3D GEOMETRY FROM PLANAR PARALLAX

Harpreet S. Sawhney
Machine Vision Group
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120
Net: sawhney@almaden.ibm.com

Abstract

Deriving 3D structure in a fixed object-centered coordinate system is an increasingly popular trend in shape from multiple views. For linear approximations to perspective projection (weak/para perspective) [12, 16, 21, 22], and for the case of image velocities [11], elegant linear methods have been devised for robust estimation. For reconstruction under arbitrary view transformations, linear projective methods [8, 9, 20] using *point correspondences* have been suggested.

In this paper, we formulate the problem of intrinsic 3D structure estimation through perspective projection using motion parallax, defined with respect to an arbitrary plane in the environment. It is shown that if an image coordinate system is warped using plane projective transformation with respect to a reference view, the residual image motion is dependent only on the epipoles and has a simple relation to the 3D structure. Our computational scheme avoids point/line correspondence and is based on hierarchical estimation and image warping [4] working directly with spatio-temporal image intensities.

1 Introduction

The trend in 3D reconstruction from image motion (or multiple views) has rapidly moved in the past few years from camera-centered depth and motion recovery to scene-centered shape and pose recovery. For linear approximations to perspective projection (weak/para perspective) [12, 16, 21, 22], and for the case of image velocities [11], elegant linear methods have been devised for robust estimation. Similarly, for reconstruction under arbitrary view transformations and uncalibrated cameras, novel linear projective methods [8, 9, 20] using *point correspondences* have been suggested.

This paper presents a linear method for 3D analysis based on the idea of motion parallax with respect to an arbitrary plane. It is shown that if an image coordinate system is warped using plane projective transformation with respect to a reference view, the residual image motion is dependent only on the epipoles and has a simple relation to the 3D structure. This approach presents a unified framework for intrinsic shape estimation for all the three commonly used models of projection – weak, para and (full) perspective. This result has strong parallels with Shashua's [20] work on deriving *projective depth* from two views under any model of projection.

2 Planar Motion Parallax

The essential principle behind planar motion parallax is that if an image coordinate system is warped so that a plane (real/virtual) is fixated between this image and a reference image, that is the plane's image motion is nulled, then the residual image motion can be factorized into a component that depends only on the non-planar shape, and another that depends only on the epipoles (i.e. only on camera displacements and not rotation). This is called *planar motion parallax*. It is a specific instance of the well-known notion of motion parallax. For general motion parallax, it can be shown [15, 17] that if two distinct points in 3D project to the same point in an image (that is are along the same view ray), then the difference in their image displacements due to a change in the viewpoint (that is the projection, in another view, of the vector joining the two) depends only on the 3D translation (perspective) or rotation (weak perspective) between the views and the relative depth of the 3D points. However, using the general motion parallax may not be practical because finding coincident points in a view is hard; for an opaque world occluding boundaries represent such points but these may be hard to detect and computing their image motion may be hard too.

The use of planar parallax instead is practical. Many cultural and other scenes naturally contain a planar surface which can serve as a coordinate system to define the structure of the rest of the scene.

Figure 1 is a geometric depiction of planar parallax. Given $\mathbf{p}$ and $\mathbf{p}'$, the projections of a 3D point in

1063-6919/94 $3.00 © 1994 IEEE

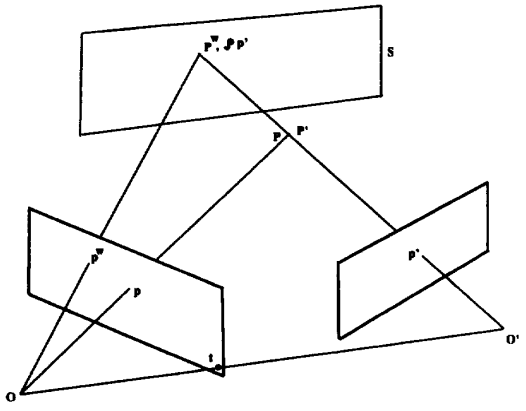

Figure 1: **Two-view Planar Parallax.**

two views, and given a reference plane S, if the planar motion transformation can be computed, then a virtual projection, $\mathbf{p^W}$, corresponding to the point of intersection of the ray $\mathbf{p'}$ and S can be computed. Alternatively the primed image coordinates ($\mathbf{p'}$) can be warped to create an image of points $\mathbf{p^W}$. Then the difference between $\mathbf{p^W}$ and $\mathbf{p}$ in the reference view is the planar parallax motion. It is clear from the figure that these parallax vectors are all oriented towards the epipole $\mathbf{t}$ (the point of intersection of the line connecting the two camera centers, $\mathbf{OO'}$, with the reference image plane).

