Pattern Recognition Letters 112 (2018) 83–90

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Market basket analysis from egocentric videos

Vito Santarcangelo[a,b], Giovanni Maria Farinella[b], Antonino Furnari[b,*], Sebastiano Battiato[b]

[a] Centro Studi S.r.l., Zona Industriale, Buccino, Italy
[b] Department of Mathematics and Computer Science, University of Catania, Italy

ARTICLE INFO

Article history:
Received 19 January 2018
Available online 8 June 2018

MSC:
41A05
41A10
65D05
65D17

Keywords:
Market basket analysis
Egocentric vision
Multimodal analysis

ABSTRACT

This paper presents Visual Market Basket Analysis (VMBA), a novel application domain for egocentric vision systems. The final goal of VMBA is to infer the behavior of the customers of a store during their shopping. The analysis relies on image sequences acquired by cameras mounted on shopping carts. The inferred behaviors can be coupled with classic Market Basket Analysis information (i.e., receipts) to help retailers to improve the management of spaces and marketing strategies. To set up the challenge, we collected a new dataset of egocentric videos during real shopping sessions in a retail store. Video frames have been labeled according to a proposed hierarchy of 14 different customer behaviors from the beginning (cart picking) to the end (cart releasing) of their shopping. We benchmark different representation and classification techniques and propose a multi-modal method which exploits visual, motion and audio descriptors to perform classification with the Directed Acyclic Graph SVM learning architecture. Experiments highlight that employing multimodal representations and explicitly addressing the task in a hierarchical way is beneficial. The devised approach based on Deep Features achieves an accuracy of more than 87% over the 14 classes of the considered dataset.

1. Introduction

Egocentric Vision is an emerging research area characterized by a number of challenges and different application domains. By exploiting the egocentric point of view, it is possible to gather hours of videos which can be processed to obtain logs of the monitored scenarios [1]. Different studies on egocentric vision have been recently published. The main goals considered in this area are related to location recognition [2], motion understanding [3], objects and actions recognition [4–6], 3D reconstruction [7,8] and summarization [9,10]. Temporal segmentation of Egocentric Vision is also fundamental to understand the behavior of the users wearing the camera [3].

Recently, the scenario of retail stores has become of particular interest and applications related to user localization and reconstruction of spaces have been investigated. Specifically, in [11], it is proposed to localize the person wearing a camera in very large indoor spaces (shopping malls with over 200 stores) by considering only a single image and a floor plan of the environment as input and by exploiting also text detection as a useful visual cue. An interesting problem in the context of retail stores concerns the monitoring of customers' paths, thereby enabling the analysis of their behaviors. Monitoring of customers' behavior is usually obtained by employing loyalty cards, counting devices connected with Bluetooth and WiFi systems [12,13], employing RFID tags [14], as well as using fixed cameras [15,16]. However, the use of wireless technologies in the retail domain shows a lot of problems due to the presence of water and liquids. In fact, wireless communications are hindered by liquids, even though these infrastructures are also expensive.

Differently from the classic approaches mentioned above, in this paper we consider to turn an ordinary cart into a "narrative shopping cart" by equipping it with a camera. The acquired egocentric videos are processed with algorithms able to turn the visual paths into customers' behaviors. It is so possible to infer the overall route travelled by shopping carts (and hence by the customers), from the cart picking to its release. Visual and audio data can be collected and processed to monitor pauses, understanding areas of interest for customers, estimating the path speed, giving alert due to busy areas of the sale point, as well as manage the inefficiencies (e.g., slowness at cash desks). We refer to the proposed behavioral monitoring in a retail store as Visual Market Basket Analysis (VMBA). This kind of analysis can be useful to enrich the classic Market Basket Analysis methods used to infer the habits of customers [17]. The research presented in this paper is part of an experimental program carried out by Centro Studi S.r.l. to infer the behavior of customers in real retail stores using egocentric shopping carts.

* Corresponding author.
E-mail address: furnari@dmi.unict.it (A. Furnari).

https://doi.org/10.1016/j.patrec.2018.06.010
0167-8655/

The main contributions of this paper are the following: (1) we present the novel problem of Visual Market Basket Analysis and propose a hierarchy of 14 behaviors to be analysed from egocentric videos; (2) we introduce a novel dataset of 15 videos acquired during real shopping sessions; (3) we investigate a multimodal classification method which combines visual features such as (GIST [18,19], Deep Features [20]), audio (MFCC [21]) and motion cues (Optical Flow [22]) within the framework of Direct Acyclic Graph SVM [23].

Experimental results highlight that (1) combining features arising from different sources (e.g., images, audio and motion) is useful to improve the performance of the overall system and (2) explicitly considering the hierarchical structure of classes is beneficial in the considered context.

2. Related work

Data mining techniques have been widely employed in the context of marketing to analyse customer behaviors and to support business decisions [24]. Classic methods to infer the customer's behavior include tracking their locations by employing dedicated hardware. For instance, [12], proposed a tracking system capable of localizing customers by means of several Ultra-Wide Band antennas positioned in the store and battery-powered tags placed on the charts. Pierdicca et al. [13] designed a low-cost system for indoor localization based on wireless embedded systems.

