Research Report

3D GEOMETRY FROM PLANAR PARALLAX

Harpreet S. Sawhney

IBM Research Division
IBM Almaden Research Center, K54/K53
650 Harry Road
San Jose, CA 95120-6099

IBM **Research Division**
Yorktown Heights, New York • San Jose, California • Zurich, Switzerland

3D GEOMETRY FROM PLANAR PARALLAX

Harpreet S. Sawhney
e-mail: sawhney@almaden.ibm.com

IBM Research Division
IBM Almaden Research Center, K54/K53
650 Harry Road
San Jose, CA 95120-6099

April 20, 1994

ABSTRACT:

Deriving 3D structure in a fixed object-centered coordinate system is an increasingly popular trend in shape from multiple views. For weak perspective projection, this problem has been formulated and solved in many different ways. We show that motion parallax with respect to a planar surface is the basis of these approaches. This paper revisits the problem of intrinsic structure derivation for the weak and paraperspective cases. Furthermore, a new derivation of a linear method for intrinsic 3D shape under *perspective projection* is presented. It is shown that the unifying concept underlying the computation of 3D geometry in an intrinsic coordinate system for all projections models is that of using an arbitrary plane in the scene (object) as a reference plane with respect to which the rest of the scene is reconstructed. The representation is adequate for recognition and new view generation tasks, and can also be used for complete metric reconstruction.

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 orthographic projection, Tomasi and Kanade [22] developed an elegant method for factoring image measurements of N feature points over F frames into F camera poses with respect to a fixed coordinate system, and N 3D point coordinates in this system. Recently, the factorization method has been generalized to *paraperspective* and *weak perspective* projections [18]. These approaches were developed for metric reconstruction over many views. For applications involving object recognition and new view generation, it may generally be not necessary to compute an absolute 3D reconstruction of an object; intrinsic shape estimates up to some arbitrary transformations may suffice.

In this paper, motion parallax with respect to an arbitrary plane in the scene is used to compute relative 3D structure of the rest of the scene. It is shown that the parallax motion, after the motion of the reference plane is compensated for, neatly separates into a component involving 3D geometry and another that depends *only* on translation under perspective projection, and rotation under weak perspective. This approach presents a unified framework for intrinsic shape estimation for all the three commonly used models of projection, weak perspective (**WP**), paraperspective (**PP**) and (full) perspective (**FP**). For weak perspective projection, under arbitrary 3D affine transformations, shape reconstruction up to a shear and scale was presented in [10], [14], [20] and [23]. We first revisit these results for **WP** and **PP**. Our formulation is different in that it makes the planar (2D affine) and the non-planar components of the image motion explicit in terms of the view transformation and the 3D shape. It is shown that the non-planar image motion component is dependent directly on the non-planar depth, and rotations in depth. Recall that under **WP**, rotations in depth are the only motion components that lead to 3D structure information.

We extend the structure-from-planar-parallax idea to perspective projection. It is shown that if the coordinate system of an arbitrary image is warped with respect to a reference image, such that the image motion of a given plane becomes zero, then the residual motion of all the points not on the reference plane leads simply to 3D structure. The non-planar residual disparity is dependent only on the translational component of motion, and the out-of-plane depth component in the reference view. An alternative but tedious derivation of this result was presented in [13]. This result has strong parallels with Shashua's [21] work on deriving *projective depth* from two views under any model of projection. The relationship is shown explicitly in the paper.

Our longer term goal is to use 3D constraints in deriving object properties in video sequences for the purposes of automated and semi-automated annotation and analysis. The idea of planar parallax based structure description for various models of projection presents a framework in which from a small number of views (minimum two), an intrinsic representation can be derived. This is possible even if the linear camera calibration parameters are unknown, a common situation in video annotation for videos captured off-line from a variety of cameras. Also, in obtaining compact descriptions of a variety of scenes, planar structures like roads, walls and buildings can be utilized as natural reference planes.

2. Planar Motion Parallax

The essential principle behind planar motion parallax is that if an image coordinate system is warped so that an environmental plane 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, 19] 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. For the problem of obstacle detection, Carlsson and Eklundh [2] and Enkelmann [4] used the specific constraint on image flow for a *ground plane*. The camera motion was modeled as the motion on the ground plane. In contrast, our method can use any arbitrary plane in the environment, (e.g. walls, ceilings, floor etc.) and is applicable for general rigid motion.

Figure 4 is a geometric depiction of planar parallax. Given $\mathbf{p}$ and $\mathbf{p}'$, the projections of a 3D point in 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. Projection Models

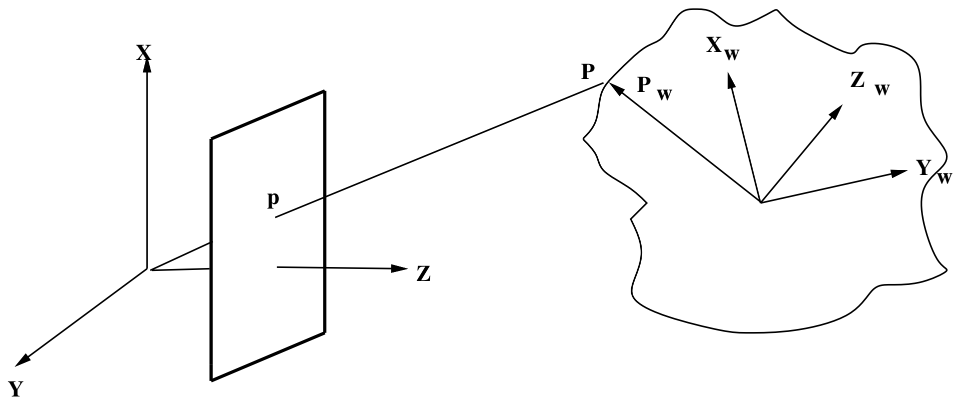

Figure 1: **The Perspective Projection.**

Consider a scene represented in a fixed scene–centered coordinate system, (X_w, Y_w, Z_w). Let (X, Y, Z) represent a camera–centered coordinate system that changes with relative motion between the camera and the scene. A point $\mathbf{P_w}$ in the w–coordinates is represented as $\mathbf{P}$ in the camera coordinates as

$$\mathbf{P} = \mathbf{R}\mathbf{P_w} + \mathbf{T}, \tag{3.1}$$

where $\mathbf{R}$ is the rotation matrix and $\mathbf{T}$ the translation vector that represent the coordinate transformation between the two coordinate systems. If $\mathbf{i}^T$, $\mathbf{j}^T$ and $\mathbf{k}^T$ are the three rows of $\mathbf{R}$, then the above equation can be written componentwise, for instance for the x component,

