An Event-Based Hierarchical Method for Customer Activity Recognition in Retail Stores

Jiahao Wen, Luis Guillen, Alfian Amrizal Muhammad, Toru Abe and Takuo Suganuma

October 13, 2020

An Event-Based Hierarchical Method for Customer Activity Recognition in Retail Stores

Jiahao Wen[1], Luis Guillen[2], Muhammad Alfian Amrizal[2], Toru Abe[3], and Takuo Suganuma[3]

[1] Graduate School of Information Sciences, Tohoku University, Japan
[2] Research Institute of Electrical Communication, Tohoku University, Japan
[3] Cyberscience Center, Tohoku University, Japan

Abstract. Customer Activity (CA) provides valuable information for marketing. CA is a collective name of customer information from on-the-spot observation in retail environments. Existing methods of Customer Activity Recognition (CAR) recognize CA by specialized end-to-end (e2e) models. Consequently, when marketing requires changing recognition targets, specialized e2e models are not reconfigurable to fit different marketing demands unless rebuilding the models entirely. Besides, redundant computation in the existing CAR system leads to low efficiency. Also, the low maintainability of the CAR system results in lots of modifications when updating methods in the system. In this research, we decompose behaviors into several primitive units called "event." We propose an event-based CAR method to achieve reconfigurability and design a hierarchy to solve issues about redundancy and maintainability. The evaluation results show that our proposed method can adapt and perform better than existing methods, which fits different marketing demands.

Keywords: Retail environments · Customer activities · Activity recognition · hierarchical activity model.

1 Introduction

1.1 Background

In retail stores, various customer information is required to support marketing. Traditional retail only uses the purchase information to analyze purchasing behavior from the records [1]. However, purchase records only reveal results instead of the process of deciding the purchase. As this process would probably reveal the reasons for purchase, it is valuable for marketing analysis. Therefore, a solution called "smart retail" installs ubiquitous cameras to collect the shopping process data. Moreover, despite the large size of real-time data, machine learning models can efficiently and accurately handle them to get information about the shopping process [1–5]. In this research, the shopping process information is named "Customer Activity" (CA). CA is a collective name of customer information from on-the-spot observation in retail environments. Usually, CA contains information, among others, of the customer's location, trajectory, behavior.

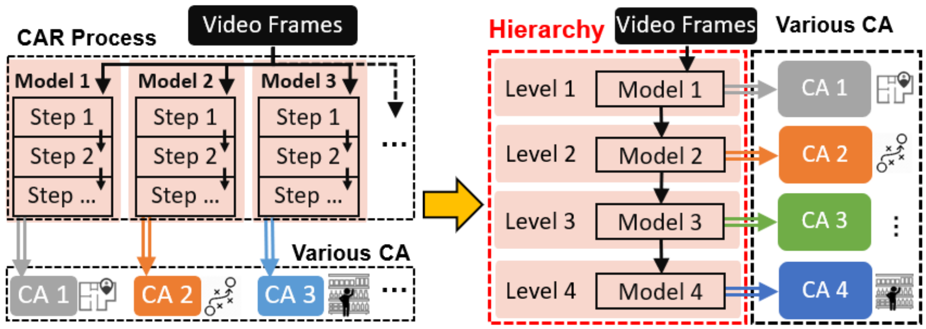

Fig. 1. Structure of Existing CAR System(left) and Proposed CAR System(right)

1.2 Problems in Existing CAR system

Many methods [6] have been proposed for Customer Activity Recognition (CAR). They recognize various CAs mainly from the visual input [1–5, 7–13]. However, none of them describe the structure of their CAR systems. Therefore, there is no way to infer the compatibility with other CAR systems, as shown on the left side of Figure. 1. Since the CAR system serves marketing, it should satisfy the malleability of different demands.

Problem 1: Models are Not Reconfigurable
The recognition targets vary in different development phases and conditions. Thus, different CAR results are required. However, these CAR methods mainly use machine learning-based end-to-end (e2e) models to recognize CA from video frames, specialized e2e models make the model impossible to be reconfigured for different target recognition. For instance, a model can recognize if a customer is selecting products. But if the requirements change so that the model is required to recognize whether a customer is selecting products either by a single or both hands, the original model is no longer useful. Consequently, new training data should be collected to train the model again, which is time-consuming.

