P_4) can be assigned $\kappa_o=1$ by choosing the appropriate scale for A_1 (or E_1). This has the effect of mapping the fifth point P_o onto the COP ($P_o \cong (1,1,1,1)$). We have, therefore, that κ (normalized) is a projective invariant, which we call "projective depth". Relative shape is described as the ratio of a point from two planes, defined by four object points, along the line to a fifth point, which is also the center of projection, that is set up such that its ratio from the two planes is of unit value. Any transformation $T \in GL_4$ will leave the ratio κ invariant. What remains is to show how κ can be computed given a second view.

Let A be the collineation between the two image planes due to π_1, i.e., $Ap_j \cong p'_j$, $j=1,2,3$, and $Av \cong v'$, where v,v' are the epipoles. Similarly, let E be the collineation due to π_2, i.e., $Ep_j \cong p'_j$, $j=1,3,4$, and $Ev \cong v'$. Note that three corresponding points and the corresponding epipoles are sufficient for computing the collineation due to the plane projecting onto the three points in both views — this is clear from the derivation in Section 3, but also can be found in [28, 29, 23]. We have that the projections of P_{π_1} and P_{π_2} onto the second image are captured by Ap and Ep, respectively. Therefore, the cross ratio of O,P_{π_1},P_{π_2},P is equal to the cross ratio of v',Ap,Ep,p', which is computed as follows:

$$p' \cong Ap - sEp,$$
$$v' \cong Ap - s'Ep,$$

then $\kappa = s/s'$, up to a uniform scale factor (which is set using a fifth point). Here we can also show that s' is a constant independent of p. There is more than one way to show that, a simple way is as follows: Let q be an arbitrary point in the first image. Then,

$$v' \cong Aq - s'_qEq.$$

Let H be a matrix defined by $H=A-s'_qE$. Then, $v' \cong Hv$ and $v'=Hq$. This could happen only if $v' \cong Hp$, for all p, and $s'=s'_q$. We have arrived to a very simple algorithm for recovering a projective invariant from two perspective (orthographic) views:

$$p' \cong Ap - \kappa Ep, \tag{2}$$

where A and E are described above, and κ is invariant up to a uniform scale, which can be set by observing a fifth correspondence p_o, p'_o, i.e., set the scale of E to satisfy $p'_o \cong Ap_o - Ep_o$. Unlike the affine case, κ is invariant for any two views from the set $\mathcal{S}$ of all possible views. Note that κ need not be normalized using a fifth point, if the first view remains fixed (we are back to the affine case). We have arrived to the following result, which is a refinement on the general result made in [4] that five corresponding points and the corresponding epipoles are sufficient for reconstruction up to a collineation in $\mathcal{P}^3$:

In 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 projective structure, up to a uniform but unknown scale, for all other points in space projecting onto corresponding points in both views. A fifth corresponding point, coming from a point in general position with the other four points, can be used to set the scale.

We have also,

Projective shape can be described as the ratio of a point P from two faces of the tetrahedron, normalized by the ratio of a fixed point (the unit point of the reference frame) from those faces.

The practical implication of this derivation is that a projective invariant, such as the one described here, is worthwhile computing for tasks for which we do not have a fixed reference view available. Worthwhile because projective depth requires an additional corresponding point, and requires slightly more computations (recover the matrix E in addition to A). Such a task, for example, is to update the reconstructed structure from a moving stereo rig. At each time instance we are given a pair of views from which projective depth can be computed (projective coordinates follow trivially), and since both cameras are changing their position from one time instant to the next, we cannot rely on an affine invariant.

5 Summary of Part I

Given a view ψ_o with image points p, there exists an affine invariant k between ψ_o and any other view ψ_i, with corresponding image points p', satisfying the following equation:

$$\mu p' = Ap - k\boldsymbol{v}',$$

where A is the collineation between the two image planes due to the projection of some plane π_1 projecting to both views, and v' is the epipole scaled such that $\mu_o p'_o = Ap_o - v'$ for some point p_o. The set of all views $\mathcal{S}_{\pi_1}$ for which the camera's center is not on π_1 will satisfy the equation above against ψ_o. The view ψ_o is a reference view.

A projective invariant κ is defined between any two views ψ_i and ψ_j, again for the sake of not introducing new notations, projecting onto corresponding points p and p', respectively. The invariant satisfies the following equation:

$$\mu p' = Ap - \kappa Ep,$$

where A is the collineation due to some plane π_1, and E is the collineation due to some other plane π_2 scaled such that $\mu_o p'_o = Ap_o - Ep_o$, for some point p_o.

Part II

6 Algebraic Functions of Views

In this part of the paper we use the results established in Section 3 to derive results of a different nature: instead of reconstruction of shape and invariants we would like to establish a direct connection between views expressed as a functions of image coordinates alone — which we will call "algebraic functions of views". With these functions one can manipulate views of an object, such as create new views, without the need to recover shape or camera geometry as an intermediate step — all what is needed is to appropriately combine the image coordinates of two reference views.

