Proceedings of the 2016 IEEE International Workshop on
Advanced Robotics and its Social Impacts
Shanghai, China, July 8-10, 2016

An Improved Approach for Model-based Detection and Pose Estimation of Texture-less Objects

Haoruo Zhang[1], Yang Cao[2], Xiaoxiao Zhu[1], Masakatsu G. Fujie[2], Qixin Cao[1]

Abstract— **Detection and pose estimation of texture-less objects still faces several challenges such as foreground occlusions, background clutter, multi-instance objects, large scale and pose changes to name but a few. In this paper, we present an improved approach for model based detection and pose estimation of texture-less objects, LINEMOD [4], in order to improve the robustness of pose estimation with partial foreground occlusions. For template creation, we modify Gradient Response Maps and propose Gradient Orientation Maps, where Non-Maximum Suppression and Dual Threshold Algorithm are applied. And we adopt image pyramid searching method for fast template matching. Next, the approximate object pose associated with each detected template is used as a starting point for fine pose estimation with Iterative Closest Point algorithm. Thirdly, we improve the accuracy of fine pose estimation by using point cloud filter. Experimental results show that our approach is more robust to estimate the pose of texture-less objects with partial foreground occlusions.**

I. Introduction

Detection and pose estimation of 3D object is of great importance to many higher level tasks, including robotic intelligent manipulation and assembly to name but a few. The common of these applications is the requirement of recognizing and accurately localizing known objects, in order that these objects can be operated by robot end-effector. And it is also a well-studied problem in computer vision, there still remain several challenges such as foreground occlusions, background clutter, multi-instance objects, large scale and pose changes. In the last few years, the main focus in the field of detection and pose estimation of 3D objects has been limited to those objects with abundant texture features. And the key is the use of a sparse representation of local features such as SIFT [1], SURF [2], ORB [3] and some recent proposals. Recently, people have started to focus on the task of detection and pose estimation of texture-less or texture-poor objects. And texture-less and uniformly colored objects occur frequently in robotic applications.

This work has been supported by National Natural Science Foundation of China (Grant No. 61273331) and YASKAWA Electric Corporation.

[1]Haoruo Zhang, Xiaoxiao Zhu and Qixin Cao are with State Key Lab of Mechanical Systems and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University (e-mail: zhr@sjtu.edu.cn)

[2]Yang Cao and Masakatsu G. Fujie are with Faculty of Science and Engineering, Waseda University.

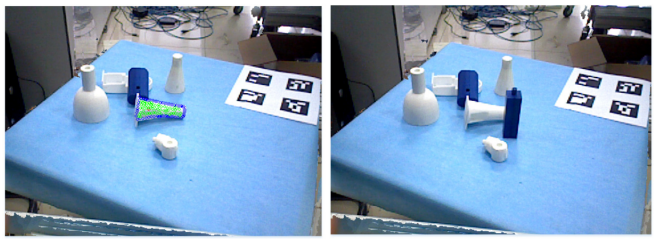

Figure 1. The effectiveness of LINEMOD method with partial foreground occlusions. **Left:** The target object that is not occluded can be detected in the scene. **Right:**The target object that is partialy occluded cannot be detected in the scene.

In the field of texture-less object detection and pose estimation, model-based or template-based techniques are superior, e.g. [4], [5], [6]. According to the method proposed by Hinterstoisser et al.[6] that is called LINEMOD, this model-based technique can be made very fast and output the accurate pose of target object. But it is not enough robust to clutter and occlusions as well as changing lighting conditions. The LINEMOD method can perform well with the dataset of Hinterstoisser et al. but it has less robustness with partial foreground occlusions and the effectiveness is presented in Fig.1. In practice, accuracy localization of target object partially occluded is a common task for robotic intelligent manipulation and assembly.

