

Investigation of Customer Behavior Analysis Based on Top-View Depth Camera

Jumpei Yamamoto

Graduate School of Engineering
Osaka Prefecture University

yamamoto-j@sig.cs.osakafu-u.ac.jp

Katsufumi Inoue

Graduate School of Engineering
Osaka Prefecture University

inoue@cs.osakafu-u.ac.jp

Michifumi Yoshioka

Graduate School of Engineering
Osaka Prefecture University

yoshioka@cs.osakafu-u.ac.jp

Abstract

Human behavior analysis based on surveillance camera is one of hot topics in security, marketing as well as computer vision and pattern recognition, and these are useful for commercial facilities such as convenience stores or book stores. In general, since surveillance camera is placed on the ceiling near store wall to monitor customer behaviors, the majority of this research utilize human model adapted to a front view of the person because the human shape has high discriminative power for human detection, pose estimation, etc. However, this approach has a problem that customers are often occluded by others in the store. To solve this occlusion problem for behavior analysis, we propose a new human behavior analysis method with top-view depth camera. In this research, for the first step to investigate the effectiveness of the analysis, we suppose the book store situation. Our proposed method is composed two behavior estimators. The first estimator is based on height of hand with depth information and the second is based on SVM with depth and PSA (Pixel State Analysis) based features. The characteristic point of our proposed method is that we utilize these estimators by cascading them. From experimental results with 10 behaviors of 3 subjects, although our research is still exploratory, we have confirmed that our proposed method allows us to obtain 89.5% of average estimation accuracy.

1. Introduction

Human behavior analysis based on surveillance camera is one of hot topics in the research area of computer vision and pattern recognition, etc. For example, in the case of commercial facilities such as convenience stores and book stores, customer behaviors analysis in front of

shelves is very useful for marketing since they are strongly related to the interests of customers for products, as well as for security to prevent stealing. In order to obtain such benefits, many researchers try to analyze customer behaviors [9, 13, 11]. In general, since surveillance cameras are placed on ceiling near store wall and they widely monitor their obliquely downward situation, customers are approximately captured from frontal view. Therefore, in order to analyze customer behaviors, the majority of this research area utilizes human model adapted to a front view of the person because the human shape has high discriminative power for human detection [17], pose estimation [18] and action recognition [9, 16]. In addition, they target only the most frontal customer in the captured images/videos with surveillance camera for behavior analysis. However, since customers are often occluded by others or the shelves in the store, this occlusion is a common problem among the methods mentioned above.

To solve this problem, recently, the customer behavior analysis with top-view depth camera [11, 14, 10, 5, 3, 8, 2, 12] has been focused on because this can relax the privacy problem. Our research also belongs to this. In the conventional researches, recognition of a few kinds of simple gestures [8, 2, 11, 12] are mainly focused on as a research targets. For example, Migniot and Ababsa [11] analyze the action of reaching for products. Unfortunately, in such researches, only gesture classification is discussed. However, in the scenario of marketing and security, we have considered that we need to analyze customer behavior more concretely. In other words, for example, in addition to the gesture recognition, we need to estimate which shelf a customer reaches for and which product the customer is interested in. These estimation may help us to consider the marketing plan and security provision.

