A Real-Time System for Object Detection and Location Reminding with RGB-D Camera

I-Kuei Chen[†], Chung-Yu Chi[†], Szu-Lu Hsu[†], and Liang-Gee Chen[*]
DSP/IC Design Lab, National Taiwan University, Taiwan
[†]{eugenegx, craig, bismarck}@video.ee.ntu.edu.tw [*]lgchen@video.ee.ntu.edu.tw

Abstract—**This paper presents a helpful application with a real-time detection system that can automatically capture the last scene where the user-defined important objects appear. The introduced method uses RGB-D information as input and has high detection rate in messy indoor environments. Additionally, we build an user-friendly using flow on object online learning and detection which may be suitable for future technologies of wearable devices.**

I. Introduction

In recent years, there has been increasing interests in using RGB-D information from video sequences for object detection. Real-time object online learning and detection are important and challenging tasks in computer vision research area. People, especially the elders, tend to forget where the important things are and spend a large amount of time searching them. In this paper, we propose a RGB-D information system for robust detection of object location and helping people find the target object immediately with a simple online training stage in advance.

The main advantage of our system is that we adopt a template-matching-based algorithm which the target object can be easily learnt online with low cost in contrast to the techniques that require controlled and strict environment in laboratory [1]. Moreover, due to the RGB-D information and the algorithm it used, this system also has high accuracy despite of the hard and cluttered situation. Tracking-by-detection method can always track the target object even if it is absence in the scene for a while. The accuracy and robustness are verified in the experiment results of several video sequences from indoor environments.

As the prevalence of wearable device such as project-glass [2], it becomes an inevitable trend, and applications available to fit in these devices turn into one of the main issues in Computer Vision research area. Also, small RGB-D cameras [3] have become an emerging product that can be used in portable devices. As the result, low complexity and efficiency of the algorithm is crucial when the applications we aim at need to be real-time.

In the proposed work, we use Point Cloud Library (PCL [4]) to capture basic 3D information in the image and to further calculate advanced data. PCL is a famous open source project, collecting large amounts of codes and algorithms to assist developers to process point cloud datasets. By PCL, it can dramatically shorten the system development duration and cut down the cost needed.

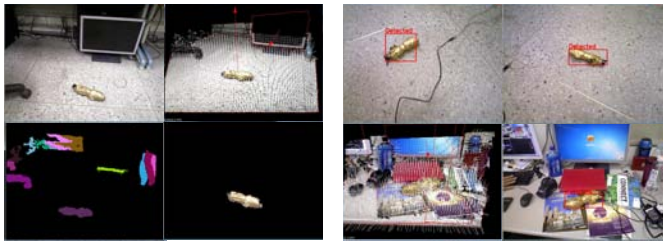
II. Proposed System Flow

Our object reminder system comprises two phases as Fig. 2. First, it begins with training phase, in which the user assigns interested objects and make the system learn online. Next, we can change the system to operating phase which performs all-time detecting and tracking for objects learnt from the previous phase. It can be used in two main scenarios: finding specific objects in current users view, and recalling the recent scenes from memory data where specific objects present.

A. Training Phase

We extract multimodal templates [5] to train a set of robust templates for each object. Good templates lead to high detection rate, so we put the interested objects in an easily-detected condition, such as a salient object on a large flat plane without clustered items. In order to increase the convenience and speed of learning, plane segmentation and cluster segmentation in point cloud [6] is firstly used to simplify the procedure of producing templates. The cluster in the middle of the frame is then chosen as the templates to learn(Fig. 1(a)).

By rotating the viewpoint for the interested object placed on the flat plane, we automatically learn various angle views of the specific object for appearance variation. Implying image cue (color gradients) and depth cue (surface normals) of the object, multimodal templates are robust to illumination change and noise, and are proved to be one of the best state-of-the-art methods in 3D object detection.



(a) (b)

Fig. 1. (a)Train an umbrella in different viewpoints. The top-right figure indicates the flat plane and the normal vectors of every points. Raw cluster segmentation is performed as the bottom-left figure, and the bottom-right figure is the segmented result which we use to extract templates. (b)Top figures refer that we can detect different viewpoints for objects when using multimodal templates. The bottom-left figure is the point clouds with normals in a messy and complicated indoor scene, while the bottom-right figure is an example of detecting things in it.

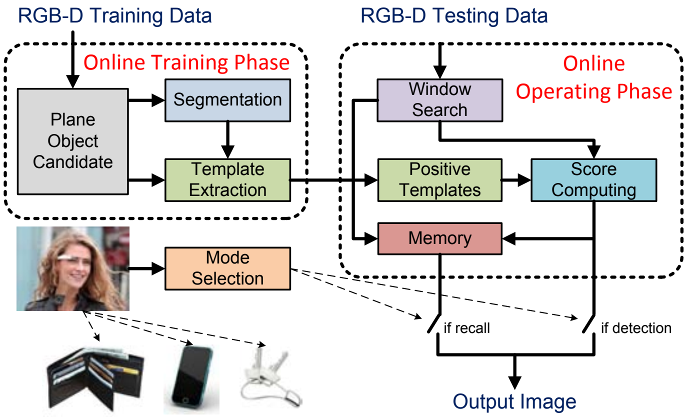


Fig. 2. Details of the proposed object detection and location reminding system and the dataflow.