In addition, there is an alternative model-based technique to 3D object detection proposed by Drost et al. [8]. This method need to describe sampled pairs of 3D points form scene and vote for corresponding object pose hypotheses. And the final pose estimation will be refined with ICP that is the same with LINEMOD method. Some similar methods e.g. [9], [10], [11], [12] adopt edge features. But the recognition capability of these method is lower than LINEMOD method, and the efficiency and performance depend directly on the complexity of current scene, which could limit real-time performance. Thus, the model-based technique has some problems need to be solved. In this paper, we will show how to overcome these problems.

Our work aims at proposing an improved approach for model-based detection and pose estimation of texture-less objects which can mostly perform well with partial foreground occlusions. And the input to this approach consists of RGB-D images provided by consumer-level RGB-D cameras such as Kinect and Xtion PRO, which provide aligned color and depth images that capture both the appearance and geometry of current scene simultaneously.

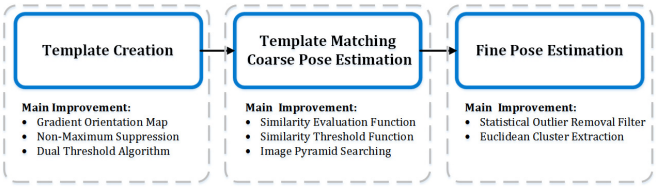

Figure 2. The main process and improvement of our proposed approach.

Our approach is based on these ideas presented in [4], [5], [6] and improves them so that it can complete the task of estimating the 6D pose objects with partial foreground occlusions. As for LINEMOD method, object templates are represented by a set of carefully selected feature points in RGB-D images instead of a raw image. And features in templates are quantized and represented as bit-vectors. And this feature map can be called Gradient Response Map (GRM) in [4]. In this way, the template matching can perform quickly with binary operations. Then, coarse pose estimation can be achieved by comparing GRM of template and scene. The best matching template and optimal position in the scene can be found, and it will be used as coarse pose estimation or a starting point for refinement with Iterative Closest Point (ICP) algorithm [7]. And it also proposes data structures optimization for fast template matching by using special SSE hardware instructions. Therefore, LINEMOD is capable of real-time matching of thousands of templates. And we use similar matching procedure in our approach, as shown in Fig.2.

The **main contribution** of our work consists of three parts. Firstly, we propose an improved feature maps that is called Gradient Orientation Maps (GOM) as binarized representation of template and scene. And the similarity evaluation function will also be improved. In this way, the output of this improved approach can be more robust with partial foreground occlusions. Secondly, because the similarity evaluation function between template and scene has higher computational cost by using GOM than GRM, we adopt image pyramid searching method to improve real-time performance during template matching. Thirdly, as for refinement of pose estimation with ICP, we improved the accurate location of target objects with partial foreground occlusions by using point cloud filter.

II. Proposed Approach

In this section, we will first describe how to create templates of texture-less target object with Gradient Orientation Maps. Then we will discuss the improved process of template matching and coarse pose estimation, how to detect target object in the scene and how we acquire coarse pose estimation which is a starting point for fine pose estimation. Finally, we will address our point cloud filter to further increase the precision of pose estimate with Iterative Closest Point algorithm.

A. Template Creation

The template of target object must have discriminant features so that objects can be detected easily. We will adopt image gradient orientations because they proved to be more discriminant than other forms of representation and robust to illumination change and noise. As for texture-less objects, image gradients are often the only reliable image cue. Without using actual values of gradient magnitudes, we only consider the orientation of the gradients. In order to increase robustness, we need to compute gradient magnitude of RGB image on R, G, and B channel separately at each location $p(r,c)$. And we use gradient orientation of the channel whose gradient magnitude is largest. In order to reduce computational cost, we quantize the gradient orientation, and divide the orientation space into n equal spacings. Then we can get the binarized representation of each gradient orientation spacing. We compute the magnitude and orientation of gradient at each location with its 3 × 3 neighborhood. We also keep only the gradients whose magnitude is larger than a default threshold, and these gradients can describe the edge feature of target objects and current scene. The binarized representation of gradient orientation at location $p(r,c)$ on template T can be formalized as:

$$biori(T, p(r,c)) = BI\{maxori(T, p(r,c))\} \tag{1}$$

$$maxori(T, p(r,c)) = ori(\max_{E\in\{R,G,B\}} \|\nabla E(p)\|_2) \tag{2}$$

Where, $biori(T, p(r,c))$ is binarized representation of gradient orientation in the template is image T at location p ; $maxori(T, p(r,c))$ is the gradient orientation of channel whose magnitude is largest; BI is binarized encode operation; $\nabla E(p)$ is the gradient of each channel in image at location p. We take the orientation angle between 0 and 180 deg and allow it to correctly handle object occluding boundaries. It will not be affected whether the target object is over a bright background or dark ground.

