From Ordinal to Euclidean Reconstruction with Partial Scene Calibration*

Daphna Weinshall
Inst. of Computer Sci.
Hebrew University
91904 Jerusalem, Israel
daphna@cs.huji.ac.il

P. Anandan
Microsoft Research
One Microsoft Way
Redmond, WA 98052
anandan@microsoft.com

Micahl Irani
Dept. of Appl. Math and CS
The Weizmann Inst. of Sci.
Rehovot, Israel
irani@wisdom.weizmann.ac.il

Abstract

Since uncalibrated images permit only projective reconstruction, metric information requires either camera or scene calibration. We propose a stratified approach to projective reconstruction, in which gradual increase in *domain* information for scene calibration leads to gradual increase in $3D$ information. Our scheme includes the following steps: (1) Register the images with respect to a reference plane; this can be done using limited scene information, e.g., the knowledge that two pairs of lines on the plane are parallel. We show that this calibration is sufficient for ordinal reconstruction - sorting the points by their height over the reference plane. (2) If available, use the relative height of two additional out-of-plane points to compute the height of the remaining points up to constant scaling. Our scheme is based on the dual epipolar geometry in the reference frame, which we develop below. We show good results with five sequences of real images, using mostly scene calibration that can be inferred directly from the images themselves.

Keywords: projective reconstruction, affine reconstruction, partial calibration, qualitative depth

1 Introduction

Given multiple images, stratified $3D$ reconstruction can be obtained depending on the available camera and scene calibration. In general uncalibrated images permit only projective reconstruction [6], which is of limited use; for example, we cannot determine from projective structure which part of the object is in front of the other. Its topological nature makes this representation useful primarily for verification, e.g., object recognition; for most other applications some scene or camera calibration is needed.

With calibrated cameras, or if self calibration is possible (when the same camera with partially fixed internal parameters is used to obtain all the images), Euclidean reconstruction can be obtained [10, 18]. Alternatively, active vision techniques, based on imposing constraints on the viewing geometry and/or the camera motion, can be used. Such externally imposed constraints may simplify the problem enough to permit affine or Euclidean reconstruction, e.g., [23] (see also [24]). However, active vision techniques cannot be universally applied, and in particular they are of little help when the sequence of images is already given (such as the case with video analysis). Similarly, techniques which obtain Euclidean reconstruction using self camera calibration cannot be used with just any sequence unless it is known that some of the internal parameters of the camera taking the sequence were kept fixed or are known.

The alternative to active vision and self camera calibration is to use scene calibration. Thus projective reconstruction can be turned into Euclidean reconstruction if the $3D$ coordinates of five points are given [15]. Computing Euclidean reconstruction from affine reconstruction requires that the $3D$ coordinates of four points are given. These results were used to formulate a stratified approach to reconstruction, characterized

*MI and DW are supported in part by DARPA through ARL Contract DAAL01-97-K-0101. This research was done while DW was on sabbatical at NECI Princeton.

by invariance to increasingly smaller groups of transformations in $\mathcal{R}^3$: projective, affine and similarity (scaled Euclidean) [7]. The usefulness of techniques using scene calibration is limited to cases when the needed $3D$ information is available. Unlike the case with camera calibration and active vision, there are no results on partial scene calibration which could give partial metric information.

Our approach fills in this gap (a related approach is independently described in [5]); we investigate partial scene calibration which can be used to obtain some metric information from uncalibrated images. Unlike previous approaches using scene calibration, in which the $3D$ information had to be given a-priori, we make use of scene information that can be inferred automatically from images, such as that lines are parallel. This information only permits ordinal reconstruction in our scheme. In addition, in order to achieve affine reconstruction we need to know the relative height of one $3D$ point; this requirement still seems easier to meet than the knowledge of the $3D$ coordinates of five $3D$ points, as required in [15].

More specifically, we compute non-invariant reconstruction in a special coordinate system. This coordinate system is defined relative to a physical (real or virtual) planar surface in the scene. The projection of the scene points onto an input camera image is decomposed into two stages: (i) the projection of each scene point through the focal-point of the camera onto the *reference plane*, and (ii) the re-projection of the reference plane image onto the camera image plane. The projection from $3D$ to the reference plane depends only on the $3D$ positions of the scene points and the focal-point of the camera. All effects of the camera's internal calibration and the orientation of the image plane are folded into the re-projection step, which is captured by a homography between the reference plane and the camera image plane.

The reconstruction of the $3D$ scene is done relative to the reference plane, within a Cartesian coordinate system whose $X-Y$ axes span the reference plane, and whose Z direction is perpendicular to it. Given multiple images of the same scene, the homography relating each image to the reference plane is determined by specifying a few pieces of geometric information about the reference plane. This homography is then applied to each camera image to determine the corresponding reference-plane image. The "Height" (Z) of each scene point is determined by analyzing the disparity between the positions of each scene point on the multiple reference plane images[1]. The basic relationships associated with multi-view parallax geometry (with respect to a reference plane) are described in our recent paper [1]. However, while [1] focuses on the geometric relationships and elaborated on one application (namely "new view synthesis"), the present paper focuses on the use of this framework for stratified reconstruction, based on partial scene calibration.

The specification of the geometric information about the reference plane amounts to "registering" the reference plane. Such registration, with no $3D$ calibration, is sufficient to determine whether points lie on the same side of the plane [17]. Here we show that ordinal information about the $3D$ scene can be obtained even with very little scene calibration data, e.g., we can compute height ordering with respect to a plane from knowing (or guessing) the existence of two pairs of parallel lines on the plane; we call this **ordinal reconstruction** (cf. [22, 13, 9]). By providing additional *domain-information* (e.g., heights of one or two out-of-plane points), we can gradually obtain more metric $3D$ information, achieving affine and Euclidean reconstruction.