From this viewpoint, for the first step to analyze cus-

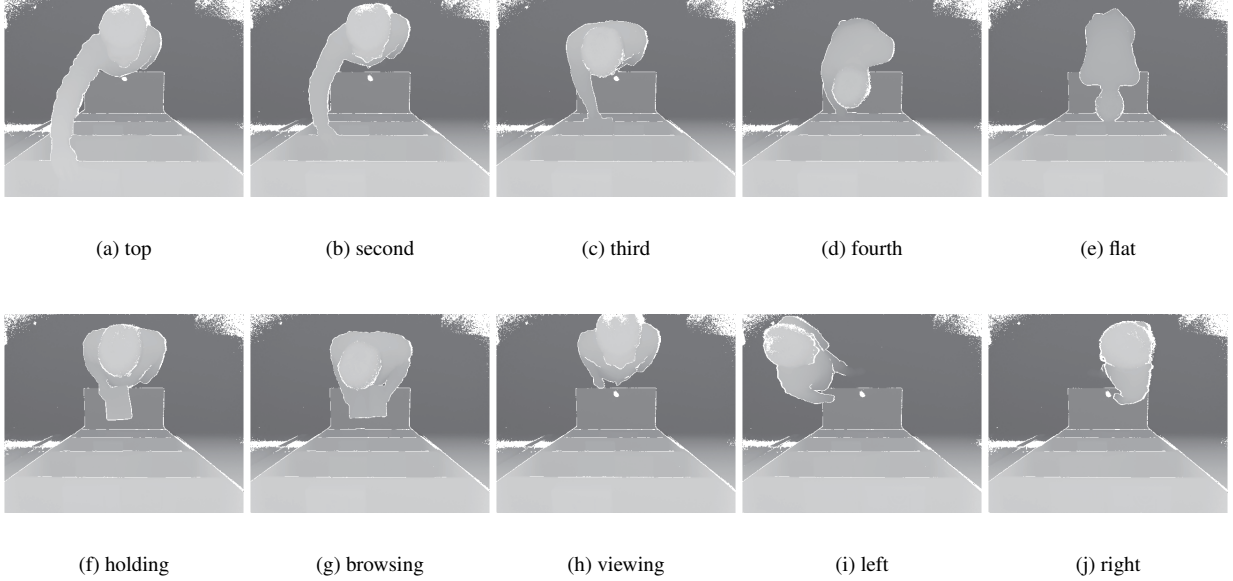


Figure 1. The 10 customer behaviors that we deal with in this researches. (a)-(d): a customer reaches for a book on top-fourth shelf, respectively. (e) flat: a customer reaches for a book on the flat table, (f) holding: a customer weighs up a book by holding it, (g) browsing: a customer browses a book, (h) browsing: a customer views shelves, (i) and (j): a customer moves to from left to right and right to left, respectively.

customer behaviors more concretely, we propose a simple yet efficient behavior analysis system with top-view depth camera and we mainly focus on the evaluation that whether the behavior of a subject is correctly estimated with our proposed method. In this research, by supposing the book store situation, we estimate 10 customer behaviors including the reaching gesture, browsing and weighing up a book, gazing into a shelf, and passing through the front of shelves. Especially, in the reaching gesture, we estimate which shelf the customer reaches for to hold the product as well as the gesture itself. The characteristic points of our proposed method are as follows; 1) we take advantage of height of hand based on depth information to distinguish the shelf that a customer reaches for, 2) a cascaded estimator, which has 2 steps, is utilized for estimate the customer behavior. In the estimator, in the former step, we estimate the behavior based on the depth information. On the other hand, in the latter step, we utilize depth information and PSA (Pixel State Analysis) [6] for extracting human silhouette features from top-view depth image, and employ SVM based estimator. Although the contribution of our method have been little in computer vision area since our study is still exploratory, we empirically confirmed that our proposed method allow us to estimate 10 customer behaviors with 89.5% in our benchmark.

2. Related Work

In the case of customer behavior analysis methods with ordinary surveillance camera, the occlusion is significant problem. To solve this, recently, the behavior analysis with top-view camera [13, 1, 8, 7, 15] or top-view depth camera [11, 14, 10, 5, 3, 2, 12] have been spotlighted as alternative ones. Although the behavior analysis performance is worse since the human shape information is obtained insufficiently with these approaches compared with the frontal view analysis, the occlusion problem is seldom occurred. Besides, the latter approach can relax the privacy and illumination problems. Therefore, we focus on the behavior analysis with top-view depth camera in this research.

In such approach, although there is still small number of researches, we review them briefly in this section. The researches with top-view depth camera is mainly divided into some fields such as human detecting, tracking, and human behavior analysis. In the human detecting and tracking cases, Rauter [14] simply searches the depth map for local maxima to localize head positions and computes trajectories for every head. Choi et al. [5], propose a sliding window based human detection method. In this method, the window is divided into many blocks, and mean and standard deviation per each block is calculated as a descriptor for human detection. Bevilacqua et al. [3] utilize background subtraction procedure based on pixel-wise parametric statistical model for human detection and blobs concatenated



Figure 2. Book shelves utilized in this research.

with human candidate pixels are tracked by using Kalman filter. Similar to [3], in this research, we also simply utilize background subtraction approach for human detection. On the other hand, in the case of human behavior analysis, Bednařík and Herman [2] propose an articulated 3D human model to recognize 6 gestures. Migniot and Ababsa [11, 12] also propose an articulated 3D human model to analyze the gesture that a customer reaches for a product. These methods estimate the position of hands by adopting the human model to top-view depth image and then hands are tracked with particle filter. Therefore, they have reported that, since large number of particle is necessary to track hands accurately, tracking process is time consuming and powerful computation machine such as GPU mounted one is required to speed up this process. Additionally, as mentioned before, although gesture classification is discussed in these conventional researches, they have insufficiently discussed the concrete gesture analysis such as which shelf a customer reaches for a product. We have considered that such analysis is important for marketing and security provision. Therefore, compared with the conventional researches, we tackle such concrete behavior analysis supposing the book store situation in this research.

