Integrating Local Affine into Global Projective Images in the Joint Image Space

P. Anandan and Shai Avidan

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA,
{anandan,avidan}@microsoft.com

Abstract. The fundamental matrix defines a nonlinear 3D variety in the joint image space of multiple projective (or "uncalibrated perspective") images. We show that, in the case of two images, this variety is a 4D cone whose vertex is the joint epipole (namely the 4D point obtained by stacking the two epipoles in the two images). Affine (or "para-perspective") projection approximates this nonlinear variety with a linear subspace, both in two views and in multiple views. We also show that the tangent to the projective joint image at any point on that image is obtained by using local affine projection approximations around the corresponding 3D point. We use these observations to develop a new approach for recovering multiview geometry by integrating multiple local affine joint images into the global projective joint image. Given multiple projective images, the tangents to the projective joint image are computed using local affine approximations for multiple image patches. The affine parameters from different patches are combined to obtain the epipolar geometry of pairs of projective images. We describe two algorithms for this purpose, including one that directly recovers the image epipoles without recovering the fundamental matrix as an intermediate step.

1 Introduction

The fundamental matrix defines a nonlinear 3D variety[1] in the joint image space, which is the 4-dimensional space of concatenated image coordinates of corresponding points in two perspective images. Each $3D$ scene point $\mathbf{X} = (X, Y, Z)$ induces a pair of matching image points (x, y, x', y') in the two images, and this stacked vector of corresponding points is a point in the joint image space. The locus of all such points forms the **joint image** for the two cameras. Since there is a one-to-one correspondence between the 3D world and the joint image, the joint image forms a 3-dimensional variety in the joint image space. Every pair of cameras defines such a variety, which is parametrized by the fundamental matrix which relates the two cameras.

The idea of the joint image space has been previously used by a few researchers – notably, by Triggs [15] who provided an extensive analysis of multi-view matching constraints for projective cameras in the joint image space, and by Shapiro [11] who analyzed the joint image of 2 affine ("para-perspective" projection) camera images. Triggs also observed that for multiple (say $m > 2$) views

[1] namely, a locus of points defined by a set of polynomial constraints, in this case the epipolar constraint, which is quadratic in the joint image space.

D. Vernon (Ed.): ECCV 2000, LNCS 1842, pp. 907–921, 2000.

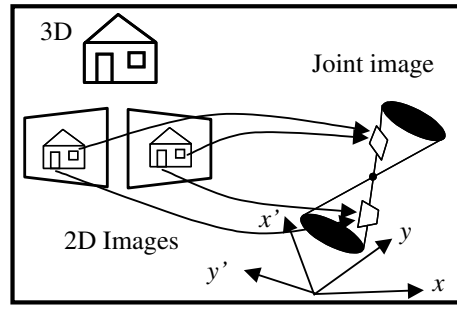


Fig. 1. Illustration of the 4D cone in the joint image space. The two perspective images view a 3D model of the house. The fundamental matrix forms a point-cone in the joint image space, whose axes are defined by x, y, x' and y'. The vertex of the cone is the joint epipole (namely the 4D point formed by stacking both epipoles). Each patch (i.e, the window or the door) in the 2D images is approximated by an affine projection which corresponds to a tangent hyperplane in the joint image space.

the "projective" joint image for any number of cameras is still a 3-dimensional submanifold of the $2m$ dimensional joint image space. This manifold is parametrized (up to a set of gauge invariances) by the m camera matrices.

Affine (or "para-perspective") projection approximates the nonlinear variety with a linear subspace. In multiple views, the affine joint image is a 3-dimensional *linear* space in the $2m$ dimensional joint image space. This insight has led to the factorization approach, which simultaneously uses correspondence data from multiple views to optimally recover camera motion and scene structure [9,11] from para-perspective images.

The non-linearity of the projective joint image makes the multi-view projective structure from motion problem harder to solve. A natural question is whether the affine approximation could be used to benefit the projective case. This question has been previously explored by a few researchers. For example, [4,2,3] use the affine model globally over the entire image to bootstrap the projective recovery. On the other hand, Lawn and Cipolla [7,8] use the affine approximations of over local regions of an image. They combine the affine parallax displacement vectors across two frames from multiple such regions in order to obtain the perspective epipole.

In this paper, we use the joint image space to show the intimate relationship between the two models and exploit it to develop a new algorithm for the recovery of the fundamental matrix and the epipoles. We establish the following results:

1. The joint image of two projective views of a 3D scene is a *point cone*[12] in the 4-dimensional joint image space. See figure 1.
2. The tangent to the projective joint image is the same as the linear space formed by the joint image obtained by using an affine projection approximation around the corresponding 3D scene point. *This is true both in 2 and in multiple views.*

The process of recovering the fundamental matrix is thus equivalent to fitting a 4D cone. For example, the 8-point algorithm [5] for recovering the fundamental matrix can be viewed as fitting a 4D cone to 8 points in the joint image space, since every pair of matching points in the 2D images gives rise to a single data point in the joint image space. Any technique for recovering the fundamental matrix can be regarded this way.

Alternatively, the projective joint image can also be recovered by using the *tangents* to it at different points. In our case, a tangent corresponds to using a local para-perspective (or "affine") projection approximation for an image patch around the corresponding 3D scene point. This leads to a practical two-stage algorithm for recovering the fundamental matrix, which is a global projective constraint, using multiple local affine constraints.

