https://doi.org/10.1051/matecconf/201823202045

3D recognition based on ordered images reconstruction

Ning Zhang[1], YongJia Zhao[2]

State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, 100191, China

Abstract. Nowadays, more and more applications require precise and quickly 3D recognition, such as augmented reality and robot navigation. In recent years, model-based methods can get accurate object or scene recognition, but it takes a lot of time to reconstruct the model. Therefore, we propose a fast 3D reconstruction method based on ordered images for robust and accurate 3D recognition. The proposed algorithm consists of two parts, the offline processing stage, and the online processing stage. First, in the offline processing stage, the sparse point cloud model of the scene or object is reconstructed based on the sequential images, optimized using the BA algorithm based on the local correlation frame, and then the local descriptor of the resulting model points is stored. Secondly, in the online processing stage, for each image frame of the camera video, a matching relationship between the stored point cloud and the 2D feature points on the image frame is established, based on which the pose of the camera can be solved accurately.

1 Introduction

Nowadays, many fields require precise and fast 3D object recognition. For example, in the field of robot navigation, it is necessary to identify an unknown scene to know its position. In augmented reality applications[1], it is necessary to accurately 3D registering and render the virtual content. This paper focuses on 3D recognition in augmented reality.

The goal of augmented reality is to seamlessly combine virtual information with the input real scene. To do this, we need to accurately calculate the pose of the input device (camera, etc.) relative to the scene. Model-based 3D tracking is the main algorithm for solving camera poses. The model can be manually set fiducial marker, a point cloud model recovered using SFM[11] (structure from motion), or a CAD model. Augmented reality systems based on artificial markers provide fast and accurate tracking, but there is must be pre-prepared artificial markers in the scene, which can cause damage to the original scene. Getting the CAD model of target object is too time consuming. So, a method based on ordered images 3D reconstruction is used to realize fast and automatic acquisition of the scene model. This approach takes full advantage of the natural features of the environment and allows scene recognition from any viewpoint.

This algorithm consists of two stages. In the offline stage, we use a calibrated monocular camera to take a series of ordered images for a specific scene or object, using an SFM to construct a sparse 3D point cloud model of an object or scene. The output of this part is a sparse point cloud model and local descriptors of points. In the online stage, for the current frame of camera, the keypoints are extracted and matched with the 3D points recovered in the offline phase. Matching pairs of 2D-3D points can be used to solve camera pose and render virtual content on the real scene.

2 Related Work

Now, the research on 3D recognition mainly focuses on two aspects: preprocessing and real-time 3D tracking. The 3D tracking problem is the basic 2D-3D position estimation problem in computer vision. Under the pinhole camera model, given the 2D-3D matching point pair, determining the camera's orientation in the world coordinate system is the PnP[7] problem. When the points are less, it can be solved by a non-iterative method, such as the DLT method. Augmented reality 3D tracking contains more matched points, but there is a certain number of mismatches. In order to solve the problem quickly, the robust PnP solving method based on RANSAC is used to eliminate the mismatch and use the iterative method to optimize the solution result.

Kato et al.'s ARToolKit[9] can accurately identify and track 2D markers, but some natural scenes can't be intervened. Skrypnyk and Lowe[3] studied the scene model reconstruction from a set of unordered images, and input images are matched to scene model during real-time 3D tracking stage. In the preprocessing stage, they use SFM[11] to restore the point cloud model of the target scene and local descriptors of 3D points. In the real-time 3D tracking stage, SIFT feature points on the input image can be matched directly with point cloud features, to obtain the corresponding point pairs between 2D and 3D,

[1] zy1703239@buaa.edu.cn, [2] zhaoyongjia@buaa.edu.cn

and then optimize the camera external parameters through RANSAC[8] and Lenvenberg-Marquardt algorithm, and use the previous frame to solve the results and dynamic weights to reduce the jitter of the parameters. Their system provides a basic framework for real-time 3D tracking based on point cloud models. Mooser[2] used the KLT method to recover the point cloud model during the preprocessing stage, but they used the projection base to generate the description of the feature points.

Our work is similar to theirs, but it increases the speed of 3D reconstruction. The traditional 3D reconstruction methods, such as colmap[12], cost a lot of time because it needs to match any two images of input image. A fast 3D reconstruction method based on ordered images is proposed. It only needs to match the adjacent images in the ordered image sequence to reduce the matching phase time and improve the reconstruction speed. The experimental results show that the proposed method can achieve same level of reconstruction effect compared with the traditional full matching method, but it has obvious advantages in reconstruction speed.

3 Algorithm architecture

The system studied in this paper for augmented reality applications will follow the system architecture shown in Figure 1.