3. Analyzed Behaviors

Before we explain the process of our proposed method, in this section, let us introduce the customer behaviors that we deal with.

As mentioned before, we suppose a book store situation for analysis of customer behaviors. In this research, we utilize the book shelves which have 4 shelves and a flat table as shown in Fig. 2. By using this environment, we deal with the following 10 behaviors. Figure 1 shows them. In

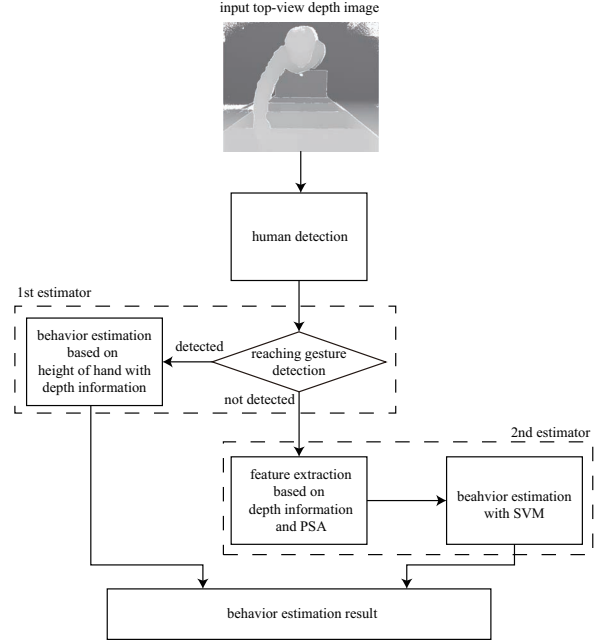


Figure 3. Overview of our proposed method.

Fig. 1, (a)-(d) show the gestures of reaching for a book on the top-fourth shelf. (e) indicates the gesture of reaching for a book on the flat table. In this paper, we describe the above five behaviors (shown in Fig. 1 (a)-(e)) as “top”, “second”, “third”, “fourth” and “flat”, respectively. In this research, each subject reaches for a book on each shelf with right hand. Besides, Fig. 1 (f), (g) and (h) show the behavior that a customer weighs up a book. The differences among them are as follows. As shown in Fig. 1 (f), a customer weighs up a book by holding it and viewing its front/back cover. Figure 1 (g) indicates the browsing behavior that a customer weighs up a book by reading it and (h) shows the viewing behavior that a customer weighs up books or searches them by standing in front of shelves. In this paper, we call these three behaviors (shown in Fig. 1 (f), (g), (h)) as “holding”, “browsing” and “viewing”, respectively. Additionally, as shown in Fig. 1 (i) and (j), we define the behavior that a customer moves to from left to right and from right to left as “left” and “right”, respectively.

4. Proposed Method

In this section, let us explain the concrete process of our proposed method. Figure 3 shows the overview of our proposed method. As shown in Fig. 3, our proposed method is composed 2 estimators; 1) reaching gesture estimation based on height of hand with depth information and 2) other behavior estimation based on PSA as well as depth information. The former estimator allows us to classify the reaching gestures as shown in Fig. 1 (a)-(d), excepting “flat” behavior (Fig. 1 (e)) because hand is often occluded by customer

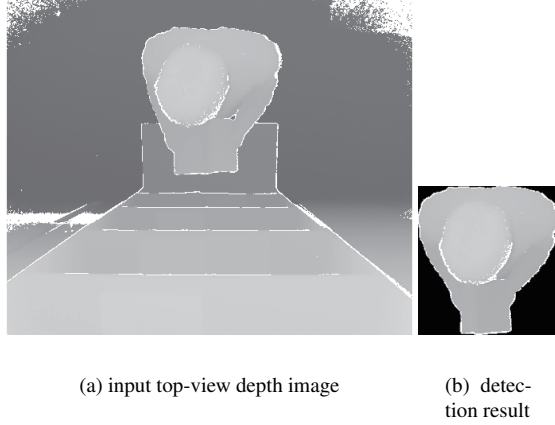


Figure 4. Example of human detection result by using background subtraction. (a) input top-view depth image and (b) result of human detection.

body in “flat” behavior. On the other hand, latter estimator allows us to classify the rest behaviors by extracting depth and PSA based features and by utilizing SVM. Why we utilize two estimator is that the depth information is very useful to estimate behaviors especially, the information of height of hand helps to distinguish the former four behaviors. In the following section, we explain the detail processes.

