METROLOGY IN UNCALIBRATED IMAGES GIVEN ONE VANISHING POINT

Hassan Foroosh Xiaochun Cao Murat Balci

School of Computer Science, University of Central Florida, Orlando, FL, 32816-3262
{foroosh,xccao,balci}@cs.ucf.edu

ABSTRACT

In this paper, we describe how 3D Euclidean measurements can be made in a pair of uncalibrated images, when only minimal geometric information are available in the image planes. This minimal information consists of a line in a reference plane, and the vanishing point orthogonal to it. Given such limited information, we show that the length ratio of two objects perpendicular to the reference plane can be expressed as a function of the camera intrinsic parameters. Assuming that the camera intrinsic parameters remain invariant between two views, we perform Euclidean metric measurements directly in the perspective images.

1. INTRODUCTION

Metrology from uncalibrated images is becoming of increasing interest for many applications, such as image-based rendering, surveillance, and forensics. This problem is straightforward if the camera is calibrated. Camera calibration can be performed by using standard methods if measurements of enough 3D points of a calibration object are available [4], or alternatively using self-calibration methods from unstructured scenes [8, 5]. However, such measurements are not always available, and also self-calibration techniques typically require bootstraping from a projective reconstruction, often leading to solving ill-conditioned non-linear problems.

Vanishing points or parallel lines have proven to be useful features for this task [1, 6, 2, 3, 7, 9]. In a seminal work, Criminisi et al. [3] proposed an approach for single view metrology, and showed that affine scene structure can be recovered from a single uncalibrated image. The limitation of their aproach is that they require three mutually orthogonal vanishing points to be available simultaneously in the image plane. Also, in order to recover the metric measurements they require three reference distances. Their advantage, however, is that they need only one image to solve the problem. In contrast, our approach requires only one vanishing point along a vertical to a reference plane. However, we require two images to solve the problem with only one reference distance. Examples of images where such scenario may apply are commonly encountered in indoor and outdoor environments, where there is a ground plane and some up-right objects, e.g. humans, street lamps, trees, etc. Furthermore, in our approach, one can directly perform measurements outside the reference plane and along non-parallel lines.

In the remaining of this paper, we denote the homogeneous image coordinates by a tilde. For instance, if $\mathbf{m}$ is the Euclidean representation of a point in the image, then $\tilde{\mathbf{m}} \sim [\mathbf{m}^T \;\; 1]^T$, where T denotes the transpose.

2. PROBLEM FORMULATION

The basic geometry of our configuration is shown in Figure 1, which consists of one line in a reference plane, and one vanishing point in the direction perpendicular to the plane. Two pairs of 3D points are observed, i.e. $(\mathbf{T}_1, \mathbf{B}_1)$ and $(\mathbf{T}_2, \mathbf{B}_2)$. With the world coordinate frame as depicted in Figure 1, these four points are all (without loss of generality) in the plane $Z = 0$. The corresponding image points are given by $\tilde{\mathbf{t}}_i = \mathbf{P}\mathbf{T}_i$ and $\tilde{\mathbf{b}}_i = \mathbf{P}\mathbf{B}_i$, where $\mathbf{P}$ is the camera projection matrix given by

$$\mathbf{P} = \mathbf{K}[\mathbf{r}_1 \;\; \mathbf{r}_2 \;\; \mathbf{r}_3 \;\; \mathbf{t}] \tag{1}$$

$\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3$ are the columns of the 3×3 rotation matrix $\mathbf{R}$, $\mathbf{t}$ is the translation vector, and $\mathbf{K}$ is the 3×3 upper triangular matrix of camera intrinsic parameters. Assuming a unit aspect ratio and a zero skew we get

$$\mathbf{K} = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where f is the focal length, and $\mathbf{c} = [u_0 \;\; v_0]^T$ is the coordinates of the principal point. In our case, since the two pars of points are all in the plane $z = 0$, the camera projection matrix reduces to the 3×3 homography given by $\mathbf{H} = \mathbf{K}[\mathbf{r}_1 \;\; \mathbf{r}_2 \;\; \mathbf{t}]$.

