Egocentric Vision for Visual Market Basket Analysis

Vito Santarcangelo[1,2], Giovanni Maria Farinella[1(✉)], and Sebastiano Battiato[1]

[1] Department of Mathematics and Computer Science, University of Catania, Catania, Italy
gfarinella@dmi.unict.it
[2] Centro Studi S.r.l., Buccino, SA, Italy

Abstract. This paper introduces a new application scenario for egocentric vision: Visual Market Basket Analysis (VMBA). The main goal in the proposed application domain is the understanding of customers behaviours in retails from videos acquired with cameras mounted on shopping carts (which we call narrative carts). To properly study the problem and to set the first VMBA challenge, we introduce the VMBA15 dataset. The dataset is composed by 15 different egocentric videos acquired with narrative carts during users shopping in a retail. The frames of each video have been labelled by considering 8 possible behaviours of the carts. The considered cart's behaviours reflect the behaviour of the customers from the beginning (cart picking) to the end (cart releasing) of their shopping in a retail. The inferred information related to the time of stops of the carts within the retail, or to the shops at cash desks could be coupled with classic Market Basket Analysis information (i.e., receipts) to help retailers in a better management of spaces and marketing strategies. To benchmark the proposed problem on the introduced dataset we have considered classic visual and audio descriptors in order to represent video frames at each instant. Classification has been performed exploiting the Directed Acyclic Graph SVM learning architecture. Experiments pointed out that an accuracy of more than 93 % can be obtained on the 8 considered classes.

1 Introduction and Motivations

Egocentric vision is a new emerging area in Computer Vision [1,2]. By exploiting wearable devices it is possible to collect hours of videos that can be processed to obtain a log of the monitored scenarios. Different papers on egocentric vision applications have been published in the recent literature. The main tasks addressed in this area are related to scene recognition [3], motion understanding [4], objects and actions recognition [5–8], 3D reconstruction [9,10] and summarization [11,12]. Among the others, context aware computing is an important research area for egocentric (first-person) vision domain [3,13,14]. Temporal segmentation of Egocentric Vision is also fundamental to understand the behavior of the users wearing a camera [4,15]. Recently, the retail scenario has

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part I, LNCS 9913, pp. 518–531, 2016.
DOI: 10.1007/978-3-319-46604-0_37

Fig. 1. Information useful for VMBA.

become of particular interest for applications related to the geo-localization of the user's positions and the reconstruction of the spaces [16]. In the retail context, one of the possible developments of interest concerns the monitoring of the paths of customers, thereby enabling to carry out an analysis of their behaviors. Nowadays customers monitoring is partially employed by using loyalty cards, counting devices connected with Bluetooth and WiFi systems, employing RFID tags [17], as well as fixed cameras (e.g., video surveillance). Differently than classic approaches and considering the potentials and spread of egocentric cameras, in this paper we consider to turn an ordinary cart in a "narrative shopping cart" by equipping it with a camera. The acquired egocentric videos are processed with algorithms able to turn the visual paths in customers' behaviour. By doing so it is possible to acquire all over the route travelled by 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, to understand the areas of personal interest, to estimate the path speed, to estimate the most busy areas of the retail by clustering routes, to register the reactions opposite to the audio announcements in the store, as well as to infer the inefficiencies (e.g., slowness at cash desk). We call this kind of behavioral monitoring in a retail "Visual Market Basket Analysis" (VMBA) since it can be useful to enrich the classic "Market Basket Analysis" methods [18] used to infer the habits of customers.

