Multiple 3D Object Recognition using RGB-D Data and Physical Consistency for Automated Warehousing Robots

Shuichi Akizuki and Manabu Hashimoto
Graduate School of Computer and Cognitive Sciences, Chukyo University,
101-2, Yagoto-Honmachi, Showa-ku, Nagoya, Aichi, Japan

Keywords: 3D Object Recognition, Hypothesis Verification, Physical Consistency, Point Cloud.

Abstract: In this research, we propose a method to recognize multiple objects in the shelves of automated warehouses. The purpose of this research is to enhance the reliability of the Hypothesis Verification (HV) method that simultaneously recognizes layout of multiple objects. The proposed method have employed not only the RGB-D consistency between the input scene and the scene hypothesis but also the physical consistency. By considering the physical consistency of the scene hypothesis, the proposed HV method can efficiently reject false one. Experiment results for object which are used at Amazon Picking Challenge 2015 have been confirmed that the recognition success rate of the proposed method is higher than the previous HV method.

1 INTRODUCTION

3D object recognition from range data is one of fundamental techniques for scene understanding, object tracking, bin-picking for industrial robots, and others. Recently, this techniques have been increased in application of logistics (Fuji et al., 2015). In this field, it is necessary to develop automatic pick-and-place systems for items stocked in the warehouse. Amazon.com, Inc. held a competition for robotic picking system for items at IEEE ICRA2015, called Amazon Picking Challenge (http://amazonpickingchallenge.org/).

Shelves in the warehouse include many boxes called bin and they are stocked many items that include many categories, materials and size. The vision system for picking is imposed to detect a specific item from bins.

In order to recognize a target object, the model matching approach is generally used. This approach detects a 6DoF pose parameter which have the best fitting score between the object model and the input scene. However, this approach cannot notice mismatching, when the matching result is generated on the pseudo surface that is made by contacting multiple objects.

In this situation, the Hypothesis Verification (HV) method (Hashimoto et al., 1999), (Aldoma et al., 2012a), (Aldoma et al., 2013) is a suitable method for detecting specific object from complex scenes. This method consists of three steps, one each for:
Step 1. Generating object hypotheses
Step 2. Generating scene hypotheses using object hypotheses
Step 3. Verifying validity of each scene hypothesis and find the best one

Step 1 and 2 are collectively called hypothesis generation step. Step 3 is verification step. In step 1, object hypotheses, pairs of the object model and the 6DoF pose parameter, are generated by using the model matching method. In step 2, many scene hypotheses which represent the input scene are generated by using object hypotheses. They are represented by combination of multiple object hypotheses. In step 3, the best scene hypotheses is decided by calculating similarity between scene hypotheses and the input scene.

Because the HV method uses scene-to-scene consistency between the scene hypothesis and the input scene for recognizing multiple objects, it can reject mismatchings on the pseudo surface. However, the HV method has two problem. One is the reliability of method to generate object hypotheses is not so high. The other is calculating cost of the step 3 is relatively high compared with other steps.

The contribution of this paper is described as follows, 1) Applying a reliable model matching method in order to generate object hypotheses instead of previous model matching methods. 2) Improving cal-

Akizuki, S. and Hashimoto, M.
Multiple 3D Object Recognition using RGB-D Data and Physical Consistency for Automated Warehousing Robots.
DOI: 10.5220/0005723806050609
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 605-609
ISBN: 978-989-758-175-5

culation efficiency of scene consistency, we have applied a new criteria, physical consistency of scene hypothesis.

2 RELATED WORKS

In this section, we will introduce state-of-the-arts model matching methods and HV methods.

Model Matching: This method generates correspondences between an object model and an input scene by matching features extracted from each of them. Pose parameters that align the object model to the input scene can be calculated by using the method (Chen and Bhanu, 2007)(Tombari and Stefano, 2010). For generating accurate correspondences, it is important to design the good feature which can well represent object shapes. From the viewpoint of quantity of data for calculating the feature, there are three types of features:
1) point cloud around a keypoint
2) point cloud associated in a segment
3) pair or triplet of keypoints

Type 1) is well-known approach uses point cloud within the support region centered in the keypoint. The SHOT (Tombari et al., 2010) feature is generated by the histogram of oriented normal vectors. The histogram is generated from multiple divided support regions, and they are linearly-combined.