In conclusion, given only partial scene calibration, we can compute $3D$ reconstruction based on a cascade of representations which are more general than affine but more specific than projective. Our approach provides intermediate structures between the Euclidean and projective structures. This fills an important gap in the reconstruction literature, using uncalibrated images and partial *scene* calibration. Our approach is useful when Euclidean shape cannot be computed because camera or scene calibration is not given and cannot be determined, but projective shape is not sufficient since some metric information is needed (e.g., in active vision tasks such as grasping and obstacle avoidance). Unlike most reconstruction algorithms, our approach uses the dual epipolar geometry for reconstruction.

The remainder of the paper is organized as follows: Section 2 describes the basic geometric relations associated with the reference-plane images. Section 3 describes how we use these relations for the stratified reconstruction of the $3D$ scene. Section 4 illustrates our approach with real-image examples.

[1] A related approach called "plane+parallax" was taken in [14, 19], but there 3D reconstruction was relative to the coordinate systems of both the reference plane and a reference image.

2 Geometry on the Reference Plane

The perspective projection of a point P_i in space to an image plane can be written as $p_{it} = M_t P_i$, where p_{it} and P_i are the point homogeneous coordinates in $2D$ and $3D$ respectively, and M_t is the 3×4 projection matrix describing the camera t. M_t depends on the orientation of the image plane and the location of the camera center P_t.

Here we break down the projection into 2 operations: the projection of the $3D$ world onto a $2D$ reference plane Π through the focal-point P_t, followed by a $2D$ projective transformation (homography) which maps the reference plane Π to the image plane of camera t.

Our purpose in this paper is to use this decomposition for the gradual reconstruction of the $3D$ scene relative to Π. The key idea in this paper is that by analyzing the images formed by projecting the scene through each camera center onto the reference plane Π, we vastly simplify the problem of $3D$ reconstruction. As explained in Section 3, the reference plane images from each camera view are obtained by registering each image to a pre-defined (affine or Cartesian) coordinate system on the reference plane.

2.1 Reference plane coordinate system

Let Π denote a (real or virtual) planar surface in the scene. We call Π the "reference plane". We define a $3D$ Cartesian coordinate system whose $X - Y$ axes span the reference plane Π, and the Z direction is perpendicular to Π. We call this system the "reference plane coordinate system".

An object, composed of n points in $3D$ space, is represented in the reference plane coordinate system by the **shape matrix P**, the following $4 \times n$ matrix:

$$\mathbf{P} = \begin{bmatrix} 0 & a_1 & a_2 & a_3 & a_4 & & X_i & X_{i+1} & \\ 0 & b_1 & b_2 & b_3 & b_4 & \cdots & Y_i & Y_{i+1} & \cdots \\ 1 & 0 & 0 & 0 & 0 & & Z_i & Z_{i+1} & \\ 0 & 1 & 1 & 1 & 1 & & W_i & W_{i+1} & \end{bmatrix} \tag{1}$$

Columns $[a_i \ b_i \ 0 \ 1]$ are the coordinates of points on the reference plane Π, $[0 \ 0 \ 1 \ 0]$ is a point at ∞ on the line through the origin which is perpendicular to the reference plane, and columns $[X_i \ Y_i \ Z_i \ W_i]^T$ are the coordinates of general points P_i on the object.

Every object can be represented this way; but the representation is *not* unique - a transformation from another projective coordinate system to this particular one is determined only up to an independent scaling of the Z axis. This ambiguity comes about because the standard projective basis requires that no four points are co-planar, whereas the basis of our system includes a plane. We need not worry about this ambiguity in the following derivations, however, because all quantities of interest involve ratios of Z values.

2.2 Projection on the reference plane

The $3D$ scene points are first projected onto the reference plane Π through the focal-point P_t of the camera. This forms a "virtual" image on Π. We refer to this image as the "reference-plane" image.

Corresponding to the object-shape matrix P defined in (1) and camera t, we define the following matrix $\mathbf{p_t}$ of reference-image points:

$$\mathbf{p_t} = \begin{bmatrix} \alpha_t & a_1 & a_2 & a_3 & a_4 & & x_{it} & x_{(i+1)t} & \\ \beta_t & b_1 & b_2 & b_3 & b_4 & \cdots & y_{it} & y_{(i+1)t} & \cdots \\ \gamma_t & 1 & 1 & 1 & 1 & & w_{it} & w_{(i+1)t} & \end{bmatrix}$$

Note that the first two coordinates of the points $1, \ldots 4$, which are physically located on the reference plane Π, are the same as the corresponding coordinates in their $3D$ representation. This is a direct consequence of our choice of the coordinate system for the $3D$ representation.

As explained before, the coordinates of the image points in the "original" input images are obtained by applying a $2D$ projective transformation to the corresponding reference plane images. That is, $\mathbf{q_t} = H_t \mathbf{p_t}$, where $\mathbf{q_t}$ is the "image matrix" corresponding to the original image t, and H_t is the $2D$ projective transformation (3×3 matrix) that relates the reference plane to the image plane of camera t.

As will become clear later, our *stratified* approach to $3D$ reconstruction does not require that the reference plane image coordinates be completely known. It is sufficient to specify them up to $2D$ affine deformation of the reference plane Π. Thus, it is sufficient to specify $\widehat{\mathbf{p}}_\mathbf{t} = G_t \mathbf{p_t}$ where $G_t = \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{21} & g_{22} & g_{23} \\ 0 & 0 & 1 \end{bmatrix}$ denotes the (possibly unknown) affine transformation of the reference plane. In the remainder of this paper, we call $\mathbf{p_t}$ the "fully normalized" reference plane image, and $\widehat{\mathbf{p}}_\mathbf{t}$ the "affine normalized" reference plane image.