$$P_x = \mathbf{i}^T \mathbf{P_w} + T_x. \tag{3.2}$$

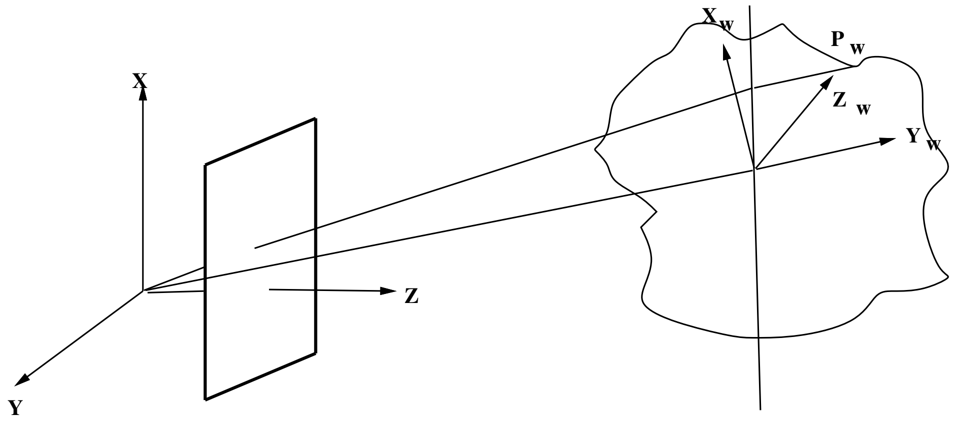

Figure 2: **The Paraperspective Projection.**

In (full) perspective projection (**FP**), depicted in figure 1, a pin–hole model of the camera is used. Assuming that the focal length of the camera is unity, the 2D image projection **p** of $\mathbf{P}$ can be written componentwise as

$$p_x = \frac{P_x}{P_z} = \frac{\mathbf{i}^T\mathbf{P_w} + T_x}{\mathbf{k}^T\mathbf{P_w} + T_z}, \qquad p_y = \frac{P_y}{P_z} = \frac{\mathbf{j}^T\mathbf{P_w} + T_y}{\mathbf{k}^T\mathbf{P_w} + T_z}. \tag{3.3}$$

That is, the x and y 2D components are the ratios, respectively, of the x and y components of $\mathbf{P}$ with its z component.

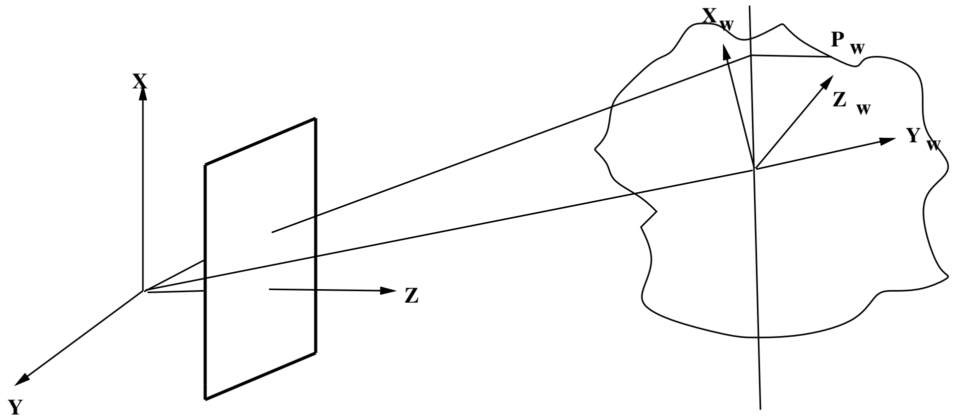

Figure 3: **The Weak Perspective Projection.**

Paraperspective (**PP**) is a particular first order approximation to **FP**. If the extent of the scene/object is small compared to its average distance from the camera, that is $|\mathbf{P_w}|^2/|P_{0_z}^2 \approx 0$ ($\mathbf{P_0}$ is the centroid of the object in the camera coordinates), and the object is significantly off–centered, that is, $|\mathbf{P_w}|^2/|P_{0_z}^2 \ll |\mathbf{P_w}||T_x|/|P_{0_z}^2$, $|\mathbf{P_w}|^2/|P_{0_z}^2 \ll |\mathbf{P_w}||T_y|/|P_{0_z}^2$, then, the 2D projections can be approximated by [18]

$$\mathbf{p} = \frac{1}{T_z}\left(\begin{bmatrix} \mathbf{i}^T - \frac{T_x}{T_z}\mathbf{k}^T \\ \mathbf{j}^T - \frac{T_y}{T_z}\mathbf{k}^T \end{bmatrix} \mathbf{P_w} + \begin{bmatrix} T_x \\ T_y \end{bmatrix} \right). \tag{3.4}$$

Note that **PP** allows for a global scale factor for all points corresponding to the z component of the centroid, and also allows for changes in the view angle for the object. Geometrically, **PP** is shown in figure 2. Each point is first projected, along the view direction of the centroid, on a frontal plane passing through the centroid. All the projections from this frontal plane are projected to the image plane resulting in a common scale factor.

If only the zeroth order term in the Taylor series expansion of the perspective equations around the z component of the object centroid is significant, then the resulting projection is called the weak perspective projection, **WP**. Geometrically (figure 3), each point is first projected along the optical axis (z direction of the camera) on to a frontal plane passing through the centroid. All projections from this plane are then projected on to the image plane. The projection equations are

$$\mathbf{p} = \frac{1}{T_z}\left(\begin{bmatrix} \mathbf{i}^T \\ \mathbf{j}^T \end{bmatrix} \mathbf{P_w} + \begin{bmatrix} T_x \\ T_y \end{bmatrix}\right). \tag{3.5}$$

For all of the models of projection above, the measured image coordinates, $\mathbf{p_c}$, in any arbitrary coordinate system on the image plane can be related to the true image coordinates, $\mathbf{p}$ through an internal camera transformation,

$$\mathbf{p_c} = \mathbf{A_c}\mathbf{p} + \mathbf{T_c}, \tag{3.6}$$

where $\mathbf{A_c}$ is a 2×2 matrix representing the x and y scale factors and skew, and $\mathbf{T_c}$ is the position of the principal point on the image plane. When this is incorporated into the projection models, the transformation from world to the measured image coordinates is a general perspective transformation [17].

4. The Weak/Paraperspective Case

The theory of planar parallax structure for **WP** (affine structure from motion) has been derived in various forms in [10], [14], [20] and [23]. The formulation presented here is different in that it makes the planar (2D affine) and the non–planar components of the image motion explicit in terms of the view transformation and the 3D shape.

For the formulation here, it is assumed that the scene to camera coordinate transformation of equation 3.1 involves a general 3D affine transformation (and not just a rigid transformation). Then the measured image projection, $\mathbf{p}$, of $\mathbf{P}$ under **WP** is

$$\mathbf{p} = \mathbf{A_c}(\mathbf{A_{23}}\mathbf{P_w} + \mathbf{T_{xy}}) + \mathbf{T_c}, \tag{4.1}$$

where $\mathbf{A_{23}}$ is the top left 2×3 sub–matrix of $\mathbf{A}$, and $\mathbf{T_{xy}}$ is the vector $[T_x\ T_y]^T$, and $\mathbf{A_c}$ and $\mathbf{T_c}$ are the internal camera parameters. Note that for each image, the camera parameters can be different. Clearly, the centroid of the image projections of a set of points is the same as the projection of the object centroid. Therefore, if $\mathbf{p}$ now refers to the difference vector between an imaged point and the centroid, the above equation can be simplified to

$$\mathbf{p} = \mathbf{A_c}\mathbf{A_{23}}\mathbf{P_w}. \tag{4.2}$$

We can choose the scene coordinate frame to be aligned with an arbitrary image frame, called the reference frame. In this frame, the projection equation simplifies to

$$\mathbf{p} = \mathbf{A_c}\mathbf{P_{w_{xy}}}. \tag{4.3}$$

Projections in any other arbitrary view, whose coordinates are denoted as $\mathbf{p'}$, can be written as

$$\mathbf{p'} = \mathbf{A'_c}\mathbf{A_{22}}\mathbf{A_c}^{-1}\mathbf{p} + \mathbf{A'_c}[a_{13}\ a_{23}]^T P_{w_z}, \tag{4.4}$$

where $\mathbf{A_{22}}$ is the top–left 2×2 sub–matrix and a_{ij} the ijth element of $\mathbf{A}$.