Other works have focused on computer vision solutions ad-hoc developed. Liciotti et al. [16] presented an integrated system employing RGB-D cameras to monitor shoppers in a retail environment. The system is able to infer the shopper's behavior and his interaction with products (e.g., product pick up). Del Pizzo et al. [15] designed a method to count people from ceiling mounted cameras (either RGB-D or traditional RGB sensors). Other authors exploited the use of text and images to localize customers in large shopping malls [11], perform fine-grained classification [25] and recognizing business places [26].

In contrast with the considered approaches, we propose to analyse egocentric images acquired by a shopping chart to aid Market Basket Analysis. The proposed method has the advantage to collect information from the point of view of the customer and hence can provide a useful picture of his behavior during the shopping activity.

3. Visual Market Basket Analysis

We introduce the problem of Visual Market Basket Analysis (VMBA) considering three different high-level behavioral categories related to the customers carrying the shopping cart : action (i.e., *stop vs moving*), location (i.e., *indoor vs outdoor*), and scene context (i.e., *cash desk, retail, pasta, fruit, gastronomy, parking, road*). Each of the three considered high-level categories provides a meaningful source of information for the VMBA problem. For instance, knowing that a cart is stopping, rather than moving in a particular area of the store, can provide insights on whether the customer is experiencing difficulties in locating the desired products. Likewise, being able to understand when the cart is inside or outside the store and, ultimately, in which part of the store it is located, allows to obtain real-time information on the distribution of customers in the store.

We propose to organize these categories hierarchically, as shown in the tree depicted in Fig. 1. Each of the paths from the root to a leaf, identifies one of 14 different behavioral classes. Given a frame of the video acquired by the camera mounted on the cart, we aim to infer a triplet identifying the path of the tree corresponding to the observed behavior (e.g., [MOVING, INDOOR, PASTA] in Fig. 1). The classification of each frame with respect to the proposed 14 classes allows to analyse the customer's behavior and can be also useful to understand if there are problems on strategic issues to be fixed by the management.

Fig. 2 shows some examples of egocentric images acquired with a shopping cart together with the temporal segmentation with respect to the 14 considered classes. From the segmented egocentric video it is possible to understand how much time customers spend at the cash desk by considering all the frames classified with the triplets [STOP, INDOOR, CASH DESK] and [MOVING, INDOOR, CASH DESK]. This may be useful, for example, to plan the opening of more cash desks in order to provide a better service to the customers. By analysing the inferred triplets, it is also possible to understand if there are carts outside the cart parking spaces in order to take appropriate actions (e.g., if a given cart is associated to the triplet [STOP, OUTDOOR, ROAD] for long time). Combining the customers' receipts with information arising from temporally segmented videos and algorithms for Visual Market Basket Analysis' re-localization [27] could help infer the order according to which products have been taken during the shopping (Fig. 2), hence increasing the amount of information usually exploited by the classic Market Basket Analysis' algorithms [17]. This opens also new research perspectives in the context of egocentric vision.

4. Proposed method

Our aim is to temporally segment the acquired egocentric videos into chapters. To this aim, each frame should be automatically labeled according to one of the considered 14 behavioral classes. In this context, a chapter is a set of consecutive frames which present the same behavior, e.g., a sequence of frames with the same label [MOVING, INDOOR, PASTA]. We propose to explicitly consider the hierarchy illustrated in Fig. 1 to classify each frame. The first level of the hierarchy is related to the action of moving or stopping, which reflect the basic actions of a customer in the retail store. The second layer of the hierarchy identifies the high level locations where the customer is moving, i.e., indoor or outdoor. The third level considers the scene-context in which the user is located during the shopping. The final classification with respect to the 14 classes can hence be obtained by considering the three classification problems and predicting a triplet [*Action, Location, Scene Context*]. To perform each of the three level classification tasks, two main components are needed: a suitable representation and a classification algorithm. In the following, we describe the representations used for the three different levels of the hierarchy in Fig. 1, as well as the classification approaches exploited in the proposed study.

4.1. Representation at the first level: actions

The first layer of the hierarchy analyses the user behavior from the point of view of the motion of the cart. To this aim we consider two possible motion classes: *stop* and *moving*. In order to infer the motion state, we considered the Mel Frequency Cepstral Coefficients (MFCC) audio features [28] and the optical flow features extracted with the classic block matching method [29]. To extract the optical flow we have divided the frames into nine blocks. Hence, we extract a 9-dimensional descriptor related to the motion of each block of a frame. The audio processing produces a 62-dimensional MFCC feature vector. We have exploited audio because there is a "visual" correlation between the audio waveform and the motion and locations of the shopping cart. The use of the optical flow features is straightforward considering the problem to be solved (i.e., *stop vs moving*). In the experiments we have compared the two considered features both separately and jointly.

4.2. Representation at the second level: location

The second layer of the hierarchy aims to identify the general location of the user: *indoor* vs *outdoor*. At this level we consider both visual and audio features. Concerning visual features, we consider the popular GIST descriptor proposed in [19] and Deep Features [20]. The GIST is able to encode the visual scene with a feature vector of 512 components. The Deep Features are obtained by the FC7 layer of Alexnet CNN architecture trained on ImageNet [21], a feature vector of 4096 components. We also consider the MFCC feature representation after visual inspection of the audio waveform. Indeed, the waveform is more pronounced in the outdoor environment than in the indoor location. In our experiments we consider audio and visual features both independently and in combination.

