A Flexible Method of Customer Activities Recognition in Retail Store

JIAHAO WEN[1] MUHAMMAD ALFIAN AMRIZAL[2]
TORU ABE[1,3] TAKUO SUGANUMA[1,3]

Abstract: Customer Activities (CA) are customers' interaction with products and services in retail stores. CA Recognition (CAR) provides valuable information for marketing. Existing methods mainly use the end-to-end (e2e) model to realize the CAR. Due to the properties of the e2e model, its deployment is not too efficient because numerous amounts of e2e models are needed to get different information for the marketing purpose. In addition, it is difficult to modify the CAR output. We propose a flexible CAR method where CAR is separated into a hierarchy of several independent recognition levels, each of which uses different e2e model. Each model can be independently updated because recognition levels are separated. Output data of each level consists of the output data from the lower level, which provides a simple way to modify the output of each level. Furthermore, outputs from different levels offer different kinds of information.

Keywords: Retail, Customer Activities, Flexible Behavior Recognition

1. Introduction

Pattern recognition is one of an essential field of Artificial Intelligence (AI). Designing an algorithm that is able to recognize and understand human activities has been a goal of Pattern recognition. Machine learning-based algorithm is becoming the most popular algorithm in the recent few years because of its fast running speed and high accuracy. The rapid development of machine learning model now enables our computers to accept raw sensor data and output the recognition result of human activities. Human activity recognition has led to the innovation in many fields, such as the retail store.

Since the rising of online shopping has occupied more and more parts of the market, the normal retail store is becoming gradually out of our sight. Thus, the innovation is necessary for the normal retail store in order to survive the competition with the online shop. As shown in Figure 1, in the traditional retail store, retailers use the records of cash registers or credit cards to analyze the purchasing behaviors of customers to support marketing plan [1]. However, those records only show the final result of the customer purchase decisions. There is no information about the process of customers' decision making which is a one of the valuable information for marketing plan because it may reveal the reason of those decisions. Therefore, a kind of retail solution called "Smart Retail" is proposed. It uses ubiquitous sensors, especially camera, to collect real-time data of the store. Then, with a popular machine learning model, the smart retail is able to analyze those real-time data to recognize Customer Activities (CA). The result of the CA Recognition (CAR) can be analyzed to get lots of valuable information for marketing plan.

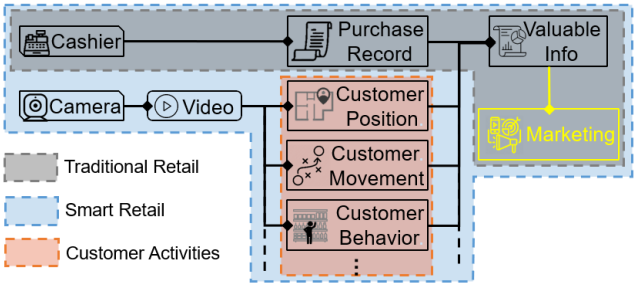


Figure 1 Smart Retail and Customer Activities.

In this research, customer behavior is concluded into a concept called Customer Activities which means the customers' interaction with products and services in retail stores. Figure 1 illustrates that CA includes customer's position, movement, behavior, etc. In other words, CA is defined as a general concept of the interaction between customer and things in a retail store.

To realize CAR, the e2e machine learning model is commonly utilized. Despite the good performance on recognition speed and accuracy, there are three problems when utilizing e2e models for CAR.

First and foremost, it is hard to modify the model's outputs. Marketing demands are constantly changing and the changes probably require a recognition of some new behaviors that are not included in the current e2e model. However, modifying the e2e model to recognize some new behaviors is time-consuming. New training data are needed to be collected and the e2e model must be re-trained with the new data. It usually takes a few hours or even several days to complete the whole modification of the e2e model.

Besides, the e2e model cannot be partially updated. A customer behavior recognition model usually has the input of video and the output of behavior. Since it is an e2e model, the processes such as people detection, motion feature extraction, behavior recognition must have been integrated into the model. Since the e2e model is an unexplainable black box, it is impossible to figure out which part belongs to which process. That is to say, replacing only the part for people detection with a better method is impossible. The model can only be entirely updated instead of partially updated.

Last but not least, outputs of an e2e model are in the same level. Marketing requires data from different levels where the customer behavior belongs to one of those levels. Due to the properties of the e2e model, an e2e model outputs data in the same level, as shown in Figure 2. Hence, many e2e models are necessary to get data from different level to support marketing.

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

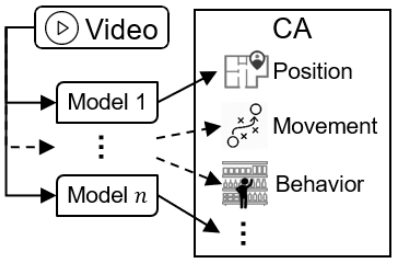


Figure 2 Existing CAR in Smart Retail.

