Egocentric Video Multi-viewer for Analyzing Skilled Behaviors based on Gaze Object

Yuki Umezawa
Graduate School of Information Science
Nagoya University
umezawa@cmc.ss.is.nagoya-u.ac.jp

Takatsugu Hirayama
Institutes of Innovation for Future Society
Nagoya University
takatsugu.hirayama@nagoya-u.jp

Yu Enokibori
Kenji Mase
enokibori@is.nagoya-u.ac.jp
mase@nagoya-u.jp
Graduate School of Informatics
Nagoya University

ABSTRACT

In many intellectual tasks, efficient succession of human physical and sensory skills is a long-standing issue. In order to analyze skilled behaviors, a useful approach is to compare same task scenes among workers or days and to understand these differences. In this paper, we propose an egocentric scene classification method based on objects which the worker turned the gaze to and a multi-viewer for egocentric videos comparison. We have experimented on proposed classification method with videos of painting watercolor.

CCS CONCEPTS

• **Human-centered computing** → *Visualization toolkits*;

KEYWORDS

Egocentric video; gaze; multi-viewer; skill analysis

ACM Reference Format:
Yuki Umezawa, Takatsugu Hirayama, Yu Enokibori, and Kenji Mase. 2018. Egocentric Video Multi-viewer for Analyzing Skilled Behaviors based on Gaze Object. In *IUI'18 Companion: IUI'18: 23rd International Conference on Intelligent User Interfaces Companion, March 7–11, 2018, Tokyo, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3180308.3180342

1 INTRODUCTION

In the field where developing skills is essential, such as sports, manufacturing, nursing care and arts, the skills are often taught to novices in subjective ways. There are problematic cases where skill succession can not be done well because experts empirically acquire the skills as tacit knowledge. Therefore, effective skill analysis and succession are important issues.

As a key media to analyze skilled behaviors, egocentric video taken with a wearable camera mounted on the head of person has attracted the attention because it shows what s/he does or pays the attention to. Also, gaze behavior of expert, which is regarded as temporal pattern of attention, is an important modality for skill analysis [1, 2]. We therefore focus on objects to which experts turn their gaze in egocentric videos.

IUI'18 Companion, March 7–11, 2018, Tokyo, Japan

ACM ISBN 978-1-4503-5571-1/18/03.
https://doi.org/10.1145/3180308.3180342

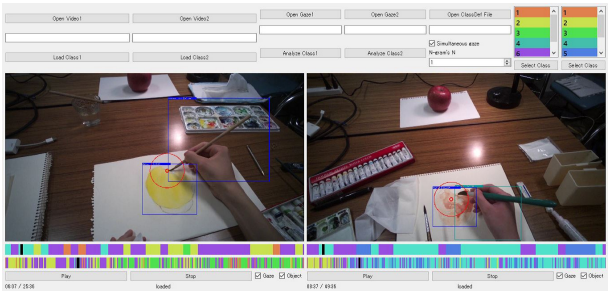

Figure 1: Multi-viewer for egocentric videos comparison.

We assume that the gaze objects correspond to a task for the reason stated above. Comparing same-task scenes among workers or days is a useful approach to analyze the skilled behaviors. In this paper, we propose an egocentric scene classification method based on the gaze objects, and a multi-view interface for egocentric videos comparison. Figure 1 shows the proposed interface.

It supports the analysis of the differences in the hand movement, the attention, and spent duration between the operators. It encourages understanding of the tips of tasks and unconscious skills. We have experimented on proposed classification method with videos of painting watercolor, and we discuss the validity of the proposed method. We also discuss the issues and improvements of the proposed interface.

2 GAZE OBJECT BASED SCENE CLASSIFICATION

We classify short video segments into some task scenes. We extract a histogram of gaze objects as a feature from each segment.

Given egocentric videos and the corresponding gaze positions, we first detect objects appeared in each of video frame using YOLO [3]. Here, we obtain a bounding box (BB) surrounding each object and also labels of them. Then, gaze counts for each object are accumulated when the gaze region with a constant radius around a gaze point is overlapped for the BB. Figure 2 shows the diagram of gaze counts calculation. Finally, we obtain a histogram of gaze counts for objects in a short video segment consisting of some frames. Figure 3 shows an example of histogram extracted as a feature.

Since we assume that any pre-defined labels of skilled behaviors can not be assigned to intellectual task scenes, we employ an unsupervised clustering, concretely k-means method in this paper.

