Kinect Sensing of Shopping Related Actions

Mirela Popa[1,2], Alper Kemal Koc[1,2], Leon J.M. Rothkrantz[1,3],
Caifeng Shan[2], and Pascal Wiggers[1]

[1] Man-Machine Interaction group, Department of Mediamatics,
Delft University of Technology, Mekelweg 4, 2628 CD, Delft, The Netherlands
[2] Video and Image Processing Department, Philips Research, HTC 36,
5656 AE, Eindhoven, The Netherlands
[3] Sensor Technology, SEWACO Department, Netherlands Defence Academy,
Nieuwe Diep 8, 1781 AC, Den Helder, The Netherlands
{m.c.popa,l.j.m.rothkrantz,p.wiggers}@tudelft.nl,
{caifeng.shan}@philips.com

Abstract. Surveillance systems in shopping malls or supermarkets are usually used for detecting abnormal behavior. We used the distributed video cameras system to design digital shopping assistants which assess the behavior of customers while shopping, detect when they need assistance, and offer their support in case there is a selling opportunity. In this paper we propose a system for analyzing human behavior patterns related to products interaction, such as browse through a set of products, examine, pick products, try on, interact with the shopping cart, and look for support by waiving one hand. We used the Kinect sensor to detect the silhouettes of people and extracted discriminative features for basic action detection. Next we analyzed different classification methods, statistical and also spatio-temporal ones, which capture relations between frames, features, and basic actions. By employing feature level fusion of appearance and movement information we obtained an accuracy of 80% for the mentioned six basic actions.

Keywords: Shopping Behavior, Action Recognition, Surveillance, Kinect.

1 Introduction

In the last decade a lot of effort has been devoted to developing methods for automatic human behavior recognition, having a great potential in enhancing human-computer interaction, affective computing, and social signal processing. Especially in the context of ambient intelligence, multimodal analysis of behavior opens up new venues of applications such as behavioral biometrics, automated care for the elderly or improved customer assistance in the marketing domain.

For assisting customers, usually human shop assistants are available, but given peak hours they are too expensive to meet the whole demand. A supporting alternative can be provided by developing digital shop assistants. By using the available surveillance systems of video cameras in shops [1], we aim at assessing customers shopping behavior and detecting when there is a need for support or a

R. Wichert, K. Van Laerhoven, J. Gelissen (Eds.): AmI 2011 Workshops, CCIS 277, pp. 91–100, 2012.

selling opportunity. The semantic interpretation of the shopping behavior is based on the way of walking (trajectory analysis) and the recognition of customer-product interaction related actions. Furthermore given a representation of the shopping area into Regions of Interest (ROI) such as products, passing, pay desk, or resting areas, detecting basic actions in a specific ROI contributes to the semantic modeling of uncertainty. Interaction of customers with the environment or between each other is recorded using different sensors (video cameras, Kinect sensor). Next, data analysis is performed by extracting different types of information, trajectory and also actions related features. Automatic recognition of basic actions represents the first step towards developing a complex system designated at shopping behavior understanding, given that every behavior is composed of basic actions. In this work we considered the following actions: pick a product, examine it, browse through a set of products, try a product to see how it fits, wave for assistance and interact with the shopping basket/cart. The complexity of the analyzed problem resides in modeling human shopping behavior, due to its diversity among different individuals. But we observed that in a specific ROI, only a limited number of basic actions are being displayed. We designed an automatic system for the assessment of shopping behavior in a hierarchical manner, by employing different levels of abstraction, from low sensory level up to the reasoning level about customers' behavior. The architecture of the proposed system is presented in Fig. 1.

