3D Pose Recognition of Complex Objects by Monocular Vision

Ping Chen[1,a,*], Xin Xu[1,b] and Zhuangzhuang Chang[1,c]

[1]College of Mechanical Engineering, Chongqing University, Shazheng Street, Chongqing, People's Republic of China
a. chempion@126.com, b. 1241801349@qq.com, c. 791150845@qq.com
**corresponding author: Ping Chen*

Keywords: Vision measurement, image registration, AP clustering, 3D object recognition.

Abstract: In order to solve the problem of long computing time and low recognition accuracy of classical monocular recognition algorithm and deep learning algorithm in complex object recognition, a method of quickly recognizing the three-dimensional pose of an object is proposed. The method transforms the recognition of three-dimensional pose of object into the calculation of rotation angle of lens and angle of projection. The experimental results indicate that the rotation angle error measured by this method is not more than 0.3 degrees, the longitudinal and latitudinal mean error is not more than 0.2 degrees, and the recognition time is about several hundred milliseconds, which is almost independent of the complexity of the object.

1. Introduction

With the development of three-dimensional model acquisition technology, object recognition and pose estimation by monocular cameras attracts attention of scholars. There are three kinds of traditional recognition approaches: the first one is view-based approaches[1], this method is not suitable for objects with complex appearance. The second approach proposed for three-dimensional object recognition is based on three-dimensional geometry[2], this method is limited to some specific shape classes and its robustness of matching result is poor. The third approach is a descriptor-based method[3], the recognition speed is relatively slow and prone to error matching. Various deep neural networks also have been proposed for three-dimensional object recognition, but it is difficult to achieve high-precision recognition[4].

In industrial production, some complex objects often need to be detected, which generally have the characteristics of complex appearance, irregular shape, weak appearance texture, and even lack of obvious color features. In this paper, a method of quickly recognizing complex object's three-dimensional posture is proposed by establishing retrieval model based on AP clustering and by introducing FMT into the registration process to realize fast and high-precision recognition. This method can be applied to the application scenario where the object to be identified is located on a stable platform, thus laying the foundation for visual guidance-based robot grasping.

2. Four-tier Retrieval Model Based on AP Clustering

When an object is placed on a conveyor belt or a detection platform, it has a limited number of stable postures. Although these postures may change in a certain range due to various errors, the range is limited. A method which can be used in practical industrial application is provided in this paper. It only needs three-dimensional CAD model and projection angles of several possible stable postures as input.

2.1.Model Sampling

By uniformly sampling the stable postures of the object and its surrounding area on the Gauss sphere, at a certain angle-interval, the projection of the three-dimensional model under various possible postures is obtained, as shown in Figure 1. Among them, yellow triangles represent virtual cameras, blue spheres represent objects, λ and φ are longitudinal and latitudinal projection angles, θ_c is rotation angle of lens, and d is the distance from the center of the virtual camera lens to the center of the object.

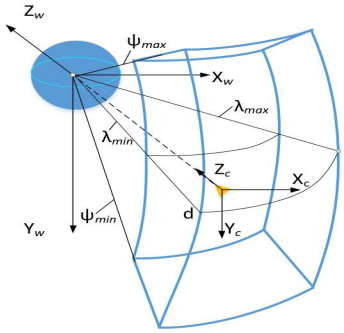

Figure 1: Uniform sampling on the Gauss sphere.

2.2.Generating Retrieval Model by Clustering

After the sampling of projection view is completed, the SIFT features of all views are extracted, saved in the form of SIFT feature matrix, and the projection angles expressed in latitude and longitude are recorded. Then, the SIFT descriptors are clustered by K-Means clustering algorithm to form a bag of words model containing K visual words[5]. The vocabulary frequency of each image is counted to form a K-dimensional histogram vector, the similarity between images is measured by the Euclidean distance between vectors. Meanwhile, in order to improve the efficiency of retrieval, a four-tier model is put forward in this paper, as shown in Figure 2. In order to generate retrieval model reasonably and efficiently, AP clustering method[6] is adopted to cluster these vectors representing images.

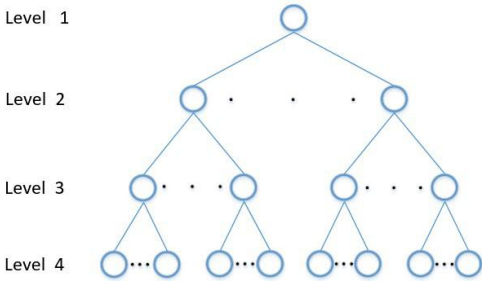

Figure 2: Four-tier retrieval model.

Images on the first level represent the stable postures. On the second and third level, there are representative images obtained by clustering in the next level. On the fourth level, there are the projection images of the stable postures and all the projection images obtained by changing the projection angle within a small area nearby.

Firstly, an $N \times N$ similarity matrix S is generated, and the clustering process is realized by information transmission between images. In the iterative process, the representative images are selected as a suitable clustering center. The goal is to maximize the sum of the similarities of all the images and the nearest representative images. In this process, *a(i,k)* and *r(i,k)* are continuously transmitted between nodes.