In this paper we introduce the problem of VMBA considering three different high-level information related to the customers which is carrying the narrative cart (Fig. 1): location (i.e., indoor vs outdoor), action (i.e., stop vs moving), and scene context (i.e., cash desk, retail, parking, road). These high-level information can be organized in a hierarchy to produce 8 different behaviors useful in the retail domain to log the storyline of the shopping of the customers that can be eventually associated to others information (e.g., receipts) for retail management purposes. The 8 classes are shown as path, from the root to the leaves, of the tree in Fig. 2. Given a frame of the video acquired with the narrative cart camera, at

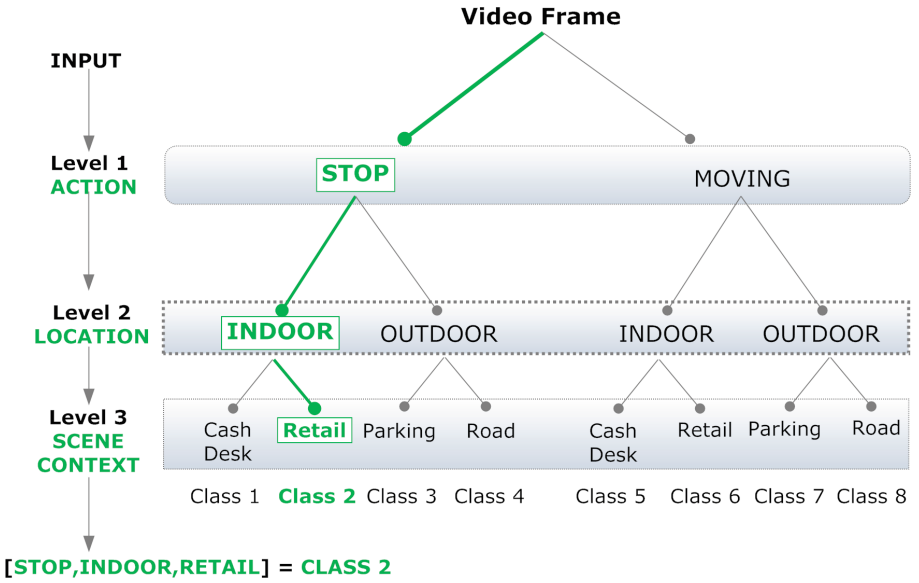

Fig. 2. Considered VMBA behavioral classes organized in a hierarchy.

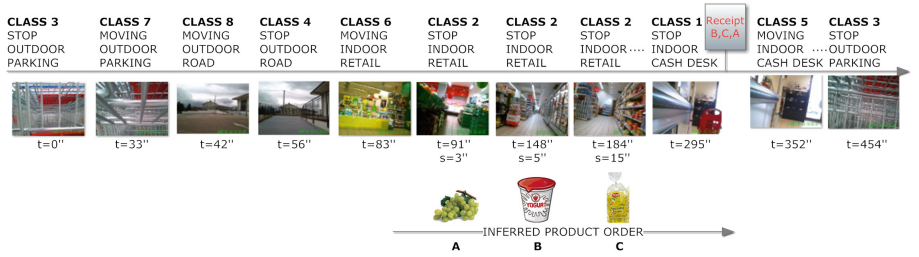

Fig. 3. VMBA timeline temporally segmented considering the 8 classes. t denotes the time, whereas s denotes the stopping time.

each instant we wish to know a triplet corresponding to a path in the tree (e.g., [STOP, INDOOR, RETAIL] in Fig. 2). By classifying each frame of the acquired egocentric videos with the proposed 8 classes (i.e., the 8 possible triplet of the hierarchy in Fig. 2), it will be simple to perform an analysis of what are the custom behaviors, and also understand if there are problem to be managed in the store. An example of a narrative cart egocentric video together with a temporal segmentation with respect to the 8 defined classes is shown in Fig. 3. For example from the segmented narrative cart video it will be simple to understand how long are the stops to the cash desk by considering the frames classified with the triplets [STOP, INDOOR, CASH DESK] and [MOVING, INDOOR, CASH DESK]. This can be useful to eventually plan the opening of more cash desks to provide a better service to the customers. By analyzing the inferred triplets of a