1. The first stage of our algorithm simultaneously uses *multiple ($m > 2$)* images to recover the 3-dimensional affine tangent image in the $2m$ dimensional joint image space. This can be done by using a factorization or "direct" type method.
2. In the second stage the two-view epipolar geometry between a reference image and *each* of the other images is independently recovered. This is done by fitting the 4D cone to the tangents recovered in Stage I from multiple image patches. We take advantage of the fact that all the tangents to the cone intersect at its vertex - the joint epipole - to compute it directly from the tangents. Thus, local affine measurements are used to *directly* estimate the epipoles *without recovering the fundamental matrix as an intermediate step.*

It is worth noting that this approach to directly recover the epipoles is a generalization of the aforementioned work by Lawn & Cipolla [7,8], as well as an algorithm by Rieger & Lawton [10] for computing the focus of expansion for a moving image sequence from parallax motion around depth discontinuities. We postpone more detailed comparison and contrast of our work with these previous papers to Section 5, since we believe that a clear understanding our method will be useful in appreciating these relationships.

2 The Affine and Projective Joint Images

This section establishes the tangency relationship between projective and affine joint images and shows that the projective joint image of two images is a 4D cone. Our derivations proceed in 3 stages: (i) We start by showing that the affine projection is a *linear* approximation to the projective case. We use the affine projection equations to derive the affine motion equations in 2 frames and the associated affine fundamental matrix. These results are already known (e.g, see [9,1,4]) but they serve to lay the ground for the remaining derivations in the paper. (ii) Next we show that for two (uncalibrated) perspective views the joint image is a 4D cone. (iii) Finally we show that the hyperplane described by the affine fundamental matrix is tangent to this 4D cone.

2.1 The Projective Joint Image

We use the following notational conventions. $\mathbf{x} = (x, y)^T$ denotes a 2D image point, while $\mathbf{p} = (x, y, 1)^T$ denotes the same point in homogeneous coordinates. Likewise $\mathbf{X} = (X, Y, Z)^T$ denotes a 3D scene point, and $\mathbf{P} = (X, Y, Z, 1)^T$ denotes its homogeneous counterpart. The general uncalibrated projective camera is represented by the projection equation:

$$\mathbf{p} \cong \mathbf{MP} = \mathbf{HX} + \mathbf{t},$$

where $\mathbf{M}$ denotes the 3×4 projection matrix, $\mathbf{H}$ (referred to as the "homography" matrix) is the left 3×3 submatrix of $\mathbf{M}$, and the 3×1 vector $\mathbf{t}$ is its last column, which represents the translation between the camera and the world coordinate systems.

Since our formulation of the joint-image space involves stacking the *in*-homogeneous coordinates $\mathbf{x}$ from multiple views into a single long vector, it is more convenient to describe the projection equations in in-homogeneous coordinates[2]. The projection of a 3D point $\mathbf{X}$ on to the 2D image point $\mathbf{x^i}$ in the i-th image is given by:

$$\mathbf{x^i} = \begin{pmatrix} \frac{\mathbf{H_1^i X}+t_1^i}{\mathbf{H_3^i X}+t_3^i} \\ \frac{\mathbf{H_2^i X}+t_2^i}{\mathbf{H_3^i X}+t_3^i} \end{pmatrix} \tag{1}$$

where $\mathbf{H_1^i}, \mathbf{H_2^i}$ and $\mathbf{H_3^i}$ are the three rows of $\mathbf{H^i}$, the homography matrix of the i-th image. Likewise (t_1^i, t_2^i, t_3^i) denote the three components of $\mathbf{t}^i$ the translation for the i-th image.

Consider the stacked vector of image coordinates from the m images – namely, the $2m$ dimensional joint-image vector $(x^1\ y^1\ x^2\ y^2 \ldots x^m\ y^m)^T$. We see from Equation 1 that each component of this vector is a *non-linear* function of the 3D position vector $\mathbf{X}$ of a scene point. Hence, the locus of all such points forms a 3-dimensional submanifold in the $2m$ dimensional space. This defines the *projective* joint image. (We have chosen to call it "projective" only to indicate that the joint image of multiple perspectively projected views of a 3D scene do not require any knowledge or assumptions regarding calibration.)

2.2 The Affine Joint Image

Around some point $\mathbf{X_0}$ on the object we can rewrite the $x-$component of Equation 1 as

$$x = \frac{\mathbf{H_1 X_0} + t_1 + \mathbf{H_1 \Delta X}}{\mathbf{H_3 X_0} + t_3 + \mathbf{H_3 \Delta X}} \tag{2}$$

where $\mathbf{\Delta X} = \mathbf{X} - \mathbf{X_0}$. (*Note that we have dropped the super-script i for ease of readability.*) Let us denote $Z_0 = \mathbf{H_3 X_0} + t_3$ and $\Delta Z = \mathbf{H_3 \Delta X}$. We divide the numerator and denominator by Z_0 to obtain

$$x = \frac{\frac{\mathbf{H_1 X_0}+t_1}{Z_0} + \mathbf{H_1}\frac{\mathbf{\Delta X}}{Z_0}}{1 + \frac{\Delta Z}{Z_0}} \tag{3}$$

[2] In this regard, our formulation is slightly different from the more general projective space treatment of [15]. Our approach turns out to be more convenient for deriving the affine approximations.