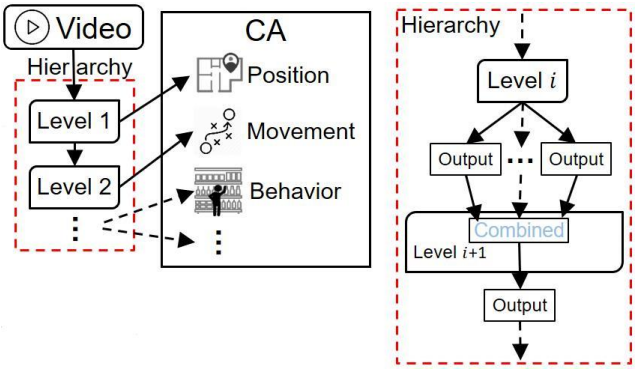


Figure 3 Solution to Existing Problems.

In addition, some of those models have repeated process, such as the behavior recognition model and the tracking model that have the same process of people detection. Since those processes are integrated into the model, that same process has to be done on both of the two models. Furthermore, to realize the smart retail, those e2e models must be run at the same time, and hence is computationally expensive.

To solve these problems, we propose a hierarchy shown in Figure 3 to organize CAR and a flexible method to use the hierarchy which makes it easier to modify the recognition outputs.

The hierarchy divides CAR into several levels. The output data of each level is combined to get the output of the higher level. It means that changing the combination of the data from this level can modify the recognition output of the higher level which makes the modification process easier than existing methods.

Each level has its own duty. Such as one of them is object detection, another one is object tracking, etc. Any method or e2e model can be used if it is able to perform the duty of this level. In other words, each level is independent to the other levels because they are doing different work. When the level of object detection is updated, it has no influence on the level of the level for object tracking. This means that the whole CAR hierarchy can be partially updated.

Each level has different duty. These different levels outputs different kinds of data, such as position, movement, behavior, etc. And marketing requires all of the different data like that instead of one of them. Thus, these different data are able to satisfy marketing demands of different kinds of data. Although there are still several models running at the same time, but they are performing different functions and thus, there are no repeated process among them.

2. Related Work

Existing methods about CAR give various results. As we focus on how to divide CAR into a multi-level hierarchy, we need to summarize existing CAR methods by their focused topic to find out what kind of CAR exists in existing methods. After the survey, we found that existing methods are mainly focused tracking object and customer behavior recognition.

2.1 Tracking objects

To acquire more details about customers, the location information should be the most basic data because more data can be analyzed only if you know where is the customer at first. Thus, researches on tracking offer the result of the position or trajectory of body, body parts or other objects. And some of them also have some other results that can be inferred from the tracking results, such as total shopping time, customer interest of each area in the store, etc.

Tracking objects can be realized from different kinds of data. As the images contain much more information than other kinds of sensor data, the visual tracking attracts much effort of researchers on it [2][3][5][8] and all methods on behavior recognition track people by images. Other sensors are also be utilized to get the position of customers [9], like Wi-Fi [10][11][12], Bluetooth [13][14][15], RFID [16] and GPS[17].

We summarized tracked objects in these methods as shown in Table 1. There are some researches on behavior recognition also have outputs about tracked objects. Thus, these researches are also taken into consideration in Table 1.

Body refers to the tracking output of the whole body. And the body part means hands and arms in existing methods. Other Objects means the objects except human that customers are possible to interact with. In [1], the pixel area of the shopping basket is detected. [22] detects the position of each products.

Those mentioned methods offer the tracking outputs of customers which are usually position or trajectory. They provide basic data that is able to reveal the simple shopping process to give some valuable information for marketing. But it is not enough for the retailer, those basic data can be analyzed to get more information. And unfortunately, it seems that due to the user-friendly e2e model, only a few researches on behavior recognition utilize the tracking results in these methods. Most researches just input the raw images and leave the tracking task to training data and e2e models, then get the recognition results.

2.2 Customer Behavior Recognition

Various researches realize the recognition of different behaviors. Table 2 lists all results about customer behaviors in existing researches. It shows most researches focus on several behaviors and none of them have the results of all those behaviors. Even if there exists a research that covers all those behaviors, no research or evidence shows that they are enough for retailers. And

Table 1 Tracked objects in Existing Methods.

Objects	Related methods
Body	[1]-[8] [10]-[20]
Body Part	[1] [22]
Other Object	[1] [22]

Table 2 Behaviors in existing methods.

Behaviors	Related methods
Pick a product from shelf	[1] [4] [6] [7] [18]
Pick nothing from shelf (Touch)	[1] [4] [7]
Return a product to the shelf	[1] [6] [7]
Put a product into cart/basket	[1]
Passing by/No interest	[1] [6]
Holding a product	[19]
Browsing a product on the hand	[4] [19]
Viewing the shelf	[1] [19]
Turning to the shelf	[1]
Fit next to you & Check how it looks & try on & take off (in clothes shop)	[4]
Emotion (Happiness, Surprise, etc.)	[20]

in some researches, the result which contains no information about behavior is defined as behavior. In other words, behaviors are not well organized.

