Structure from Motion in the Presence of Noise

Gang Liu, Reinhard Klette and Bodo Rosenhahn

CITR, Department of Computer Science
The University of Auckland, Auckland, New Zealand

Abstract

Structure from motion (SfM) aims on estimating 3D structure from 2D image sequences. This paper focuses on a stability analysis and studies the error propagation of image noise. To stabilize SfM, we further present two optimization schemes by using a-priori knowledge of the scene.

Keywords: structure from motion, Gaussian noise, collinearity and coplanarity optimization

1 Introduction

Structure from motion (SfM) is a research topic in computer vision and photogrammetry, with applications in areas such as e-commerce, real estate, games and special effects. It aims at recovering 3D (shape) models of (usually rigid) objects from an (uncalibrated) sequence or set of 2D images.

The original approach [5] of structure from motion consists of the following steps: (1) extract corresponding points from pairs of images, (2) compute the fundamental matrix, (3) specify the projection matrix, (4) generate a dense depth map, and (5) build a 3D model. A brief introduction of some of those steps will be presented in Section 2.

Errors are inevitable to every procedure, and it is also true for SfM. We may expect computation errors appearing in each step. Errors caused by noise (such as sensor measuring errors) are limited to Step 1, and numerical errors may appear in all steps.

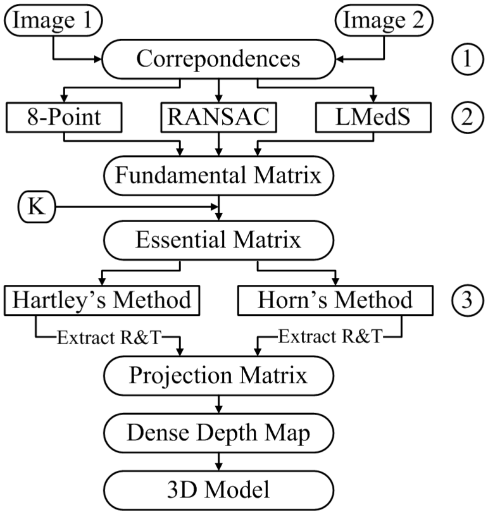

Figure 1: Basic steps of SfM.

To analyze the influence of noise, we perform SfM in a way as shown in Figure 1. At Step 1, Gaussian noise is introduced into coordinates of detected correspondences. At Step 2, three different methods are compared to specify which one is the best to compute the fundamental matrix. At Step 3, a quantitative error analysis is performed. Additionally, to improve the stabilization of SfM, two optimizations are proposed using information from the 3D scene; see Section 3. Section 4 presents experimental results, and Section 5 concludes the paper with a brief summary.

2 Modules of SfM

This section gives a brief introduction for some of the SfM steps (and related algorithms). For extracting correspondent points, we recall a method proposed in [16]. Then, three methods for computing the fundamental matrix are briefly introduced. To specify a projection matrix from a fundamental matrix, we apply two common methods based on [3, 4]. In this step we also use the knowledge of intrinsic camera parameters, which can be obtained through Tsai calibration [12]; this calibration is performed before or after taking the pictures for the used camera. It allows to specify the effective focal length f, the size factors k_u and k_v of CCD cells (for calculating the physical size of pixels), and the coordinates u_0 and v_0 of the principal point (i.e., center point) in the image plane.

2.1 Corresponding points

We need a number of at least seven pairs of corresponding points to determine the geometric relationship between two images, caused by viewing the same object from different view points. One way to extract those points from a pair of images is as follows [16]:

(1) extract candidate points by using the Harris corner detector [2], (2) utilize a correlation technique to find matching pairs, and (3) remove outliers by using a LMedS (i.e., least-median-of-squares) method.

Due to the poor performance of the Harris corner detector on specular objects, this method is normally not suitable for objects with smooth surfaces.

