Semantic Assessment of Shopping Behavior Using Trajectories, Shopping Related Actions, and Context Information

M.C. Popa\textsuperscript{a,b,*}, L.J.M. Rothkrantz\textsuperscript{a,c}, C. Shan\textsuperscript{b}, T. Gritti\textsuperscript{b}, P. Wiggers\textsuperscript{a}

\textsuperscript{a}Man-Machine Interaction group, Department of Mediamatics, Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands

\textsuperscript{b}Video and Image Processing Department, Philips Research, HTC 36, 5656 AE, Eindhoven, The Netherlands

\textsuperscript{c}Sensor Technology, SEWACO Department, Netherlands Defence Academy, Nieuwe Diep 8, 1781 AC, Den Helder, The Netherlands

Abstract

The possibility of automatic understanding of customers’ shopping behavior and acting according to their needs is relevant in the marketing domain, attracting a lot of attention lately. In this work, we focus on the task of automatic assessment of customers’ shopping behavior, by proposing a multi-level framework. The framework is supported at low-level by different types of cameras, which are synchronized, facilitating efficient processing of information. A fish-eye camera is used for tracking, while a high-definition one serves for the action recognition task. The experiments are performed on both laboratory and real-life recordings in a supermarket. From the video recordings, we extract features related to the spatio-temporal behavior of trajectories, the dynamics and the time spent in each region of interest (ROI) in the shop and regarding the customer-products interaction patterns. Next

\footnotesize{*Corresponding Author. Phone: +31(0)15 2781537. Fax: +31(0)15 2787141. Email address: m.c.popa@tudelft.nl (M.C. Popa)
we analyze the shopping sequences using a Hidden Markov Model (HMM). We conclude that it is possible to accurately classify trajectories (93%), discriminate between different shopping related actions (91.6%), and recognize shopping behavioral types by means of the proposed reasoning model in 95% of the cases.

Keywords: Shopping Behavior, Semantic Analysis, Trajectory Analysis, Action Recognition, Hidden Markov Models.

1. Introduction

In recent years there is an increasing interest towards developing intelligent software solutions to enhance the user experience. Fields such as affective computing, gaming industry, surveillance, or marketing could benefit a lot by enabing systems which could act according to the user preferences or intentions. In the marketing domain it is of great interest to build a satisfactory relation with the customer, by assessing his emotional state and intentions. The shopping experience could be enhanced by facilitating easy access to the products for which the customer shows interest or by offering timely assistance whenever a customer needs help in finding or selecting a merchandise. For assisting customers, usually human shop assistants are available, but given peak hours they are too expensive to meet the whole demand or they are not always well trained or willing to adapt to the different types of customers. Following the conclusions presented in [1], the employee affective delivery plays an important role in the customer’s level of satisfaction. Therefore, a supporting alternative can be provided by developing an automatic behavior assessment system. Using the available surveillance sys-
tems of video cameras in shops [2], we aim at a semantic interpretation of the customers’ shopping behavior and at detecting when there is a need for support or a selling opportunity.

The modeling of the shopping behavior is based on different types of information. The customers’ way of walking provides a first indication of the type of behavior, while the different customer-product interaction patterns can reveal the level of interest of the customer for a specific product. Our assessment model is context sensitive and is based on the segmentation of the shopping area into Regions of Interest (ROIs) such as products, passing, pay desk, or resting areas. Features such as the time spent in each ROI together with the transitions between different ROIs can help at a better modeling of the shopping behavior, as an action can have different meanings in different ROIs. For example, standing in the products area can mean visual inspection of the available products, the same action at the pay desk ROI, denotes waiting, while in the passing ROI can be regarded as orientation or waiting for another person. It is very important to detect if a customer spends too much time in a specific ROI, in order to take appropriate actions on short or long term, such as sending a shop assistant to offer help or optimizing the products arrangement. Customers display different shopping behaviors depending on several factors, such as whether they are experienced with the shopping area or not, their purpose, their mood, if they are accompanied or alone. Different shopping behaviors can be noticed not only between different shopping trips, but also during the same one, fact which adds even more complexity to the behaviour modeling task.

Our contributions in this paper consist of designing a framework for auto-
matic assessment of shopping behavior built in a hierarchical manner, by employing different levels of abstraction, from low sensory level up to the semantic level. At the sensor level different video cameras are synchronized and used in a collaborative manner. A fish eye camera is used to detect people and to track them along the shop, while a high-definition camera is employed for action recognition. Analyzing a customer’s actions is not relevant all the time, but only when he is in a specific ROI. Given the position of a customer, entering or exiting from a ROI can be detected, and the action recognition module can be started. The high-level semantic interpretation of the different shopping behavioral types is realized using a reasoning model, which combines the intermediary outputs of the trajectory analysis, action recognition, and ROI detection modules. The architecture of the proposed system is presented in Fig. 1. The proposed framework is tested both in a laboratory set-up and also in a real-life scenario.