Type 2) is the semi-global feature that uses relatively large support region compared with Type 1) features. This type of features describe the rough geometric aspect on surface rather than the fine geometric aspect on surface. In this category, there is the OUR-CVFH (Aldoma et al., 2012b) that uses distribution of normal vectors and location of point clouds associated a segments. The GRF (Akizuki and Hashimoto, 2015a) generates the Reference Frame by using orientations of line segments sampled from outlines of the segment.

Type 3) is the category of the low-dimensional feature. The Point Pair Feature (Drost et al., 2010) and the Vector Pair (Akizuki and Hashimoto, 2015b) are describe geometric relation of points, such as distance of points or angle between normal vectors. Thanks to the low dimensional feature, it can quickly generate object hypotheses.

HV Method: This approach simultaneously recognize multiple objects by calculating consistency between the scene hypothesis and the input scene. The scene hypothesis is generated by combining object hypotheses recognized by the model matching method. Therefore, the HV method regards the multiple object recognition problem as a combinatorial optimization problem of object hypotheses. Because scene hypothesis represents layout of objects in the scene, false object hypotheses can be rejected. Important thing here is, what kind of information is suitable for calculating scene consistency. Methods (Hashimoto et al., 1999) and (Aldoma et al., 2012a) used depth data in order to evaluate shape consistency. The method (Aldoma et al., 2013) have developed reliability of consistency by employing color consistency in addition to shape consistency.

3 PROPOSED HV METHOD USING RGB-D DATA AND PHYSICAL CONSISTENCY

3.1 Overview

Previous HV methods have used RGB-D information for calculating the scene consistency. The proposed HV method uses not only RGB-D score but also physical consistency for calculating the scene consistency. Physical consistency represents the naturalness of scene hypothesis like whether recognized objects are intrude each other or not. By employing physical consistency on scene consistency, scene hypotheses that have impossible layout can be early reject. So, we have developed efficiency of recognition.

The proposed HV method consists two steps, 1) generating object hypotheses and 2) verification, same as general HV methods. In module 1), object hypotheses are generated by the low-dimensional feature based model matching method. In module 2), Shape, color and physical consistency between scene hypotheses and the input scene is calculated. Figure 1 shows the overview of the algorithm of the proposed HV method.

3.2 Generating Object Hypotheses

This module generates object hypotheses by using the model matching method. Object hypotheses are defined as $H\{h_1,\ldots,h_n\}$. h_i consists pair (M_{h_i}, T_{h_i}). M_{h_i} and T_{h_i} means the object model and the 6DoF pose parameter, respectively. In the model matching, object hypotheses are generated by allowing to detect some false positives. From the viewpoint of processing time, the proposed method have employed the Vector Pair Matching (VPM) method (Akizuki and Hashimoto, 2015b) as the model matching. Overview of the algorithm of the VPM is explained below.

