Detecting Customers' Buying Events on a Real-Life Database

Mirela C. Popa[1], Tommaso Gritti[2], Leon J.M. Rothkrantz[1], Caifeng Shan[2], and Pascal Wiggers[1]

[1] Man-Machine Interaction Group, Delft University of Technology, Delft, The Netherlands
m.c.popa@tudelft.nl

[2] Video and Image Processing Group, Philips Research, Eindhoven, The Netherlands
{tommaso.gritti,caifeng.shan}@philips.com

Abstract. Video Analytics covers a large set of methodologies which aim at automatically extracting information from video material. In the context of retail, the possibility to effortlessly gather statistics on customer shopping behavior is very attractive. In this work, we focus on the task of automatic classification of customer behavior, with the objecting to recognize buying events. The experiments are performed on several hours of video collected in a supermarket. Given the vast effort of the research community on the task of tracking, we assume the existence of a video tracking system capable of producing a trajectory for every individual, and currently manually annotate the input videos with trajectories. From the annotated video recordings, we extract features related to the spatio-temporal behavior of the trajectory, and to the user movement, and analyze the shopping sequences using a Hidden Markov Model (HMM). First results show that it is possible to discriminate between buying and non-buying behavior with an accuracy of 74%.

Keywords: Trajectory analysis, Optical flow, Hidden Markov Models, Shopping Behavior.

1 Introduction

There is an increasing amount of research in the area of video analytics and semantic interpretation as an application to automatic surveillance, traffic monitoring, video games, marketing, etc. In the field of marketing it is of primary concern to identify the most appealing products and services for customers and to maximize their impact on the shopping behavior. Computer vision provides multiple techniques which enable surveillance [5], action recognition, and behavior interpretation of customers. Tracking people inside the shop can have many applications, such as global shopping behavior recognition, region of interest detection both individually and for a group of customers, measured at a specific moment or over time intervals. We plan to use the existing surveillance systems to observe the shopping behavior of people [4], to get a better understanding of

A. Berciano et al. (Eds.): CAIP 2011, Part I, LNCS 6854, pp. 17–25, 2011.

their needs. The action recognition module can provide cues regarding customers' interest in products and can help interpreting different interaction patterns, such as grasping a product immediately, after a period time or even after more visits at the same place. In this paper we propose an automatic surveillance system for detecting customers' buying behavior based on tracking and motion information and tested on real-life recordings in a shopping mall. Its applicability resides in identifying different buying patterns in terms of number of interactions and time spent in the vicinity of a product but also in finding for which products categories the customers have trouble deciding. As a result, appropriate actions could be taken such as new products arrangements and more efficient usage of the store space. Next we provide an overview of related studies, then the design of our system is presented, followed by the data acquisition process and the experimental results section. Finally we formulate our conclusions and give directions for future work.

1.1 Related Work

People tracking, behavior analysis, and prediction were investigated by Kanda et al. in [3]. 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. Hu et al. [2] used the Motion History Image (MHI) along with the foreground image obtained by background subtraction and the histogram of oriented gradients (HOG) [1] to obtain discriminative features for action recognition. Next a multiple-instance learning framework SMILE-SVM was build to improve the performance. This approach 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 merchandize from the shelf. These approaches are suitable for action recognition under varying conditions in complex scenes such as background clutter or partial occluded crowds; still they require supervised learning based on a large reliable dataset. 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 MHI and accumulated motion image (AMI) [8] description and SVM classification. It remains to be proved and tested whether this method will be applicable in an uncontrolled real-life scenario which deals with occlusions and different types of settings. Another issue regards the variability of performing an action in relation with the dataset size, which in this case is limited to 4 persons.

2 Proposed Methodology

Based on observations made in real shops we proposed a number of shopping behavior models as described in [4]. There are many individual differences in shopping behavior of people. Some shoppers know what they want and the location of that product (*goal oriented*), others prefer to inspect the offer (*looking*

around), some are helpless and would need support (*disoriented*), while others are actively looking for assistance, finally some shoppers are just looking for interesting products or just enjoy being in a shop (*fun-shopper*). We assume the ultimate goal of shopping is to buy a required product. Next the design of our system for automatic assessment of customers' buying behavior is presented. We propose a modular approach and we describe next the functionality of each module. A diagram of the proposed system is shown in Fig. 1.