Problem 2: System has Redundant Computation
As shown in Figure. 1, all models share the same visual input. Therefore, there might be some similar steps in different models during processing video frames. For example, common models for customer behavior recognition include locating customers in the frame, which is also computed by the people detection model. Thus, the computation of locating customers is repeated. This leads to redundancy in the CAR system, which results in low efficiency.

Problem 3: System has Low Maintainability
To achieve better performance of the CAR system, it is necessary to update the system to utilize different methods in different situations. But, the existing CAR systems do not have a clear division of tasks in their CAR models, and the output CA is not uncategorized, which leads to the existence of some common steps in

different models. Therefore, updating methods in the system should modify all these common steps, which results in the system's low maintainability.

1.3 Solutions

In this paper, we categorize output CA in existing methods. According to the CA category, we design a four-level hierarchy, as shown in Figure. 1, which specifies the division of CAR tasks. While the hierarchy solves Problems 2 and 3, we propose a reconfigurable behavior recognition method based on the hierarchy to solve Problem 1.

Solution to Problem 1: Though CAR of every type of CA should be reconfigurable, in this research, we only achieve a reconfigurable method for customer behavior recognition. To design a reconfigurable model in this structure, we decompose a complex behavior into the permutation and combination of several primitive units, called "event." The proposed method recognizes events instead of behaviors. Therefore, redefining the permutation and combination of events can easily reconfigure behavior outputs.

Solution to Problem 2: Visual data is processed level by level in the hierarchy in Figure. 1. Each level receives its previous level's output as its input. Therefore, every level processes different types of data, which means no repeated process. Thus, redundant computation can be avoided by this hierarchy.

Solution to Problem 3: According to Table. 1, we categorize various CAs into several levels, which provides a clear division of tasks. This avoids common steps between models. Therefore, updating any level does not influence the other levels. In other words, the hierarchy solves Problem 3.

We evaluate the hierarchy by adapting to three different marketing demands. The results show that our proposed hierarchy can fit different marketing demands easier than existing methods.

2 Related Work

The related work on CAR is categorized in Table. 1 by the output of those methods. It reveals that current methods mostly detect/recognize the object's location, movement, and customer behavior. The column of "Content" refers to the detailed classification of outputs. These methods achieved good performance in their particular cases. However, once the CAR system is built with these methods to get outputs of each category, it forms the structure on the left side of Figure. 1, because of incompatible methods and sharing the same visual data as input. Therefore, it leads to the three problems explained above.

For methods in the category "Customer Behavior," their models require time-consuming training, and the trained models are specialized, which means Problem 1 "Models are not reconfigurable."

Table 1. Outpus CAs in Existing Methods

Category	Content	Related Methods
Object's Location	Body (region of the whole body)	[1–3, 7–19]
	Body Part (hands, arms)	[1, 5]
	Other Object (product, basket, etc.)	[1, 5]
Object's Movement	Object/Pixel's motion feature	[1–3, 5, 10, 11]
	Object's trajectory	[13]
Customer Behavior	Passing by / No interest	[1, 10]
	Viewing the shelf	[1, 3]
	Turn to the shelf	[1]
	Pick a product from shelf	[1, 2, 10, 11, 13]
	Pick nothing from shelf (Touch)	[1, 2, 11]
	Return a product to the shelf	[1, 10, 11]
	Put a product into cart/basket	[1]
	Holding a product	[3]
	Browsing a product on the hand	[2, 3]
	Fit next to you & Check how it looks & try on & take off (in clothes shop)	[2]

On the other hand, most methods in the category "customer behavior" contain the computation of locating and tracking customers, which has been done in methods of the other two categories. This leads to Problem 2 "System has redundant computation."

In addition, these methods do not have a clear division of tasks, which results in incompatibility with other methods. In [3], it even regards the location of hand which is position information as customer behavior because the output CAs are not categorized. Therefore, updating methods in a system consists of these methods is difficult. This refers to Problem 3 "System has low maintainability."

3 Proposal

To solve Problems 2 and 3, we design a hierarchical structure to specify the division of CAR tasks and categorize CAs. After the hierarchy, we introduce our proposed method for reconfigurable behavior recognition, which is the solution to Problem 1.