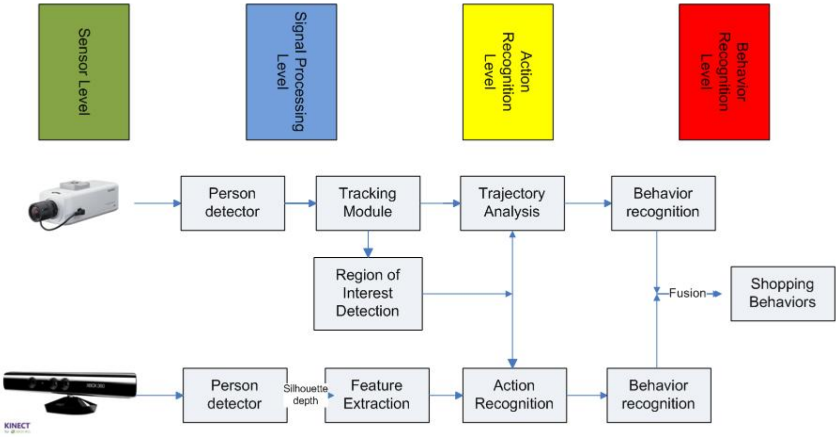


Fig. 1. System Architecture

Assessment of shopping behavior based on trajectory analysis is presented in details in [1], the discrimination between customers' buying and non-buying events is discussed in [2], while the scope of this paper consists in researching the most suitable methods towards basic actions detection. We used the Kinect Sensor developed by Microsoft [3], for recordings, due to its advantages. Regarding characteristic features for action recognition we investigated both appearance and movement features and we proposed fusing them in order to benefit of the most complete information. Different classification methods are tested for finding the most suitable one for our problem. By considering a segmentation of action data into constant size segments and using a sliding-window approach, we investigated the suitability of using our system in real-time conditions. The outline of the paper is as follows. In the next section related work is reported, then the action recognition module is presented in details, followed by the description of the data acquisition process. Next, the discussion of the experimental results is provided and finally we formulate our conclusions and give directions for future work.

2 Related Work

Action recognition based on body pose estimation can be applied in many fields such as: "surveillance, medical studies and rehabilitation, robotics, video indexing, and animation for film and games" [4]. There are two main approaches used for analyzing human motions and actions: model-based and appearance-based.

Model-based approaches employ a kinematics model for representing body parts relations with respect to each body action. In [5] Akita et al. decomposed the human body into six parts: head, torso, arms, and legs, and built a cone model with the six segments corresponding to counterparts in stick images. This type of approaches are very dependent on the reliability of body parts detection and tracking. A method with a remarkable computational performance for human body pose estimation was proposed in [6] by Shotton et al., by employing the Kinect Sensor.

Appearance-based methods build a mapping from a set of image features to an action class, without explicitly representing the kinematics of human body. A good example is template matching, which is widely used as an appearance-based action recognition algorithm. Bobick et al. introduced in [7] temporal templates constructed from Motion-Energy Images (MEI) and Motion-History Images (MHI), for the recognition of human movements. Silhouette-based features were successfully applied for action recognition, by considering either distances to the local maxima points of the silhouette contour [8], or edge features extracted from the contour of the silhouette image [9]. In [10] another method based on silhouettes is introduced for detecting the interaction of people with an object. They can detect if someone has something in his/her hand, or just left it somewhere.

It is also important which method is used for the classification task. Classification methods can be divided into two categories, one is based on the stochastic model such as Hidden Markov Models (HMMs) [11], [12], and [13] while the other one is based on a statistical model such as Support Vector Machines [14] and [15], Nearest Neighbor Classifier (NNC) [16], or Linear Discriminant Analysis [17]. Those examples show that there could be different choices for the classification method and they are all proved to be successful with a good feature set.

In [14], Schindler et al. address the problem of finding the number of frames required to perform an action or to represent an action. The proposed sequences of actions contain a number of frames from 1 to 10 and it is proven that they lead to the same performance as processing the whole video sequence.

Analyzing the presented works on action recognition provided us with an overview of the problems existing in this field and also with potential solutions. We present next our approach towards basic action detection in the shopping context.