If image $\mathbf{p_t}$ is fully normalized, it can be shown (or, more easily, verified) that the projection matrix M_t associated with it is:

$$M_t = \begin{bmatrix} \delta_t & 0 & \alpha_t & 0 \\ 0 & \delta_t & \beta_t & 0 \\ 0 & 0 & \gamma_t & \delta_t \end{bmatrix}$$

We impose the constraint that the focal point of the camera cannot be possibly seen in the (reference-plane) image; thus it is the null vector of M_t. We denote the $3D$ coordinates of the focal point P_t of the camera t by $[X_t\ Y_t\ Z_t\ W_t]$, then

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \propto \begin{bmatrix} \delta_t & 0 & \alpha_t & 0 \\ 0 & \delta_t & \beta_t & 0 \\ 0 & 0 & \gamma_t & \delta_t \end{bmatrix} \cdot \begin{bmatrix} X_t \\ Y_t \\ Z_t \\ W_t \end{bmatrix} \quad \Longrightarrow \begin{cases} 0 = \delta_t X_t + \alpha_t Z_t \\ 0 = \delta_t Y_t + \beta_t Z_t \\ 0 = \delta_t W_t + \gamma_t Z_t \end{cases}$$

and the solution is:

$$[\delta_t\ \alpha_t\ \beta_t\ \gamma_t] \propto [-Z_t\ X_t\ Y_t\ W_t]$$

The normalized reference-plane image $\mathbf{p_t}$ is therefore obtained by the following projection:

$$\mathbf{p_t} \propto \begin{bmatrix} -Z_t & 0 & X_t & 0 \\ 0 & -Z_t & Y_t & 0 \\ 0 & 0 & W_t & -Z_t \end{bmatrix} \mathbf{P}$$

Let $p_{it} = [x_{it}\ y_{it}\ w_{it}]^T$ denote the reference-plane image position of the object point P_i projected through the camera focal-point P_t (onto Π). Then:

$$p_{it} = \begin{bmatrix} x_{it} \\ y_{it} \\ w_{it} \end{bmatrix} \propto \begin{bmatrix} -Z_t & 0 & X_t & 0 \\ 0 & -Z_t & Y_t & 0 \\ 0 & 0 & W_t & -Z_t \end{bmatrix} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ W_i \end{bmatrix} \tag{2}$$

Observe that each point on the reference-plane image imposes 2 constraints relating to the coordinates of the object point P_i and the camera center P_t, which follow immediately from (2):

$$\frac{x_{it}}{w_{it}} = \frac{-Z_t X_i + X_t Z_i}{-Z_t W_i + W_t Z_i}, \qquad \frac{y_{it}}{w_{it}} = \frac{-Z_t Y_i + Y_t Z_i}{-Z_t W_i + W_t Z_i} \tag{3}$$

In this equation, the $3D$ coordinates of the $3D$ point P_i and the focal point P_t of camera t are **dual**: the focal point of the camera $[X_t\ Y_t\ Z_t\ W_t]$ is interchangeable with the $3D$ point $[X_i\ Y_i\ Z_i\ W_i]$.

From (3) we get the epipolar and dual-epipolar geometry on the reference plane.

2.3 Epipolar Geometry on the Reference Plane:

The epipolar geometry is obtained by the elimination of the $3D$ point coordinates from (3). In the following derivation we follow the method described in [8]. First, we rewrite (3) and note that each camera t imposes 2 constraints on the $3D$ point coordinates

$$\begin{pmatrix} w_{it} Z_t & 0 & (x_{it} W_t - w_{it} X_t) & -x_{it} Z_t \\ 0 & w_{it} Z_t & (y_{it} W_t - w_{it} Y_t) & -y_{it} Z_t \end{pmatrix} \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ W_i \end{bmatrix} = 0$$

Two cameras (e.g., t and s) give 4 constraints. Since they have a non-trivial solution $[X_i\ Y_i\ Z_i\ W_i]$, the determinant of their matrix must equal 0. This yields a relation between the focal points $P_t = [X_t\ Y_t\ Z_t\ W_t]$ and $P_s = [X_s\ Y_s\ p_{it} = [x_{it}\ y_{it}\ w_{it}]$ and $p_{is} = [x_{is}\ y_{is}\ w_{is}]$ of the point P_i in the two corresponding reference-plane images. This relation can be rewritten as:

$$p_{it}^T\ F_{ts}\ p_{is} = 0, \quad F_{ts} = \begin{bmatrix} 0 & Z_sW_t - Z_tW_s & -Z_sY_t + Z_tY_s \\ -Z_sW_t + Z_tW_s & 0 & Z_sX_t - Z_tX_s \\ Z_sY_t - Z_tY_s & -Z_sX_t + Z_tX_s & 0 \end{bmatrix} \tag{4}$$

and F_{ts} is the "fundamental" matrix.

(4) gives the **epipolar geometry *on the reference plane***—a relation between the coordinates of point P_i in the reference-plane image from camera t to its coordinates in the reference-plane image from camera s. The "fundamental matrix" F_{ts} is anti-symmetric, with essentially 3 unknowns: $(Z_sX_t - Z_tX_s)$, $(Z_sY_t - Z_tY_s)$ and $(Z_sW_t - Z_tW_s)$. We also observe that these are the coordinates of the point p_{ts}, which can be obtained by substituting the object point P_i by the camera-center P_s in (2). Thus, p_{ts}, which determines the fundamental matrix F_{ts}, is the projection of the focal point P_s through P_t on the reference plane—in other words, p_{ts} is the *epipole*.

The two epipoles: p_{st} - the projection of the focal point of camera s through the focal point of camera t, and p_{ts} - projection of the focal point of camera t through the focal point of camera s, are the same on the reference plane; they are defined by the intersection of the line going through the focal points of the two cameras and the reference plane. Thus the epipolar geometry on both image s and image t, when mapped to the reference plane, is the same: it is defined by the same anti-symmetric matrix F_{ts}, the epipole is the same, and the pencils of epipolar lines are the same. All this remains true even if the images are not normalized at all, that is, they are aligned with the reference plane up to some homography only.