4.3. Representation at the third level: scene context

The third layer of the hierarchy is related to the analysis of the context in which the shopping cart is located. Among the considered seven classes, i.e., *cash desk, retail, pasta, gastronomy, fruit, parking and road*, the first five are related to the indoor environment, whereas the other two are related to the outdoor location. Also for this level of description we have considered the GIST features, given their property of capturing the "shape of the scene" for context discrimination [19]. As for the previous layer, we have also tested deep features [20].

4.4. Classification methods

After representing a frame of the egocentric video as described in previous sections, a classifier has to be used in order to infer one of the 14 behavioral classes. To benchmark the VMBA problem and assess the impact of the proposed hierarchical organization shown in Fig. 1, the following different classification modalities have been considered:

- combination of the results obtained by three different SVM classifiers trained to perform classification with respect to each of the levels of the tree shown in Fig. 1;
- A single SVM classifier trained to discriminate among the 14 considered classes;
- a Direct Acyclic Graph SVM learning architecture (DAGSVM) [23] which reflects the hierarchy in Fig. 1 on each node.

5. VMBA15 dataset

We acquired a dataset of 15 different egocentric videos during real shopping sessions in a retail store. The dataset is referred to as VMBA15. All videos have been acquired using a narrative cam[1] mounted in the front part of the shopping cart. Fig. 3 shows some samples extracted from the dataset. The duration of each video is comprised between 3 and 20 min and has a resolution of 640 × 480 pixels. Audio has been also recorded since it can be useful to discriminate indoor vs outdoor environments. Each video has been manually labeled at 1 fps according to the 14 different behavioral classes arising from the possible paths root-leave of the hierarchy shown in Fig. 1. This implies labeling each frame according to action (i.e., "stop" or "moving"), location (i.e., "indoor" or "outdoor") and scene context (i.e., "cash desk", "retail", "pasta", "fruit", "gastronomy", "parking" and "road"). The dataset contains a total of 7839 labeled samples. Table 1 reports the number of samples for each of the 8 scenes contained in the egocentric videos, while Table 2 reports the number of samples for each of the 14 considered class. Each class C_i is related to a path in the tree of Fig. 1. Specifically: C1 [STOP INDOOR Cash Desk], C2 [STOP INDOOR Retail], C3 [STOP INDOOR Cash Desk], C4 [STOP INDOOR Fruit], C5 [STOP INDOOR Gastronomy], C6 [STOP OUTDOOR Parking], C7 [STOP OUTDOOR Road], C8 [MOVING INDOOR Cash Desk], C9 [MOVING INDOOR Retail], C10 [MOVING INDOOR Cash Desk], C11 [MOVING INDOOR Fruit], C12 [MOVING INDOOR Gastronomy], C13 [MOVING OUTDOOR Parking], C14 [MOVING OUTDOOR Road]. The labeled dataset is available for research purposes at http://iplab.dmi.unict.it/vmba15.

[1] www.vehomuvi.com.

Table 1
Number of samples per context label for each egocentric video.

VID	Park	Road	Cash desk	Retail	Gastronomy	Fruit	Pasta	Total
1	23	89	8	117	43	13	29	322
2	10	84	9	209	16	32	30	390
3	10	96	12	163	36	27	23	367
4	13	106	10	156	20	35	3	343
5	9	107	10	159	24	32	4	345
6	39	93	128	123	4	6	35	428
7	19	119	7	81	3	4	27	241
8	20	75	14	210	14	7	33	373
9	22	85	160	646	5	7	73	998
10	13	75	7	48	48	14	5	210
11	41	89	52	355	28	19	100	684
12	27	104	26	376	38	37	39	647
13	51	137	63	129	34	12	79	495
14	25	46	53	832	4	6	32	998
15	6	73	84	761	3	61	10	998

Table 2
Number of samples per class for each video. Each class C_i is related to a path in the tree of Fig. 1. See text for the details.

VID	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
1	0	3	2	6	18	6	0	8	114	27	7	25	17	89
2	0	15	4	2	0	0	0	9	194	26	30	16	10	84
3	0	4	4	3	12	0	0	12	159	19	24	24	10	96
4	0	15	3	0	5	0	0	10	141	0	35	15	13	106
5	0	12	4	2	4	0	0	10	147	0	30	20	9	107
6	69	20	28	3	2	4	2	59	103	7	3	2	35	91
7	0	11	13	0	0	3	0	7	70	14	4	3	16	119
8	6	20	7	2	4	12	0	8	190	26	5	10	8	75
9	142	150	8	2	2	13	0	18	496	65	5	3	9	85
10	0	5	3	2	2	0	0	7	43	2	12	46	13	75
11	42	79	31	7	2	25	0	10	276	69	12	26	16	89
12	0	24	22	4	3	4	0	26	352	17	33	35	23	104
13	56	41	57	6	10	23	4	7	88	22	6	24	28	133
14	50	380	7	2	2	14	0	3	452	25	4	2	11	46
15	81	482	0	18	1	3	27	3	279	10	43	2	3	46

6. Experimental settings and results

The experiments have been performed splitting the dataset randomly in three parts, each composed of five egocentric videos. All experiments are repeated three times using two parts (10 videos) for training and one part (5 videos) for testing. The reported results are obtained by averaging over the three runs. We begin our analysis comparing the investigated representations when used to solve independently one of the three considered classification tasks. This test has been performed by exploiting a SVM classifier with an RBF kernel for each level separately. This experiment is useful to determine the best representation (or a combination of them) to be employed at each level of the hierarchy.