Considering a shallow portion of the 3D scene around $\mathbf{X_0}$, i.e, $\epsilon = \Delta Z/Z_0 \ll 1$, we can use the approximation $1/(1+\epsilon) \approx 1-\epsilon$ to obtain

$$x \approx (x_0 + \mathbf{H_1}\frac{\mathbf{\Delta X}}{Z_0})(1 - \frac{\Delta Z}{Z_0}) \tag{4}$$

Expanding this equation, replacing ΔZ with $H_3\mathbf{\Delta X}$, and neglecting second-order terms $\mathbf{\Delta X}^2$ gives us the first-order Taylor expansion for the perspective projection equation for x:

$$x \approx x_0 + \mathbf{H_1}\frac{\mathbf{\Delta X}}{Z_0} - x_0\frac{\mathbf{H_3}\mathbf{\Delta X}}{Z_0} \tag{5}$$

Performing a similar derivation for y we get the *affine* projection equations that relate a 3D point to its 2D projection:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} + \frac{1}{Z_0}\begin{pmatrix} \mathbf{H_1} - x_0\mathbf{H_3} \\ \mathbf{H_2} - y_0\mathbf{H_3} \end{pmatrix}\mathbf{\Delta X} \tag{6}$$

Since these equations are linear in $\mathbf{\Delta X}$, they define a 3-dimensional *linear* variety in the $2m$ dimensional joint-image space. Stacking all such equations for the m views gives us the parametric representation of the linear affine joint-image. Also, in each of the m images these equations represent the first-order Taylor expansion of the perspective projection equations around the image point $\mathbf{x}_0$. This means that the $2m$ dimensional *affine* joint-image is the tangent to the *projective* joint image at the point represented by the $2m$ dimensional vector $(x_0^1\ y_0^1\ x_0^2\ y_0^2 \ldots x_0^m\ y_0^m)^T$.

2.3 Two View Affine Motion Equations

Next, we want to derive the affine motion equations that relate matching points across two images. Such equations have also been previously used by a number of researchers (e.g., see [11,9]). We present them here in terms of the homography matrix H and a matching pair of image points (x_0, y_0) and (x_0', y_0') in two views, around which the local affine approximation to perspective projection is taken.

Let us align the world coordinate system with that of the first camera so its homography becomes the 3×3 identity matrix and translation is a zero 3-vector. Let $\mathbf{H}, \mathbf{t}$ be the homography matrix and translation, respectively of the second camera. We will not present the entire derivation here, but simply note that the key step is eliminating ΔX and ΔY from equation 6. Let $\gamma = \frac{\Delta Z}{Z_0} = \frac{Z-Z_0}{Z_0}$ be the "relative" depth of a (x, y) in the reference image. Then, (after some algebraic manipulation), we can show that the corresponding point (x', y') in the second image is given by

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \mathbf{A}\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} + \gamma\mathbf{t}_a \tag{7}$$

where $\mathbf{A}$ is a 2×3 affine matrix

$$\mathbf{A} = (\mathbf{G} \ \mid \ \mathbf{x}_0' - \mathbf{G}\mathbf{x}_0) \tag{8}$$

$\mathbf{G}$ is a 2×2 matrix

$$\mathbf{G} = \frac{Z_0}{Z_0'} \begin{pmatrix} H_{11} - x_0' H_{31} & H_{12} - x_0' H_{32} \\ H_{21} - y_0' H_{31} & H_{22} - y_0' H_{32} \end{pmatrix} \tag{9}$$

and

$$\mathbf{t}_a = \frac{Z_0}{Z_0'} \begin{pmatrix} (\mathbf{H_1} - x_0' \mathbf{H_3}) \mathbf{p_0} \\ (\mathbf{H_2} - y_0' \mathbf{H_3}) \mathbf{p_0} \end{pmatrix}. \tag{10}$$

Equation 7 can be interpreted as follows. $\mathbf{A}$ defines a 2D affine transformation of the image which captures the motion of points for whom $\Delta Z = 0$, i.e., they are on a "frontal" plane at depth Z_0 in the first image. Off-plane points undergo an additional parallax motion $\gamma \mathbf{t}_a$. The parallax magnitude γ is determined by the relative depth $\gamma = \frac{\Delta Z}{Z_0}$. The direction of parallax is $\mathbf{t}_a$, and is the same for all points, i.e, the "affine epipole" is at infinity[3].

We observe that Equation 7 (affine motion) is valid for any view relative to a reference view. The 2D affine transformation matrix $\mathbf{A}$ and the vector $\mathbf{t}_a$ vary across the images while the local shape γ varies across all points, but is fixed for any given point across all the images. The fact that γ is constant over multiple view enables us to simultaneously recover all the 2D affine transformation matrices A, the affine parallax vectors $\mathbf{t}_a$ (see Section 3).

The two-view affine epipolar constraint: The affine motion equations defined in Equation 7 also imply an affine epipolar constraint [11]:

$$\mathbf{p'}^T \mathbf{F_a} \mathbf{p} = 0,$$

where:

$$\mathbf{F_a} = \begin{pmatrix} 0 & 0 & t_2 \\ 0 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ 0 & 0 & 1 \end{pmatrix}, \tag{11}$$

is the "affine" fundamental matrix. Note that this matrix is of the form

$$\begin{pmatrix} 0 & 0 & f_3 \\ 0 & 0 & f_4 \\ f_1 & f_2 & f_5 \end{pmatrix},$$

Let us denote $\mathbf{f} = (f_1, \ldots, f_5)^T = (-t_2 A_{11} + t_1 A_{21}, -t_2 A_{12} + t_1 A_{22}, t_2, -t_1, 1)^T$. Also let $\mathbf{q} = (x, y, x', y')$ be the stacked vector of the two pairs of image coordinates in the 4-dimensional joint image space. The affine epipolar constraint says that the affine joint image consists of all points $\mathbf{q}$ which lie on a hyperplane given by the equation

$$\mathbf{f}^T \begin{pmatrix} \mathbf{q} \\ 1 \end{pmatrix} = 0. \tag{12}$$

[3] It can also be shown that this vector lies along the direction of the line connecting the point p_0' to the epipole defined by t in the second *perspective* image.