Algebraic functions of two views include the expression

$$p'^T Fp = 0, \tag{3}$$

where F is known as the "Fundamental" matrix (cf. [4]) (a projective version of the well known "Essential" matrix of [16]), and the expression

$$\alpha_1 x' + \alpha_2 y' + \alpha_3 x + \alpha_4 y + \alpha_5 = 0 \tag{4}$$

due to [10], which is derived for orthographic views. These functions express the epipolar geometry between the two views in the perspective and orthographic cases, respectively. Algebraic functions of three views were introduced in the past only for orthographic views [31, 21]. For example,

$$\alpha_1 x'' + \alpha_2 x' + \alpha_3 x + \alpha_4 y + \alpha_5 = 0.$$

These functions express a relationship between the image coordinates of one view as a function of image coordinates of two other views — in the example above, the x coordinate in the third view, x'', is expressed as a linear function of image coordinates in two other views, similar expressions exist for y''.

We will use the affine-depth invariant result to derive algebraic functions of three perspective views. The relationship between a perspective view and two other perspective views is shown to be trilinear in image coordinates across the three views. The relationship is shown to be bilinear if two of the views are orthographic — a special case useful for recognition tasks. We will start by addressing the two-view case. We will use Equation 1 to relate the entries of the camera transformation A and v' (of Equation 1) to the fundamental matrix by showing that $F = v' \times A$. This also has an advantage of introducing an alternative way of deriving expressions 3 and 4, a way that also puts them both under a single framework.

6.1 Algebraic Functions of Two Views

Consider Equation 1, reproduced below,

$$\rho \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = A \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} - k\boldsymbol{v}'.$$

By simple manipulation of this equation we obtain:

$$k = \frac{v'_1 - x'v'_3}{x'\boldsymbol{a}_3 \cdot p - \boldsymbol{a}_1 \cdot p} = \frac{v'_2 - y'v'_3}{y'\boldsymbol{a}_3 \cdot p - \boldsymbol{a}_2 \cdot p} = \frac{y'v'_1 - x'v_2}{x'\boldsymbol{a}_2 \cdot p - y'\boldsymbol{a}_1 \cdot p}, \quad (5)$$

where $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ are the row vectors of A and $v' = (v'_1, v'_2, v'_3)$. After equating the first two terms, we obtain:

$$x'(v'_2\boldsymbol{a}_3 \cdot p - v'_3\boldsymbol{a}_2 \cdot p) + y'(v'_3\boldsymbol{a}_1 \cdot p - v'_1\boldsymbol{a}_3 \cdot p) + (v'_1\boldsymbol{a}_2 \cdot p - v'_2\boldsymbol{a}_1 \cdot p) = 0. \quad (6)$$

Note that the terms within parentheses are linear polynomials in x, y with fixed coefficients (i.e., depend only on A and v'). Also note that we get the same expression when equating the first and third, or the second and third terms of Equation 5. This leads to the following result:

The image coordinates (x, y) and (x', y') of two corresponding points across two perspective views satisfy a unique equation of the following form:

$$x'(\alpha_1 x + \alpha_2 y + \alpha_3) + y'(\alpha_4 x + \alpha_5 y + \alpha_6) + \alpha_7 x + \alpha_8 y + \alpha_9 = 0, \quad (7)$$

where the coefficients α_j, $j = 1, ..., 9$, have a fixed relation to the camera transformation A and v' of Equation 1:

$$\begin{aligned}
\alpha_1 &= v'_2 a_{31} - v'_3 a_{21}, \\
\alpha_2 &= v'_2 a_{32} - v'_3 a_{22}, \\
\alpha_3 &= v'_2 a_{33} - v'_3 a_{23}, \\
\alpha_4 &= v'_3 a_{11} - v'_1 a_{31}, \\
\alpha_5 &= v'_3 a_{12} - v'_1 a_{32}, \\
\alpha_6 &= v'_3 a_{13} - v'_1 a_{33}, \\
\alpha_7 &= v'_1 a_{21} - v'_2 a_{11}, \\
\alpha_8 &= v'_1 a_{22} - v'_2 a_{12}, \\
\alpha_9 &= v'_1 a_{23} - v'_2 a_{13}.
\end{aligned}$$

