Shopping Behavior Recognition using a Language Modeling Analogy

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

\textsuperscript{a}Section of Interactive Intelligence, Department of Intelligence Systems, 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

Automatic understanding and recognition of human shopping behavior has many potential applications, attracting an increasing interest in the marketing domain. The reliability and performance of the automatic recognition system is highly influenced by the adopted theoretical model of behavior. In this work, we address the analogy between human shopping behavior and a natural language. The adopted methodology associates low-level information extracted from video data with semantic information using the proposed behavior language model. Our contribution on the action recognition level consists of proposing a new feature set which fuses Histograms of Optical Flow (HOF) with directional features. On the behavior level we propose combining smoothed bi-grams with the maximum dependency in a chain of conditional probabilities. The experiments are performed on both laboratory and real-life datasets. The introduced behavior language model achieves an

\textsuperscript{*}Corresponding Author. Phone: +31(0)15 2781537. Fax: +31(0)15 2787141. Email address: m.c.popa@tudelft.nl (M.C. Popa)
accuracy of 87% on the laboratory data and 76% on the real-life dataset, an
improvement of 11% and 8% respectively over the baseline model, by incor-
porating semantic knowledge and capturing correlations between the basic
actions.

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

1. Introduction

Creating machines that can understand human behavior has been a goal
of Artificial Intelligence (AI) since its beginning. However, even today’s most
successful algorithms and systems have only a partial knowledge of the mean-
ing of a person’s behavior. Advances in technology have known a tremendous
progress in the past years, enabling tracking, motion detection, action recog-
nition, and to some extent behavior understanding. The area of application
domains to which awareness of the users’ behavior could contribute, ranges
from affective computing, gaming industry, surveillance, to elderly care or
marketing. In the marketing domain it is of great interest to build a satisfac-
tory relation with the customer, by assessing his/her emotional state [1] 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
product. Furthermore, understanding customers’ behavior has many other
advantages, such as gathering statistics regarding shoppers’ preferences in
terms of products, moments of the day to visit the shop, or areas in the
shop, which could lead to an optimized marketing strategy both for the com-
pany and the customers. Automatic recognition and assessment of human shopping behavior is definitely preferred over the traditional methods represented by questionnaires, interviews, or human observations. Still this task is a challenging one and needs to be modeled accordingly.

We propose in this paper a representation of the shopping behavior analogous to a language processing model. Following a top-down approach, behavior can be decomposed into activities, activities into basic actions, and finally basic actions into human postures. In the same manner, spoken language is composed of sentences, sentences are formed of words and words at their turn can be split into phonemes. Still it is well known that not every combination of phonemes forms a word (e.g. “ghklf”) and that some sequences are more likely than others. Furthermore, regarding words, not any combination of words makes sense and can form a sentence (e.g. “man world try give home”). Similar rules apply also for sequences of basic actions, some combinations being very improbable to happen (e.g. “pick an item, take it off, check how it looks on you”), while others are very likely because they are performed in a meaningful order (“pick an item, check its characteristics, try it on, check how it looks in the mirror, and take it off”). Furthermore, in speech recognition, context plays an important role, contributing to the overall recognition accuracy. Visual scene understanding and human behavior recognition is also affected by contextual knowledge and could benefit a lot from this type of information.

Therefore, we propose in this paper an approach towards shopping behavior modeling based on a probabilistic language processing model, which combines in an efficient manner the characteristics of the computational model.
with semantic information as depicted in Fig. 1.