TABLE I
THE AVERAGED PRECISION, RECALL, AND F-SCORE OF EACH OBJECT.

Objects	Camera	Bottle	Umbrella
Precision	98.92%	91.05%	88.37%
Recall	92.00%	83.50%	76.00%
F-score	95.34%	87.09%	81.72%

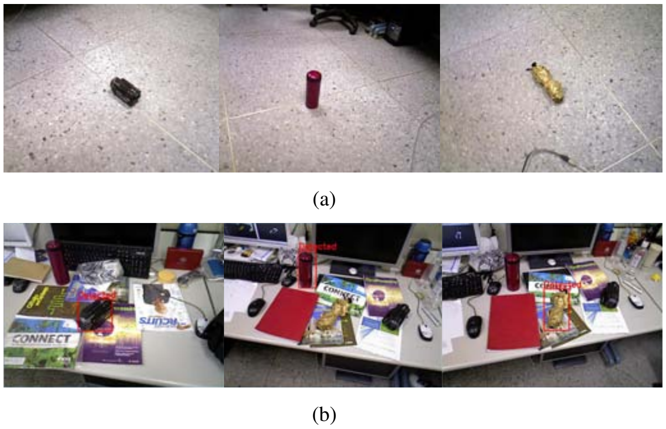
B. Operating Phase

After acquiring templates, our system performs multimodal detection on real-time RGB-D input video stream. We utilize LINE-MOD [5] method, which applies window search on the current view, and then the response scores of each window are compared. The window with the maximal response is considered detection result as Fig. 1(b).

In every incoming frame, we continuously execute LINE-MOD detection of trained objects as tracking. If the interested object suddenly disappears, the system captures the last frame it remains and saves to memory. Considering the real user situation, the detection of objects may flicker due to strong shakes of wearable devices or sensor noises. Thus it should be more practical that we keep saving critical frames and replace the oldest with newest under limited memory. The users can then simply recall the most recent scenes in which important objects exist if users forget where they are.

III. Experiment Results

We use a 3D input sensor, which generates RGB-D data with 640×480 resolution, and run on a 3.07GHz CPU. It takes about 0.31 seconds to train a template and about 0.19 seconds to detect from a scene. We specified three interested objects as Fig. 3(a), named *Camera*, *Bottle*, and *Umbrella*. We put them on flat plane and rotated viewpoint every 30 degrees to train new templates, around total 12 templates per object. To test the robustness of our system, we also produced two complicated and cluttered indoor scenes and placed three mentioned objects in them.


(a)

(b)

Fig. 3. (a)Interested objects from left to right *Camera*, *Bottle*, and *Umbrella*. (b)Successful detection of three objects in the sequences tested.

Fig. 3(b) is the result of detection of three objects in cluttered scene individually. Through the experiment, we get the centric point $P_d{=}(x_d, y_d)$ of each detection box. On the other hand, we locate the golden 2D center position $P_g{=}(x_g, y_g)$ of each object by human labeling. The precision is defined as the proportion of true existence in detected results. By measuring the recall $R = \frac{\text{frames where object is detected}}{\text{frames where object exists}}$, it can be proved that our real-time system is quite robust for object finding(Table. I), where F-score equals to

$$\text{F-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{1}$$

We define the term "detected" as D, where we empirically and reasonably set threshold as 20 pixels in 2D image.

$$D = \begin{cases} \text{true}, & \text{if } \|P_d - P_g\|_2 \leq \text{threshold} \\ \text{false}, & \text{if } \|P_d - P_g\|_2 > \text{threshold} \end{cases} \tag{2}$$

IV. Discussion and Conclusion

We have presented a method to exploit RGB-D information for real-time object detection and tracking. Our novel approach is able to correctly detect target objects in real-time under messy situations. Due to fast growing demand of depth information, it is believed that the technology of depth sensor is improving rapidly. If the day comes, our user-friendly and intuitive application must have great potentials to be developed into a commercial product.

References

[1] K. Lai, L. Bo, X. Ren, D. Fox, "A large-scale hierarchical multi-view RGB-D object dataset," *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1817-1824, 2011.

[2] Google$^{\text{TM}}$, "Glass," Internet: www.google.com/glass/, Jul. 15, 2013

[3] PrimeSense$^{\text{TM}}$, "Capri 1.25," Internet: www.primesense.com/tag/capri-1-25/, Jul. 15, 2013

[4] Radu Bogdan Rusu and Steve Cousins, "3D is here: Point Cloud Library (PCL)," *IEEE ICRA*, Shanghai, China, May 9-13, 2011.

[5] S. Hinterstoisser, S. Holzer, C. Cagniart,S. Ilic, K. Konolige, N. Navab, V. Lepetit, "Multimodal templates for real-time detection of textureless objects in heavily cluttered scenes," *Computer Vision (ICCV), IEEE International Conference on*, pp. 858-865, 2011.

[6] A. Trevor, S. Gedikli, R. Rusu, H. Christensen, "Efficient Organized Point Cloud Segmentation with Connected Components," *3rd Workshop on Semantic Perception Mapping and Exploration (SPME)*, Karlsruhe, Germany. May, 2013.