3 Planar Parallax under Perspective

Perspective projection is a model of projection that accounts for the pin-hole projection of a scene onto an image plane. The linear internal calibration parameters can be modeled as an affine transformation of the image coordinates. This transformation can be combined with an arbitrary rigid transformation due to camera/object motion by modeling the 3D transformation as an affine transformation. Thus, in the following formulation, the 3D transformation between any two time instants is modeled as a 3D affine transformation; a 3D rigid transformation is a special case.

A reference view and any other arbitrary view are chosen to present the motion parallax analysis. The 3D coordinate transformation between the primed coordinates, $\mathbf{P'}$, in view 2 and the reference coordinates, $\mathbf{P}$, in view 1 is written as an arbitrary 3D affine transformation:

$$\mathbf{P'} = \mathbf{A'P} + \mathbf{T'}. \tag{1}$$

Let $\mathbf{N^T P} = d$ represent a plane in the reference coordinate system. Substituting this in the above equation, one can write the *plane projective transformation* as [10]:

$$\mathbf{P'} \approx [\mathbf{A'} + \mathbf{T'N}^T/d]\mathbf{P}, \tag{2}$$

where $\approx$ denotes equality up to an unknown arbitrary scale. Note that this represents the general 8-parameter projective relationship for plane-to-plane projection.

We are interested in developing a direct method to compensate for the image transformation corresponding to the above planar transformation. That is, the second image is to be registered with respect to the reference image by a warping transformation corresponding to the plane projective transformation. Thus, it is necessary to express the warping transformation for the coordinates of the second image. From equation 2, the warped projective coordinates of the second image are

$$\mathbf{P^W} \approx [\mathbf{A'} + \mathbf{T'N}^T/d]^{-1}\mathbf{P'} \approx [\mathbf{I} - \mathbf{TN}^T/d]^{-1}\mathbf{A'}^{-1}\mathbf{P'}, \tag{3}$$

$\mathbf{T} = -\mathbf{A'}^{-1}\mathbf{T'}$ is the displacement of the second frame's origin in the reference coordinates.

Using the identity $[\mathbf{I}+\mathbf{uv}^T]^{-1} = [\mathbf{I}-\alpha\mathbf{uv}^T]$ (see [6]), where $\alpha = (1/(1+\mathbf{v}^T\mathbf{u}))$, $(\mathbf{v}^T\mathbf{u} \neq -1)$, the above relationship can be written as the following projective transformation

$$\mathbf{P^W} \approx [\mathbf{I} + \beta\mathbf{TN}^T/d]\mathbf{A'}^{-1}\mathbf{P'}, \tag{4}$$

(with $\beta = 1/(1-\mathbf{N}^T\mathbf{T}/d)$, and $\mathbf{N}^T\mathbf{T}/d \neq 1$) because in general $\mathbf{T}$ (i.e. the second camera center) does not lie on the reference plane, which would lead to the degenerate case of the plane projecting as a line in one image.

Equation (4) represents the warping transform applied to the second image coordinates to account for the plane projective transformation. This warping transformation will exactly register points in the second image lying on the plane with their projections in the reference frame. However, the points not lying on the plane will have some residual displacement.