3 Action Recognition Module

An important problem for action recognition is to find the appropriate image representation in order to extract meaningful features for action recognition. The role of image segmentation is to get such an image representation in which the salient object or person is given prominence. In video with static background, the silhouette is a good image representation, which can be obtained by background subtraction; still in many cases this option is not feasible or is not optimum due to varying lighting conditions. Therefore by employing the Kinect sensor developed by Microsoft, we are able to obtain silhouette data of each person in the scene, independent of the background or the ambient light conditions, given that the senor has been placed at a distance of at least 1 meter away from the person(s). Kinect sensor has an RGB camera and two IR cameras which enable extracting the depth map of the scene and contribute to improving the detection rate.

Next, given silhouette information, we aim at investigating the most appropriate feature descriptors for action recognition. Moment invariants seem a good approach as they are compact in description, are capable of selecting different levels of detail [18], and represent a global shape descriptor. We start from first order geometric moments, m_{10} and m_{01}, or x and y axis projections (1), computed for the image intensity function f(x, y), which in our case is a binary function, having values of 1 for the pixels belonging to the person.

$$m_{pq} = \sum_{I}\sum x^p y^q f(x,y) \text{'} \; m_{10} = \sum_{I}\sum xf(x,y) \text{'} \; m_{01} = \sum_{I}\sum yf(x,y) \tag{1}$$

An example of a silhouette image and the corresponding projections is presented in Fig. 2.

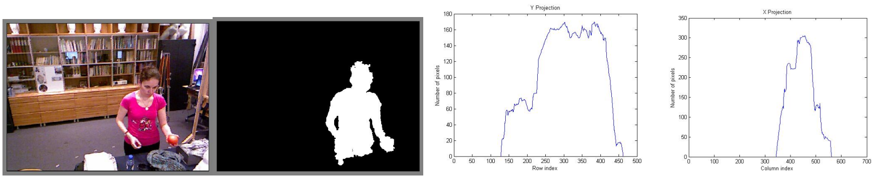


Fig. 2. (a) Original image. (b) Silhouette image. (c) Y projection. (d) X projection.

Given that the position of the person in the image frame can be changing over time, we obtain invariant features under translation, by computing central moments (2) with respect to the silhouette image centroid $(\overline{x}, \overline{y})$.

$$\overline{x} = \frac{m_{10}}{m_{00}}, \quad \overline{y} = \frac{m_{01}}{m_{00}}, \quad \mu_{pq} = \sum_{I}\sum (x-\overline{x})^p (y-\overline{y})^q f(x,y) \tag{2}$$

Next we investigated different feature sets, by considering several orders of central moments. The optimum feature set for our problem, given in (3), contains statistical measures such as the mean, the variance, the standard deviation, the skewness, and the index for peak values computed for both x and y projections. The third order central moment (μ_{30}, μ_{03}), or the skewness, characterizing the degree of asymmetry of a distribution around its mean, can be interpreted in our case as an indicator of a body limb being apart from the rest of the body.

$$\begin{gathered} featV = [mean_x, \mathrm{var}_x, \sqrt{\frac{1}{n-1}\mu_{20}}, skew_x, \max_{i=1}^{N}(x_i - \overline{x}), mean_y, \mathrm{var}_y, \sqrt{\frac{1}{m-1}\mu_{02}}, skew_y, \max_{j=1}^{M}(y_j - \overline{y})] \\ skew_x = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^3}, skew_y = \frac{\frac{1}{m}\sum_{j=1}^{m}(y_j - \overline{y})^3}{\left(\sqrt{\frac{1}{m}\sum_{j=1}^{m}(y_j - \overline{y})^2}\right)^3} \end{gathered} \tag{3}$$

Besides spatial patterns described by the featV feature set and computed on a frame basis, we also compute motion related features by employing frame differencing between two consecutive silhouette images (see an example in Fig. 3), using the intensity difference function: $m_i(x,y) = f_i(x,y)-f_{i-1}(x,y)$, i=2,n. The resulting motion image matrix has values in the (-1, 0, 1) set. Pixels with value 0 represent the constant human area, without movement, the -1 valued pixels represent the area from where the movement originated, while the pixels with value 1 are depicting the regions to which the person moved to. The extracted motion feature vector featV' is obtained by applying eq. (1), (2), and (3) to the motion image.