Above techniques are similar to GRM in LINEMOD method. Then, GRM will spread the orientations around their locations to obtain a new representation of the original image and pre-computing response maps. The similarity evaluation function in LINEMOD method is also based on these response maps. But in practice, the output is not robust enough to detect target objects with partial foreground occlusions. In our improved approach, we won't spread the orientations around their locations, but will adopt Non-Maximum Suppression method and Dual Threshold Algorithm to thinning edge and get the optimal edge of target object and current scene. And the edge is composed of binarized representation of gradient orientation at each location p. This technique is motivated by canny edge detection approach proposed by J. Canny et al. [13]. And such feature map which has the optimal edge can be called Gradient Orientation Maps (GOM). Fig. 3 shows the plain introduction about it.

The Non-Maximum Suppression (NMS) method is used to solve the problem of local maximum searching, where a local maximum is greater than all its neighbors excluding itself. For a given n, the neighborhood of any pixel consists in the 1D case of the n pixels to its left and right side and in the 2D case of the quadratic region centered on the pixel under consideration [14].

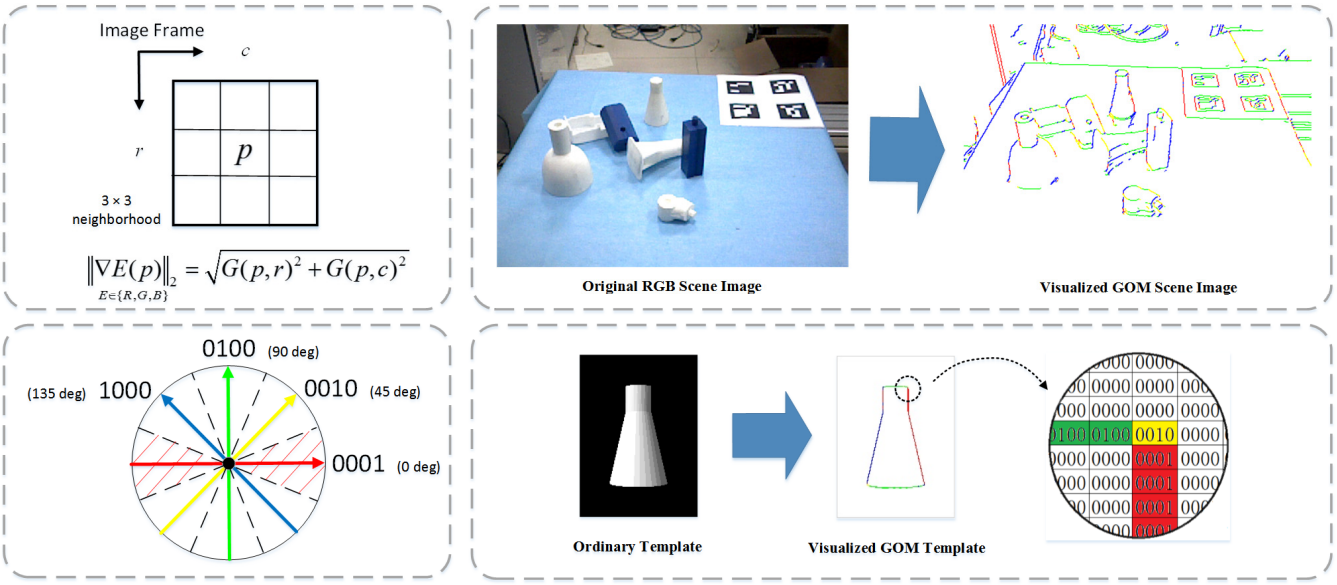

Figure 3. The plain introduction of Gradient Orientation Maps (GOM) in our approach. **Top-left:** The computation of gradient magnitude at the location p in image. **Bottom-left:** The quantization of gradient orientation and binary encoding for different orientations (the number of quantized spacings $n = 4$). **Top-right:** The convertion from original RGB data to GOM data for current scene image. **Bottom-right:** The convertion from original RGB data to GOM data for template.