In [19], “top, second, third, fourth” which represents a customer reaches for a book on the top-fourth floor of the shelf. They are defined as independent behaviors which means they cannot appear with other behaviors at the same time. But it is obviously that they are just the position information, not behavior.

[1] has a behavior called “Picking and putting” which is the combination of “Pick a product” and “Put into cart”. Actually, it would be better if “Picking and putting” is separated into two independent behaviors.

All of the existing researches on customer behavior recognition share the problem of such no well-organized output behaviors because none of them mentioned the reason why they choose the recognition of those behaviors. It causes a serious problem that if the results are not designed at the view of users (retailers), it is hard to say that those results are valuable information to supporting marketing plan.

Another problem is that these researches mostly utilize e2e models based on different machine learning methods to recognize customer behaviors, such as SVM [1] [19], HMM [4] [18], etc. [20]. And for the general behavior recognition which is not in retail environment, RNN-based model is also utilized [21]. Those e2e models cause the three problems mentioned in the "Introduction" part. They own the fixed number of outputs which is not easy to be modified when the changeful marketing demand requires modification. And it is impossible to partially update an e2e model because it is a black box. Also, outputs of an e2e model are in the same level while marketing requires data from different levels.

Except those researches have definition for each behavior, some researches focus on describing the degree of a kind of behavior. For instance, Merad et al. equip each customer with a glasses and track hands movement by the images from glasses, then give the prediction of the indecisiveness degree of customers [22]. Similar to above researches, they share the same problems.

To sum up, all existing methods of customer behavior recognition adapt to their assumed conditions. They are not flexible enough due to the property of their e2e models. And no prove shows that their recognition results are valuable information for marketing because the not well-organized output behaviors.

3. Proposal

As our purpose is to divide CAR into a multi-level hierarchy and make it flexible, our idea is firstly dividing CAR in existing methods into several parts and then integrate them into a multi-level hierarchy. With the integrated hierarchy, a clearer view of existing CAR is provided. We are able to modify the hierarchy to solve existing problems and find a method to make behavior outputs flexible.

3.1 Integrated Hierarchy of Existing Methods

We list the results of all existing methods and find that they have similar flow on the recognition about retail. They detect the targets in each frame with the results of their position, then track and analyze the change of each target, finally recognize behaviors from the previous results of changes of targets. Therefore, we integrate results of existing methods into a hierarchy with three levels: Position, Movement and Behavior.

A. Level 1: Position

Table 1 shows the output of the level “Position”. These outputs are summarized from the output of existing methods. Since the object detection in a single frame is an important part for CAR, and nearly all researches have outputs about objects’ position in each frame, we use the name “Position” at the most basic level. It means the position of objects in each frame. Therefore, the duty of this level is to get the input of images, detect objects in each frame and output their positions.

Most researches analyze the change of the human pixel area to recognize behavior instead of detect the precise position of body parts like hands, arms. Thus, all those researches actually utilize the position data of body pats but they did not have the detection for any specific body parts. Thus, the listed related researches [1] [22] are those researches which have the detection outputs for specific body parts.

B. Level 2: Movement

With the output of “Position”, existing methods analyze “Position” from consecutive frames to track and get motion features of each object. Therefore, the level “Movement” is one level higher than “Position”. Thus, this level has the duty of extracting data from several consecutive frames of each target such as the moving direction, trajectory, change of pixels, etc. And outputs of "Movement" are divided into three parts as shown in Table 3.

Motion feature contains the direction of targets or pixels, such as moving direction of the body which shows the direction of an object, histograms of optical flow (HOF) which calculates the direction of each pixel, etc.

Trajectory means the connections of each position of a tracked object. Different from motion feature which is usually

Table 3 Outputs of the level “Movement”.

Output	Related Existing Method
Motion Feature	[1] [4] [6] [7] [19] [22]
Trajectory	[18]
Human-Object Relation	[1] [22]

limited by the length of the video, a trajectory is a sequence of several positions which has no limitation of video length.

Human-Object Relation means the relation between human and objects. This relation includes the relative position, similarity of moving direction or trajectory. For instance, [1] calculates whether hands and arms inside or outside a basket, [22] analyzes whether the product is moving with, near or away from hands. A problem in this part is that it is inferred from position and direction information, but we cannot reach the next level "Behavior" only with the information of relation. The other parts of "Movement" are still necessary to reach the level "Behavior". Hence, the Human-Object Relation is placed in this level instead of a level in the middle of "Movement" and "Behavior".

C. Level 3: Behavior

This level includes the customer behavior and emotion outputs in existing methods as shown in Table 4. And all contents in this level are inferred from its lower level "Movement".

