Efficient 3D object tracking approach based on convolutional neural network and Monte Carlo algorithms used for a pick and place robot

Y. Zhang, C. Zhang, R. Nestler, M. Rosenberger, G. Notni

ABSTRACT

Currently, Deep Learning (DL) shows us powerful capabilities for image processing. But it cannot output the exact photometric process parameters and shows non-interpretable results. Considering such limitations, this paper presents a robot vision system based on Convolutional Neural Networks (CNN) and Monte Carlo algorithms. As an example to discuss about how to apply DL in industry. In the approach, CNN is used for preprocessing and offline tasks. Then the 6-DoF object position are estimated using a particle filter approach. Experiments will show that our approach is efficient and accurate. In future it could show potential solutions for human-machine collaboration systems.

Keywords: image processing, 3d object detection, robot vision, deep learning, pick and place robot, 3D tracking

1. INTRODUCTION

In recent years, Deep Learning (DL) shows us powerful capabilities for image processing. But it cannot output exact photometric process parameters, such as the 6-DoF object position, and it shows non-interpretable results. Due to such reasons this paper presents a robot vision system based on Convolutional Neural Networks (CNN) and Monte Carlo algorithms. As an example to discuss about how to apply DL in industry. A robust, efficient and accurate 3D object detection system could show potential solutions for human-machine collaboration systems in future. The approach solves the 3D object detection problem in real time 6-DoF pick-and-place operations for a wide variety of objects in clutter. First thanks to PointNet++, keypoints of target objects are extracted from their template point clouds. Since the difficulty and time cost of point cloud analysis a CNN based on YOLO architecture is used to segment the scene point cloud. Then outliers will be removed using a cluster analysis based on Euclidean distance. After such preprocessing objects are tracked using a particle filter approach. Distance of nearest point, color feature and geometric feature are used as three criteria for Maximum Likelihood Estimation to find the 6-DoF position of objects. An experiment was designed to verify, due to the new approach the difficulty within point cloud analysis can be faster and easier realized. Another experiment will show that the three criteria used for Maximum Likelihood Estimation are essential. In the end the efficiency (20 fps) and accuracy (mean error of position: 3.57 mm and orientation: 0.103 rad) of 6-DoF position detection and success rate (97.3%) of object grabbing in a concrete pick-and-place application will be shown.

2. RELATED WORK

In the past 20 years the research on object detection can be divided into two categories: traditional methods and novel neural network approach. Many researches are on manually defined feature or descriptor for 2D image analysis, e.g.: [1][2][3]. For the field of 3D feature someone has made such studies: [4][5][6][7]. Image processing based on multispectral images is generally used for the analysis of Remote Sensing image [8][9]. Others someone have studied on edge detection based on multispectral image analysis [10]. All the above studies are on the features or descriptor. Then with the help of feature map and some functional modules (such as classifier SVM) can be some functional applications realized: e.g.: [11][12][13].

In recent years due to the rise of neural network development many researches based on this novel method have emerged. [14][15] are the researches on 3D object detection based on RGB-D image processing. For 3D object detection tasks can be solved by the neural network based on bird view from Lidar view or point cloud from stereo camera [16][17][18].

It is worth mentioning that three papers of Ross Girshick [19][20][21]. In these papers he studies to combine the traditional method with neural network to solve the task of 2D object detection. Then in the researches [22][23] his idea has extended to more efficient methods. Inspired by them, in this case we also discuss to perform the 3D object detection based on classic method and neural network approach.

Photonics and Education in Measurement Science 2019, edited by Maik Rosenberger, Paul-Gerald Dittrich, Bernhard Zagar, Proc. of SPIE Vol. 11144, 1114414
 · doi: 10.1117/12.2530333

3. METHOD OVERVIEW

In this section we explain the 3D object detection system based on deep learning approach and classic method. This system is to support a pick and place robot. Due to the combination of the two approach the difficult problems of 3D image processing are solved.