3.1 Hierarchy

In Table. 1, we categorize the output CAs and notice that existing CAR methods mainly output object location, object trajectory, and customer behavior. Apart from these three types of outputs, we add a new data type called "event," representing primitive units of behaviors. With "event" and three types of data of existing methods, we design a hierarchy in Figure. 2. It includes "Single Frame," "Consecutive Frame," "Event," and "Behavior." Among them, level 1 "Single

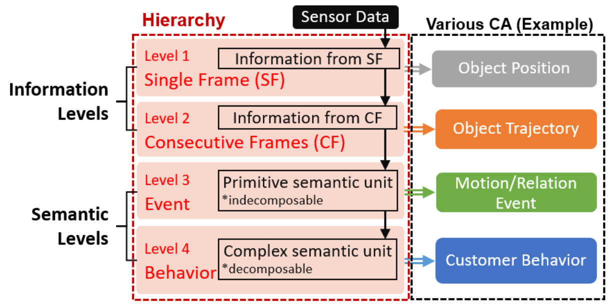

Fig. 2. Proposed Hierarchy

Frame," which is in the top place, is the most basic. Data flows through levels from top to bottom. Therefore, each level receives output CA from its previous level as the input.

In the hierarchy, we assume that level 1 and 2 are information levels, which extract objective information from the sensor data. Level 3 and 4 are semantic levels, which learn to understand the information from level 1 and 2 subjectively. Therefore, level 1 and 2 output data of successive values, which represent the data observed from the objective world, while level 3 and 4 provide discrete values, which compress the redundant part of data together to show the comprehension of the objective information.

Level 1 "Single Frame:" This most basic level of the hierarchy refers to the information that can be extracted from a single frame. For example, the output CA can be the objects' position. The object refers to any object in the frame, such as the content in Table. 1. Thus, the model for extracting information from a single frame is supposed to be implemented into this level. For the case of object detection, a model of detecting objects from sensor data should be implemented.

Level 2 "Consecutive Frame": This level refers to the information extracted from consecutive frames, such as the content in Table. 1. Compared to level 1, this level tends to find the relation of data, which has been extracted in level 1, between several frames. Usually, the pixel-based method, such as optical flow estimation, and the coordinate-based tracking method are implemented into this level.

Level 3 "Event:" This level is inspired by [20] that defines human behavior as a composition of multiple events. An event refers to a single low-level spatiotemporal entity that cannot be further decomposed. Thus, we utilize the name "event" and define it as a primitive semantic unit that cannot be decomposed. In this research, an event can be the motion of one object or relation between several objects. Moreover, behavior is regarded as the composition of several events. For instance, the behavior "Pick up a product from the shelf" can be decomposed as the motion "put hands into the shelf" happens at first, then, the motion "take out hands from the shelf" and the relation "a product is following hands" occur concurrently. These motions and relations are called "Event." Therefore, there are two types of events, motion events (ME) represent the motion of an object, and relation events (RE) describe the relation of two objects.

Level 4 "Behavior:" Compared to the primitive unit "event," the behavior is defined as a complex semantic unit that can be decomposed as several events. As its name implies, this level is supposed to do tasks about behavior recognition.

Compared to the existing CAR system's structure, this hierarchy specifies a clear division of tasks. Level 1 and 2 are responsible for extracting objective information from sensor data. Then, level 3 and 4 comprehend the objective information to get a semantic understanding of the data. Each level outputs one type of CA. Figure. 2 shows the example outputs of each level. Each level processes different data types, and the computation for a particular type of CA is integrated into one level. Thus, this design avoids redundant computation and makes the methods in different levels independent. Therefore, updating the methods of any level does not influence the other levels, which shows high maintainability.

Besides, the proposed hierarchy does not specify methods in each level. Therefore, as long as the method completes each level's work, any method can be utilized in this level.

3.2 Event-Based Method: Level 3

The hierarchy solves Problems 2 and 3. In the hierarchy, we define the behavior recognition as recognizing events and matching a particular combination of events to recognize the behavior. To solve Problem 1, we propose a reconfigurable customer behavior recognition method based on events. We assume the input is objects' trajectories, which is the objective information from level 2. Therefore, our method is supposed to comprehend trajectory information to get MEs and REs. Understanding objective information to get semantic units requires a process to transfer successive values into discrete values. Thus, we separate the method into trajectory segmentation and symbolization.