4.1. Human Detection

In this research, we suppose the top-view depth camera is installed and fixed on the ceiling above book shelves. Therefore, basically the camera field of view is stable. From this, similar to [3], we simply utilize background subtraction for human detection because the depth values in the human area are significantly changed if a human is existed in the field. Figure 4 shows an example of human detection in “browsing” behavior case.

4.2. Estimation of behavior based on height of hand with depth information

In this section, let us explain the behavior estimation method based on height of hand with depth information. As mentioned before, this is the first estimator of our proposed method. Thanks to depth information, in the estimator, we can easily predict the customer behavior of reaching gesture as shown in Fig. 1 (a)-(d). This is based on the observation that the height of each shelf is a known information since the camera is fixed on the ceiling above the shelves and the height of each shelf is calculated with depth camera in advance. The detail process is as follows.

First, we calculate the height of each shelf with a top-view depth camera as well as the area of each shelf is detected in the depth image. Let d_1, d_2, d_3 and d_4 be the depth

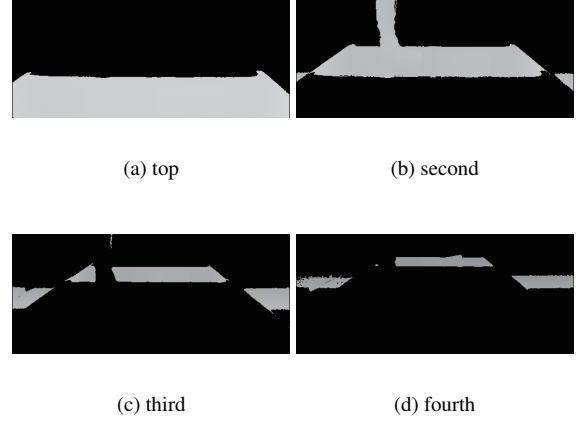


Figure 5. Example of reaching gesture. In this case, a subject reaches for a book on the second shelf.

values of top, second, third and fourth shelf, respectively. If a customer reaches for a book on the second shelf, in the depth image, the pixels having the approximately same depth values with the second shelf are increased. In addition, the third and fourth shelves are occluded by customer’s hand. From this observation, we calculate the increasing and decreasing pixels having similar depth values of d_1, d_2, d_3 and d_4 . In the above example case, as shown in Fig. 5, the pixels having similar depth value of d_2 are increased and those having similar depth values of d_3 and d_4 are decreased due to the hand based occlusion. Therefore, from this depth value checking process, we can predict that a customer reaches for a book on the second shelf. Additionally, if the pixels having similar depth value of d_1, d_2, d_3 and d_4 are not increased/decreased, we regard that a customer does not reach for a book on the shelves. In this case, we move to the second estimator to predict other customer behaviors. As mentioned above, thanks to the depth information, although our proposed method is very simple, it allows us to estimate reaching behavior precisely.

4.3. Estimation of behaviors based on SVM

After the first estimator is applied, if the reaching gesture excepting “flat” behavior is not detected, we utilize the second estimator. This is composed two parts; 1) feature extraction part based on depth information and PSA and 2) behavior estimation part with SVM. In the following, we explain the detail process of second estimator.

First, we normalize the size of human detected depth image as 256×256 . In the following, we describe the normalized human detected image as $I_{\text{norm}}(i)$ where i is the frame ID of captured movie with a top-view of depth camera. From $I_{\text{norm}}(i)$, we extract two features. In order to extract the depth based features, we divide $I_{\text{norm}}(i)$ into $m \times m$ blocks. Then, we calculate the average depth value



Figure 6. Example of pixel states by analyzing with PSA. Green, red and black pixels indicate stationary, transient and background states, respectively.

from each block. From this process, we can obtain the depth based features whose number of dimension is m^2 . This feature is very simple yet effective to describe the pose information.

On the other hand, we extract PSA features from $I_{\text{norm}}(i)$ and s frames before and after $I_{\text{norm}}(i)$, namely, $2s + 1$ frames are utilized for PSA feature extraction. PSA is a one of pixel analysis methods that can classify whether each pixel is stationary or transient by observing its intensity over time. In our proposed method, by utilizing PSA [6], we determine whether each pixel of $I_{\text{norm}}(i)$ is stationary or transient as well as background by using $2s + 1$ frames. Figure 6 shows an example of pixel states by analyzing with PSA. In this figure, green, red and black pixels indicate stationary, transient and background states, respectively. After determined the pixel state, similar to depth feature extraction, we also divided $I_{\text{norm}}(i)$ into $m \times m$ blocks. Then, we create a pixel state histogram, which has 3 bins, by checking each pixel in the window of $n \times n$ ($n = \{1, 2, \dots, m\}$) blocks. By sliding this window, finally, we can obtain PSA features whose number of dimension is $3 \times \sum_{j=1}^m j^2$.