If all the scene points lie on a plane $\mathbf{g}^T\mathbf{P_{wxy}} + P_{w_z} = 0$, then the relation between projections in the reference view and an arbitrary view can be written as:

$$\mathbf{p'} = \mathbf{A'_c}[\mathbf{A_{22}} - \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix} \mathbf{g}^T]\mathbf{A_c}^{-1}\mathbf{p} \tag{4.5}$$

This is the well known result that projections of a plane under weak perspective transformation are related through an affine transformation.

Let $\mathbf{g}^T\mathbf{P_{wxy}} + P_{w_z} = 0$ be a reference plane defined in the scene coordinate system. Consider an arbitrary point (not lying on the reference plane) in the reference view. The view ray for this point intersects the reference plane at some point. So, the reference image projection for both these points (the original point and its planar intersection) is the same. P_{w_z} for the arbitrary point can be written as a sum of the z–component of the corresponding planar point, say $P^{pl}_{w_z}$, and the out–of–plane z–component, $P^{np}_{w_z}$. Since, the planar component satisfies equation 4.5, the projection relation for the arbitrary point in two views is

$$\mathbf{p'} = \mathbf{A'_c}[\mathbf{A_{22}} - \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix} \mathbf{g}^T]\mathbf{A_c}^{-1}\mathbf{p} + \mathbf{A'_c}[r_{13}\ r_{23}]^T P^{np}_{w_z}. \tag{4.6}$$

We write this more compactly as

$$\mathbf{p'} = \mathbf{A_{im}}\mathbf{p} + Z_{np}\mathbf{b} \tag{4.7}$$

where the new symbols have the obvious correspondence with those in equation 4.6. The first part of the transformation is due to the affine planar component and the second is due to the non–planar component. Thus, the image motion of points has been decomposed into a planar component with respect to a reference plane in a reference view, and an out–of–plane component. Furthermore, the non–planar motion is decomposed into a component that depends on the relative structure of the scene in the reference view, and another component that depends on the view transformation. Also note that for any arbitrary view, the vector $\mathbf{b}$ is fixed for all the points. So, the non–planar vectors for each point in the view are parallel. Their magnitude is directly proportional to the out–of–plane depth component. If an additional reference point is chosen, then the relative magnitude of the non–planar displacement of any other point with respect to this reference point gives a view–invariant representation of the structure of the scene.

In the formulation above, the scene-to-camera transformations are arbitrary 3D affine transformations. Therefore, the view–invariant representation of structure derived above is invariant to general affine view transformations. Also, since the projection transformations in equations (3.5) and (3.4) for **WP** and **PP** are similar in that they have a linear and a translational part, the above decomposition of image motion into planar and non–planar components is valid for **PP** too.

Any three points present both in the reference view and an arbitrary view can be used to define a 2D affine transformation corresponding to the reference plane passing through those three points. The correspondence of a fourth point specifies the invariant structure completely. Three coordinates are being specified from two views to compute an invariant structure representation of the object. The in–plane 2D affine coordinates can be computed in the reference view itself and the third coordinate is computed using the additional view. Alternately, if the image plane coordinate system in an arbitrary view is warped to account for the planar 2D affine transformation of the reference plane, then the residual motion is due only to the non–planar structure. This represents the one parameter of scene structure with respect to an intrinsic coordinate system of the reference plane.

5. The Perspective Case

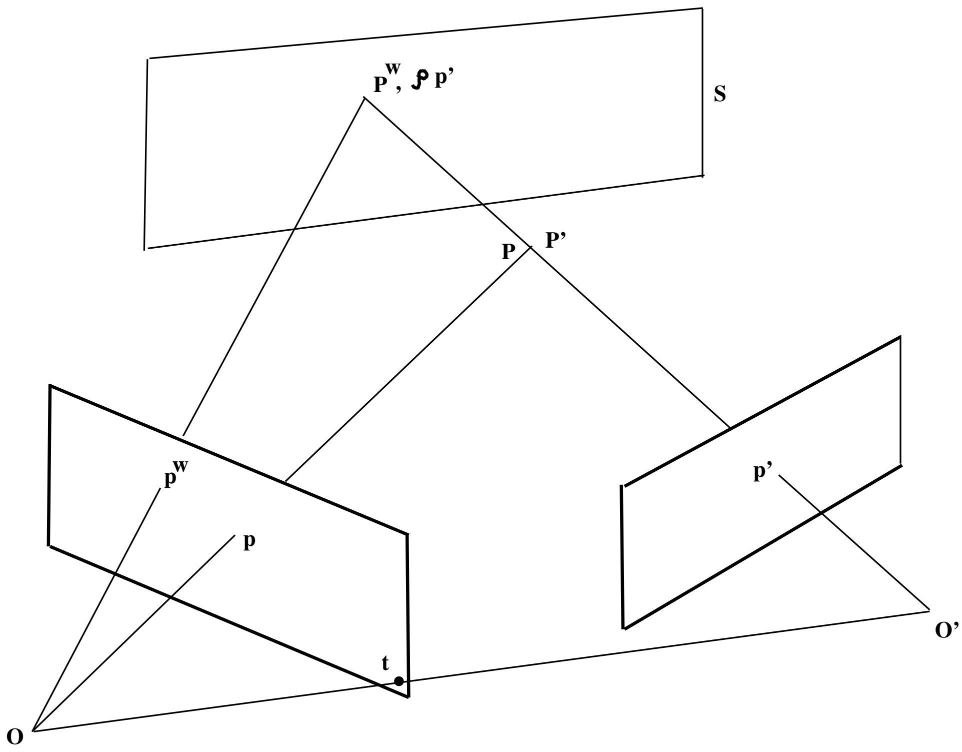

Figure 4: **Two-view Planar Parallax.**

In the following formulation, a reference view and any other arbitrary view are chosen to present the motion parallax equations. 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:

$$\mathbf{P} = \mathbf{A}\mathbf{P}' + \mathbf{T} \tag{5.1}$$

Let $\mathbf{N}'^{\mathbf{T}}\mathbf{P}' = d'$ represent a plane in the second coordinate system. Substituting this in the above equation, one can write the view transformation for the plane as [9]:

$$\mathbf{P}^{\mathbf{w}} \quad = \quad \mathbf{A}[\mathbf{I} + \mathbf{A}^{-1}\mathbf{T}\mathbf{N}'^{\mathbf{T}}/d']\mathbf{P}' \tag{5.2}$$