Given this configuration, we can determine the line $\mathbf{b}_1\mathbf{b}_2$ of the reference plane, and also the vanishing point along the vertical direction to the reference plane (i.e. along the y-axis) by intersecting the images of the two vertical objects,

This work was partially supported by ONR grant #N00014-04-1-0512, and Sun Microsystem's grant #EDUD-7824-030482-US.

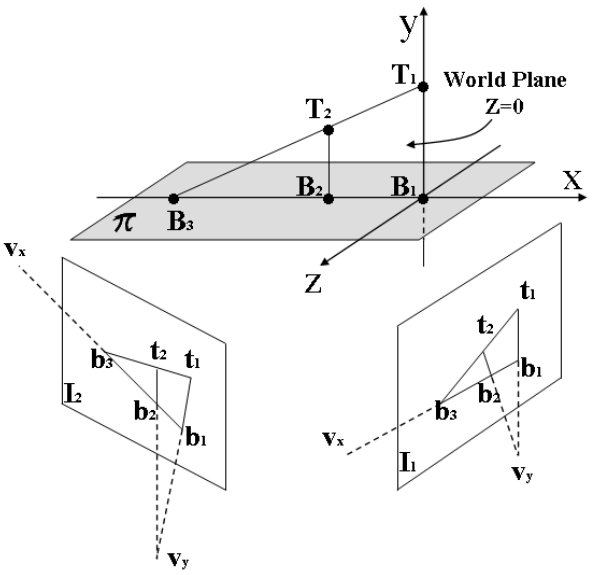

Fig. 1. Our configuration: Two objects $\mathbf{T}_1\mathbf{B}_1$ and $\mathbf{T}_2\mathbf{B}_2$ are both perpendicular to the plane π.

i.e. $\mathbf{b}_1\mathbf{t}_1$ and $\mathbf{b}_2\mathbf{t}_2$. However, since in general the two vertical objects are not of the same height, we can not find the vanishing point along the x-axis by intersecting $\mathbf{b}_1\mathbf{b}_2$ and $\mathbf{t}_1\mathbf{t}_2$. The only information known about $\mathbf{v}_x$ is that it must lie along the line $\mathbf{b}_1\mathbf{b}_2$. However, since the points $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, and $\mathbf{v}_x$ are collinear (see Fig. 1), they define a cross-ratio given by

$$\frac{d(\mathbf{p},\mathbf{b}_2)d(\mathbf{v}_x,\mathbf{b}_1)}{d(\mathbf{v}_x,\mathbf{b}_2)d(\mathbf{b}_3,\mathbf{b}_1)} = \frac{d(\mathbf{B}_3,\mathbf{B}_2)d(\mathbf{V}_x,\mathbf{B}_1)}{d(\mathbf{V}_x,\mathbf{B}_2)d(\mathbf{B}_3,\mathbf{B}_1)} \quad (3)$$

$$= \frac{d(\mathbf{B}_2,\mathbf{T}_2)}{d(\mathbf{B}_1,\mathbf{T}_1)} \quad (4)$$

where $d(,)$ denotes the Euclidean distance between two points.

The first equality follows from the invariance of the cross-ratio under projective transformation, and the second one follows from the geometry of the scene. Therefore the value of the cross ratio in the image plane determines the ratio of the lengths of the two vertical objects in the world. Our problem is thus reduced to finding the unknown components of this cross-ratio. At the first glance, this seems rather impossible, since given the minimal geometric information in the image domain, $\mathbf{v}_x$ can not be determined. The key idea that this paper presents for solving this problem is that all the unknown components of this cross-ratio can be expressed in terms of the camera intrinsic parameters. The solution is thus obtained by minimizing the symmetric transfer errors of geometric distances [5] as described below.

3. PROPOSED SOLUTION

We know that $\mathbf{v}_x$ is along the line $\mathbf{b}_1\mathbf{b}_2$, i.e.

$$\tilde{\mathbf{v}}_x^T(\tilde{\mathbf{b}}_1 \times \tilde{\mathbf{b}}_2) = \tilde{\mathbf{v}}_x^T\mathbf{l}_x = 0 \quad (5)$$

where $\times$ denotes the vector cross-product, and $\mathbf{l}_x$ is the projection of the world x-axis into the image plane.