narrative cart video, it will be simple to understand if there are carts outside the cart parking spaces in order to take actions (e.g., if there is a long sequence of the triplet [STOP, OUTDOOR, ROAD] which does not change for long time). A lot of other considerations for a better management of the retail can be done by considering the narrative cart egocentric videos when those have been temporal segmented by classifying each frame with the 8 possible triples (i.e., behavioral classes). By combining the receipt with the temporal segmented video and with algorithms for visual re-localization [19] it will be simple to establish the order in which the products have been taken, hence increasing the information that are usually exploited by the classic "Market Basket Analysis" algorithms [18] and opening new research perspectives (Fig. 3). To set the first VMBA challenge we propose a new dataset of 15 sequences (VBMA) obtained by collecting and labeling real video sequences acquired in a retail. The proposed dataset is available for the research community upon request to the authors. We benchmark the dataset by considering a Direct Acyclic Graph SVM approach [20] coupled with classic descriptors for the representation of visual content (GIST [21]), motion (Optical Flow [22]), and audio (MFCC [23]). Experiments show that a classification accuracy of more than 93 % can be obtained on the proposed VMBA dataset when the 8 behavioral classes are considered.

The reminder of the paper is organized as follows: in Sect. 2, we describe the approach used to perform the benchmark study. Section 3 presents the dataset acquired with the narrative cart in a retail of Southern of Italy. Section 4 discuses the results. Finally, Sect. 5 concludes the paper and gives hints and open challenges.

2 Proposed Approach

The main goal of this paper is the segmentation of egocentric videos acquired with narrative carts in chapters automatically labeled with one of the 8 possible classes defined by the path of the tree presented in Fig. 2. A chapter is a set of consecutive frames of the video which present the same behavior (e.g., a sequence of frames with the same label [stop, indoor, cash desk]). The proposed behavioral tree has three layers (Fig. 2). The first layer is related to the action of moving or stopping the narrative carts (which reflect the user's basic actions in a retail). The second level of the tree identifies the high level location where the user is acting (indoor vs outdoor). The third level is related to the scene context during the shopping. Regarding the scene context we have considered four classes: parking, road, cash desk and retail. Looking at the VMBA hierarchy defined in Fig. 2, it is straightforward to understand that the different scene contexts are observed depending on the main location where the user is acting (indoor vs outdoor). Hence the scene contexts cash desk and retail can be observed only in indoor location, whereas the scene contexts parking and road can be observed only in outdoor locations. To perform the classification of a narrative cart video frame in order to automatically assign to it a triplet [action, location, scene context], two main "ingredients" are needed: the representation of the frame and

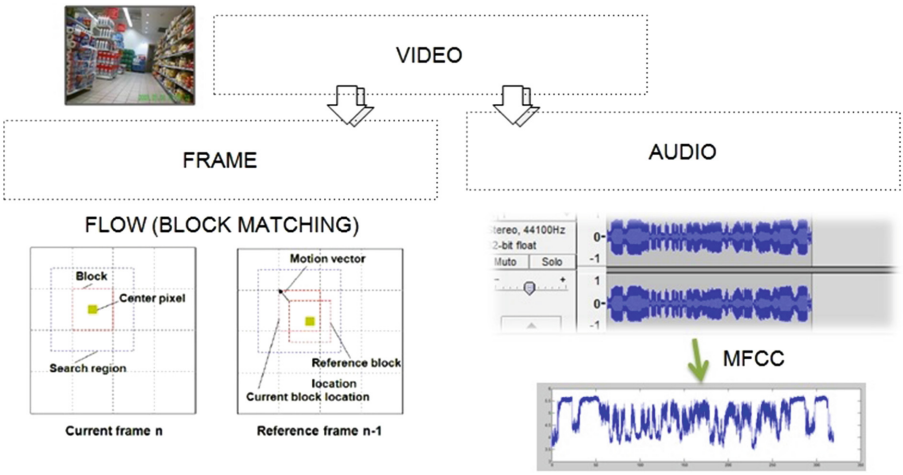

Fig. 4. Features used at the first level.

the classification modality. To benchmark the problem we have employed classic simple features for the representation and standard discriminative classifiers. In the following subsections we will detail the representations used for the three different layers of the tree in Fig. 2 and the classification method employed.