2.2 Fundamental matrix

A fundamental matrix is an algebraic representation of epipolar geometry [15]. It can be calculated if we have at least seven correspondences (i.e., pairs of corresponding points), for example using linear methods (such as the *8-Point Algorithm* of [8]) or nonlinear methods (such as the *RANSAC Algorithm* of [1], or the *LMedS Algorithm* of [16]).

In the case of a linear method, the fundamental matrix is specified through solving an overdetermined system of linear equations utilizing the given correspondences. In the case of a nonlinear method, subsets (at least seven) of correspondences are chosen randomly and used to compute candidate fundamental matrices, and then the best is selected, which causes the smallest error for all the detected correspondences.

According to our experiments, linear methods have a more time efficient and provide reasonably good results for large (say more than 13) numbers of correspondences. Nonlinear methods are more time consuming, but less sensible to noise, especially if correspondences also contain outliers.

2.3 Projection matrix

We express a projection matrix P as follows [3]:

$$P = K[R \mid -RT]$$

where K is a matrix of the intrinsic camera parameters, and R and T are the rotation matrix and translation vector (the extrinsic camera parameters). Since the intrinsic parameters are specified by calibration, relative rotation and translation can be successfully extracted from the fundamental matrix F. When recovering the projection matrices in reference to the first camera position, the projection matrix of the first camera position is given as $P_1 = K_1[I \mid 0]$, and the projection matrix of the second camera position is given as $P_2 = K_2[R \mid -RT]$.

The method proposed by Hartley and Zisserman for computing rotation matrix R and translation vector T (from the essential matrix E) is as follows:

1. Compute E by using $E = K_2^T F K_1$, where

   $$K_i = \begin{pmatrix} fk_u & 0 & u_0 \\ 0 & fk_v & v_0 \\ 0 & 0 & 1 \end{pmatrix}$$

   (note: $K_1 = K_2$ if we use the same camera at view points 1 and 2),

2. Perform a singular value decomposition (SVD) of E by following the template $E = Udiag(1,1,0)V^T$,

3. Compute R and T (for the second view point), where we have two options, namely

   $$R_1 = UWV^T \qquad R_2 = UW^TV^T$$

   $$T_1 = u_3 \qquad T_2 = -u_3$$

   where u_3 is the third column of U and

   $$W = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Another method for computing R and T from E (also only using elementary matrix operations) is given in [4], which leads to almost identical results as the method by Hartley and Zisserman.

2.4 Dense depth map

At this point, the given correspondences allow only a few points to be reconstructed in 3D. A satisfactory 3D model of a pictured object requires a dense map of correspondences. The epipolar constraint (as calculated above) allows that correspondence search can be restricted to one-dimensional epipolar lines, it supports that images are at first rectified following the method in [10], and that correspondence matching is then done by searching along a corresponding scan line in the rectified image. We also require a recovered base line between both camera positions to calculate a dense depth map.

3 Optimization with Prior Knowledge

Since computations of fundamental and projection matrix are sensitive to noise, it is necessary to apply a method for reducing the effect of noise (to stabilize SfM). We utilize information about the given 3D scene, such as knowledge about collinearity or coplanarity. In practice, this knowledge is information of a set of points which lie in a line, e.g. an edge of a building, or a plane, e.g. a tabletop.

3.1 Knowledge about collinearity

It is not hard to detect collinear points on man-made objects, such as buildings or furniture. Assuming ideal central projection (i.e., no lens distortion or noise), then collinear points in object space are mapped onto one line in the image plane. We assume that lens distortions can be ignored. Linearizing points which are supposed to be collinear can then be seen as a way to remove noise.