And in GOM, we consider the smallest feasible neighborhood next, where the central pixel only needs to be compared to its direct neighbors. So, the Non-Maximum Suppression in GOM will be implemented in 3 × 3 neighborhood. Firstly, we will get the magnitude and orientation of gradient at each location $p(r,c)$ in input image. And then, the magnitude of gradient at neighboring location p_n and p_m will be computed. The pixel location p, p_n and p_m are in the same line whose orientation is the same with gradient orientation at location p. Once when the magnitude of gradient at location p is larger than any one of neighboring location p_n and p_m, the pixel location p will be a part of optimal edge and gradient orientation at location p will be added in GOM. The Dual Threshold Algorithm is also an edge optimization method which is widely used in image edge detection. It is crucial to choose an appropriate threshold in detecting edges of the image, which has a serious effect on quality of the image edges. In our approach, we acquire an edge image by using the high threshold τ_h at first, and this edge image may have several unclosed contours. Then, we will compute the magnitude of gradient at location p_s which is belong to 3 × 3 neighborhood of each edge in original edge image. In case that the magnitude of gradient at location p_d is between the high threshold τ_h and the low threshold τ_l , the pixel location p_d will be a part of optimal edge and gradient orientation at location p_d will be added in GOM. And the process can be easy to understand in Fig.4. Finally, we can get the complete Gradient Orientation Maps which has more robustness than GRM in practice.

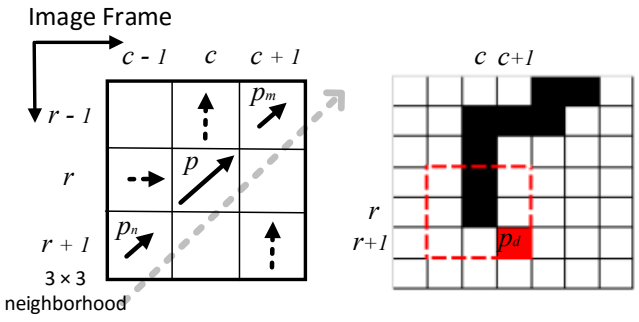

Figure 4. **Left:** The main principle of Non-Maximum Suppression. (↗) shows the gradient orientation and the length of (↗) represent gradient magnitude. According to the computation in 3 × 3 neighborhood of location p, p is a local maximum location. **Right:** The main principle of Dual Threshold Algorithm. If and only if the gradient magnitude at location p_d is between two threshold, p_d will be added in GOM.

B. *Template Matching and Coarse Pose Estimation*

Rely on Gradient Orientation Maps, we can acquire the optimal feature represent of template of target object w and current scene. After that, we will define similarity evaluation function which can evaluate the similarity between template of target object w and current scene with location $p(r,c)$ in scene image. Then, the optimal location can be found by using template matching where the evaluation score of similarity is highest in scene image. And our similarity evaluation function S can be formalized as:

$$S\{I,T,p(r,c)\} = \sum_{q \in D(T)} \{biori(I, p+q) \,\&\, biori(T,q)\} \quad (3)$$