2.1 Representation at the First Level: Actions

The first level analyzes the customer behavior from the point of view of the motion of the narrative cart by considering two possible states: stop and moving. In order to understand such states from the egocentric video, in our benchmark we tested the MFCC audio features [24] and the optical flow features computed with the classic block matching approach [25] (Fig. 4). For the optical flow we have considered the frames divided in 9 blocks, so for each frame we have got a 9-dimensional features vector. The audio processing produced a feature vector of 62 components. We have decided to consider the audio because there is a visual correlation between the audio waveform with the narrative cart motion and locations (Fig. 5). The exploitation of the optical flow feature is straightforward since the problem under consideration. In our experiments we have tested the two considered features separately and jointly.

2.2 Representation at the Second Level: Location

The second level of the tree in Fig. 2 has the scope to identify the high level location where the user is acting: indoor vs outdoor. As for the first level we have considered MFCC features after visual inspection of waveform (Fig. 5). Indeed the waveform is more pronounced in the outdoor environment than in the indoor location. To benchmark the VMBA problem addressed in this paper for indoor

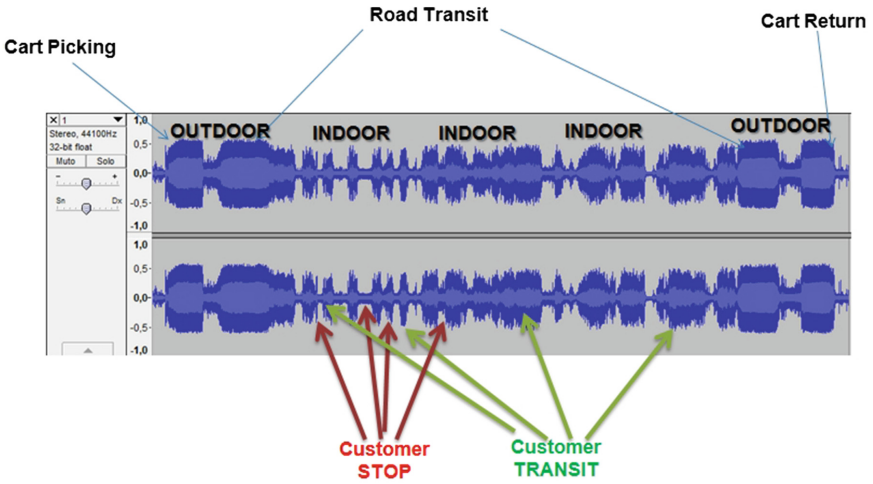

Fig. 5. Audio waveform and behaviors.

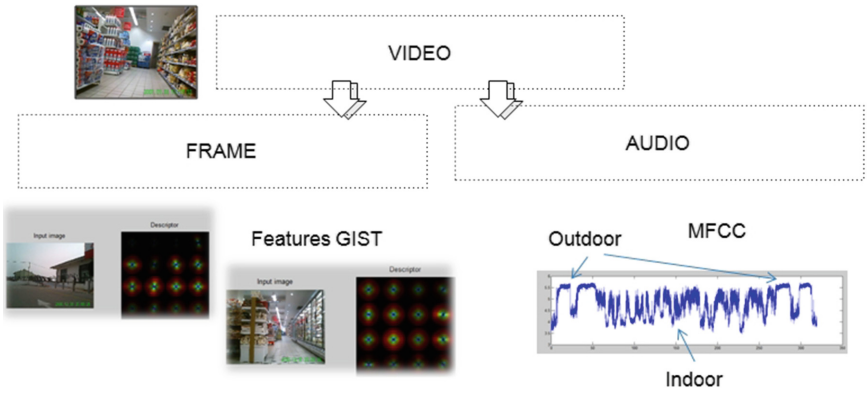

Fig. 6. Features used at the second level.

vs outdoor locations discrimination we have also tested the popular GIST visual descriptor [26], which is able to encode the scene context with a feature vector composed by 512 components (Fig. 6). In our experiments we tested indoor vs outdoor classification by considering audio and visual features independently and combined.