Substituting equation (1) in equation (4), we get

$$\mathbf{P^W} \approx [I+\beta\mathbf{TN}^T/d]\mathbf{A'}^{-1}(\mathbf{A'P}+\mathbf{T'}) \approx (\mathbf{P}+\gamma\mathbf{T}), \tag{5}$$

where $\gamma = -1 + \beta\mathbf{P}^T\mathbf{N}/d - \beta\mathbf{T}^T\mathbf{N}/d = (\mathbf{P}^T\mathbf{N} - d)/(-(\mathbf{T}^T\mathbf{N} - d)) = d_N/(-T_d)$; d_N is the perpendicular distance of $\mathbf{P}$ from the plane, and T_d is the distance of the translation vector $\mathbf{T}$ from the plane.

In order to see that the parallax vectors between the warped points, $\mathbf{P^W}$, and the actual points, $\mathbf{P}$, are directed towards the epipole, it is easily shown from equation (5) that $\mathbf{T} \cdot (\mathbf{P^W} \times \mathbf{P}) = 0$. This is a projective relationship that shows that the projection plane

normals defined by all the parallax vectors lie on a great circle on the unit sphere, and the translation vector is normal to the plane of this circle. Lawn and Cipolla [3] use this structure of the motion parallax field for the *special case* of image velocities (closely spaced viewpoints) to compute the epipole. They approximate the planar flow locally as an affine transformation. However, they do not relate the parallax field to the intrinsic structure of the scene. Also, the derivation here is valid for an arbitrary view transformation of the type in equation (1), that includes the case of small displacements. We now derive the relationship between the polar parallax field and scene structure.

Given a view ray $\mathbf{p}'$ in an arbitrary view, with the knowledge of the plane projective transformation of equation (3), its warped coordinates with respect to the reference view can be computed. For points that do lie on the plane, the warping transformation leads to their real projection in the reference view. For the non-planar points, the planar motion parallax vector (the difference between the virtual planar projection and the actual projection) is given by (figure 1):

$$\begin{aligned}\mathbf{p}-\mathbf{p}^{\mathbf{w}} &= \frac{1}{P_z}\mathbf{P}-\frac{1}{P_z^w}\mathbf{P}^{\mathbf{w}} \\ &= \frac{1}{P_z}\mathbf{P}-\frac{1}{P_z+\gamma T_z}(\mathbf{P}+\gamma\mathbf{T}) \\ &= (1/(1+\frac{P_z}{d_N}/\frac{T_z}{-T_d}))(\mathbf{p}-\mathbf{t}), \qquad (6)\end{aligned}$$

where the lower case bold letters represent the respective image vectors with their z-components unity. Note that an internal homogeneous camera transformation, $(\mathbf{A_C})$, can be applied to each of the image vectors, $\mathbf{p}, \mathbf{p}^{\mathbf{w}}, \mathbf{t}$, in equation (6) without changing its form. Thus, the equation is valid for an arbitrary unknown internal camera transformation.

When T_z is zero, the parallax equation becomes:

$$\mathbf{p}-\mathbf{p}^{\mathbf{w}} = (-d_N/P_z)[T_x\ T_y\ 0]^T \qquad (7)$$

In this case, the parallax motion vectors are all parallel, oriented towards the epipole at infinity.

For an alternative but more tedious derivation of a similar result, without using image warping, see Lee [14]. A geometric derivation of the planar parallax under perspective projection result and a similar algebraic derivation has also been recently done independently by Kumar and Anandan [13].

Recall that in traditional structure from motion algorithms, decomposing the image motion into rotational and translational components is hard because of inherent ambiguities [1, 2]. This problem may be circumvented in the planar parallax approach because the rotations affect only the plane projective transformation of equation (2) and not the parallax motion.

3.1 View-Invariant Representation

Let $\eta = 1/(1 + P_z/d_N\frac{T_z}{d})$ (equation 6) be the (signed) ratio of the parallax magnitude to $|(\mathbf{p}-\mathbf{t})|$, and let $\tau = 1/\eta - 1$. If a point P_0 not lying on the fixated plane is chosen as a reference then for any other point P_i: $\tau_i/\tau 0 = \frac{P_{i_z}}{d_{i_N}}/\frac{P_{0_z}}{d_{0_N}}$. This ratio represents a view-independent "coordinate" of the structure of the environment that does not lie on the reference plane. Given any arbitrary viewpoint, if the new view can be warped using the transformation corresponding to the reference plane, then the relative parallax magnitude is always the τ-ratio above.