![Figure 1: System Overview.](image)

The outline of the paper is as follows. In Section 2 we give an overview of
related work. Next we describe the proposed shopping behavior models. We continue in Section 4, by presenting the computational framework, the integrated modules, namely trajectory analysis and action recognition in terms of feature extraction and classification approach. The proposed reasoning model is introduced in Section 4.6. Next, we provide a descriptions of the used datasets and the experimental results in Section 5. Finally, we formulate our conclusions and give directions for future work.

2. Related Work

Organizations, companies, and retailers design products, experiences, and ambiance in order to influence people’s behavior, as a means to increase their profitability and popularity. In order to achieve this goal, human behavior needs to be understood and modeled accordingly. There were many attempts to explain and categorize human behavior. Social psychology explains human behavior as a result of the interaction of mental states and immediate social situations. Examples of proposed models are: the Social Cognitive Theory [3], the Cost-benefit model [4] or the Theory of Reasoned Action [5]. The model developed by Fogg named FBM [6] seems a valid approach towards describing human behavior. This model has three basic components: motivation, ability and triggers and asserts that in order for a behavior to happen: ”a person must have sufficient motivation, sufficient ability, and an effective trigger”. In the context of shopping behavior, it is reasonable to assume that the same components play an important role. The technical perspective upon human shopping behavior is focused on developing efficient algorithms which can enable automatic behavior recogni-
tion. An attempt towards human behavior analysis while shopping was investigated by Sicre and Nicolas in [7]. They propose a finite-state-machine model for simple actions detection, while the interaction between customers and products is based on Motion History Image (MHI) [8] and Accumulated Motion Image (AMI) [9] description and Support Vector Machines (SVM) classification. An interesting idea towards enhancing shopping experience in a retail store was proposed in [10], by Meschtscherjakov et al. Customers are made aware of the activity of other customers in the shop through a dynamic map, similar to concepts of online-shops, as sales rank. The results of this study proved that customers are interested in areas which are frequently visited by other customers. This pilot study can be implemented using a surveillance system inside a shop and detecting the most visited area(s) by the customers. Further on, this information can be useful to re-arrange products or promotions inside the shop. Willem et al. presents in [11] their work towards uncommon behavior detection, by measuring deviations from defined normal behavior. By clustering trajectories from a shopping mall corridor and the CAVIAR dataset, the most common paths are detected.

Computer vision supports shopping behavior analysis by providing multiple techniques which enable surveillance, trajectory analysis, or action recognition. People tracking, behavior analysis and prediction were investigated by Kanda et al. in [12]. Accumulated people’s trajectories over a long period of time provided a temporal use-of-space analysis, facilitating the behavior prediction task performed by a robot. Still this approach lacks a path-planning process, which is important for notifying the target person of the robot’s presence.
Between the several main research directions which contribute to human behavior assessment, we mention human action recognition. Laptev et al. [13] propose space-time interest points (STIP) in combination with a multi-channel SVM classifier to recognize realistic human actions in unconstrained movies. This method has several advantages such as robustness to occlusions and under different illumination conditions. Another successful approach towards action recognition is presented in [14]. The authors define the integral video to efficiently calculate 3D spatio-temporal volumetric features and train cascaded classifiers to select features and recognize human actions. In the shopping environment, Hu et al. [15] use the MHI along with foreground image obtained by background subtraction and the histogram of oriented gradients (HOG) [16] to obtain discriminative features for action recognition. The proposed approach is improved by building a multiple-instance learning framework SMILE-SVM. This method proved its effectiveness on a real world scenario from a surveillance system in a shopping mall aimed at recognizing customers’ interest in products defined by the intent of getting the merchandise from the shelf.

We presented two different views on human behavior, which can be combined for achieving a deeper understanding of human behavior. Social psychological view offers insights into the motivations and reasons why people behave in a certain manner, enabling behavior modeling, while computer vision provides the means to analyze and recognize shopping behavior in an automatic manner. To the best of our knowledge no study has proposed an automatic system for customers’ shopping behavior assessment, based on trajectory analysis, action recognition, and context related features. In order
to provide an answer to the main research question addressed in this paper, we will present in the following Sections the considered shopping behavior models together with the proposed reasoning model for shopping behavior interpretation.

3. Behavioral Models

There are many ways in which human behavior can be investigated, starting from the traditional methods which involve questionnaires, interviews, or focus groups, online research, scanner data, and ending with the more advanced automatic techniques which are non-intrusive such as audio-video recordings.