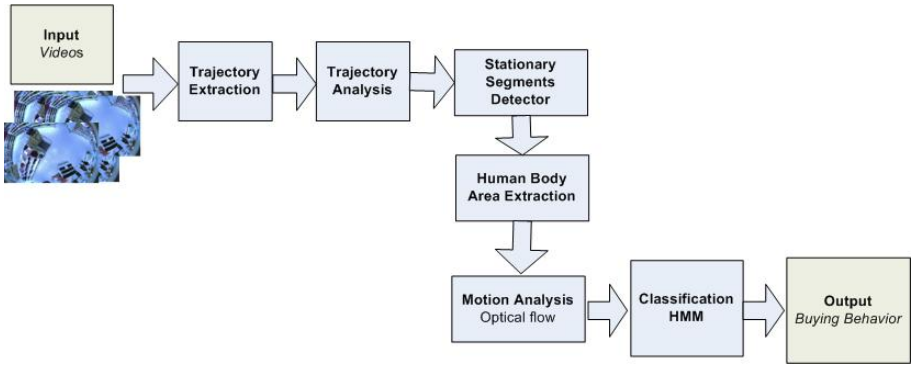

Fig. 1. System Overview

2.1 Trajectory Extraction

First the trajectory extraction module is employed. Currently the customers' trajectories are manually labeled, given that our goal consisted in the high-level analysis of behavior. In our future work, trajectories will be extracted by adopting person detection and tracking. For this task we used our frame based annotation tool which enables both person and event annotation.

2.2 Trajectory Analysis

Global motion analysis provides a first insight into customers' shopping behavior. Therefore the Trajectory Analysis module is employed to extract relevant trajectory features. We started from the feature set $f_T = [x, y, x^{'}, y^{'}, x^{''}, y^{''}]$ proposed in [11], described by position (x,y), velocity $(x^{'}, y^{'})$, and acceleration $(x^{''}, y^{''})$. We decided to exclude spatial features (x,y) in order to prevent learning of a preferred shopping path, while our interest resided in capturing motion characteristics. The curvature k of a trajectory was considered due its properties such as invariance under planar rotation and translation of the curve [9].

$$k = (y^{''}/x^{'})^2/(\sqrt{}((y^{'}/x^{'})^2 + 1))^3$$

Based on experiments we noticed that the best feature set for encapsulating trajectory information was 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 $\Delta y = y(t) - y(t-1)$.

2.3 Stationary Segments Detector

The next module of our system is responsible for detecting segments of interest and potential buying segments. The detection is performed using the features defined in the previous section. Due to non-linearity of persons' motion and errors introduced by the manual annotation, we used Gaussian smoothing of velocity. In this way each velocity value v_μ is approximated by:

$$v_\mu = \sum_{t=\mu-\sigma}^{t=\mu+\sigma} v_t * N(t; \mu, \sigma^2) \quad (2)$$

where N is a density function for normal distribution with the mean μ and variance σ^2.

2.4 Human Body Area Extraction

For each trajectory segment detected by the previous module, the human body area is extracted. To this aim, for every frame in a segment, we estimate a binary mask corresponding to a human in a given trajectory point, according to [6]. We then combine all binary masks belonging to a segment into one area. The combined binary mask is used to extract image content from every frame. The extracted image content is rectified along the radial direction (see an example in Fig. 2), to remove the influence of the orientation, i.e. so that all people are in the upright position.

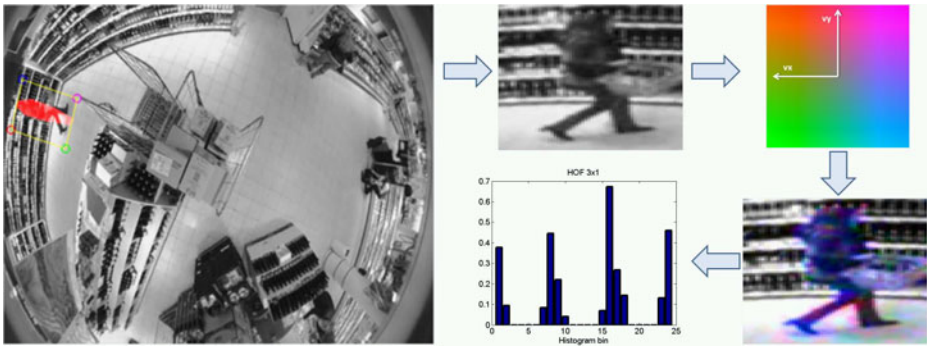

Fig. 2. Overview of the Human Body Area Extraction and Motion Analysis modules. From left to right, clockwise: human binary mask from [6], highlighted in red; rectified area defined by the combination of binary mask in the stationary segment; optical flow and corresponding color coding; histogram of optical flow.