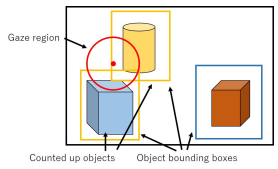

Figure 2: Diagram of gaze counts calculation.

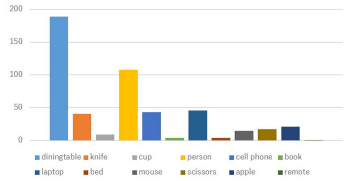

Figure 3: Histogram of gaze counts.

1	2	3	4	5	6
apple	**person**	**person**	**donut**	**oven**	**diningtable**
110.94	168.73	130.61	142.37	210.60	99.63
diningtable	**diningtable**	**apple**	**diningtable**	**refrigerator**	**person**
101.05	104.43	127.42	125.79	140.05	62.41
person	**donut**	**diningtable**	**person**	**diningtable**	**apple**
33.69	27.45	98.97	74.84	82.92	25.13
knife	**book**	**knife**	**apple**	**person**	**knife**
21.61	22.11	43.12	20.52	53.14	22.52
donut	**apple**	**donut**	**knife**	**cake**	**cup**
17.62	21.13	33.07	12.84	31.12	20.63

Table 1: Representative gaze objects.

The histograms for the segments in multiple egocentric videos are classified into k clusters.

3 CLASSIFICATION EXPERIMENT

3.1 Data and parameters

We asked seven experimental participants with normal vision to wear Tobii Pro glass 2 eye tracker[1] and draw a watercolor painting of an apple. The participants included three skilled painters and four novices. The total length of egocentric videos with gaze data taken in this experiment was approximately 3.5 hours. As the object detection method, we used YOLO trained with Microsoft COCO data set[2]. Regarding the scene classification, we obtained six clusters.

3.2 Result and Discussion

Table 1 shows the representative gaze objects that are included in each scene class and have the top five scores based on the principle of tf-idf, and the scores are shown in the table. The score is calculated as follows,

$$S(c, o) = C(c, o) \cdot \ln \frac{\max_i C(i, o)}{\frac{1}{n}\sum_{i=1}^{n} C(i, o)} + 0.5 \qquad (1)$$

where c is the scene class label, o is the object label, n is the number of scene class, $C(c, o)$ is the number of o shown in cluster centroid of c obtained from the k-means result.

Figure 4 shows the examples of frames included in segments belonging to the same scene class.

From Table 1, it can be seen that the scene was classified based on the participant's gaze object, and the video segments classified into the same class show similar scene as for participant's behavior as shown in Figure 4(a). However, there were segments including different gaze objects and behaviors even though they belong to the same class as shown in Figure 4(b). From this result, gaze action on objects could express some skilled behavior.

[1]Tobii Pro glass 2: http://www.tobiipro.com/ja/product-listing/tobii-pro-glasses-2/
[2]Microsoft Common Objects in Context: http://mscoco.org/

(a) Class 4

(b) Class 6 Different behavior scenes

Figure 4: Examples of frames included in segments belonging to the same scene class (left: participant A, right: participant B).

4 MULTI-VIEWER FOR EGOCENTRIC VIDEO COMPARISON

We created a comparative multi-viewer of egocentric videos to help comparing among the video segments classified into the same task scene as shown in Figure 1. This interface can simultaneously play back the two videos. By selecting a scene class label from the upper right box, only the segments belonging to the selected scene class are played back continuously.

Although the current interface have a function that displays the representative gaze objects included in each scene class, it provides only minimum information about each class characteristic. In order to understand the difference of skills from the videos, we need to extend it to exploit temporal gaze transition patterns in the future.

5 CONCLUSION

We proposed the egocentric scene classification method based on gaze objects and the multi-viewer for egocentric videos comparison. We will conduct a quantitative evaluation for the egocentric scene classification and interface usability toward realizing effective skill analysis and succession. The interface design to help a sequential pattern mining of the gaze objects is also a future work.

6 ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Number 26280074.

REFERENCES

[1] A. L. Yarbus, "Eye Movements and Vision," New York: Plenum Press, 1967.
[2] A. Iwatsuki, T. Hirayama, J. Morita, K. Mase, "Skilled gaze behavior extraction based on dependency analysis of gaze patterns on video scenes," *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research and Applications*, pp. 299-302, 2016.
[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.