Our methodology towards behavior modeling consists of two steps. We started by participation observation of shopping trips of customers in an unobtrusive manner. Based on 20 hours of observations collected by the researchers in shops we defined the following types of shoppers: goal oriented, looking around, disoriented, looking for support, fun-shopper, and duo-shopper. We introduced these types in [2], while the focus of this work consists of the customer-product interaction analysis and of the design and implementation of a computational framework for detecting and assessing shopping behavior. Our study goes beyond the individual shopping behavior, towards social interactions during which people could display certain behaviors not in relation with products but as a cause of other circumstances. For example a mother with children could wonder around not because she is disoriented but to take care of her children, or a person going directly to the coffee corner, is not an example of a goal-oriented shopper but he intends to
get free coffee.

In a second step we validated our assumptions by watching video recordings of shopping trips in a real shop. This brought us a new insight and helped us at refining the proposed models, by revealing different ways in which an action can be performed and also which are the most common combinations of behaviors.

We consider human behavior from a general point of view, in order to understand what triggers it, which are the motivations behind it and finally which are the causes that prevent it from happening. We start from the general human behavior model proposed in [6] and adapt it to the shopping context. It is reasonable to assume that the same components introduced by Fogg: motivation, ability, and triggers play an important role also in the case of shopping behavior. There is always a motivation for which people go shopping such as: need, relaxation, curiosity, desire to be up to the latest trend and so on. Complementary to motivation there is also the ability of performing the actual action of buying a product. From this perspective, ability is highly correlated with the amount of money a person has, but also with other factors such as time and effort required to obtain a certain item. Considering this view we can better understand why a behavior happens or not, even though certain criteria were met. On one hand a person might have the motivation to buy a product, be willing to spend time and effort to find it, but if he doesn’t have enough money, finally he will not buy it. On the other hand, a person might have both the motivation and the money to buy a product, but if it is too difficult to find it, the person might decide to give up. Especially this type of situation we aim at avoiding by providing
an automatic system able to recognize different types of shopping behavior and to trigger an alarm when a customer displays a *disoriented* or *looking for support* behavior. Still, recognizing the other types of shopping behavior is important for gathering statistics about customer preferences in terms of products, areas visited in the shop, or preferred shopping paths and could contribute at enhancing the shopping experience.

Each type of shopping behavior has a number of characteristic features which are presented in Fig. 2 below. The features are displayed next to each behavior along with the possible transitions from one behavior to another one.

![Diagram of different shopping behavior types and the dynamics between them.](image)

We assume the *goal oriented* type of shopper knows what he wants and where to find the product(s) of his interest. If the goes directly to a product
display, takes the product and then heads towards another place, we assume he will not need assistance. The disoriented type of shopping behavior is representative for a customer who does not seem to know what he wants or where to find it. He is going from one place to another one without an apparent plan. A shop assistant might help him at finding a particular product, at selecting something appropriate for him/her, or to make a choice. The customer which does not know where to find a product of his interest and asks for help is called the looking for support type of shopper. On the other side a looking around shopper is assumed to be inspecting the offer, without needing something in particular. He/she might become at a certain moment interested in a product, decide to buy it and in that case his behavior will be similar with the one depicted by the goal oriented shopper. Another type of shopping behavior is the fun-shopper who visits a shop for getting acquainted with the latest promotions or novelties and is rather attracted by a crowd, an exposition, or an event in the shop than by products. Shopping is not only an individual activity but also a social one, case in which people are shopping together with their partner, children, or friends. We call this the duo-shopper type of behavior, which is characterized by trajectories usually one in the proximity of the other one(s) and by interactions between the shoppers, regarding selection or appreciation of products. During a shopping trip, the behavior can remain constant or it can change according to the different situations or products encountered. Any behavioral type can suffer transformations, therefore, we model the behavior, segment based. At the end of the shopping trip we draw a conclusion using a reasoning model. By segment, we mean the transitions from one ROI to another one.
and the behavior displayed by the customer while he is present in that ROI.

While modeling the shopping area we considered the following representative ROIs: entrance/exit, passing, products, pay desk, and resting areas, adding that for a clothes shop also mirrors and fitting rooms ROIs are present (see Fig. 3a). The proposed representation of the relevant ROIs might be too fine-grained, still it needs to be mentioned that the analysis of the customer’s actions is performed only when he is stationary in a specific ROI. A diagram of possible transitions from one ROI to the other ones is depicted in Fig. 3b.

Figure 3: (a) Segmentation of the shopping area into Regions of Interest (ROIs). (b) Transitions between different ROIs.