Since each point provides one homogeneous constraint on these unknowns, 2 points (which are not on the reference plane) are sufficient to compute the epipolar geometry in full. However, because of this redundancy of the fundamental matrix, it is not possible to extract from it the coordinates of the focal points of cameras t and s.

2.4 Dual Geometry on Reference Plane

The dual epipolar geometry is obtained by eliminating the coordinates of the camera's focal point from (3). Once again, we first rewrite (3) so that each point P_i imposes 2 constraints on the coordinates of the focal point of the camera t:

$$\begin{pmatrix} w_{it}Z_i & 0 & (x_{it}W_i - w_{it}X_i) & -x_{it}Z_i \\ 0 & w_{it}Z_i & (y_{it}W_i - w_{it}Y_i) & -y_{it}Z_i \end{pmatrix} \begin{bmatrix} X_t \\ Y_t \\ Z_t \\ W_t \end{bmatrix} = 0$$

Two points give 4 constraints, which have a non-trivial solution. Following the same reasoning as before for points P_i,P_j, we get a relation between the $3D$ coordinates of the two points $[X_i\ Y_i\ Z_i\ W_i], [X_j\ Y_j\ Z_j\ W_j]$, and the image coordinates p_{it}, p_{jt} of the points in image t. This relation can be rewritten as

$$p_{it}^T\ G_{ij}\ p_{jt}, \quad G_{ij} = \begin{bmatrix} 0 & Z_jW_i - Z_iW_j & -Z_jY_i + Z_iY_j \\ -Z_jW_i + Z_iW_j & 0 & Z_jX_i - Z_iX_j \\ Z_jY_i - Z_iY_j & -Z_jX_i + Z_iX_j & 0 \end{bmatrix} \tag{5}$$

and G_{ij} is the **dual** "fundamental" matrix.

Similar to the case of the fundamental matrix F, the *dual* fundamental Matrix G is determined by the coordinates of the **dual** epipole p_{ij} [11], which is obtained by projecting the scene point P_i through P_j onto the reference plane. These dual relations are similar to those previously described in [3, 21, 4], where they were derived with respect to the quantities from the actual image points; here the dual relations are derived in the context of the reference-plane images.

2.5 Using affine normalization

When the homography H_t between the actual *observed* image $\mathbf{q_t}$ and the corresponding virtual image $\mathbf{p_t}$ on the reference plane is *completely* determined by the given scene calibration data, then for each point in the observed image¡ we can compute the corresponding reference-image point coordinates $p_{it} = H_t^{-1} q_{it}$. These quantities can then be used in the constraints derived above. However, if the homography is known only up to a $2D$ affine transformation G_t of Π, then we can use the affine normalized coordinates $\widehat{p_{it}} = G_t p_{it}$.

If we replace p_{it} by $\widehat{p}_{it}$ in all the derivations above, everything remains true but with respect to a different $3D$ coordinate system: the $3D$ point coordinates are now taken with respect to a $3D$ coordinate system where the $X-Y$ plane is transformed by the same affine transformation G_t. The coordinates of point P_i in this new coordinate system are $[X_i', Y_i', \alpha Z_i, W_i]$, where X_i', Y_i' are obtained from X_i, Y_i by G_t, and α is some scale factor. Thus when using image coordinates from $\widehat{\mathbf{p}}_t$, we can still use all the expressions developed in Section 2.2, noting to replace (X_i, Y_i) by (X_i', Y_i').

3 Stratified Reconstruction

In this section we show how the relations established in Section 2 between the image positions of scene points on the reference plane can be used to recover $3D$ information about the scene with very little scene information. We will describe an approach in which gradual increase in *domain* information for scene calibration leads to gradually increasing $3D$ information.

Registering the Input Images to the Reference Plane Π: As mentioned in Section 1 our approach is based on registering each input image to a pre-specified affine or Cartesian coordinate system on the reference plane. For practical reasons, we do this in three stages: (i) determine the homography H_{st} between each image "s" in the sequence and an arbitrarily selected reference image "t" from the same sequence, (ii) specify or infer the domain information needed to register the reference image t to the reference plane Π; based on this specification determine the $2D$ transformation $H_{t\pi}$ that aligns image t with Π, and (iii) concatenate H_{st} and $H_{t\pi}$ to determine the transformation $H_{s\pi}$ that registers s with Π. We refer to this process of registering the images to the reference plane as "registering the reference plane".

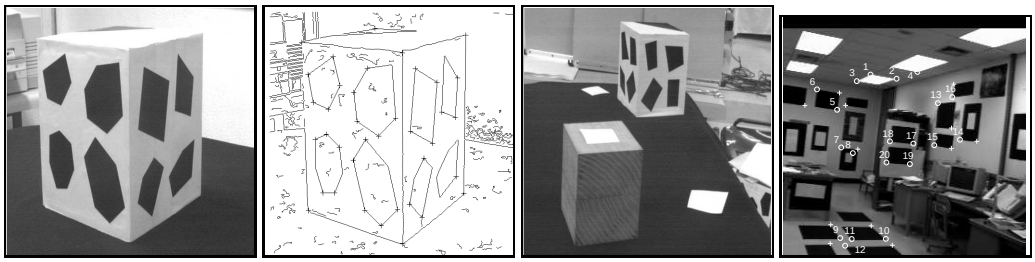
Affine vs. Euclidean Normalization: As noted in Section 2.5, if the positions of the image points on the reference plane Π are known only up to a $2D$ affine transformation of Π, it only affects the X and Y coordinates of the $3D$ scene points. The Z component is not affected by the unknown $2D$ affine transformation G_t. Here the minimal scene calibration required for some $3D$ inference should be sufficient for the registration of the reference plane up to an affine transformation.