Where $biori(T, q)$ is binarized representation of gradient orientation in the template T at location q; $D(T)$ is pixel position domain of template T, and $biori(I, p+q)$ is binarized representation of gradient orientation in scene image I at location $p + q$; & is logic AND operation bit-by-bit between two binarized representations. According to this formula, similarity evaluation function S will be computed at each location in scene image. And the input is GOM of template T and scene image I. Meanwhile, in order to reduce false recognition rate, we will present similarity threshold function Φ, which denotes the ratio between the similarity of current

matching target in scene image I and the maximum of similarity of current template T. It can be formalized as:

$$\Phi\{I,T,p(r,c)\} = S\{I,T,p(r,c)\}/countNonZero(T) \quad (4)$$

Where $countNonZero(T)$ denotes the maximum of similarity of template T. We compute similarity by using GOM of template T in practice, and every pixel in GOM of template T is binarized representation of gradient orientation. And the function *countNonZero*( ) is to compute the number of pixel that is not zero in input matrix. Thus, if current matching target is exact the same with template T, the output of similarity evaluation function will be exact the same with $countNonZero(T)$. And similarity threshold function not only can export the percent of similarity of current matching target, but also can reflect the percentage of occluded portion of target object. And the effectiveness of our approach will be demonstrated later.

To get the output of the similarity evaluation function in LINEMOD method, cosine of the angle between two gradient orientations need to be computed. Our approach only need to perform logic calculation. But because the number of features in GOM is larger than GRM, the time consuming in computation will still increase. To solve this problem, we adopt image pyramid searching method for fast template matching. Image pyramid is regarded as a good method in image processing and it offers a flexible, convenient multi-resolution format that mirrors the multiple scales of processing in visual system, is presented by Adelson et al. [15]. In our approach, we take Gaussian pyramid as image pyramid to establish optimal searching strategy. In a Gaussian pyramid, subsequent images are weighted down using a Gaussian blur and scaled down, and each pixel containing a local average that corresponds to a pixel neighborhood on a lower level of the pyramid. And [16], [17] applied it in object recognition task. As for searching strategy, our goal is to find the optimal location where the similarity score is largest in the scene image with regard to each template. Firstly, we can start at the top level with global search, and it will perform fast due to low resolution. Then, we will accept the top 10 percent of location hypotheses at the top layer of scene image pyramid. After that, we will perform template matching at the next layer with local search, and the search region is 10 × 10 neighborhood of location hypotheses produced at the higher layer. At last, the output is the optimal location at the bottom layer.

Our ultimate goal is to estimate the 6 DOF pose for target object. The pose is defined as the rigid body transformation (3D translation and 3D rotation) that maps point cloud from object space into camera space. [18], [19] As for RGB-D image, the color data is aligned with the depth data. So, the optimal location in 2D RGB image can be associated with the corresponding location in 2D depth image. Due to the output of 2D template matching, we can estimate the 3D translation of the target object by using aligned depth data. In practice, firstly we will extract depth data from the depth image at the corresponding location. Then, the depth data can be converted into 3D point cloud. The geometric center of this point cloud can be computed easily. Fig.5 shows this process. We will

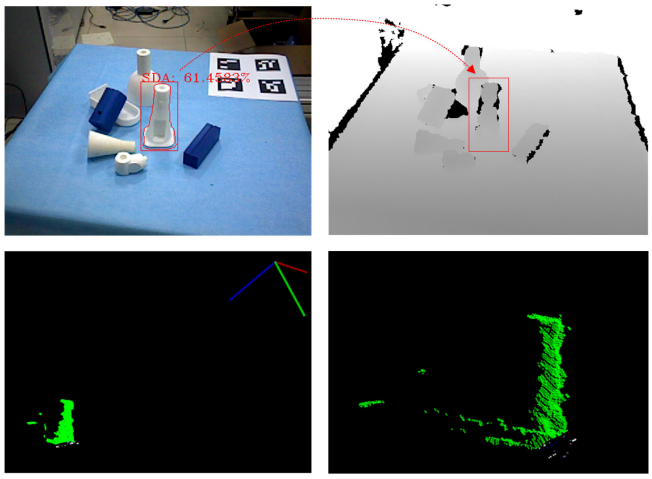

Figure 5. The process about how to get the coarse pose estimation. **Top-left:** The output of 2D template matching in RGB scene image. **Top-right:** The optimal location of target object in depth scene image can be found by using aligned data. **Bottom-left:** The original point cloud converted from the depth data in Top-right, and the reference frame of original point cloud is camera coordinate system. **Bottom-right:** Thr original point cloud viewed at high resolution.

take coordinate value of the geometric center as 3D coarse translation estimation. In our approach, each template has its 3D coarse rotation estimation, which is recorded during the process of template creation. Therefore, coarse pose estimation of target object can be accepted.