This implicit form (as opposed to the parametric form described by Equation 6) will be useful later for us (see Section 5) to relate to the perspective Fundamental matrix[4]. (Of course, this approximation is only reasonable for points $\mathbf{p}$ and $\mathbf{p}'$ which lie *near* the matching pair of points $\mathbf{p}_0$ and $\mathbf{p'}_0$.)

2.4 The Fundamental Matrix as a Point-Cone

Given two views, the well-known epipolar constraint equation can be written in our notation as:

$$\mathbf{p'}^T \mathbf{F} \mathbf{p} = 0 \tag{13}$$

where $\mathbf{p}$ and $\mathbf{p}'$ denote the 2D image location of a scene point in two views specified in homogeneous coordinates, and $\mathbf{F}$ is the 3×3 *fundamental matrix*. In the joint image space, this equation can be written as:

$$\frac{1}{2}(\mathbf{q}^T \quad 1)\mathbf{C}\begin{pmatrix}\mathbf{q}\\ 1\end{pmatrix} = 0 \tag{14}$$

where as before, $\mathbf{q} = (x, y, x', y')^T$ is the stacked vector of the image coordinates of a matching point, and $\mathbf{C}$ is the 5×5 matrix defined below.

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & F_{11} & F_{21} & F_{31} \\ 0 & 0 & F_{12} & F_{22} & F_{32} \\ F_{11} & F_{12} & 0 & 0 & F_{13} \\ F_{21} & F_{22} & 0 & 0 & F_{23} \\ F_{31} & F_{32} & F_{13} & F_{23} & 2F_{33} \end{bmatrix}. \tag{15}$$

This equation describes a quadric in the 4 dimensional joint image space of (x, y, x', y'). We now analyze the shape of this quadric.

Theorem: *The joint-image corresponding to two uncalibrated perspective views of a 3D scene is a* **point cone.**

Proof: First, we show that the rank of the 5×5 matrix $\mathbf{C}$ is 4. To do this, we rewrite $\mathbf{C}$ as the sum of two matrices $\mathbf{C_1}$ and $\mathbf{C_2}$, where

$$\mathbf{C_1} = \begin{bmatrix} 0 & 0 & \mathbf{F}_1^T \\ 0 & 0 & \mathbf{F}_2^T \\ 0 & 0 & \mathbf{0}^T \\ 0 & 0 & \mathbf{0}^T \\ 0 & 0 & \mathbf{F}_3^T \end{bmatrix}, \tag{16}$$

and

$$\mathbf{C_2} = \mathbf{C_1}^T$$

where $\mathbf{F}_i^T$ denotes the $i-th$ row of $\mathbf{F}^T$ (equivalently, $\mathbf{F_i}$ is the $i-th$ column of $\mathbf{F}$) and $\mathbf{0}$ is a 3D zero vector. Now $\mathbf{C_1}$ and $\mathbf{C_2}$ are both Rank 2, since the 3×3 submatrices contained in them are in fact the fundamental matrix $\mathbf{F}$ and $\mathbf{F}^T$

[4] The parameters $(f_1, \ldots, f_5)$ can be derived in terms of $\mathbf{H}$, $\mathbf{p}_0$ and $\mathbf{p}'_0$. However, these expressions are somewhat tedious and do not shed any significant insight into the problem. Hence, we have not elaborated them here.

which are of Rank 2. Also, it can be easily verified that $\mathbf{C_1}$ and $\mathbf{C_2}$ are linearly independent of each other. Hence $\mathbf{C} = \mathbf{C_1} + \mathbf{C_2}$ is of Rank 4.

According to [12] a quadric defined by a 5×5 matrix $\mathbf{C}$ which is of Rank 4 represents a 4D cone called a **point cone**, which is simply a term to describe a 4-dimensional cone. Since the projective joint image is defined by our $\mathbf{C}$ which is rank 4, the joint image has the shape of a point cone. QED

Let $\mathbf{e} = (e_1 \ e_2 \ 1)^T$ and $\mathbf{e'} = (e'_1 \ e'_2 \ 1)$ denote the two epipoles in the two images in homogeneous coordinates. Let us define the point $\mathbf{q_e} = (e_1 \ e_2 \ e'_1 \ e'_2 \ 1)^T$ in the joint image space as the "joint epipole".