6.1. First level: actions

Table 3 reports the results of the *stop vs moving* classification (i.e., first level). Both audio and visual features obtain good performance. However, the optical flow features outperform audio fea-

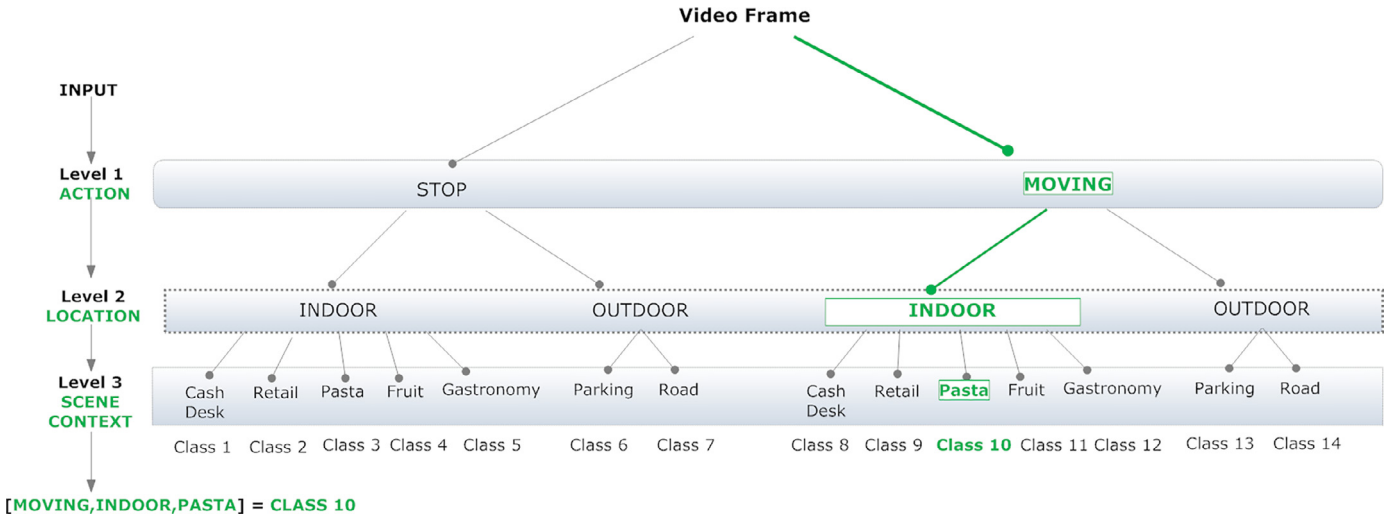

Fig. 1. Visual Market Basket Analysis (VMBA) behavioral classes organized in a hierarchy. Best seen in digital version.

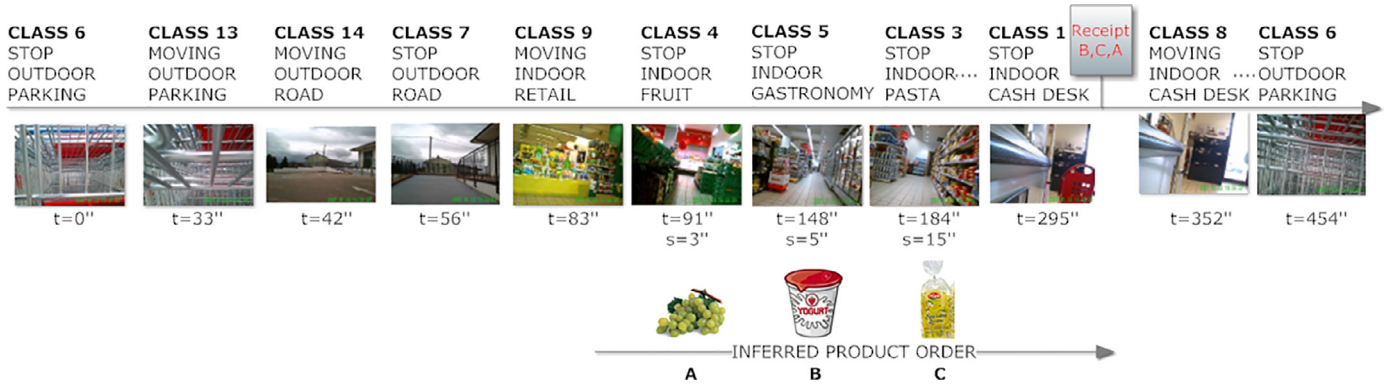

Fig. 2. VMBA timeline temporally segmented considering the 14 classes. *t* denotes the time, whereas *s* denotes the stopping time.

Fig. 3. Some frames extracted from the egocentric videos of the proposed VMBA15 dataset. Each frame is related to one of the 14 considered classes. Images are presented in the same order as Fig. 1 from left to right and from top to bottom. Notice that some classes are characterized by similar visual content but different actions (frames in the top row are related to the "stop" categories, whereas frames in the second row are related to the "moving" categories).

