Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Scene Regions Guided Pose Estimation Using an Improved Voting Method in Cluttered Scenes*

Guohui Tian, Zhengwei Jia
School of Control Science and Engineering
University of Shandong
Jinan, Shandong Province, China
g.h.tian@sdu.edu.cn

***Abstract*— In the process of service robot performing home services, one of the key parts is robotic grasping. Meanwhile, accurate object pose estimation is essential for grasping. In home environments, estimating the poses of household textureless objects simply and effectively in cluttered and occluded scenes is challenging. This paper proposes a method by using the color information of object to extract the 3D scene regions where the object may exist. Point Pair Features voting approach is applied to obtained voting array in extracted 3D scenes. Then a novel votes adjustment method is proposed to recalculate the voting number to reduce the effects of occlusion. Our algorithm is evaluated on Linemod Occluded dataset and the experimental results show that the proposed algorithm can effectively improve the accuracy of pose estimation when there is object occlusion in a cluttered scene and improve the ranking of correct pose in candidate poses. Meanwhile, the average calculation time is shortened.**

***Index Terms*— 6D pose estimation; point pair feature; scene region extraction; service robot; robotic grasping**

I. INTRODUCTION

With the aggravation of aging, the research on home service robots has begun to attract more and more attention. When the robot manipulates the object, the 6D pose of the object needs to be obtained in the world coordinate system, so the pose estimation of the object is indispensable.

Six-dimensional pose estimation of object has been extensively studied. Initially, the pose of the object was restored from a single RGB image using local features, but these methods did not work well for textureless objects in the home. Point Pair Feature(PPF)[6] method calculates the pose of the object in the 3D scene without using color information, so it can cope with textureless objects. However, PPF has low accuracy and long calculation time.

There are many cluttered scenes in the home environment where objects are occluded. The occluded object has fewer features, while there are rich environmental features in a cluttered scene. These characteristics make it difficult to match the target object features in the scene. PPF method had been improved in [8] and [9] to improve performance in cluttered and objects occluded environments. However, these two methods still restore the pose of the object in the entire scene, so the calculation cost was relatively large. Convolutional neural network(CNN) was used to locate the location of an object in [1], thereby reducing the size of the 3D scene. But, CNN requires a large amount of tagged data for training, and it does not perform well on unfamiliar objects or scenes. These characteristics make CNN have higher cost of use and poor adaptability.

The purpose of this study is to construct an easy-to-use and adaptable object pose estimation system. Important things are how to extract precise object feature and how to increase target object's information proportion in cluttered scenes. Therefore, we use the region where the target object may exist in the scene to guide the pose estimation, and at the same time we adopt an effective votes number calculation method to increase the accuracy of the result.

In order to reduce the proportion of background information in the scene, the color information of the object is used to extract the scene regions where the object may exist. This method can be used in a variety of scenes and occlusion situations. To further eliminate the effect of occlusion, a novel pose votes adjustment is applied to increase the number of votes on the occluded target object.

The contributions of this paper are as follows:

- We propose an easy-to-use and adaptable method to extract the regions of the scene where the target object may exist. This can reduce the cost of calculation and increase the data proportion of target object.
- We propose a novel method of candidate pose votes adjustment, which can reduce the impact of occlusion.

The rest of this paper is organized as follows. In Section II, we introduce the related work. Section III describes the pipeline of our object pose estimation algorithm. In Section IV and V, our method of region extraction and votes adjustment will be introduced. In Section VI, we explain the results of verification experiments of our algorithm. Finally, we conclude in Section VII.

*This work is Supported by National Key R&D Program of China(2018YFB1307101), and National Natural Science Foundation of China(U1813215)

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

II. Related Work

With the development of deep learning, the method of deep learning began to be applied to the object 6D pose estimation. PoseCNN[2] proposed a two stage, three branch convolutional neural network to estimates the 3D translation of an object. Tekin et al.[3] is a single-shot approach by using CNN. In that study, CNN predicted the eight corner point and a center point of object's 3D bounding box, and transformation matrix was calculated using a PnP algorithm. These methods can give a good result, but these are critically dependent on training data so that these methods are currently not very adaptable and easy to use.