Because the computational cost of point cloud processing is large. As shown in Figure 1 our system consists of three functional modules: (1) Target point cloud creation, (2) Scene point cloud segmentation and (3) 6DoF object detection. The first two modules are offline and online preprocessing, respectively. Some deep learning approach are used in these two modules. Then the third module is the core functional block. In this module we have used classic method to find the exact photometric process parameters (6DoF object position).

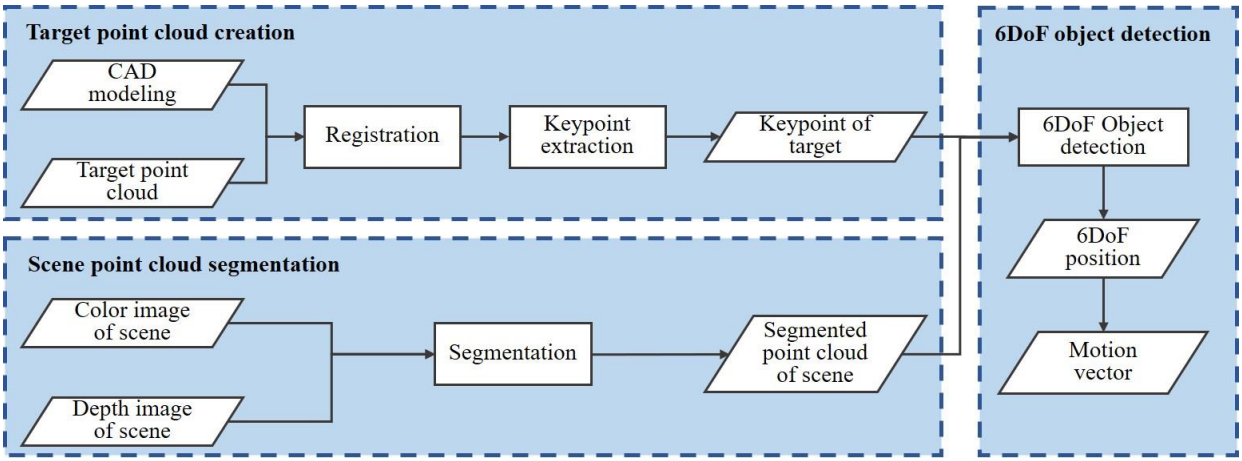

Figure 1. 3D object detection system overview

3.1 Target point cloud creation

First, the target point cloud can be obtained by CAD modeling or by a 3D sensor. The ideal point cloud can be easily created by CAD modeling. But the disadvantage with this approach is that the real color information cannot be obtained. Because the quality of color recording by the sensor is dependent by many factors such as light intensity, reflection, diffraction etc. It is very difficult to simulate this process. By 3D sensor the real color information can be obtained. However, for the whole 360 degrees panoramic point cloud cannot be obtained by a single 3D sensor. To solve this problem an expensive 3D camera group is necessary. For these reasons we present a new approach to create the target point cloud.

First, we use a single camera from different view point to get multiple point cloud with color information. Then we use the ideal colorless point cloud from CAD as a template. With the help of this template the colored point clouds are aligned. Then a voxel-grid-based filter is used to remove the redundant information generated during the aligning process.

Finally, we use the PointNet++ [24] approach to extract the keypoints from the target point cloud. With this operation the computation cost to object detection will be reduced and the result will be better.

3.2 Scene point cloud segmentation

The important two aims of this part are: (1) Classification of objects in scene, (2) Rough object localization (bounding box). Thanks to YOLO2 [22], based on the analysis of color images the object will be segmented from scene. Then only the pixels in the bounding box will be converted to point cloud, of course with classification information. Finally, outliers can be segmented by Euclidean clustering, as shown in figure 2.