Table 3
Results for stop vs moving classification. Per-row best results are reported in bold numbers.

	Flow	MFCC	Combined
Accuracy%	92.50	87.04	**94.50**
TP rate%	73.03	61.54	**84.76**
TN rate%	**99.18**	95.21	97.65
FP rate%	**0.82**	4.79	2.35
FN rate%	26.97	38.46	**15.24**

tures with a margin of about 5%. The combination of audio and optical flow features allows to increase accuracy to the value of 94.50%. The obtained results point out that the combination of audio and visual features are the best suited for the first level.

6.2. Second level: location

Table 4 reports the results of the *indoor vs outdoor* classification (second layer). In this case, deep features outperform both GIST and audio features with a good margin obtaining an accuracy of 97.26%. Hence, we employ the Deep Features descriptor alone in the second level.

6.3. Third level: scene-context

For the *scene-context* classification (third level), we obtain an accuracy of 85.05% using the GIST descriptor and 90.34% with the use of Deep Features descriptor. Note that, in this case, a multi-class SVM with RBF kernel has been trained to discriminate be-

Table 4
Results for indoor vs outdoor classification. Per-row best results are reported in bold numbers.

	GIST	MFCC	DEEP
Accuracy%	95.79	88.00	**97.26**
TP rate%	89.3	49.51	**94.34**
TN rate%	97.8	97.66	**98.33**
FP rate%	**2.20**	2.34	1.66
FN rate%	10.7	50.49	**5.64**

Table 5
Confusion matrix for scene context classification with the GIST descriptor.

	Predicted						
	Parking	Road	Cash desk	Retail	Gastronomy	Fruit	Pasta
Parking	**54.25**	22.88	1.96	20.26	0.00	0.00	0.65
Road	7.25	**84.46**	6.57	0.99	0.54	0.00	0.19
Cash desk	1.23	13.35	**78.67**	6.75	0.00	0.00	0.00
Retail	1.82	0.02	1.19	**92.46**	1.72	0.80	1.99
Gastronomy	0.00	3.89	0.00	27.16	**68.95**	0.00	0.00
Fruit	0.00	0.00	0.00	37.10	0.00	**62.90**	0.00
Pasta	0.56	0.56	0.00	28.65	0.00	0.00	**70.23**

Table 6
Confusion matrix for scene context classification with Deep Features.

	Predicted						
	Parking	Road	Cash desk	Retail	Gastronomy	Fruit	Pasta
Parking	**85.51**	1.87	0.0	7.94	0.0	0.0	0.0
Road	0.38	**94.08**	2.86	1.34	0.0	0.0	0.0
Cash desk	0.0	12.59	**68.53**	17.83	0.35	0.0	0.0
Retail	0.0	0.12	0.32	**98.2**	0.28	0.2	0.84
Gastronomy	0.0	2.6	0.0	38.31	**59.09**	0.0	0.0
Fruit	0.0	0.67	0.0	35.57	0.00	**61.74**	2.01
Pasta	0.0	0.0	0.0	35.78	0.0	0.0	**64.22**

Table 7
Results of the three considered classification approaches.

	Combination	Multi-Class SVM	DAGSVM
Accuracy%	80.10	67.50	87.71

Table 8
Confusion Matrix of the DAGSVM approach.

Classes	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
C1	**63.4**	14.5	0	0	0.21	0	9.8	4.9	3.4	0	0	0.1	0	3.7
C2	0.25	**93.6**	0.7	0.1	0.23	0	0.1	0.1	4.6	0.2	0.1	0.1	0	0.1
C3	0	34.5	**61.2**	0	0	0	0	0	1.3	3.1	0	0	0	0
C4	0	32.1	1.8	**62.9**	0	0	0.62	0	1.4	0.1	0.9	0	0	0.1
C5	0	34.2	0	0	**59.7**	0	2.1	0	2.1	0	0	1.2	0	0.7
C6	0	8.5	0	0	0	**83.1**	2.4	0	2.8	0.4	0	0	2.5	0.2
C7	1.7	1.1	0	0	0	0.4	**93.8**	0.3	0.1	0	0	0	0	2.6
C8	7.6	2.1	0	0	0.1	0	1.9	**61.2**	14.6	0	0	0.3	0	12.2
C9	0.1	3.1	0.1	0.1	0	0	0	0.2	**95.4**	0.6	0.1	0.3	0	0.1
C10	0	6.1	3.2	0	0	0	0	0	30.2	**60.3**	0	0	0	0
C11	0	3.2	0.3	1.7	0	0	0.1	0	32.7	1.9	**59.6**	0	0	0.4
C12	0	0.9	0	0	1.3	0	0.5	0	37.5	0	0	**58.1**	0	2.1
C13	0	2.3	0	0	0	3.6	0.3	0	9.2	0	0	0	**82.5**	1.9
C14	0.5	0.8	0	0	0	0	2.3	2.1	1.9	0	1.3	0	0.3	**90.8**