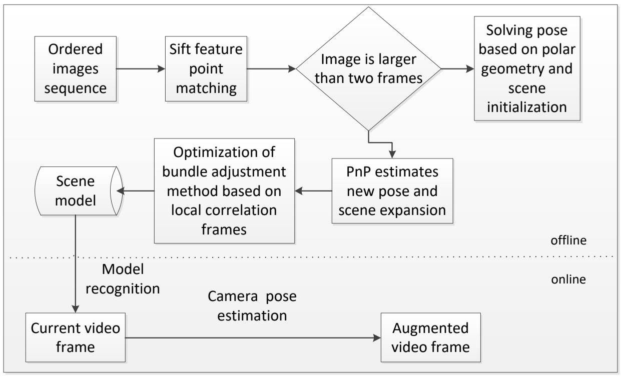
Ordered images sequence
Sift feature point matching
Image is larger than two frames
Solving pose based on polar geometry and scene initialization
PnP estimates new pose and scene expansion
Optimization of bundle adjustment method based on local correlation frames
Scene model
offline
online
Model recognition
Camera pose estimation
Current video frame
Augmented video frame

Figure 1. Augmented Reality System Architecture

This article will use the framework shown above to implement an augmented reality application based on the 3D point cloud model.

1) Most of the existing 3D reconstruction algorithms are aimed at unordered pictures. In fact, we are shooting a series of ordered pictures for a specific scene. For the ordered pictures, this paper proposes a robust algorithm for monocular 3D reconstruction. Compared to existing unordered 3D reconstruction, the complexity of the algorithm can be reduced.
2) In the optimization of 3D scene structure, compared with the traditional bundle adjustment method, this paper uses a bundle adjustment method based on local correlation frame optimization.
3) In the feature matching process between the point cloud model and the image feature points, the RANSAC-based method is used to solve the initial value of the camera pose, and then the Lenvenberg-Marquardt algorithm is used to optimize the camera pose.

4 Detailed description of the algorithm

4.1. Ordered image sequence reconstruction

Compared with the 3D reconstruction of the unordered image sequence, we think that ordered images have a good correlation, which were captured consequently by a monocular camera. Therefore, we use a monocular camera to continuously take a series of ordered pictures to restore the 3D structure of the object or scene. We need to calibrate the camera’s internal parameters. The entire scene reconstruction can be summarized as the following steps:

1) Feature extraction of input ordered image sequences using SIFT features.
2) The Brute-Force Matcher is used to match the features between the two images, and the RANSAC algorithm is used to remove the outliers.
3) Use polar geometry to initialize camera pose and triangulation for 3D scene initialization.
4) Use the robust PnP algorithm to calculate the new camera pose, and improve the usage strategy, then expand the scene by considering the commonality of the feature points.
5) Appling the Bundle Adjustment method based on local correlation frames to optimize the restored pose and 3D points.

4.1.1 Feature extraction and matching author names and author affiliations

A robust and efficient SIFT[5] feature matching algorithm is adopted in this paper. We get initial matches of two images based on Brute-Force Matcher, but there are many mismatchs in them. In order to remove these noises, after completing the SIFT feature matching, this paper uses RANSAC algorithm to reject the mismatched points. Figure 2 shows the matching results after the RANSAC rejects the mismatch point. Obviously, most of the matching points are correct after the RANSAC iteration rejects the mismatch point.

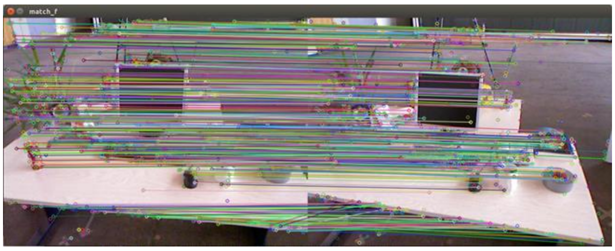

Figure 2. SIFT feature matching results after RANSAC optimization

4.1.2 Motion and structure recovery

Now we have got matches between two images, for the initial two images, there are polar geometry[10] constraint:

$$x_2^T E x_1 = p_2^T F p_1 = 0 \tag{1}$$

We can decompose the camera's motion R and T based on the estimated essential matrix E.

Now we know the transformation matrix between the two cameras, and the coordinates of each pair of matching points. The 3D reconstruction is to restore the coordinates of the matching points in space by using these information. In the previous derivation, we have the equation:

$$s_2 x_2 = K\left(R_2 X + T_2\right) \tag{2}$$

The above equation can't be solved directly, so it is transformed into a homogeneous equation:

$$\hat{x}_2 K\left(R_2 \quad T_2\right)\begin{pmatrix} X \\ 1 \end{pmatrix} = 0 \tag{3}$$

Use SVD decomposition to solve the zero space of the matrix on the left side of X, and then normalize the last element to 1, you can find X. This type of reconstruction is also known as triangulation.