2.5 Motion Analysis

We assume buying behavior can be characterized by motion patterns, such as picking a product and putting it in the shopping basket. The motion analysis module is applied to each segment, by estimating optical flow in the rectified

Fig. 3. Motion Flow Visualization. Overlay color is according to color coding shown in Fig. 2.

areas between every two consecutive frames. Normalized histograms of motion vectors in 8 directions are extracted from the whole image patch and also by considering a image patch segmentation into three regions corresponding to the approximate position of the head, body and legs of a person. 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 [10]. An example of a buying event is depicted in Fig. 3.

2.6 Classification

Classification techniques can be divided into two groups, namely supervised and unsupervised. From the supervised group (e.g. Hidden Markov Models (HMMs), SVM, and Gaussian Mixture Models (GMMs)) we chose a HMM-based classification method due to its characteristics such as incorporating dynamics of motion features during time and ability to capture temporal correlations. The extracted features (trajectory and optical flow) were fed to a HMM and the maximum likelihood rule was used to decide the label corresponding to each interesting trajectory segment.

3 Experimental Results

3.1 Data Acquisition

In order to test our system in a realistic environment, we recorded video material in a supermarket, at different time intervals, using a fish-eye camera attached to the ceiling. An example of the acquired type of images is shown in Fig. 2.

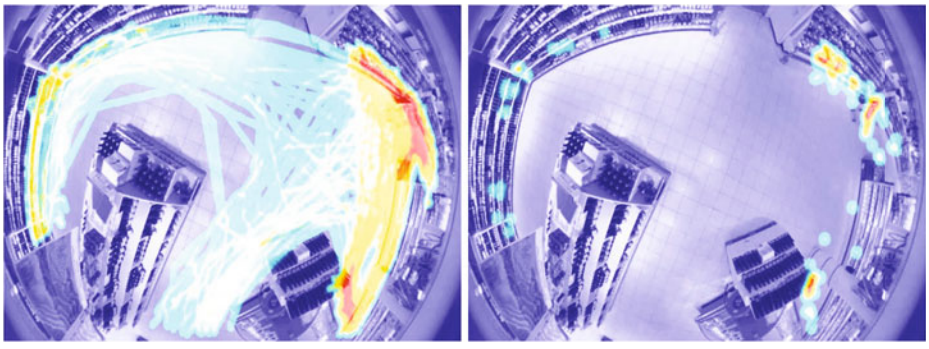

Fig. 4. Trajectory density map (left), and buying event density (right), computed on the dataset adopted for the experimental evaluation. Color coding shows in red areas with higher density.

We collected and manually annotated approximately 5 hours of recordings resulting in 270 customers' trajectories, from which 100 trajectories contained buying segments. We define as a *buying event* an action of a customer who picked a product and put it in the shopping basket or just took it with him/her. The total number of annotated buying events was 130, since some of the trajectories contained more than one buying event. A density map of the annotated trajectories and for the buying events is shown in Fig. 4. We present next the experimental results obtained using the recorded data.

3.2 Experiments

We performed a number of tests in order to find the best feature descriptor and HMM topology for our buying behavior analysis system as described in Section 2. We investigated different detection methods of potential buying trajectory segments. The aim was to detect automatically the segments containing buying events. From our observations of buying behavior we noticed that the action of buying a product usually happens after the customer stopped for a period of time in the products area. By employing a stationary detector as described in Section 2.3 based on slow velocity and duration of staying in the vicinity of a product of at least one second we were able to detect 90% of all the buying actions, meaning 118 segments out of 130. The rest of 10% were associated with a different type of behavior (goal oriented) characterized by a customer which knows what he wants and picks that product very quickly and then continues his shopping trip. By applying the stationary detector the number of analyzed video frames (N) was reduced to 67% of which 17% corresponded to buying segments and 50% to non-buying ones.

In order to refine our analysis, we employed motion analysis in the detected stationary trajectory segments. Normalized histograms of optical flow (HOF) was selected as feature. Adopting a quantization of the optical flow directions

in 8 bins proved to be the best tradeoff compared to average length and angle. Furthermore, we investigated the influence of computing optical flow histogram 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, HOF3x1 for the case of subdivision of the image in 3 vertical subregions, and HOF3x3 for the subdivision of 9 subregions. 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). We found out that the best accuracy of 74% was obtained for a HMM model (left-to-right) with 6 states and 2 GMMs, for HOF3x3, using a 9-fold cross validation testing approach. The ROC curves obtained for the different HOF features are shown in Fig. 5. The improvement in accuracy of HOF3x3 over HOF3x1 and HOF features indicates that such separation allows to better discriminate actions, possibly because of different body parts movement which are related to the buying actions.