Note that this represents the general 8–parameter projective relationship for plane–to–plane projection. Using the identity $[\mathbf{I} + \mathbf{A}^{-1}\mathbf{T}\mathbf{N}'^{\mathbf{T}}/d']^{-1} = [\mathbf{I} - \beta\mathbf{A}^{-1}\mathbf{T}\mathbf{N}'^{\mathbf{T}}/d']$ (see [8]), where $\beta = (1/(1+\mathbf{N}'^{\mathbf{T}}\mathbf{A}^{-1}\mathbf{T}/d'))$, the above relationship can be written as the following projective transformation:

$$\rho\mathbf{p}' \quad = \quad [\mathbf{I} - \beta\mathbf{A}^{-1}\mathbf{T}\mathbf{N}'^{\mathbf{T}}/d']\mathbf{A}^{-1}\mathbf{P}^{\mathbf{w}} \tag{5.3}$$

$\mathbf{p}'$ is the image plane vector $(p'_x, p'_y, 1)$ in the reference view, and ρ is an unknown scale factor that assures that $\rho\mathbf{p}'$ lies on the reference plane. This is shown in figure 4. Equation 5.1 can be written in terms of $\mathbf{p}'$ and an unknown scale factor k as:

$$\mathbf{P} = \mathbf{A}k\mathbf{p}' + \mathbf{T} \tag{5.4}$$

Substituting for $\mathbf{p}'$ from equation 5.3, after some algebraic manipulations (using equation (5.2)), we get:

$$\mathbf{P}^{\mathbf{w}} \quad = \quad \frac{\rho}{k}[\mathbf{I} + \mathbf{T}\mathbf{N}^{\mathbf{T}}/d'](\mathbf{P} - \mathbf{T}) \quad = \quad \frac{\rho}{k}[\mathbf{P} + \frac{d_N}{d'}\mathbf{T}] \tag{5.5}$$

where $\mathbf{N} = \mathbf{A}^{-1^T}\mathbf{N}'$ is the plane normal in the reference view, and $d_N = \mathbf{P}^T\mathbf{N} - \mathbf{T}^T\mathbf{N} - d'$ is the perpendicular distance of $\mathbf{P}$ from the plane. In order to see that the parallax vectors between the warped points, $\mathbf{P}^{\mathbf{w}}$, and the actual points, $\mathbf{P}$, are directed towards the epipole, it is easily shown from equation (5.5) that

$$\mathbf{T} \cdot (\mathbf{P}^{\mathbf{w}} \times \mathbf{P}) = 0. \tag{5.6}$$

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 [12] 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 which is the focus of this paper. Also, the derivation here is valid for an arbitrary view transformation of the type in equation (5.1), 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 (5.3), the projection of the virtual planar point (intersection of $\mathbf{p}'$ with the plane) in the reference view can be computed using equation (5.3). In other words, points in any arbitrary view can be transformed (or warped) so that they project to the corresponding virtual planar points in the reference view. 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 4):

$$\mathbf{p} - \mathbf{p^w} = (1/(1 + \frac{P_z}{d_N}/\frac{T_z}{d'}))(\mathbf{p} - \mathbf{t}) \tag{5.7}$$

where the lower case bold letters represent the respective image vectors with their z–components unity, d_N is as defined above. Note that the internal camera transformation, $(\mathbf{A_c}, \mathbf{T_c})$, can be applied to each of the image vectors, $\mathbf{p}, \mathbf{p^w}, \mathbf{t}$, in equation (5.7) without changing its form. Thus, the equation is valid for arbitrary camera parameters that can change from view-to-view.

A few of the steps leading to equation (5.7) are:

$$\begin{aligned}
\mathbf{p} - \mathbf{p^w} &= \frac{1}{P_z}\mathbf{P} - \frac{1}{P_z^w}\mathbf{P^w} \\
&= \frac{1}{P_z}\mathbf{P} - (1/(P_z + \frac{d_N}{d'}T_z))(\mathbf{P} + \frac{d_N}{d'}\mathbf{T}) \\
&= (1/(1 + \frac{P_z}{d_N}/\frac{T_z}{d'}))(\mathbf{p} - \mathbf{t})
\end{aligned}$$

In equation (5.7), d', the distance of the reference plane from the origin of the second image coordinate system, can be replaced by $-T_d$, the distance of the translation vector from the reference plane. In fact, $d' = \mathbf{P'}^T\mathbf{N}' = \mathbf{P}^T\mathbf{N} - \mathbf{T}^T\mathbf{N} = -T_d$. Thus,

$$\mathbf{p} - \mathbf{p^w} = (1/(1 + \frac{P_z}{d_N}/\frac{T_z}{-T_d}))(\mathbf{p} - \mathbf{t}). \tag{5.8}$$

When T_z is zero, the parallax equation becomes:

$$\mathbf{p} - \mathbf{p^w} = (-d_N/P_z)[T_x\ T_y\ 0]^T \tag{5.9}$$

In this case, the parallax motion vectors are all parallel, oriented towards the epipole at infinity as in the case of weak/paraperspective in equation 4.7. How can one decide when a **WP** assumption is good enough ? Given two frames, the first being the reference frame, say the parallax vectors turn out to be all parallel. Then, the parallax vectors should be computed with the other frame as the reference frame. If the parallax vectors are again parallel, then either there is no rotation (and $T_z = 0$ translation), or **WP** is a valid assumption. The former is not an interesting special case. The latter is true because with rotation, under perspective projection, the epipole in the second frame will not be at infinity, leading to parallax vectors that have a finite focus of expansion/contraction. In contrast, for **WP**, the parallax vectors are always parallel independent of which frame is chosen as the reference frame.

The above derivation of planar parallax is in terms of the parallax vector between the projections of the "pseudo-points" corresponding to the planar projection, and the actual projections of the points in the reference image (figure 4). A very similar derivation with identical results but using an explicit warping of the *coordinate system* of the second image frame, according to the planar transformation with respect to the reference frame, is shown in the appendix.

For an alternative but more tedious derivation of the above results see [13]. 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 [11].

We have shown that the parallax vector defined with respect to an arbitrary plane is directed towards the epipole in the reference image. Thus, the parallax vector field is due only to the translational component of the 3D view transformation, as is expected of any motion parallax field. The effect of rotations on the image motion has been eliminated by choosing a warping transformation corresponding to a plane in the environment. In the warped coordinate system, the motion disparity of the plane is zero. In other words, the points on the plane have been fixated through a coordinate transformation. The residual image motion is due only to the non–planar component of the environment, and translational motion. Recall that in traditional structure from motion algorithms, decomposing the image motion into rotational and translational components is hard because of inherent ambiguities [1, 3]. This problem has been circumvented in the planar parallax approach because the rotations affect only the plane projective transformation of equation (5.2) and not the parallax motion. If the planar transformation is not decomposed into its rotational and translational component, then this method does not suffer from the inherent ambiguities.

In summary, the above derivation shows that the motion parallax vectors, defined as the residual motion vectors after the motion of a reference plane has been subtracted, lie along the epipolar lines, and hence all intersect at the epipole. Therefore, their direction directly encodes the epipolar information. Furthermore, their relative magnitude depends only on the 3D structure of the scene with respect to the reference plane.