On the other hand, since $\tilde{\mathbf{v}}_x$ and $\tilde{\mathbf{v}}_y$ are the vanishing points along two orthogonal directions, we have [5]

$$\tilde{\mathbf{v}}_x^T\omega\tilde{\mathbf{v}}_y = 0 \quad (6)$$

where ω is the image of the absolute conic (IAC) [5]. Since we assumed a unit aspect ratio and zero skew for our camera, we have

$$\omega = \frac{1}{f^2}\begin{bmatrix} \mathbf{I}_{2\times2} & -\mathbf{c} \\ -\mathbf{c}^T & \mathbf{c}^T\mathbf{c}+f^2 \end{bmatrix} \quad (7)$$

where $\mathbf{I}_{2\times2}$ is a 2×2 identity matrix.

On the other hand, except for a degenerate case, $\mathbf{l}_x$ is not the vanishing line of the refernce plane, i.e. $\mathbf{l}_x \neq \omega\mathbf{v}_y$. The two lines, however, intersect at $\mathbf{v}_x$. Therefore

$$\tilde{\mathbf{v}}_x = \mathbf{l}_x \times \frac{1}{f^2}\begin{bmatrix} \mathbf{I}_{2\times2} & -\mathbf{c} \\ -\mathbf{c}^T & \mathbf{c}^T\mathbf{c}+f^2 \end{bmatrix}\tilde{\mathbf{v}}_y \quad (8)$$

Since $\tilde{\mathbf{v}}_y$ and $\mathbf{l}_x$ are known, this last equation defines the unknown vanishing point $\tilde{\mathbf{v}}_x$ as a function of the focal length f and the principal point $\mathbf{c} = [u_0 \;\; v_0]^T$. On the other hand, since cross-ratio is a projective invariant, given a pair of images of the scene, we can write the following constraint

$$\left.\frac{d(\mathbf{p},\mathbf{b}_2)d(\mathbf{v}_x,\mathbf{b}_1)}{d(\mathbf{v}_x,\mathbf{b}_2)d(\mathbf{p},\mathbf{b}_2)}\right|^1 = \left.\frac{d(\mathbf{p},\mathbf{b}_2)d(\mathbf{v}_x,\mathbf{b}_1)}{d(\mathbf{v}_x,\mathbf{b}_2)d(\mathbf{p},\mathbf{b}_2)}\right|^2 \quad (9)$$

where the superscripts indicate the image in which the cross-ratio is taken. This equation is quadratic in f^2, and provides four relations between f and $\mathbf{c}$, two of which can be eliminated using the so called cheirality constraint [5], i.e. the fact that objects in the scene must lie in front of the camera. To remove the remaining ambiguity, we will express also $\mathbf{H}$ as a function of f and $\mathbf{c}$, and minimize the symmetric transfer errors of geometric distances.

For this purpose, note that for a camera with unit aspect ratio and zero skew the 3×3 homography $\mathbf{H}$ can also be written as [4]

$$\mathbf{H} = \begin{bmatrix} r_{31}\tilde{\mathbf{v}}_x & r_{32}\tilde{\mathbf{v}}_y & t_z\tilde{\mathbf{b}}_1 \end{bmatrix} \quad (10)$$

where t_z is the z-component of the translation vector, r_{ij} are the components of the rotation matrix, and for our choice of the world reference frame depicted in Figure 1, $\mathbf{b}_1$ is the projection of the world origin into the image plane.