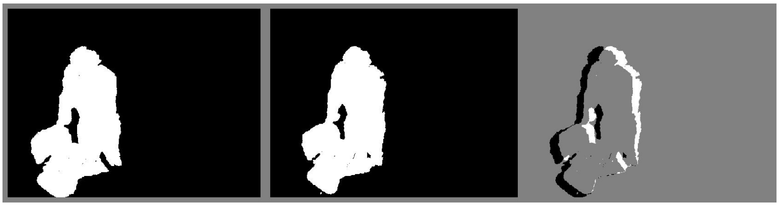


Fig. 3. (a) Silhouette frame image i (b) Silhouette frame image i+1. (c) Frame differencing

Besides silhouettes, depth information is also valuable as it contains information about the distances between the body and the limbs and the distance relative to the sensor. New feature sets are extracted from depth data in the same manner as for silhouettes, with the distinction that in the previous equations (1),(2), and (3) the binary function f(x, y) is replaced by a distance function d(x, y) representing the distance of each pixel from the Kinect sensor (see Fig. 4).

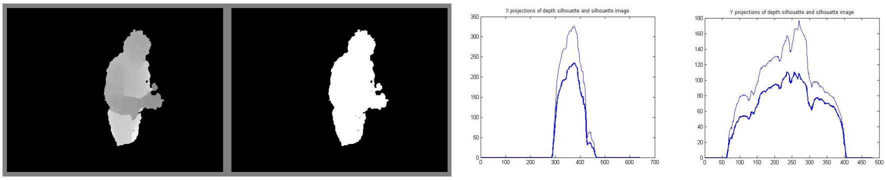


Fig. 4. (a) Silhouette frame image i. (b) Silhouette frame image i+1. (c) Depth based frame differencing on x axis. (d) Depth based frame differencing on y axis.

Finally, as appearance and movement features are complimentary, we obtain new feature sets by fusing the information coming from the two sources, both for silhouettes and depth information.

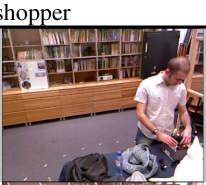
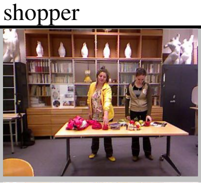
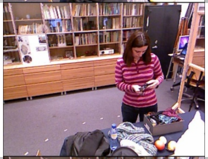
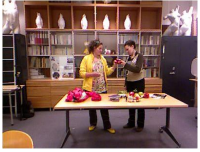
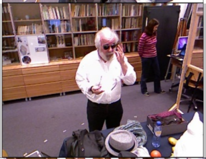
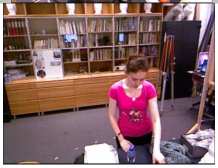
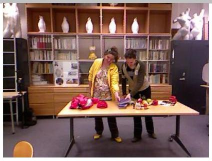
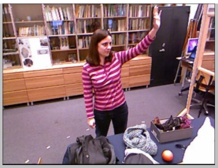
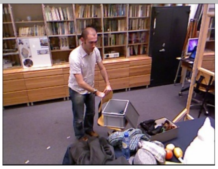
The next step, after extracting relevant feature characteristics, consists of applying different pattern recognition methods in order to discriminate between the different action classes. For the classification task we used some of the most commonly employed classifiers in literature, both statistical ones: SVM, K-NN, LDC and also stochastic ones as, namely HMMs. Due to their properties, such as incorporating dynamics of motion features during time and ability to capture temporal correlations, we use HMMs both at the action recognition level as well as for detecting sequences of actions. In the following section we present our data collection process, the performed experiments and the analysis of results.