5.1. View–Invariant Representation

The magnitude of the parallax vector is a function of the non–planar distance, d_N of $\mathbf{P}$, and the z–components of $\mathbf{P}$ and $\mathbf{T}$. Let $\eta = 1/(1 + P_z/d_N \frac{T_z}{d})$ be the magnitude 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:

$$\tau_i/\tau 0 = \frac{P_{i_z}}{d_{i_N}} / \frac{P_{0_z}}{d_{0_N}} \tag{5.10}$$

That is, the ratio of the magnitudes of the non–planar parallax motion components are dependent only on the relative structure of the environmental points not lying on the reference plane. 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 magnitude of the residual parallax vectors is always that given in equation (5.10). Thus, fixation with respect to the reference plane not only compensates for the effects of rotations, but also provides an environment centered reference surface with respect to which the complete shape of the environment can be specified.

5.2. Affine-Invariant 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 (5.7), and then solving for the rotation and the plane parameters from equation (5.2), or (ii) by solving for the plane and motion parameters directly from equation (5.2) [7]. 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 $\frac{P_{0_z}}{d_{0_N}}$ for a reference non-planar point to be unity, all the other points can be reconstructed using their respective ratios $\frac{P_{i_z}}{d_{i_N}}$ and their view rays $\mathbf{p}$. In particular, say for a given point, $\frac{P_z}{d_N} = \alpha$, 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

$$\mathbf{P} = ((\alpha d)/\mathbf{p}^T(\alpha\mathbf{N} - \mathbf{z}))\mathbf{p}, \tag{5.11}$$

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 (5.11) represents the 3D geometry of the scene up to an arbitrary 3D affine transformation. To see this, consider 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. 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

$$\mathbf{P_c} = \mathbf{A}\mathbf{P_w} + \mathbf{T}. \tag{5.12}$$

Note that the internal camera transformation, $\mathbf{A_c}, \mathbf{T_c}$, 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 ${\mathbf{P_c}}^T\mathbf{N} = d$. With $\mathbf{P_0}$ and the plane thus defined, the ratio $\frac{P_{0_z}}{d_{0_N}}$ is fixed.

For any other non-planar point (fifth and more), say, the ratio $\frac{P_z}{d_N}$ is α. Then, as in equation (5.11),

$$\mathbf{P_c} = ((\alpha d)/{\mathbf{p_c}}^T(\alpha\mathbf{N} - \mathbf{z}))\mathbf{p_c}, \tag{5.13}$$

defines the 3D $\mathbf{P_c}$. 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 in [21], [6] and [16].

Note that in the reconstruction above using four scene points, no explicit reconstruction of the 3D motion is required. Also, any arbitrary view, when warped for the reference plane with respect to the reference view, will lead to the same 3D reconstruction in a canonical coordinate system defined by the four chosen points.

The derived representation contains the necessary representations both for motion and structure reconstruction and matching. If the absolute metric motion and structure information is required, then the translation can be computed from the parallax field leading subsequently to the rotation [13] and the absolute coordinates in any particular view reference frame. If the goal is recognition and matching, then absolute metric structure need not be computed. From two views, given four planar and one out–of–plane point, the intrinsic structure representation can be computed. For any arbitrary view, if six points (four in-plane and two out-of-plane) can be matched, then a synthetic image for the new viewpoint can be created using the intrinsic structure. This can then be compared with the given view. Similarly for new view generation, if the mapping of the six points is available, then a new view can be created without knowing the absolute structure and the motion between the views. For new view generation, at least two out-of-plane points are required because the epipole needs to be computed. The parallax vectors for the two non-planar points can be used for this. In practice, many more parallax vectors should be used to solve for the epipole using least-squares.

5.3. Interpretation of the Parallax Magnitude

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

$$\mathbf{P_c} = ((\alpha d)/\mathbf{p_c}^T(\alpha\mathbf{N} - \mathbf{z}))\mathbf{p_c} = (1/(1 - \frac{1}{\alpha}))\mathbf{p_c}. \tag{5.14}$$

Recall that the magnitude of the parallax vector from equation 5.8 is proportional to $(1/(1 + \frac{P_z}{d_N}/\frac{T_z}{-T_d}))$. $\frac{T_z}{T_d}$ can be conveniently set to unity to fix the overall scale. Then, the magnitude becomes $(1/(1 - \frac{P_z}{d_N}))$ which is the same as $(1/(1 - \frac{1}{\alpha}))$ in equation 5.14 because $\alpha = \frac{P_z}{d_N}$. Therefore, in the coordinates of the reference image, the length of the parallax vector is directly *the depth of the corresponding point.* Of course, the structure reconstruction is valid up to an arbitrary 3D affine transformation as shown above. This is similar to Koenderink's [10] and Shashua's [20] affine structure from motion under weak perspective, and Shashua's [21] projective depth under perspective projection.

An important point to note here is that the "affine depth" computed in equation 5.14 does not require the explicit computation of the epipoles as in Shashua's method.

5.4. Relationship to Shashua's Projective Depth

In [21], 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 5 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

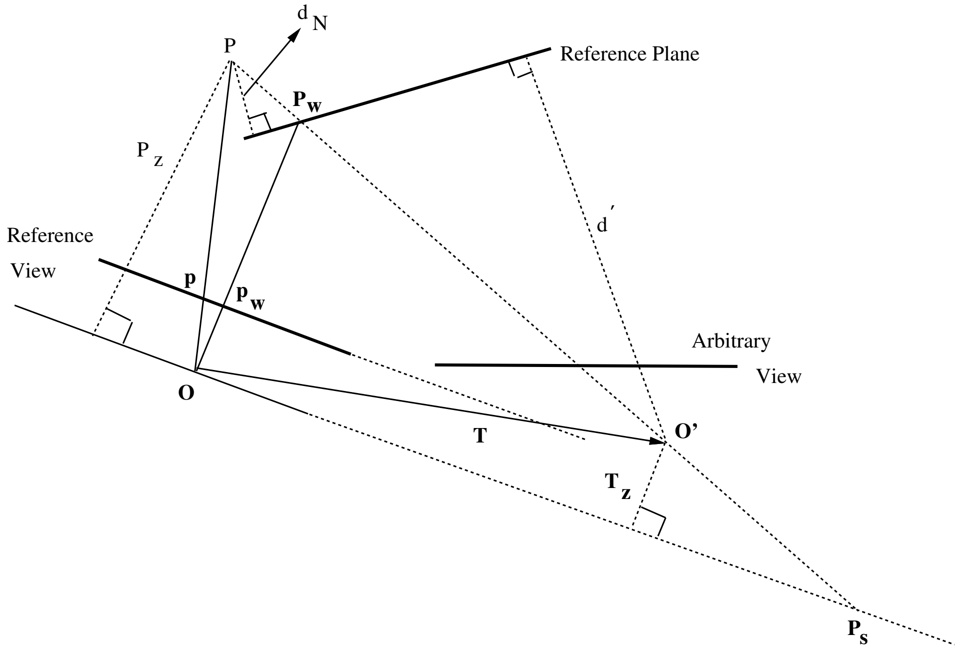

Figure 5: **Relationship with Shashua's two-plane cross-ratio.**

$\mathbf{O}'$, 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 $(\mathbf{PP_s}/\mathbf{O'P_s})/(\mathbf{PP_w}/\mathbf{O'P_w})$. From similar traingles in figure 5, this is exactly the ratio $((P_z/T_z)/(d_N/d))$ given by the motion parallax equation (5.7). 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.