$3D$ Reconstruction From the Dual Epipolar Geometry: (5) establishes the relationship between reference-plane image positions of two points in one view and the dual Fundamental matrix G, which in turn depends on the coordinates of the dual-epipole. Since this equation is homogeneous, it can be divided by an arbitrary scale factor. In particular, we can divide (5) by $Z_i Z_j$ and obtain a new form for G, which depends on the *scaled* homogeneous coordinates of the dual-epipole

$$p_{ij} \stackrel{\text{def}}{=} [\frac{X_j}{Z_j} - \frac{X_i}{Z_i}, \frac{Y_j}{Z_j} - \frac{Y_i}{Z_i}, \frac{W_j}{Z_j} - \frac{W_i}{Z_i}]. \tag{6}$$

(5) provides one constraint on these 3 variables. Looking separately at the pairs of $3D$ points $\{P_i, P_j\}$, $\{P_i, P_k\}$, $\{P_j, P_k\}$, we get from each image 3 constraints on the 9 homogeneous coordinates of the three dual-epipoles p_{ij}, p_{jk} and p_{ki}. Thus, 2 images $t = 1, 2$ give 6 homogeneous constraints on these 9 variables. In addition, it can be easily verified from (6) that $p_{ij} + p_{jk} + p_{ki} = 0$ which gives 3 more constraints. Taken together these 9 homogeneous constraints are sufficient to compute the coordinates of the three dual-epipoles up to a single scale factor.

Note that each additional image provides 3 more homogeneous constraints on the same 9 variables. Thus, if given more than 2 images, the computation of the dual epipoles requires the solution of an over-determined



a) b) c) d)

Figure 1: a) one frame from the medium depth sequence; b) corners extracted from a); c) one frame from the large depth sequence; d) one frame from the lab sequence.

linear system of equations; such system can be solved in a least-squares sense using standard tools such as SVD.

Of particular interest is the third coordinate of the dual-epipoles. Assuming, w.l.o.g, $W_i = 1$, the third coordinate of the three dual-epipoles are: $(\frac{1}{Z_i} - \frac{1}{Z_j}), (\frac{1}{Z_j} - \frac{1}{Z_k}), (\frac{1}{Z_k} - \frac{1}{Z_i})$. Since we can compute these quantities only up to a scale factor, we can only determine their ratios. We arbitrarily select two out-of-plane points as "reference points" and denote them by indices 1 and 2. Then for any other point i we can compute

$$u_i = (\frac{1}{Z_i} - \frac{1}{Z_1})/(\frac{1}{Z_2} - \frac{1}{Z_1}) \quad \Longrightarrow \quad Z_i = \frac{Z_1}{1 + \alpha u_i} \tag{7}$$

where $\alpha = \frac{Z_1}{Z_2} - 1$.

Stratified Reconstruction: (7) can be used to compute the height of points relative to the reference plane Π. With increasing amount of scene information we get increasing specificity:

Ordinal reconstruction: from (7) it follows that u_i is monotonically related to the height Z_i. Hence, given only the knowledge whether α is positive or negative, the ordering of the height of all other points relative to Π can be determined without *any additional information.* This type of $3D$ information is useful in a number of visual reasoning tasks such as navigation and grasping, in order to determine potential obstructions or provide micro-management control commands.

Affine reconstruction: if the ratio (Z_2/Z_1) is given, α can be determined, and the height of all other scene points relative to Π can be determined up to an unknown scale factor.

Absolute depth: if, in addition, the height of Z_1 is known, then the absolute height of all points relative to Π can be determined.

Euclidean reconstruction: the remaining elements of the dual epipoles can be used to compute the X and Y coordinates of the object points. Given the height of Z_1, Z_2, the X, Y coordinates can be determined up to image translation. Given (X_1, Y_1), (X_i, Y_i) can be determined absolutely.

4 Experiments

We compute the stratified reconstruction using four sequences of real images. In the first three cases, corners were automatically extracted (see Fig 1b) and then automatically tracked over the sequence; for comparison we were given all the $3D$ coordinates of the points. In the last two sequences, a reference plane in the scene was stabilized first, then a dense parallax flow field was recovered [11, 14]. The reconstruction was then based on this parallax data. In this case, the images were obtained using a hand-held video camera in a casual manner, without making controled measurements of the camera or scene parameters.

u_i est.	u_i act.
12.9521	12.6736*
8.5741	6.4453
7.7989	6.2027
2.5462	3.1829
2.4668	2.4414*
2.0456	2.3201
2.2506	1.9965
1.4493	1.9965
0.9668	0.9498*
0.9641	0.8550
0.6754	0.6091

u_i est.	u_i act.
0.5658	0.5729
0.3900	0.4712*
0.0666	0.3356
0.1853	0.2679
0.0527	0.2170
0.1875	0.1458*
0.0972	0.0096
0.0341	-0.0188
-0.1484	-0.1468
-0.0954	-0.1622*
-0.4950	-0.5036

u_i est.	u_i act.
-0.5205	-0.5166
-0.4302	-0.5292
-0.5136	-0.5456*
-0.1878	-0.5576
-0.5954	-0.5844
-0.6176	-0.6407
-0.6576	-0.6473*
-0.7028	-0.7083
-0.7537	-0.7083
-0.8160	-0.7083

Z_i est.	Z_i act.
0.8827	0.9
3.1798	3.2
5.0657	5.1
6.5621	6.3
7.3215	7.5
8.7334	9.15
12.2156	12.6
14.1592	14

Table 1: The estimated vs. actual reconstructed $3D$ data, ordinal (u_i) and Euclidean (Z_i), for the Medium Depth Sequence. For conciseness, the height Z_i is only shown for every fourth point, and the corresponding ordinal values are marked by *.