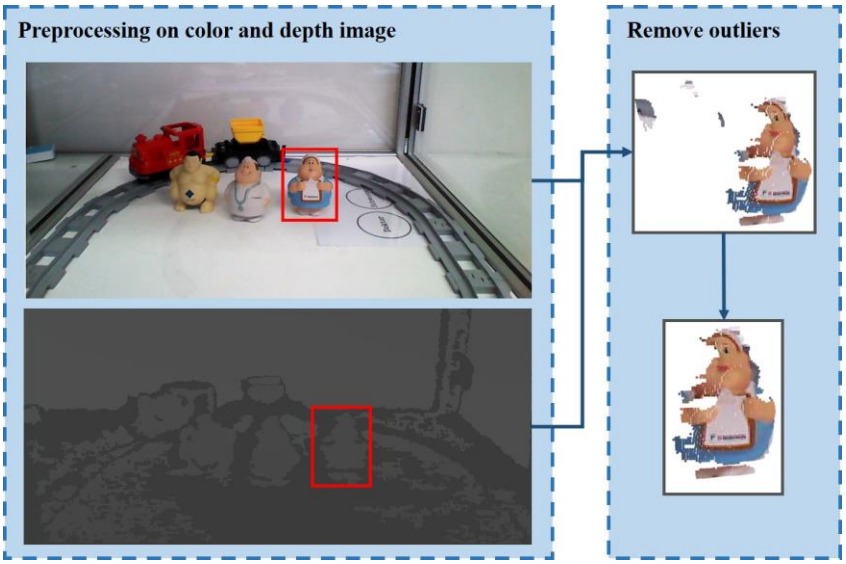

Figure 2. 2D, 2.5D and 3D Preprocessing

3.3 6DoF object detection

As introduced in the first chapter, with deep learning the exact photometric process parameters cannot be output. Here we used a classic method (particle filter approach [25][26][27]) to detect 6DoF object position. Particle filter approach is an efficient method to object tracking. But for the better result the quality of prior information and the criteria used for Maximum Likelihood Estimation are essential. The second module (Scene point cloud segmentation) support the prior information of high quality.

For the criteria for likelihood estimation we have following discussion. The normal vector as a basic geometric feature does not have the invariance for rotation, so it is not used here. The computation cost for 3D descriptor such as SHOT [7] is usually too larger and unable to meet real time requirements. Finally, distance of nearest point and curvature are chosen as the spatial and geometric criteria and HSV color as the color feature for Maximum Likelihood Estimation to find the 6-DoF position of objects.

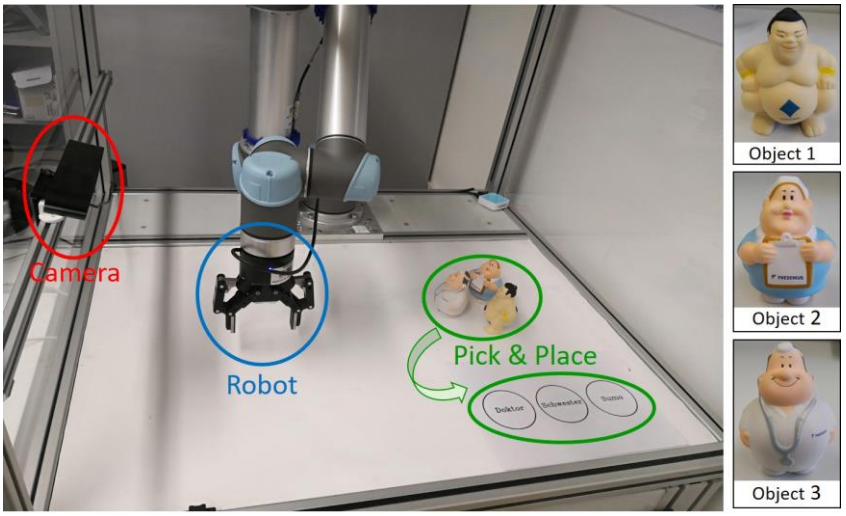

Figure 3. Experiment environment

4. EXPERIMENTS

Our experiment environment is based on a pick and place robot application. Due to hand eye calibration approach the position and orientation (pose) of the camera frame with respect to the robot frame is identified. Then the pose of the object in the camera frame is obtained by our 3D object detection system. Finally, the pose of the robot grasper for target object grasping is calculated. As shown in figure 3, we have three different rigid body objects as test object.