The proposed computational approach is a bottom-up one, consisting of recognizing basic actions related to shopping on the low-level, detecting sequences of actions on the intermediary level which are translated into activities, and finally recognizing different types of shopping behavior as combination of activities, on the high-semantic level. Previously we investigated different methods towards basic action recognition [2], [3], while the focus of this work is on shopping behavior modeling. The outline of the paper is as follows. In Section 2 we give an overview of related work. Next we provide in Section 3 the theoretical foundations of the proposed shopping behavior model. We introduce in Section 4 the computational model, consisting of the feature extraction methods and classification technique. Next, we provide a description of the used datasets and the experimental results in Section 5. Finally, we formulate our conclusions and give directions for future work.
Automatic analysis of human behavior has been attracting a lot of attention lately due to its potential applications in a variety of domains and posed scientific challenges. To achieve such a challenging task, several research fields focus on modeling human behavior and several methods for modeling behavior have been investigated. Probabilistic graphical models represent one possible alternative due to their properties, such as the ability to capture uncertainty and of generalising to complex activities. Common graphical models for interpreting behavior include static [4] and dynamic Bayesian networks [5], latent Dirichlet allocation model [6], context free-grammars [7], and stochastic context-free grammars [8]. Besides probabilistic graphical models, there are also rule-based methods [9] for behavior interpretation, which require manual user intervention in the form of behavior rules. Given several types of information representative for behavior modeling, a methodology is needed to combine them, by capturing the temporal and spatial relations between the different data streams. An alternative is represented by language modeling which is a powerful tool for dealing with temporal ordering problems and its efficiency has been proven in automatic speech recognition [10]. The common basis shared by language and action has been introduced in [11]. Pastra and Aloimonos propose a biologically inspired generative grammar of action, based on the universal grammar introduced by Chomsky in [12]. The use of language modeling has been proposed in [13], for the recognition of atomic human actions from video sequences. Their contribution consists of proposing a hybrid framework of variable-length Markov models (VLMM) and Hidden Markov Models (HMM), which proved to be successful at rec-
ognizing human postures. The applicability of the method is tested on the
human action recognition level, while it would be very interesting to see how
it performs at the behavioral level. Chen et al. propose in [14], a similar con-
cept and model behavior as a language using a different type of input data,
namely accelerometer readings. The complexity of the analyzed problem is
kept simple, by including three types of actions (“walking”, “jogging”, and
“cycling”) and its efficiency is evaluated for different n-gram models. Tzouk-
ermann et al. use language models in [15] for the optimal representation of
textual information and integrate them with the overall vision system. This
work illustrates the power of language models at using temporal information
contained in the text at increasing the annotation accuracy.

The previously presented works on language modeling highlight its capabili-
ties at dealing with temporal and semantic information, still none of them, to
the best of our knowledge applies it in a complex environment for behavior
representation and recognition. We introduce in the next Section, our pro-
posed behavior language model, applied in the context of shopping behavior
representation.

3. Language Behavioral Model

We investigated the applicability of both deterministic models and proba-
bilistic models (Bayesian networks) [16] for shopping behavior assessment. In
this paper we propose a probabilistic model inspired by speech recognition,
namely a behavior language model. In language processing, a set of words is
analysed in order to assess if it forms a grammatically correct sentence in a
given language. The typical use is then to choose from a set of hypotheses
the most plausible one. Our goal consists of defining a set of behavior models \( B = \{ B_1, B_2, \ldots, B_m \} \) and analysing the ability of the proposed language model to recognize a new behavior sample \( B_i \) as one of the already defined ones. Each behavior model is formed of a number of terms or behavioural cues, such as trajectory types, basic shopping related actions, and their relation with the special areas in the shop which are referred to as Regions of Interest (ROIs). This section describes the methodology of the proposed model.

We model behavior by means of a probabilistic approach, which was preferred over a deterministic one. In the case of a deterministic model, all rules need to be defined in advance, limiting the flexibility of the model in the presence of unseen data. On the other hand a probabilistic approach is capable of working with real world data and to handle noisy sensory input data. The mechanism used to select a behavior model is as follows: the probability of a behavior sample is estimated against each defined behavior model and the one with the highest probability is selected as the behavior class of the sample.

Next, we describe the approach adopted for estimating the probability of a behavior sample \( B_i \) under the assumption of a general behavior model \( M \), namely \( P(B_i|M) \). We exploit the relations between the constituent terms \( B_i = \{ t_1, t_2, \ldots, t_n \} \) of a behavior model, by formulating several assumptions. Initially, we assume term independence, meaning that the behavior terms do not depend on each other, relation described by the following equation:

\[
P(B_i|M) = P(t_1, t_2, \ldots, t_n|M) = \prod_{i=1}^{n} P(t_i|M)
\]  

(1)
Even though, this formula would be preferred due to the low computational complexity, it is not a valid one in practice, as there are definitely dependencies between the constituent terms of a behavior sample. Furthermore accepting this assumption would mean that the order of the behavior terms is not important, which is also not true, as ordering contributes to the semantic meaning of a sequence of terms. We incorporate the dependence assumption into (1), using conditional probabilities:

$$P(B_i|M) = P(t_1, t_2, \ldots, t_n|M) = P(t_1|M) \times P(t_2|t_1, M) \times P(t_3|t_1, t_2, M) \times \ldots \times P(t_n|t_1, t_2, \ldots, t_{n-1}, M),$$

