Customers' Activity Recognition in Intelligent Retail Environments

Emanuele Frontoni¹, Paolo Raspa¹, Adriano Mancini¹,
Primo Zingaretti¹, and Valerio Placidi²

¹ Dipartimento di Ingegneria dell'Informazione,
Università Politecnica delle Marche
Ancona, Italy

{frontoni,raspa,mancini,zinga}@dii.univpm.it

² Grottini Lab, Porto Recanati, Italy
valerio.placidi@grottinilab.com

Abstract. This paper aims to propose a novel idea of an embedded intelligent system where low cost embedded vision systems can analyze human behaviors to obtain interactivity and statistical data, mainly devoted to customer behavior analysis. In this project we addressed the need for new services into the shop, involving consumers more directly and instigating them to increase their satisfaction and, as a consequence, their purchases. To do this, technology is very important and allows making interactions between costumers and products and between customers and the environment of the shop a rich source of marketing analysis.

We construct a novel system that uses vertical RGBD sensor for people counting and shelf interaction analysis, where the depth information is used to remove the affect of the appearance variation and to evaluate customers' activities inside the store and in front of the shelf, with products. Also group interactions are monitored and analyzed with the main goal of having a better knowledge of the customers' activities, using real data in real time.

Even if preliminary, results are convincing and most of all the general architecture is affordable in this specific application, robust, easy to install and maintain and low cost.

Keywords: RGBD, depth images, activity recognition, interactions, retail environments, people counting, shelf interaction map.

1 Introduction

The concept of shop is changing during years and today it means many different things: shop became not only the place where customers go to buy a specific product, but also the place where customers go to spend part of their time; competition between different producers is made also in the shop structure, meaning of communication, graphics, position of products, shelf availability and many other aspects [21].

This project focuses on developing automated RGBD video analysis techniques for tracking customers in retail stores with the goal of recognizing various customer activities that may be relevant for marketing analysis with a particular focus on the interactions

between shoppers and products on shelves. The approach uses low-level segmentation and tracking methods, mapping the customers to the floor plan, identifying events such as picking up products in front of store shelves, group formations by customers, etc.

The aim of every retail environment is to attract customer attention: with technology and strategy, shop's owner has to attract customer attention on particular products, giving also the chance to interact physically with technology, making the experience absolutely complete and involving. The main goal of this new concept of retail environment is simple: make more business with respect to the old shop idea.

First of all, in some cases technology works as a virtual shop assistant, giving information to the customer directly; in this way shop's owners save money with staff.

Second, and more important, is the fact that technology let to provide many statistic data that, opportunely elaborated, can give much information about customers, customizing the shop to their needs and adapting product expositions to their behaviours. Data extracted in this way are objective, while often this kind of decision is made looking to subjective data.

Several aspects of these problems are currently solved using artificial intelligence and, in particular, vision. Visual analysis in dynamic scenes is currently one of the most active research topics in computer vision [1, 5, 9].

In general, the processing framework of visual surveillance in dynamic scenes includes the following stages: modelling of environments, detection of motion, classification of moving objects, tracking, understanding and description of behaviours, human identification, and fusion of data from multiple cameras [3, 6, 7].

As a result, more and more research has been conducted in this field involving a great quantity of different vision based approaches.

Automatic recognition of human actions and interactions, however, remains a very challenging problem. The key difficulty stems from the fact that the imaged appearance of a person performing a particular action can vary significantly due to many factors such as camera viewpoint, person's clothing, occlusions, variation of body pose, object appearance and the layout of the scene.

In addition, motion cues often used to disambiguate actions in video [9, 17, 20] are not available in vertical acquired RGBD images.

In this work, we seek to recognize common human actions, such as "picking up a product", "interacting with the shelf" or "moving in a group" in challenging realistic retail images. A number of previous works [2, 4, 8, 10] focus on exploiting body pose as a cue for action recognition. In particular, several methods address joint modeling of human poses, objects and relations among them [11-16].

Reliable estimation of body configurations for people in arbitrary poses, however, remains a very challenging research problem. Less structured representations, e.g. [18, 19] have recently emerged as a promising alternative demonstrating state-of-the-art results for action recognition in static images.

In this paper we will introduce the general framework of our system and we will discuss how vision and, in particular, people and object tracking, are involved in different tasks of the whole system. We will show successful applications for retail environment management and for products management, showing the visual tracking methodology and results of real scenarios with real time performances. Software

analysis, tests and experiments developed are presented to show the feasibility of the proposed approach in large-scale applications, using only RGBD vision sensors.