tween the seven possible scene contexts without using priors given by the previous level in the hierarchy (i.e., indoor vs outdoor). We also report the confusion matrices with respect to the considered seven scene contexts in Tables 5 and 6. When using GIST descriptors, the main classification errors are related to the confusion between the "parking", "road" and "retail" classes (first row of Table 5). The confusion between parking and retail classes is probably due to the encoding of the scene information by the GIST descriptor. Indeed, when the cart is in the parking space, the scene is mainly composed by vertical and horizontal edges which can be confused with the vertical and horizontal edges of some scenes in the retail (see Fig. 4). As it is shown in Table 6, As it will discussed in the experiments, explicitly enforcing a hierarchy using a DAGSVM classifier helps reducing ambiguities by enforcing a prior on the main location (*indoor vs outdoor*). As it is shown in Table 6, results improve using Deep Features descriptor. For instance, the classification errors for the "parking" and "road" classes are reduced. Misclassification between the "retail" classes are mainly due to the visual similarity of different retail sectors in the frames. Other classification problems are due to strong occlusions caused by people in the scene as shown in the examples in Fig. 5.

Fig. 4. On the left a typical observation of the narrative cart when in the parking space. On the right an example of a frame of indoor acquired by the narrative cart in retail (labeled as pasta). The distribution of vertical and horizontal edges generates confusion in the classification when using GIST features.

Fig. 5. Some examples of frames with occlusions.

6.4. Overall classification

The experiments presented in the previous sections pointed out that the best features to be employed in the first level of the hierarchy are the combination of MFCC and FLOW, whereas the DEEP Features descriptor can be employed in the second and third levels. Since the main goal is the classification with respect to the 14 possible triplets corresponding to the leaves of the tree in Fig. 1, after selecting the features for the three levels independently, we have compared the three different classification modalities discussed in Section 4.4, namely: (1) combination of the results of three different SVM classifiers trained to address classification at each of the three considered levels; (2) a multiclass SVM classifier which discards the proposed hierarchy and classifies samples into the 14 possible classes; a DAGSVM [23] classifier which reflects the hierarchy proposed in Fig. 1 to perform classification.

The results of the three different approaches are reported in Table 7. Best results are obtained by the DAGSVM approach, which obtains an accuracy of 87.71%. It is worth noting that the simple concatenation of MFCC features with the FLOW and DEEP descriptors does not allow the multi-class SVM to reach the best accuracy (67.50%). The proposed DAGSVM approach also outperforms the concatenation of three different classifiers trained separately at each level (80.10%). Some visual examples for the assessment of the output given by the proposed DAGSVM-based approach are available in Fig. 6.

Table 8 reports the confusion matrix related to the performances of the DAGSVM. Class C1 [STOP, INDOOR, CASH DESK] obtains a True Positive Rate (TPR) of 63,4%, with the largest misclassification obtained on the C2 class [STOP, INDOOR, RETAIL]. The best performances are obtained with the C9 class [MOVING, IN-

Table 9
Number of Shopping Charts predicted for each Class at a given time. Ground Truth Predictions are reported in parenthesis. Reported times are in MM:SS format.

Time	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
00:00	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	14 (14)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
03:10	0 (0)	2 (2)	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	6 (5)	5 (6)	0 (0)	0 (0)	0 (0)	1 (1)
06:30	0 (0)	2 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	4 (3)	0 (0)	0 (0)	0 (0)	1 (2)	2 (2)
09:50	0 (1)	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)	0 (0)	3 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
13:10	0 (1)	3 (2)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
16:30	1 (2)	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (1)

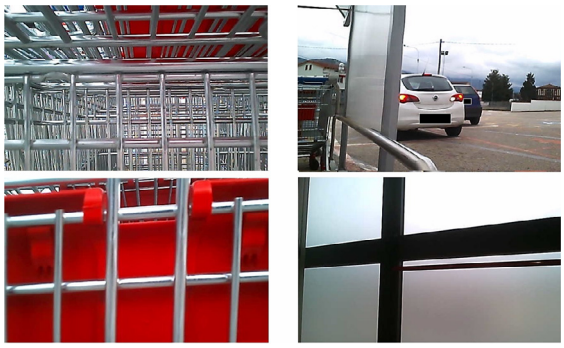

Fig. 6. Examples of frames correctly classified by the proposed DAGSVM approach but misclassified by the other two compared approaches. The frames on the left columns are related to the parking space of the carts, but are recognized as retail by the combined approach. The frames on the right are related to indoor, but are recognized as outdoor by the combined approach.

DOOR, RETAIL]. C3 class [STOP, INDOOR, PASTA] is misclassified with the generic C2 class that represents the generic retail environment. Similar performances are obtained for the classes C4 and C5 related to the fruit and gastronomy departments. For the class C6 [STOP, OUTDOOR, PARKING], the performance obtained is of 83,10%, due to the misclassification with retail scenarios (the cart grid is mistaken for windows and shelves). Class C7 [STOP, OUTDOOR, ROAD] is classified with 93.8% of accuracy. Considering the results, the proposed approach can be used for rough localization of customers in the sale point.

To demonstrate the system, in Table 9, we report some sample predictions at different times in a retail store. The system analyses 15 input videos and classifies them at different instants to infer their behavior. Specifically, each row reports the number of predicted charts belonging to a given category. Ground truth predictions are reported in parenthesis. A video demo of the proposed method is also available at our web page: http://iplab.dmi.unict.it/vmba15.