C. Fine Pose Estimation

Fine pose estimation refers to the accurate computation of translation and rotation parameters that the target object's position and orientation in camera space, on condition that coarse pose estimation parameters are provided. In our approach, objects are represented with 3D point cloud models, which can originate from digital scans or CAD drawings. Above all, we will upload the 3D point cloud model of target object. And we will adopt Iterative Closest Point (ICP) algorithm [7] to compute the fine pose estimation. ICP algorithm represents the gold standard method for geometrically aligning two sets of points whose coarse relative pose is provided. When the target object is partially occluded, the point cloud extracted from current scene point cloud may cover several occlusion points. ICP will generates large numbers of incorrect correspondences because of these occlusion noise and outliers. In order to solve this problem, we will use Statistical Outlier Removal Filter and Euclidean Cluster Extraction method in Point Cloud Library (PCL). The procedure is visualized in Fig.6. The Statistical Outlier Removal Filter is based on the computation of the distribution of point to neighbor's distances in the input dataset. For each point, the filter will compute the mean distance from it to all its neighbors. By means of assuming that the resulted distribution is Gaussian with a mean and a standard deviation, all points whose mean distances are outside an interval defined by the global distances mean and standard deviation will be regarded as noise and outliers. And Euclidean Cluster Extraction [20] is a clustering method dividing an unorganized point cloud into smaller parts by considering

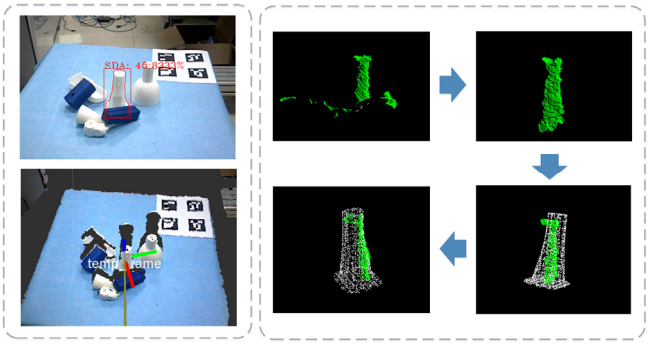

Figure 6. The procedure of fine pose estimation is visualized. **Top-left:** The output of 2D template matching in RGB scene image with partial foreground occlusions. **Bottom-left:** The output of fine pose estimation is visualized in RVIZ (3D simulation envionment in Robot Operation System). **Right:** The procedure of fine pose estimation relies on Statistical Outlier Removal Filter, Euclidean Cluster Extraction method and Iterative Closest Point algorithm.

Euclidean distance between nearest point. In practice, we will remove those parts whose number of points is not largest. Depends on these techniques, we can acquire the pure point cloud of target object. Then, ICP can mostly perform well for fine pose estimation. The process is repeated until the error is less than a certain threshold or maximum number of iterations is reached.

III. Experimental Result

The proposed approach aims to improve the current method of model-based detection and pose estimation of texture-less objects. Mainly, our goal is to overcome the problem that these methods all have less robustness with partial foreground occlusions. In order to verify the effectiveness of our approach, we build the object instance detection dataset which consists of 6 texture-less objects. In this paper, we take 3D printed model as target object due to the pervasive application of Material Increase Manufacturing.

Our training templates are rendered from the CAD models in OpenGL. Each object is represented by 1326 templates. And these templates show the feature of target object from different viewpoints. During the process of template creation, we record the rotation estimation for each template.