6. Experimental Results

We demonstrate the application of planar motion parallax on images of a rotating box. Two frames from a sequence are shown in figures 6 and 7. The box was held by a gripper and was rotated around an axis going through the opposite face centers. The magnitude of rotation between the two frames is approximately 4^o. The background is stationary. A SONY B/W AVC-D1 camera with effective FOV 24 by 23 degrees was used to capture 512×484 images. These were reduced to 256×242 for the experiments. The range of depths in the scene is about 550 to 700 mm.

All the processing on the images is done using direct methods developed by the Sarnoff Research Center group [5]. 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 8. The difference between this warped image and the reference image, *BOX1*, is shown in figure 9. 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 **WP** might be a good enough model as was noted by Daphna Weinshall in [23]. In the difference image (figure 9), 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 [5] is applied between the reference image and the affine warped image, *BOX2AFFW*. The flow vectors are shown in figure 10. (Due to the display program used, the vector display is upside down.) This process of registration using a general flow algorithm almost completely cancels the residual translation motion as shown in the difference image in figure 11. The residual non-planar motion vectors produced by this registration correspond to the equations (4.7) and (5.7). 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 figures 12 and 13. A shaded plot (whose hard copy reproduction is not too good, unfortunately) is shown in figure 14. The viewpoint has been chosen to make the computed shape fairly explicit. (All the surface plots use 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 plots clearly show that the qualitative estimates of the planar facets of the box and the overall shape have been recovered fairly well. We are in the process of computing quantitative estimates and comparing these with the ground truth. Also, we are experimenting with more general scenes.

7. Conclusions

A new derivation for 3D structure estimation using planar parallax has been presented. It is shown that this approach unifies the ideas of intrinsic 3D structure from weak perspective and perspective projections. It is also shown that the plane-relative depth estimate obtained from our method is closely related to Shashua's projective depth. Our goal is to apply these formalisms to derive relative arrangement of objects and surfaces in scenes for the purposes of scene annotation in video sequences and for recognition and new view generation. Metric structure is not very important for these applications. Thus, the plane relative derived structure (even with unknown camera calibration) should be adequate for the tasks. We are in the process of developing a system for more elaborate experimentation with the formalism presented in this paper.

Acknowledgements

Many of the ideas in this paper evolved through discussions with Rakesh (Teddy) Kumar and P. Anandan of the David Sarnoff Research Center, Princeton, NJ, and with R. Manmatha, a visiting student from the University of Massachusetts, Amherst. Thanks to Chitra Dorai, a visiting student from Michigan State University, for her help in implementing the affine registration code. Thanks to Teddy for his help in the experiments.

References

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

[2] S. Carlsson and J. Eklundh. Obj. det. using model based pred. and motion parallax. In *1st ECCV*, pages 297–306, 1990.

[3] K. Daniilidis and H. H. Nagel. The coupling of rotation and translation in motion estimation of planar surfaces. In *Proc. Computer Vision and Pattern Recognition Conference*, pages 188–193, 1993.

[4] W. Enkelmann. Obst. detection by evaluation of opt. flow fields from image seqs. *IVC*, 9(3):160–168, 1991.

[5] J. R. Bergen et al. Hierarchical model–based motion estimation. In *Proc. 2nd European Conference on Computer Vision*, pages 237–252, 1992.

[6] O. D. Faugeras. What can be seen in three dimensiona with an uncalibrated stereo rig ? In *Proc. 2nd European Conference on Computer Vision*, pages 563–578, 1992.

[7] O. D. Faugeras and F. Lustman. Let us suppose the world is piece-wise planar. In *Proc. The Third International Symposium on Robotics Research*, 1987.

[8] R. Hartley and R. Gupta. Computing matched epipolar projections. In *Proc. Computer Vision and Pattern Recognition Conference*, pages 549–555, 1993.

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

[10] J. J. Koenderink and Andrea J. van Doorn. Affine structure from motion. *Journal of the Optical Society of America A*, 81:377–385, 1991.

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

[12] J. Lawn and R. Cipolla. Epipole estimation using affine motion parallax. Technical Report CUED/F-INFENG/TR 138, Cambridge University Engineering Department, 1993.

[13] Chia-Hoang Lee. Structure and motion from two perspective views via planar patch. In *Proc. 2nd Intl. Conf. on Computer Vision*, pages 158–164, 1988.

[14] Chia-Hoang Lee and T. Huang. Finding point correspondences and determining motion of a rigid object from two weak perspective views. *Computer Vision Graphics and Image Processing*, 52:309–327, 1990.

[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] R. Mohr, F. Veillon, and L. Quan. Relative 3D reconstruction using multiple uncalibrated images. In *Proc. Computer Vision and Pattern Recognition Conference*, pages 543–548, 1993.

[17] J. L. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision*. The MIT Press, MA, 1992.

[18] C. J. Poelman and T. Kanade. A paraperspective factorization method for shape and motion recovery. Technical Report CMU–CS–92–208, Carnegie Mellon University, 1992.

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

[20] Amnon Shashua. Correspondence and affine shape from two orthographic views: Motion and recognition. Technical Report AI Memo No. 1327, Massachusetts Institute of Technology, 1991.

[21] Amnon Shashua. Projective depth: A geometric invariant for 3D reconstruction from two perspective/orthographic views and for visual recognition. In *Proc. 4th Intl. Conf. on Computer Vision*, pages 583–590, 1993.

[22] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992.

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

Appendix

Let the 3D affine transformation between a point in the reference view, $\mathbf{P}$, and a second view, $\mathbf{P}'$, be

$$\mathbf{P}' = \mathbf{A}'\mathbf{P} + \mathbf{T}'. \tag{7.1}$$

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 [9]:

$$\mathbf{P}' \quad \approx \quad [\mathbf{A}' + \mathbf{T}'\mathbf{N}^T/d]\mathbf{P} \tag{7.2}$$

If the second frame's coordinate system is warped with respect to this plane projective transformation, then the warped projective coordinates are

$$\mathbf{P^w} \quad \approx \quad [\mathbf{A}' + \mathbf{T}'\mathbf{N}^T/d]^{-1}\mathbf{P}'. \tag{7.3}$$

This can be written as

$$\begin{aligned} \mathbf{P^w} \quad &\approx \quad [\mathbf{I} - \beta\mathbf{A}'^{-1}\mathbf{T}'\mathbf{N}^T/d]\mathbf{A}'^{-1}\mathbf{P}' \\ &\approx \quad [\mathbf{I} + \beta\mathbf{T}\mathbf{N}^T/d]\mathbf{A}'^{-1}\mathbf{P}', \end{aligned} \tag{7.4}$$

where $\beta = 1/(1 - \mathbf{N}^T\mathbf{T}/d)$, and $\mathbf{T} = -\mathbf{A}'^{-1}\mathbf{T}'$ is the translation (displacement of the second frame's origin) in the reference coordinate system.

Equation (7.4) represents the warping transform applied to the second image coordinates to take 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 (7.1) in equation (7.4), we get

$$\begin{aligned} \mathbf{P^w} \quad &\approx \quad [I - \beta\mathbf{A}'^{-1}\mathbf{T}'\mathbf{N}^T/d]\mathbf{A}'^{-1}(\mathbf{A}'\mathbf{P} + \mathbf{T}') \\ &\approx \quad [\mathbf{I} + \beta\mathbf{T}\mathbf{N}^T/d](\mathbf{P} - \mathbf{T}) \\ &\approx \quad (\mathbf{P} + \gamma\mathbf{T}), \end{aligned} \tag{7.5}$$

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.