4.1 Optimization with the help of preprocessing

In this experiment the average deviation of the object position predicted by our system in the first 100 frames with and without preprocessing are separately tested. In this way, the importance of preprocessing will be shown. This test will be performed at 25 different locations. Gaussian distributions with variance of (σ = 0.01 m for position prediction, σ = 0.4 rad for orientation prediction) are used as the sampling range for particle resampling. And number of samples (particles) is defined in five different scales (200-10000).

Table 1. Mean error of position (m) and (rad) orientation with and without preprocessing (prep.)

Number of samples		**200**	**500**	**800**	**1k**	**3k**	**5k**	**10k**
Mean error of position prediction	**With prep.**	10.65	8.43	3.87	3.85	**3.57**	3.76	3.65
	Without prep.	23.63	24.52	23.94	23.87	16.68	15.78	**13.24**
Mean error of orientation prediction	**With prep.**	0.235	0.164	0.116	0.106	0.104	**0.103**	0.104
	Without prep.	0.873	0.826	0.727	0.733	0.724	0.635	**0.524**

Table 2. Time expenses (ms) for object detection operation with and without preprocessing (prep.)

Number of samples		**200**	**500**	**800**	**1k**	**3k**	**5k**	**10k**
Time expenses for preprocessing		15.63	15.63	15.63	15.63	15.63	15.63	15.63
Time expenses for object detection	**With prep.**	15.25	19.65	24.98	30.45	54.25	106.48	225.24
	Without prep.	86.54	146.58	198.76	259.14	406.75	594.64	1356.21

The result is shown in table 1. First, the test result show that the prediction results of the system is much increased due to the preprocessing. Secondly, there is no longer a significant error reduction when the particle number is above 1000. So too larger resampling is not recommended for better efficiency.

The time expenses of computation with and without preprocessing is also recorded in this experiment. Due to the larger scale segmentation with the help of preprocessing the efficiency increase is much greater than the time expenses of preprocessing itself. As shown in table 2. The importance of preprocessing is verified once again.

4.2 Success rate of object grabbing

In this experiment same three different objects are used. The whole experimental environment is divided into 9 areas. The success rate of the capture in each area is tested. The result is shown in table 3.

Table 3. Success rate of grabbing

Success rate			
Pick		Place	
Object 1	99.0%	Object 1	98.7%
Object 2	98.8%	Object 2	98.3%
Object 3	96.8%	Object 3	92.2%
Average	98.2%	Average	96.4%
Total	97.3%		

5. CONCLUSION

With the help of two deep learning approach (PointNet++, YOLO) the efficiency, stability and robustness are greatly improved. A method that deep learning in industrial image processing applied is proposed: preprocessing (segmentation) and offline tasks (keypoint extraction). Classic method such as particle filter approach is used to compensate for the disadvantage of deep learning. Finally, traditional methods and novel deep learning combine to bring the optimal performance for this 3D object detection system.

6. ACKNOWLEDGMENT

The author would like to acknowledge the help from Prof. Gunther Notni, Dr. Maik Rosenberger and Mr. Chen Zhang.