In this last form of the world-to-image homography, $\mathbf{v}_y$ and $\mathbf{b}_1$ are known, and $\mathbf{v}_x$ is already expressed in terms

of the intrinsic parameters using (8). Also, since the world origin is visible in the images, t_z in our case can not be close to zero. Therefore, without loss of generality, we can set $t_z = 1$ for one of the images, and determine the t_z for other images by forcing the corresponding image points to be back-projected to the same 3D point. Therefore the only remaining components of $\mathbf{H}$ that need to be expressed in terms of the intrinsic parameters are the elements of the rotation matrix. If we use (10) and the fact that the rotation matrix must be orthonormal $\mathbf{R}^T\mathbf{R} = \mathbf{I}$, then we can show that

$$\mathbf{r}_1 = \left(\|\tilde{\mathbf{v}}_x - \tilde{\mathbf{c}}\|^2 + f^2\right)^{-\frac{1}{2}} \begin{bmatrix} \mathbf{v}_x - \mathbf{c} \\ f \end{bmatrix}, \quad (11)$$

$$\mathbf{r}_2 = \left(\|\tilde{\mathbf{v}}_y - \tilde{\mathbf{c}}\|^2 + f^2\right)^{-\frac{1}{2}} \begin{bmatrix} \mathbf{v}_y - \mathbf{c} \\ f \end{bmatrix}, \quad (12)$$

and $\mathbf{r}_3 = \frac{\mathbf{r}_1 \times \mathbf{r}_2}{\|\mathbf{r}_1 \times \mathbf{r}_2\|}$. Therefore, we now have the homography $\mathbf{H}$ defined in terms of only the principal point up to a global scalar. This implies that, given $\mathbf{H}_1$ and $\mathbf{H}_2$ as the world-to-image homographies of a pair of images, the inter-image homography $\mathbf{H_c} = \mathbf{H}_1\mathbf{H}_2^{-1}$ depends on the position of the principal point up to an ambiguity caused by the quadratic relation between f^2 and $\tilde{\mathbf{c}}$. This ambiguity is resolved immediately if one of the solutions for f^2 leads to a complex value for f. Otherwise, the true values of the intrinsic parameters and the correct relation between them should minimize the symmetric transfer error of the geometric image distances between the pairs of corresponding points $(\mathbf{m}_i, \mathbf{m}'_i)$.

$$\mathbf{c} = \arg\min_{\Gamma} \sum_i d(\tilde{\mathbf{m}}_i, \mathbf{H_c}^{-1}\tilde{\mathbf{m}}'_i))^2 + d(\tilde{\mathbf{m}}'_i, \mathbf{H_c}\tilde{\mathbf{m}}_i)^2 \quad (13)$$

where Γ is the 2D search space of the solution for (u_0, v_0), $\mathbf{H_c} = \mathbf{H}_1\mathbf{H}_2^{-1}$ is the inter-image homography (which is only a function of the principal point $\mathbf{c}$ up to an ambiguity due to f).

For each relation expressing f in terms of $\tilde{\mathbf{c}}$, we can search for a principal point in the neighborhood of the image center that would minimize the error in (13). The solution is therefore found quickly by discretizing the search space and an exhaustive search. Once $\mathbf{c}$ and hence f are obtained, $\mathbf{v}_x$ can be computed from equation (8), from which one can get the length ratio from equation (9). Therefore given a known reference length in the scene one can transform these length ratios to actual metric measurements. Note also that The rotation matrix can also be computed from equations (11) and (12), and the ranslations are obtained from the last column of the projection matrix by fixing the scale for the first camera and following the approach discussed above Given all internal and external parameters one can then also perform metric measurements outside the reference plane or in directions other than the vertical to the reference plane by using the optimal triangulation algorithm [5].

4. EXPERIMENTAL RESULTS

4.1. Computer Simulation

The simulated camera had a focal length of $f = 1000$, unit aspect ratio, zero skew, and the principal point close to the center of the image. The image resolution is 720×360. We observed the two vertical objects with height 100 and 50 units in different camera poses. First, we evaluated the performance versus noise level over 50 independent trials. Gaussian noise with zero mean and a standard deviation of $\sigma \leq 1.5$ was added to the projected image points. The estimated ratio was then compared with the ground truth and shown in figure 2. The relative error of estimated ratio is 1.46% for a typical noise of 1.5 pixels, and keeps increasing till 2.87% when added a noise of up to 4.5 pixels, which is larger than the typical noise in practice.

We also examined the performance with respect to the number of image pairs, again using 50 independent trials. We show the results using four and nine views also in figure 2. With nine views, the relative error of estimated ratio are not beyond 1% until much noise ($\sigma \geq 1.2$) is added. The more imagess we have, the more accurate are the measurements, since data redundancy compensates for the noise in the data.