2.3 Representation at the Third Level: Scene Context

The third level of the hierarchy in Fig. 2 is related to the analysis of the scene context considering four different classes: cash desk, retail, parking and road. As described before, the first two contexts are related to the indoor environment, whereas the other two describe in more details the outdoor location. For this level

Fig. 7. Narrative carts.

of description we have used the GIST descriptor [26] again since its property in capturing the shape of the scene for context discrimination.

2.4 Classification Methods

After representing a frame of the egocentric video as described in previous sections, a classifier has to be employed to infer one of the 8 considered classes (i.e., one of the 8 possible triplet corresponding to a path of the tree in Fig. 2). In this paper we benchmarked three different classification modalities:

- combination of the results obtained by three different SVM classifiers in correspondence of the three different levels of the hierarchy;
- a single Multi-Class SVM trained on the 8 possible classes;
- a Direct Acyclic Graph SVM learning architecture (DAGSVM) [20] which reflects the hierarchy in Fig. 2 on each node.

Experiments reported in Sect. 4 demonstrate that good classification accuracy can be obtained considering the hierarchical classification performed by DAGSVM.

3 VMBA15 Dataset

To set the first VMBA challenge and perform the benchmark on the considered problem we acquired a dataset composed by 15 different egocentric videos with narrative carts in a retail of the Southern of Italy during real shopping. To this aim we have mounted a narrative cam veho muvi pro [27] into the front of a classic shopping cart as depicted in Fig. 7. Each narrative cart video has a

Fig. 8. Some visual examples of frames related to the egocentric videos of the VMBA15 dataset. The eight scenes represent the eight possible classes with order from top to bottom, left to right. Notice that some classes are characterized by similar visual content but different actions, such as in the case of the image at the first row of the second column (CLASS 2) and the third image in the second column (CLASS 6). The images at the first row are related to CLASS 1 (left) and CLASS 2 (right), and share the same location (INDOOR) but show different scene context (RETAIL vs CASH DESK).

Table 1. Number of samples per class for each egocentric video.

VIDEO	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	Total
1	0	13	2	0	8	193	17	89	322
2	0	17	4	0	9	266	10	84	390
3	0	19	4	0	12	226	10	96	367
4	0	20	3	0	10	277	13	106	429
5	0	20	4	0	10	213	9	107	363
6	69	10	28	2	59	134	35	91	428
7	0	3	13	0	7	102	16	119	260
8	6	36	7	0	8	233	8	75	373
9	142	186	8	0	18	550	9	85	998
10	0	5	3	0	7	106	13	75	209
11	42	90	31	0	10	406	16	89	684
12	0	36	22	0	26	436	23	104	647
13	56	80	57	4	7	130	28	133	495
14	50	396	7	0	3	485	11	46	998
15	81	528	0	27	3	310	3	46	998

duration between 3 to 20 min and resolution of 640×480 pixels. Audio has been also recorded since it can be useful to discriminate indoor vs outdoor environment. From each narrative cart video we have sampled and manually labeled frames and audio at 1 fps considering the 8 possible paths of the tree shown in Fig. 2. The total number of sample is 7961 (see Table 1 for more retails about the dataset). Some examples of frames extracted from the VMBA15 dataset are shown in Fig. 8. The labeled data is available upon request to the authors.