No interaction behavior refers to the behavior without any interaction with other objects, such as passing by, viewing, etc. On the contrary, interaction behavior is the behavior that has the interaction with other objects, like picking/return a product, put into cart, etc. Except the research on emotion recognition [20], the other researches on customer behavior recognition provide the outputs of those two kinds of behavior. And emotion outputs are provided in [20] such as happiness, surprise, neutral, etc. Details about emotion recognition are not mentioned in [20], Thus, we are not able to know any more about it. As its output belongs to behavior recognition, we put it into the level "Behavior".

Outputs in existing methods are integrated into a hierarchy as shown in Figure 4. It shows how data is processed step by step to become recognized behaviors. As marketing demands seldom require the modification of the basic levels "Position" and "Movement", the flexibility is unnecessary in these levels. But when it comes to "Behavior" where modification is usually required, according to the problem that it is hard to modify the behavior output of an e2e model. Hence, our solution is based on modifying this part of the hierarchy.

3.2 Proposed Hierarchy

Similar to the discovery of the neural network, in order to design a flexible method of behavior recognition, we can go back to the essence of the word "behavior" to find out how we define behavior. Most definitions of behavior explain what is behavior but it does not reveal the process of how we recognize behavior. However, there is a definition defines human behavior as a composition of multiple events and an event refers to a single low-level spatiotemporal entity that cannot be further decomposed [23]. This definition reveals the process that we recognize events primarily and then combine them together to be the behavior. Therefore, we modify the integrated hierarchy and propose a new hierarchy as shown in Figure 5 which enable the recognition of customer activities to be flexible.

Table 4 Outputs of the level "Behavior".

Output	Related Existing Method
No interaction Behavior	[1] [6] [7] [18] [19]
Interaction Behavior	[1] [4] [6] [7] [18] [19] [22]
Emotion	[20]

Comparing to the original integrated hierarchy, two new level "Event" and "Intention" are added and the emotion is removed because there is no detail about it and it is also not the focus of our research.

Inspired by the definition in [23], we insert a level "Event" which has the duty to get the basic entity of "Behavior". The combination of events can be defined as a behavior.

A new level "Intention" is added higher than "Behavior". And with the data from "Behavior" and additional data "Personal Information", the intention is able to be inferred. That is a part of our future work because we think that "Behavior" should not be the end of this hierarchy of CAR, there must be a higher level reveals more valuable information such as the intention of customers.

This design of hierarchy which divides CAR into five individual levels has two advantages.

Each level has its own duty and it is independent of other levels. Thus, each level can use the best method to performs its duty and there is nearly no influence on the other levels when the method of this level is changed. For example, the level "Position" needs to perform the duty of object detection in each frame and "Movement" needs to perform the duty of object tracking during consecutive frames. We are able to use the best method of object detection and object tracking. When the detecting method in "Position" changes, because it still performs its duty, the method in "Movement" still works well.

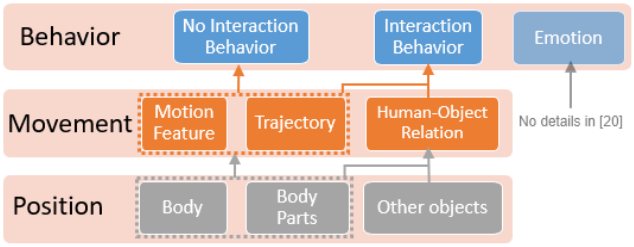


Figure 4 Integrated Hierarchy of Existing Method.

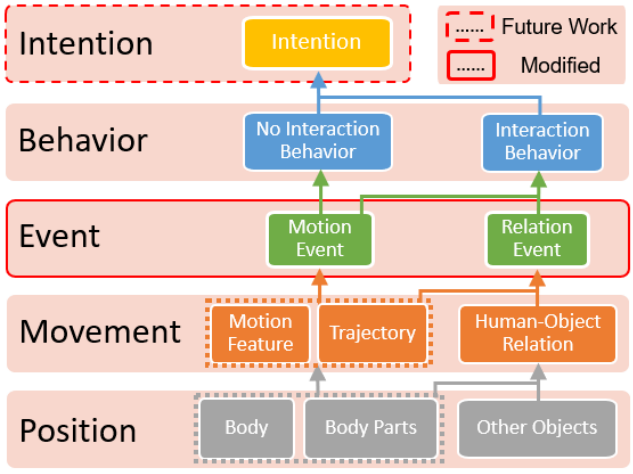


Figure 5 Proposed Hierarchy.

Then, each level in the hierarchy output its outputs. Outputs from "Position" and "Behavior" are different data. Therefore, with a model instructed based on this hierarchy, five different levels are able to provide varieties of different data for marketing which is better than a single e2e model in existing methods.

3.3 Proposed Method

Since a new level Event is added, we also proposed a method to realize the duty of this level and make the output of Behavior level flexible.

A. Level 3: Event

Before it goes to the behavior, we need to figure out the process from Movement to Event. Since the behavior consists of events which are its basic entities, we would like to symbolize the data from Movement to get rid of some limitations when using specific value such as the value is harder to be understood than a symbol.