Linemod[4][5] has been extended a lot of work because of well performance in cluttered scenes. This method selected image gradients and surface normals as features to match templates. First, this method created many templates for the color gradient and surface normal vector on the object 3D model at various angles. Templates was used to match the scene during the detection process to obtain a rough pose of the object. Then, ICP(Iterative Closest Point) was used to refine the rough pose. However, Linemod does not handle the occlusion very well.

PPF(Point Pair Feature) has been extensively studied in industry. Drost et al.[6] first proposed PPF and subsequently made some improvements[7]. [6] used PPF to match similar point pairs between 3D scene and 3D object by using efficient Hough-like voting. This method can handle texture-less object and cluttered scenes. In the work [8] and [9], strategy of sampling and voting schemes was studied to improve performance in cluttered scenes. Besides, [8] studied a variety of candidate pose screening methods, so that it can deal with occlusion better. Kiforenko et al.[10] performed a detailed performance evaluation of the Point Pair Features and the results of [11] showed that the performance of PPF is excellent. However, these previous studies matched point pairs in the whole 3D scene and applied multiple refine methods, which made the algorithm more complex and computationally intensive.

In paper [1], the author studied a method for estimating the pose of a particular industrial component by using a neural network to locate the position of object. This method can obtain the exact position of the object, greatly improving the accuracy of the pose estimation. But there are many kinds of objects and scenes in the home environment, and there may be strange objects. In this case, using neural networks to locate the target objects will result in higher costs. Because, neural network or deep learning does not have strong generalization performance for strange environments and objects. So, our study attempts to improve the adaptability and reduce the complexity of algorithm.

III. Pipeline of Our Object Pose Estimation

Fig. 1 shows the pipeline of our method. First, we extract the color points from the 3D model of the object, and use the machine learning to calculate the color clusters. Scene regions are extracted by using these clusters separately, and the extracted multiple scene regions are combined to obtain the final scene mask. The mask is used to scope the 3D scene to reduce unimportant data of scene. Then we can match the pose of the object on the reduced 3D scene.

The 3D model of the object and the reduced 3D scene need to be downsampled, which not only reduces the data density and reduces the amount of calculation, but also retains the key information. The downsampling method we use is similar to [9]. Due to the different sized objects, using a fixed sampling step size during downsampling does not accommodate the characteristics of different objects. So, we set a fixed factor to calculate the sampling step size for different objects. But now there is another problem. If a object has a flat part, the point density of the plane part is the same as other places after uniform downsampling. When there is a plane part in the real scene, it is easy to produce a plane mismatch. In order to reduce the number of points in the insignificant region, we increase the downsampling factor and cluster in each step according to whether the angle between the point normal vectors is less than 30 degrees. The points in the same cluster are used as similar points. In addition, the sampling step size of the scene is the same as that of the model.

When calculating point pair features on a scene, if the distance between the two points is greater than the diameter of the object, the current point is skipped directly, as in [9]. But this in turn leads to another problem. In the extracted 3D scene, if there is a plane or area where the points are concentrated, and there are fewer reference points on the object or the item is occluded. In this case, if the PPF is matched with the object diameter as the search radius, the reference point on the object will get fewer votes than the reference point on area with more points. Therefore, we have designed a method for correcting the number of votes, as shown in Fig. 1, we will describe this method in detail later.

After calculating a pose for each reference point in the scene, the poses are clustered according to the similarity within the set translation and rotation thresholds, as in [15], and the total number of votes for each cluster is recalculated. The combined pose is output from large to small according to the number of votes.