Trajectory Segmentation: It means dividing the trajectory into several segments. If a part of the trajectory has a similar direction, it is regarded as a segment. For MEs and REs, different methods of trajectory segmentation are applied due to their different properties.

For MEs, they require that each segment has a similar direction. Thus, we apply DynMDL [21] to divide the trajectory by direction. For REs, the segment should include a similar part of two trajectories. Therefore, we design a value named "difference" to describe the similarity of two trajectories. We accumulate each corresponding point's distance in two trajectories. The average accumulated distance is considered as the difference between the two trajectories. The trajectories are divided according to their difference value. Additionally, to reduce computation, we only handle the trajectories of some specific objects.

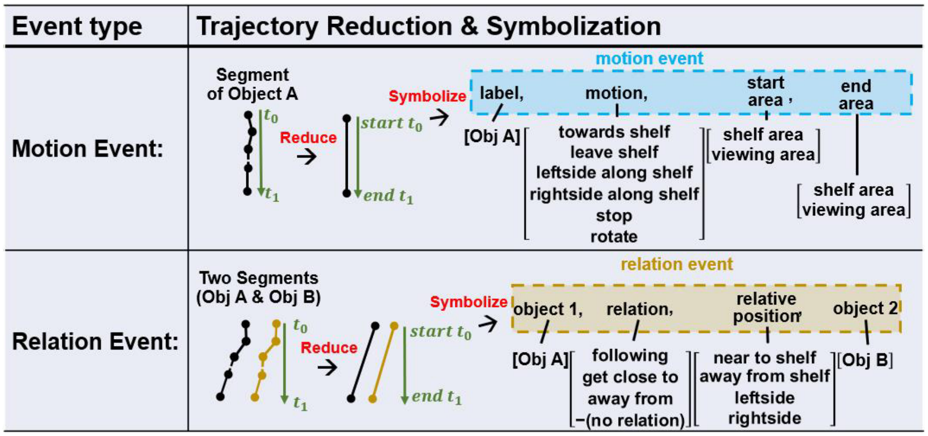

Fig. 3. Trajectory Reduction and Symbolization

Trajectory Symbolization: The object moves in a similar direction in each segment. To compress the information and make it readable, we symbolize the segment into a four-dimension vector. As shown in Figure. 3, each segment is firstly reduced to a vector of the start and end. Then, the vector is symbolized to get MEs and REs.

(ME)Label: An object's identity, which can be the name or id of an object.

(ME)Motion: An object's moving direction. Thus, it is determined by the vector's direction. The "left-side/right-side" is one's left/right side when facing the shelf. These symbols represent the direction relative to the shelf because a behavior consists of events, and all the behaviors in existing methods are related to the product shelf.

(ME)Start Area & End Area: The located area of the reduced vector's endpoints. Each frame is divided into two areas, viewing area (VA) and shelf area (SA). VA includes the area where the customer is close enough to the shelf to interact with products on the shelf. SA includes the region of the whole shelf and all products on the shelf.

(RE)Label of object 1 & 2: The label of two objects.

(RE)Relation: The relation of two objects. We design three types of relations. However, only the "following" is implemented currently, which means an object is following another one. Another two relations refer to an object is getting close to / away from another one.

(RE)Relative Position: The relative position of two objects. The product shelf is the reference in the comparison. Compared to object 2, object 1 can be nearer to / away from the shelf. When the distance to the shelf is similar, object 1 can be on the left-side/right-side of object 2 from the shelf's view.

With MEs and REs, trajectory information from level 2 is comprehended as readable symbols which are discrete values. Compared to unexplainable numerical features, these readable symbols make it easier to explain or define behaviors.

Table 2. Formulated Definition of behavior "pick a product"

Behavior	Event Pattern ([*] referes to any value)
Pick a product (Order:1→2&3)	1. ME: [hand], [towards shelf], [*→SA] 2. ME: [hand], [leave shelf], [SA→*] 3. RE: [product A], [following], [*], [hand]

3.3 Event-Based Method: Level 4

To achieve reconfigurable behavior recognition, we decompose behaviors into events and match these events instead of recognizing behaviors from pixel features. Therefore, changing the composed events can easily reconfigure behavior recognition. In level 4, we predefine behaviors as event patterns, and match the composed events from the results of level 3.