The approximating line for a set of ”noisy collinear points” is identified by least-square line fitting [7] (minimizing perpendicular offsets). Assume a set of points $P = \{(x_i, y_i) | i = 1, \ldots, n\}$ which determines a line $l(\alpha, \beta, \gamma) = \alpha x + \beta y + \gamma$. The coefficients α, β and γ are calculated as follows:

$$\alpha = \frac{\mu_{xy}}{\sqrt{\mu_{xy}^2 + (\lambda^* - \mu_{xx})^2}}$$

$$\beta = \frac{\lambda^* - \mu_{xx}}{\sqrt{\mu_{xy}^2 + (\lambda^* - \mu_{xx})^2}}$$

and $\gamma = -(\alpha \overline{x} + \beta \overline{y})$, where

$$\lambda^* = \frac{1}{2}(\mu_{xx} + \mu_{yy} - \sqrt{(\mu_{xx} - \mu_{yy})^2 + 4\mu_{xy}})$$

$$\mu_{xx} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$\mu_{yy} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$\mu_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{and} \quad \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Points' positions are modified through perpendicular projection onto the calculated line.

3.2 Knowledge about coplanarity

Coplanar points can be expected on rigid structures such as on walls or on a tabletop. For a set of points, all incident with the same plane, there is a 3×3 matrix H called *homography* which defines a perspective transform of those points into the image plane [11].

3.2.1 Homography

Consider we have an image sequence (generalizing the two-image situation from before) and p_{ki} is the projection of 3D point P_i into the kth image, i.e. P_i is related to p_{ki} as follows:

$$p_{ki} = \omega_{ki} K_k R_k (P_i - T_k) \tag{1}$$

where ω_{ki} is an unknown scale factor, K_k denotes the intrinsic matrix (for the used camera), and R_k and T_k are the rotation matrix and translation vector. Following Equation (1), P_i can be expressed as follows:

$$P_i = \omega_{ki}^{-1} R_k^{-1} K_k^{-1} p_{ki} + T_k \tag{2}$$

Similarly, for point p_{li} lying on the lth image, we have

$$P_i = \omega_{li}^{-1} R_l^{-1} K_l^{-1} p_{li} + T_l \tag{3}$$

From Equations (2) and (3), we get

$$p_{ki} = \omega_{ki} K_k R_k (\omega_{li}^{-1} R_l^{-1} K_l^{-1} p_{li} + T_l - T_k) \tag{4}$$

With $R_{kl} = R_k R_l^{-1}$ we define $H_{kl}^{\infty} = K_k R_{kl} K_l^{-1}$. We also have epipole $e_{kl} = K_k R_k (T_l - T_k)$. Equation (4) can then be simplified to

$$p_{ki} = \omega_{ki} \omega_{li}^{-1} (H_{kl}^{\infty} p_{li} + \omega_{li} e_{kl}) \tag{5}$$

H_{kl}^{∞} is what we call the homography which maps points at infinity ($\omega_{li} = 0$) from image l to image k. Consider a point P_i on plane $\widehat{n}^T P_i - d = 0$. Then, from Equation (3), we have

$$\widehat{n}^T P_i - d = \widehat{n}^T \omega_{li}^{-1} R_l^{-1} K_l^{-1} p_{li} + \widehat{n}^T T_l - d = 0$$

Then we have

$$\omega_{li} = \frac{\widehat{n}^T R_l^{-1} K_l^{-1} p_{li}}{d - \widehat{n}^T T_l}$$

what can be rewritten as follows:

$$\omega_{li} = d_l^{-1} \widehat{n}^T R_l^{-1} K_l^{-1} p_{li}$$

where $d_l^{-1} = d - \widehat{n}^T T_l$ is the distance from the camera center (principal point) of the lth image to the plane $(\widehat{n}, d)$. Substituting ω_{li} into Equation (5), finally we have

$$p_{ki} = \omega_{ki} \omega_{li}^{-1} (H_{kl}^{\infty} + d_l^{-1} e_{kl} \widehat{n}^T R_l^{-1} K_l^{-1}) p_{li}$$

Let

$$H = \omega_{ki} \omega_{li}^{-1} (H_{kl}^{\infty} + d_l^{-1} e_{kl} \widehat{n}^T R_l^{-1} K_l^{-1})$$

This means: points lying in the same plane have identical H which can be utilized as coplanarity constraint; see [11].