Equation 7 can also be written as ${p'}^t F p = 0$, where the entries of the matrix F are the coefficients α_j, and therefore, $F = v' \times A$. We have, thus, obtained a new and simple relationship between the elements of the "fundamental" matrix F and the elements of the camera transformation A and v'. It is worth noting that this result can be derived much easier, as follows. First, the relationship ${p'}^t F p = 0$ can be derived, as observed by [4], from the fact that F is a correlation mapping points p onto their corresponding epipolar lines l' in the second image, and therefore $p' \cdot l' = 0$. Second[1], since $l' \cong v' \times Ap$, we have $F = v' \times A$. It is known that the rank of the fundamental matrix is 2; we can use this relationship to show that as well:

$$F = v' \times A = \begin{bmatrix} v'_2\boldsymbol{a}_3 - v'_3\boldsymbol{a}_2 \\ v'_3\boldsymbol{a}_1 - v'_1\boldsymbol{a}_3 \\ v'_1\boldsymbol{a}_2 - v'_2\boldsymbol{a}_1 \end{bmatrix},$$

where $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$ are the row vectors of A. Let $\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3$ be the row vectors of F, then it is easy to verify that

$$\boldsymbol{f}_3 \cong \alpha\boldsymbol{f}_1 + \beta\boldsymbol{f}_2,$$

by setting

$$\alpha v'_2 = \beta v'_1.$$

Next, we can use the result $F = v' \times A$ to show how the orthographic case, treated by [10], fits this relationship. In the framework of Equation 1, we saw that with orthographic views we have A being affine in $\mathcal{P}^2$, i.e., $a_3 \cdot p = 1$, and $v'_3 = 0$. After substitution in Equation 6, we obtain the equation:

$$\alpha_1 x' + \alpha_2 y' + \alpha_3 x + \alpha_4 y + \alpha_5 = 0, \quad (8)$$

where the coefficients α_j, $j = 1, ..., 5$ have the following values:

$$\begin{aligned}
\alpha_1 &= v'_2, \\
\alpha_2 &= -v'_1, \\
\alpha_3 &= v'_1 a_{21} - v'_2 a_{11}, \\
\alpha_4 &= v'_1 a_{22} - v'_2 a_{12}, \\
\alpha_5 &= v'_1 a_{23} - v'_2 a_{13}.
\end{aligned}$$

These coefficients are also the entries of the fundamental matrix, which can also be derived from $F = v' \times A$ by setting $v'_3 = 0$ and $\boldsymbol{a}_3 = (0, 0, 1)$.

The algebraic function 7 can be used for re-projection onto a third view, by simply noting that the function between view 1 and 3, and the function between view 2 and 3, provide two equations for solving for (x'', y''). This was proposed in the past, in various forms, by [20, 3, 19]. Since the algebraic function expresses the epipolar geometry between the two views, however, a solution can be found only if the COPs of the three cameras are non-collinear (cf. [28, 27]) — which can lead to numerical instability unless the COPs are far from collinear. The alternative, as shown next, is to derive directly algebraic functions of three views. In that case, the coordinates (x'', y'') are solved for separately, each from a single equation, without problems of singularities.

6.2 Algebraic Functions of Three Views

Consider Equation 1 applied between view 1 and 2, and between view 1 and 3:

$$\begin{aligned}
\mu p' &= Ap - k\boldsymbol{v}' \\
\nu p'' &= Bp - k\boldsymbol{v}''.
\end{aligned} \quad (9)$$

Here we make use of the result that affine-depth k is invariant for any view in reference to the first view. We can isolate k again from Equation 9 and obtain:

$$k = \frac{v''_1 - x''v''_3}{x''\boldsymbol{b}_3 \cdot p - \boldsymbol{b}_1 \cdot p} = \frac{v''_2 - y''v''_3}{y''\boldsymbol{b}_3 \cdot p - \boldsymbol{b}_2 \cdot p} = \frac{y''v''_1 - x''v''_2}{x''\boldsymbol{b}_2 \cdot p - y''\boldsymbol{b}_1 \cdot p}, \quad (10)$$

where $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$ are the row vectors of B and $v'' = (v''_1, v''_2, v''_3)$. Because of the invariance of k we can equate terms of Equation 5 with terms of Equation 10 and obtain trilinear functions of image coordinates across three

[1] This was a comment made by Tuan Luong.

views. For example, by equating the first two terms in each of the equations, we obtain:

$$x''(v'_1 \boldsymbol{b}_3 \cdot p - v''_3 \boldsymbol{a}_1 \cdot p) + x''x'(v''_3 \boldsymbol{a}_3 \cdot p - v'_3 \boldsymbol{b}_3 \cdot p) \quad + \\ x'(v'_3 \boldsymbol{b}_1 \cdot p - v''_1 \boldsymbol{a}_3 \cdot p) + v''_1 \boldsymbol{a}_1 \cdot p - v'_1 \boldsymbol{b}_1 \cdot p = 0. \quad (11)$$

This leads to the following result:

The image coordinates (x, y), (x', y') and (x'', y'') of three corresponding points across three perspective views satisfy a trilinear equation of the following form:

$$x''(\alpha_1 x + \alpha_2 y + \alpha_3) + x''x'(\alpha_4 x + \alpha_5 y + \alpha_6) \quad + \\ x'(\alpha_7 x + \alpha_8 y + \alpha_9) + \alpha_{10} x + \alpha_{11} y + \alpha_{12} = 0, \quad (12)$$

where the coefficients α_j, $j = 1, ..., 12$, have a fixed relation to the camera transformations between the first view and the other two views.