IV. Scene Regions Extraction based on Color Clustering

Currently, there are many methods of image segmentation to obtain a region of interest, such as CNN based method, color based method and edge detection based method. Among them, the method of convolutional neural network has received extensive attention, such as 2D object detection

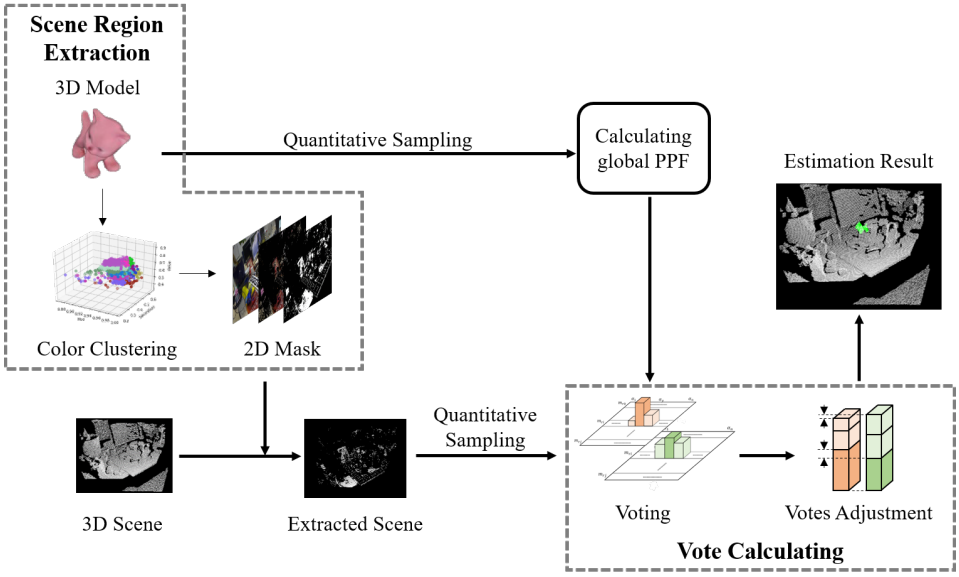


Fig. 1: Complete structural diagram of our method.

and semantic segmentation based on Fast R-CNN[12]/R-CNN[13]. But these methods of deep learning require a lot of support with labeled data and will generate a large training cost. Therefore, deep learning based method cannot be easily applied to variety home environment with a wide variety of household goods.

Although we study the pose estimation method of textureless objects in home, there are still differences in color information between objects and between objects and the environment. Meanwhile, color based scene region selection can be applied to various environments and objects. Compared with the deep learning method, color based method is more flexible and has a stronger generalization.

We select the HSV color space to represent the color data of object model. Because, RGB color space is susceptible to lighting and shadows and not suitable for digital processing. In Fig. 2, we show the distribution of the RGB and HSV color spaces of two objects from the Linemod Occluded dataset[14][5]. As can be seen from the figure, the distribution of the RGB points is linear, and the points with similar colors exhibit a narrow distribution, which is disadvantageous for extracting color features. In contrast, the distribution of points in the HSV color space is more concentrated.

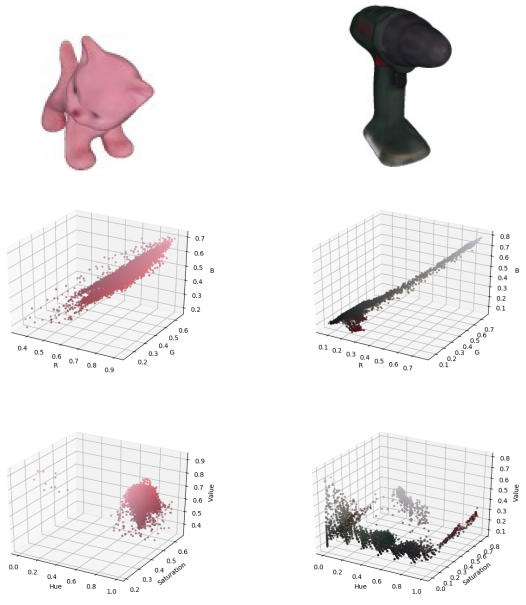
Fig. 2b shows that color of object is not purely uniform, so only counting the threshold of HSV will not produce satisfactory results. Therefore, we apply a K-means algorithm to cluster the colors of object to get multiple clusters. The result of clustering Cat, Driller colors is shown in Fig. 3. We set the number of clusters to 16, so that the points in the cluster can be brought together better while including all the object points. To cope with the noise of the model color, we designed a simple filter to filter out the points that are most likely to be noise, as show in Fig. 3a. Compared with Fig. 2a, the remote points are filtered out.