4 Experimental Settings and Results

We have performed experiments by randomly splitting that dataset in three parts composed by five egocentric video each. The experiments have been repeated three times considering 10 videos for the training and 5 video for the tests. The final results are obtained by averaging among the three runs. As first we have compared the different features employed at the different levels of the hierarchy independently by exploiting a SVM classifier with RBF kernel. This was useful to understand which are the best features (or combination of them) to be employed at each level for the final classification of each frame with respect to the 8 classes. In Table 2 are reported the results of the stop vs moving classification (i.e. First Level). Both audio and visual feature achieve good performance, however, visual feature (the flow) outperforms the audio features with a margin of about 5 %. Interestingly the combination by concatenation of audio and visual features improve the results and obtains an accuracy of 94.50 % in discriminating stop vs moving actions. The obtained results pointed out that the combination of MFCC and flow features has to be used at the first level.

Also in the case of the discrimination of the main location where the narrative cart is moving (or stopping), the visual features outperform audio features with a good margin by obtaining 95.79 % of accuracy (see Table 3). Differently than in the first level, the combination of audio and visual features do not improve the indoor vs outdoor classification. Hence for the second level we decided to employ the GIST descriptor alone.

For the third level of the hierarchy we have obtained an accuracy of 92.42 % with the GIST descriptor. Note that in this case a multi-class SVM with RBF kernel has been trained to discriminate this four possible scene contexts without considering the prior indoor vs outdoor. The results respect to the four scene contexts are reported in Table 4. The main confusion is related to the class parking and retail (first column in Table 4). This is probably due to the encoding of the scene information by the GIST descriptor. Indeed, when the narrative cart is in the parking space, the scene is mainly composed by vertical and horizontal edges that can be confused with the vertical and horizontal edges of some scenes in the retail (see Fig. 9). As demonstrated by the results reported later, this problem is mitigated when the classification is performed by the DAGSVM approach since it introduces a prior on the main location (indoor vs outdoor). One more problem in the classification is due to strong occlusions as the one in the examples reported in Fig. 10.

The aforementioned experiments pointed out that the best features to be employed in the hierarchy are the combination of MFCC and FLOW for the first level, whereas the GIST descriptor for the second and third level. Since the main goal is the classification with respect to the 8 possible triplets generated by the hierarchy in Fig. 2, after selecting the features for the three levels independently we have compared the three classification modalities described in Sect. 2.4. For the combination of three different classifiers (one for each level) we have considered the concatenation of the labels given by three different SVM (with RBF kernel) when trained independently on the best selected features of

Table 2. STOP vs MOVING classification

	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

Table 3. INDOOR vs OUTDOOR classification

	GIST	MFCC	COMBINED
Accuracy %	95.79	88.00	91.77
TP RATE %	89.3	49.51	67.49
TN RATE %	97.8	97.66	97.1
FP RATE %	2.20	2.34	2.90
FN RATE %	10.7	50.49	32.51

Table 4. Scene Context classification

PREDICTED				
	PARKING	ROAD	RETAIL	CASH DESK
PARKING	54.25 %	17.01 %	25.09 %	3.64 %
ROAD	0.55 %	88.94 %	9.5 %	1.01 %
RETAIL	0.17 %	1.09 %	98.46 %	0.28 %
CASH DESK	0.13 %	7.47 %	17.21 %	75.19 %

Fig. 9. On the left a typical scene of the narrative cart when in the parking space. On the right an example of a frame acquired by the narrative cart in retail. The distribution of vertical and horizontal edges could generate confusion in the classification.

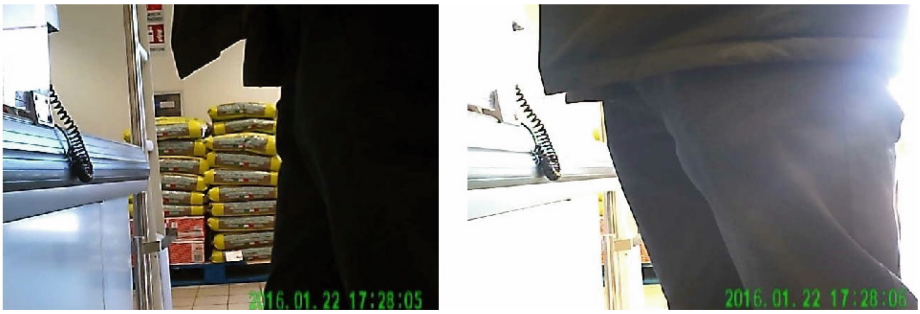

Fig. 10. Some examples of frames with occlusions (at the cash desk).

