International Journal of Computer Applications Technology and Research
Volume 12–Issue 04, 43-46, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1204.1010

A Pose Estimation Method Combining Instance Segmentation and Point Pair Features

Yu Xin
College of Communication Engineering
Chengdu University of Information Technology
Chengdu China

Hao Peng
College of Communication Engineering
Chengdu University of Information Technology
Chengdu China

Abstract: For the traditional pose estimation method based on point pair features, both the preprocessing of scene point cloud and the construction of point pair features of scene point cloud have serious time-consuming problems, which cannot meet the needs of actual industry. Therefore, this paper proposes a pose estimation method that combines instance segmentation and point pair features. Therefore, this paper proposes a pose estimation method that combines instance segmentation and point pair features. First, use the Mask R-CNN-based instance segmentation network to obtain the location of the target object in the two-dimensional image of the scene; then, obtain the local point cloud data of the space where the target object is located from this position and the depth information of the scene; finally, the local the point cloud data is used as the scene point cloud based on point pair feature pose estimation to perform feature matching with the point cloud of the target object.

Keywords: point cloud data; point pair features; Mask R-CNN; pose estimation

1. INTRODUCTION

3D feature descriptors are key factors in 3D pose estimation, they encode the relationship between point cloud data into low-dimensional feature vectors. Global feature descriptors are constructed by using the overall geometric information of the model, such as Shape distribution feature descriptors, Spherical harmonics feature descriptors, SPR feature descriptors, etc. However, global feature descriptors are less stable in the presence of occlusions and simple object shapes. The local feature descriptor describes the local structure and shape information of the target object through the domain relationship and normal vector of the 3D point cloud, such as the 3DSC feature descriptor, the FPFH fast point feature histogram, and the SHOT feature descriptor. However, in the subsequent feature matching, local feature descriptors need to perform seriously time-consuming hypothesis verification, which will cause the matching efficiency to become very low. In order to make full use of the respective advantages of the global feature descriptor and the local feature descriptor, Drost et al. proposed the point pair feature descriptor PPF (Point Pair Feature), whose main idea is to extract point pair features from the point cloud of the target object and the scene, The pose of the object in the scene point cloud is estimated according to the relationship between the point pair features and the pose combined with the Hough voting strategy. However, the pose estimation algorithm based on PPF also has certain defects, such as being susceptible to background and noise interference and a large amount of calculation when voting on the scene point cloud. Improved PPF algorithms for everyday objects in cluttered scenes by Birdal et al. and Hinterstoisser et al. Wu et al. used the PPF algorithm to perform robotic trash bin picking and achieved a recognition rate of 93.9%. Choi et al. introduced boundary points with directions and boundary line segments into an algorithm for estimating planar industrial objects, which performed a higher recognition rate and faster speed than traditional PPF. In this paper, a point-to-feature pose estimation algorithm combined with instance segmentation is proposed. This method only selects the local point cloud data containing the range of the target point cloud as the matching scene point cloud for pose estimation, which not only reduces the number of points also predicts in advance the spatial extent of the point cloud of the target object.

2. METHOD

2.1 Original PPF Algorithm

Drost et al. used "global description, local matching" to describe the algorithm, so we can know that the algorithm can be divided into two stages. The first stage is for the model to train the global feature descriptor of the target object offline, that is, the point pair feature descriptor. The second stage is to match the feature description of the target global point pair trained in the first stage with the scene point cloud, establish a local coordinate system from the relationship between the point pair features, and calculate the three-dimensional rigid body transformation of the target object in the scene point cloud. Then, by voting on all possible pose transformations, the pose of the target object is obtained, and finally the obtained pose is optimized to complete the pose estimation of the target object in the scene point cloud. Figure 1 shows the technical flow chart of pose estimation based on point pair features.