We present in the next section our approach towards building a computational framework for assessing shopping behavior, based on different behavioral cues.

4. Computational Framework

In order to answer the main research question addressed in this paper (How to design an automatic system for shopping behavior assessment?),
we propose a modular approach and describe next the functionality of each module. A diagram of the proposed framework is shown in Fig.1.

The framework is organised on several levels of abstraction. At the low-level, different types of video cameras are employed, contributing to the efficient gathering of information. A fish-eye camera, mounted on the ceiling, captures the whole scene and is used for people tracking. It has the disadvantage of distorting the image, especially on the borders, region which is very relevant in our case, as it corresponds with the products shopping areas. Therefore, new cameras are installed in the products ROIs facilitating recognition of the customers’ actions. The intermediary level includes the trajectory analysis and the action recognition modules. The position of a customer is analyzed continuously for detecting his presence in the defined ROIs. Every time a customer is stationary in the products ROI, an event is triggered to start the action recognition module, enabling efficient data processing, given that the two types of cameras are synchronized, recording the video feed at twenty frames per second (fps). Finally, the high-semantic level is responsible for combining the intermediary outputs with the context-related features and drawing a conclusion regarding the customer’s shopping behavioral type, by means of the reasoning module. Next, we present in more details each module.

4.1. People Tracking

Tracking algorithms were investigated in the last period, leading to better fitted and online-learning methods. Our purpose consisted in finding a reliable tracker capable of coping with the constraints imposed by our data. Mean shift [17] together with the more recent Predator [18] algorithms were
both considered. The first algorithm described in [2] uses color histograms and the Bhattacharyya distance, while the second one is build on the Lucas-Kanade tracker, and provides long tracking by employing a P-N learning algorithm. Both methods require an initialization phase in which the properties of the object to be tracked are computed. We reduce the manual intervention of the user, by incorporating context properties, which imply that every customer enters the shop in a specific ROI. When a person is detected in the entering ROI, the tracker algorithm is started. People detection is realized using the algorithm presented in [23]. An example of the tracking results is depicted in Fig. 4.

![Tracking of a customer inside the shop.](image)

**Figure 4:** Tracking of a customer inside the shop.

### 4.2. Trajectory Analysis

The output of the tracking module consists of trajectories, which are further analyzed in order to make a distinction between the different types of shoppers. Each trajectory point \((x,y)\) obtained in image coordinates needs to be mapped in the ground-plane coordinates, in order to compensate for the distortion introduced by the wide angle fish-eye camera. We use the following formulas to compute the normalized image coordinates \((x_n, y_n)\), as:
in which \( x_i \) and \( y_i \) are the input image coordinates, \( Height_I \) and \( Weight_I \) the image height and width, \( x_0 \) and \( y_0 \) are the coordinates of the centre of the image adjusted for a possible principal point shift \( p_x \) and \( p_y \):

\[
x_0 = \frac{Width_I}{2} + p_x \quad \text{and} \quad y_0 = \frac{Height_I}{2} + p_y
\]

Next, trajectory features are computed using the normalized coordinates.

Surveillance applications [19] are usually based on feature sets:

\[
f_T = [x_n, y_n, x'_n, y'_n, x''_n, y''_n],
\]

described by position \( (x_n, y_n) \), velocity \( (x'_n, y'_n) \), and acceleration \( (x''_n, y''_n) \), which were considered as a starting point.

Another characteristic of trajectories, which could reveal customers’ shopping intentions is the trajectory orientation, that can be described by including curvature related features. The curvature \( k \) of a trajectory was considered due to its properties such as invariance under planar rotation and translation of the curve [20]:

\[
k = \frac{|x_n y''_n - y'_n x''_n|}{(x'^2 + y'^2)^{3/2}}
\]

Conducted experiments proved that the best feature set for encapsulating trajectory information is the following one:

\[
f_T = [x', y', x'', y'', \sqrt{\Delta x^2 + \Delta y^2}, k]
\]

where \( \Delta x = x(t) - x(t-1) \) and \( \Delta y = y(t) - y(t-1) \).