Note that the x coordinate in the third view, x'', is obtained as a solution of a single equation in coordinates of the other two views. The coefficients α_j can be recovered as a solution of a linear system, directly if we observe 11 corresponding points across the three views (more than 11 points can be used for a least-squares solution), or with fewer points by first recovering the elements of the camera transforms as described in Section 3. Then, for any additional point (x, y) whose correspondence in the second image is known (x', y'), we can recover the corresponding x coordinate, x'', in the third view by substitution in equation 12.

In a similar fashion, after equating the first term of Equation 5 with the second term of Equation 10, we obtain an equation for y'' as a function of the two other views:

$$y''(\beta_1 x + \beta_2 y + \beta_3) + y''x'(\beta_4 x + \beta_5 y + \beta_6) \quad + \\ x'(\beta_7 x + \beta_8 y + \beta_9) + \beta_{10} x + \beta_{11} y + \beta_{12} = 0. \quad (13)$$

Taken together, Equations 5 and 10 lead to 9 algebraic functions of three views, six of which are separate for x'' and y''. The other four functions are listed below:

$$x''(\cdot) + x''y'(\cdot) + y'(\cdot) + (\cdot) = 0, \quad (14)$$
$$y''(\cdot) + y''y'(\cdot) + y'(\cdot) + (\cdot) = 0, \quad (15)$$
$$x''x'(\cdot) + x''y'(\cdot) + x'(\cdot) + y'(\cdot) = 0, \quad (16)$$
$$y''x'(\cdot) + y''y'(\cdot) + x'(\cdot) + y'(\cdot) = 0, \quad (17)$$

where $(\cdot)$ represent linear polynomials in x, y. The solution for x'', y'' is unique without constraints on the allowed camera transformations. If we choose Equations 12 and 13, then v'_1 and v'_3 should not vanish simultaneously, i.e., $v' \cong (0, 1, 0)$ is a singular case. Also $v'' \cong (0, 1, 0)$ and $v'' \cong (1, 0, 0)$ give rise to singular cases. One can easily show that for each singular case there are two other functions out of the nine available ones that provide a unique solution for x'', y''. Note that the singular cases are pointwise, i.e., only three epipolar directions are excluded, compared to the much stronger singular case when the algebraic function of two views is used separately, as described in the previous section.

Taken together, the process of generating a novel view can be easily accomplished without the need to explicitly recover structure (affine depth), camera transformation (matrices A, B and epipoles v', v'') or epipolar geometry (just the epipoles or the Fundamental matrix) — for the price of using more than the minimal number points that are required otherwise (the minimal is six between the two model views and the novel third view).

The connection between the general result of trilinear functions of views to the "linear combination of views" result [31] for orthographic views, can easily be seen by setting A and B to be affine in $\mathcal{P}^2$, and $v'_3 = v''_3 = 0$. For example, Equation 11 reduces to:

$$v'_1 x'' - v''_1 x' + (v''_1 \boldsymbol{a}_1 \cdot p - v'_1 \boldsymbol{b}_1 \cdot p) = 0, \quad (18)$$

which is of the form:

$$\alpha_1 x'' + \alpha_2 x' + \alpha_3 x + \alpha_4 y + \alpha_5 = 0.$$

In the case where all three views are orthographic, then x'' is expressed as a linear combination of image coordinates of the two other views — as discovered by [31].

In the next section we address another case, intermediate between the general trilinear and the orthographic linear functions, which we find interesting for applications of visual recognition.

6.2.1 Recognition of Perspective views From an Orthographic Model

Consider the case for which the two reference (model) views of an object are taken orthographically (using a tele lens would provide a reasonable approximation), but during recognition any perspective view of the object is allowed. It can easily be shown that the three views are then connected via a bilinear function (instead of trilinear): A is affine in $\mathcal{P}^2$ and $v'_3 = 0$, therefore Equation 11 reduces to:

$$x''(v'_1 \boldsymbol{b}_3 \cdot p - v''_3 \boldsymbol{a}_1 \cdot p) + v''_3 x''x' - \\ v''_1 x' + (v''_1 \boldsymbol{a}_1 \cdot p - v'_1 \boldsymbol{b}_1 \cdot p) = 0,$$

which is of the following form:

$$x''(\alpha_1 x + \alpha_2 y + \alpha_3) + \alpha_4 x''x' + \\ \alpha_5 x' + \alpha_6 x + \alpha_7 y + \alpha_8 = 0. \quad (19)$$

Similarly, Equation 13 reduces to

$$y''(\beta_1 x + \beta_2 y + \beta_3) + \beta_4 y''x' + \\ \beta_5 x' + \beta_6 x + \beta_7 y + \beta_8 = 0. \quad (20)$$

A bilinear function of three views has two advantages over the general trilinear function. First, only seven corresponding points (instead of 11) across three views are required for solving for the coefficients (compared to the minimal six if we first recover A, B, v', v''). Second, the lower the degree of the algebraic function, the less sensitive the solution should be in the presence of errors in measuring correspondences. In other words, it is likely (though not necessary) that the higher order terms, such as the term $x''x'x$ in Equation 12, will have a higher contribution to the overall error sensitivity of the system.