(2)

Still, the expansion of a behavior sample probability into a chain of conditional probabilities is very hard to be computed, due to the sparse nature of training data. The probability of a term $t_j$, given that a sequence of $(j-1)$ terms has been observed can be computed in a reliable way, only if in the training dataset we have enough combinations of behavior terms, which is definitely very hard to achieve. There are several possible approximations to this exact formula which can be considered. A first solution implies ignoring the higher order dependencies in each term, and selecting from the conditioning variables, one particular variable which reflects the strongest dependency. In other words we select in the right hand side of the equation the variable which maximizes the probability value, and approximates each probability term in (2) by:

$$P(t_i|t_1, t_2, \ldots, t_{i-1}, M) \approx \max_{1 \leq k < i} P(t_i|t_k, M)$$

(3)

This approximation represents a partial solution between the independent and the complete dependant probability models. However, it lacks the ability
to incorporate the ordering attribute. Therefore, we propose using \textit{n-gram} models, which imply that the n-th term in a sequence will be described only in relation with the last (n-1) terms. In speech processing, the most common choices of n are two or three, and the resulting models were proved to be powerful enough in spite of their simplicity. We chose for a bi-gram model and applied it in equation (2). This operation helps reducing the complexity and results in:

\[ P(t_1, t_2, \ldots, t_n|M) = P(t_1|M) \prod_{i=2}^{n} P(t_i|t_{i-1}, M), \]  

(4)

A bi-gram model does not solve completely the problem, as it only uses local information. In some cases, the i-th term might not depend on the last previous term, but on the last four ones, while in other cases it might even not depend on any previous terms. Still, based on the observations gathered on our data we noticed that in most of the cases this assumption holds and represents a satisfactory solution. The advantage of incorporating history information up to the second level, consists of efficiently decomposing a sequence of terms into smaller components, while the disadvantage is that we lose any distant-range dependencies between the behavior terms. Both proposed solutions in equations (3) and (4) have advantages, still they are unable to meet all demands (dependency and ordering) in an efficient manner. Therefore, we propose using a combined solution, which either selects from the both proposed approximations the best one:

\[ P(t_i|t_1, t_2, \ldots, t_{i-1}, M) \approx \max(P(t_i|t_{i-1}, M), \max_{1 \leq k < i} P(t_i|t_k, M)) \]  

(5)
or, it approximates them using a weighted average:

$$P(t_i|t_1, t_2, \ldots, t_{i-1}, M) \approx (w_1 \times P(t_i|t_{i-1}, M)) +$$

$$+ w_2 \times \max_{1 \leq k < i} P(t_i|t_k, M) / 2$$

(6)

In order to satisfy the constraint imposed on the probabilities to sum up to one, we also apply a normalization after applying the above specified formulas.

Next we discuss how to compute the model parameters. A simple solution is to count the relative frequencies in a training data set, given that it contains enough examples of different combinations of the behavior terms.

$$P(t_i|M) = \frac{C(t_i, M)}{\sum_{t \in M} C(t, M)}$$

(7)

where $C(t_i, M)$ is the number of occurrences of behavior term $t_i$ in the behavior model $M$. The joint probabilities are computed using the same principle:

$$P(t_i|t_j, M) = \frac{C(t_i, t_j, M)}{n_2(M)}$$

$$P(t_i|t_j, t_k, M) = \frac{C(t_i, t_j, t_k, M)}{n_3(M)}$$

(8)

where $C(t_i, t_j, M)$ and $C(t_i, t_j, t_k, M)$ are the number of occurrences of behavior term sequences: $(t_j, t_i)$ and $(t_k, t_j, t_i)$ in the behavior model $M$, and $n_i(M)$ represents the total number of ordered sequences formed of $i$ terms which appear in model $M$. Even for a very large dataset, the problem of unseen events is an important one, and for estimating frequencies in a reliable way we need to compensate for unseen behavior terms combinations. An often applied solution to this problem are the so called smoothing techniques [17] which discount the relative frequencies of seen events and distribute the
We also applied smoothing for estimating the model parameters defined in (7) and (8):

\[
P(t_i|M) = \lambda \frac{C(t_i, M)}{\sum_{t \in M} C(t, M)} + (1 - \lambda) \frac{C(t_i, D)}{\sum_{t \in D} C(t, D)}
\]

\[
P(t_i|t_j, M) = \lambda \frac{C(t_i, t_j, M)}{n_2(M)} + (1 - \lambda) \frac{C(t_i, t_j, D)}{n_2(D)}
\]

(9)

where \(C(t_i, D)\), and \(C(t_i, t_j, D)\) are the number of occurrences of behavior term sequences: \(t_i\), and \((t_j, t_i)\) in the database \(D\), and \(n_i(D)\) represents the total number of ordered sequences formed of \(i\) terms which appear in the database \(D\), where \(D\) contains all behavior models.

