Recognizing Personal Contexts from Egocentric Images

Antonino Furnari, Giovanni M. Farinella, Sebastiano Battiato
Department of Mathematics and Computer Science - University of Catania
Viale Andrea Doria, Catania, 95125, Italy
{furnari,gfarinella,battiato}@dmi.unict.it

Abstract

Wearable cameras can gather first-person images of the environment, opening new opportunities for the development of systems able to assist the users in their daily life. This paper studies the problem of recognizing personal contexts from images acquired by wearable devices, which finds useful applications in daily routine analysis and stress monitoring. To assess the influence of different device-specific features, such as the Field Of View and the wearing modality, a dataset of five personal contexts is acquired using four different devices. We propose a benchmark classification pipeline which combines a one-class classifier to detect the negative samples (i.e., images not representing any of the personal contexts under analysis) with a classic one-vs-one multi-class classifier to discriminate among the contexts. Several experiments are designed to compare the performances of many state-of-the-art representations for object and scene classification when used with data acquired by different wearable devices.

1. Introduction and Motivation

Wearable devices capable of continuously acquiring images from the user's perspective have become more and more used in the last years. Part of this success is due to the availability of commercial products which, featuring small size and extended battery life, are affordable both in terms of costs and usability. The egocentric data acquired using wearable cameras jointly offers new opportunities and challenges [1]. The former are related to the relevance of the egocentric data to the activity performed by the users, which makes its analysis interesting for a number of applications [2, 3, 4, 5]. The latter concern the large variability exhibited by the acquired data due to the inherent camera instability, the non-intentionality of the framing, the presence of occlusions (e.g., by the user's hands), as well as the influence of varying lighting conditions, fast camera movements and motion blur [1]. Figure 1 shows some examples of the typical variability exhibited by egocentric images.

Despite the recent industrial interest in these technologies, researchers have explored the opportunities offered by wearable cameras ever since the 90s. Applications include recognizing human activities [2, 3, 5], improving user-machine interaction [6], context modelling [7, 8], video temporal segmentation and indexing [9], and video summarization [10, 11]. Wearable and mobile devices have been also employed in applications related to assistive technologies, such as, food-intake monitoring [12], providing assistance to the user on object interaction [4, 13], estimating the physiological parameters of the user for stress monitoring and quality of life assessment [14], providing assistance to disabled or elders through lifelogging and activity summarization [15, 16].

Figure 1. Some egocentric images of personal contexts. Each column reports four different shots of the same context acquired using wearable cameras during regular user activity. The following abbreviation holds: coffee v. machine - coffee vending machine.

Visual contextual awareness is a desirable property in wearable computing. As discussed in [2], wearable computers have the potential to experience the life of the user in a "first-person" sense, and hence they are suited to provide serendipitous information, manage interruptions and tasks or predict future needs without being directly commanded by the user. In particular, being able to recognize the personal contexts in which the user operates at the instance level (i.e., recognizing a particular environment such as "my office"), rather than at the category-level, (e.g., "an office"), can be interesting in a number of assistive-related

scenarios in which contextual awareness may be beneficial. Possible applications could include daily routine analysis, stress monitoring and context-based memory reinforcement for people with memory impairment. Other applications could focus on assessing the mobility of elders inside their homes in the context of ageing-in-place, as well as providing assistance on the possible interactions with the objects available in a specific environment.

In this paper, we study the problem of recognizing personal contexts from egocentric images. We define a personal context as:

> *a fixed, distinguishable spatial environment in which the user can perform one or more activities which may or may not be specific to the context*

According to the definition above, a simple example of personal context consists in an office desk, in which the user can perform a number of activities, such as typing at the computer or reading some documents. In addition to the general issues associated with egocentric data (e.g., occlusions, fast camera movement, etc.), recognizing contexts of interest for a person (i.e., personal contexts) poses some unique challenges:

- few labelled samples are generally available since it is not feasible to ask the user to collect and annotate huge amounts of data for learning purposes;
- the appearances of personalized contexts are characterized by large intra-class variability, due to the different views acquired by the camera as the user moves in the environment;
- personalized contexts belonging to the same category (e.g., two different offices) tend to share similar appearances;
- given the large variability of visual information that will be acquired by an always-on wearable camera, the gathering of representative negative samples for learning purposes (i.e., images depicting scenes which do not belong to any of the considered contexts to be recognized) is not always feasible.