Compared to the case when all views are assumed orthographic, this case is much less of an approximation. Since the model views are taken only once, it is not unreasonable to require that they be taken in a special way, namely, with a tele lens (assuming we are dealing with object recognition, rather than scene recognition). If that requirement is satisfied, then the recognition task is general since we allow any perspective view to be taken during the recognition process.

7 Applications

Algebraic functions of views allow the manipulation of images of 3D objects without necessarily recovering 3D structure or any form of camera geometry (either full, or weak — the epipoles).

The application that was emphasized throughout the paper is visual recognition via alignment. In this context, the general result of a trilinear relationship between views is not encouraging. If we want to avoid implicating structure and camera geometry, we must have 11 corresponding points across the three views — compared to six points, otherwise. In practice, however, we would need more than the minimal number of points in order to obtain a least squares solution. The question is whether the simplicity of the method using trilinear functions translates also to increased robustness in practice when many points are used — this is an open question.

Still in the context of recognition, the existence of bilinear functions in the special case where the model is orthographic, but the novel view is perspective, is more encouraging. Here we have the result that only seven corresponding points are required to obtain recognition of perspective views (provided we can satisfy the requirement that the model is orthographic) compared to six points when structure and camera geometry are recovered. The additional corresponding pair of points may be indeed worth the greater simplicity that comes with working with algebraic functions.

There may exist other applications where simplicity is of major importance, whereas the number of points is less of a concern. Consider for example, the application of model-based compression. With the trilinear functions we need 22 parameters to represent a view as a function of two reference views in full correspondence. Assume both the sender and the receiver have the two reference views and apply the same algorithm for obtaining correspondences between the two views. To send a third view (ignoring problems of self occlusions that could be dealt separately) the sender can solve for the 22 parameters using many points, but eventually send only the 22 parameters. The receiver then simply combines the two reference views in a "trilinear way" given the received parameters. This is clearly a domain where the number of points are not a major concern, whereas simplicity, and probably robustness due to the short-cut in the computations, is of great importance.

Related to image coding is a recent approach of image decomposition into "layers" as proposed in [1, 2]. In this approach, a sequence of views is divided up into regions, whose motion of each is described approximately by a 2D affine transformation. The sender sends the first image followed only by the six affine parameters for each region for each subsequent frame. The use of algebraic functions of views can potentially make this approach more powerful because instead of dividing up the scene into planes (it would have planes if the projection was parallel, in general its not even planes) one can attempt to divide the scene into objects, each carries the 22 parameters describing its displacement onto the subsequent frame.

Another area of application may be in computer graphics. Re-projection techniques provide a short-cut for image rendering. Given two fully rendered views of some 3D object, other views (again ignoring self-occlusions) can be rendered by simply "combining" the reference views. Again, the number of corresponding points is less of a concern here.

8 Summary of Part II

The derivation of an affine invariant across perspective views in Section 3 was used to derive algebraic functions of image coordinates across two and three views. These enable the generation of novel views, for purposes of visual recognition and for other applications, without going through the process of recovering object structure (metric or non-metric) and camera geometry.

Between two views there exists a unique function whose coefficients are the elements of the Fundamental matrix and were shown to be related explicitly to the camera transformation A, v':

$$x'(\alpha_1 x + \alpha_2 y + \alpha_3) + y'(\alpha_4 x + \alpha_5 y + \alpha_6) + \alpha_7 x + \alpha_8 y + \alpha_9 = 0.$$

The derivation was also useful in making the connection to a similar expression, due to [10], made in the context of orthographic views.

We have seen that trilinear functions of image coordinates exist across three views, one of them shown below:

$$x''(\alpha_1 x + \alpha_2 y + \alpha_3) + x''x'(\alpha_4 x + \alpha_5 y + \alpha_6) + x'(\alpha_7 x + \alpha_8 y + \alpha_9) + \alpha_{10} x + \alpha_{11} y + \alpha_{12} = 0.$$

In case two of the views are orthographic, a bilinear relationship across three views holds. For example, the trilinear function above reduces to:

$$x''(\alpha_1 x + \alpha_2 y + \alpha_3) + \alpha_4 x'' x' + \alpha_5 x' + \alpha_6 x + \alpha_7 y + \alpha_8 = 0.$$

In case all three views are orthographic, a linear relationship holds — as observed in [31]:

$$\alpha_1 x'' + \alpha_2 x' + \alpha_3 x + \alpha_4 y + \alpha_5 = 0.$$

9 General Discussion

For purposes of visual recognition, by alignment, the transformations induced by changing viewing positions is at most affine. In other words, a pin-hole uncalibrated camera is no more than an "affine engine" for tasks for which a reference view (a model) is available. One of the goals of this paper was to make this claim and make use of it in providing methods for affine reconstruction and for recognition.