3.2.2 Coplanarity optimization

Coplanar points satisfy the relation described by homography. We use this relation for modifying “noisy coplanar points,” using the equation

$$p_{ki} = H_{kl} p_{li}$$

Here, H_{kl} is the homography between kth and lth image in the sequence, and p_{ki}, p_{li} are projections of point P_i on the kth and lth image, respectively.

4 Experiments and Analysis

This section shows at first experiments of the performance of different methods for computing the fundamental matrix, and second the performance of the remaining SfM steps. Finally we evaluate effect of the optimizations mentioned in Section 3.

4.1 Computation of fundamental matrix

Three algorithms (8-Point, RANSAC and LMedS) are compared with each other in this section. To specify the most stable one in presence of noise, Gaussian Noise (with mean 0 and deviation $\delta = 1$ pixel) and one outlier are propagated to given correspondences. Performances of the three algorithms are characterized in Figure 2: the 8-Point Algorithm is more sensitive than the other two.

4.2 Optimizations

To test the effect of the optimizations mentioned in the previous section, the results of splitting essential matrices (rotation matrices and translation vectors) are compared to each other. Two images of a calibration object are used as test images (shown in Figure 3). The data got from calibration (intrinsic parameters and extrinsic parameters of camera) are used as the ground truth.

The roll angle α, pitch angle β and yaw angle γ are used to compare the rotation matrices in a quantitative manner. These angles can be computed from a rotation matrix R by the following equations [9]:

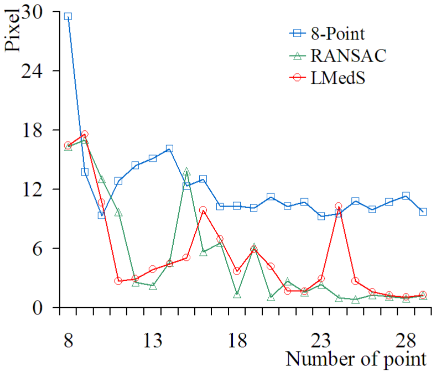

Figure 2: Performance of three algorithms in presence of noise.

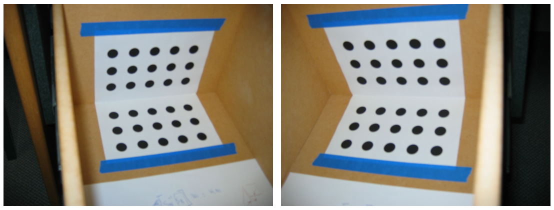

Figure 3: The first (left) and second (right) candidate images.

$$\begin{aligned} \alpha &= \operatorname{atan2}\left(\frac{r_{23}}{sin(\gamma)}, \frac{r_{13}}{sin(\gamma)}\right) \\ \beta &= \operatorname{atan2}\left(\frac{r_{32}}{sin(\gamma)}, \frac{-r_{31}}{sin(\gamma)}\right) \\ \gamma &= \operatorname{atan2}\left(\sqrt{r_{31}^2 + r_{32}^2}, r_{33}\right) \end{aligned}$$

where r_{ij} is the element of R at ith row and jth column, and

$$atan2(y,x) = \begin{cases} atan(\frac{y}{x}) & (x > 0) \\ \frac{y}{|y|} \cdot (\pi - atan(|\frac{y}{x}|)) & (x < 0) \\ \frac{y}{|y|} \cdot \frac{\pi}{2} & (y \neq 0, x = 0) \\ \text{undefined} & (y = 0, x = 0) \end{cases}$$

Since splitting the essential matrix only results in a translation vector up to a scale factor [3], all translation vectors (including the ground true one) are transformed into a normalized vector (length equal to one unit) to compare with each other in a quantitative manner.