4.1 Medium depth sequence

An object (box) 15cm wide at about 60cm from the camera (Fig 1a) was photographed from 5 different points of view. First, we affine-normalized all the images: using two identified pairs of parallel lines (e.g., the sides on one of the faces of the box), we applied a $2D$ homography which made the two lines parallel in the image and confined to the same image position in all the frames. We then arbitrarily selected two out-of-plane points as reference points, and computed the dual-epipole between them, as well as the dual-epipole between each of the other data points and the two reference points. The dual-epipoles were computed by using all the 5 frames and a least-squares solution to (5). Using these dual-epipoles, we computed ordinal reconstruction as described in Section 3. The results - each computed u_i value vs. the real u_i value - are shown in Table 1.

Given the heights of the two reference points, we can continue further and transform the ordinal reconstruction into Euclidean reconstruction by computing the actual height Z_i at each point. The results - some of the computed Z_i vs. the actual Z_i - are shown in Table 1.

Looking closely at these results, clearly the ordinal reconstruction by u_i is accurate at most of the points, with only a few points close in height switched around. The heights are also quite accurate for most points, as the few representative heights in Table 1 show.

4.2 Large depth sequence

Objects spanning 60cm at about 60cm from the camera (Fig 1c) were photographed from 5 different points of view. Unlike the previous sequence, there is much larger depth of field in this sequence, and thus there are larger projective distortions. Such distortions make metric reconstruction difficult. Once again, we first affine-normalized all the images and followed a procedure similar to the previous experiment to recover ordinal height (u_i) for 24 tracked points. The estimated vs. the actual $3D$ data are shown in Table 2.

Given the heights of the 2 reference points, we can continue further and transform the ordinal reconstruction into Euclidean reconstruction by computing the actual height Z_i at each point. The results - the computed Z_i vs. the actual Z_i - are shown in Table 2.

Here, too, the ordinal reconstruction by u_i is almost always accurate. Of the 30 features, only 3 pairs of neighbors in height were swapped with each other (their height was 4 vs. 3.5, 11 vs. 12 and 10.2 vs. 10.5). The Z values are still good, but less accurate when compared with the previous experiment.

4.3 Lab sequence

This sequence includes 16 images of a robotic laboratory, obtained by rotating a robot arm 120^o (one frame is shown in Fig. 1d). 32 corner-like points were tracked. This sequence has the largest depth of field - the

u_i est.	u_i act.
0.0077	0*
-0.1529	-0.3068
-0.167	-0.3068
-2.4833	-3.079
-32.2960	∞*
15.8281	∞
10.8487	8.2697
8.4828	6.3750

u_i est.	u_i act.
3.8829	3.7841*
2.4482	2.5179
2.4848	2.402
2.6841	2.25
2.1878	2.0795*
1.7263	1.6045
1.6723	1.6045
1.4574	1.4024

u_i est.	u_i act.
1.4512	1.3885*
1.3658	1.3817
1.1322	1.1103
1.1175	1.1103
1.2088	1.0893*
1.072	1.0568
1.0901	1
1.089	1

Z_i est.	Z_i act.
-20.4185	-20
-0.2296	0
2.138	2.2
4.1373	4.4
6.9691	7.4
8.9946	10.5

Table 2: Estimated vs. Actual $3D$ Data, ordinal (u_i) and Euclidean (Z_i), for the Large Depth Sequence. For conciseness, the height Z_i is only shown for every fourth point, and the corresponding ordinal values are marked by *.

Z_i est.	Z_i act.
0.2045	-0.3386
-0.2048	0.3409
2.2095	1.1965
8.5974	1.5249
8.1943	2.8690

Z_i est.	Z_i act.
9.4100	5.6942
7.1877	6.4347
9.4263	9.3823
9.1724	9.4161
9.9265	9.4161

Z_i est.	Z_i act.
8.3978	10.0444
9.0416	10.0446
9.1713	10.9708
9.5846	10.9993
9.9490	11.2723

Z_i est.	Z_i act.
9.2478	11.4074
9.3495	12.1890
9.2442	13.6701
13.2664	13.9289
16.5506	14.6616

Z_i est.	Z_i act.
15.3879	15.8765
16.1396	16.5352
17.4406	16.6093
13.6005	16.6761
16.5688	17.5328

Table 3: Estimated vs. Actual Heights for the Lab Sequence.

depth values of the points in the first frame (relative to the camera) ranged from 13 to 33 feet. Moreover, a wide-lens camera was used, causing distortions at the periphery which were not compensated for. This is therefore the most difficult sequence so far. Once again, following the same procedure as in the other two experiments, we estimated ordinal and Euclidean reconstruction in 25 points. The estimated Z_i vs. actual Z_i for these points are shown in Table 3.

Note that although there are gross errors for some points, the computed height in most points is fairly close to their true height.

4.4 Dense Height Maps

The previous three examples gave quantitative illustrations of stratified reconstruction at a sparse set of points in the scene. Here we include two examples of dense reconstruction of ordinal heights. In both cases, the images were acquired using a hand-held video camera. No quantitative information about the scene structure, the camera imaging parameters, or its motion were available.

Given two input images, we used the method described in [20] to first estimate the homography that aligns a dominant planar surface in the scene between the two images. The residual parallax displacements between the points were then computed using the method described in [14]. These displacements were used as input for the stratified reconstruction of the scene.

The first example uses the "Toys" image sequence previously used in [12], see Figure 2a. The scene consists of a few toys standing on a rug.