3.2 Affine Reconstruction

If the internal camera parameters are known, and the 3D transformation between views is a rigid transformation (that is, the matrix $\mathbf{A}$ is a rotation matrix $\mathbf{R}$), then the reference plane can be reconstructed in a Euclidean frame attached to the reference view, and subsequently the whole scene can be reconstructed. The plane can be reconstructed in two ways: (i) by solving for the translation from the epipolar constraint of equation (6), and then solving for the rotation and the plane parameters from equation (2), or (ii) by solving for the plane and motion parameters directly from equation (2) [5]. The latter case may be unstable because it relies on higher order information (more than affine) in the image displacements; these generally are unreliable for commonly used small field-of-view cameras [1]. After solving for the plane, by choosing the ratio P_{0_z}/d_{0_N} for a reference non-planar point to be unity, all the other points can be reconstructed using their respective ratios P_{i_z}/d_{i_N} and their view rays $\mathbf{p}$. In particular, say for a point, $P_z/d_N = \alpha$, then since $\mathbf{P} = \lambda\mathbf{p}$, the two constraints define an intersection of the view ray with a plane. This intersection defines λ uniquely. If the reference plane is given by $\mathbf{P}^T\mathbf{N} = d$, then

$$\mathbf{P} = ((\alpha d)/\mathbf{p}^T(\alpha\mathbf{N}-\mathbf{z}))\mathbf{p}, \qquad (8)$$

where $\mathbf{z}$ is the unit vector along the optical axis in the reference view.

However, when the internal camera parameters are unknown, and euclidean reconstruction is not required, then the reconstructed $\mathbf{P}$ of equation (8) represents the 3D geometry of the scene up to an arbitrary 3D affine transformation. To see this, assume that three points on the reference plane, and a fourth

reference point not on the plane have been chosen arbitrarily and specified a set of 3D coordinates (say the standard affine basis). The coordinates of these four points are related to their true 3D coordinates (in some coordinate system) through a 12-parameter 3D affine transformation. This is left unspecified in the reconstruction.

Let three points on the reference plane and a non-planar reference point, ($\mathbf{P_0}$), be given some arbitrary 3D coordinates. Assume that these coordinates define the scene points in the coordinate system of the reference view. Thus, these coordinates are related to their true world coordinates through a transformation $\mathbf{P} = \mathbf{AP_W} + \mathbf{T}$. Note that the internal camera transformation, $\mathbf{A_C}$, relating the ideal pin-hole model image coordinates to the measured image coordinates has been absorbed in the 3D affine transformation. The planar points define a plane $\mathbf{P}^T\mathbf{N} = d$. With $\mathbf{P_0}$ and the plane thus defined, the ratio P_{0_z}/d_{0_N} is fixed.

For any other non-planar point (fifth and more), say, the ratio P_z/d_N is α (as computed by the planar image warping and residual motion computation algorithm to be illustrated in the results section). Then, as in equation (8), $\mathbf{P} = ((\alpha d)/\mathbf{p}^T(\alpha\mathbf{N} - \mathbf{z}))\mathbf{p}$, defines the 3D $\mathbf{P}$. However, in this case, the 3D geometry can be specified only up to an unknown affine transformation that brings the arbitrarily selected four reference points into registration with the known corresponding points in the scene. Therefore, all the scenes related through a 3D affine transformation are indistinguishable in this approach. This is similar to the affine and projectively invariant reconstruction methods [8, 7, 20].