In Figure. 4, yellow blocks show that each behavior is predefined as a particular pattern of events. "→" specifies the chronological order of events. Events on the left side of the arrow happen at first. "&" means adjacent events occur concurrently. To arrange all events, we use the timeline because events happen in chronological order. Figure. 4 shows that we match event patterns from events in the timeline to recognize behaviors.

Table. 2 shows the formulated definition of behavior "pick a product." Firstly, ME 1 occurs, which is explained as "A hand is moving towards shelf from any area to the shelf area." Then, ME 2 happens, which represents "A hand is leaving the shelf from the shelf area to any area." Concurrently, RE 3 happens, which means "The product A is following hand at any relative position." Figure. 4 shows an example of matching "pick a product" from timelines. For the timeline, the latest event $i+2$ matches ME 2, and $i+3$ matches RE 3. Since the latest events are matched, the algorithm continues matching previous events. Then, ME 1 is matched by event i. The behavior "Pick a product" is recognized.

Behaviors are decomposed as events. Therefore, even with the same events, we can recognize another behavior by changing the order of the composed events.

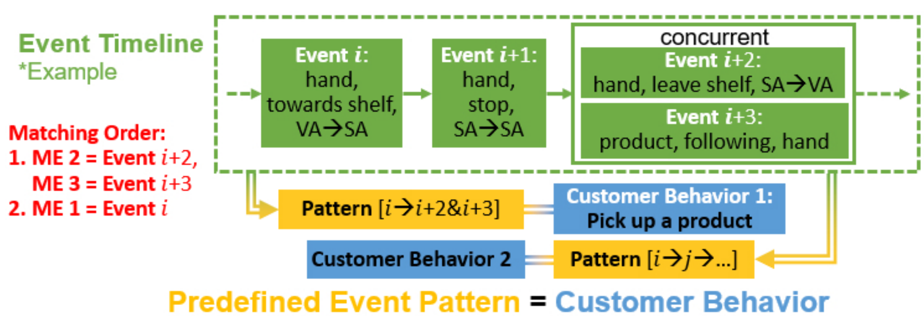

Fig. 4. Define & Match Customer Behavior by Event Pattern

Namely, with this event-based method, we achieve a reconfigurable customer behavior recognition, which solves Problem 1.

4 Evaluation

4.1 Implementation Experiment

For level 1, we use top-view video as the input to avoid occlusion. Unfortunately, we did not find any existing dataset about the top-view video for customer behaviors. Therefore, we build a retail laboratory environment to collect top-view videos to evaluate our proposed method. Figure. 5 shows that an RGB camera takes images from the above of the shelf. Camera frames are divided into the viewing area and shelf area. Videos are collected at a public activity, where random participants are requested to pick at least one provided product from the shelf. Besides, we assume that each participant does not interact with another person. Eventually, 19 videos from 19 participants, including 10648 frames with the FPS of 30, are collected and manually labeled for evaluation. Considering the various products, we choose four kinds of products which include two hard objects and two soft objects. Finally, person, hand, and four products are labeled.

4.2 Evaluation Steps

To evaluate our proposed event-based method without the influence of level 1 and 2, we use the labeled bounding boxes as the output of level 1 instead of training an object detection model. Namely, level 1 is skipped in evaluation steps. To evaluate our proposed method's performance in meeting different marketing demands, we designed three steps to simulate the change in marketing demands.

Step 1. Recognize six behaviors of existing methods
Six common behaviors are chosen from existing methods in step 1 of Table. 3. The