However, the color of captured object in the actual scene is easily affected by the illumination and the shadow. Therefore, the color threshold obtained by pure object color clustering does not satisfactorily meet the requirements of scene extraction, as show in Fig. 4b. This method cannot completely extract the region where the target object is located. Therefore, we set a coefficient of expansion to deal with these problems. We expand the thresholds of the Hue, Saturation, and Value components of each cluster separately. Since Hue is a loop variable, if the threshold is more than one period after expansion, another cluster is added, so the number of clusters may exceed the set number. Fig. 4c shows the result of region extraction by using expanded thresholds. In this region, we can apply point pair feature to match the object 6D pose.

V. Votes Adjustment

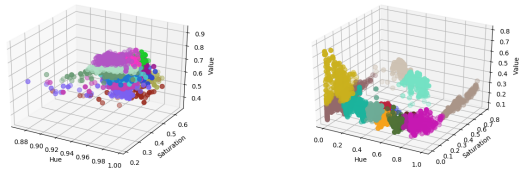
When we calculate point pair features on 3D scenes, we only select points that distance is less than the object diameter from the reference point, same as [9]. This reduces unnecessary time costs. However, this still does not solve the problem of high misrecognition rate when the effective three-dimensional information is relatively low when the object is occluded, as shown in Fig. 5a. In this scene, we display the first eight poses according the votes in different colors, and we can see that no pose matches the actual position of the object in the scene.

There are more 3D points belonging to a plane in the extracted scene region. Therefore, when points in the plane are used as the reference points to calculate the point pair



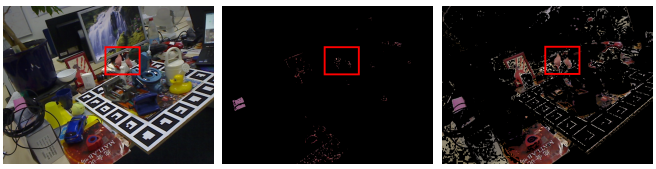
(a) Cat (b) Driller

Fig. 2: Two objects color space distribution. First row: Two objects 3D model; Second row: RGB color space, the coordinate axes R, G and B represent red channel, green channel and blue channel, respectively, their values range from 0 to 1; Final row: HSV color space, hue, saturation and value are in range from 0 to 1.



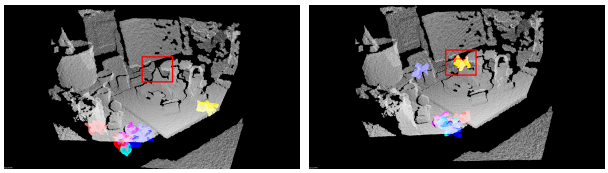
(a) Cat (b) Driller

Fig. 3: K-means clustering results. Different clusters are represented by different colors. (a): Clustering result of Cat, some points are filtered out. (b): Clustering result of Driller.



(a) Scene (b) No expansion (c) With expansion

Fig. 4: Comparison of the results of different clustering methods. (a): The target object Cat is partially occluded in this cluttered scene. (b): The result of image region extraction when cluster expansion is not used. (c): The result of image region extraction after cluster expansion.



(a) No pose matched (b) Two poses matched

Fig. 5: Pose estimation results before and after the votes adjustment is used. (a): The first eight estimated poses are mismatched to the plane. (b): There are two poses that match to the correct position, so the final pose can be selected by using Iterative Closest Point(ICP) algorithm.

features with the radius that is equal to object diameter, the voting number of the pose with the highest voting number will be greater than the voting number of selected pose calculated by the reference points lie on the object, as show in Fig. 6. So, we need to increase the number of votes in the correct pose.