3.3 Parallax Magnitude as Affine Depth

In particular, the image plane can be chosen as the reference plane. Therefore, the plane normal $\mathbf{N} = \mathbf{z}$ and $d = 1$. Given these and α as above,

$$\mathbf{P} = ((\alpha d)/\mathbf{p}^T(\alpha\mathbf{N} - \mathbf{z}))\mathbf{p} = (1/(1 - \frac{1}{\alpha}))\mathbf{p}. \quad (9)$$

Recall that the magnitude of the parallax vector from equation 6 is proportional to $(1/(1+\frac{P_z}{d_N}/\frac{T_z}{-T_d}))$. T_z/T_d can be conveniently set to unity to fix the overall scale. Then, the magnitude becomes $(1/(1 - P_z/d_N))$ which is the same as $(1/(1 - 1/\alpha))$ in equation 9 because $\alpha = P_z/d_N$. Therefore, in the coordinates of the reference image, the quantity $(1/(1+\frac{P_z}{d_N}/\frac{T_z}{-T_d}))$ of equation 6 is directly *the affine depth of the corresponding point* because the structure reconstruction is valid up to an arbitrary 3D affine transformation. This is similar to Koenderink's [12] and Shashua's [19] affine structure from motion under weak perspective, and Shashua's [20] projective depth under perspective projection.

3.4 Shashua's Projective Depth

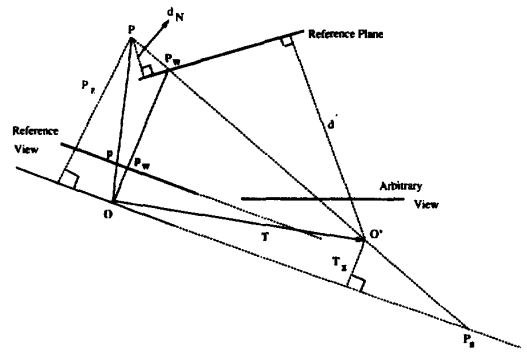

Figure 2: **Relationship with Shashua's two-plane cross-ratio.**

In [20], Shashua presented an elegant method for computing an affine/projective 3D structure invariant from two views under perspective projection. He called this invariant the *projective depth*. His method essentially computed the location of an arbitrary scene point by defining a cross-ratio using the point, the principal point in a reference view, and the intersection of the view ray with two reference planes defined in the reference view coordinate system. In the reference view all these four points project to a single point, but in any other view, the cross-ratio can be computed using image measurements, namely the image correspondence of the scene point in the second view, the epipole in the second view, and the projections of the two planar intersections. The cross-ratio is a projective invariant. Hence, for any view, knowing the epipole and the planar projections, the projection of any point can be reconstructed.

Figure 2 depicts the relationship between our method and that of Shashua. A point $\mathbf{P}$ is viewed in a reference view and an arbitrary view with centers of projection, $\mathbf{O}$ and $\mathbf{O'}$, respectively. Instead of using two reference planes in the scene to define a cross-ratio, our method uses one reference plane and the $z = 0$ plane in the reference view. So the cross-ratio is defined using the line $\mathbf{PP_WO'P_S}$ as shown in the figure. $\mathbf{P_W}$ is the intersection of the view ray $\mathbf{O'P}$ with the reference plane, and $\mathbf{P_S}$ is its intersection with the $z = 0$ plane. A cross-ratio for the point $\mathbf{P}$ can be defined as $(\mathbf{PP_S/O'P_S})/(\mathbf{PP_W/O'P_W})$. From

similar traingles in figure 2, this is exactly the ratio $((P_z/T_z)/(d_N/d))$ given by the motion parallax equation (6). The equation shows how this ratio can be computed using the image measurements based on the planar parallax, similar to the computation using planar projections for two reference planes in Shashua's case.

4 Experimental Results

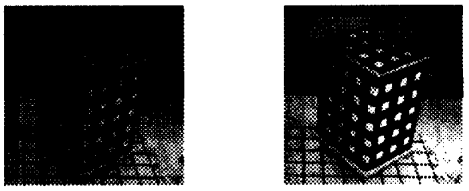

Figure 3: **Frames 1 and 2 of the box scene.**

We demonstrate the application of planar motion parallax on images of a rotating box. Two frames of the box, rotated 4° between frames with the background stationary, are shown in figure 3. The effective FOV is 24° by 23° and the range of depths is about 550 to 700 mm.