Mainly, we compare our method to the LINEMOD [4] method of Hinterstoisser et al. and the method of Drost et al. [8]. We use the same valuation approach of pose estimation error as in [4]. The target object is considered to be correctly localized if $e \leq k_m d$, where e is root-mean-square error of the transformation from the estimated pose to the ground truth, k_m is a fixed coefficient and d is approximate diameter of the target object. Table 1 shows the recognition rates ($k_m = 0.1$) of 6 texture-less objects with partial foreground occlusions by using different methods.

The average recognition rate of our approach is 86.2%, and that is the percentage of correctly localized objects. However, the average recognition rate of LINEMOD [4] and Drost et al. [8] are 46.9% and 43.2% respectively. Our approach has more robustness with regard to partial foreground occlusions than other model-based method. The effectiveness of our approach is shown in Fig.7. The score in the scene is the output of similarity evaluation function. And the recognition rate of our approach with various k_m for different target object is shown in Fig.8.

TABLE I. Recognition Rates [%] for Our Dataset (Partial Foreground Occlusions) ($k_m = 0.1$)

Target Object	Detection and Pose Estimation Approach		
	Our approach	*LINEMOD*	*Drost et al.*
Simple_base	**89.4**	56.8	38.7
Shell	**80.1**	40.8	39.5
SDA_base	**86.5**	54.9	28.5
Robot_base	**93.7**	34.6	46.0
Robot_part1	**84.4**	48.5	50.2
Robot_part2	**82.9**	45.7	56.4
Average	**86.2**	46.9	43.2

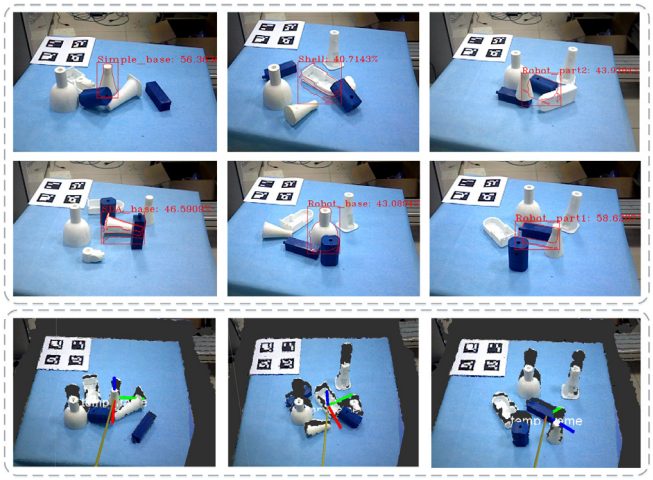

Figure 7. The effectiveness of our approach with partial foreground occlusions. **Top:** The result of 2D template matching. **Bottom:** The result of fine pose estimation.(cropped)

The RGB-D sensor that we utilize is Asus Xtion PRO. And our experiment is run on a PC with an Intel Core i5 CPU, 2.90GHz and 8GB memory. Those methods are programmed on Linux Operation System. In our experiment, several parameters need to be defined. For template creation, we adopt Dual Threshold Algorithm to improve feature maps of template with high threshold $\tau_h = 150$ and low threshold $\tau_l = 50$. As for fast template matching, the layers of image pyramid is 3. In the process of fine pose estimation, we also need initialize the point cloud filter. For Statistical Outlier Removal Filter, the number of neighbors to analyze for each point is set to 50, and the standard deviation multiplier to 0.01. The cluster tolerance of Euclidean Cluster Extraction method is set to 0.02, and minimum size of cluster points is 10.