Fig. 5. Camera installation and input image

Table 3. Event Patterns of Six Common Behaviors

Step	Behavior	Event pattern ([*] referes to any value)
Step 1	Walking: walking in VA (1)	1. ME: [person], not [stop], [VA→VA]
	Viewing: stop and view products (1)	1. ME: [person], [stop], [VA→VA]
	Browse: browse a product on hands (1,2)	1. ME: [person], [*], [VA→VA] 2. RE: [A], [following], [*], [hand]
	Pick: pick a product out of shelf (1→2,3)	1. ME: [hand], [towards shelf], [*→SA] 2. ME: [hand], [leave shelf], [SA→*] 3. RE: [A], [following], [*], [hand]
	Touch: pick nothing out of shelf (1→2)	1. ME: [hand], [towards shelf], [*→SA] 2. ME: [hand], [leave shelf], [SA→*]
	Return: return a product to shelf (1,2→3)	1. ME: [hand], [towards shelf], [*→SA] 2. RE: [A], [following], [*], [hand] 3. ME: [hand], [leave shelf], [SA→*]
Step 2	Selecting: select products in SA (1)	1. ME: [person], [*], [*→SA]
Step 3	Select by one hand (1,2)	1. ME: [person], [*], [*→SA] 2. ME: [hand], [*], [VA→VA]
	Select by both hands (1)	1. ME: [person], [*], [*→SA]

symbol "A" refers to the product A. For existing methods, these six behaviors should be annotated to train models to do recognition. In our proposed method, only four events shown in Table. 3 in red color are required to define these behaviors. ME 1 in "browse" can be regarded as the union set of ME 1 in "walking" and ME 1 in "viewing." The F1 score of the proposed method reaches 94.71% (Total 153 behaviors, precision=95.32%, recall=94.11%).

Step 2. Add a new behavior "selecting"
Suppose that the marketing demand requires recognizing a new behavior "selecting," which is defined as the customer is selecting products with hands in SA, but there is still nothing in his hand. Existing methods should collect new training data to train their models again. However, we only need to define this new behavior as Table. 3. "Selecting" is defined as "The whole person's region moves from any area to shelf area." We use the region of the whole person (including arms) because the region of the hand may be invisible when it is occluded by the shelf. The F1 score of the proposed method reaches 95.10% (Total 178 behaviors, precision=95.79%, recall=94.45%).

Step 3. Extract details from "selecting"
In this case, details are required about "selecting," such as selecting products by one hand or both hands. Existing methods require re-training the model with new training data. However, we only need to redefine event patterns as Table. 3. If a hand is found outside the shelf when the whole person's region (including arms) is still in SA, it means another hand is inside the shelf, namely "select by one hand." Otherwise, it is "select by both hands." The F1 score reaches 94.94% (Total 181 behaviors, precision=95.64%, recall=94.30%).

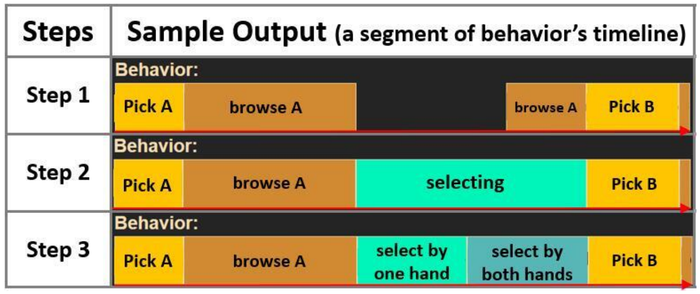

Fig. 6. Behavior Timeline of Each Step

Figure. 6 shows the behavior timeline of each step within the same period. We can see that the proposed method reveals more details to meet the change in marketing demands. However, the specialized models in existing methods cannot be reconfigured to meet different demands. To sum up, the evaluation results show good accuracy and reconfigurable behavior recognition, it indicates that our proposed method can be reconfigured to meet different marketing demands.

5 Conclusion

The existing CAR system cannot adapt to different marketing demands when the recognition target changes or the recognition methods require updates. The specialized models are not reconfigurable, making it hard to adapt to the change of recognition target. The redundant computation in the system causes low efficiency. The low maintainability of the system results in the difficulty in fitting the update of recognition methods. In this research, we design a four-level hierarchy and an event-based method for reconfigurable customer behavior recognition to solve these three problems. The evaluation shows that the proposed method can easily reconfigure behavior outputs compared to existing methods, which adapt to different marketing demands.