In order to register the images to the reference plane (*upto 2D affine transformation*), we used the following approach. Using commonly available image manipulations tools on a PC, we drew two pairs of lines on the reference image, that were visually judged to be parallel to the "grooves" of the rug. These are shown as grey lines painted on the carpet in Figure 2a. Note that the vertical pair in particular is not parallel in the image itself, although it represents parallel lines on the rug. We then interactively warped the images (using homographies) until the lines appeared parallel in the image. Any of a family of homographies that are equivalent upto a 2D affine transformation is sufficient for this purpose. We arbitrarily picked one to achieve the intended effect. The result of this process is shown in Figure 2b.

We then computed the ordinal height u_i (see (7)) using the parallax displacements (also appropriately

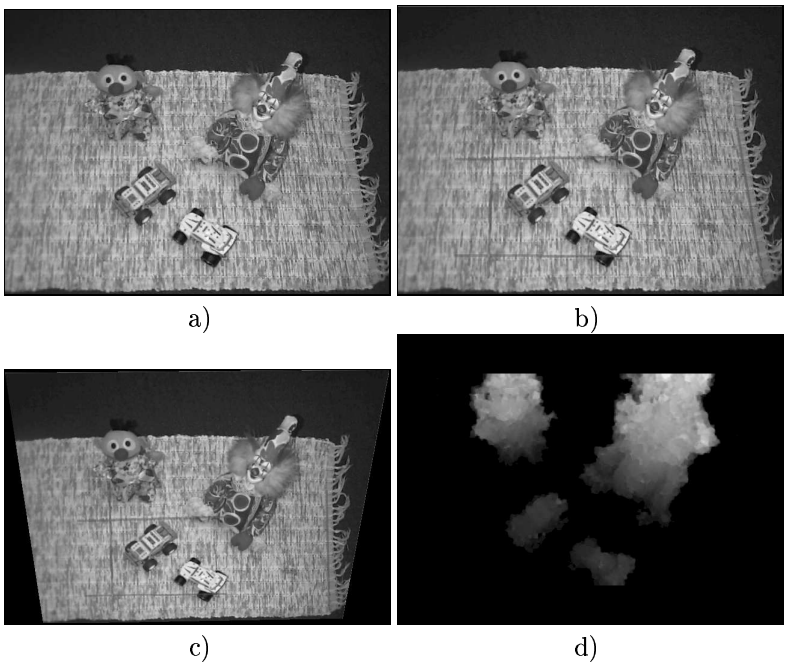


Figure 2: a) one frame from the toys sequence; b) same image as (a) but with lines drawn on it (see text); c) the result of projecting (b) onto the reference plane (the rug); d) ordinal height.

warped to reflect the projection onto the reference plane). However, since $u_i = \infty$ for points on the reference plane itself (i.e., when $Z_i = 0$ in (7)), we display $\frac{1}{u_i}$ in Figure 2c. As evident from this image, the points on the rug are dark (corresponding to an ordinal height of 0); also, a gradual increase in height along the two dominant objects is noticeable. Since we did not have the actual heights of any of the objects in the scene, we stopped at this stage of our stratified reconstruction.

The second example uses a new sequence, which we will refer to as the "Doll" sequence (see Figure 3a). In this case, the carpet and the floor constituted the reference plane. The grid lines on the carpet and the floor served as the basis for registering the images to the reference plane. As in the case of the "Toys" example, we interactively warped the images until these lines appeared parallel in the image, and arbitrarily picked one homography that achieved this effect. Once again, we display $\frac{1}{u_i}$ in Figure 3c.

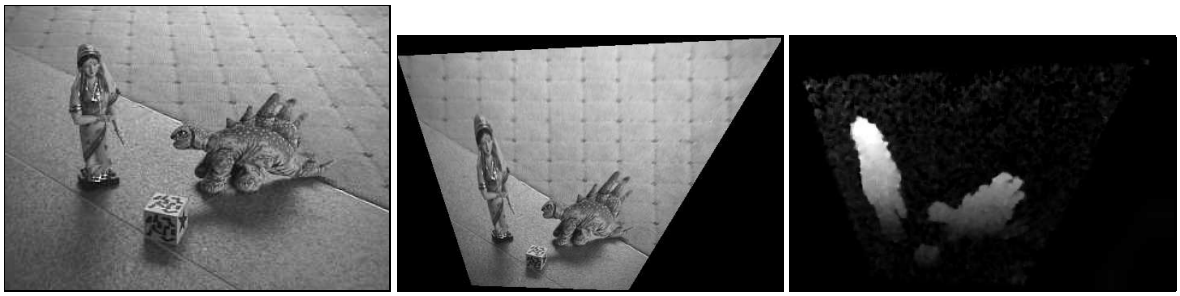
Note that in both these examples, even with the minimal calibration (two sets of parallel lines) we obtain a reconstruction that looks qualitatively consistent with the actual structure of the scene. As noted earlier, this would be useful in a number of visual reasoning tasks, such as navigation and grasping.

4.5 Discussion

The computation of height (a *metric* quantity) relative to the reference plane consistently gave good results while using noisy sequences with large perspective distortions. The algorithm is fast, robust and easy to implement since it only solves a linear system of equations. Height was computed with minimal scene calibration: (i) **ordinal reconstruction** required knowledge of the reference frame up to affine transformation - 2 degrees of freedom[2], and (ii) **exact height** required knowledge of the height of two points not on the reference plane - 2 additional d.o.f. Thus we used 2-4 d.o.f. of scene calibration information, much less than required by other reconstruction algorithms which use scene calibration.

For example, the scene calibration needed by the reconstruction algorithm described in [15] includes the specification of the $3D$ coordinates of 5 reference points (supplied by an oracle) - 15 d.o.f. We used this method in [2] to accomplish reconstruction with the first two sequences described above. Although this computation relied on 15, rather than 4, pre-determined pieces of calibration data, and involved a complex

[2]The $2D$ affine transformation G_t has 6 degrees of freedom, whereas a general $2D$ projective transformation (homography) has 8 d.o.f.; thus $2D$ affine plane calibration requires the specification of 2 d.o.f.



a) b) c)

Figure 3: a) one frame from the doll sequence; b) the result of projecting (a) onto the reference plane (the rug); c) ordinal height.

non-linear algorithm, the reconstruction results in [2] are no better than our results here (e.g., a relative error of about $5\% - 10\%$ per datapoint using the second sequence).