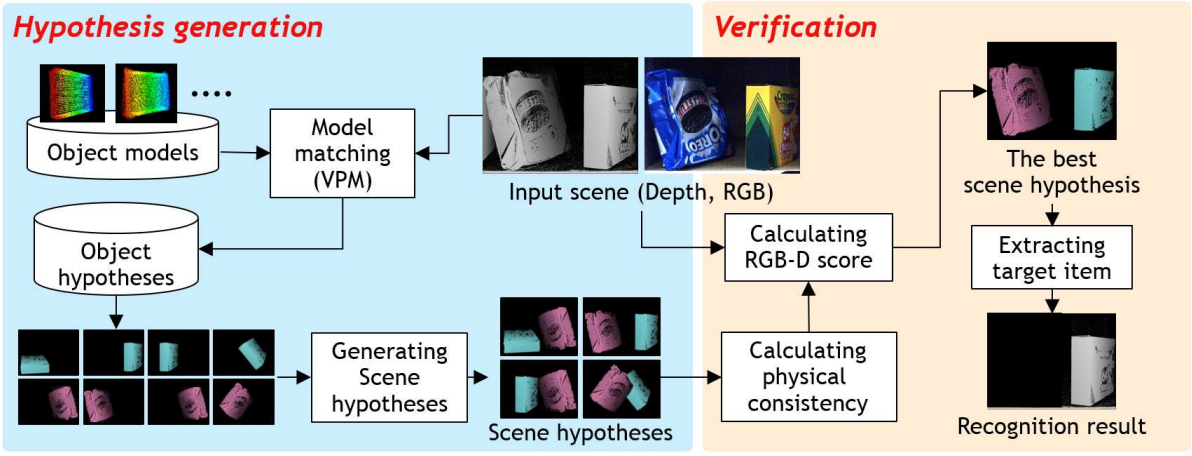

Figure 1: The overview of the algorithm of the proposed HV method consists of main two module, hypothesis generation (left) and verification (right).

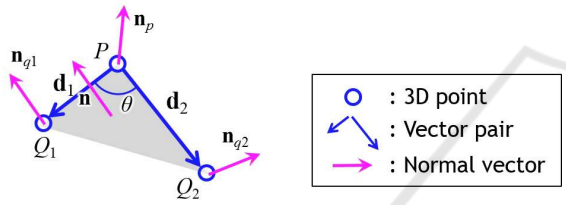

Figure 2: The structure of the vector pair. The blue circles represent 3D points, the pair of blue arrows represents the vector pair, and the pink arrows represent normal vectors of a point or triangle.

Vector Pair Feature: Three 3D points are necessary to determine 3D pose parameter of an object. The VPM method treats these points as a vector pair which consists of two 3D vectors with common start point. Structure of vector pair is shown in Figure 2.

Here, the vector pair consists of a start point P and two end points Q_1, Q_2. Displacement vectors $P-Q1$ and $P-Q2$ are represented by $\mathbf{d_{q1}}$ and $\mathbf{d_{q2}}$, respectively. And vector pair has a feature vector $\mathbf{v} = (s_p, s_{q1}, s_{q2})$ which is calculated from distribution of surface normal vector. s_p, s_{q1} and s_{q2} are calculated by Equation 1.

$$s_p = \mathbf{n} \cdot \mathbf{n}_p,\ s_{q1} = \mathbf{n} \cdot \mathbf{n}_{q1},\ s_{q2} = \mathbf{n} \cdot \mathbf{n}_{q2}$$
$$\textit{where}, \qquad \mathbf{n} = \mathbf{d}_{q1} \times \mathbf{d}_{q2} / \|\mathbf{d}_{q1}\mathbf{d}_{q2}\| \qquad (1)$$

$\mathbf{n_p}$, $\mathbf{n_{q1}}$ and $\mathbf{n_{q2}}$ represent surface normal vector of point P, Q_1 and Q_2, respectively. $\mathbf{n}$ represents normal vector of $\triangle PQ_1Q_2$. This feature also have Reference Frame (RF). Vector $P-Q1$ and $P-Q2$ are orthogonal, therefore, $\mathbf{y} = (P-Q_1)/|P-Q_1|$ and $\mathbf{x} = (P-Q_2)/|P-Q_2|$ and $\mathbf{z} = \mathbf{y} \times \mathbf{x}$ consist each axis of the RF.

The VPM method uses few number of distinctive vector pairs which have high observability for reliable matching. The observability factor of the vector pair is calculated by simulating the visible state of the vector pair from various viewpoints.

Matching Module: This module consists following three steps: 1) Correspondences between scene vector pairs and distinctive vector pairs are extracted from the object model. At this time, the vector pair has RF, 6DoF pose parameters that aligns corresponding vector pairs are calculated. 2) Second step is the voting process. Calculated pose parameters are voted to a voting space consisted of axes that represents each pose parameter, x, y, z, roll, pitch and yaw. 3) Third step calculates shape consistency. The object model is transformed by pose parameters which have high voting value. And shape consistency between the transformed object model and the input scene is calculated. Pairs the object model and the pose parameter which have the score exceeding threshold value are registered to the object hypotheses.