In this study, we perform a benchmark of different state-of-the-art methods for scene and object classification on the task of recognizing personal contexts. To this aim, we built a dataset of egocentric videos containing five personalized contexts which are relevant to the tasks of routine analysis and stress monitoring. Figure 1 shows some examples of the acquired data. In order to build a plausible training set, the user is only asked to take a ten-seconds video of the personal context of interest to be monitored by moving the camera around to cover the different views of the environment. To assess the influence of device-specific factors such as wearing modality and Field Of View (FOV), we acquire the dataset using four different devices. In order to compare the performances of different state-of-the-art representations, we propose a benchmark classification scheme which combines in cascade a one-class classifier to detect the negative samples and a multi-class classifier to discriminate among the personal contexts. The experiments are carried by training and testing the benchmark classification scheme on data arising from different combinations of devices and representations.

The remainder of the paper is organized as follows: in Section 2 we discuss the related works; Section 3 presents the devices and the data used in the experiments; Section 4 summarizes the considered state-of-the-art representation techniques; in Section 5 we define the experimental settings, whereas in Section 6 we discuss the results; Section 7 concludes the paper and gives insights for further works.

2. Related Works

The notion of personal context presented in Section 1 is related to the more general concept of visual context, which has been thoroughly studied in the past decade. In particular, in [17] is described a procedure for organizing real world scenes along semantic axes, while in [18] is proposed a computational model for classifying real world scenes. Efficient computational methods for scene understanding have also been proposed for mobile and embedded devices [19, 20]. More recently, Convolutional Neural Networks (CNNs) have been successfully applied to the problem of scene classification [21]. Our work is also related to the problem of recognizing human activities from egocentric data, which has been already studied by Computer Vision researchers. In [3], daily routines are recognized in a bottom-up way through activity spotting. In [2], some basic tasks related to the Patrol game are recognized from egocentric videos in order to assist the user. In [5], Convolutional Neural Networks and Random Decision Forests are combined to recognize human activities from egocentric images. Also systems for recognizing personal contexts have already been proposed. In [7], personal locations are recognized based on the approaching trajectories. In [8], images of sensitive spaces are detected for privacy purposes combining GPS information and an image classifier. In [22], an unsupervised system for discovering recurrent scenes in large sets of lifelogging data is proposed.

Differently than the aforementioned works, we systematically study the performances of the state-of-the-art methods for scene and object representation and classification on the task of personal context recognition. We assume that only visual information is available and that the quantity of labelled data is limited (see challenges in Section 1).

3. Wearable Devices and Egocentric Dataset

We built a dataset of egocentric videos acquired by a single user in five different personal contexts. Given the availability of diverse wearable devices on the market, we selected four different cameras in order to assess the influence of some device-specific factors, such as the wearing modality and the Field Of View (FOV), on the considered task. Specifically, we consider the smart glasses Recon Jet (RJ), two ear-mounted Looxcie LX2, and a wide-angular chest-

	Wearing Modality			Field Of View	
	Glasses	Ear	Chest	Narrow	Wide
RJ	✓			✓	
LX2P		✓		✓	
LX2W		✓			✓
LX3			✓		✓

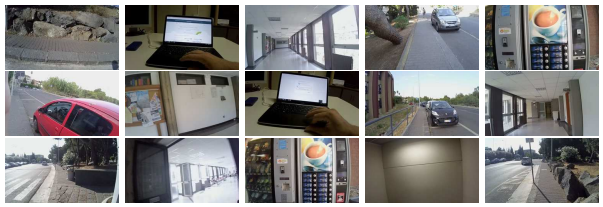
Table 1. A summary of the main features of the devices used to acquire the data. The technical specifications of the cameras are reported at the URL: *http://iplab.dmi.unict.it/PersonalContexts/*

mounted Looxcie LX3. The Recon Jet and the Looxcie LX2 devices are characterized by narrow FOVs ($70°$ and $65,5°$ respectively), while the FOV of the Looxcie LX3 is considerably larger ($100°$). One of the two ear-mounted Looxcie LX2 is equipped with a wide-angular converter, which allows to extend its Field Of View at the cost of some fisheye distortion, which in some cases requires dedicated processing techniques [23, 24]. The wide-angular LX2 camera will be referred to as LX2W, while the perspective LX2 camera will be referred to as LX2P. Table 1 summarizes the main features of the cameras used to acquire the data. Figure 2 (a) shows some sample images acquired by the devices under analysis.

The considered five personal contexts arise from the daily activities of the user and are relevant to assistive applications such as quality of life assessment and daily routine monitoring: car, coffee vending machine, office, TV and home office. Since each of the considered context involves one or more static activities, we assume that the user is free to turn his head and move his body when interacting with the context, but he does not change his position in the room. In line with the considerations discussed in Section 1, our training set is composed of short videos ($\approx$ 10 seconds) of the personal contexts (just one video per context) to be monitored. During the acquisition of the context, the user is asked to turn his head (or chest, in the case of chest-mounted devices) in order to capture a few different views of the environment. The test set consists in medium length (8 to 10 minutes) videos of normal activity in the given personal contexts with the different devices. Three to five testing videos have been acquired for each context. We also acquired several short videos containing likely negative samples, such as indoor and outdoor scenes, other desks and other vending machines. Figure 2 (b) shows some negative samples. Most of the negative-videos are used solely for testing purposes, while a small part of them is used to extract a fixed number (200 in our experiments) of frames which are used as "optimization negative samples" to optimize the performances of the one class classifier. The role of such negative samples is detailed in Section 5. At training time, all the frames contained in the 10-seconds video shots are used, while at test time, only about 1000 frames per-class uniformly sampled from the testing videos are used.