7. Conclusion

This paper has introduced the problem of Visual Market Basket Analysis (VMBA). To set the first VMBA challenge, a new egocentric video dataset (VBMA15) has been acquired in a retail store with cameras mounted on shopping carts. The VBMA15 dataset has been labeled considering 14 different classes arising from a hierarchical organization of *Actions, Location and Scene Contexts*. A first benchmark has been performed considering classic representations and classification modalities. Experiments pointed out that audio, motion and global visual features are all useful in the VMBA application domain when coupled with a Direct Acyclic Graph based SVM leaning architecture.

Future works will investigate the design of a framework based on deep learning trainable in an end-to-end fashion to address Market Basket Analysis from Egocentric Videos processing and fusing information coming from the three levels. Moreover, the analysis will be extended to data collected in more retail stores.

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. We thank Centro Study S.r.l. for the grant of two Ph.D. scholarship and for their support with the collection of the dataset used in this work.

References

[1] A. Ortis, G.M. Farinella, V. DAmico, L. Addesso, G. Torrisi, S. Battiato, Organizing egocentric videos of daily living activities, Pattern Recognit. 72 (Supplement C) (2017) 207–218.
[2] A. Furnari, G.M. Farinella, S. Battiato, Recognizing personal locations from egocentric videos, IEEE Trans. Hum. Mach. Syst. 47 (1) (2017) 6–18.
[3] V. Poleg, S. Peleg, Temporal segmentation of egocentric videos, Int. Conf. Comput. Vis. Pattern Recognit. (2014).
[4] D. Damen, T. Leelasawassuk, W. Mayol-Cuevas, You-do, i-learn: discovering task relevant objects and their modes of interaction from multi-user egocentric video, Br. Mach. Vis. Conf. (2014).
[5] A. Fathi, J. Rehg, Modeling actions through state changes, Int. Conf. Pattern Recognit. (2013).
[6] A. Fathi, X. Ren, J. Rehg, Learning to recognize objects in egocentric activities, Int. Conf. Comput. Vis. Pattern Recognit. (2011).
[7] Y. Lee, J. Ghosh, K. Grauman, Discovering important people and objects for egocentric video summarization, Int. Conf. Comput. Vis. Pattern Recognit. (2012).
[8] P. Poleg, T. Halperin, C. Arora, S. Peleg, Egosampling: fast-forward and stereo for egocentric videos, Int. Conf. Comput. Vis. Pattern Recognit. (2015).
[9] Z. Lu, K. Grauman, Story-driven summarization for egocentric video, Int. Conf. Comput. Vis. Pattern Recognit. (2013).
[10] B. Xiong, G. Kim, L. Sigal, Storyline representation of egocentric videos with an application to story-based search, Int. Conf. Comput. Vis. (2015).
[11] S. Wang, S. Fidler, R. Urtasun, Lost shopping! monocular localization in large indoor spaces, Int. Conf. Comput. Vis. (2015).
[12] M. Contigiani, R. Pietrini, A. Mancini, P. Zingaretti, Implementation of a tracking system based on uwb technology in a retail environment, in: Mechatronic and Embedded Systems and Applications (MESA), 2016 12th IEEE/ASME International Conference on, IEEE, 2016, pp. 1–6.
[13] R. Pierdicca, D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini, P. Zingaretti, Low cost embedded system for increasing retail environment intelligence, in: Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on, IEEE, 2015, pp. 1–6.
[14] Z. Ali, R. Sonkusare, RFID Based smart shopping and billing, Int. J. Adv. Res. Comput. Commun. Eng. Vol. 2 (12) (2013).
[15] L. Del Pizzo, P. Foggia, A. Greco, G. Percannella, M. Vento, A versatile and effective method for counting people on either RGB or depth overhead cameras, in: Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on, IEEE, 2015, pp. 1–6.
[16] D. Liciotti, M. Contigiani, E. Frontoni, A. Mancini, P. Zingaretti, V. Placidi, Shopper analytics: A customer activity recognition system using a distributed RGB-D camera network, in: International Workshop on Video Analytics for Audience Measurement in Retail and Digital Signage, Springer, 2014, pp. 146–157.
[17] P. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison-Wesley Companion Book Site, 2006.
[18] G.M. Farinella, D. Ravi, V. Tomaselli, M. Guarnera, S. Battiato, Representing scenes for real-time context classification on mobile devices, Pattern Recognit. (4) (2015).
[19] A. Oliva, A. Torralba, Modelling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.
[20] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012).
[21] L. Muda, M. Begam, I. Elamvazuthi, Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques, J. Comput. (3) (March 2010).