7. REFERENCES

[1] Dalal, N., and Triggs, B., "Histograms of oriented gradients for human detection." In international Conference on computer vision and Pattern Recognition (CVPR'05), 886-893 IEEE Computer Society (2005).
[2] Scovanner, P., Ali, S., and Shah, M., "A 3-dimensional sift descriptor and its application to action recognition." In Proceedings of the 15th ACM international conference on Multimedia, 357-360 ACM. (2007).
[3] Gupta, S., and Mazumdar, S.G., "Sobel edge detection algorithm." International journal of computer science and management Research 2.2: 1578-1583 (2013).
[4] Rusu, R.B., Blodow, N., Beetz, M., "Fast Point Feature Histograms (FPFH) for 3D Registration." Robotics and Automation, ICRA'09. IEEE International Conference, 3212-3217 (2009).
[5] Rusu, R.B., Blodow, N., Marton, Z.C. and Beetz, M., "Aligning point cloud views using persistent feature histograms." Intelligent Robots and Systems, IROS'08. IEEE/RSJ International Conference, 3384-3391 (2008).
[6] Rusu, R.B., Marton, Z.C. Blodow, N., and Beetz, M., "Persistent Point Feature Histograms for 3D Point Clouds." Proc 10th Int Conf Intel Autonomous Syst (IAS-10), Baden-Baden, Germany, 119-128 (2008).
[7] Tombari, F., Salti, S. and Stefano, L.D., "Unique signatures of histograms for local surface description." European 61 conference on computer vision. Springer, Berlin, Heidelberg, 356-369 (2010).
[8] Eugenio, F., Marcello, J. and Martin, J., "High-resolution maps of bathymetry and benthic habitats in shallow-water environments using multispectral remote sensing imagery." IEEE Transactions on Geoscience and Remote Sensing 53.7: 3539-3549 (2015).
[9] Chang, Y., Yan, L., Fang, H. and Luo, C., "Anisotropic spectral-spatial total variation model for multispectral remote sensing image destriping." IEEE Transactions on Image Processing 24.6: 1852-1866 (2015).
[10] Rosenberger, Maik. "Multispectral edge detection algorithms for industrial inspection tasks." 2014 IEEE International Conference on Imaging Systems and Techniques (IST) Proceedings, 232-236 (2014).
[11] Cai, Z. and Chao, S., "Implementation of a 3D ICP-based scan matcher." University of Freiburg, Tech. Rep (2010).
[12] Llorca, D. F., Arroyo, R., and Sotelo, M. A., "Vehicle logo recognition in traffic images using HOG features and SVM. " In 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), 2229-2234 (2013).
[13] Pang, Y., Yuan, Y., Li, X. and Pan, J., "Efficient HOG human detection." Signal Processing 91.4, 773-781 (2011).
[14] Kehl, W., Manhardt, F., Tombari, F., Ilic, S., and Navab, N., "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. " In Proceedings of the IEEE International Conference on Computer Vision, 1521-1529 (2017).
[15] Lahoud, J., and Ghanem, B., "2d-driven 3d object detection in rgb-d images. " In Proceedings of the IEEE International Conference on Computer Vision, 4622-4630 (2017).
[16] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J., "3d shapenets: A deep representation for volumetric shapes. " In Proceedings of the IEEE conference on computer vision and pattern recognition, 1912-1920 (2015).
[17] Zhou, Y., and Tuzel, O., "Voxelnet: End-to-end learning for point cloud based 3d object detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4490-4499 (2018).

[18] Qi, C. R., Liu, W., Wu, C., Su, H., and Guibas, L. J., "Frustum pointnets for 3d object detection from rgb-d data." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 918-927 (2018).
[19] Girshick, R., Donahue, J., Darrell, T., and Malik, J., "Rich feature hierarchies for accurate object detection and semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, 580-587 (2014).
[20] Girshick, R., "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, 1440-1448 (2015).
[21] Ren, S., He, K., Girshick, R., and Sun, J., "Faster r-cnn: Towards real-time object detection with region proposal networks." In Advances in neural information processing systems, 91-99 (2015).
[22] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., "You only look once: Unified, real-time object detection." In Proceedings of the IEEE conference on computer vision and pattern recognition, 779-788 (2016).
[23] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg, A. C., "SSD: Single shot multibox detector." In European conference on computer vision, 21-37 Springer, Cham. (2016).
[24] Qi, C. R., Su, H., Mo, K., and Guibas, L. J., "Pointnet: Deep learning on point sets for 3d classification and segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 652-660 (2017).
[25] Anderson, H.L., "Monte Carlo and the MANIAC. " Los Alamos Science 14, 96-108 (1986).
[26] Del Moral, P., "Non-linear filtering: interacting Particle resolution." Markov processes and related fields 2.4, 555-581 (1996).
[27] Li, T., Fan, H., and Sun, S., "Particle Filtering: Approach, and Application for Multitarget Tracking." Acta Automatica Sinica, (2015).