4 Experiments

4.1 Data Acquisition

In order to define shopping behavioral models, we made observations of people during shopping in supermarkets, retail and clothes shops. From our observations we noticed the most common shopping scenarios which can be furthermore decomposed into basic shopping actions. We asked five students and researchers to play these scenarios in a controlled environment; still we gave them the freedom to show individual behavior. Each considered scenario (looking for a favorite item, examine the differences between similar products, get the attention of the shop assistant, and finally purchase the desired products) is composed of basic actions. In the definition and validation process of the proposed set of basic actions, not only the researchers but also experts in the field have been involved. In the table below an explanation and an example for each considered basic action is provided. It needs to be mentioned that we considered different view angles and also single and duo-shopper cases.

Table 1. Basic shopping actions description

Actions	Description	Key Pose(s) – single shopper	Key Pose(s) – duo shopper
Browsing	The person goes through the products on the table using the hands.		
Examining	The person holds one or more products in her/his hand, looking at them more closely.		
Trying on	The person is putting on a product, like a jacket, scarf, vest, or sun glasses etc.		
Picking	The person leans to the table, picks a product, and takes the hand back.		
Looking for help (waving one hand)	Having one or two arms in the air, the person waves at some direction, trying to get the attention of an assistant.		
Shopping Cart Interaction	The person is holding the shopping cart, drives it in shop (pushes or pulls it) and interacts with it.		

4.2 Experimental Results

For testing the proposed approach towards basic actions recognition, we used a 5-fold cross validation method and the error rates of the employed classifiers for each considered feature set are presented in Table 2. As introduced in Section 4.1 we considered 6 basic shopping actions, and the total number of samples was 94.

Appearance features extracted from silhouette data (1) seem to be performing quite good achieving an accuracy of 77,50%, better than the movement related features (2). By employing feature level fusion of both types of information, appearance and movement, we obtain an improvement in accuracy of 2,5% using the Linear Bayes Normal classifier.

Table 2. Error rates of different classifiers using the proposed feature sets

feature set	SVC	LDC	K-NN
Silhouette projections(1)	24,26	22,39	23,71
Silhouette projection differencing (2)	33,33	33,83	34,33
Fusion of (1) and (2)	23,48	**20,04**	22,56
Depth projections (3)	28,63	27,96	29,63
Depth frame differencing (4)	27,63	26,88	28,71
Fusion of (3) and (4)	26,83	24,33	25,60

Regarding depth related feature sets, the fusion of appearance and movement information proved to be beneficial, leading to the best result of 75% accuracy, still they were not performing better than the silhouette features, showing that the pose information was more discriminative than the distance to the sensor.

For a better understanding of the results, not only in terms of performance but also regarding the discriminative power of the proposed features for each particular action, we also include the confusion matrix for the best performing classifier (see Table 3).

Table 3. Confusion matrix of LDC classifier for the silhouete fusion feature set

%	Browse	Examine	Pick	Try-on	Wave	Shopp. cart
Brow.	**84,31**	7,84	5,88	0	0	1,96
Exam.	4,65	**75,58**	8,13	8,13	1,16	2,32
Pick	6,25	6,25	**75**	12,50	0	0
Try on	0	8,82	0	**82,35**	8,82	0
Wave	0	0	12,50	12,50	**75**	0
Shopp. cart	0	0	0	0	0	**100,00**

From the table above we can notice that some actions are more difficult to be classified, being easily confused with other ones, for example ‘browsing’ with ‘examining’ and ‘picking’, as they contain similar hand movements, or ‘waving’ which is also similar with ‘try on’ and ‘picking’ actions.