Corollary: **:** *The vertex of the projective joint image point cone is the joint epipole.*

Proof: From the definition of $\mathbf{C}$ is easy to verify that $\mathbf{Cq_e} = \mathbf{Fe} + \mathbf{F}^T\mathbf{e'}$. But since the epipoles are the null-vectors of the $\mathbf{F}$ matrix, we know that $\mathbf{Fe} = \mathbf{F}^T\mathbf{e'} = 0$. Hence, $\mathbf{Cq_e} = 0$. This means that the joint epipole is the null-vector for $\mathbf{C}$, and once again according to [12], this means that the point $\mathbf{q_e}$ which denotes the joint epipole is the vertex of the point cone. QED

2.5 The Tangent Space of the Projective Joint Image

As per Equation 14, in the case of 2 views, the projective joint image variety is a level set of the function

$$f(\mathbf{q}) = \frac{1}{2}(\mathbf{q}^T \ 1)\mathbf{C}\begin{pmatrix}\mathbf{q}\\ 1\end{pmatrix}$$

corresponding to level zero. Hence,

$$\nabla f = \mathbf{C}\begin{pmatrix}\mathbf{q}\\ 1\end{pmatrix} \tag{17}$$

defines its orientation (or the "normal vector") at any point $\mathbf{q}$. Considering a specific point $\mathbf{q}_0$, let $\mathbf{p}_0 = (x, y, 1)^T$ and $\mathbf{p}'_0 = (x', y', 1)^T$ be the corresponding image point in the two views (in homogeneous coordinates). By looking into the components of $\mathbf{C}$, it can be shown that:

$$\Pi_{\mathbf{q}_0} = \nabla f_{\mathbf{q_0}} = \mathbf{C}\begin{pmatrix}\mathbf{q_0}\\ 1\end{pmatrix} = \begin{pmatrix}\mathbf{F}_1^T\mathbf{p}'_0\\ \mathbf{F}_2^T\mathbf{p'}_0\\ \mathbf{F}_1\mathbf{p}_0\\ \mathbf{F}_2\mathbf{p}_0\\ \mathbf{F}_3\mathbf{p}_0 + \mathbf{F}_3^T\mathbf{p}'_0\end{pmatrix} \tag{18}$$

We have denoted the normal by $\Pi_{\mathbf{q}_0}$. The equation of the tangent hyperplane at $\mathbf{q_0}$ is:

$$\Pi_{\mathbf{q}_0}^T\begin{pmatrix}\mathbf{q}\\ 1\end{pmatrix} = 0 \tag{19}$$

In other words, all points $\mathbf{q}$ in the joint-image space that satisfy the above equation lie on the tangent hyperplane. Note that the joint epipole $\mathbf{q_e}$ lies on

the tangent plane[5] since

$$\mathbf{\Pi}_{\mathbf{q_0}}^{\mathbf{T}} \begin{pmatrix} \mathbf{q_e} \\ 1 \end{pmatrix} = (\mathbf{q_0^T}\ \mathbf{1})\mathbf{C}^{\mathbf{T}} \begin{pmatrix} \mathbf{q_e} \\ 1 \end{pmatrix} = (\mathbf{q_0^T}\ \mathbf{1})\mathbf{C} \begin{pmatrix} \mathbf{q_e} \\ 1 \end{pmatrix} = \mathbf{0}, \tag{20}$$

since $\mathbf{C}$ is symmetric. We already showed in Section 3 that the tangent hyperplane to the projective joint image is given by the local affine joint image. Hence, the components of $\mathit{\Pi}_{\mathbf{q_0}}$ given above must be the same as $(f_1, \ldots, f_5)$ defined in Section 3, which are estimated from the local affine patches. This fact will be useful in our algorithm described in Section 3.3.

3 Algorithm Outline

We use the fact that the local affine projection approximation gives the tangent to the 4D cone for recovering the epipolar geometry between a reference view and all other views. Our overall algorithm consists of two stages:

1. Estimate the affine projection approximation for multiple local patches in the images. For each patch, use *all* the images to estimate the affine projection parameters. This is equivalent to computing the linear 3D subspace of the multiview joint image in the $2m$-dimensional joint image space of the m input images. This is described in Section 3.1.
2. Determine the epipolar geometry between each view and the reference view by integrating the tangent information computed for different local affine patches in Step 1 above. This is described in Sections 3.2 and 3.3.

We actually present two different methods for Step 2. The first method (see 3.2) samples the joint image around the location of different affine tangents in the 4-dimensional joint image space to obtain a dense set of two-frame correspondences, and then applies the standard 8-point algorithm to recover the fundamental matrix between each view and the reference view.

The second described in Section 3.3 is more novel and interesting and uses the fact that all tangent planes to the cone must pass through its vertex (i.e., the joint epipole) to directly recover the epipole between each views and the reference view without computing the fundamental matrix as an intermediate step.

3.1 Local Affine Estimation

Any algorithm for estimating affine projection can be used. For example, each affine patch could be analyzed using the factorization method. However, within a small patch it is usually difficult to find a significant number of features. Hence, we use the "direct multi-frame" estimation by [6] that attempts to use all available brightness variations with the patch. The algorithm takes as input three or more images and computes the shape and motion parameters to minimize a brightness error. Specifically, let $\mathbf{a}_j, \mathbf{t}_j$ be the affine motion parameters from

[5] This is not surprising, since every tangent plane to a cone passes through its vertex, which in this case is the joint epipole.