[22] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, Flownet: learning optical flow with convolutional networks, in: International Conference on Computer Vision, 2015.
[23] J. Platt, N. Cristianini, J. Shawe-Taylor, Large Margin Dags for Multiclass Classification, MIT Press, 2000, pp. 547–553,.
[24] M.J. Berry, G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Support, John Wiley & Sons, Inc., New York, NY, USA, 1997.
[25] X. Bai, M. Yang, P. Lyu, Y. Xu, J. Luo, Integrating scene text and visual appearance for fine-grained image classification, arXiv:1704.04613, (2017).
[26] S. Karaoglu, R. Tao, T. Gevers, A.W. Smeulders, Words matter: scene text for image classification and retrieval, IEEE Trans. Multimedia 19 (5) (2017) 1063–1076.
[27] A. Kendall, M. Grimes, R. Cipolla, Posenet: a convolutional network for real–time 6-dof camera relocalization, Int. Conf. Comput. Vis. (2015).
[28] M. Sahidullah, G. Saha, Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition, Speech Commun. (2012) 543–565.
[29] J. Barron, S. Beauchemin, Performance of optical flow techniques, Int. J. Comput. Vis. 12 (1994) 43-77. issue

Vito Santarcangelo Master Computer Science Engineer (cum laude), since 2015 is a Computer Science Ph.D. candidate at the University of Catania under the advisorship of Prof. Sebastiano Battiato and Dr. Giovanni Maria Farinella. His main research interests concern Computer Vision for First Person Vision, DOOH applications and Computer Security. He is an Applied Research Engineer of Centro Studi (Orizzonti Holding Group) and CEO of the iInformatica company. He is author of 8 granted patents. He is also Lead Auditor ISO 27001:2013. In 2016 and 2017 he has been Technical Program Committee of the conference IARIA SecurWare.

Giovanni Maria Farinella is Assistant Professor at the Department of Mathematics and Computer Science, University of Catania, Italy. He received the (egregia cum laude) Master of Science degree in Computer Science from the University of Catania in April 2004. He was awarded the Ph.D. in Computer Science from the University of Catania in October 2008. From 2008 he serves as Professor of Computer Science for undergraduate courses at the University of Catania. He is also an Adjunct Professor at the School of the Art of Catania in the field of Computer Vision for Artists and Designers (Since 2004). From 2007 he is a research member of the Joint Laboratory STMicroelectronics - University of Catania, Italy. His research interests lie in the field of Computer Vision, Pattern Recognition and Machine Learning. He is author of one book (monograph), editor of 5 international volumes, editor of 4 international journals, author or co-author of more than 100 papers in international book chapters, international journals and international conference proceedings, and of 18 papers in national book chapters, national journals and national conference proceedings. He is co-inventor of 4 patents involving industrial partners. Dr. Farinella serves as a reviewer and on the board programme committee for major international journals and international conferences (CVPR, ICCV, ECCV, BMVC). He has been Video Proceedings Chair for the International Conferences ECCV 2012 and ACM MM 2013, General Chair of the International Workshop on Assistive Computer Vision and Robotics (ACVR - held in conjunction with ECCV 2014, ICCV 2015, ECCV 2016 and ICCV 2017), and chair of the International Workshop on Multimedia Assisted Dietary Management (MADiMa) 2015/2016. He has been Speaker at international events, as well as invited lecturer at industrial institutions. Giovanni Maria Farinella founded (in 2006) and currently directs the International Computer Vision Summer School (ICVSS). He also founded (in 2014) and currently directs the Medical Imaging Summer School (MISS). Dr. Farinella is an IEEE Senior Member and a CVF/IAPR/GIRPR/AIxIA/BMVA member. Dr. Farinella was awarded the PAMI Mark Everingham Prize in October 2017.

Antonino Furnari is a postdoctoral fellow at the University of Catania under the supervision of Dr. Giovanni Maria Farinella. He received his PhD in Mathematics and Computer Science in 2016 from the University of Catania, where he was supervised by Prof. Sebastiano Battiato and Dr. Giovanni Maria Farinella. His main research interests concern Computer Vision and Pattern Recognition, with particular focus on First Person Vision (http://iplab.dmi.unict.it/fpv/).

Sebastiano Battiato is Full Professor of Computer Science at University of Catania. He received his degree in computer science (summa cum laude) in 1995 from University of Catania and his Ph.D. in Computer Science and Applied Mathematics from University of Naples in 1999. From 1999 to 2003 he was the leader of the ?Imaging? team at STMicroelectronics in Catania. He joined the Department of Mathematics and Computer Science at the University of Catania in 2004 (respectively as assistant professor, associate professor in 2011 and full professor in 2016). He is currently Chairman of the undergraduate program in Computer Science, and Rector's delegate for Education (postgraduates and Ph.D.). He is involved in research and directorship of the IPLab research lab (http://iplab.dmi.unict.it). He coordinates IPLab participation to large scale projects funded by national and international funding bodies, as well as by private companies. Prof. Battiato has participated as principal investigator in many international and national research projects.His research interests include image enhancement and processing, image coding, camera imaging technology and multimedia forensics. He has edited 6 books and co-authored about 200 papers in international journals, conference proceedings and book chapters. Guest editor of several special issues published on International Journals. He is also co-inventor of 22 international patents, reviewer for several international journals, and he has been regularly a member of numerous international conference committees. Chair of several international events (ICIAP 2017, VINEPA 2016, ACIVS 2015, VAAM2014-2015-2016, VISAPP2012-2015, IWCV2012, ECCV2012, ICIAP 2011, ACM MiFor 2010–2011, SPIE EI Digital Photography 2011-2012-2013, etc.). He is an associate editor of the SPIE Journal of Electronic Imaging. He is the recipient of the 2011 Best Associate Editor Award of the IEEE Transactions on Circuits and Systems for Video Technology. He is director (and co-founder) of the International Computer Vision Summer School (ICVSS), Sicily, Italy. He is a senior member of the IEEE.