An affine reconstruction follows immediately from Equation 1 and the realization that A is a collineation of some plane which is fixed for all views. The reconstructed homogeneous coordinates are $(x, y, 1, k)$ where

$(x, y, 1)$ are the homogeneous coordinates of the image plane of the reference view, and k is an affine invariant. The invariance of k can be used to generate novel views of the object (which are all affinely related to the reference view), and thus achieve recognition via alignment. We can therefore distinguish between affine and non-affine transformations in the context of recognition: if the object is fixed and the transformations are induced by camera displacements, then k must be invariant — space of transformations is no more than affine. If, however, the object is allowed to transform as well, then k would not remain fixed if the transformation is not affine, i.e. involves more than translation, rotation, scaling and shearing. For example, we may apply a projective transformation in $\mathcal{P}^3$ to the object representation, i.e., map five points (in general position) to arbitrary locations in space (which still remain in general position) and map all other points accordingly. This mapping allows more "distortions" than affine transformations allow, and can be detected by the fact that k will not remain fixed.

Another use of the affine derivations was expressed in Part II of this paper, by showing the existence of algebraic functions of views. We have seen that any view can be expressed as a trilinear function with two reference views in the general case, or as a bilinear function when the reference views are created by means of parallel projection. These functions provide alternative, much simpler, means for manipulating views of a scene. The camera geometries between one of the reference views and the other two views are folded into 22 coefficients. The number 22 is perfectly expected because these camera geometries can be represented by two camera transformation matrices, and we know that a camera transformation matrix has 11 free parameters (3×4 matrix, determined up to a scale factor). However, the folding of the camera transformations are done in such a way that we have two independent sets of 11 coefficients each, and each set contains foldings of elements of both camera transformation matrices (recall Equation 11). This enables us to recover the coefficients from point correspondences alone, ignoring the 3D structure of the scene. Because of their simplicity, we believe that these algebraic functions will find uses in tasks other than visual recognition — some of those are discussed in Section 7.

This paper is also about projective invariants, making the point of when do we need to recover a projective invariant, what additional advantages should we expect, and what price is involved (more computations, more points, etc.). Before we discuss those issues, it is worth discussing a point or two related to the way affine-depth was derived. Results put aside, Equation 1 looks suspiciously similar, or trivially derivable from, the classic motion equation between two frames. Also, there is the question of whether it was really necessary to use the tools of projective geometry for a result that is essentially affine. Finally, one may ask whether there are simpler derivations of the same result. Consider the classic motion equation for a calibrated camera:

$$z' p' = zRp + t.$$

Here R is an orthogonal matrix accounting for the rotational component of camera displacement, $\boldsymbol{t}$ is the translation component (note that $\boldsymbol{t} \cong \boldsymbol{v}'$), z is the depth from the first camera frame, and z' is the depth value seen from the second camera frame. Divide both sides of the equation by z, assume that R is an arbitrary non-singular matrix A, and it seems that we have arrived to Equation 1, where $k = -1/z$. In order to do it right, one must start with an affine frame, map it affinely onto the first camera, then map it affinely onto the second camera, and then relate the two mappings together — it will then become clear that k is an invariant measurement. This derivation, which we will call an "affine derivation", appears to have the advantage of not using projective geometry. However, there are some critical pieces missing. First, and foremost, we have an equation but not an algorithm. We have seen that simple equation counting for solving for A and k, given $\boldsymbol{t}$, from point correspondences is not sufficient, because the system of equations is singular for any number of corresponding points. Also, equation counting does not reveal the fact that only four points are necessary: three for A and the fourth for setting a mutual scale. Therefore, the realization that A is a homography of some plane that is fixed along all views — a fact that is not revealed by the affine derivation — is crucial for obtaining an algorithm. Second, the nature of the invariant measurement k is not completely revealed; it is not (inverse) depth because A is not necessarily orthogonal, and all the other results described in Section 3.2 do not clearly follow either.

Consider next the question of whether, within the context of projective geometry, affine-depth could have been derived on geometric grounds without setting up coordinates, as we did. For example, although this was not mentioned in Section 3, it is clear that the three points p', Ap, v' are collinear — this is well known and can be derived from purely geometric considerations by observing that the optical line $\overline{OP}$ and the epipolar line $\overline{p'v'}$ are projectively related in $\mathcal{P}^1$ (cf. [28, 29, 22]). It is less obvious, however, to show on geometric grounds only that the ratio k is invariant independently of where the second view is located, because ratios are not generally preserved under projectivity (only cross-ratios are). In fact, as we saw, k is invariant but up to a uniform scale, therefore, for any particular optical line the ratio is not preserved. It is for this reason that algebra was introduced in Section 3 for the derivation of affine-depth.