the reference image to image j. (Each $\mathbf{a}_j$ is a 6-vector corresponding to the 6 elements of the 2D affine matrix in equation 8 and $\mathbf{t}_j$ corresponds to the vector $\mathbf{t_a}$ in equation 10). Also, let γ_i be the "relative depth" of pixel i in the reference image. We minimize the following error function:

$$E(\{\mathbf{a}_j\},\{\mathbf{t}_j\},\{\gamma_i\}) = \sum_j \sum_i (\nabla \mathbf{I}_i^T \mathbf{u}_i^j + I_{t_i}^j)^2,$$

where $\nabla \mathbf{I}_i$ is the gradient of pixel i at the reference image and $I_{t_i}^j$ is the temporal intensity difference of the pixel between frame j and the reference frame 0, and

$$\mathbf{u}_i^j = \mathbf{x}_i^j - \mathbf{x}_i^0 = \mathbf{Y}_i \mathbf{a}_j + \mathbf{t}_j \gamma_i,$$

is the *displacement* of the pixel i between frame 0 (the reference frame) and frame j. The matrix $\mathbf{Y}$ has the form:

$$\mathbf{Y} = \begin{pmatrix} x\ y\ 1\ 0\ 0\ 0 \\ 0\ 0\ 0\ x\ y\ 1 \end{pmatrix} \tag{21}$$

Taking derivatives with respect to $\mathbf{a}_j, \mathbf{t}_j$ and γ_i and setting to zero we get:

$$\begin{aligned} \nabla_{\mathbf{a_j}} \mathbf{E} &= \sum_i \sum_j \mathbf{Y}_i \nabla \mathbf{I}_i (\nabla \mathbf{I}_i^T (\mathbf{Y}_i \mathbf{a}_j + \mathbf{t}_j \gamma_i) + I_{t_i}^j) \\ \nabla_{\mathbf{t_j}} \mathbf{E} &= \sum_i \sum_j \gamma_i \nabla \mathbf{I}_i (\nabla \mathbf{I}_i^T (\mathbf{Y}_i \mathbf{a}_j + \mathbf{t}_j \gamma_i) + I_{t_i}^j) \\ \nabla_{\gamma_\mathbf{i}} \mathbf{E} &= \sum_j \mathbf{t}_j \nabla \mathbf{I}_i (\nabla \mathbf{I}_i^T (\mathbf{Y}_i \mathbf{a}_j + \mathbf{t}_j \gamma_i) + I_{t_i}^j) \end{aligned} \tag{22}$$

Because $\mathbf{t}_j$ and γ_i are coupled in $\nabla \mathbf{E}$ we take a back-and-forth approach to minimizing E. At each step we fix $\mathbf{a}_j, \mathbf{t}_j$ and compute γ_i^* as:

$$\gamma_i^* = \frac{\sum_j \mathbf{t}_j^T \nabla \mathbf{I}_i (\nabla \mathbf{I}_i^T \mathbf{Y}_i \mathbf{a}_j + I_{t_i}^j)}{\sum_j \mathbf{t}_j^T \nabla \mathbf{I}_i \nabla \mathbf{I_i}^T t_j} \tag{23}$$

and minimize the new error function

$$E^*(\{\mathbf{a}_j\},\{\mathbf{t}_j\}) = \sum_j \sum_i (\nabla \mathbf{I}_i^T (\mathbf{Y}_i \mathbf{a}_j + \mathbf{t}_j \gamma_i^*) + I_{t_i}^j)^2 \tag{24}$$

The new parameters $\mathbf{a}_j, \mathbf{t}_j$ are used to recalculate γ_i^* and so on. This entire process is applied within the usual coarse-to-fine estimation framework using Gaussian pyramids of the images.

3.2 From Local Affine Approximation to the Global Fundamental Matrix

In this method, we use the *affine* motion parameters of all the patches between the reference and a given image to recover the projective fundamental matrix

between these two images. For a particular image and particular patch we consider the affine motion parameters $\mathbf{a}, \mathbf{t}$, as recovered in the previous subsection, to calculate the tangent plane. For increased numerical stability, we uniformly sample points on a 3D "slab" tangent to the cone. This is done using the equation:

$$\mathbf{p}' = \mathbf{ap} + \gamma\mathbf{t} \tag{25}$$

where $\mathbf{p} = (x, y, 1)^T$, x, y are uniformly sampled within the patch and γ takes the values $-1, 0, 1$. This sampling procedure has the effect of "hallucinating" [14] matching points between the two images, based on the affine parameters of the patch. These hallucinated matching points correspond to virtual 3D points within a shallow 3D slab of the 3D scene. This is repeated for every patch. We use Hartley's normalization for improved numerical stability and compute the fundamental matrix, from the hallucinated points, using the 8-point algorithm. Note that since the matching points are hallucinated, there is no need to use robust estimation techniques such as LMeDS or RANSAC.

3.3 Estimating the Joint Epipole Directly from the Local Affine Approximations

While the previous method was useful to recover the global projective F matrix, it does not take full advantage of the tangency relationship of the local affine joint images to the global projective and joint image. Here we present a second method that uses the fact that the tangent hyperplane defined by the local affine patches to the 4D cone must pass through the vertex of the cone (see Section 2.5). Thus, we can use the affine motion parameters (which specify the tangent plane) to recover the epipoles directly, without recovering the fundamental matrix as an intermediate step. Let $\mathbf{a}, \mathbf{t}$ be the affine motion parameters of a given patch and let $\mathbf{f}$ be the hyperplane normal vector given in (see equation 11). Then in the joint image space, the tangent $\pi_{\mathbf{q_0}}$ to the patch is given by:

$$\mathbf{\Pi_{q_0}} = \mathbf{f} = (t_1a_{21} - t_2a_{11}, t_1a_{22} - t_2a_{12}, t_2, -t_1, t_1a_{23} - t_2a_{13})^T \tag{26}$$

where $a_{13} = ((a_{11} - 1)x_0 + a_{12}y_0 + a_{13}))$ and $a_{23} = (a_{21}x_0 + (a_{22} - 1)y_0 + a_{23})$ are modified to account for the relative position of the patch in the global coordinate system of the image, and (x_0, y_0) is the upper-left corner of the patch.