We present in the next section the computational approach used to map sensory input data into behavior terms and the applicability of the proposed model to a real problem, namely shopping behavior representation in two case scenarios.

4. Computational Approach

The input to the proposed language behavior model is provided by the computational model. By computational model we mean the methodology used to recognize basic shopping related actions using feature representation techniques and classification methods. Next, the set of recognized basic actions conveys different semantic meanings based on their combination and time ordering, dynamics captured by the behavior language model. The design of the proposed approach towards behavior modeling is depicted in Fig. 2, being organized in a hierarchical manner.

Our methodology towards modeling and recognizing behavior is composed of several steps.
A) **Shopping related actions definition.** First, we define the scope of our problem by identifying the set of shopping related actions, depending on the type of shop. 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. While, 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.

**B) Action recognition features.** Next, we introduce a method towards recognition of the proposed shopping related actions, based on analyzing motion patterns in the vicinity of a person. People detection is performed using the algorithm presented in [18]. Next the motion analysis module is applied, by estimating optical flow in the people corresponding 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 [19]. Motion related patterns are described by computing normalized histograms of motion vectors. The histogram size varies from 4 to 16, in a search for finding the optimal number of bins. Next, 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 (see Fig. 3a). 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.

Besides motion vectors histograms, which describe motion locally in each image patch, we also propose using directional features which encapsulate global characteristics of motion. Directional features are obtained by computing in-wards and out-wards motion between blocks, as depicted in Fig. 3b. The two types of features are complementary and therefore, a new feature vector is obtained by fusing them.

**C) Action classification.** For classification of motion related features we selected a Hidden Markov Model (HMM) classification method due to its char-
Figure 3: (a) Computational scheme of motion related features extracted from the segmented image patch and concatenated into histograms. (b) Directional motion features.

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.
D) Language Behavior Model. At the higher semantic level, the recognized basic actions by HMMs are combined in sequences and a conclusion is drawn regarding the different types of shopping behavior. At each moment in time, the probability of observing each basic action is computed and the one with the maximum likelihood is selected. Given that the action recognition task is prone to errors, one or several miss-classified sample(s) can influence the accuracy of the behavior recognition task. The proposed solution to this problem consists of applying a semantic model on top of the computational model. By incorporating semantic information represented by the most likely action(s) to be observed if a set of other actions were already detected, we restrict the scope of the problem and reduce the search space of the possible hypotheses. We used the Viterbi algorithm [20] for finding the approximate most likely sequence of basic actions, given the probabilities of observations for each time step and the most likely transitions from one basic action to another one. In the implementation process, we adapted the Hidden Markov Model Toolkit (HTK) [21], in order to make it suitable for our recognition task, meaning that we had to define proper configuration, prototype, and network files. The discussion of the obtained results is presented in the next section.

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) captured with a fish-eye camera in Fig. 4a, for testing a proof-of-concept prototype.
The laboratory has the appearance of a shop still permitting easy installation of any video devices, at the locations showed in Fig. 4b. We used a HD camera on the shelf to perform action recognition from side-view, an example of the acquired type of images is presented in Fig. 4c.

Figure 5: Examples of shopping sequences.
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 prod-
ucts, and find the reason why a crowd of people is in a certain place). Ex-
amples of shopping sequences existent in the database are shown in Fig. 5.
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,
480 shopping related action and 80 shopping behavior samples. An example
of the acquired type of images is shown in Fig. 6.

Figure 6: (a) View of the real-life database by a fish-eye camera. (b) Examples of shopping
sequences.

