3D RECONSTRUCTION FROM IMAGE SEQUENCE TAKEN WITH A HANDHELD CAMERA*

Shaoxing Hu[a], Jingwei Qiao[a], Aiwu Zhang[b], Qiaozhen Huang[a]

[a]School of Mechanical Engineering and Automation, Beijing University of Aeronautics and Astronautics, Beijing, P.R. China -JingweiQiao@gmail.com

[b]Key Lab of 3D Information and Application of Ministry of Education ,Capital Normal University, Beijing, P.R. China - husx@buaa.edu.cn

KEY WORDS: Image Sequence, Matching, Fundamental Matrix, Algorithm, 3D Reconstruction

ABSTRACT:

This paper shows in theory and in practice how to implement a 3D reconstruction algorithm. It uses image sequence taken with a handheld camera as input to reconstruct a scene up to an unknown scale factor. The camera's motion and intrinsic parameters are all unknown. We especially address to apply SIFT algorithm to find distinctive features. Feature matching is done through a Euclidean-distance based nearest neighbor searching. The fundamental matrix is then estimated by existing correspondences and sequentially used to refine matching. By recovering project matrix from fundamental matrix we get projective reconstruction. And finally we demonstrate how the projective reconstruction can be successively upgraded to affine and Euclidean reconstruction.

1. INTRODUCTION

Concerning the reconstruction of 3D scene, there are commonly two approaches: (1) laser scanners and (2) image-based approach. These scanners are robust and accurate, but they are also costly, and have certain restrictions on the size and on the surface properties of the objects. They are also unable to capture the color information of the objects. The image-based method is built on the knowledge that has been acquired in computer vision and photogrammetry in the past 30 years. Despite of less accuracy, it reconstructs 3D model provided at least two images or a sequence of images that can be easily obtained by optical imaging devices (e.g., CCD camera). So it is relatively low cost which is extremely important for such applications in virtual reality, simulation and entertainment. Especially as the development of electronic technique in the past few years, digital cameras are more and more universalized and produce quality images. This excellent imaging device has attracted researchers' attention.

The aim of our work is to create a system that allows users to make the reconstruction without other expensive and complex devices but a domestic handheld camera. The user acquires images by freely moving the camera with not very large viewport transformation around a static environment whose motions are unknown and whose intrinsic parameters are also unknown and may vary. The software system will do the 3D reconstruction following several steps with few human operations.

In this work, a computer vision and photogrammetric approach for 3D reconstruction is described. The process consists of four parts:

1) Image sequence acquisition and feature points extraction;
2) Feature matching and refinement;
3) Projective, Affine and Euclidean reconstruction;
4) Point cloud generation and modeling.

Figure 1 shows the architecture of the reconstruction. In the process, we put special emphasis on finding distinctive features and robustly matching them. For the capturing of images, we using a Canon A560 digital camera and no a priori information on camera internal and external parameters are assumed; all the required parameters are recovered from the images.

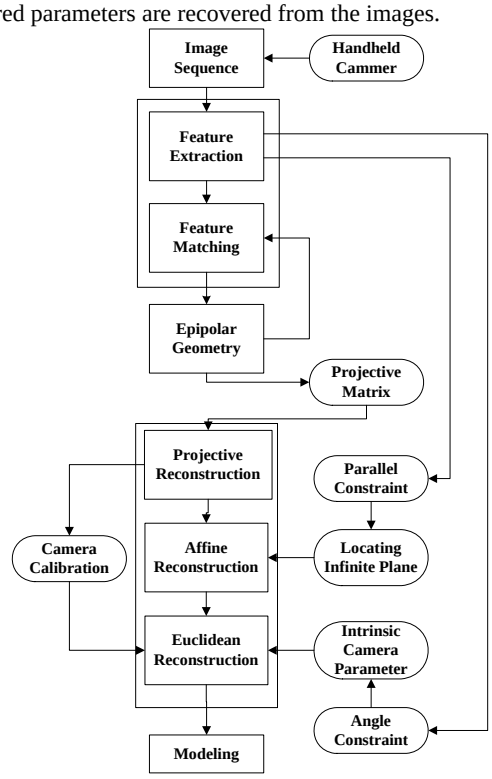


Figure 1. Architecture of Reconstruction

* This work was supported by the National Natural Science Foundation (NSFC40601081), Beijing Municipal Science & Technology Commission (2006B57), and National Education Department Doctor Fund (20070006031)

2. FEATURE POINT EXTRACTION

Starting from a collection of images, the first step consists in detecting image feature points or so-called interest points in each image. Feature points are distinctive and invariant to different transformations and have high information content. Many algorithms are available, such as Harris operator (Harris, 1988), SUSAN operator (Stephen, 1997). Since the images are of relative small viewport change and captured in different time and place, we use SIFT algorithm (Lowe, 2004) in our application to extract and describe feature points.