Next we investigated the real-time recognition of basic actions, when no indication is provided regarding the starting or ending time of one action. Using a sliding window of 0.5s (10 frames) and an overlap of 0.25s we trained prototypes for each action class. The features extracted for each segment were similar to the ones presented in Section 3, with the difference that instead of considering the total number of frames of one action, we computed them for each 10 consecutive frames. Using the best performing classifier (LDC), we tested each segment against all trained action prototypes and the one with the highest probability was selected as the segment label. Finally the accuracy was computed by dividing the ratio of correctly classified action segments over the total number of segments and we obtained an average of 75%. This experiment proved that by applying a sliding window approach, the accuracy remains high, while the continuous recognition of actions is achieved.

5 Conclusions and Future Work

In this paper we proposed an automatic system for assessing customers' shopping behavior based on action recognition. We made recordings, using the Kinect sensor, which enabled extraction of people silhouettes under different lighting conditions. Feature level fusion of appearance and movement information proved to be beneficial, achieving an average accuracy of 80% on six basic actions. By applying a sliding window of 0.5s and training prototypes for each considered action, we were able to recognize basic actions in continuous scenarios.

There are still many ways in which the proposed system can be improved. Currently we only tested our system on pre-defined scenarios in a controlled environment on a limited number of samples. Next, we plan to use interest-points models and to assess the performance of our system on real-life recordings of customers in a supermarket.

Acknowledgments. 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. 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)
2. Popa, M.C., Gritti, T., Rothkrantz, L.J.M., Shan, C., Wiggers, P.: Detecting Customers' Buying Events on a Real-Life Database. In: Real, P., Diaz-Pernil, D., Molina-Abril, H., Berciano, A., Kropatsch, W. (eds.) CAIP 2011, Part I. LNCS, vol. 6854, pp. 17–25. Springer, Heidelberg (2011)
3. Microsoft Corp. Redmond WA. Kinect for Xbox 360
4. Moeslund, T.B., Hilton, A., Kruger, V.: A Survey of Advances in Vision-based Human Motion Capture and Analysis. Computer Vision and Image Understanding (2006)
5. Akita, K.: Image sequence analysis of real world human motion. Pattern Recognition 17(1), 73–83 (1984)
6. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-Time Human Pose Recognition in Parts from Single Depth Images. In: CVPR (2011)
7. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. PAMI (2001)
8. Ekinci, M., Gedikli, E.: Silhouette Based Human Motion Detection and Analysis for Real-Time Automated Video Surveillance. Turkish Journal of Electrical Engineering and Computer Sciences 13(2), 199–230 (2005)
9. Jiang, H., Drew, M.S., Li, Z.-N.: Action Detection in Cluttered Video with Successive Convex Matching. IEEE Transactions on Circuits and Systems for Video Technology 20(1) (2010)
10. Haritaoglu, I., Cutler, R., Harwood, D., Davis, L.S.: Detection of People Carrying Objects Using Silhouettes. In: International Conference on Computer Vision, Corfu, Greece (1999)

11. Moore, D.J., Essa, I.A., Hayes, M.H.: Exploiting Human Actions and Object Context for Recognition Tasks. In: IEEE International Conference on Computer Vision, Corfu, Greece (1999)
12. Brand, M., Oliver, N., Pentland, A.: Coupled Hidden Markov Models for Complex Action Recognition. In: Proceedings of IEEE Computer Vision and Pattern Recognition (1996)
13. Yamato, J., Ohya, J., Ishii, K.: Recognizing Human Action in Time-Sequential Images using Hidden Markov Model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 379–385 (1992)
14. Schindler, K., Van Gool, L.: Action Snippets: How Many Frames Does Human Action Recognition Require? In: IEEE Computer Society Conference on Computer Vision (2007)
15. Shüldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: A Local SVM Approach. In: Proceedings of the 17th International Conference on Pattern Recognition (2004)
16. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as Space-Time Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(12) (2007)
17. Weinland, D., Ronfard, R., Boyer, E.: Free Viewpoint Action Recognition using Motion History Volumes. Computer Vision Image Understanding 104(2), 249–257 (2006)
18. Prismall, S.P.: Object reconstruction by moments extended to moving sequences, PhD thesis, Department Electronic and Computer Science, University of Southampton (2005)