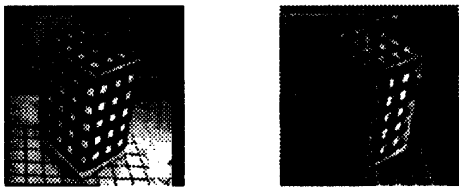

Figure 4: **(i) Frame 2 warped using the affine transformation cooresponding to the left face, and (ii) Difference between frame 2 affine warped and frame 1.**

All the processing on the images is done using direct methods developed by the Sarnoff group [4]. No point or discrete feature correspondence is assumed. First, the left face of the box is specified as the reference plane in the first image (called *BOX1*). The second image is registered with respect to the first using the image flow corresponding to the reference plane. That is, the coordinate system of the *whole* of the second image is transformed according to the planar flow estimate. It was found that a general 6-parameter affine transformation was sufficient for this. The second image warped corresponding to the planar affine transformation (called *BOX2AFFW*) is shown in figure 4 along with the difference between this warped image and the reference image, *BOX1*. Clearly, the motion of the reference plane has been nulled and the residual motion is only due to the parts of the scene not lying on the reference plane. For the *BOX* scene

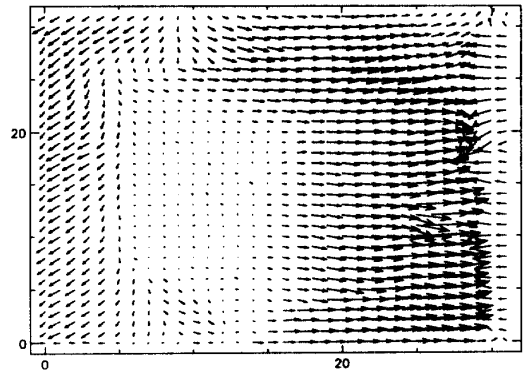

Figure 5: **The non-planar residual flow.**

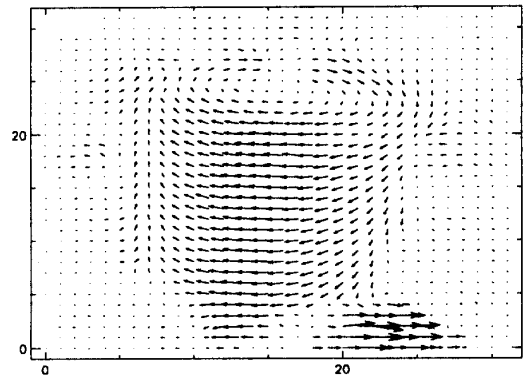

Figure 6: **Raw flow b/w frames 1 and 2.**

WP might be a good enough model as was noted by Daphna Weinshall in [22]. In the difference image (figure 4), it is apparent from the "motion blur" that the residual motion is almost translatory. This is very clear when the reference image and the difference image are shown as a sequence on a CRT display.

Subsequently, a general flow algorithm [4] is applied between the reference image and the affine warped image, *BOX2AFFW*. The flow vectors are shown in figure 5. This process of registration using a general flow algorithm almost completely cancels the residual translation motion. The residual non-planar motion vectors produced by this registration correspond to equation (6). In order to highlight the epipolar flow obtained after planar warping, we show the "raw" flow

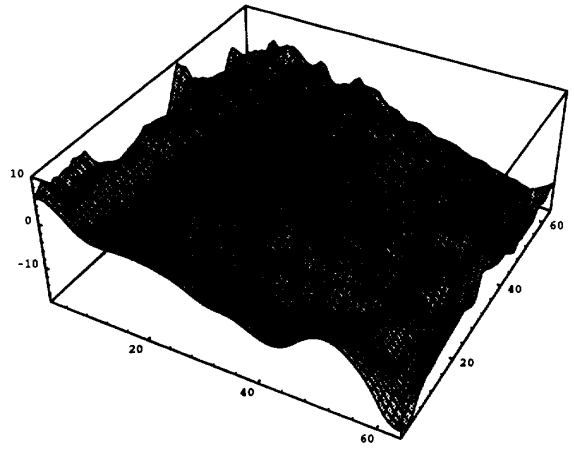

Figure 7: **Grided surface plot of the box.**

between frames 1 and 2 in figure 6. Clearly, in the region of the box, the original flow is rotational but after planar warping, the residual flow is mostly along the horizontal axis because the effective translation is mostly along the x axis.