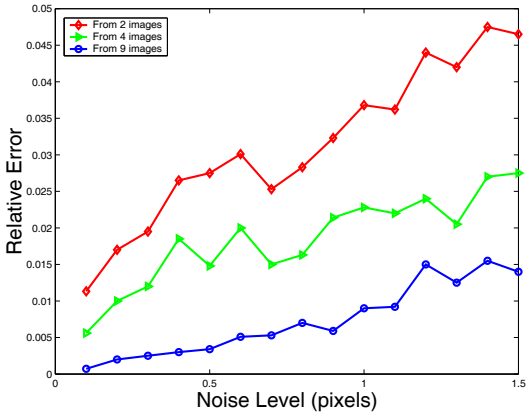

Fig. 2. Performance vs noise (in pixels) using two, four, and nine views. The results shown here are averaged over 50 independent trials.

4.2. Real Data

The proposed method was also tested on many real data sets, some of which are shown below. To test the accuracy, we compared the computed results with ground truth measurements. In all experimental results the relative error did not exceed 4%. For instance in Figure 5, the standing person's height is used as a reference. The estimated stick's height is

100.75cm, while the ground truth is 99.4cm. The distance between the bottom point of the standing person and bottom of the stick is 116cm, the estimated one is 119.47cm. The approach can be used to measure heights of the objects that are not accessible for direct measurement. For instance, we estimated the tree's height as 263.62cm as shown in Figure 5. Other estimated distances which might be difficult to measure in practice are also shown below.

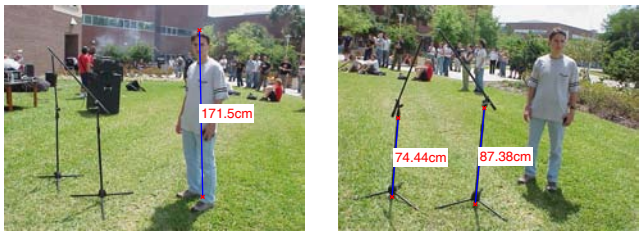

Fig. 3. Standing person has known height. Heights of the two microphone stands are computed using our approach.

Fig. 4. Standing person has known height. Heights of the two metal bars are computed using our approach.

5. CONCLUSION

We have explored new solutions for metrology from uncalibrated images that require minimal geometric information. This work therefore extends the work of Criminisi et al. [3], whereby external geometric constraints are relaxed by trading off the intrinsic constraints. The results show the high accuracy and the effectiveness of the eapproach as compared to the ground truth. The approach can be made further robust by using additional feature points or extra images, in which case one can use bundle adjustment to improve the accuracy.

6. REFERENCES

[1] B. Caprile and V. Torre. Using vanishing points for camera calibration. *International Journal of Computer Vision*, 4(2):127–140, 1990.

Fig. 5. Measurements both along and not along the vertical lines.

Fig. 6. Measurements outside the world plane.

[2] R. Cipolla, T. Drummond, and D. Robertson. Camera calibration from vanishing points in images of architectural scenes. In *Proc. of the British Machine Vision Conference*, pages 382–391, 1999.

[3] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *International Journal of Computer Vision*, 2001.

[4] O.D. Faugeras. *Computer Vision: a Geometric Viewpoint*. MIT Press, 1993.

[5] R.I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[6] D. Liebowitz and A. Zisserman. Combining scene and auto-calibration constraints. In *Proc. International Conference on Computer Vision*, pages 293–300, 1999.

[7] F. Lv, T. Zhao, and R. Nevatia. Self-calibration of a camera from video of a walking human. In *Proc. of the 16th International Conference on Pattern Recognition*, pages 562–567, 2002.

[8] S.J. Maybank and O.D. Faugeras. A theory of self-calibration of a moving camera. *International Journal of Computer Vision*, 8(2):123–152, 1992.

[9] K.-Y. Wong, R.S.P. Mendonca, and R. Cipolla. Camera calibration from surfaces of revolution. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 2003.