3.3 Verification

This module generates a reconstructed scene that well describe the input scene by evaluating the RGB-D score and the physical consistency of scene hypotheses. First of all, scene hypotheses are generated by combining some object hypotheses. $X = \{h_1, \ldots, h_m\}$ represents the combination of object hypotheses. A depth image I^D and a color image I^C are generated by projecting transformed M_{h_i} associated h_l within the X to each image plane.

If the proposed method chooses correct combination of object hypotheses, the consistency of the generated scene hypothesis and the input scene will be high. We have regarded this problem as a combinatorial optimization problem, and the proposed method

detects the X that maximizes the cost function defined in Equation 2. We have employed the Genetic Algorithm to solve this problem.

$$Score = P(X)Score^{RGBD}(X) \quad (2)$$

$P(X)$ evaluates the physical consistency of the scene hypothesis X. $Score^{RGBD}$ evaluates the shape, color consistency of the hypothesis scene and the input scene. Methods to calculate the physical consistency and the shape, color consistency are explained, as follows.

Physical Consistency $P(X)$**:** This function evaluates whether the hypothesis scene can physically exist or not by using Equation 3.

$$P(X) = \prod_{i,j \in S} C(h_i, h_j) \quad (3)$$

In this equation, function C is a binary function. If h_i and h_j does not overlap, it returns 1. On the other hands, if overlap is occurred, C returns 0. S represents a set of index pair of h. It does not includes same index pair. In order to evaluate overlap of object hypotheses, we employed the fast collision detection method proposed in (Gottschalk et al., 1996). This method detects whether paired Oriented Bounding Box (OBB)s contact or not. By using the method, the proposed method can evaluate overlap of paired object hypothesis. At this time, the size of OBB is little bit smaller than the actual object size (e.g. 90%). By evaluating collision of small sized OBBs, P(X) can estimate whether paired object hypotheses overlap or not.

RGB-D Score $Score^{RGBD}(X)$**:** RGB-D score of scene hypothesis X is evaluate by Equation 4.

$$Score^{RGBD}(X) = \sum_{i=1}^{N} f^D(i) f^C(i) \quad (4)$$

Function f^D, the shape consistency of scene hypothesis, is defined by Equation 5.

$$\begin{cases} 1 & if |I^D_S(i) - I^D_{Hyp}| < th \\ 0 & else \end{cases} \quad (5)$$

I^D_{Hyp}, I^D_S are depth image of the scene hypothesis and the input scene, respectively. In the equation 5, f^D returns 1, if the difference of same pixel of two images is lower than th.

Function f^C, the color consistency of scene hypothesis, is defined by Equation 6.

$$f^C(i) = 1 - |I^C_S(i) - I^C_{Hyp}(i)| \quad (6)$$

(a) Items

(b) Shelf

(c) Input scene

Figure 3: The overview of the dataset. (a) Target items. Attached number means item ID. (b) Shelf. (c) Examples of input scene. Left shows depth image. Right shows RGB image. Top row is the scene of item 1 and 6. Bottom row is the scene of item 2 and 5.

I^C_{Hyp}, I^C_S are color image of the scene hypothesis and the input scene, respectively. This function evaluates similarity of hue value. In this equation, value $I^C_{Hyp}(i)$, $I^C_S(i)$ represent hue value of i th pixel.

4 EXPERIMENTS

4.1 Dataset

In order to evaluate recognition performance of the proposed method, we have prepared 25 items which are used on Amazon Picking Challenge 2015. In this experiment, we have selected 20 items which can be stably acquired depth data. We put randomly chosen two items in the bin of shelf, and captured the depth data and the RGB data. We prepared 300 input scenes. Figure 4 shows items, the shelf and an example of captured data. We also prepared the ground truth data for all input scene. They are masked images which have pixel-level label for each object.

4.2 Result

We evaluated method's performance by recognizing items in the bin. In this experiment, we compared our method with the method (Aldoma et al., 2012a). Methods are implemented by using Point Cloud Li-

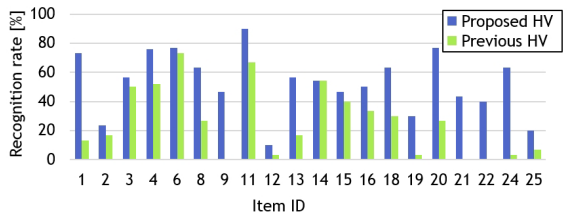

Figure 4: Recognition success rate for each item.

brary (Rusu and Cousins, 2011). In order to avoid influence of performance of the method to generate object hypothesis, we applied common model matching method, the VPM.