Let $\mathbf{q_e} = (e_1\ e_2\ e_1'\ e_2'\ 1)^T$ be the joint epipole composed from the two epipole then we have that (refer to equation 18):

$$\mathit{\Pi}_{q_i}^T \begin{pmatrix} \mathbf{q_e} \\ 1 \end{pmatrix} = 0 \tag{27}$$

This equation is true for every patch in the image, thus given several patches we can recover the joint epipole $\mathbf{q_e}$ by finding the null space of:

$$\begin{bmatrix} \mathit{\Pi}_{q_1}^T \\ \mathit{\Pi}_{q_2}^T \\ . \\ \mathit{\Pi}_{q_n}^T \end{bmatrix} \tag{28}$$

Once the epipole $\mathbf{e}'$ is known we can recover the homography using the following equation:

$$\mathbf{p}'^T \mathbf{F} \mathbf{p} = \mathbf{p}'^T \begin{pmatrix} 0 & 1 & e'_2 \\ 1 & 0 & -e'_1 \\ -e'_2 & e'_1 & 0 \end{pmatrix} \mathbf{H}\mathbf{p} = 0 \tag{29}$$

In this equation the epipole $(e_1, e_2, 1)$ is known from the step just described above. Given $\mathbf{p} = (x, y, 1)^T$ and $\mathbf{p}' = (x', y', 1)^T$ are hallucinated matching points that are sampled as in the method described in Section 3.2, the only unknown is the homography $\mathbf{H}$. This equation defines a homogeneous linear constraint,

$$\mathbf{s}^T \mathbf{h} = 0$$

in the 9 unknown parameters of the homography $\mathbf{H}$. Here $\mathbf{s}$ is a 9×1 vector which depends on the point, and $\mathbf{h}$ is the homography parameters stacked as a 9-dimensional vector. Every hallucinated matching point provides one such constraint. Given a set of N hallucinated matching points indexed by i, the vector $\mathbf{h}$ must lie on the null space of the matrix formed by stacking the vectors $\mathbf{s}_i^T$ into a $N \times 9$ matrix.

Note that the null space of this equation is a 4-dimensional space as described in [13], since any planar homography consistent with the given camera geometry and an arbitrary physical plane in 3D will satisfy these equations. Hence, we are free to choose any linear combination of the null vectors to form a legitimate homography matrix $\mathbf{H}$.

4 Experiments

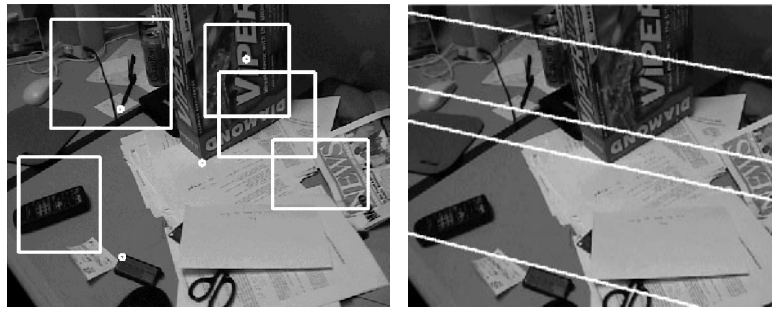
We performed a number of experiments on real images. In all the cases we used the progressive scan Canon ELURA DV cam-corder that produces RGB images of size 720×480 pixels. In all the experiments we used the "direct multi-frame" estimation technique to recover the affine model parameters of a local patch across multiple images. We collected several such patches and used them as tangents to compute either the fundamental matrix or the epipole directly.

4.1 Recovering the Fundamental Matrix

This experiment consists of 6 images. We manually selected 5 patches in the first image and recovered the affine motion parameters for each patch for all the images in the sequence. We then hallucinated matching points between the first and last images and used them to compute the fundamental matrix. To measure the quality of our result, we have numerically measured the distance of the hallucinated points to the epipolar lines generated by the fundamental matrix and found it to be about 1 pixels. The results can be seen in figure 2.

4.2 Recovering the Joint Epipole

We conducted two experiments, one consisting of 6 images, the other consisting of 8 images. We manually selected a number of patches in the first image and



(a) (c)

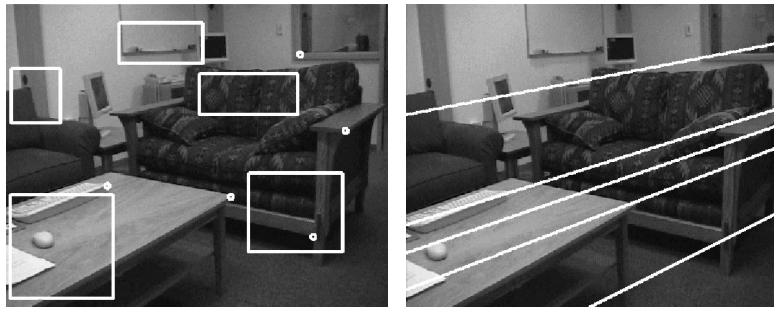
Fig. 2. First (a) and last (b) images in a 6-frame sequence. The rectangles represent the manually selected patches. A "direct multi-frame" algorithm was used to estimate the affine motion parameters of every patch throughout the sequence. "Hallucinated" matching points were used to compute the fundamental matrix between the reference and last image. We show the quality of the fundamental matrix on a number of hand-selected points. Note that the epipolar lines pass near the matching points with an error of about 1 pixel.

recovered the affine motion parameters for each patch for all the images in the sequence . The patches were used to recover the joint-epipole. From the joint epipole we obtained the epipole and used it to recover the 4-dimensional space of all possible homographies. We randomly selected a homography from this space, and together with the epipole, computed the fundamental matrix between the first and last images. The fundamental matrix was only used to generate epipolar lines to visualize the result. The results can be seen in figures 3 and 4.

