Uncalibrated Two-View Metrology

Bojian Liang[1], Zezhi Chen[1,2] and Nick Pears[1],
[1]Department of Computer Science, University of York, YO10 5DD, UK
[2]ISN National Key Lab., Xidian University, Xi'an 710071, P.R.China
{bojian, chen, nep}@cs.york.ac.uk

Abstract

A method of visual metrology from uncalibrated cameras is proposed in this paper, whereby a camera, which captures two images separated by a (near) pure translation, becomes a height measurement device. A novel projective construction allows accurate affine height measurements to be made relative to a reference plane, given that the reference plane planar homography between the two views can be accurately recovered. To this end a planar homography estimation method is presented, which is highly accurate and robust and based on a novel reciprocal-polar (RP) image rectification. The absolute height of any pixel or feature above the reference plane can be obtained from this affine height once the camera's distance to the reference plane, or the height of a second measurement in the image is specified. Results from our data show a mean absolute error of 6.9mm and with two outliers removed this falls to 1.5mm.

1. Introduction

There are several different methods to implement Euclidean metrology in the literature. For example, 3D world structure can be computed from uncalibrated views of a scene given sufficient correspondences in general position and this has already been used to answer specific, metric questions about the scene. The approach by Tomasi and Kanade [5] is known as the factorization method; Triggs [6] extends the factorization method to the projective camera model by using epipolar constraints to calculate depth scale factors; Heyden et al. [3] upgrades the affine approximations to projective results by iterative optimization. Criminisi et al. [1] proposed methods to make measurements of world planes from their (single) perspective images. Reid and Zisserman [4] give a method for locating 3D position of a soccer ball from monocular image sequences.

Our goal is to be able to measure height from two uncalibrated images separated by a (near) pure translation. The basic idea is to use the plane-and-parallax cue computed via a projective cross-ratio construct in conjunction with a planar homography relation which encodes the motion of a reference plane. Seven salient aspects of this method are as follows: (i) A new projective construction to compute affine height via the plane-and-parallax cue. (ii) A new reciprocal-polar (RP) image rectification renders all coplanar, co-radial image motion to a pure shift, which allows correlation based image motion recovery even over large perspective image distortions induced by large camera motions. (iii) Since the image motion recovery is correlation based, no corner feature correspondences are required on the reference plane, just local intensity variation. (iv) Again, since the method is correlation based, all coplanar pixels with local intensity variation contribute to the homography estimation, not just a potentially low density of corner feature matches. (v) We show that the magnitude of image motion in the $\frac{1}{r}$ dimension of the rectified image pair follows a sinusoidal form along the θ dimension over a maximum of π radians, for the four DOF class of planar homographies called elations. (vi) RANSAC/LS estimation of the phase of the sinusoid yields a highly accurate vanishing line orientation of the reference plane (and simultaneously a segmentation of the reference plane) and this, along with the focus of expansion (FOE), allows an accurate reference plane homography relation to be obtained. (vii) The limiting aspect of the method is that it is only applicable to (near) pure translation. However, in many application, such a mobile robot navigation, deliberative (near) translation motions can be executed to probe the environment in terms of heights, and measured heights of near zero are obstacle free navigable zones. Other applications include measurement of interior scenes for interior design purposes, and outdoor architectural measurements.

2. Planar Homography Recovery

Given two views of a plane separated by a pure translation, there are two relations between the two views: first, through epipolar geometry defined by F and second, through the homography induced by the plane. We note that the FOE, v, is the epipole $\mathbf{e}$ and also $\mathbf{e}'$ so that the fun-

damental matrix is given by, $\mathbf{F} = [\mathbf{e}']_\times = [\mathbf{e}]_\times = [\mathbf{v}]_\times$, where $[\bullet]_\times$ is a skew-symmetric matrix corresponding to the vector. The relationship between the planar homography (**H**) and the epipolar geometry (expressed by the dual epipole or FOE, $\mathbf{v} = (x_v, y_v, 1)^T$) is

$$\mathbf{H} = \mathbf{I} - k\mathbf{v}\mathbf{l}^T. \quad (1)$$

Where **l** is the vanishing line $(a_v, b_v, 1)^T$ corresponding to the plane and *k* is a constant scalar. In order to recover the homography based on the form in equation 1, we check for (near) pure translation between two views, by intersecting all lines defined by all corner correspondences from the image pair (not necessarily reference plane corners). If most lie in a small area (for example, 90% of intersection should lie within a 50 pixels radius), then pure translation is assumed and the FOE is computed as follows:

(1). Extract interest point correspondences: Compute interest points in each images by using Harris detector, KLT algorithm or SUSAN method and compute a set of interest point matches based on proximity and similarity of their intensity neighbourhood.