After feature extraction process, we simply utilize 1 vs. 1 rule SVM to distinguish 6 behaviors as shown in Fig. 1 (e)-(j). We concatenate depth based features and PSA based features and the concatenated one is utilized as an input of SVM.

5. Experiments

We have evaluated our proposed method with 3 subjects. Each subject was captured each behavior one by one. In the experiments, for the first step to evaluate whether the behavior of a subject is correctly estimated with top-view camera, we suppose only one subject is captured in the depth image.

5.1. Experimental Settings

In this research, we utilized Microsoft Kinect for Windows v2 as a top-view depth camera whose resolution is 512×424 and it was installed on the ceiling above the shelves as shown in Fig. 2. Besides, subject A, B and C

whose heights were 183[cm], 177[cm] and 172[cm], respectively were employed in this experiment. Each subject was instructed to act each behavior one by one and was captured with Kinect. We totally collected 21,340 frames for 3 subjects. In the experiments, we set the parameters as follows; $m = 8$ and $s = 7$. Therefore, the number of dimensions of depth and PSA based features were 64 and 612, respectively. We employed the computational machine with Intel core i7-4790K CPU and 8[GB] RAM. Besides, we utilized LIVSVM [4] as SVM software library and RBF kernel for behavior estimation. In this research, 3 cross validation is employed for estimation, i.e., the frames captured with 2 subjects and that with 1 subjects were employed for training and for testing, respectively, and we conducted the following experiments by changing the testing subject.

5.2. Effectiveness of Features

Before we show the effectiveness of our proposed method, first, we have evaluated the effectiveness of depth based features and PSA based features. In this experiment, in order to investigate it, we utilized three variation of features, i.e., only depth based features, only PSA based features and their combination. In addition, in this experiment, we did not utilize the cascaded estimator, namely, all behaviors were estimated with SVM.

Figure 7 shows the experimental results. As shown in Fig. 7, we have confirmed that PSA features have a discriminative power insufficiently to distinguish the reaching gestures because the PSA features measure whether pixels are stationary or transient and depth information is not included in the features. In addition, from Fig. 7, the performance of “flat” behavior is the worst compared with others. In this experiment, there were many cases that the testing “flat” behaviors were erroneously estimated as “fourth” behaviors since the silhouette of both behaviors were very similar. Therefore, we have considered that more discriminative features are necessary to improve the estimation performance. Although we have some problems mentioned above, from Fig. 7, we have confirmed that the feature combination allows us to estimate the behaviors accurately compared with others and we have obtained the average accuracy of 65.1%, 65.0% and 71.1% with depth based feature, PSA based feature and their combination, respectively.

5.3. Effectiveness of the cascaded estimator

Next, we have evaluated our proposed cascaded estimator. In this experiment, we compared the behavior estimation performance with our proposed method and that with the method without cascade architecture utilized in the experiment mentioned in the previous section. In addition, the combination of depth and PSA based features were utilized as input for SVM.

Figure 8 shows the experimental results. As shown in

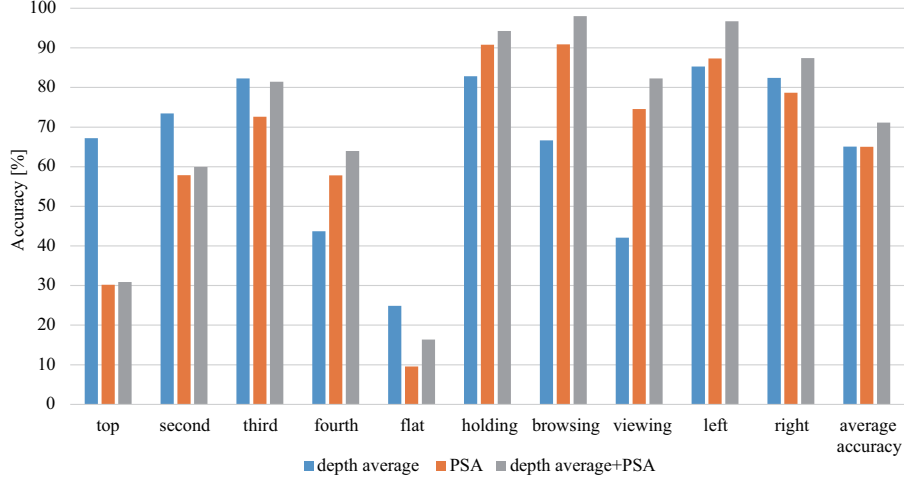


Figure 7. Experimental results of behavior estimation accuracy with each feature.