5 Discussion and Summary

We have shown that the fundamental matrix can be viewed as a point-cone in the 4D joint image space. The cone can be recovered from its tangents, that are formed by taking the affine (or "para-perspective") approximation at multiple patches in the 2D images. In fact, the tangency relationship between affine and projective joint images extend to multiple images. These observations lead to a novel algorithm that combine the result of multi-view affine recovery of multiple local image patches to recover the global perspective epipolar geometry. This leads to a novel method for recovering the epipoles *directly* from the affine patches, without recovering the fundamental matrix as an intermediate step.

As mentioned earlier, our work generalizes those of Rieger & Lawton [10] and Lawn & Cipolla [7,8]. Rieger & Lawton used the observation that the difference in image flow of points (namely, parallax) on two sides of a depth discontinuity is only affected by camera translation, and hence points to the focus-of-expansion (FOE) . They use multiple such parallax vectors tom recover the FOE. Their approach has been shown to be consistent with human psychological evidence.



(b) (c)

Fig. 3. First (a) and last (b) images in a 6-frame sequence. The rectangles represent the manually selected patches. A "direct multi-frame" algorithm was used to estimate the affine motion parameters of every patch throughout the sequence. The affine parameters are used to recover the epipoles *directly*. The fundamental matrix is recovered from the epipoles and an arbitrary legitimate homography only to visualize the epipolar lines. We show the quality of the fundamental matrix on a number of hand-selected points. Note that the epipolar lines pass near the matching points.

As mentioned earlier, Lawn and Cipolla use the affine approximation for two-frame motion within local regions. However, they do not require that the region contain discontinuities. Our algorithm generalizes their approach by using multiple views simultaneously. The use of multiple views allows the use of the (local) rigidity constraint over all the views. We expect that this will increase the robustness of the affine structure from motion recovery. This generalization comes naturally as a result of treating the problem in the joint-image space. In particular, the identification of the tangency relationship between the affine and projective cases and realization that the two-view projective joint image is a cone are the key contributions of the paper that enable this generalization.

References

1. R. Basri Paraperspective Equiv Affine In *International Journal of Computer Vision*, Vol. 19, pp. 169–179, 1996
2. Rikard Berthilsson, Anders Heyden, Gunnar Sparr : Recursive Structure and Motion from Image Sequences using Shape and Depth Spaces In *Computer Vision and Pattern Recognition*, Puerto Rico, 1997.
3. Chen, Q. and Medioni, G. Efficient Iterative Solutions to M-View Projective Reconstruction Problem, In *Computer Vision and Pattern Recognition*, Puerto Rico, 1999.
4. S. Christy and R. Horaud Euclidean shape and motion from multiple perspective views by affine iterations In *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI, 18(10):1098-1104, November 1996.
5. 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.

(b) (c)

Fig. 4. First (a) and last (b) images in a 6-frame sequence. The rectangles represent the manually selected patches. A "direct multi-frame" algorithm was used to estimate the affine motion parameters of every patch throughout the sequence. The affine parameters are used to recover the epipoles *directly*. The fundamental matrix is recovered from the epipoles and an arbitrary legitimate homography only to visualize the epipolar lines. We show the quality of the fundamental matrix on a number of hand-selected points. Note that the epipolar lines pass near the matching points.

6. Hanna, K.J. and Okamoto, N.E. Combining Stereo And Motion Analysis For Direct Estimation Of Scene Structure In *Proceedings of the International Conference on Computer Vision*, pages 357–365, 1993.
7. J.M. Lawn and R. Cipolla. Robust egomotion estimation from affine motion-parallax In *Proceedings of the European Conference on Computer Vision*, pages I:205–210, 1994.
8. J.M. Lawn and R. Cipolla. Reliable extraction of camera motion using constraints on the epipole. In *Proceedings of the European Conference on Computer Vision*, pages II:161-173, 1996.
9. C. J.Poelman, T. Kanade A paraperspective factorization method for shape and motion recovery In *Proceedings of the European Conference on Computer Vision*, pages 97–108, 1994.
10. J.H. Rieger, D.T. Lawton Processing differential image motion In *Journal of the Optical Society of America* Vol. 2, Februay 1985.
11. Shapiro, L.S. Affine Analysis Of Image Sequences, Cambridge University Press, Cambridge
12. J.G. Semple, G.T. Kneebone Algebraic Projective Geometry, Oxford university press
13. A. Shashua and S. Avidan. The rank4 constraint in multiple view geometry. In *Proceedings of the European Conference on Computer Vision*, Cambridge, UK, April 1996.
14. R. Szeliski and P. Torr Geometrically constrained structure from motion: Points on planes. In *European Workshop on 3D Structure from Multiple Images of Large-Scale Environments* (SMILE), pages 171-186, Freiburg, Germany, June 1998.
15. B. Triggs Matching constraints and the joint image In *Proceedings of the International Conference on Computer Vision* (ICCV), 1995.