Consider next the difference between the affine and the projective frameworks. We have seen that from a theoretical standpoint, a projective invariant, such as projective-depth κ in Equation 2, is really necessary when a reference view is not available. For example, assume we have a sequence of n views $\psi_o, \psi_1, ..., \psi_{n-1}$ of a scene and we wish to recover its 3D structure. An affine framework would result if we choose one of the views, say ψ_o, as a reference view, and compute the structure as seen from that camera location given the correspondences $\psi_o \Longrightarrow \psi_i$ with all the remaining views — this is a common approach for recovering metric structure from a sequence. Because affine-depth is invariant, we have $n-1$ occurrences of the same measurement k for every point, which can be used as a source of information for a least-squares solution for k (or naively, simply average the $n-1$ measurements). Now consider the projective framework. Projective-depth κ is invariant for any two views ψ_i, ψ_j of the sequence. We have therefore $n(n-1)$ occurrences of κ which is clearly a stronger source of information for obtaining an over-determined solution. The conclusion from this example is that a projective framework has practical advantages over the affine, even in cases where an affine framework is theoretically sufficient. There are other practical considerations in favor of the projective framework. In the affine framework, the epipole v' plays a double role — first for computing the collineation A, and then for computing the affine-depth of all points of interest. In the projective framework, the epipoles are used only for computing the collineations A and E but not used for computing κ. This difference has a practical value as one would probably like to have the epipoles play as little a role as possible because of the difficulty in recovering their location accurately in the presence of noise. In industrial applications, for example, one may be able to set up a frame of reference of two planes with four coplanar points on each of the planes. Then the collineations A and E can be computed without the need for the epipoles, and thus the entire algorithm, expressed in Equation 2, can proceed without recovering the epipoles at all.

Acknowledgments

A draft of this report was read by Eric Grimson, Tuan Luong and Nassir Navab; I am grateful to them for their comments which have been incorporated in the report.

Appendix

A Proof of Proposition

Proposition 1 *Given an arbitrary view $\psi_o \in \mathcal{S}_\pi$ generated by a camera with COP at initial position O, then all other views $\psi \in \mathcal{S}_\pi$ can be generated by a rigid motion of the camera frame from its initial position, if in addition to taking pictures of the object we allow any finite sequence of pictures of pictures to be taken as well.*

Lemma 1 *The set of views $\mathcal{S}_\pi$ can be generated by a rigid camera motion, starting from some fixed initial position, followed by some collineation in $\mathcal{P}^2$.*

Proof: We have shown that any view $\psi \in \mathcal{S}_\pi$ can be generated by satisfying Equation 1, reproduced below:

$$p' \cong Ap - k\boldsymbol{v}'.$$

Note that $k = 0$ for all $P \in \pi$. First, we transform the coordinate system to a camera centered by sending π to infinity: Let $M \in GL_4$ be defined as

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

We have:

$$\begin{aligned} p' &\cong Ap - k\boldsymbol{v}' \\ &= [A, -v'] \begin{pmatrix} x \\ y \\ 1 \\ k \end{pmatrix} \\ &\cong [A, -v'] M^{-1} \begin{pmatrix} x_b \\ y_b \\ z_b \\ 1 \end{pmatrix} \\ &= S \begin{pmatrix} x_b \\ y_b \\ z_b \end{pmatrix} + u, \end{aligned}$$

where $x_b = x/(x+y+1+k), y_b = y/(x+y+1+k)$ and $z_b = 1/(x+y+1+k)$. Let R be a rotation matrix in 3D, i.e., $R \in GL_3$, $det(R) = 1$, and let B denote a collineation in $\mathcal{P}^2$, i.e., $B \in GL_3$, and let w be some vector in 3D. Then, we must show that

$$p' \cong BR \begin{pmatrix} x_b \\ y_b \\ z_b \end{pmatrix} + Bw.$$

For every R, B and w, there exists S and u that produce the same image, simply be setting $S = BR$ and $u = Bw$. We must also show that for every S and u there exists R, B and w that produce the same image: Since S is of full rank (becasue A is), then the claim is true by simply setting $B = SR^T$ and $w = B^{-1}u$, for any arbitrary orthogonal matrix R. In conclusion, any view $\psi \in \mathcal{S}_\pi$ can be generated by some rigid motion R, w starting from a fixed intial position, followed by some collineation B of the image plane. $\square$

We need to show next that any collineation in $\mathcal{P}^2$ can be expressed by a finite sequence of views taken by a rigidly moving camera, i.e., calibrated camera. It is worthwhile noting that the equivalence of projective transformations (an algebraic concept) with a finite sequence of projections of the plane onto itself (a geometric concept) is fundamental in projective geometry. For example, it is known that any projective transformation of the plane can be obtained as the resultant of a finite sequence of projections [32, Thm. 10, pp. 74]. The question, however, is whether the equivalence holds when projections are restricted to what is generally allowed in a rigidly moving camera model. In other words, in a sequence of projections of the plane, we are allowed to move the COP anywhere in $\mathcal{P}^3$; the image plane is allowed to rotate around the new location of the COP and scale its distance from it along a distinguishable axis (scaling focal length along the optical axis). What is not allowed, for example, is tilting the image plane with respect to the optical axis (that has the effect of changing the location of the principal point and the image scale factors — all of which should remain constant in a calibrated camera). Without loss of generality, the camera is set such that the optical axis is perpendicular to the image plane, and therefore when the COP is an ideal point the projecting rays are all perpendicular to the plane, i.e., the case of orthographic projection.