The SIFT features are invariant to image scale and rotation. They are also robust to changes in illumination, noise, occlusion and minor changes in viewpoint. It is also shown that SIFT descriptors are invariant to minor affine changes (Lowe, 2004). The algorithm is performed in the following four stages:

1) Extrema detection in scale space
2) Refining keypoints location
3) Keypoint orientation Assignment
4) Generation of keypoint descriptor

By the four steps keypoints are detected and described in SIFT descriptors that are computed over respective scales. Typically, a SIFT descriptor is of length 128 (8 orientation bins and 4 by 4 cells for voting). Extracted feature points are marked in Figure 2.

Figure 2. Feature pints detected by SIFT algorithm. Six images from a sequence of indoor scene. The image size is 800*600 pixels, and about 1000 features are detected in each image. Because of redundant keypoints at different scales there will be several duplicate points around a keypoint when transferred to original images. However feature point descriptors at multi-scales makes matching result invariant to scales.

3. MATCHING CORRESPONDENCES

3.1 Initial Matching

General steps of matching feature points are illustrated in Figure 3. We have got keypoints and now we are matching them across neighboring image pairs.

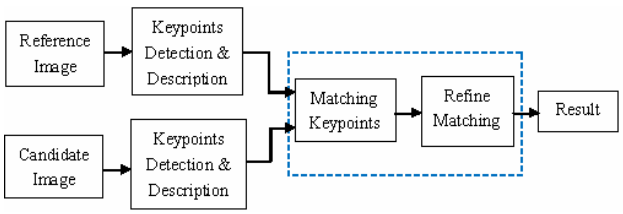


Figure 3. General steps of feature point matching

Euclidean distance between keypoints in different images is calculated to measure the similarity.

$$D(X,Y) = \| X - Y \| = \sqrt{\sum_{i=1}^{d} (X_i - Y_i)^2} \quad (1)$$

Small D indicates that the two keypoints are close and thus of high similarity.

To matching a keypoint in candidate image, we find the closest and the second closest keypoints in reference image using nearest neighbor searching strategy. If the ratio of them is smaller than a threshold, the keypoint and the closest matched keypoint are accepted as correspondences; or else, the keypoint cannot be matched.

3.2 Refining matching

The above simple matching bases only on similarity of keypoints and inevitably produces mismatches. For image pairs, the epipolar geometry provides a constraint for identifying mismatches between feature points: corresponding keypoints are constrained to lie on epipolar lines. This relationship can be expressed as following:

$$u^T F v = 0 \quad (2)$$

where, F is the so-called fundamental matrix. u is an image point, u= [u1, u2, 1] T, and v is the corresponding point in another image, v= [v1, v2, 1] T.

So we firstly using the initial matching result to retrieval the fundamental matrix and then apply it to refine matching. The 3×3 matrix F can be computed just from image points and at least 7 correspondences are needed to compute it.
Equation (2) can be written as:

$$u^T f = 0 \quad (3)$$

where $u = [u_1 u_2, v_1 u_2, v_1 v_2, v_2, u_1, v_1, 1]^T$

$f = [F_{11}, F_{12}, F_{13}, F_{21}, F_{22}, F_{23}, F_{31}, F_{32}, F_{33}]$

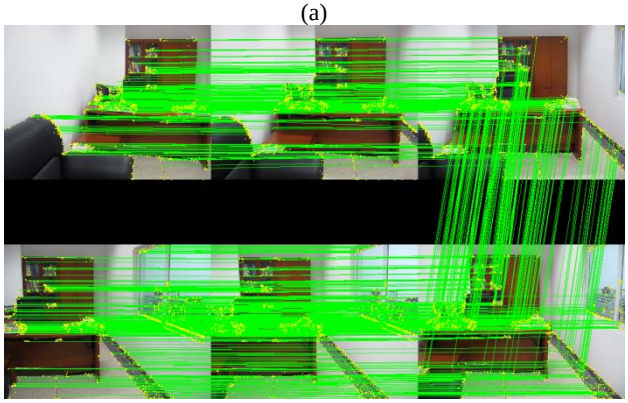
Thus we know that if we are given eight matches we will be able to determine a unique solution for F, defined up to a scale factor. This approach is known as the eight point algorithm (Richard, 1997).

Many solutions have been published to compute fundamental matrix, but to cope with possible blunders, a robust method of estimation is required. In general RANSAC-like algorithm (Fischler, 1981) and least median of the squares (LMedS) (Zhang, 1994; Scaioni, 2001) are very powerful in presence of mismatches. LMedS solves non-linear minimization problems and yield the smallest value for the median of the squared residuals.

The computed epipolar geometry is then used to refine the matching process, which is now performed as guided matching

along the epipolar lines. This geometric constraint restricts the searching area and allows a higher threshold for the matching process. A maximal distance from the epipolar line is also set as threshold to accept a point.

(a)



(b)

Figure 4. (a) Initial matching. There are about 5% mismatches. (b) Refined matching. By epipolar geometry constraint almost all correspondences are correctly matched.

After these processes, the number of matched points between the image pair is around 58% of the extracted feature points.