(2). RANSAC robust estimation: Repeat the following process for m samples, where m is determined adaptively by using binning technique[8]:

(a). Select a random sample of at least 2 correspondences and compute the FOE by using the simultaneous equations $(\mathbf{x}_i \times \mathbf{x}'_i) \bullet \mathbf{v} = 0$, where $\mathbf{x}_i \longleftrightarrow \mathbf{x}'_i$ $(i = 1, 2, \cdots, n)$ is any pair of matching points in two images.

(b). Calculate the epipolar distance $f_m(\mathbf{v})$ for each putative correspondence. $f_m(\mathbf{v}) = \lambda_i \left|\mathbf{x}_i'^T \mathbf{F} \mathbf{x}_i\right|$, where **F** is the fundamental matrix and $\lambda_i = \left(\frac{1}{\sqrt{(\mathbf{F}\mathbf{x}_i)_1^2 + (\mathbf{F}\mathbf{x}_i)_2^2}} + \frac{1}{\sqrt{(\mathbf{F}^T\mathbf{x}'_i)_1^2 + (\mathbf{F}^T\mathbf{x}'_i)_2^2}}\right)$

(c). Compute the number of inliers consistent with **v** by the number of correspondences for which $f_m(\mathbf{v})$ <*threshold*. Choose the FOE with the largest number of inliers.

(3).Optimal estimation: re-estimate the FOE from all correspondences classified as inliers, by minimizing the objective function $f_m(\mathbf{v})$. Repeat steps (2)-(3) until the number of correspondences are stable.

Once the FOE has been computed, we shift image coordinates using transformation $\mathbf{T}_c$, so that each image is centred on the FOE and the homography expressing the FOE centered image motion is:

$$\mathbf{H}' = \mathbf{I} - k\mathbf{v}'\mathbf{l}'^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ s & \mu & q \end{bmatrix}. \quad (2)$$

where $q = 1 - k\mathbf{v}^T\mathbf{l}$, $s = -ka_v$ and $\mu = -kb_v$. If the translation direction is parallel to the ground, such as would be obtained for mobile robot motion, $q = 1$, since the FOE lies on the vanishing line for this special motion. Otherwise, the FOE is at a distance $d = \left|(1-q)/(k\sqrt{a_v^2 + b_v^2})\right|$ from the vanishing line. If we now define r, θ as the polar coordinates of a pixel in the FOE centered frame and we let $\rho = \frac{1}{r}$, some algebraic manipulation of the homogenous coordinate relation $\lambda\mathbf{x}'_2 = \mathbf{H}'\mathbf{x}'_1$ yields the key equation:

$$f(\theta) = \rho_2 - q\rho_1 = s\frac{x'_1}{r_1} + \mu\frac{y'_1}{r_1} = k_{s\mu}\sin(\theta + \alpha). \quad (3)$$

where $k_{s\mu} = \sqrt{s^2 + \mu^2}, \sin\alpha = \frac{s}{k_{s\mu}}, \cos\alpha = \frac{\mu}{k_{s\mu}}, \tan\alpha = \frac{a_v}{b_v}$.. Eq.(3) indicates that the magnitude of radial image motion in (ρ, θ) or RP image space is sinusoidal with respect to that motion's orientation, under the FOE centered homography. The relation indicates that we need to find three constants $(q, k_{s\mu}, \alpha)$ in order to recover the homography and that the computation should be implemented in RP image space, in order to use correlation irrespective of the perspective distortion which occurs over large translations. Here, we will assume that $q = 1$, i.e the camera motion is parallel to the reference plane and the homography is the special case of a 4 DOF elation (it is, however, straightforward to estimate q in the case of more general 5 DOF homologies). For each of the original image pair, an RP image is generated, an example of a single image is shown in fig 1. Then, for each pixel in regular (Cartesian) image space, we find its motion by correlation between the two RP image along the ρ direction (i.e. fixed θ). The position of the maximum value of the correlation is related as a value of $f_i(\theta)$. For different values of θ, we know that a point's motion traces out a sinusoid, irrespective of that point's distance from the FOE (r or ρ value). Thus we need to find the phase and amplitude of that sinusoid and this is done by RANSAC [2] followed by cycles of least squares on the inliers until the inliers are stable.