In order to decide whether the recognition is success or not, we evaluated the F measure calculated by comparing the recognized object region and the ground truth. If the F measure exceeding 0.5, then we decided recognition is success.

Figure 5 shows the recognition performance of each method. Both method are used the VPM for generating object hypotheses, so the results depended on the algorithm of the verification. Average recognition rate of the proposed HV method and the previous HV method are 52.8% and 25.8 %, respectively. It have been confirmed that the reliability of recognition is higher than the previous HV method.

Recognition rate of the ID 2, 12, 19, and 25 are relatively lower than the other method. These items are thin compared with others, so the area of appearance in the input scene was small. As a result, these items are not recognized by the VPM.

5 CONCLUSION

In this research, we have proposed the method to enhance the reliability of the Hypothesis Verification (HV) method that simultaneously recognizes layout of multiple objects. The proposed method have employed not only the RGB-D consistency between the input scene and the scene hypothesis but also the physical consistency. By considering the physical consistency of the scene hypothesis, the proposed HV method can efficiently reject false one. In addition, the method have applied a reliable model matching method, the VPM. As for future work, we will develop the method to recognize thin objects.

ACKNOWLEDGEMENTS

This work was partially supported by Grant-in-Aid for Scientific Research (C) 26420398.

REFERENCES

Akizuki, S. and Hashimoto, M. (2015a). A proposal of the global reference frame for surface flatness-independent 3d object detection. In *Proc. Joint Conference of IWAIT and IFMIA*.

Akizuki, S. and Hashimoto, M. (2015b). Stable position and pose estimation of industrial parts using evaluation of observability of 3d vector pairs. 27(2):174–181.

Aldoma, A., Tombari, F., di Stefano, L., and Vincze, M. (2012a). A global hypotheses verification method for 3d object recognition. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision*, pages 511–524.

Aldoma, A., Tombari, F., Prankl, J., Richtsfeld, A., di Stefano, L., and Vincze, M. (2013). Multimodal cue integration through hypotheses verification for RGB-D object recognition and 6dof pose estimation. In *IEEE International Conference on Robotics and Automation*, pages 2104–2111.

Aldoma, A., Tombari, F., Rusu, R. B., and Vincze, M. (2012b). OUR-CVFH - oriented, unique and repeatable clustered viewpoint feature histogram for object recognition and 6dof pose estimation. In *Pattern Recognition - Joint 34th DAGM and 36th OAGM Symposium*, pages 113–122.

Chen, H. and Bhanu, B. (2007). 3d free-form object recognition in range images using local surface patches. 28(10):1252–1262.

Drost, B., Ulrich, M., Navab, N., and Ilic, S. (2010). Model globally, match locally: Efficient and robust 3d object recognition. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 998–1005.

Fuji, T., Kimura, N., and Ito, K. (2015). Architecture for recognizing stacked box objects for automated warehousing robot system. In *Proceedings of the 17th Irish Machine Vision and Image Processing conference*, pages 50–56.

Gottschalk, S., Lin, M. C., and Manocha, D. (1996). Obbtree: A hierarchical structure for rapid interference detection. In *SIGGRAPH*, pages 171–180.

Hashimoto, M., Sumi, K., and Usami, T. (1999). Recognition of multiple objects based on global image consistency. In *Proceedings of the British Machine Vision Conference*, pages 1–10.

Rusu, R. B. and Cousins, S. (2011). 3d is here: Point cloud library (PCL). In *IEEE International Conference on Robotics and Automation, ICRA*. IEEE.

Tombari, F., Salti, S., and di Stefano, L. (2010). Unique signatures of histograms for local surface description. In *European Conference on Computer Vision ECCV*, pages 356–369.

Tombari, F. and Stefano, L. D. (2010). Object recognition in 3d scenes with occlusions and clutter by hough voting. In *Proc. Fourth Pacific-Rim Symposium on Image and Video Technology*, pages 349–355.