The comparison of rotation matrices and translation vectors are shown in Figure 4 and Figure 5. The errors are the mean errors of ten iterations when different number of correspondences are given. The noise propagated is Gaussian noise (with mean 0 and deviation $\delta = 1$ pixel). The method used to compute the fundamental matrix is the 8-Point Algorithm, which is more sensitive to noise than RANSAC and LMedS Algorithm. The method of Hartley and Zisserman is used to split the essential matrix.

According to the results shown in Figure 4 and Figure 5, the coplanarity knowledge gives a better optimization than collinearity knowledge. One possible reason is that the collinearity optimization is performed on uncalibrated images, in which the true correlation of collinear points are not strictly lying in a straight line.

For arbitrary images, the effect of optimizations can be seen from Figure 6 through looking at relative positions of epipolar lines computed from different data sets. It shows that the two optimization strategies bring positive effects on reducing the influence of noise, and the coplanarity optimization performs better than the collinearity optimization.

Figure 7 shows an example of the reconstructed CITR-building, located at Tamaki campus in Auckland.

5 Summary

Modules relating to structure from motion have been discussed in this paper. According to experiments, structure from motion is sensitive to noise

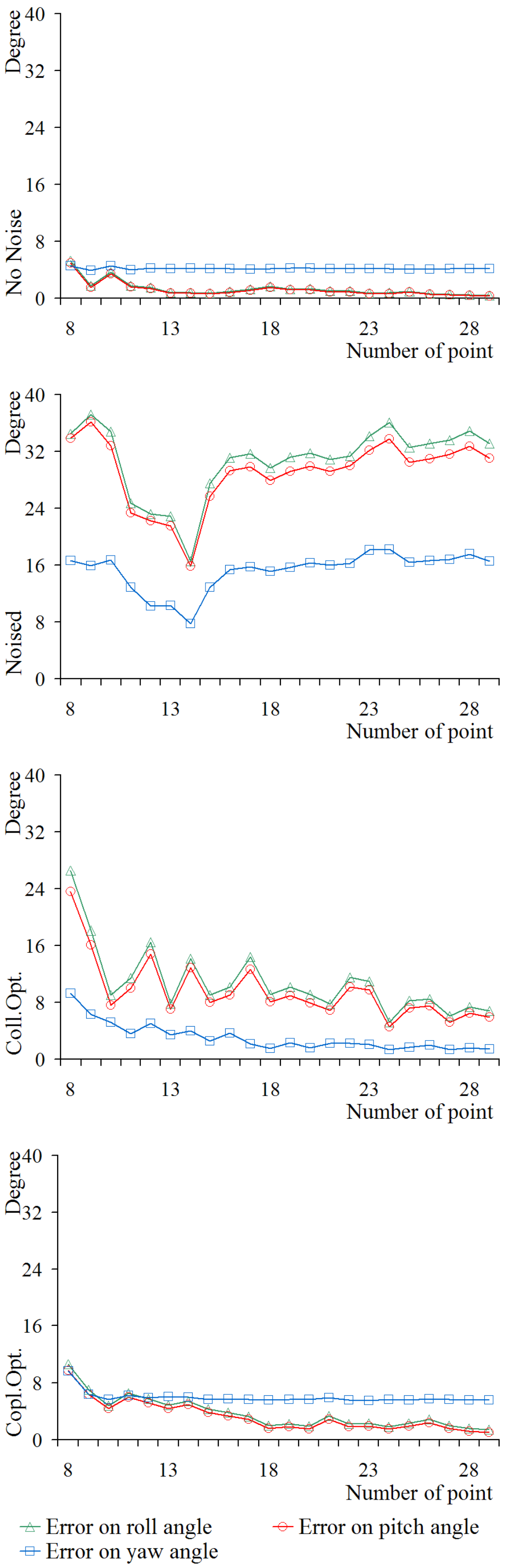

Figure 4: Errors at rotation matrices. The first diagram shows the errors from non-noisy data (regarded as ground truth); the second diagram shows the errors from noisy data; the third and fourth diagram indicate the errors from the noisy data after being respectively optimized with collinearity and coplanarity knowledge.