5.2. Experiments

The success rate of the shopping behavior recognition system is highly
dependant on the reliability and performance of the action recognition mod-
ule. Therefore we perform a number of tests in order to find the best feature
descriptors and HMM parameters. 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 net-
work 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

We first investigated the efficiency of the proposed action recognition
method by computing optical flow histograms in separate regions of the rec-
tified image patch and concatenating them to allow for an increased level
of detail. The results obtained for the different types of HOF features
(HOF, HOF3x3, HOF4x4, HOF5x5, and HOF6x6) introduced in Section 4
were computed and the best result was obtained for the HOF5x5 descriptor,
achieving an overall accuracy of 85.7% for a HMM model (left-to-right) with
16 states and 2 GMMs. Next, we found out that by fusing HOF4x4 with
directional features the overall accuracy was improved with 3%. We need to
mention that directional features were fused with HOF features only for the
cases with a segmentation into an even number of blocks (e.g. HOF4x4 and
HOF6x6). The improvement in accuracy of the fused vector (HOF4x4 and
directional features) in comparison with HOF5x5 is explained by its ability
to combine both local motion information inside each considered block and
global motion patterns computed between the right and left side of the image
patch. The comparative results are presented in Fig. 7 below.

Secondly, we investigated the performance of the proposed language be-
Figure 7: Statistics on shopping related actions for HOF5x5 and HOF4x4+Directional features.

<table>
<thead>
<tr>
<th>Shopping Actions</th>
<th>Nr. Of Samples</th>
<th>Accuracy HOF5x5</th>
<th>Accuracy HOF4x4+ Directional features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pick</td>
<td>116</td>
<td>91.2%</td>
<td>93.5%</td>
</tr>
<tr>
<td>Put</td>
<td>74</td>
<td>79%</td>
<td>82.3%</td>
</tr>
<tr>
<td>Browse</td>
<td>67</td>
<td>90.5%</td>
<td>92.6%</td>
</tr>
<tr>
<td>Fit next to you</td>
<td>35</td>
<td>85%</td>
<td>89%</td>
</tr>
<tr>
<td>Check how it looks</td>
<td>48</td>
<td>87%</td>
<td>90%</td>
</tr>
<tr>
<td>Try on</td>
<td>30</td>
<td>92.5%</td>
<td>94.3%</td>
</tr>
<tr>
<td>Take off</td>
<td>30</td>
<td>74.5%</td>
<td>78.5%</td>
</tr>
<tr>
<td>Overall Acc.</td>
<td>-</td>
<td>85.7%</td>
<td>89%</td>
</tr>
</tbody>
</table>

...behavior model. We applied it on 120 behavior samples composed of different sequences of basic shopping actions. Initially, we implemented and tested a general behavior model without adding any semantic information. We analyzed the obtained results, in terms of accuracy at different levels of granularity: action level, activities level, and finally behavior level. The recognition rates for each level are presented in Fig. 8a. At a second stage, we tested the performance of a bi-gram model, which incorporates information regarding the most likely action to be detected, based on the previous detected one. The improvement in accuracy of the bi-gram model compared to the general one was substantial causing an increase of 8%. Despite its simplicity, this model is powerful enough to capture first level dependencies between basic actions and to filter out many infrequent combinations. For example, due to the similarity between the actions pairs (pick - put) and (try on - take off),...
they are confused in 8% and respectively 10% of the cases. The benefit of using semantic knowledge translates into decreasing the likelihood of combinations which are less probable to happen, such as (browse, put back an item) or (pick, take off).

Figure 8: (a) Statistics on different levels of granularity of actions, activities, and behavior recognition measures. (b) ROC curves corresponding to the proposed language behavior models.

<table>
<thead>
<tr>
<th>Level of granularity</th>
<th>Baseline Model</th>
<th>Bi-gram Model</th>
<th>Bi-gram + Dependency Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actions</td>
<td>85%</td>
<td>92.2%</td>
<td>94.8%</td>
</tr>
<tr>
<td>Activities</td>
<td>80%</td>
<td>86%</td>
<td>89.5%</td>
</tr>
<tr>
<td>Behavior</td>
<td>76%</td>
<td>84.3%</td>
<td>87%</td>
</tr>
</tbody>
</table>

The problem of modeling shopping behavior is not completely solved by using a bi-gram model. Next, we analyzed the miss-classified samples and observed that there are errors due to dominant dependencies, up to the third or fourth history level, which cannot be captured using a bi-gram model. Applying a tri-gram model was not an option due to the sparsity of data and low frequency of many tri-grams. Therefore, we modeled the dominant dependency attribute by incorporating an additional constrained for corresponding set of actions (pick - put) and (try on - take off). The probability of observing the second action is increased, only if the first one was already observed, otherwise its likelihood is decreased to a negligible amount. The de-
pendency attribute increased the discriminative power of the model, leading to an improvement of 3%, by correctly detecting the previous miss-classified sequences (e.g. “pick, check how it looks, take off” or “try on, check how it looks, put it back”). The ROC curves corresponding to the proposed language behavior models are presented in Fig. 8b, while the reported accuracies are obtained for a false positives rate (FPr) of 25%. The influence of the language behavior model can be noticed not only on the high-semantic level, but also on the action and activity recognition levels which benefit from the additional semantic knowledge, achieving a significant improvement.