We propose a new way to calculate the number of votes. A two dimensional accumulator array is created for voting. Each row represents a reference point m_r of model and each column represents a discretized rotation angle α. Adjacent rotation angles in the same reference point have similarities in the pose of the object. Therefore, the number of votes of adjacent rotation angles can be used as a reference for the final number of votes.

The pose P, with highest number of votes N_y, is calculated on the extracted region by using method of [6]. The model reference point corresponding to the selected pose is m_{ri}, and the corresponding α angle is α_y. The number of votes for the angles α_{y-1} and α_{y+1} at the reference point m_{ri} are N_{y-1} and N_{y+1}, respectively. So, the number of votes N_P of pose P is calculate by using (1). If α_y is α_0 or α_n , (2) or (3) can be used.

Compared with the previous results, after using the new method proposed by us, two of the first eight poses sorted by number of votes successfully match the true pose of the target object, as show in Fig. 5b.

$$N_P = N_y + N_{y-1} + N_{y+1} \tag{1}$$

$$N_P = 2 * N_y + N_{y+1} \tag{2}$$

$$N_P = 2 * N_y + N_{y-1} \tag{3}$$

VI. Experiments

A. Dataset

Linemod Occluded dataset was used to test our algorithm. This dataset included 8 textureless objects and 1214 actual scene images and depth maps. Different objects had different degrees of occlusion in different scenes. And each scene had a ground pose of the object used to test the performance

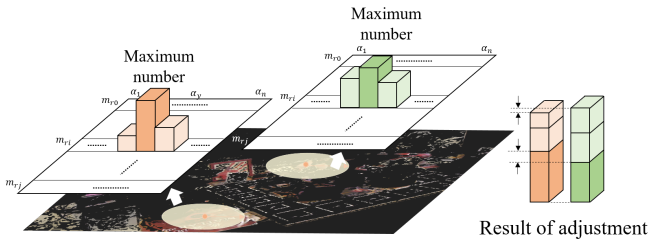


Fig. 6: An example of votes adjustment. The number of votes for the object obtained at the left reference point is greater than that of the right reference point. However, when using our calculation method, the number of votes on the right reference point is greater than the number on the left. The height of the three green columns is greater than the height of the three orange columns.

of the algorithm. The object pose was represented by a four-dimensional matrix that represented the transformation relationship from the model coordinate space to the scene coordinate space. The color information used in the scene region extraction came from the object model, and the Red, Green, Blue components ranged from 0 to 1.

B. Parameters setting

In this experiment, all the images in Linemod Occluded dataset were tested and the program parameters on all test pictures were the same. In the process of object color clustering, the expansion factors of the H, S, and V color components were set to 0.2, 0.2, and 0.12, respectively. In addition, the downsampling factor of the object 3D model was 0.1, and the downsampling factor of the 3D scene was $(0.1 * D_m)/D_s$. Where D_m was the diameter of the object and D_s was the diagonal length of the scene. One of every 5 points in the scene after downsampling was used as a reference point to calculate the point pair feature. And, the rotation angle α was divided into 45 parts by 2π.

Our algorithm was implemented in C++ and OpenMP was used to parallel our program. The configuration of the computer used for the experiment was Intel Core i9-7900X CPU and 64G RAM.

C. Results

We evaluated all result by using ADD metric[5] without considering the case of symmetry and the similar situation of different views of object. The evaluation metric is as shown in (4). Where x represents the point coordinates on the three-dimensional model, and $[R|T]$, $[\hat{R}|\hat{T}]$ represent the true transformation matrix and the estimated transformation matrix, respectively. When m is less than $0.1 * D_m$, the result is considered correct.

$$m = \underset{\mathbf{x}\in\mathcal{M}}{avg} \|(\mathbf{R}\mathbf{x} + \mathbf{T}) - (\tilde{\mathbf{R}}\mathbf{x} + \tilde{\mathbf{T}})\| \tag{4}$$

The three experimental results is shown in Table I. The first experiment was used the basic Point Pair Feature method as described in section III without region extraction and votes adjustment. Only region extraction method was used in second experiment while region extraction and votes adjustment both were applied in the last experiment. All experiments did not use the ICP algorithm.