Through the above method, we get the 3D point cloud of the initial two images. Then, the third image is added, the features are matched with the second image. Of these matching points, there must be a part of the matching points between image two and image one. That is to say, the spatial coordinates of some of these matching points are known, while at the same time knowing the pixel coordinates of these points in the third image. PnP (Perspective-n-Point) is a method for solving 3D to 2D matching point pair motion.

After obtaining the transformation matrix of the camera three by the above method to solve the transformation matrix, we also need to fuse the newly obtained spatial point with the previous 3D point cloud. There is no need to add space points that already exist, only the matching between image two and image three is added. Add an image and repeat the above method.

4.1.3 Bundle Adjustment

We use the Bundle Adjustment to optimize the camera pose and 3D coordinates. Its purpose is to optimize the projection matrix P_i and the three-dimensional space point X_{ij} and the feature point image coordinates $P_i X_{ij}$ after re-projection. If the image error is zero mean Gauss, the Bundle Adjustment method is a maximum likelihood estimator, and the cost function is as follows:

$$f\left(C, X\right) = \frac{1}{MN_T} \sum_{i=N-N_T+1}^{N} \sum_{j=1}^{M} r\left(x_{ij} - proj_i\left(X_j\right)\right)^2 \qquad \left(X_j \in X\right) \tag{4}$$

Among the equation, $M = |X|$, $C = \left\{R^i, T^i \middle| N - N_0 < i \le N\right\}$ is the set of pose parameters of the camera to be optimized, N_0 is the number of images of the camera pose to be optimized, considering the feature point projection of N_t images, X is the coordinate set of scene results to be optimized. X_{ij} is the two-dimensional coordinates of the j_{th} 3D point in the i_{th} image. $r(X)$ is 2-norm.

In the case that there are many image sequences, the scene structure is large, and the camera poses more, the algorithm complexity of the Bundle Adjustment method is too high. It is time consuming to perform Bundle Adjustment for all structures and poses after each new image reconstruction. Therefore, this paper process a bundle adjustment method based on local correlation frame. The so-called local association frame is a reconstructed image set that is closely related to the current newly added image frame.

Assume that the current scene already contains the number of images as N , the number of image frames to be optimized is N_t , and the local associated frame is M . When $N \le 20$, this paper uses the traditional bundle adjustment method to optimize, when $N > 20$, the bundle adjustment method based on the local correlation frame is used. Then have the following relationship:

$$N_t = \begin{cases} N & N \le 20 \\ M & N > 20 \end{cases} \tag{5}$$

4.2 Model recognition and camera tracking

The second part of the marker-free augmented reality system is to identity the model and calculate the pose of the current camera in real time. The online phase can be divided into the following steps:

1) Extracting SIFT features from the current frame of the live video stream.
2) The Flann[4] algorithm is used to match the current frame image feature points with the sparse point cloud model reconstructed in the offline stage to obtain the corresponding relationship between the 3D points and the 2D points.
3) Optimize the camera pose by RANSAC and Lenvenberg-Marquardt algorithm, and reduce the jitter of the parameters by using the result of the previous frame solution.

In order to track the 3D scene, we first need to build a k-d tree for the feature points in the 3D point cloud model reconstructed in the offline phase. K-d tree can help us search and match descriptors quickly. We perform SIFT feature extraction and descriptor calculating on the current frame image of the real-time video stream, and perform descriptor matching with the pre-established k-d tree. To improve the matching effect, we use Flann combined with k-nearest neighbor search[6] for feature matching.

Flann (Fast Library for Approximate Nearest Neighbors) is an abbreviation for Fast Nearest Neighbor Search Library, a collection of algorithm for nearest neighbor search for large data sets and high dimensional features. Since the descriptor of the SIFT feature point is a 128-dimensional high-dimensional vector, Flann can be used for fast matching. For each feature point in the current frame image, we perform a European distance k-neighbor search with the k-d tree. If the ratio of the nearest distance to the next closets distance is less than 0.8, we think this is a reliable match. Through the above method, we obtain the matching pair of 2-D image points and 3-D model points. We use a robust PnP method to calculate the camera pose corresponding to the current frame. The RANSAC algorithm is used to calculate the camera pose, which can eliminate unreliable 2D-3D matching and obtain robust calculation results.

5 Experiments

The experiments in this section run on a desktop system with the Intel® Pentium(R) CPU G4600 @ 3.60GHz × 4 and the graphics card GeForce GTX 1060. The input device of the system is Logitech HD camera C270, the input video resolution is 640*480. The virtual and real fusion adopts OpenGL API, and the output device is ordinary LCD.