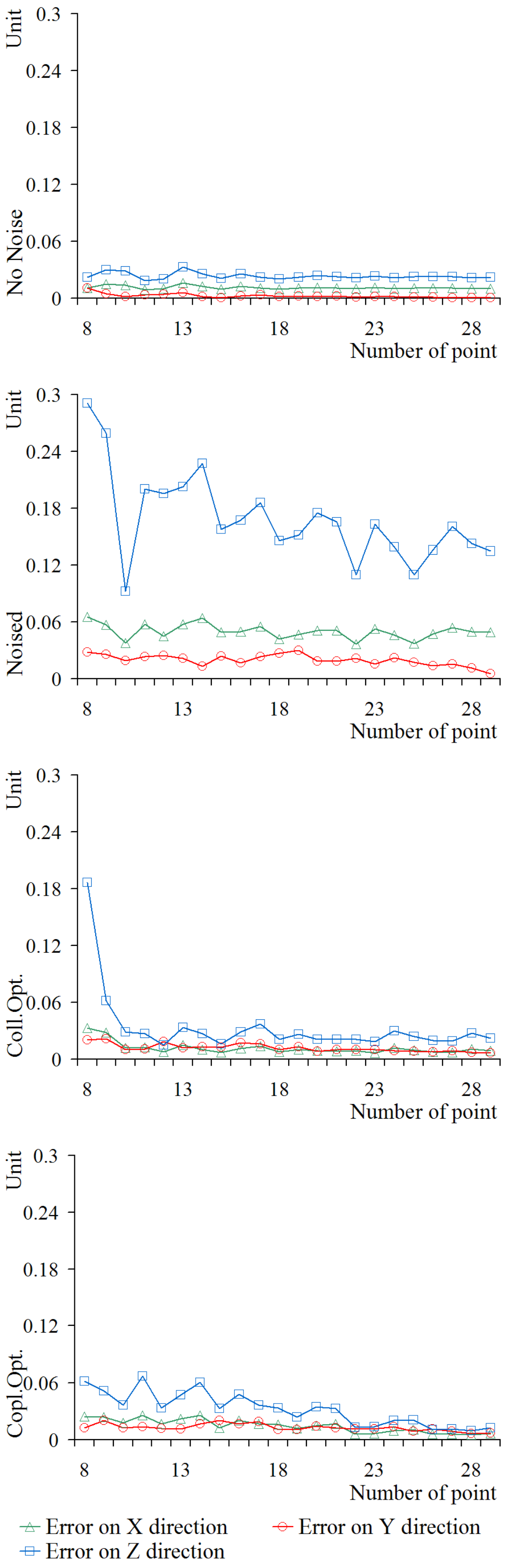

Figure 5: Errors at translation vectors. The first diagram shows the errors from non-noisy data (regarded as ground truth); the second diagram shows the errors from noisy data; the third and fourth diagram indicate the errors from the noisy data after being respectively optimized with collinearity and coplanarity knowledge.