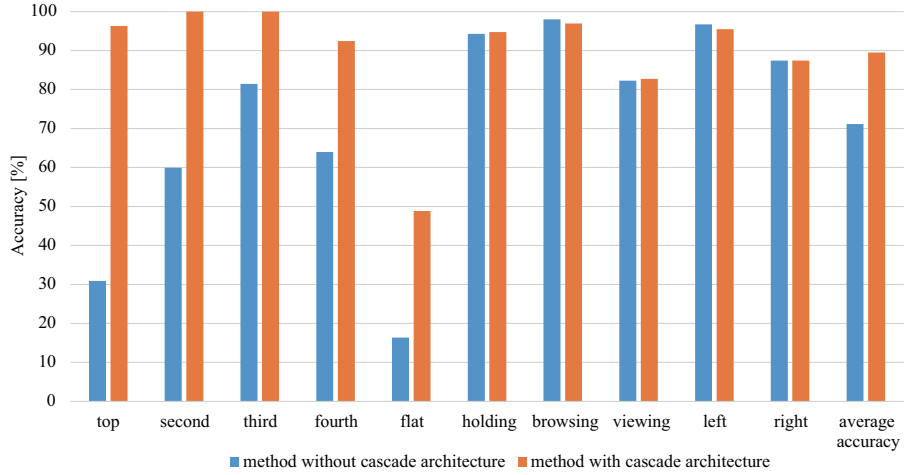


Figure 8. Experimental results with the method with and without cascaded architecture

Fig. 8, we have confirmed that our proposed method enables us to improve the behavior estimation performance. Especially, the reaching behaviors (“top”, “second”, “third” and “fourth”) were significantly improved the performance. This is because we utilize the depth information of height of shelves efficiently. From the experimental results, we have achieved 89.5% of average estimation accuracy with our proposed method. Even though the performances were totally improved, however, we achieved still 48.9% of the estimation accuracy for “flat” behavior as shown in Fig. 8. In this experiment, there were many cases that the testing “flat” behaviors were erroneously estimated as “browsing” behaviors though they were distinguished from “fourth” behaviors by introducing the cascade architecture. We have considered that this error is occurred based on the image normalization and the similarity of silhouettes of “flat” and “browsing” behaviors. In other words, we have considered that since the silhouettes of “flat” and “browsing” behaviors

after image normalization are relatively similar compared with others, this similarity affects the performance. Similar to the experiment mentioned in the previous section, in order to solve the above problem, we have considered that the features having more discriminative power are required and this is one of our future work.

5.4. Investigation of behavior estimation accuracy for each subject

Finally, we have evaluated the behavior estimation accuracy for each subject with our proposed method. In this experiment, we have investigated the effect of behavior difference among individuals for estimation.

Figure 9 shows the experimental results. As shown in Fig. 9, we have achieved 83.7%, 88.2% and 96.6% of behavior estimation accuracy with subject A, B and C, respectively. The estimation performance of subject A was worse compared with other subjects. Especially, the performance

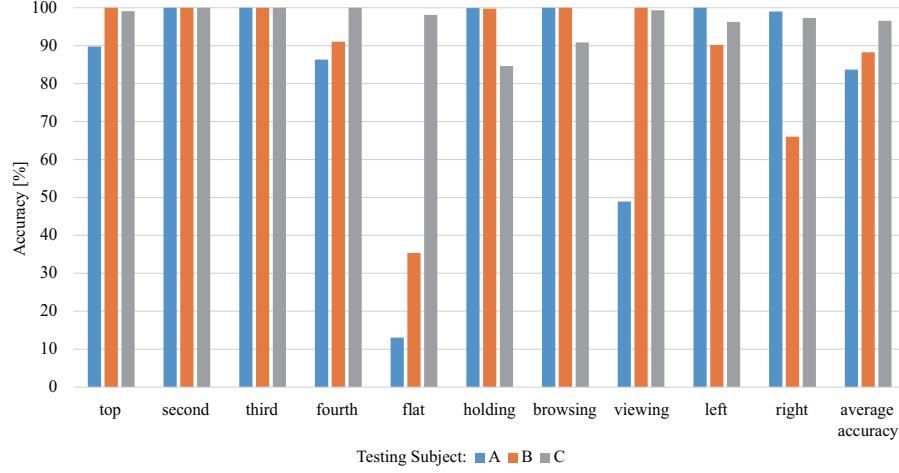


Figure 9. Experimental results of behavior estimation accuracy with each subject.