The output of Event is divided into motion event and relation event as shown in Table 5.

The motion event takes the input of object's trajectory from Movement and summarize it into four symbols. These symbols refer to object's label, object's motion which can be move, stop or rotate, object's direction which can be moving up, down, left, right or "-" when it is not moving and finally object's start position and its end position. In the implementation, each frame is separated into several areas, and the start-end position means the object moved from which area to which area.

The relation event comparing two motion events to calculate the relation between these two objects. It also has four symbols refer to the label of object 1, the relation of their direction which can be is following near object 2, moving close to object 2, moving away from object 2 or "-" means no clear relation, the relative position which can be object 1 on the left side of, on the right side of, above, or below object 2 and finally the label of object 2.

Table 5 Symbolize Data from Movement.

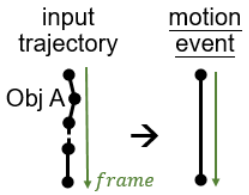
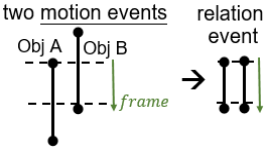
Event	Proposed Method	Output
Motion Event		[**Obj A**], [**Move**/Stop/Rotate], [Up/**Down**/Left/Right/-], [**Start-End Position**]
Relation Event		[**Obj A**], [**Is following**/Close to/ Away from/-], [**Left side**/Right side/ above/below], [**Obj B**]

Table 6 Example of behavior "Pick a Product".

Behavior	Event (sequence: 1→2) [*] refers to any value
Pick a Product	1. [Hand], [move], [down], [shelf area] (motion)
	2. [Hand], [move], [up], [shelf area] (motion)
	2. [Product A], [is following], [*], [hand] (relation)

B. Level 4: Behavior

Since events are the basic entities of behaviors, this method defines behaviors as a sequence of events. Table 6 is an example of the behavior "Pick a Product" which is a sequence of two motion events and a relation event. When event 1 is followed by event 2 (motion and relation), it is regarded as the behavior "Pick a Product".

It is easy to understand this sequence by these symbols. When a hand moves down (to the shelf) in the shelf area with no product, then the hand moves up (away from the shelf) still in the shelf area with the Product A following, that is a behavior called "Pick a Product". And the symbol "*" is written at the place of relative position in relation event. It means that we do not care about this value because we do not need to specify this value during a customer is picking a product.

With the level Event and Behavior are realized by the proposed method, it brings several advantages:

A. "Behavior" is easy to be modified

Since the behavior is defined as an event sequence where events are combined sequentially like "Pick a Product" in Table 6, when the modification of "Behavior" is required, assuming that events are enough, the only thing we need to do is to change the combination of events without any new training step.

B. Reduce the training data size in "Event"

To recognize the behavior "Pick a Product", we need five kinds of data: object's class, moving direction, located area, relation kind and related object's class. Though five kinds of data are required for one behavior, nearly all behaviors in Table 2 can be defined by these five kinds of data. That is to say, we are able to recognize at least thirteen kinds of behaviors except emotion in Table 2 if we can get these five kinds of data. Since marketing demands seldom require the change of the data in the level "Event", an e2e model can be trained to get these data. During collecting the training data, it only needs to label for these data without labeling for every required behavior which reduces the size of training data.

C. Allows the customization of behaviors by users

Because the "Event" level is represented by simple symbols, users can easily customize their wanted behaviors by themselves just as the definition in Table 6.

In our proposal, we propose a hierarchy divides CAR into five levels and a flexible method of the level "Event" which provides flexibility to the level "Behavior".

Thus, from what has been discussed above, we may safely draw the conclusion that our proposed hierarchy and method can modify the behavior recognition easily, output different kind of data to support marketing and allows partial update.

4. Implementation

To verify the feasibility of the proposed method, we set up a retail environment as shown in Figure 6 assuming that every person is alone, namely no interaction with other people.

To avoid occlusion, a top-view camera is installed on the top of a shelf. And each frame is manually divided into two areas for the symbolizing process in "Event". Viewing Area (VA) refers to the area that a customer is close enough to interact with products

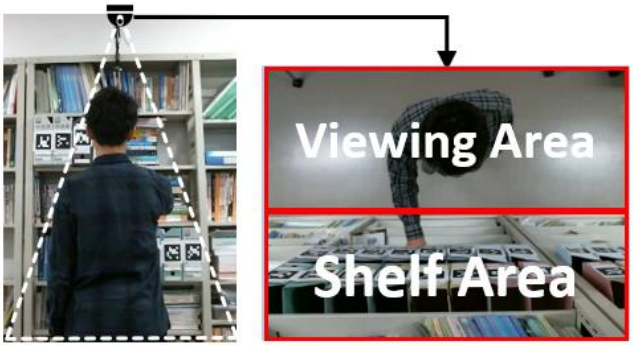


Figure 6 Camera Installation and Input Image.