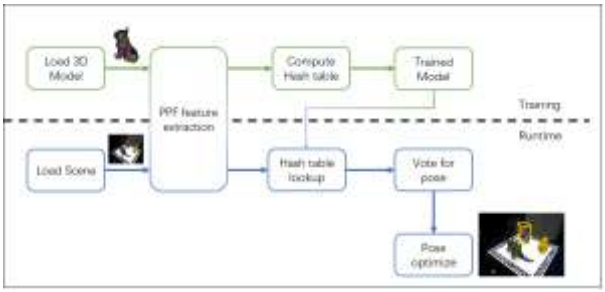

Figure. 1 Original PPF algorithm process

2.2 Our Algorithm

In this paper, we propose a method combining deep learning for instance segmentation and point-to-feature based voting. This method is divided into two stages. In the first stage, the advanced Mask R-CNN network is used to find the target

International Journal of Computer Applications Technology and Research
Volume 12–Issue 04, 43-46, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1204.1010

object in the scene RGB image, and the type ID and area of the target object are returned. Then, in the second stage, the detected target object is located. The region adopts point-to-feature voting method to obtain the 3D pose of the target object. This method in this chapter combines the advantages of the two methods. The instance segmentation network based on deep learning can quickly filter out the data in complex real scenes, so that the search space is reduced when feature matching, and the method of point-to-feature voting is retained. Recover the robustness of the target pose. Figure 2 shows the network flow diagram of the method proposed in this paper.

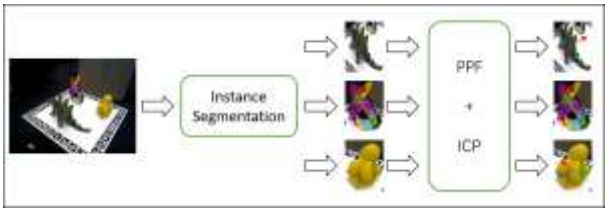

Figure. 2 Our algorithm process

2.2.1 Instance Segmentation Stage

Our data set comes from the HomebrewedDB data set of the BOP challenge competition. This data set is mainly used for 6D pose estimation of 3D objects. This data set contains a total of 33 model objects and 13 real scenes composed of these 33 model objects. As shown in Figure 3, the target sizes in the scene pictures in this dataset are almost the same, so it is necessary to enhance the data of scene target object instance segmentation before training the Mask R-CNN network.

Figure. 3 HomebrewedDB dataset scene images

Since the HomebrewedDB dataset has the target 3D model point cloud data, it is possible to obtain pictures from different angles of the target through the 3D point cloud data of the target, and then randomly paste these pictures from different angles into the VOC2017 dataset according to different proportions. On the picture, as shown in Figure 4.

Figure. 4 Synthesize objects containing model objects on the VOC2017 dataset

2.2.2 Pose Estimation

After segmenting the scene point cloud by instance, we obtain the local point cloud data of the space where the target object is located, and then calculate the matching between the point pair features of this local point cloud and the global feature description of the target object. In order to ensure that the error in calculating the point pair feature descriptor is small, we use the method of consistent normal vectors to process the target object and local point cloud, as shown in Figure 5.

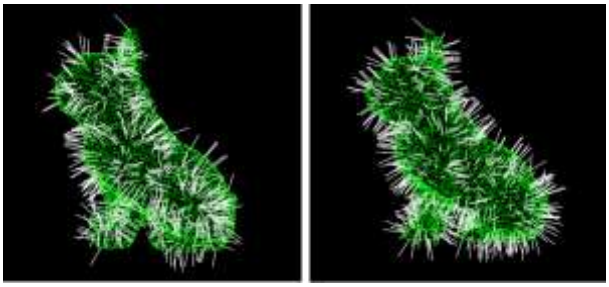

Figure. 5 Consistency processing of model point cloud normal vectors

3. EXPERIMENT

3.1 Experimental Environment

This experiment was implemented in the Pytharm development tool in the Windous 11 operating system, with Python version 3.8. The deep learning framework is Python, version 1.13.1, corresponding to CUDA version 11.6. The experimental GPU is NVIDIA GeForce RTX 3060, and point cloud data is processed using OpenCV library version 4.6 and PCL library version 1.12.1.