The paper is more focused on the application of vision to retail environment and on the general application of visual tracking as a main information source; main novelties are in the original and wide application of the tracking system and activity analysis focused on shelf interactions. The other interesting point of the project is in the real experimental platform described in the result section and in the combination of vision based statistical approach with marketing analysis data.

The paper is organized as follows: Section 2 introduces the general concept of the system and describes the peculiarity of the particular methodology described in this paper. The experimental results are provided in Section 3. Conclusions are given in the last Section.

2 Metodology

2.1 Customers Counting and Single/Group Behaviors

The goal of the system is the use of vertical mounted RGBD camera (Fig.1a) to detect costumers activities and in particular their interactions with shelf and products. The first part of the activity recognition method is based on the water filling algorithm to identify people from a vertical depth image even when they are moving in group inside the store (e.g. a mam with child or a group of friends). This last aspect of group behaviors will be better investigated in future tests in real scenario and results on this side are not reported in this paper even if the methodology will be the same described here following.

The system needs to identify every person moving in the scene (Fig. 1b) to track his activities and in particular to analyze the interaction of this certain person with the shelf. Section 2.2 will give more details on this second aspect.

Fig. 1. Example of a user interacting with the shelf (a) and the approach for people detection and activity analysis in our test environment. Test case (a) will be described in next section.

We take the depth image as a function f , where $f(x, y)$ stands for the depth information of pixel (x, y) . Due to the noise of RGBD sensor, $f(x, y)$ can be non-derivable or even discontinuous. Finding people in depth image equals to finding local minimum regions in f . Mathematically, the problem can be defined as finding the region A and N that satisfy the following constraint:

$$E_A(f(x, y)) + \eta \leq E_{NA}(f(x, y)) \quad (1)$$

where $A \in N$, A is the local region and N is its neighborhood. $E(.)$ is an operation to pool the depth information in the region to a real value that reflects the total depth information in the region. η is a pre-defined threshold to ensure that depth in A should lower than N_A with a margin.

Note that A and N can be of arbitrary shape, and finding all the regions in image can be very time consuming. In [21] is possible to find detail about the water filling algorithm that can effectively find all the suitable regions as reported in the example of Fig.2.

Fig. 2. Example of the water filling algorithm detecting customer with ID 1 together with the shelf line estimated by the depth sensor and the reference system. This second aspect will be better described in section 2.3

2.2 Customer Interaction Aspects

The second methodological aspect of this work is the idea to have an estimation of the interaction with products on the shelf and a classification of the interaction type. In particular we define *positive* interaction when the product is picked up from the shelf and bought and *negative* one when the product is put back on the shelf after a pickup. This is very useful information when we try to identify customers' activities in front of shelves. This allows also the description of the activity also with an interaction map that will be discussed in the result section.

To solve the hand problem we designed a template matching method dealing with hand and products templates. A sample of two of these templates is reported in Fig. 3.

The template frame is identified when the hand is entering or exiting the shelf line identified by the depth sensor.

Fig. 3. Example of hand template and hand/product template, the first one entering the shelf line, the second exiting with product

In general template matching requires similarity measures between the features of a template and the query image. Image intensity patches are often compared by normalized cross-correlation whereas Hausdorff and chamfer measures are popular with edge-based features. A common approach is to have a number of prototype shape templates and search for them in the image. Templates used for hand with products and hands without were collected in real scenarios. Chamfer score and a closely related measure Hausdorff score have been used in many shape matching schemes. We use truncated chamfer score, which makes it more robust to outliers [10, 11].

3 Results

The system was installed in a real store using 5 sensors mounted in the ceiling of particular locations of the store (main entrance, high interest product categories, etc.).

The hardware was selected to respect the low cost aspect of the project and in particular the idea to use commercial components that are suitable for low maintenance and low cost requirements. The Raspberry processing unit was used to compute the RGBD video streaming at 10 frames/second using OpenCV and our software with drivers for the Asus Pro Live RGBD sensor.

The processing unit is devoted to all image-processing elaborations and only synthetic data are transmitted over the WiFi network of the store to a web server that collect numerical data in respect to privacy limitation.

Fig. 4. a: The system installed in the ceiling of a real store (5 of this unit was installed for test in particular areas of the store. b: the Raspberry used as processing unit (one for each RGBD camera with WiFi modules for synthetic data transmission) c: the Asus RGBD sensor.

The system on the web server collects several data suitable for customer activity analysis together with the interaction map of the shelf. Here following Tab. 1 presents a list of parameters that are grabbed from the real environments that was on line for 30 day. The table is related to a complete day analysis and is reported to give an idea of possible data comparisons that can be performed both between different categories (locations inside the store) and different periods.

Also change detection can be easily applied to this data to better understand what is the customers' reaction when a change is performed in the layout of the store, in the communication on the shelf or in the packaging or planogram of a certain category.