5 Summary

Since uncalibrated images only permit projective reconstruction, no metric information (such as relative depth) can be deduced without some calibration, either camera calibration (external and internal parameters of the camera) or scene calibration (the $3D$ affine or Euclidean coordinates of some $3D$ landmarks). Camera calibration, however, is not always possible: it requires a partially fixed camera (unsuitable, e.g., if the images are taken by many cameras) or some control over the camera motion (unsuitable, e.g., with video data). Scene calibration requires a priori knowledge of known $3D$ points, and is typically employed after the projective reconstruction; therefore it is typically ill-advised to use directly a least squares linear reconstruction algorithm as we do here, since there is no suitable least squares error in $3D$ projective space.

Our contribution in this paper is three fold: (1) We use partial scene calibration, as little as two parallel lines on a plane; this information need not be given a priori, and can be inferred from the images directly. (2) We perform the calibration prior to the reconstruction; this allows the use of a robust least squares structure computation from many frames. (3) We obtain a hierarchy of intermediate representations, from ordinal to Euclidean, which increasingly depend on the amount of scene calibration available.

References

[1] M. Irani, P. Anandan, and D. Weinshall. From Reference Frames to Reference Planes: Multi-view Parallax Geometry and Applications In *Proc. 5th ECCV*, Freiburg, Germany, June 1998.

[2] B. Boufama, D. Weinshall, and M. Werman. Shape from motion algorithms: a comparative analysis of scaled orthography and perspective. In *Proc. 3rd ECCV*, pages 199–204, Stockholm, Sweden, 1994.

[3] S. Carlsson. Duality of reconstruction and positioning from projective views. In *Workshop on Representations of Visual Scenes*, 1995.

[4] S. Carlsson and D. Weinshall. Dual Computation of Projective Shape and Camera Positions from Multiple Images. *IJCV*, 27(3), 1998.

[5] A. Criminisi, I. Reid, and A. Zisserman. Duality, rigidity, and planar parallax. In *Proc. 5th ECCV*, Freiburg, Germany, June 1998.

[6] O.D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *Proc. 1st ECCV*, pages 563–578, Santa Margarita Ligure, May 1992.

[7] O.D. Faugeras. Stratification of three-dimensional vision:projective, affine and metric representations. JOSA, 12(3):465-484, March 1995.

[8] O. Faugeras and B. Mourrain. On the geometry and algebra of the point and line correspondences between N images. In *Proc. ECW*. Xidian University Press, 1995.

[9] C. Fermüller and Y. Aloimonos. Ordinal Representations of Visual Space. In *Proc. DARPA Image Understand Workshop*. 1996.

[10] Richard Hartley. Euclidean Reconstruction from Uncalibrated Views. In *Applications of Invariance in Computer Vision*, J.L. Mundy, D. Forsyth, and A. Zisserman (Eds.), Springer-Verlag, 1993.

[11] M. Irani and P. Anandan. Parallax geometry of pairs of points for 3D scene analysis. In *Proc. 3rd ECCV*, Cambridge, UK, April 1996.

[12] M. Irani and P. Anandan. A unified approach to moving object detection in 2D and 3D scenes. *IEEE Trans. on PAMI*, in press.

[13] J. J. Koenderink and A. J. van Doorn. Affine structure from motion. *JOSA*, 8(2):377–385, 1991.

[14] R. Kumar, P. Anandan, and K. Hanna. Direct recovery of shape from multiple views: a parallax based approach. In *Proc 12th ICPR*, 1994.

[15] R. Mohr, , L. Quan, F. Veillon, and B. Boufama. Relative 3D reconstruction using multiple uncalibrated images. RT 84–IMAG–12 LIFIA, Uni. of Grenoble, 1992.

[16] H. S. Sawhney, J. Oliensis, and A. R. Hanson. Description and reconstruction from image trajectories of rotational motion. In *Proc. 3rd ICCV*, Osaka, Japan, 1990.

[17] L. Robert and O.D. Faugeras. Relative 3D Positioning and 3D Convex Hull Computation from a Weakly Calibrated Stereo Pair. J. Imaging and Vision Compting, 13(3), 1995.

[18] M. Pollyfeys, R. Koch, and L. van Gool. Self-calibration and Metric Reconstruction in Spite of Varying and Unknown Internal Camera Parameters. In *Proc. 6th ICCV*, Mumbai, India, January 1998.

[19] A. Shashua and N. Navab. Relative affine structure: Theory and application to 3D reconstruction from perspective views. In *Proc. CVPR*:483–489, Seattle, 1994.

[20] R. Szeliski and H. Shum. Creating full view panoramic mosaics and texture-mapped models. In *Proc. SIGGRAPH 97*, pp. 251-258, Los Angels, 1997.

[21] D. Weinshall, M.Werman, and A. Shashua. Shape Tensors for Efficient and Learnable Indexing. In *Workshop on Representations of Visual Scenes*, 1995.

[22] D. Weinshall. Qualitative depth from stereo, with applications. *Computer Vision, Graphics, and Image Processing*:49(222-241), 1990.

[23] A. Zisserman. Active Visual Navigation using Non-Metric Structure. In *Proc. 5th ICCV*, Boston, USA, 1995.

[24] Z. Zhang, R. Weiss and A. Hanson. Obstacle Detection Using Qualitative and Quantitative 3D Reconstruction. In *IEEE PAMI*, 19(2):15-26, January 1997.