3.2 Instance Segmentation Results

After using the Mask R-CNN network to train the cat, dinosaur, and rabbit models in the HomebrewDB dataset, 340 scene images were tested in the HomebrewDB dataset. The recognition accuracy P and recall R of the three target objects were calculated, and the recognition effect was comprehensively evaluated using the harmonic mean F, as shown in Table 1.

Table 1. Mask R-CNN training results

Model	P	R	F
Cat	61.19%	72.21%	66.70%
Dinosaur	69.34%	78.52%	73.93%
abbit	67.14%	73.68%	70.41%
Total	65.89%	74.80%	70.34%

The target recognition results for some example images are shown in Figure 6, which intuitively shows that the target recognition position matches the actual position of the image, indicating that this training result can effectively recognize the position of the target.

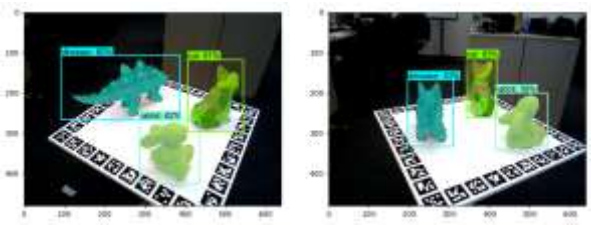

Figure.6 Target object recognition results

International Journal of Computer Applications Technology and Research
Volume 12–Issue 04, 43-46, 2023, ISSN:-2319–8656
DOI:10.7753/IJCATR1204.1010

3.3 Pose Estimation Results

Compare the efficiency of target pose estimation of traditional point to feature and point to feature target pose estimation based on instance segmentation in the real environment. First, use the same Downsampling scale coefficient of 20% for model and scene point cloud to Downsampling them, respectively process the scenes in HomebrewdDB dataset in the form of traditional filtering and instance segmentation, and then input them into the improved point to feature pose estimation method in Chapter 3 Perform point to point feature calculation on three rabbit models, and verify their average error ADD after matching with the scene. If the average error is less than 0.05, it is determined that the matching is successful. Finally, the recognition rates and recognition time consumption of these three models in 340 real scene point clouds from the HomebrewedDB dataset were calculated, as shown in Table 2.

Table 2. Comparing the traditional PPF algorithm with our algorithm on three models

Model	Drost-PPF		Our	
	R	Time (s)	R	Time (s)
Cat	0.615	17.3	0.690	5.5
Dinosaur	0.560	34.5	0.727	7.7
Rabbit	0.673	13.2	0.672	4.1

At the same time, compared with other algorithms that also test the target model pose estimation on the HomebrewedDB dataset, these algorithms have different advantages in recognition accuracy and recognition efficiency. The algorithm in this paper has obvious advantages in target pose estimation efficiency and accuracy. The improvement, especially the improvement in efficiency is very obvious, as shown in Table 3.

Table 3. Comparing the traditional PPF algorithm with our algorithm on three models

Method	Data Type	R	Time (s)
Drost-PPF	D	0.615	20.7
CRT-6D	RGB	0.603	3.8
PointVoteNet2	RGB-D	0.556	21.9
Pix2Pose	RGB-D	0.711	10.5
Our	RGB-D	0.693	6.3

The following visually shows the 3D target recognition and pose estimation effect of the algorithm proposed in this paper. The red point cloud is the pose of the target in the scene, as shown in Figure 7.