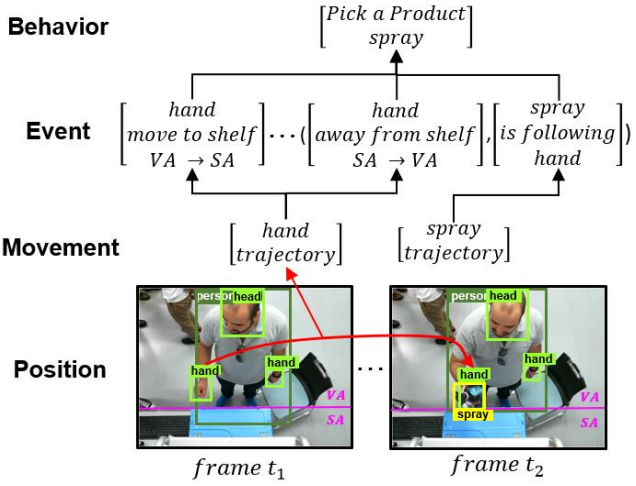


Figure 7 Recognition Flow of the Implemented Model

on the shelf. Shelf Area (SA) refers to the area includes the shelf and all products on it.

To test our proposed hierarchy, we should implement specific methods into each level. In our implemented model, the recognition flow is shown in Figure 7. The level "Intention" is not included in our experiment because it is the future work.

For "Position", we utilize a pre-trained Mask-RCNN[24] model to detect person, hand, bottle, spray, wet tissue, pear water.

For "Movement", if the object's position in the current frame is close to that in the previous frame, it will be tracked as the same object. And the same object's positions are recorded to be a trajectory. When a new input of object's position comes, calculate its direction vector and add it to the saved trajectory as shown in Figure 8. If the vector has similar direction with the saved trajectory, add it to this trajectory and output nothing. Otherwise, output the saved trajectory and overwrite it with the vector.

For "Event", the trajectories from "Movement" are symbolized by the proposed method in Table 5. The region of Viewing Area (VA) and Shelf Area (SA) is preset in Figure 7. The direction of moving "down" refers to moving "to shelf" and moving "up" refers to moving "away from shelf".

For "Behavior", it recognizes behaviors by finding the predefined event sequence like the example of Table 6. About seven kinds of common behaviors in existing methods are defined by events in this implemented model as shown in Table 7.

Except the outputs of each level, we also need to provide valuable information for marketing. Thus, an additional method

Table 7 Recognized Behaviors in Implemented Model.

Behavior	Definition
Walking	Walking in VA
Viewing	Stopped in front of the shelf and View the shelf
Selecting	Raise up the hand without any product in SA
Holding	Holding a product on the hand
Pick Something	Pick a product out of the shelf
Pick Nothing	Hand into the shelf and pick nothing out
Return	Return a product back to the shelf

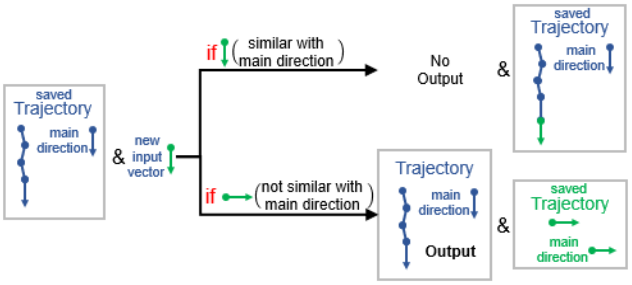


Figure 8 Implemented Method of Movement.

which is about how to use the outputs from levels is implemented to calculate Preference Score. It reveals customer's attention on each product. Each product has a Preference Score and each behavior has a manually preset point. Once the customer has the behavior interacted with a product, the point of that behavior will be accumulated to the Preference Score of that product. For instance, once the behavior "Pick a spray" happened, a point of +40 is added to the score of spray. And if the behavior "Return a spray" happened, a negative point of -50 is added to the score of spray because returning is a kind of negative behavior. The point number here is determined randomly.

5. Evaluation

As our purpose is to prove that the proposed model is more flexible than existing methods. There is a plan has two steps.

First, verify the implemented model is feasible by experiment. Second, if the feasibility of the implemented model is verified and all behaviors in existing methods can be defined as event sequences, it can be theoretically proved that the proposed model can recognize all behaviors in existing methods which means that the proposed model is more flexible.

5.1 Feasibility

The experiment is carried out on a public activity "Open Campus". We have got total 49 random visitors as the participants (customers) for the experiment of the implemented model. All of them are requested to imagine shopping in front of a shelf with products and pick at least one of them out of the shelf by hands.

Figure 9 is the experimental result of our experiment. During the experiment, the implemented model works well when Mask RCNN maintains a good accuracy. And it performs poor when there are lots of mistakes in the detection results of Mask RCNN.

There is only results of videos and more experimental results such as accuracy are on the schedule of our future work. Those videos show that our implemented model is feasible to realize CAR if the method of "Position" maintains a good accuracy.