In the result, it is obvious that the introduction of the scene extraction method greatly improves the accuracy of the results in the Top1 and Top10 candidate poses and greatly reduces the calculation time than basic PPF method, especially on textureless object such as Ape, Can, Cat, Duck. Because basic PPF method matches the pose of the object across the cluttered 3D scene, the background information affects the accuracy of the results and the calculation time becomes more longer. Driller and Glue have a relatively large number of textures, so the accuracy of such objects in the experiment is not obvious, but the calculation time has decreased. Although Eggbox has less texture as Ape, it is an object that contains many similar parts. When we use the VSD(Visible Surface Discrepancy)[11] metric, Eggbox's Top1 and Top10 pose estimation accuracy can reach 39.74% and 46.81%, respectively, in the case of using the scene extraction method.

In addition, the weight of the correct pose of our method in the first 200 poses of the candidate poses is higher than the basic PPF method, and the gap has a tendency to become larger, as show in Fig. 7. These indicate that our method can better increase the votes proportion of target object and advance the ranking of the correct pose. So, fewer candidate poses will be verified when we use the screening method.

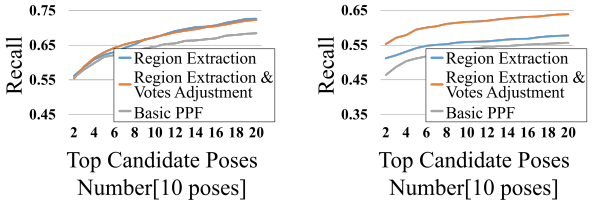
It can be seen in Fig. 8 that the growth of the recall rate in the first 200 candidate poses is accelerated after the votes adjustment method is used. In addition, the result of the votes adjustment method in Fig. 7b is higher than the other two methods. This is because the proposed method of votes adjustment can further increase the proportion of correct pose votes in the cluttered and occluded scenes.

VII. Conclusion

In this paper, a novel method is proposed to estimate household objects by using scene regions guidance and votes adjustment in cluttered scenes. The color information of the object is used to extract the scene regions in which the target object exists. Then, the 3D pose of the target object is calculated using the votes adjusted point pair feature method on the extracted 3D scene regions. Our algorithm is evaluated on all scenes of Linemod Occluded dataset. The experimental results show that the scene region guidance method can effectively improve the pose estimation accuracy and reduce calculation time of the textureless objects in the cluttered environment and the method of votes adjustment can improve the ranking of the correct pose in the candidate poses of some objects.

TABLE I: Experiment Result. Top1 indicates the correct rate of the first pose. Top10 indicates the ratio of correct poses in the top 10 poses. Time represents the average time used for each scene. Number is the number of tested scenes for the corresponding object.

	Basic PPF			Region Extraction			Region Extraction & Votes Adjustment			
Object	Top1	Top10	Time/s	Top1	Top10	Time/s	Top1	Top10	Time/s	Number
Ape	28.03%	53.68%	12.64	**60.68%**	68.89%	0.79	59.66%	**69.40%**	**0.77**	1170
Can	34.30%	58.24%	7.37	52.44%	75.89%	0.43	**56.01%**	**76.72%**	**0.42**	1207
Cat	20.72%	33.70%	3.85	**38.84%**	46.42%	**0.76**	37.74%	**47.01%**	0.77	1187
Driller	35.09%	51.65%	3.24	**36.16%**	**52.80%**	**1.82**	34.43%	51.15%	1.83	1214
Duck	17.15%	36.40%	9.85	**30.27%**	**47.24%**	0.88	27.56%	46.72%	**0.76**	1143
Eggbox	9.02%	43.40%	7.17	9.62%	46.21%	2.70	**11.83%**	**52.94%**	**2.25**	1175
Glue	**3.99%**	9.52%	7.08	2.99%	**11.41%**	5.02	1.99%	7.97%	**5.01**	903
Hole Puncher	53.14%	74.13%	9.93	**60.25%**	**77.36%**	5.40	54.63%	73.72%	**5.36**	1210
Average	25.18%	45.09%	7.64	**36.41%**	**53.28%**	2.23	35.48%	53.20%	**2.15**	-