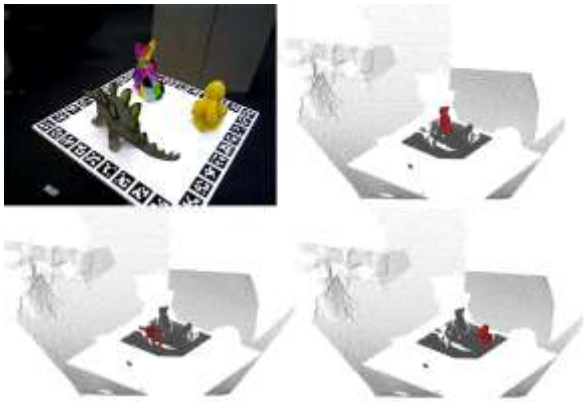

Figure. 7 Model pose estimation results

4. CONLUSION

This paper proposes a pose estimation algorithm based on Mask R-CNN network instance segmentation after processing the target object in the scene. Compared with the traditional pose estimation algorithm based on point pair features, this algorithm greatly reduces the input in real scenes. The number of scene point clouds; at the same time, when constructing the point pair feature between the target object and the scene point cloud, ensure that the normal vector direction of the point cloud is consistent. The method in this paper effectively solves the problem of low efficiency and accuracy of pose estimation due to the huge amount of scenic spot cloud data in the real scene.

5. REFERENCES

[1] Osada R, Funkhouser T, Chazelle B, et al. Shape distributions[J]. ACM Transactions on Graphics (TOG), 2002, 21(4): 807-832.

[2] Kazhdan M, Funkhouser T, Rusinkiewicz S. Rotation invariant spherical harmonic representation of 3 d shape descriptors[C]//Symposium on geometry processing. 2003, 6: 156-164.

[3] Wahl E, Hillenbrand U, Hirzinger G. Surflet-pair-relation histograms: a statistical 3D-shape representation for rapid classification[C]//Fourth International Conference on 3-D Digital Imaging and Modeling, 2003. 3DIM 2003. Proceedings. IEEE, 2003: 474-481.

[4] Frome A, Huber D, Kolluri R, et al. Recognizing objects in range data using regional point descriptors[C]//Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part III 8. Springer Berlin Heidelberg, 2004: 224-237.

[5] Rusu R B, Blodow N, Beetz M. Fast point feature histograms (FPFH) for 3D registration[C]//2009 IEEE international conference on robotics and automation. IEEE, 2009: 3212-3217.

[6] Tombari F, Salti S, Di Stefano L. Unique signatures of histograms for local surface description[C]//Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part III 11. Springer Berlin Heidelberg, 2010: 356-369.

[7] Drost, B.; Ulrich M.; Navab N.; Ilic S. Model globally, match locally: Efficient and robust 3d object recognition. In Proceedings of the 2010 IEEE Conference on

Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13-18 June 2010; pp. 998-1005.

[8] Birdal T.; Ilic S. Point pair features based object detection and pose estimation revisited. In Proceedings of the 2015 International Conference on 3D Vision (3DV), Lyon, France, 19-22 October 2015; pp. 527-535.

[9] Hinterstoisser, S.; Lepetit, V.; Rajkumar, N.; Konolige, K.; Going further with point pair features. In Proceedings of the European Confer ence on Computer Vision (ECCV 2016), Amsterdam, The Netherlands, 8-16 Octorber 2016; pp. 834-848.

[10] Wu, C.H.; Jiang, S.Y.; Song, K.T. CAD-based pose estimation for random bin-picking of multiple objects using a RGB-D camera. In Proceedings of the 2015 15th International Conference on Control, Automation and Systems (ICCAS), Busan, South Korea, 13-16 October 2015; pp. 1645-1649.

[11] Choi, C.; Taguchi, Y.; Tuzel, O.; Liu, M.Y. Voting-based pose estima tion for robotic assembly using a 3d sensor. In Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA), Saint Paul, MN, USA, 4-18 May 2012; pp. 1724-1731

[12] Castro P, Kim T K. CRT-6D: Fast 6D Object Pose Estimation with Cascaded Refinement Transformers[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 5746-5755.

[13] Sundermeyer M, Hodan T, Labbe Y, et al. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects[J]. arXiv preprint arXiv:2302.13075, 2023.

[14] Park K, Patten T, Vincze M. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7668-7677.