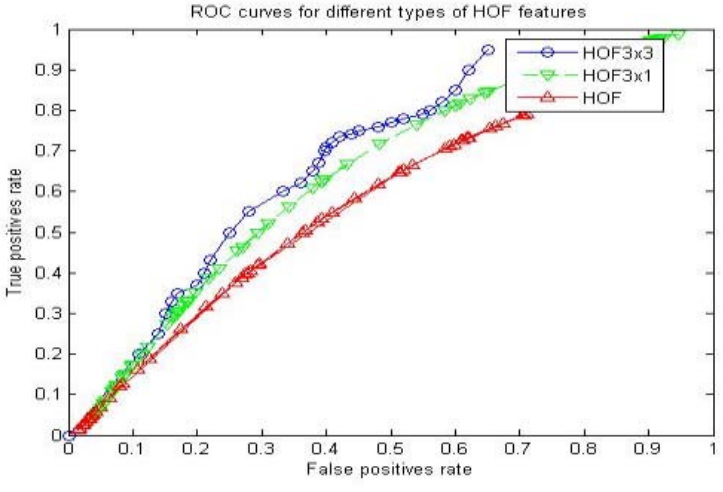

Fig. 5. ROC curves for HOF features

The number of HMM training iterations plays an important role at determining the best tradeoff between true positives and false positives rate. In our case 10 iterations were enough to learn the two HMM models, while using a bigger number of iterations led to overfitting the non-buying HMM model in detriment of the buying one. The miss-classified false negatives are due to the challenging type of data, such as occluded buying sequences either by another customer or by the person herself, while the false positives contain examples of customers' picking a product and then putting it back or just interacting with the shopping

basket without putting a new product. Still given the difficulty and variability of the recorded data we consider that our approach is quite good at detecting customers' buying behavior under varying conditions.

4 Conclusion and Future Work

We presented an approach towards understanding customers' shopping behavior applied to real-life recordings in a supermarket. We designed and implemented a first running prototype for detecting customers' buying behavior. We used global features extracted from trajectories to detect potential buying segments and we extracted optical flow from an interesting area for action recognition. We achieved a best accuracy of 74% by using HOF3x3 features. As future work we plan to improve and refine the action recognition module by using different types of features such as interest-points models and an extended set of shopping related actions. We also aim at extending the system by fusing the data from cameras at different location and view angles.

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

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, California, vol. 1, pp. 886–893 (June 2005)
2. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action detection in complex scenes with spatial and temporal ambiguities. In: Proceedings of International Conference on Computer Vision, ICCV 2009 (October 2009)
3. Kanda, T., Glas, D.F., Shiomi, M., Ishiguro, H., Hagita, N.: Who will be the customer? A social robot that anticipates people's behavior from their trajectories. In: Int. Conf. on Ubiquitous Computing, UbiComp 2008 (2008)
4. Popa, M.C., Rothkrantz, L.J.M., Yang, Z., Wiggers, P., Braspenning, R., Shan, C.: Analysis of Shopping Behavior based on Surveillance System. In: 2010 IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC 2010), Istanbul, Turkey (2010)
5. Valera, A., Velastin, S.A.: Intelligent distributed surveillance systems: A Review. IEEE Proc. Vision, Image, and Signal Processing 152(2), 192–204 (2005)
6. Zhang, Z., Venetianer, P.L., Litpon, A.J.: A Robust Human Detection and Tracking System Using a Human-Model-Based Camera Calibration. In: The Eighth International Workshop on Visual Surveillance (2008)
7. Sicre, R., Nicolas, H.: Human behavior recognition at a point of sale. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Chung, R., Hammound, R., Hussain, M., Kar-Han, T., Crawfis, R., Thalmann, D., Kao, D., Avila, L. (eds.) ISVC 2010. LNCS, vol. 6455, pp. 635–644. Springer, Heidelberg (2010)
8. Kim, W., Lee, J., Kim, M., Oh, D., Kim, C.: Human action recognition using ordinal measure of accumulated motion. EURASIP Journal on Advances in Signal Processing (2010)

9. Junejo, I.N., Javed, O., Shah, M.: Multi Feature Path Modeling for Video Surveillance. In: 17th Int. Conf. on Pattern Recognition (ICPR 2004), vol. 2 (2004)
10. Liu, C.: Beyond Pixels: Exploring New Representations and Applications for Motion Analysis, Doctoral Thesis. Massachusetts Institute of Technology (May 2009)
11. Moris, B.T., Trivedi, M.M.: A survey of vision-based trajectory learning and analysis for surveillance. IEEE Trans. on Circuits and Systems for Video Technology 18(8), 1114–1127 (2008)