4. RECONSTRUCTION FROM IMAGE POINTS

4.1 Projective Reconstruction

Up to this stage of the process, we have got a number of reliable correspondences between each pair of contiguous images in the ordered subsequence and also estimated respective image pairs' fundamental matrices. The set of N-1 fundamental matrices (given N images in the sequence) is then used to derive a first approximation of the set of N projective matrices Pj. For the first two images, we can compute Projective matrices P1 and P2 as follows (Faugeras, 1992a; Beardsley, 1997):

Decomposite fundamental matrix F= [e2] ×M21 where [e2] × is an antisymmetric matrix. e2 can be uniquely calculated up to a scale factor by solving linear equation $e_2^T F = 0$. M21 can be many solutions but we just need to employ one of them.
So we then get the projection matricesP1 and P2:

$$P_1 = (I \quad 0)$$
$$P_2 = (M_{21} \quad e_2) \tag{4}$$

Once P1 and P2 are computed, the initial projective structure can be recovered:

$$u = P_1 X$$
$$v = P_2 X \tag{5}$$

where, X is a 3D point
If there are more than one image pairs in the sequence, an updating procedure needs to be defined, so that all image pairs contribute to the constructed 3D points. In order to update this reconstruction with another image, the projective matrix P for the new image must be computed. In case P is known, we can reconstruct those X that are visible in at least two views of the previous images. Then iterating this process, we can compute the reconstruction from all the images.

4.2 Toward Euclidean Reconstruction

For simple reasons, the projective reconstruction may not be sufficient for visualization. Therefore we need a method to upgrade the reconstruction to metric one. This can be achieved by computing an upgrading matrix with a precondition of information about the intrinsic parameters of the camera. In the most general case they are constant but unknown. So the core issue now lies in the camera calibration. Here a self-calibration method using sense constraints is proposed.

In this approach which can be fully automated we first retrieve the affine stratum by extracting pairs of parallel lines in the scene and computing the plane at infinity. Retrieving the affine structure amounts to define a projective basis set and maps it on a reference set. Five reconstructed points of the scene are required for this computation: one for the origin, three to define coordinate axes and planes, and a final one (not in the coordinate planes) representing the scaling effect along the 3 coordinates axes.

Then we go to the Euclidean stratum from the affine one. We extract pairs of orthogonal lines in the scene. The use of three pairwise orthogonal directions permit to rectify the affine coordinate basis up to three scale factors. To retrieve the Euclidean structure, up to a global scale of the scene, thee skew parameters are introduced to account for the non-orthogonality of the reference affine basis. A standard iterative technique then leads us to the solution.

Thus, we now have a way of computing a Euclidean reconstruction of the scene without any knowledge of the camera parameters or of the scene coordinates. Only information about point and line matches, parallelism, and angular relations have been used.

5. RESULT

The experimental result for the reconstructed coarse 3D point structure is shown in Figure 5. Our next work is to reconstruct more accurate points and build a model with texture.

Figure 5. Result of 3D reconstruction from the indoor scene sequence.

REFERENCE

Beardsley P.A., Zisserman A., 1997. Sequential updating of projective and affine structure from motion. Int. J. Computer Vision, 23, pp. 235-259

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2), pp. 91-110

Faugeras O.D., Luong T., 1992b. Camera self-calibration: theory and experiments. In Proc 2nd ECCV, Santa Margherita Ligure, Italy, Lecture Notes in Computer Science, Vol. 588, pp. 321-334,

Faugeras O.D., 1995. Stratification of 3-D vision：projective, affine, and metric representations [J]. Journal of the Optical Society of America, 12(3), pp. 465-484

Faugeras, O.D., 1992a. What can be seen in three dimensions with an uncalibrated stereo rig? In Proc. 2nd European Conference on Computer Vision, Santa Margherita Ligure, Italy, pp. 563-578

Fischler M., Bolles R., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Comm. Assoc. Comp., 24 (6), pp. 381-395

Gruen A., Beyer H.A.,1992. System calibration through self-calibration, in Proceedings of the Workshop on Calibration and Orientation of Cameras in Computer Vision, Washington D.C.

Harris C, Stephens M, 1988. A combined corner and edge detector. Alvey Vision Conference, pp. 147-151

Hartley R., 1997. Kruppa's equations derived from the fundamental matrix. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2), pp. 133～135

Richard I., Hartley, 1997. In defense of the eight-point algorithm. IEEE Transaction on Pattern Recognition and Machine Intelligence, 19 (6), pp. 580-593.

Scaioni M., 2001. The use of least median squares for outlier rejection in automatic aerial triangulation. Proc. of 1st Int. Symposium on Robust Statistics and Fuzzy Techniques in Geodesy and GIS, ETH Zurich, pp. 233-238

Stephen M. Smith, 1997. SUSAN - A new approach to low level image processing. International Journal of Computer Vision, 23(1), pp. 45-78

Zhang Z., Deriche R., 1994. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. TR 2273, INRIA