IV. Conclusion

In this paper, an improved approach for model-based

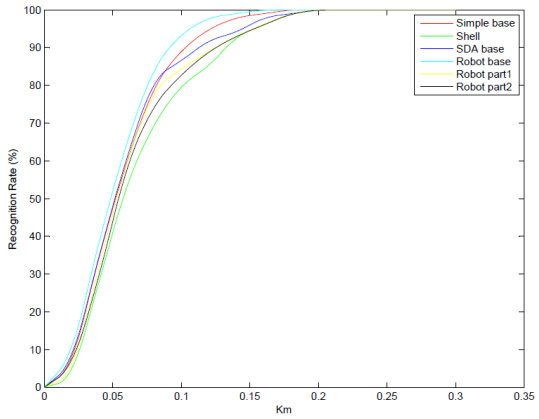

Figure 8. The recognition rate of our approach with various k_m.

detection and pose estimation of texture-less objects is presented. Current model-based method like LINEMOD [4] has less robustness with partial foreground occlusions. In LINEMOD [4] method, the gradient orientation features in GRM cannot represent the optimal contour of target object. Essentially, the feature representation cannot keep a high robustness with partial foreground occlusions, though it can perform well with heavily cluttered background. We present GOM as feature maps which can represent the optimal edge. Besides, image pyramid searching is also applied for fast template matching. As for fine pose estimation, the accuracy is improved by point cloud filter. And experiments on our dataset demonstrate that the proposed method has higher recognition rate compared with other model-based method with regard to partial foreground occlusions.

ACKNOWLEDGMENT

This work has been supported by National Natural Science Foundation of China (Grant No. 61273331) and YASKAWA Electric Corporation.

REFERENCES

[1] D. G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91–110, 2004.

[2] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In Proc. ECCV, pages 404–417, 2006.

[3] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In Proc. ICCV, pages 2564–2571, 2011.

[4] Hinterstoisser, S., Cagniart, C., Ilic, S., Sturm, P., Navab, N., Fua, P., Lepetit, V.: Gradient response maps for real-time detection of texture-less objects. IEEE Trans. on PAMI (2012)

[5] Hinterstoisser, Stefan, et al. "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes." Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011.

[6] Hinterstoisser, Stefan, et al. "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes." Computer Vision–ACCV 2012. Springer Berlin Heidelberg, 2012. 548-562.

[7] P. Besl and N. McKay, "A Method for Registration of 3-D Shapes," PAMI, vol. 14, no. 2, pp. 239–256, 1992.

[8] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in CVPR, 2010, pp. 998–1005.

[9] H. Cai, T. Werner, and J. Matas, "Fast detection of multiple textureless 3-D objects," in ICVS, ser. LNCS, 2013, vol. 7963, pp. 103–112.

[10] D. Damen, P. Bunnun, A. Calway, and W. Mayol-Cuevas, "Realtime Learning and Detection of 3D Texture-less Objects: A Scalable Approach," in BMVC, Sep 2012, pp. 1–12.

[11] C. Choi and H. Christensen, "3D Pose Estimation of Daily Objects Using an RGB-D Camera," in IROS, 2012, pp. 3342–3349.

[12] Choi, Changhyun, and Henrik I. Christensen. "3D textureless object detection and tracking: An edge-based approach." Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on. IEEE, 2012.

[13] Canny, John. "A computational approach to edge detection." Pattern Analysis and Machine Intelligence, IEEE Transactions on 6 (1986): 679-698.

[14] Neubeck, Alexander, and Luc Van Gool. "Efficient non-maximum suppression." Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Vol. 3. IEEE, 2006.

[15] Adelson, Edward H., et al. "Pyramid methods in image processing." RCA engineer 29.6 (1984): 33-41.

[16] Lai, Kevin, et al. "RGB-D object recognition: Features, algorithms, and a large scale benchmark." Consumer Depth Cameras for Computer Vision. Springer London, 2013. 167-192.

[17] Cheng, Ming, et al. "Global contrast based salient region detection." Pattern Analysis and Machine Intelligence, IEEE Transactions on 37.3 (2015): 569-582.

[18] Zhai, Meiyu, and Gregory B. McKenna. "Elastic modulus and surface tension of a polyurethane rubber in nanometer thick films." Polymer 55.11 (2014): 2725-2733.

[19] Zhai, Meiyu, and Gregory B. McKenna. "Viscoelastic modeling of nanoindentation experiments: A multicurve method." Journal of Polymer Science Part B: Polymer Physics 52.9 (2014): 633-639.

[20] R. B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, May 9-13 2011.