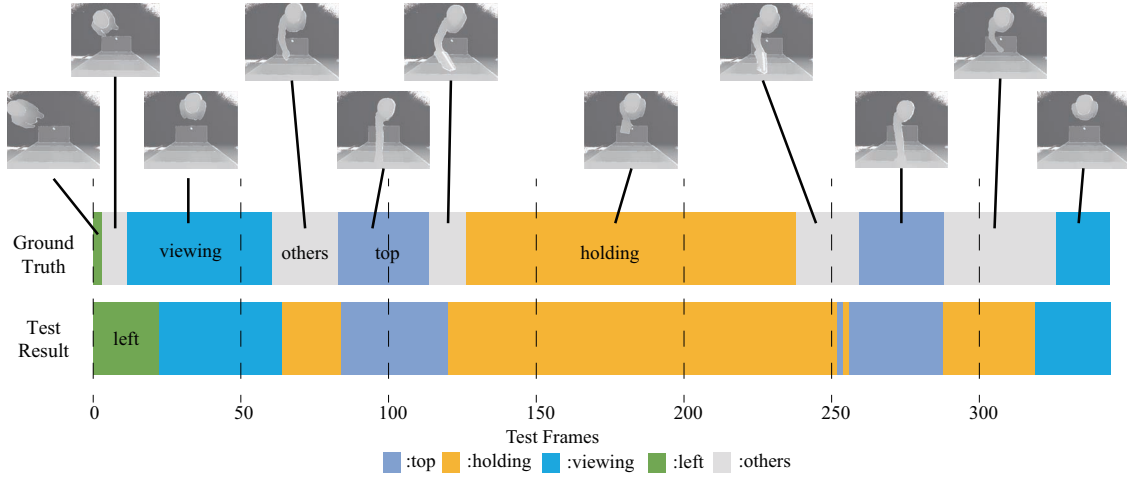


Figure 11. Experimental result by sequentially analyzing behaviors with subject C.

of “flat” behavior for subject A was significantly worse. As shown in Fig. 10, the “flat” behavior silhouette of subject A after size normalization was different from the rest, i.e., the head position of subject A in the normalized image was different from the rest. Since we utilize the block division strategies for feature extraction, our proposed method is not robust to rotation of human body. Therefore, from the experimental results, we have confirmed that the performance of our proposed method is affected by the personal habit. In the experiments, only 3 subjects are utilized in this experiment. Therefore, to solve the personal habit problem, we have considered that we need to investigate the performance of our proposed method with more subjects and this is also one of our future work.

Moreover, we have investigated the estimation performance by sequentially analyzing the behaviors. In this experiment, for testing, we employed subject C who allows us to obtain best performance in the experiment mentioned

above and we prepared the following scenario; 1) subject C came to shelves from left side and viewed the shelves, 2) after that subject C reached for a book on the top shelf and checked the book covers by holding it, 3) then returned it to the top shelf. In this scenario, there were ambiguous behaviors that were intermediate ones from a behavior to another. We define such ambiguous behaviors as “others” behavior.

Figure 11 shows the experimental result by sequentially analyzing behaviors with subject C. As shown in Fig. 11, we have confirmed that our proposed method enables us to roughly obtain the correct estimation for 10 behaviors. However, since our proposed method has no estimation strategy for ambiguous behaviors, we obtained totally 63.5% of behavior estimation accuracy. Therefore, we need to improve our proposed method to be able to deal with such ambiguous behaviors. This is also one of our future work.

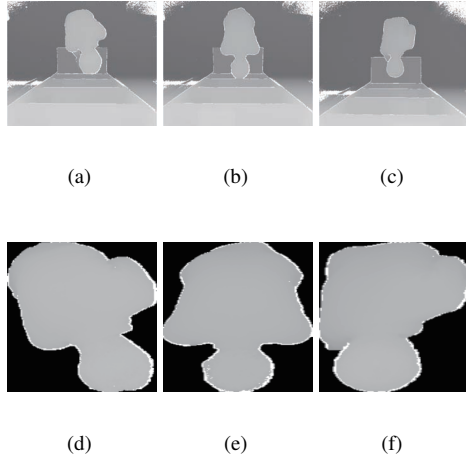


Figure 10. Examples of “flat” behavior of each subject. (a), (b) and (c) show the top-view depth image of “flat” behaviors of subject A, B and C, respectively. (d), (e) and (f) show the normalized human region depth image of subject A, B and C, respectively.