4.3. Analysis of Regions of Interest

The detection of the customers’ presence inside each of the relevant regions of interest in the shop is meaningful in the framework flow of activities,
since it facilitates the starting of the action recognition module. We defined a number of ROIs: \( R_1 \) (entrance), \( R_2 \) (pay desk), \( R_3 \) (resting), \( R_4 \) (mirror), \( R_5 \) (fitting room), and \( R_{6\ldots8} \) corresponding to the different products areas. The customers’ presence in each ROI is not always easy to detect if we impose strict boundaries, as they could move back and forth while interacting with the products. Therefore, we propose an additional step, consisting of detecting stationary segments. By stationary segment we mean a time interval in which the customer stops inside of a ROI, has a low speed and the area inside he moves is not wider than 1 meter. The features defined in Section 4.2 are used for stationary segments detection. Due to non-linearity of persons’ motion and errors introduced by the tracking algorithm, we use Gaussian smoothing of velocity \( v_n = (x'_n, y'_n) \). We obtain a new velocity value \( v_\mu \) according to the approximation:

\[
v_\mu = \sum_{n=\mu+\sigma}^{n=\mu-\sigma} v_n \ast N(n; \mu, \sigma^2) \tag{5}
\]

where \( N \) is a density function for normal distribution with the mean \( \mu \) and variance \( \sigma^2 \).

Besides, we also extract the following meta-features which are used at the high-semantic level: duration of each stationary segment, number of visits to the same ROI, and the sequence of ROIs visited during one shopping trip.

### 4.4. Action Recognition

Depending on the type of shop, we define different sets of shopping related actions. In a clothes shop, the relevant actions are: *browse through products*, *pick an item* (e.g. coat, shirt, or blouse), *check the item characteristics* such as texture or the price tag, *fit the item next to him/her*, and *try on the item*
in front of the mirror. In a supermarket, another set of shopping actions are most often observed: interaction with products, such as grabbing it from the products display or putting it back, interaction with the shopping basket, and also interaction with the shopping cart.

We propose an approach towards customers’ actions recognition, using the images acquired with the fish-eye camera. The information about the customers’ position is provided by the tracking module. Next, we compute for every frame in a segment a binary mask corresponding to a human in a given trajectory point, according to [21]. By segment we mean the video frames for which the customer is stationary in a products ROI. Then, we combine all binary masks belonging to a segment into one area. The combined binary mask is used to extract image content for every frame.

Figure 5: Overview of the Action Recognition module. From left to right, clockwise: human binary mask from [21], highlighted in red; rectified area defined by the combination of binary mask in the stationary segment; optical flow and corresponding color coding; histogram of optical flow.

The extracted image content is rectified along the radial direction, to
remove the influence of the orientation, i.e. so that all people are upright. Next the motion analysis module is applied to each segment, by estimating optical flow in the rectified areas between every two consecutive frames. We tested several optical flow algorithms both in terms of accuracy and also execution time such as Lucas-Kanade or Horn-Schunk and the best results were obtained using the method proposed by Liu [22].

Motion related patterns are described by computing normalized histograms of motion vectors in 8 directions. We consider different segmentations of the image patch in: 3x1, 1x3, 3x3, 4x4, 5x5, and 6x6 in order to find the best level of granularity. An example of the different steps involved in the analysis is presented in Fig. 5. Fish-eye cameras are already present in real-life scenarios (e.g. supermarkets) for surveillance purposes, while by means of the proposed method they can also be employed for action recognition tasks. Still, this approach has also limitations, such as dealing with different positions of the customers in the scene, occlusions, or reduced quality of the video due to the applied transformations (un-distortion, rectification, and resizing). Therefore, we propose using an additional type of camera, a high-definition one, which captures the side-view of the customers. In the case of the new type of images, we apply a state-of-art method, which was successfully tested in [13] for action recognition in movies. Space-time interest points (STIP) are computed for each video segment and described using 5-bin histograms of optical flow (HOF) and 4-bin histograms of oriented gradients (HOG). The assumption used by this method is that each action is composed of atomic actions, which can be described by STIP feature descriptors (see Fig. 6), and are named visual words.
In order to learn how to discriminate between the considered set of actions, we use a bag-of-words approach. Each action is described as a histogram of visual words, which are obtained by clustering all descriptors using the k-means algorithm. Each features descriptor is assigned to a visual word, using the minimum Euclidean distance.

4.5. Classification Techniques

Given the wide range of available classification approaches, we aim at finding the best model in terms of learning from instances (a set of examples in the training set) and also generalization to new instances. The choice of which specific algorithm to use is highly dependant on the dataset at hand, being a critical step. Classification techniques can be divided into two groups, namely supervised and unsupervised. From the supervised group, we consider both spatial methods (e.g. SVM, Fisher, or k-NN) for discriminating between histograms of visual words, and a spatio-temporal classifier (e.g. Hidden Markov Models (HMMs)) which is used in combination with
trajectory and optical flow features. We selected a HMM-based classification method due to its characteristics such as incorporating dynamics of motion features during time, ability to capture temporal correlations and to deal with noise in measured observations. The learning phase of the HMM consists of adapting the model parameters (number of states, number of Gaussian Mixtures, or the topology which describes possible state-transitions of the underlying Markov process), to the training data by maximizing the a-posteriori probability $P(O|\lambda)$, where $O$ is an observation sequence and $\lambda$ is the HMM model. We present in Section 5.2 our findings regarding the best performing HMM model for each type of features.