Fig. 11. Examples of frames correctly classified by the proposed DAGSVM approach. These frames are misclassified by the other two compared approaches. The frame on the left is related to the parking space of the carts, but is recognized as retail by both the combined approach and Multi-Class SVM. The frame on the right is related to outdoor, but is recognized as indoor by both the combined approach and Multi-Class SVM.

Table 5. Results of the classification considering the 8 classes

	Combination	Multi-Class SVM	DAGSVM
Accuracy %	87.36	69.54	93.47

the three levels. For the multi-class SVM with 8 classes we have trained a SVM with RBF kernel on the concatenation of MFCC, GIST and FLOW features. Finally, we have trained a DAGSVM [20] reflecting the hierarchy in Fig. 2. Each node of the DAG is composed by a SVM with RBF kernel in which the best features to solve the problem at each node are exploited. The results of the three different approaches are reported in Table 5.

The final results are in favor of the DAGSVM approach which obtain an accuracy of 93.47 %. It is worth to note that a concatenation of the MFCC features with the FLOW and GIST descriptors does not allow a multi-class SVM to reach good accuracy (69.54 %). Finally, the results of the combination

of the three different classifiers stated at the second place in the classification a ranking (87.36 %). The visual examples for the assessment of the output given by the proposed DAGSVM-based approach are available in Fig. 11 and at the following URL: http://iplab.dmi.unict.it/epic2016.

5 Conclusion

This paper introduces 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 with cameras mounted on shopping carts. The VBMA15 dataset has been labeled considering 8 different classes corresponding to a hierarchical organization of actions, location and scene contexts. A first benchmark has been performed considering different 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. Our future works will be devoted to a complete formalization of the VMBA problem and to the augmentation of both dataset and labels to reflect the domain. We will also consider different retails to introduce a more realistic variability in the dataset. Moreover, recently appeared learning mechanisms to encode audio and flow features will be considered [3,22,28]. Also we will take into account egocentric camera with low capabilities for the recognition of the scene context [21]. The problem of relocalization [19] for the narrative carts will be explored to infer the order of the products acquired in a retail in order to combine visual location recognition and receipts. Finally, structure from motion based techniques [29] will be considered to automatically reconstruct the 3D shape of the store and to track the paths of the narrative carts to extract information useful for retails management.

Acknowledgment. The authors would like to thank Antonino Furnari, for his support in the development of this work.

References

1. Betancourt, A., Morerio, P., Regazzoni, C.S., Rauterberg, M.: The evolution of first person vision methods: a survey. IEEE Trans. Circuits Syst. Video Technol. **25**(5), 744–760 (2015)
2. Mann, S., Kitani, K.M., Lee, Y.J., Ryoo, M.S., Fathi, A.: An introduction to the 3rd workshop on egocentric (first-person) vision. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 827–832 (2014)
3. Furnari, A., Farinella, G.M., Battiato, S.: Recognizing personal contexts from egocentric images. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 393–401 (2015)
4. Poleg, Y., Arora, C., Peleg, S.: Temporal segmentation of egocentric videos. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2537–2544 (2014)