6. Conclusion

In this paper, we proposed a new human behavior analysis method with a cascaded architecture, which have height of hand with depth information based estimator and depth information and PSA based estimator. From the experimental results supposing the book store situation, we have confirmed that our proposed method allows us to estimate 10 behaviors with 89.5% of average accuracy.

Future work is to investigate the more discriminative features in order to improve the estimation performance and to evaluate our proposed method with more subjects. Additionally, our proposed method is relatively specialized for book store situation. Therefore, we need to evaluate our proposed method in other situation such as convenience store. Furthermore, in the case of concrete scenario at store, there are possibilities that two or more customers are captured simultaneously and ambiguous behaviors that we have not defined are performed. However, our proposed method is not considered the situations. Consequently, we need to improve our proposed method to estimate behaviors of two or more subjects simultaneously and ambiguous behaviors. This is also one of our future work.

References

- [1] S. Ardeshtir and A. Borji. Egocentric Meets Top-view. In *arXiv:1608.08334v2*, 2016.
- [2] J. Bednařík and D. Herman. Human gesture recognition using top view depth data obtained from Kinect sensor. In *Proc. of Excel@FIT*, 2015.
- [3] A. Bevilacqua, L. D. Stefano, and P. Azzari. People tracking using Time-of Flight depth sensor. In *Proc. of International Conference on Video and Signal Based Surveillance*, pages 89–89, 2006.
- [4] C. C. Chang and C. J. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Thchnol.*, 2(3):1–27, 2011.
- [5] T.-W. Choi, D.-H. Kim, and K.-H. Kim. Human Detection in Top-View Depth Image. *Contemporary Engineering Sciences*, 9(11):547–552, 2016.
- [6] H. Fujiyoshi and T. Kanade. Layered Detection for Multiple Overlapping Objects. *IEICE Transactions on Information and Systems*, 87(12):2821–2827, 2004.
- [7] N. Hu, G. Englebienne, and B. Kröse. Posture Recognition with a Top-view Camera. In *Proc. of International Conference on Intelligent Robots and Systems*, pages 2152–2157, 2013.
- [8] M. Kolesnik and T. Kuleba. Detecting, Tracking and Interpretation of a Pointing Gesture by an Overhead View Camera. In *Pattern Recognition, Lecture Notes in Computer Vision*, pages 429–436, 2001.
- [9] J. Liu, Y. Gu, and S. Kamijo. Recognition of Customer Behavior on the Front of Shelf from Surveillance Camera. In *Proc. of IEICE Technical Report, PRMU2015-82*, pages 97–102, 2015.
- [10] F. MalaWski. Top-view People Counting in Public Transportation using Kinect. *Challenges of Modern Technology*, 5(4):17–20, 2014.
- [11] C. Migniot and F.-E. Ababsa. 3D Human Tracking from Depth Cue in a Buying Behavior Analysis Context. In *Proc. of International Conference on Computer Analysis of Images and Patterns*, pages 482–489, 2013.
- [12] C. Migniot and F.-E. Ababsa. 3D Human Tracking in a Top View Using Depth Information Recorded by the Xtion Pro-Live Camera. In *Proc. of 9th International Symposium on Visual Computing*, pages 603–612, 2013.
- [13] O. Ozturk, T. Yamasaki, and K. Aizawa. Tracking of Humans and Estimation of Body/Head Orientation from Top-view Single Camera for Visual Focus of Attention Analysis. In *Proc. of International Conference on Computer Vision Workshop*, pages 1020–1027, 2013.
- [14] M. Rauter. Reliable Human Detection and Tracking in Top-View Depth Images. In *Proc. of Computer Vision and Pattern Recognition Workshop*, pages 529–534, 2013.
- [15] J. Serrano-Cuerda, J. C. Castillo, and A. Fernández-Caballero. Indoor Overhead Video Camera for Efficient People Counting. *Jurnal Teknologi*, 63(3):17–22, 2013.
- [16] A. Yao, J. Gall, G. Fanelli, and L. V. Gool. Does Human Action Recognition Benefit from Pose Estimation? In *Proc. of 22nd British Machine Vision Conference*, pages 67.1–67.11, 2011.
- [17] S. Yoshinaga, A. Shimada, and R. ichiro Taniguchi. Real-time people counting using blob descriptor. In *Proc. of the 1st International Conference on Security Camera Network, Privacy Protection and Community Safety*, pages 143–152, 2009.
- [18] zhichen Zhao, H. Ma, and S. You. Single image action recognition using semantic body part actions. In *arXiv:1612.04520v1*, 2016.