(a) Driller (b) Eggbox

Fig. 7: The trend of the correct pose ratio in the first 20-200 candidate poses of all the tested scenes.

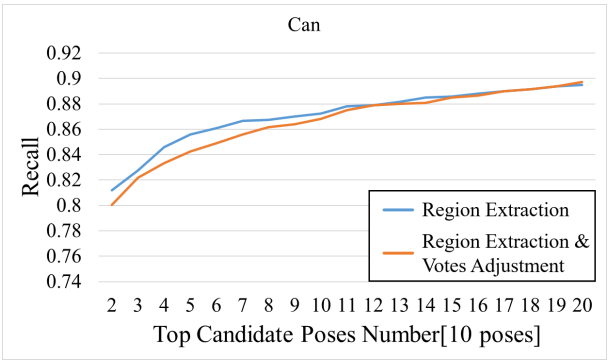


Fig. 8: The comparation between scene extraction and vote adjustment & scene extraction method in the ratios of the correct poses in the first 20-200 candidate poses of all the tested scenes.

REFERENCES

[1] D. Liu et al., "2D Object Localization Based Point Pair Feature for Pose Estimation," 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO), Kuala Lumpur, Malaysia, 2018, pp. 1119-1124.

[2] Y. Xiang, T. Schmidt, V. Narayanan and D. Fox, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes," arXiv preprint, arXiv:1711.00199, 2017.

[3] B. Tekin, S. N. Sinha and P. Fua, "Real-Time Seamless Single Shot 6D Object Pose Prediction," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 292-301.

[4] S. Hinterstoisser et al., "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," 2011 International Conference on Computer Vision, Barcelona, 2011, pp. 858-865.

[5] S. Hinterstoisser et al., "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," 2012 Asian Conference on Computer Vision, Berlin, Heidelberg, 2012, pp. 548-562.

[6] B. Drost, M. Ulrich, N. Navab and S. Ilic, "Model globally, match locally: Efficient and robust 3D object recognition," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 2010, pp. 998-1005.

[7] B. Drost and S. Ilic, "3D Object Detection and Localization Using Multimodal Point Pair Features," 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission, Zurich, 2012, pp. 9-16.

[8] J. Vidal, C. Lin and R. Martł, "6D pose estimation using an improved method based on point pair features," 2018 4th International Conference on Control, Automation and Robotics (ICCAR), Auckland, 2018, pp. 405-409.

[9] S. Hinterstoisser, V. Lepetit, N. Rajkumar and K. Konolige, "Going Further with Point Pair Features," 2016 Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 834-848.

[10] L. Kiforenko, B. Drost, F. Tombari, N. Krger and A. G. Buch, "A performance evaluation of point pair features. Comput," Computer Vision and Image Understanding, vol. 166, 2018, pp. 66-80.

[11] T. Hodan et al., "BOP: Benchmark for 6D object pose estimation," 2018 European Conference on Computer Vision (ECCV), 2018, pp. 19-34.

[12] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1440-1448.

[13] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, 2014, pp. 580-587.

[14] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton and C. Rother, "Learning 6d object pose estimation using 3d object coordinates," 2014 European Conference on Computer Vision (ECCV), 2014, pp. 536-551.

[15] T. Birdal and S. Ilic, "Point Pair Features Based Object Detection and Pose Estimation Revisited," 2015 International Conference on 3D Vision, Lyon, 2015, pp. 527-535.