Suppose that we have two value of $f_{i,j}$ measured at two angles, $\theta_{i,j}$, then

$$\tan\alpha = \frac{f_j\sin\theta_i - f_i\sin\theta_j}{f_i\cos\theta_j - f_j\cos\theta_i}. \quad (4)$$

When we have the phase angle, α, which corresponds to the orientation of the vanishing line, we can compute $k_{s\mu}$ from Eq.(3). Thus a pair of $(\theta, f(\theta))$ measurements gives us a putative sinusoid in the RANSAC process in which the number of inliers is counted. Once the RANSAC process has terminated, we examine the dominant (high consensus) sinusoids to find a phase (vanishing line orientation), which is close to zero. Thus, with some weak assumptions regarding camera orientation with respect to the reference plane, we can reliably detect and segment pixels which belong to

the reference plane. Finally, we refine the sinusoid parameters from the RANSAC process and hence the homography $\mathbf{H}'$, using standard LS on the inliers. Using the model $\mathbf{AZ} = \mathbf{b}$, where

$$\mathbf{A} = \begin{bmatrix} \cos\theta_1 & \sin\theta_1 \\ \vdots & \vdots \\ \cos\theta_n & \sin\theta_n \end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix} s \\ \mu \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix}$$

allows us to compute the sinusoid related parameters (s, μ) and hence homography $\mathbf{H}'$ directly using a standard pseudo-inverse: $\mathbf{Z} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$. The homography expressed in the original (non FOE centered) frame can be explicitly expressed as $\mathbf{H} = \mathbf{T}_c^{-1}\mathbf{H}'\mathbf{T}_c$.

3. Affine Height Measurements

The approach described above allows pixels to be classified as either belonging to the reference plane or not. Those that do are within a threshold of a recovered sinusoidal model. For those non ground plane regions, we would like to know their height above the reference plane. Our aim is to compute the relative height, h_r, of corner point **A** in fig. 3, as a fraction of the height, h_c, of the camera optical centre **O** above the reference (ground) plane, when the camera undergoes pure translation **t** and the motion direction is parallel to the reference plane. Point **A** is the actual position of the corner point relative to the camera before the translation and the **C** is the position of the corner after the translation. Points $\mathbf{A}'$ and $\mathbf{C}'$ are the projections of these actual corner positions onto the ground plane. Points **a** and **c** are the image positions of the corner at positions **A** and **C** respectively, and **b** is the predicted image position of the corner point by using H matrix of ground plane, if the corner point were to lie in the ground plane. Image point **b** is computed as $\mathbf{b} = \mathbf{Ha}$, using the recovered homography.

Figure 3 shows that point **a** lies below the vanishing line. Because $\mathbf{AC}//\mathbf{A}'\mathbf{C}'$, using similar triangles, and denoting the distance between point x and y as $d(x, y)$, we can get

$$h_r = \frac{h}{h_c} = 1 - \frac{d(\mathbf{O}, \mathbf{C})}{d(\mathbf{O}, \mathbf{C}')} = 1 - \frac{d(\mathbf{A}, \mathbf{C})}{d(\mathbf{A}', \mathbf{C}')}. \quad (5)$$

For pure translation, $d(\mathbf{A}, \mathbf{C}) = d(\mathbf{A}', \mathbf{B}')$, so that

$$h_r = 1 - \frac{d(\mathbf{A}', \mathbf{B}')}{d(\mathbf{A}', \mathbf{C}')}. \quad (6)$$

Now, the four image points $(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{v})$, where $\mathbf{v}$ is the FOE, and their corresponding four ground plane points $(\mathbf{A}', \mathbf{B}', \mathbf{C}', \infty)$ are collinear. The cross ratio for this set of points remains invariant under perspective projection transform and so we can get the height of the corner point relative to the height of the optical centre is

$$h_r = 1 - \frac{d(\mathbf{a}, \mathbf{b})d(\mathbf{c}, \mathbf{v})}{d(\mathbf{a}, \mathbf{c})d(\mathbf{b}, \mathbf{v})}. \quad (7)$$

(a) (b)

Figure 1. (a) Original image, $I(x, y)$ (b) corresponding RP image, $I(\rho, \theta)$

If the point a lies above the vanishing line we have

$$h_r = 1 + \frac{d(\mathbf{a}, \mathbf{b})d(\mathbf{c}, \mathbf{v})}{d(\mathbf{a}, \mathbf{c})d(\mathbf{b}, \mathbf{v})}. \quad (8)$$

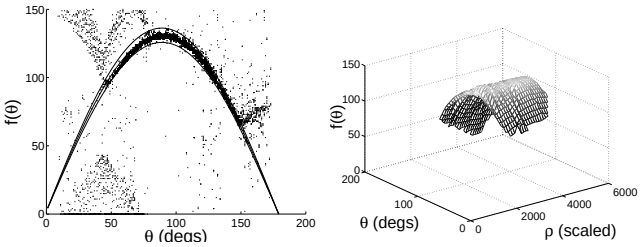
If points **a** coincide lie on the vanishing line, then $h_r = 1$.