4.6. Reasoning Mechanism

All partial information regarding a customer shopping behaviour is integrated in a common reasoning framework and a decision is taken whether a customer would need assistance or regarding his general behavior. We propose a fusion method which combines classification outputs of the trajectory analysis and action recognition modules with meta-features and assigns different weights to each component.

Each trajectory segment, consisting of the corresponding features to a transition from one ROI to another one, is assigned to one of the three considered behavioral classes goal oriented, disoriented, or looking around. In this work we focus on these three most common behaviour categories, still it needs to be mentioned that they do not describe all the possible shopping behaviors. The decision regarding the trajectory label is taken by selecting the HMM model with the highest probability. Next the customer’s interaction with products indicates his/her level of interest, which is decided based on the
type of actions detected while the customer is stationary inside the products ROI. First level of interest corresponds to the browsing action, the second one implies picking an item, checking its characteristics, such as price tag or texture, looking at it more carefully, and/or putting back one item; the third level of interest is considered when a customer is fitting one item next to him, action which denotes that he is interested if that product is suitable to him; the fourth level implies taking the selected item to the mirror ROI and checking how it looks and finally the fifth level of interest corresponds to trying on one item in front of the mirror or in the fitting room. A meaningful action is the asking for help, during which the customer is trying to get the attention of a shop assistant by raising or waiving his hand. If this action is detected the looking for support label is assigned to the output of the action recognition module, receiving the highest weight as an immediate action is required. By meta-features we mean the type of ROIs, the time spent in each of them, and the number of returns in the same ROI. Long time in the products ROI or several returns to the same products ROI could mean need for support and this behavior will be treated with higher priority. Finally the outputs of each component are combined and a decision regarding the behavioral type is taken using the following rules, listed in the order of importance:

1. looking for support → send shop assistant

2. (disoriented | long time in products ROI | several returns to the same products ROI)&(stationary in the pay desk ROI) → disoriented buying

3. (disoriented | long time in products ROI | several returns to the same products ROI) → disoriented, might need assistance
4. \((\text{goal oriented} \mid \text{short time in products ROI}) \& (\text{stationary in the pay desk ROI}) \rightarrow \text{goal oriented buying}\)

5. \((\text{looking around} \mid \text{short time in products ROI} \mid \text{long time in products ROI}) \& (\text{stationary in the pay desk ROI}) \rightarrow \text{looking around buying}\)

6. \((\text{looking around}) \& (\text{goal oriented}) \rightarrow \text{looking around}\)

For each rule which contains \text{time spent in the products ROI}, a decision regarding the level of interest is received from the action recognition module and is considered while deciding the final behavioral type. Still, due to space constraints, we do not include all the possible combinations of trajectory types with the levels of interest and the meta-features. We present next the experimental results obtained on the recorded data.

5. Experimental Results

5.1. Datasets

In order to test our system we use two types of recordings. First we made recordings in our laboratory (ShopLab), see Fig. 7a for testing a proof-of-concept prototype. The laboratory has the appearance of a shop still permitting easy installation of any video devices, which were installed at different points, at the locations depicted in Fig. 7b. The employed video cameras serve different purposes: the fish-eye camera attached to the ceiling is used for tracking, while the HD camera on the shelf serves at action recognition from side-view. All types of shopping behaviors presented in Section 3 were recorded by asking 20 participants, students and researchers to do different tasks (e.g. inspect the shop, select a product which they like, look for a product which was not in the shop, with and without asking for assistance,
shop together with a friend/colleague and interact with each other regarding the choice of products, and find the reason why a crowd of people is in a certain place.

Figure 7: (a) View of the ShopLab by a fish-eye camera. (b) Map of the shopping area indicating the position of the video cameras. (c) Trajectory density map computed on the real-life dataset adopted for the experimental evaluation. Color coding shows in red areas with higher density.

Next we aimed at a realistic environment and we recorded video material in a supermarket, at different time intervals, using a fish-eye camera attached to the ceiling. The dataset consists of 5 hours of recordings, 270 trajectories and 480 shopping related action samples. An example of the acquired type of images is shown in Fig. 7c.