In this case, because all of the scene that is not on the reference plane is on one side of it, if we plot the magnitude of flow as a function of the image plane xy-coordinate system, then this will represent the intrinsic structure of the box up to an arbitrary 3D affine transformation. That is, the reference plane is the image plane and the parallax magnitude is the non-planar depth with respect to this plane. The intrinsic shape estimate is shown as a surface plot in figure 7. The viewpoint has been chosen to make the computed shape fairly explicit. (The surface plot uses some arbitrary scale and coordinate system specific to the plotting programs.) Note that in the regions corresponding to the background, the flow is arbitrary because the background was stationary and only the box was moving.

The surface plot clearly shows that the qualitative estimates of the planar facets of the box and the overall shape have been recovered fairly well. We have also applied the above processing to an outdoor scene with a moving camera and self-moving objects. The parallax flow separates the various components of motion qualitatively but due to lack of space, those results cannot be presented here (see [18]).

Acknowledgements

I thank Teddy and Anandan (Sarnoff) for many stimulating discussions, Chitra (MSU) for her help in implementations, and Manmatha (UMass) for his help in producing nice pictures.

References

[1] G. Adiv. Inherent ambiguities in recovering 3D information from a noisy flow field. *IEEE PAMI*, 11(5):477–489, 1989.

[2] K. Daniilidis and H. H. Nagel. The coupling of rotation and translation in motion estimation of planar surfaces. In *CVPR*, pages 188–193, 1993.

[3] J. Lawn et al. Epipole estimation using affine motion parallax. Technical Report CUED/F-INFENG/TR 138, Camb. Univ. Engg. Dept., 1993.

[4] J. R. Bergen et al. Hierarchical model-based motion estimation. In *2nd ECCV*, pages 237–252, 1992.

[5] O. D. Faugeras et al. Let us suppose the world is piece-wise planar. In *3rd Intl. Symp. on Rob. Res.*, 1987.

[6] R. Hartley et al. Computing matched epipolar projections. In *CVPR*, pages 549–555, 1993.

[7] R. Mohr et al. Relative 3D reconstruction using multiple uncalibrated images. In *CVPR*, pages 543–548, 1993.

[8] O. D. Faugeras. What can be seen in three dimensiona with an uncalibrated stereo rig ? In *2nd ECCV*, pages 563–578, 1992.

[9] R. I. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *2nd ECCV*, pages 579–587, 1992.

[10] J. C. Hay. Optical motions and space perception: An extension of Gibson's analysis. *Psych. Rev.*, 73:550–565, 1966.

[11] A. D. Jepson and D. J. Heeger. Linear subspace methods for recovering translational direction. Technical Report RBCV-TR-92-40, U. Toronto, 1992.

[12] J. J. Koenderink and A. J. van Doorn. Affine structure from motion. *JOSA A*, 81:377–385, 1991.

[13] Rakesh Kumar and P. Anandan. Personal Communication.

[14] Chia-Hoang Lee. Structure and motion from two perspective views via planar patch. In *ICCV*, pages 158–164, 1988.

[15] H. C. Longuet-Higgins and K. Prazdny. The interpretation of a moving retinal image. In *Proc. Royal Society of London B*, pages 385–397, 1980.

[16] C. J. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. Technical Report CMU-CS-92-208, CMU, 1992.

[17] J. H. Rieger and D. T. Lawton. Processing differential image motion. *JOSA A*, 2(2):354–360, 1985.

[18] H. S. Sawhney. Motion video analysis using planar parallax. In *SPIE Conf. on Image and Video Databases, San Jose, CA*, pages 231–242, 1994.

[19] A. Shashua. Correspondence and affine shape from two orthographic views. Technical Report AI Memo No. 1327, MIT, 1991.

[20] A. Shashua. Projective depth: A geometric invariant for 3D reconstruction from two perspective/orthographic views and for visual recognition. In *ICCV*, pages 583–590, 1993.

[21] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *IJCV*, 9(2):137–154, 1992.

[22] D. Weinshall. Model based invariants for 3D vision. *IJCV*, 10(1):27–42, 1993.