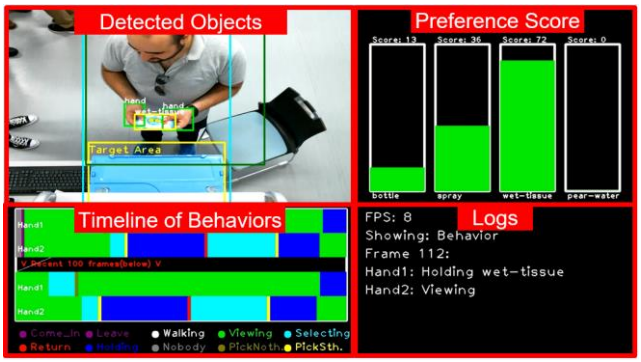


Figure 9 Experiment to verify feasibility.

5.2 Flexibility

In the evaluation of flexibility, we prove that our model is more flexible by proving that our proposed model can recognize all behaviors in existing methods.

As shown in Table 8, it defines all behaviors in existing methods by event sequence except "Fit next to you" and "Check how it looks" which have no definition in the paper. And all the symbols in events are mentioned in Table 5 in the proposal.

Since the feasibility is basically verified and all behaviors in existing methods are defined in Table 8, we are able to conclude that the proposed model should be able to recognize all those behaviors in existing methods. Therefore, the proposed hierarchy and method is more flexible than existing methods.

6. Conclusion

To adapt to the changeful market demands in smart retail, we proposed a hierarchy which divides CAR into five levels and a flexible method to utilize the hierarchy to recognize CA.

The proposed method is capable of easily modifying the results of behavior recognition. Different levels in the hierarchy output different data to support marketing. And the independent levels in the hierarchy allow the partial update of the CAR hierarchy and the usage of the best methods in each level of CAR.

Besides, except the advantage of flexibility, the symbolization step in "Event" level makes defining behavior easy which also means it is easy to modify the behavior. The symbol is friendly for users to customize their wanted behaviors according to their marketing plan.

Not only for smart retail, this hierarchy should also able to be generalized to a hierarchy for general behaviors after some improvement. But before that generalization, we would like to make it perfect in the field of customer activities recognition firstly. Our future work is planned to be mainly focused on the level "Intention" and the recognition of customer groups. The level "Intention" which probably reveals customers interest on products will provide lots of valuable information. And the improvement of realizing customer group recognition allows the proposed method to be more feasible in the real retail environment.

Table 8 All Behaviors' Event Combination

Behavior	Event (sequence: 1→2) [*] refers to any value
Pick a Product	1. [Hand], [move], [to shelf], [SA] or [VA→SA] (motion)
	2. [Hand], [move], [away from shelf], [SA] or [SA→VA] (motion)
	2. [Product A], [is following], [*], [Hand] (relation)
Pick Nothing	1. [Hand], [move], [to shelf], [SA] or [VA→SA] (motion)
	2. [Hand], [move], [away from shelf], [SA] or [SA→VA] (motion)
Return a Product	1. [Hand], [move], [to shelf], [SA] or [VA→SA] (motion)
	1. [Product A], [is following], [*], [Hand] (relation)
	2. [Hand], [move], [away from shelf], [SA] or [SA→VA] (motion)
Put into cart/basket	1. [Hand], not [Stop], [*], [VA] (motion)
	1. [Hand], [close to], [*], [cart] or [basket] (relation)
	1. [Product A], [is following], [*], [Hand] (relation)
	2. [Hand], [*], [*], [VA] (motion)
	2. [Hand], [away from], [cart] or [basket] (relation)
Passing by	1. [Person], not [Stop], [*], [VA] (motion)
Holding a Product	1. [Product A], not [Rotating], [*], [VA] (motion)
	1. [Product A], [is following], [*], [Hand] (relation)
Browsing a Product	1. [Product A], [Rotating], [-], [VA] (motion)
	1. [Product A], [is following], [*], [Hand] (relation)
Viewing the shelf	1. [Person], [Stop], [-], [VA] (motion)
Turning to the shelf	1. [Head], [*], [Left side] or [Right side], [Hand] (relation)
	2. [Head], [*], [Above] or [Below], [Hand] (relation)
Try on	1. [Product A], [is following], [*], [Hand] (relation)
	2. [Product A], [is following], [*], [Head] (relation)
Take off	1. [Product A], [is following], [*], [Head] (relation)
	2. [Product A], [is following], [*], [Hand] (relation)

Reference

[1] J. Liu, Y. Gu and S. Kamijo, "Customer Behavior Recognition in Retail Store from Surveillance Camera", IEEE International Symposium on Multimedia, pp. 154-159 (2015).