5.2. Experiments

The success rate of the shopping behavior reasoning system is highly dependant on the reliability and performance of the integrating modules. Therefore we perform a number of tests in order to find the best feature descriptors and HMM parameters for each processing module. The performance of a HMM is highly dependant on its topology. In order to determine the best topologies for our HMM models we performed an extensive search,
by employing a diverse number of states (1-10), number of Gaussian Mixtures (1-20), and also network topologies (left-to-right, ergodic model). Next, we applied the same testing approach (10-fold cross validation) for each experiment. We present separately the experiments performed on each dataset.

5.2.1. Laboratory data experiment

A. Trajectories recognition

We investigated different trajectory feature sets and determined that speed and curvature are the most beneficial ones, followed closely by acceleration and Euclidean distance, see Section 4.2. It needs to be mentioned that curvature should be first normalized in order to be in same range with the other features. Regarding classification results, 140 trajectory segments were successfully recognized with an accuracy of 93% into one of the defined classes goal oriented, looking around, and disoriented. The best performing HMM had 8 states, 3 Gaussian Mixtures and a left-to-right topology. The required number of states is bigger for coping with both shorter and longer trajectory segments, while for describing the data distribution a few number of GMMs is sufficient.

<table>
<thead>
<tr>
<th>Trajectory type</th>
<th>Nr. of samples</th>
<th>Accuracy Exp.1</th>
<th>Accuracy Exp.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goal oriented</td>
<td>62</td>
<td>92%</td>
<td>95%</td>
</tr>
<tr>
<td>Disoriented</td>
<td>37</td>
<td>84%</td>
<td>92%</td>
</tr>
<tr>
<td>Looking around</td>
<td>41</td>
<td>88%</td>
<td>92.5%</td>
</tr>
</tbody>
</table>

Table 1: Trajectory types recognition results
The *disoriented* type of trajectories is more difficult to be recognized at the segment level as its characteristics are better highlighted at the global level when more extended information regarding the shopping trip is available. By adding history features to each trajectory segment, such as the number of visits to the same ROI, we were able to increase the recognition rate of the *disoriented* class by 8% and the overall accuracy by 4%. A detailed presentation of the achieved results is included in Table 1.

**B. Action recognition**

In order to refine our analysis, we employed action detection in the stationary segments corresponding to the products ROI. STIP features were computed for all the action samples in the training set, resulting in 200k features. For identifying the optimal size of the visual words codebook we considered several dimensions (30, 50, 100, 200, 300), and the best result was obtained for 100 clusters. Each action sample was represented as a histogram of visual words, which were afterwards classified in one of the considered classes, using different classification techniques: k-Nearest Neighbors (k-NN) classifier, Fisher’s linear discriminant (Fisher), Gaussian Naive Bayes (NB), and linear and non-linear (radial basis function kernel) Support Vector Machines (SVM). The number of shopping related actions and the associated accuracies are presented in Table 2. The best accuracy of 91.6% was obtained using the Fisher classifier.

The integration of the action recognition module into the automatic processing framework, requires real-time capabilities. We used a sliding window approach, by segmenting the data into constant size segments of 1s and we re-trained the prototypes for each action class. The distinction was that
<table>
<thead>
<tr>
<th>Shopping Actions</th>
<th>Nr. of samples</th>
<th>Accuracy k-NN</th>
<th>Accuracy Fisher</th>
<th>Accuracy nBayes</th>
<th>Accuracy SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pick</td>
<td>116</td>
<td>80%</td>
<td>91%</td>
<td>91%</td>
<td>90%</td>
</tr>
<tr>
<td>Browse</td>
<td>67</td>
<td>82.6%</td>
<td>95%</td>
<td>95%</td>
<td>83%</td>
</tr>
<tr>
<td>Check products</td>
<td>140</td>
<td>91.5%</td>
<td>89%</td>
<td>87%</td>
<td>91%</td>
</tr>
<tr>
<td>Fit an item</td>
<td>26</td>
<td>91.5%</td>
<td>87%</td>
<td>85%</td>
<td>90%</td>
</tr>
<tr>
<td>Try on an item</td>
<td>21</td>
<td>97%</td>
<td>96%</td>
<td>96%</td>
<td>95%</td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td></td>
<td>88%</td>
<td><strong>91.6%</strong></td>
<td>90.8%</td>
<td>89.8%</td>
</tr>
</tbody>
</table>

Table 2: Statistics on shopping related actions

Instead of using all frames corresponding to an action, we only considered histograms of visual words for each 20 consecutive frames with an overlap of 0.5s (10 frames) and in combination with a HMM classifier. This resulted in a decrease in accuracy of 6% due to losing discriminative information which might not be present in all segments, but only scarcely in the whole action sample.