Table 1. List of parameters provided by the system for activity analysis in a day of test in a real scenario

Parameter	Value
1. Total number of people	465
2. Average attention time	36 sec
3. People passing by (no attention)	43%
4. Average group number	1,2
5. Number of shelf interactions	590
6. Interaction per person	2,1
7. Average X, Y (to evaluate attention)	134,20
8. Direction Left-Right	72%
9. Direction Right-Left	28%

Fig. 5 reports an example of the interaction map. This particular result, that is one of the novelties of this paper, is performed on the web application using data of hand interaction and their x, y, z labeled with the product/no-product features and with the id of the person that is performing the interaction. In particular the sequence of the interactions can be used to represent over the shelf image an intensity map that proves information about the interaction sequence: red areas are customers that picked up the products and put id back on the shelf; green areas represent positive interaction where customers bought the product.

Fig. 5. Example of a user interaction map. Red area represents negative interactions.

Also in this case change detection is the suggested way to perform comparisons between different planograms and packaging, or on-shelf communication.

Further results analysis are currently being developed mainly integrating data coming from 5 sensors. The system demonstrated to be affordable and robust and, until today, it collected data for 30 days with no evidence of particular errors. Even if preliminary and open to huge further investigations, results are convincing and also the discussion with marketing people involved in the project demonstrate a lot of interest for the application here described.

A subset of results (6 hours of store activities with 143 people passing by in a 1.33mt shelf) was compared with ground truth manually recorded by visual merchandising people, showing an accuracy in Table 1 data of 98,5% for parameters 1,2,8 and 9. Parameters 5 and 6 were compared with manual grabbed data with an accuracy of 97,2%.

4 Conclusion and Future Works

In this paper we presented a novel application of visual activity recognition, using a combination of techniques to better understand consumer behaviors also in the ability to interact with the environment and in particular with shelves and products; the proposed approach was tested in a real environment with interesting results in the field of retail marketing analysis.

The paper presented also an integrated architecture for mixing together different kind of vision based applications such as people and products tracking. Future work about this project is, first of all, the optimization of these systems in terms of stability, performances and robustness to environmental inconveniences, considering that in a shop there is no special worker able to reset or modify these systems in case of trouble.

The final step of this project is to provide the shop with a full automated user interaction model that can also give information about shelf refill, out of stock, crowded areas and walkability map. The goal is also to have a fast system to evaluate variations in the layout of the store, of the shelf and in on-shelf communication, to allow marketing people to work on the bases of real test with costumers in a real scenario.

Acknowledgements. Authors would like to thank all the staff of the company Grottni Lab S.r.l and Daniele Liciotti for their support and the provided test location.

References

- [1] Ascani, A., Frontoni, E., Mancini, A., Zingaretti, P.: Feature group matching for appearance-based localization. In: IEEE/RSJ, International Conference on Intelligent RObots and Systems, IROS 2008, Nice (2008)
- [2] Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV (2009)

- [3] Brox, T., Bourdev, L., Maji, S., Malik, J.: Object segmentation by alignment of poselet activations to image contours. In: CVPR (2011)
- [4] Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of key-points. In: WS-SLCV, ECCV (2004)
- [5] Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: ICCV (2009)
- [6] Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: SMiCV, CVPR (2010)
- [7] Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009)
- [8] Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE PAMI (2009)
- [9] Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Pose search: retrieving people using their pose. In: CVPR (2009)
- [10] Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE PAMI 31(10), 1775–1789 (2009)
- [11] Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR (2011)
- [12] Lampert, C., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
- [13] Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008)
- [14] Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR, pp. II: 2169–II: 2178 (2006)
- [15] Li, L., Su, H., Xing, E., Fei-Fei, L.: Object bank: A high-level image representation for scene classification and semantic feature sparsification. In: NIPS (2010)
- [16] Maji, S., Bourdev, L., Malik, J.: Action recognition from a distributed representation of pose and appearance. In: CVPR (2011)
- [17] Moeslund, T.B., Hilton, A., Kruger, V.: A survey of advances in vision-based human motion capture and analysis. CVIU 103(2-3), 90–126 (2006)
- [18] Yao, B., Fei-Fei, L.: Grouplet: A structured image representation for recognizing human and object interactions. In: CVPR (2010)
- [19] Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR (2010)
- [20] Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV 73(2), 213–238 (2007)
- [21] Mancini, A., Frontoni, E., Zingaretti, P., Placidi, V.: Smart vision system for shelf analysis in intelligent retail environments. In: ASME/IEEE International Conference on Mechatronic and Embedded Systems and Applications (MESA 2013), Portland, Oregon (2013)