[2] K. Lee, C. Y. Choo, H. Q. See, Z. J. Tan, Y. Lee, "Human detection using Histogram of oriented gradients and Human body ratio estimation", International Conference on Computer Science and

Information Technology, pp. 18-22 (2010).
[3] S. Zhang, X. Wang, “Human detection and object tracking based on Histograms of Oriented Gradients”, International Conference on Natural Computation (ICNC), pp. 1349-1353 (2013).
[4] M. C. Popa, L. J. M. Rothkrantz, P. Wiggers, C. Shan, “Shopping behavior recognition using a language modeling analogy”, Pattern Recognition Letters, vol. 34, Iss. 15, pp. 1879-1889 (2013).
[5] M. Ahmad, I. Ahmed, K. Ullah, I. Khan, A, Khattak, A. Adnan, “Person detection from overhead view: A survey”, International Journal of Advanced Computer Science and Applications, vol. 10, Iss. 4, pp. 567–577 (2019).
[6] E. Frontoni, P. Raspa, A. Mancini, P. Zingaretti, “Customers’ Activity Recognition in Intelligent Retail Environments”, New Trends in Image Analysis and Processing (ICIAP), vol. 8158, pp. 509-516 (2013).
[7] D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini, P. Zingaretti, V. Placidi, “Shopper Analytics: A Customer Activity Recognition System Using a Distributed RGB-D Camera Network”, Video Analytics for Audience Measurement, vol. 8811, pp. 146–157 (2014).
[8] M. Sturari, D. Liciotti, R. Pierdicca, E. Frontoni, A. Mancini, M. Contigiani, P. Zingaretti, “Robust and affordable retail customer profiling by vision and radio beacon sensor fusion”, Pattern Recognition Letters, vol. 81, pp. 30-40 (2016).
[9] D. A. Mora Hernandez, O. Nalbach, D. Werth, “How Computer Vision Provides Physical Retail with a Better View on Customers”, IEEE Conference on Business Informatics (CBI), vol. 1, pp. 462-471 (2019).
[10] B. Fang, S. Liao, K. Xu, H.Cheng, C. Zhu, H. Chen, “A novel mobile recommender system for indoor shopping”, Expert Systems with Applications, vol. 39, no. 15, pp. 11992-12000 (2012).
[11] Z. Zheng, Y. Chen, S. Chen, L. Sun, D. Chen, “Location-aware POI recommendation for indoor space by exploiting WiFi logs”, Mobile Information Systems, vol. 2017 (2017).
[12] Y. Chen, Z. Zheng, S. Chen, L. Sun and D. Chen, "Mining Customer Preference in Physical Stores From Interaction Behavior," in IEEE Access, vol. 5, pp. 17436-17449 (2017).
[13] S. S. Chawathe, “Beacon Placement for Indoor Localization using Bluetooth”, IEEE Conference on Intelligent Transportation Systems, pp. 980-985 (2008).
[14] E. Lacic, D. Kowald, M. Traub, G. Luzhnica, J. Simon, E. Lex, “Tackling cold-start users in recommender systems with indoor positioning systems”, RecSys Psters (2015).
[15] P. Christodoulou, K. Christodoulou, A. S. Andreou, “A Real-time Targeted Recommender System for Supermarkets”, International Conference on Enterprise Information Systems (ICEIS), vol. 2, pp. 703-712 (2017).
[16] W. T. So, K. Yada, “A Framework of Recommendation System Based on In-store Behavior”, Multidisciplinary International Social Networks Conference (MISNC), pp. 33 (2017).
[17] D. V. d. S. Silva, R. d. S. Silva, F. A. Durão, “Recstore: Recommending stores for shopping mall customers”, Brazillian Symposium on Multimedia and the Web, pp. 117-124 (2017).
[18] M. C. Popa, T. Gritti, L. J. M. Rothkrantz, C. Shan, P. Wiggers, “Detecting Customers’ Buying Events on a Real-Life Database”, Computer Analysis of Images and Patterns, pp. 17–25 (2011).
[19] J. Yamamoto, K. Inoue, M. Yoshioka, “Investigation of Customer Behavior Analysis Based on Top-View Depth Camera”, IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 67-74 (2017).
[20] A. Generosi, S. Ceccacci and M. Mengoni, “A deep learning-based system to track and analyze customer behavior in retail store”, IEEE International Conference on Consumer Electronics (ICCE), pp. 1-6 (2018).
[21] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, T. Darrell, “Long-Term Recurrent Convolutional Networks for Visual Recognition and Description”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677-691 (2017).
[22] D. Merad, P. Drap, Y. Lufimpu-Luviya, R. Iguernaissi, B. Fertil, “Purchase behavior analysis through gaze and gesture observation”, Pattern Recognition Letters, vol. 81, pp. 21-29 (2016).
[23] J. Candamo, M. Shreve, D. B. Goldgof, D. B. Sapper and R. Kasturi, “Understanding Transit Scenes: A Survey on Human Behavior -Recognition Algorithms”, IEEE Transactions on Intelligent Transportation Systems, vol. 11, no. 1, pp. 206-224 (2010).
[24] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN", IEEE International Conference on Computer Vision (ICCV), pp. 2980-2988 (2017).