5. Agrawal, P., Carreira, J., Malik, J.: Learning to see by moving. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 37–45 (2015)
6. Damen, D., Leelasawassuk, T., Haines, O., Calway, A., Mayol-Cuevas, W.W.: You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In: British Machine Vision Conference (2014)
7. Fathi, A., Rehg, J.M.: Modeling actions through state changes. In: Computer Vision and Pattern Recognition, pp. 2579–2586 (2013)
8. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In: Computer Vision and Pattern Recognition, pp. 3281–3288 (2011)
9. Poleg, Y., Halperin, T., Arora, C., Peleg, S.: Egosampling: Fast-forward and stereo for egocentric videos. In: Computer Vision and Pattern Recognition, pp. 4768–4776 (2015)
10. Lee, Y.J., Ghosh, J., Grauman, K.: Discovering important people and objects for egocentric video summarization. In: Computer Vision and Pattern Recognition, pp. 1346–1353 (2012)
11. Xiong, B., Kim, G., Sigal, L.: Storyline representation of egocentric videos with an applications to story-based search. In: International Conference on Computer Vision, pp. 4525–4533 (2015)
12. Lu, Z., Grauman, K.: Story-driven summarization for egocentric video. In: Computer Vision and Pattern Recognition, pp. 2714–2721 (2013)
13. Xu, Q., Li, L., Lim, J.H., Tan, C.Y.C., Mukawa, M., Wang, G.: A wearable virtual guide for context-aware cognitive indoor navigation. In: International Conference on Human-computer Interaction with Mobile Devices and Services, pp. 111–120 (2014)
14. Starner, T., Schiele, B., Pentland, A.: Visual contextual awareness in wearable computing. In: Second International Symposium on Wearable Computers, Digest of Papers, pp. 50–57, October 1998
15. Ortis, A., Farinella, G.M., D'Amico, V., Addesso, L., Torrisi, G., Battiato, S.: Organizing egocentric videos for daily living monitoring. Submitted to the ACM MM International Workshop on Lifelogging Tools and Applications (2016)
16. Wang, S., Fidler, S., Urtasun, R.: Lost shopping! monocular localization in large indoor spaces. In: International Conference on Computer Vision, pp. 2695–2703 (2015)
17. Ali, Z., Sonkusare, R.: Rfid based smart shopping: an overview. In: International Conference on Advances in Communication and Computing Technologies, pp. 1–3 (2014)
18. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining, 1st edn. Addison-Wesley Longman Publishing Co., Inc., Boston (2005)
19. Kendall, A., Grimes, M., Cipolla, R.: Posenet: a convolutional network for real-time 6-dof camera relocalization. In: International Conference on Computer Vision, pp. 2938–2946 (2015)
20. Platt, J.C., Cristianini, N., Shawe-taylor, J.: Large margin dags for multiclass classification. In: Advances in Neural Information Processing Systems, vol. 12, pp. 547–553 (2000)
21. Farinella, G., Raví, D., Tomaselli, V., Guarnera, M., Battiato, S.: Representing scenes for real-time context classification on mobile devices. Pattern Recogn. **48**(4), 1086–1100 (2015)
22. Dosovitskiy, A., Fischery, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., Smagt, P.V.D., Cremers, D., Brox, T.: Flownet: learning optical flow with convolutional networks. In: International Conference on Computer Vision, pp. 2758–2766 (2015)

23. Muda, L., Begam, M., Elamvazuthi, I.: Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. CoRR abs/1003.4083 (2010)
24. Sahidullah, M., Saha, G.: Design, analysis and experimental evaluation of block based transformation in mfcc computation for speaker recognition. Speech Commun. **54**(4), 543–565 (2012)
25. Barron, J.L., Fleet, D.J., Beauchemin, S.S., Burkitt, T.A.: Performance of optical flow techniques. In: Computer Vision and Pattern Recognition, pp. 236–242 (1992)
26. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vision **42**(3), 145–175 (2001)
27. Veho Muvi Cam: Narrative cam. www.vehomuvi.com. Accessed April 2016
28. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
29. Wu, C.: Towards linear-time incremental structure from motion. In: International Conference on 3D Vision, pp. 127–134 (2013)