D. Semantic analysis of shopping behavior

After we obtained the intermediary outputs of the trajectory analysis and action recognition modules, we combined them with the meta-features consisting of: the sequence of visited ROIs, time spent in each of them and number of visits in the same ROI and applied the reasoning model presented.
in Section 4.6 for discovering the shopping behavioural types. We tested the proposed model on 40 samples of shopping behavior and we were able to correctly recognize the three main types: goal oriented, disoriented, and looking around together with their sub-types: buying or non-buying in 95% of the cases. The miss-classified samples were due to the incorrect inputs provided by the intermediary modules. The segment based analysis is beneficial as it can reveal different behavioral types, leading to a better oriented service for the customers. The proposed reasoning model enables the interpretation of customers’ behavior: a customer which spends a lot of time in the products ROI, without interacting with the products, could have problems finding what he needs, while another customer which returns several times to the same ROI, picks products and then leaves, might behave like this because he remembered he also needs another product, and not because he is disoriented. Therefore, the assessment of the level of interest is an important component in the reasoning process.

5.2.2. Real-life data experiment

If in the case of the ShopLab data recordings we had available a map of the shop and a segmentation into regions of interest, we cannot assume the same information is always available for a supermarket. Therefore we adopt another approach, which aims at detecting first the stationary segments based on the customers’ speed followed by applying action analysis in that segment. We considered 5 types of interaction patterns: stationary, product interaction, shop basket interaction, shop cart interaction and walking relevant for detecting the customer’s level of interest. Normalized histograms of optical flow (HOF) was selected as feature. Adopt-
ing a quantization of the optical flow directions in 8 bins proved to be the best trade-off compared to average length and angle.

Figure 8: Performance of different interaction patterns for HOF, HOF3x3, HOF4x4, HOF5x5, and HOF6x6 features.

Furthermore, we investigated the influence of computing optical flow histograms in separate regions of the rectified image patch, and concatenating them to allow for an increased level of detail. We refer to HOF for the case of a single histogram, HOF3x3 for the subdivision of the image patch in 3 vertical and 3 horizontal subregions, HOF4x4 for the 16 subregions, HOF5x5 for the subdivision in 25 subregions and finally HOF6x6 for the 36 subregions. We found out that the best accuracy of 80.7% was obtained for the HOF5x5, for a a HMM model (left-to-right) with 16 states and 4 GMMs. The results obtained for the different HOF features are shown in Fig. 8,
with the specification that for the \textit{stationary and walking} patterns there is not a high variation between the different types of features achieving 90\% and respectively 88\% recognition rates. The improvement in accuracy of HOF5x5 over the other descriptors indicates that such separation allows to better discriminate between actions, possibly because of different involved body parts and types of motion corresponding to the interaction patterns. The rather low accuracy of 69\% of the \textit{shopping basket interaction} pattern is explained by the limited amount of motion, which is not very visible, being often occluded by the customer’s body. There are several challenges posed by the real-life data which explain the drop in performance in comparison with the laboratory data, such as low quality of the images, distortion in the products area, occlusion and also a quite complex range of modalities in which an action can be performed. For example, the \textit{interaction with products} can be realized in many ways, depending on the customer (e.g. slowly or fast), but also on the specific situation, such as stopping in the middle of the action for another inspection of the product, or grabbing a product, then putting it back and picking another one. All these variations need to be captured by the model, fact which adds even more complexity to the analyzed problem.

6. Conclusions and Future Work

We presented an approach towards semantic interpretation of customers’ shopping behavior applied to both laboratory and real-life recordings in a supermarket. We designed and implemented a first running prototype for automatic assessment of shoppers’ behavior by proposing a framework on multiple levels of abstraction. We made efficient use of different types of
cameras, by extracting trajectories and also action related features. We achieved a best accuracy of 93% for trajectories classification and 91.6% for five types of shopping related actions in laboratory conditions. In the case of the real-life database we were able to discriminate between customer’s interaction with products, with the shopping cart and with the shopping basket with an accuracy of 81%. Different types of shopping behaviors were successfully recognized in 95% of the cases, by employing a reasoning model which combined intermediary features (trajectory and optical flow) with context related ones.

As future work we plan to refine the reasoning process by implementing a probabilistic approach based on Gaussian Mixture Models and to extend the action recognition task by combining different camera views. Furthermore, we plan to extend the analysis of the real-life dataset, to the recognition of all proposed behavioral types. Next, we plan to do a complete analysis of the customers’ shopping behavior, by considering also the social interaction inside a shopping environment.

Acknowledgement

This work was supported by the Netherlands Organization for Scientific Research (NWO) under Grant 018.003.017 and the Visual Context Modeling (ViCoMo) project.

References

ogy, pp. 1001-1008.


Proceedings of the International Conference on Computer Vision (ICCV '09).