Therefore,

$$\begin{aligned} \mathbf{p} - \mathbf{p^w} \quad &= \quad \frac{1}{P_z}\mathbf{P} - \frac{1}{P_z^w}\mathbf{P^w} \\ &= \quad \frac{1}{P_z}\mathbf{P} - \frac{1}{P_z + \gamma T_z}(\mathbf{P} + \gamma\mathbf{T}) \\ &= \quad (1/(1 + \frac{P_z}{d_N}/\frac{T_z}{-T_d}))(\mathbf{p} - \mathbf{t}), \end{aligned} \tag{7.6}$$

which is the same as the earlier derived equation (5.8).

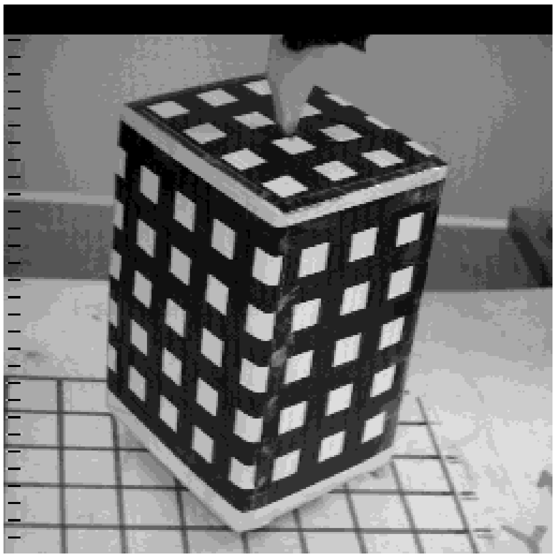

Figure 6: **Frame 1 of the box scene.**

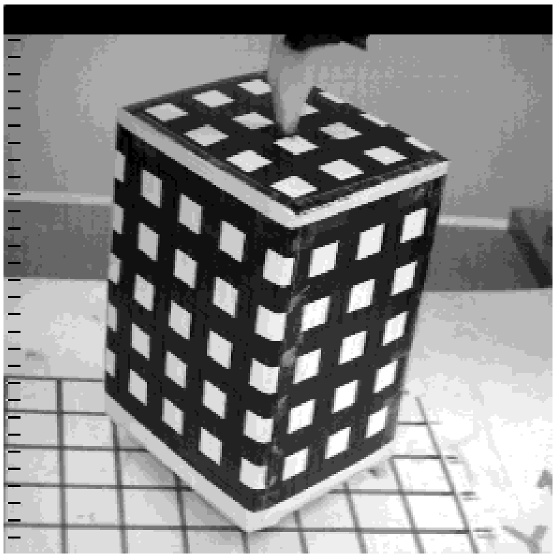

Figure 7: **Frame 2 of the box scene.**

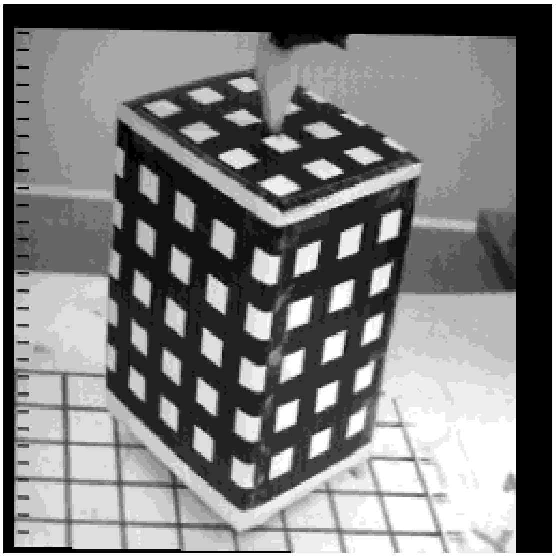

Figure 8: **Frame 2 warped using the affine transformation cooresponding to the left face.**

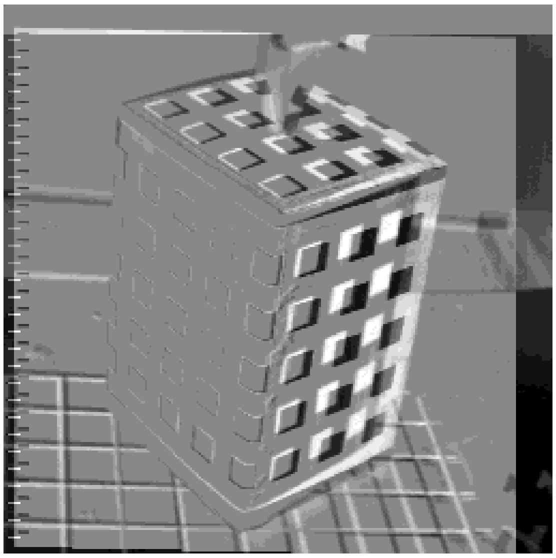

Figure 9: **Difference between frame 2 affine warped and frame 1.**

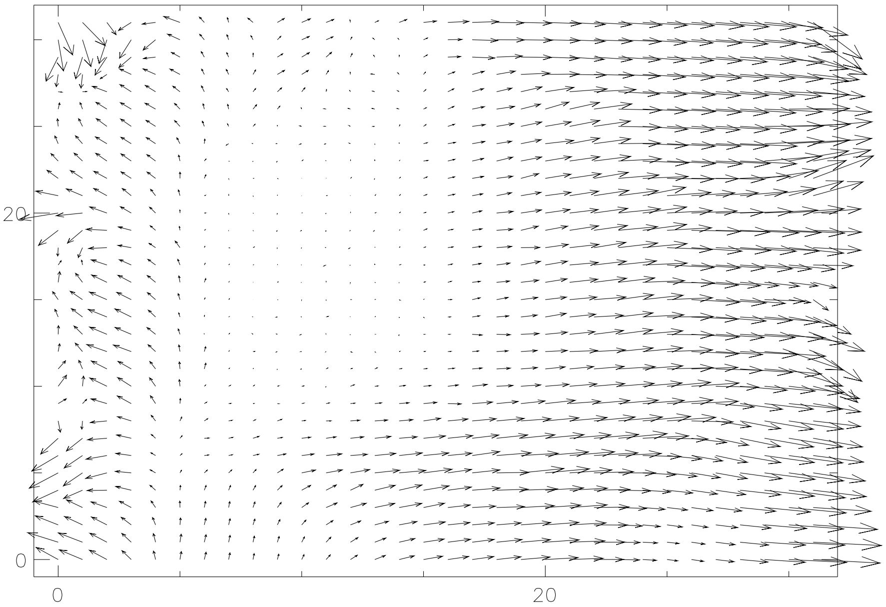

Figure 10: **The non-planar residual flow.**

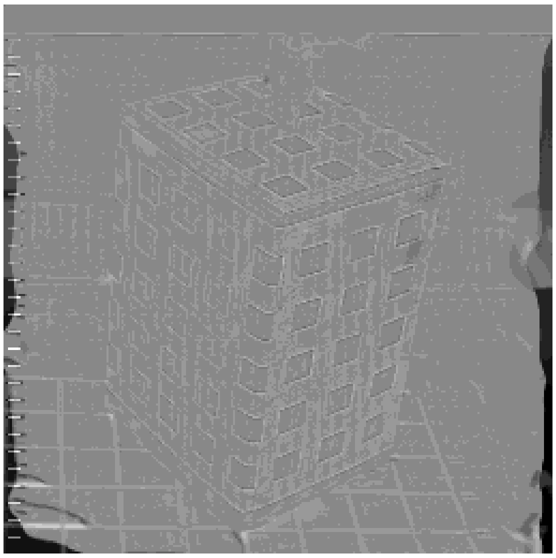

Figure 11: **Difference between affine warped and general flow warped frame 2 and frame 1.**