First, we performed 3D reconstruction of 11 images for the offline phase. The reconstruction results of all 11 images are shown in Figure 3.

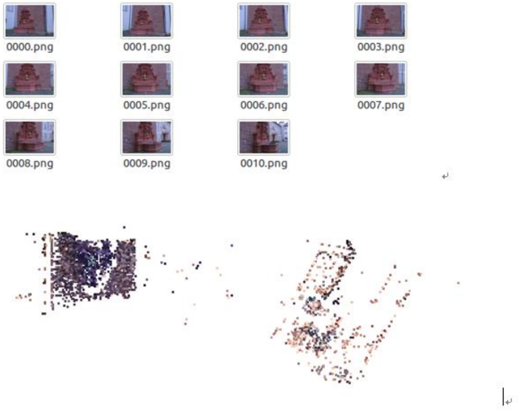

Figure 3. Reconstruction of 11 images

We calculate the re-projection error of all reconstructed 3D points on 2D image feature points. We use Bundle Adjustment to optimize the reconstruction results and calculate the re-projection error of the 3D points. The point cloud model after BA optimization is shown in Figure 4.

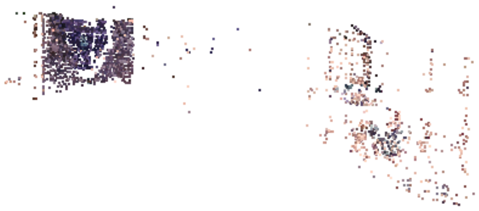

Figure 4. Point cloud model after BA

In order to evaluate the reconstruction effect, we compare the time and reprojection error with the best 3D reconstruction scheme Colmap.

Table 1. Reconstruction time and RMS error

	Time(minutes)	RMS
Colmap(gpu)	0.659	0.344747
Colmap(cpu)	17.954	0.344747
Ours(cpu)	0.841	0.971903

Table 1 shows the results of the comparison. In the case of using only cpu, the reconstruction time of this paper is greatly reduced, and the reprojection error is only a little worse than colmap, and is limited to one pixel.

Then you can combine virtual and real. The camera pose is calculated for the current frame of the input video sequence and the virtual content is superimposed. We modeled and rendered a teapot model for a desktop scene, at a rate of 3 frames/sec, as shown in Figure 5.

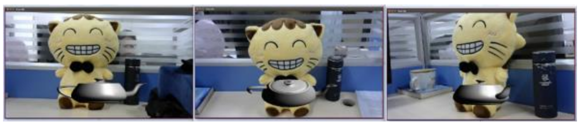

Figure 5. Desktop scene

6 Conclusion

This paper constructs a markerless augmented reality system based on point cloud model. We can recognize and track specific scenes from any viewpoint and then render the virtual content. Experiments show that ordered image sequence reconstruction can quickly and accurately reconstruct the scene model. The reconstructed model can be used to identify scene from any viewpoint, robustly solve camera pose and render virtual content.

References

1. H. Luo, T. Xue, and X. Yang. "Real-Time Dense Monocular SLAM for Augmented Reality", ACM on Multimedia Conference, 1237-1238, 2017
2. J. Mooser, S. You, U. Neumann, "Applying robust structure from motion to markerless augmented reality", IEEE Workshop on Applications of Computer Vision, 2009
3. I. Skrypnyk, David G. Lowe, "Scene modelling, recognition and tracking with invariant image features", ISMAR '04: Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality, 110-119, 2004
4. Marius Muja, David G. Lowe, "Fast Approximate Nearest Neighbors with Automatic Algorithm Configuration", International Conference on Computer Vision Theory and Applications (VISAPP'09), 331-340, 2009
5. David G. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 60(2):91-110, 2004
6. J.S. Beis, David G. Lowe, "Shape indexing using approximate nearest-neighbour search in high-dimensional spaces", CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition, 1000-1006, 1997

7. F. Morenonoguer, V. Lepetit, and P. Fua. “Accurate Non-Iterative O(n) Solution to the PnP Problem”, Proc.int.conf.on Computer Vision, 1-8, 2007
8. M. Fischler, R. Bolles, “Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography”, Communications of the Association for Computing Machinery, 24(6):381-395, 1981
9. K. Hirokazu, M. Billinghurst. “Marker Tracking and HMD Calibration for a Video-Based Augmented Reality Conferencing System”, IEEE and ACM International Workshop on Augmented Reality IEEE Computer Society, 85, 1999
10. R. Hartley, A. Zisserman, “Multiple View Geometry in Computer Vision”, Cambridge University Press, 2004
11. S. Agarwal, N. Snavely, I. Simon, “Building rome in a day”, International Conference on Computer Vision (ICCV), 54(10):72-79, 2009
12. J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016