The equivalence between a sequence of perspective/orthographic views of a plane and projective transformations of the plane is shown by first reducing the problem to scaled orthographic projection by taking a sequence of two perspective projections, and then using a result of [30, 11] to show the equivalence for the scaled orthographic case. The following two auxilary propositions are used:

Lemma 2 *There is a unique projective transformation of the plane in which a given line* $\boldsymbol{u}$ *is mapped onto an ideal line (has no image in the real plane) and which maps non-collinear points* A, B, C *onto given non-collinear points* A', B', C'.

Proof: This is standard material (cf. [7, pp. 178]). $\square$

Lemma 3 *There is a scaled orthographic projection for any given affine transformation of the plane.*

Proof: follows directly from [30, 11] showing that any given affine transformation of the plane can be obtained by a unique (up to a reflection) 3D similarity transform of the plane followed by an orthographic projection. $\square$

Lemma 4 *There is a finite sequence of perspective and scaled orthographic views of the plane, taken by a calibrated camera, for any given projective transformation of the plane.*

Proof: The proof follows and modifies [7, pp. 179]. We are given a plane α and a projective transformation T. If T is affine, then by Lemma 3 the proposition is true. If T is not affine, then there exists a line $\boldsymbol{u}$ in α that is mapped onto an ideal line under T. Let A, B, C be three non-collinear points which are not on $\boldsymbol{u}$, and let their image under T be A', B', C'. Take a perspective view onto a plane α' such that $\boldsymbol{u}$ has no image in α' (the plane α' is rotated around the new COP such that the plane passing through the COP and $\boldsymbol{u}$ is parallel to α'). Let A_1, B_1, C_1 be the images of A, B, C in α'. Project α' back to α by orthographic projection, and let A_2, B_2, C_2 be the image of A_1, B_1, C_1 in α. Let F be the resultant of these two projections in the stated order. Then F is a projective transformation of α onto itself such that $\boldsymbol{u}$ has no image (in the real plane) and A, B, C go into A_2, B_2, C_2. From Lemma 3 there is a viewpoint and a scaled orthographic projection of α onto α'' such that A_2, B_2, C_2 go into A', B', C', respectively. Let L be the resultant of this projection (L is affine). $\hat{T} = FL$ is a projective transformation of α such that $\boldsymbol{u}$ has no image and A, B, C go into A', B', C'. By Lemma 2, $T = \hat{T}$ (projectively speaking, i.e., up to a scale factor). $\square$

Proof of Proposition: follows directly from Lemma 1 and Lemma 4. $\square$

References

[1] E.H. Adelson. Layered representations for image coding. Technical Report 181, Media Laboratory, Massachusetts Institute of Technology, 1991.

[2] E.H. Adelson and J.Y.A. Wang. Layered representation for motion analysis. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 361–366, New York, NY, June 1993.

[3] E.B. Barrett, M.H. Brill, N.N. Haag, and P.M. Payton. Invariant linear methods in photogrammetry and model-matching. In J.L. Mundy and A. Zisserman, editors, *Applications of invariances in computer vision.* MIT Press, 1992.

[4] 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.

[5] 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.

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

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

[8] J. Harris. *Algebraic Geometry, A First Course.* Springer-Verlag, Graduate Texts in Mathematics., 1992.

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

[10] T.S. Huang and C.H. Lee. Motion and structure from orthographic projections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11:536–540, 1989.

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

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

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

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

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] T. Poggio. 3D object recognition: on a result of Basri and Ullman. Technical Report IRST 9005-03, May 1990.

[22] Q.T.Luong and O.D. Faugeras. Determining the fundamental matrix with planes: Instability and new algorithms. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, pages 489–494, New York, NY, June 1993.

[23] 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.

[24] J.G. Semple and G.T. Kneebone. *Algebraic Projective Geometry.* Clarendon Press, Oxford, 1952.

[25] J.G. Semple and L. Roth. *Introduction to Algebraic Geometry.* Clarendon Press, Oxford, 1949.

[26] 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.

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

[28] 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.

[29] 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.

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

[31] 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.

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

[33] D. Weinshall. Model based invariants for3-D vision. *International Journal of Computer Vision*, 10(1):27–42, 1993.

[34] 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.