Note that h_r can be interpreted as the height of point **A** in units of height h_c. The absolute distance can be obtained from this distance ratio once the camera's height h_c is specified. However, it is usually more practice to determine the distance via a second measurement in the image, that of a known reference length. Note that this approach only needs the H-matrix of ground plane, and the tracked image correspondences **a** and **c** of the feature to determine the height above the ground plane. The main advantageous compare with other method (such as the method of Criminisi 2000[1] and Wilczkowiak 2001[7] is that this method without need camera calibration and any geometry constraints of a scene, and it can be used to compute the height from any isolated point to the reference plane.

4. Experimental results

A large amount of synthetic data and real images were selected and intensive experimental work was carried out in order to test the robustness and the accuracy of the method. Here we present results that validate the method as a viable approach. Fig. 1 (a) shows a sample image and its RP rectification is shown in fig. 1 (b). Fig. 2 (a) shows the RP image motion of all image points, and clearly shows the sinusoidal form of the reference plane pixels as well as partial sinusoids of other planes in the scene. Fig. 2 (b) shows the reference plane image motion extracted by RANSAC. The third dimension in the plot clearly illustrates that the sinusoidal form is close to being constant irrespective of the radial distance of an image point from the FOE.

In the two visual metrology experiments (one indoor, one outdoor) presented here, $q = 1$ was assumed, since two VGA frames (640x480 resolution) were captured with the translation direction parallel to the ground plane, and in



(a) (b)

Figure 2. (a) all image motion in RP space and (b) extracted reference plane motion in RP space

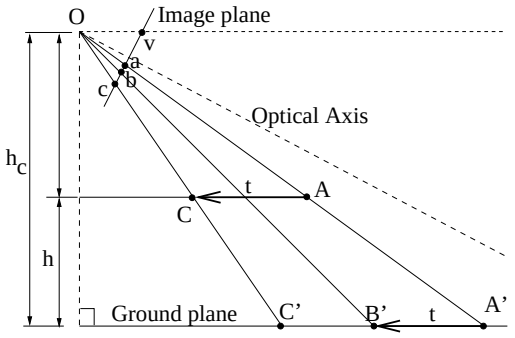


Figure 3. Measuring the height of point A.

both cases the height A-B was used as the reference height. For the indoor experiment, we computed $\mu = -5.6e-4$, $tan\alpha = 0$, and for the outdoor experiment we computed $\mu = -2.94e-4$, $tan\alpha = 0.027$. Results are shown in table 1, where 'TM' are the manual (tape measure) measurements and 'VM' are the visual metrology results. We find a mean absolute error of 6.9mm and mean relative error of 0.35%. If we remove the two rather inaccurate measurements (a)EF and (a)PQ, the remaining measurements have a mean absolute error of 1.5mm and a 0.1% mean relative error.

Table 1. Visual Metrology results in centimetres

Seg.	TM	VM	Seg.	TM	VM
(a)CD	30.0	29.88	(b)CD	233.1	233.08
(a)EF	227.7	229.36	(b)EF	149.8	149.51
(a)LM	208.4	208.23	(b)GH	258.7	258.55
(a)GH	252.5	252.73	(b)MN	233.1	232.67
(a)NO	121.1	120.94	(b)OP	149.8	149.82
(a)PQ	210.3	214.91	(b)QR	none	651.51

5. Conclusions

The main contributions in this work are (in the order of algorithm execution) (1) Robust FOE estimation, (2) RP rectification, (3) planar image motion estimation, planar homography estimation and plane segmentation by robust estimation of a sinusoid in RP image motion space, and (4) a projective construction allowing affine height above a plane to be measured from an uncalibrated image pair. Intensive experimental work was carried out in order to test the accuracy of the method proposed in this paper. The results show that our algorithm performs very well to outliers and noise and provides a practical method of visual metrology.

(a) (b)

Figure 4. (a) Indoor metrology, (b) Outdoor metrology

References

[1] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *Int. Journ. of Computer Vision*, 40(2):123–148, 2000.

[2] M. A. Fischer and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. Assoc. Comp. Mach*, 24(6):381–395, 1981.

[3] A. Heyden and R. Berthilsson. An iterative factorization method for projective structure and motion from image sequences. *Image and Vision Computing*, 13(17):981–991, 1999.

[4] I. Reid and A. Zisserman. Goal-directed video metrology. *Proc. ECCV*, 2:647–658, April 1996.

[5] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: A factorisation approach. *Int. Journ. of Computer Vision*, 9(2):137–154, 1992.

[6] B. Triggs. Factorization methods for projective structure and motion. *Int. Conf. Computer Vision and Pattern Recognition*, pages 845–851, San Francisco, California, USA. 1996.

[7] M. Wilczkowiak, E. Boyer, and P. F. Sturm. Camera calibration and 3d reconstruction from single images using parallelepipeds. *ICCV*, pages 142–148, July 2001.

[8] Z. Zhang. Determining the epipolar geometry and its uncertainty: a review. *Int. Journ. of Computer Vision*, 27(2):161–195, 1998.