The *a(i,k)* reflects the suitability of image i choosing k as representative image. The *r(i,k)* reflects the suitability of image k as representative image of i. The values of *a(i,k)* and *r(i,k)* are expressed as:

$$r(i,k) = s(i,k) - \max_{k \neq k'} \{a(i,k') + s(i,k')\} \tag{1}$$

$$a(i,k) = \begin{cases} \min\{0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\{0, r\{i',k\}\}\}, i \neq k \\ \sum_{i \neq k} \max\{0, r(i',k)\}, i = k \end{cases} \tag{2}$$

The *a(i,k)* and *r(i,k)* of all images updated continuously through iterative process until they converge to produce several representative images that satisfy the requirements. At the same time, the rest of the images are allocated to the corresponding clustering.

3. Pose Recognition Method for Complex Objects

After the retrieval model is formed, match the photograph with retrieval model layer by layer to find the correct matching result. Compared with the projection view obtained from sampling, the camera may rotate an uncertain angle around its physical center. The value of this angle must be measured to determine the real pose of the object.

In this paper, FMT is adopted in the registration process to transform rotation and scaling in Cartesian coordinate system into translation in a new coordinate system. According to rotation angle and scaling factor, the photograph is rotated and scaled. Then, the photograph corrected by rotation and scaling is sequentially matched with the fourth layer image represented by the template image for the SIFT features to obtain the numbers of matching points. The RANSAC[7] algorithm is used to filter the matching results and the image with the most matching points with the photograph taken by the camera after RANSAC screening is the final matching result.

4. Experiment and Result

The matching time is almost independent of the complexity of the object. The main factors influencing the matching time are the number of three-dimensional objects and the number of possible stable postures.

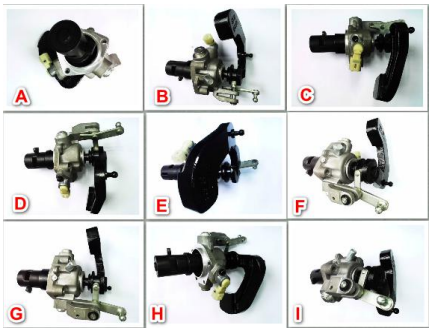

Figure 3: Nine postures of a shift tower for automobiles.

In order to validate the method proposed in the paper, the shift tower shown in Figure 3 is taken as the experimental object. An industrial camera is mounted on a six-axis robot in the mode of eye-in-hand to take photos. The photos are cropped to the same size as the projection view.

As is shown in Figure 3, the shift tower shown in has nine stable postures, therefore, level 1 contains nine pictures. For each stable posture, a total of 1681 pictures are extracted at the sampling interval of 0.1° in the latitude and longitude range of ±2°.

It should be noted that the formation of the retrieval model requires clustering all projection views together. In this process, the dictionary size of the bag of words model, which is the K value, has a great influence on the correctness of the retrieval model.It is taken as the criterion to select the current K value if the projection angles of all level 4 pictures belonging to the same level 1 are all near the same stable posture.

The posture of the lens relative to the object is obtained by controlling the posture of the robot, which is compared with the posture of the object calculated by the method of this paper to determine the precision of the method in this paper. Experiments indicate that the matching time is about 800 milliseconds. The mean error of pose information in longitude and latitude direction is not more than 0.2 degrees, the error of rotation angle of the lens is less than 0.3 degrees. The position information of the object can also be obtained by phase correlation. The recognition results under different illumination conditions are obtained by controlling the brightness of LED light source, which confirms that this method has good robustness to illumination change.

5. Conclusions

A high robust recognition algorithm for complex three-dimensional object's pose by monocular camera is proposed in this paper. The SIFT feature descriptor with good versatility and stability is adopted to form the bag-of-word model. Taking full advantage of the characteristics of AP clustering, a four-tier retrieval model is formed by multi-tier clustering. The problem of matching with large angle rotation is solved by combining FMT. This method is simple, effective and practical, and the accuracy and speed of pose recognition are better than traditional algorithms.

Acknowledgments

This research was funded by Chongqing Science and Technology Major Theme Project, grant number cstc2018jszx-cyztzxX0032.

References

[1] Ulrich, M., Wiedemann, C., Steger, C. (2012) Combining Scale-Space and Similarity-Based Aspect Graphs for Fast 3D Object Recognition. IEEE Trans. Pattern Anal. Mach. Intell, 34(10), 1902-1914.

[2] B Schnabel, R., Wahl, R., Klein, R. (2007) Efficient RANSAC for Point-Cloud Shape Detection. Comput. Graph. Forum, 26(2), 214-226.
[3] Lowe, D.G. (2004) Distinctive Image Features from Scale-Invariant Keypoints. Int'l J. Computer Vision, 60, 91-110.
[4] Billings, G., Johnson-Roberson, M. (2019) SilhoNet: An RGB Method for 6D Object Pose Estimation. Computer Vision and Pattern Recognition.
[5] Josef, S., Andrew, Z. (2003) Video Google: A Text Retrieval Approach to object Matching in Videos. 9th IEEE International Conference on Computer Vision, NICE, FRANCE, OCT 13-16; IEEE COMPUTER SOC, LOS ALAMOS, USA, 1470-1477.
[6] Brendan, J., Delbert, D. (2007) Clustering by Passing Messages Between Data Points. Science, 315, 972-976.
[7] Fischler, M.A., Bolles, R.C. (1981) 'Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography', Commun. ACM, 24, (6), pp. 381–395.