References

1. Liu, J., Gu, Y., Kamijo, S.: Customer behavior recognition in retail store from surveillance camera. In: 2015 IEEE International Symposium on Multimedia (ISM). pp. 154–159. IEEE (2015)
2. Popa, M.C., Rothkrantz, L.J., Wiggers, P., Shan, C.: Shopping behavior recognition using a language modeling analogy. Pattern Recognition Letters **34**(15), 1879–1889 (2013)
3. Yamamoto, J., Inoue, K., Yoshioka, M.: Investigation of customer behavior analysis based on top-view depth camera. In: 2017 IEEE Winter Applications of Computer Vision Workshops (WACVW). pp. 67–74. IEEE (2017)
4. Generosi, A., Ceccacci, S., Mengoni, M.: A deep learning-based system to track and analyze customer behavior in retail store. In: 2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin). pp. 1–6. IEEE (2018)

5. Merad, D., Drap, P., Lufimpu-Luviya, Y., Iguernaissi, R., Fertil, B.: Purchase behavior analysis through gaze and gesture observation. Pattern Recognition Letters **81**, 21–29 (2016)
6. Hernandez, D.A.M., Nalbach, O., Werth, D.: How computer vision provides physical retail with a better view on customers. In: 2019 IEEE 21st Conference on Business Informatics (CBI). vol. 1, pp. 462–471. IEEE (2019)
7. Lee, K., Choo, C.Y., See, H.Q., Tan, Z.J., Lee, Y.: Human detection using histogram of oriented gradients and human body ratio estimation. In: 2010 3rd International Conference on Computer Science and Information Technology. vol. 4, pp. 18–22. IEEE (2010)
8. Zhang, S., Wang, X.: Human detection and object tracking based on histograms of oriented gradients. In: 2013 ninth international conference on natural computation (ICNC). pp. 1349–1353. IEEE (2013)
9. Ahmad, M., Ahmed, I., Recognition in Intelligent Ullah, K., Khan, I., Khattak, A., Adnan, A.: Person detection from overhead view: a survey. Int J Adv Comput Sci Appl **10**(4) (2019)
10. Frontoni, E., Raspa, P., Mancini, A., Zingaretti, P., Placidi, V.: Customers' activity recognition in intelligent retail environments. In: International Conference on Image Analysis and Processing. pp. 509–516. Springer (2013)
11. Liciotti, D., Contigiani, M., Frontoni, E., Mancini, A., Zingaretti, P., Placidi, V.: Shopper analytics: A customer activity recognition system using a distributed rgb-d camera network. In: International workshop on video analytics for audience measurement in retail and digital signage. pp. 146–157. Springer (2014)
12. Sturari, M., Liciotti, D., Pierdicca, R., Frontoni, E., Mancini, A., Contigiani, M., Zingaretti, P.: Robust and affordable retail customer profiling by vision and radio beacon sensor fusion. Pattern Recognition Letters **81**, 30–40 (2016)
13. Popa, M.C., Gritti, T., Rothkrantz, L.J., Shan, C., Wiggers, P.: Detecting customers' buying events on a real-life database. In: International Conference on Computer Analysis of Images and Patterns. pp. 17–25. Springer (2011)
14. Zheng, Z., Chen, Y., Chen, S., Sun, L., Chen, D.: Location-aware poi recommendation for indoor space by exploiting wifi logs. Mobile Information Systems **2017** (2017)
15. Chen, Y., Zheng, Z., Chen, S., Sun, L., Chen, D.: Mining customer preference in physical stores from interaction behavior. IEEE Access **5**, 17436–17449 (2017)
16. Lacic, E., Kowald, D., Traub, M., Luzhnica, G., Simon, J., Lex, E.: Tackling cold-start users in recommender systems with indoor positioning systems. In: RecSys Posters (2015)
17. Christodoulou, P., Christodoulou, K., Andreou, A.S.: A real-time targeted recommender system for supermarkets (2017)
18. So, W.T., Yada, K.: A framework of recommendation system based on in-store behavior. In: Proceedings of the 4th Multidisciplinary International Social Networks Conference. pp. 1–4 (2017)
19. de Sousa Silva, D.V., de Santana Silva, R., Ara, F., et al.: Recommending stores for shopping mall customers with recstore. Journal of Information and Data Management **9**(3), 197–197 (2018)
20. Candamo, J., Shreve, M., Goldgof, D.B., Sapper, D.B., Kasturi, R.: Understanding transit scenes: A survey on human behavior-recognition algorithms. IEEE transactions on intelligent transportation systems **11**(1), 206–224 (2009)
21. Leal, E., Gruenwald, L.: Dynmdl: A parallel trajectory segmentation algorithm. In: 2018 IEEE International Congress on Big Data (BigData Congress). pp. 215–218 (2018)