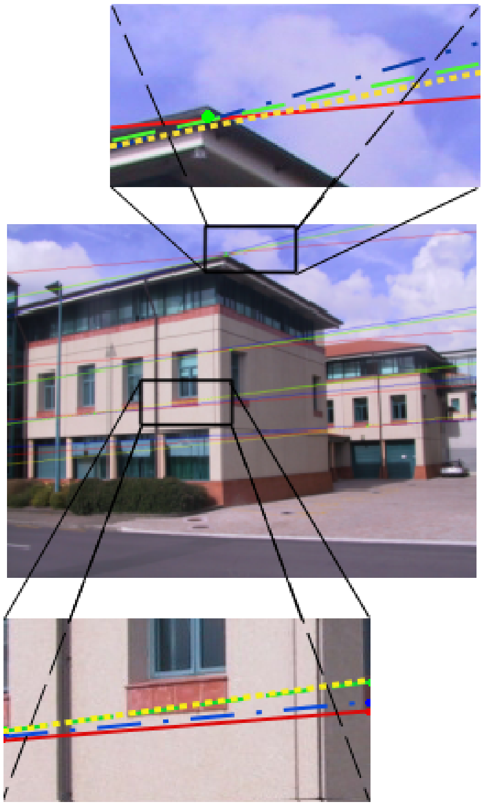

Figure 6: Epipolar lines result from different data sets: green (dashed) lines from data without generated noise, the red (straight) lines from noisy data, the blue (dash-dotted) lines or yellow (dotted) lines from noisy data and the use of collinearity or coplanarity knowledge, respectively.

and it is necessary to improve its stability. Two optimizations, using collinearity and coplanarity knowledge, have been proposed, and the relating experiments show that the two proposed optimizations, especial the coplanarity one, bring positive effects on reducing influences of noise.

Acknowledgments: The authors thank Daniel Grest and Kevin Koeser from the University of Kiel for the BIAS-Library and useful hints.

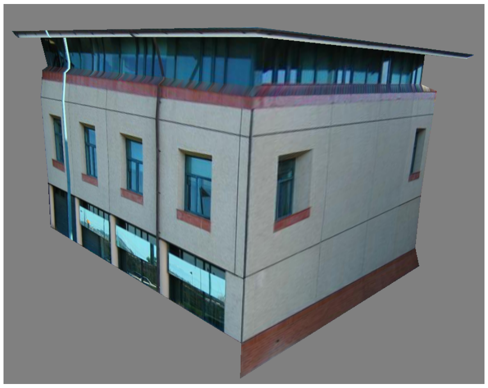

Figure 7: Example: reconstructed CITR–building.

References

[1] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, **24**:381–385, 1981.

[2] C. Harris and M. Stephen. A combined corner and edge detector. In Proc. *Alvey Vision Conf.*, pages 147–151, 1988.

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

[4] B. K. P. Horn. Recovering baseline and orientation from essential matrix. http://www.ai.mit.edu/people/bkph/papers/essential.pdf, 1990.

[5] T. Huang. Motion and structure from feature correspondences: a review. *Proc. IEEE*, **82**:252–268, 1994.

[6] R. Klette, K. Schlüns, and A. Koschan. *Computer Vision – Three-dimensional Data from Images*. Springer, Singapore, 1998.

[7] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice Hall, Englewood Cliffs, 1974.

[8] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, **293**:133–135, 1981.

[9] R. Murray, Z. Li, and S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Boca Raton, 1994.

[10] M. Pollefeys, R. Koch, and L. V. Gool. A simple and efficient rectification method for general motion. In Proc. *Int. Conf. Computer Vision*, pages 496–501, 1999.

[11] R. Szeliski and P. H. S. Torr. Geometrically constrained structure from motion: points on planes. In Proc. *European Workshop 3D Structure Multiple Images Large-Scale Environments*, pages 171–186, 1998.

[12] R. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE J. Robotics and Automation*, **3**:323–344, 1987.

[13] J. Weng, N. Ahuja and T. S. Huang. Optimal Motion and Structure Estimation. *PAMI*, **15**:864–884, 1993.

[14] J. Weng, T. S. Huang and N. Ahuja. Motion and structure from two perspective views: algorithms, error analysis, and error estimation. *PAMI*, **11**:451–476, 1989.

[15] Z. Zhang. Determining the epipolar geometry and its uncertainty: a review. *Int. J. Computer Vision*, **27**:161–198, 1998.

[16] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence J.*, **78**:87–119, 1995.