5.2.2. Real-life data experiment

In order to test the applicability of the proposed language behavior model to a real-life database, we followed the same methodology as presented in the previous sub-section. We first investigated the best feature descriptor for discriminating between the different customer-product interaction patterns (pick, check products, put, interaction with the shopping basket, interaction with the shopping cart). We found out that the best accuracy between the different HOF features was obtained for the HOF5x5 descriptor of 80.7%, for a HMM model (left-to-right) with 16 states and 4 GMMs. 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 72% of the 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. The fusion between HOF4x4 and directional features was beneficial also in this case,
leading to an overall improvement of 2.3%. The comparative performances
of the two best feature descriptors are depicted in Fig. 9.

Figure 9: Statistics on customer-product interaction patterns for HOF5x5 and
HOF4x4+Directional features.

![Figure 9: Statistics on customer-product interaction patterns for HOF5x5 and HOF4x4+Directional features.](image)

Next, we applied both a baseline and a bi-gram model to 80 shopping
behavior samples containing combinations of customer-product interaction
patterns. The bi-gram model gave a 6% improvement over the baseline
model, proving the efficiency of incorporating semantic information. The
overall recognition rates obtained for the three levels of granularity are pre-
sented in Fig. 10a together with the ROC curves corresponding to the three
behavior models in Fig. 10b. The error analysis showed that there are distant
dependencies which cannot be captured efficiently by a second-order model.

By applying the dominant dependency attribute we were able to constrain
the detection of an interaction pattern (put, put in the shopping basket, or put in the shopping cart) only if another one was already detected (e.g. pick). The obtained improvement was of 2%, by correctly modeling the order of actions. Still, the improvement was not so significant, due to the confusion between the three interaction patterns.

The conducted experiments proved that the proposed language behavior model was successful also in the case of a difficult and complex environment. 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, factors which have a direct influence on the action recognition level. Furthermore, real-life sequences contain many variations of the order of interaction patterns (e.g. usually customers interact with the shopping cart and then

<table>
<thead>
<tr>
<th>Level of granularity</th>
<th>Baseline Model</th>
<th>Bi-gram Model</th>
<th>Bi-gram + Dependency Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actions</td>
<td>77.4%</td>
<td>82%</td>
<td>84%</td>
</tr>
<tr>
<td>Activities</td>
<td>73%</td>
<td>77.4%</td>
<td>79%</td>
</tr>
<tr>
<td>Behavior</td>
<td>68%</td>
<td>74%</td>
<td>76%</td>
</tr>
</tbody>
</table>
walk away, others walk away first and then interact with the shopping cart, while some are just grabbing a product without any interaction with the shopping cart or basket), fact which increases the complexity of the analysed problem.

6. Conclusions and Future Work

We presented a language modeling approach towards recognition of customers’ shopping behavior applied to both laboratory and real-life recordings in a supermarket. We decomposed the problem on several levels of abstraction and made efficient use of the different types of information, both sensory and semantic. On the action recognition level, we proposed a new feature set composed of both HOF4x4 and directional features and achieved an overall accuracy of 89% for the recognition of seven shopping related actions. In the case of the real-life database, the obtained accuracy was of 83% for five customer-product interaction patterns. Next, on the behavior level, we introduced a model for shopping behavior representation, able to deal with continuous data, meaning different combinations of shopping related actions, and not only restricted to the discrete case in which several defined patterns would be considered. We implemented and tested different variants of language models. By means of a smoothed bi-gram model, we were able to improve the baseline model with 8%, proving the importance of including semantic knowledge. Furthermore, by combining the bi-gram model with the maximum dependency in a chain of conditional probabilities, a second improvement of 3% was obtained, leading to an overall recognition accuracy of 87% for the ShopLab experiment and 76% for the real-life shop data.
As future work we plan to refine the analysis of shopping behavior by integrating a module for appreciation of products using facial expression analysis. Next, we also aim at conducting 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**