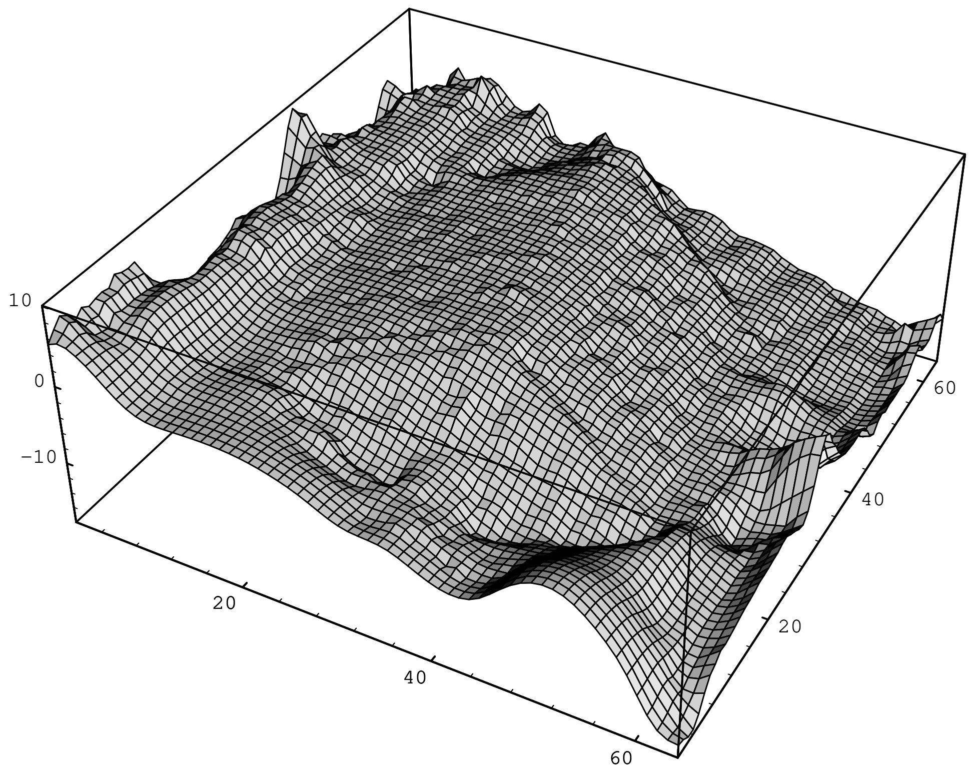

Figure 12: **Grided surface plot of the box.**

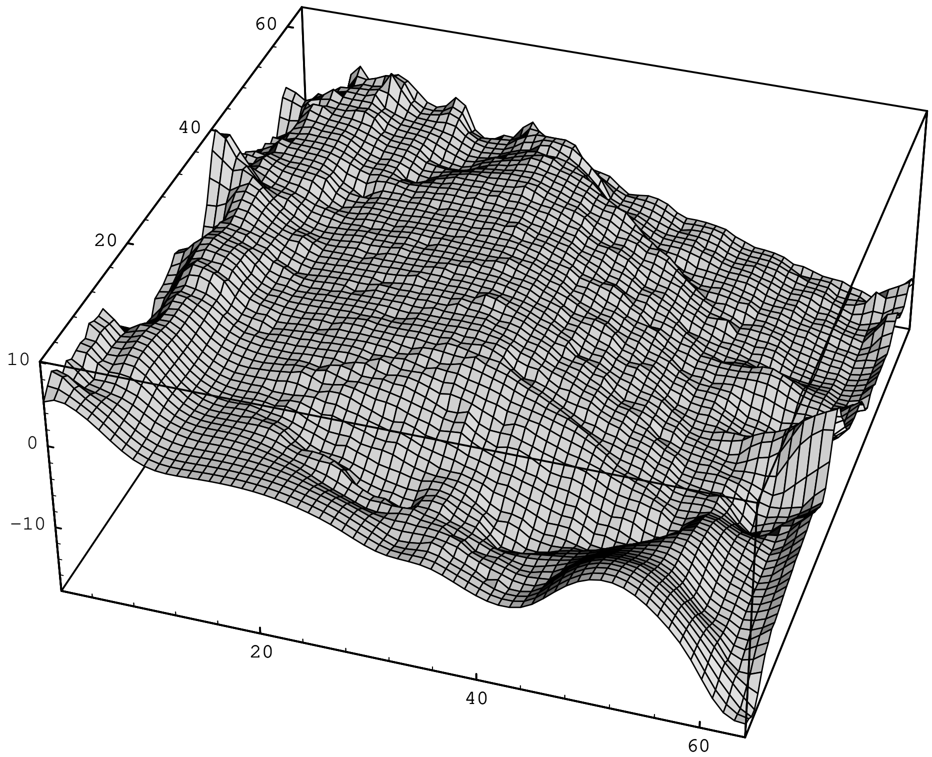

Figure 13: **Grided surface plot of the box.**

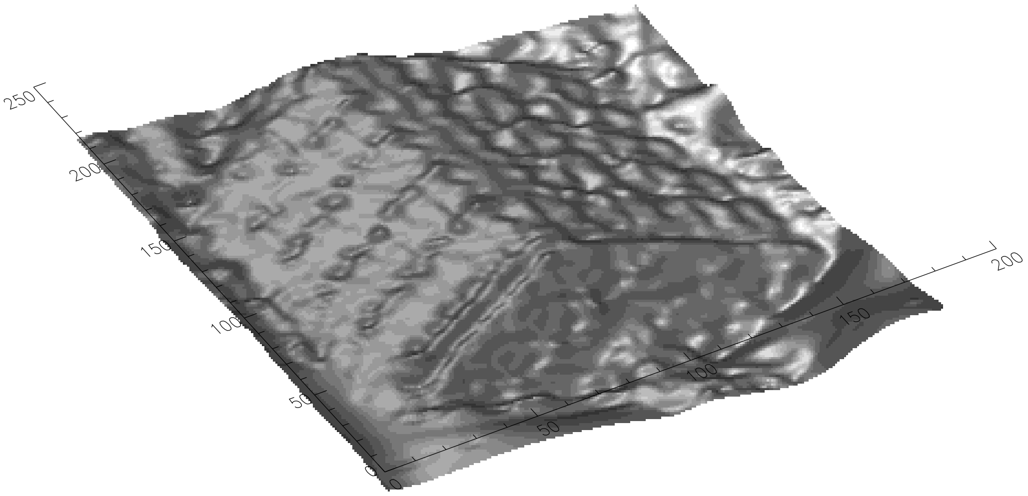

Figure 14: **Shaded surface plot of the box.**