In order to perform fair comparisons across the different devices, we built four independent, yet compliant, device-specific datasets. Each dataset comprises data acquired by a single device and is provided with its own training and test sets. Figure 2 (a) shows some sample images included in the dataset. The device-specific datasets are available for download at the URL: *http://iplab.dmi.unict.it/PersonalContexts/*.

(a) positive samples



(a) negative samples

Figure 2. (a) Some sample images of the five personal contexts acquired using the considered wearable devices. Images from the same contexts are grouped by columns, while images acquired using the same device are grouped by rows. The following abbreviation holds: coffee v. machine - coffee vending machine. (b) Some negative samples used for testing purposes.

4. Representations

We assume that the input image I can be mapped to a feature vector $\mathbf{x} \in \Re^d$ which can be further used with a classifier through a representation function Φ. Specifically, we consider three different classes of representation functions Φ: holistic, shallow and deep. All of these representations have been used in the literature for different tasks related to scene understanding [18, 21] and object detection [25, 26]. In the following subsections we discuss the details of the considered representations and the related parameters.

4.1. Holistic Representations

Holistic representations aim at providing a global descriptor of the image to capture class-related features and discard instance-specific variabilities. Holistic representations have been used mainly for scene classification [18, 20]. In this work, we consider the popular GIST descriptor proposed in [18] and use the standard implementation and parameters provided by the authors. In these settings, the GIST descriptors have dimensionality $d = 512$.

4.2. Shallow Representations

Representations based on encoding of local features (e.g., dense-multiscale or keypoint-based SIFT descriptors) have recently been referred to as shallow representations as opposed to the deep representations provided by CNNs [26]. We consider the Improved Fisher Vector (IFV) scheme to encode dense-multiscale SIFT features extracted from the input image according to the approach discussed in [25, 26]. The IFV can be considered the state-of-the-art in shallow representations for object classification [25, 26]. Motivated by the geometric variability exhibited by egocentric images (e.g., image rotation), in addition to the dense-multiscale extraction scheme proposed in [25], we also consider a keypoint-based extraction scheme with the aim of improving the rotational and translational invariance properties of the representation. In this case, the SIFT descriptors are computed according to the keypoint locations and scale extracted by a standard SIFT keypoint detector. When dense SIFT features are extracted, the input images are resized to a normalized height of 300 pixels (keeping the original aspect ratio), while no resizing is performed when sparse SIFT keypoints are considered. Following [25], the SIFT descriptors are component-wise square-rooted and decorrelated using Principal Component Analysis (PCA) in order to reduce their dimensionality to 80 components. We also consider the spatially-extended local descriptors discussed in [26] in our experiments. This variant simply consists in concatenating the normalised spatial coordinates of the location from which the descriptor is extracted with the PCA-reduced SIFT descriptors, obtaining a 82-dimensional vector as detailed in [26]. As discussed in [25], we train a Gaussian Mixture Model (GMM) with a number of centroids $K = 256$ on the PCA-decorrelated descriptors extracted from all the images of the training set (excluding the negatives). We also performed experiments using a larger number of centroids equal to $K = 512$. The IFV representation is obtained concatenating the average first and second order differences between the local descriptors and the centres of the learned GMM (the reader is referred to [25] for the details). Differently from [26], we do not L2 normalize the IFV descriptor in order to employ different normalization methods as discussed in Section 5. The dimensionality of the IFV descriptors depends on the number of clusters K of the GMM and on the number of dimensions D of the local feature descriptors according to the formula: $d = 2KD$. Using the parameters discussed above, the number of dimensions of our IFV representations ranges from a minimum of 40960 to a maximum of 83968 components. The VLFeat library[1] is used to perform all the operations involved in the computation of the IFV representations.

[1] VLFeat: http://www.vlfeat.org/.

4.3. Deep Representations

Convolutional Neural Networks (CNNs) have demonstrated state-of-the-art performances in a series of tasks including object and scene classification [21, 26, 27]. They allow to learn multi-layer representations of the input images which are optimal for a selected task (e.g., object classification). CNNs have also demonstrated excellent transfer properties, allowing to "reuse" a representation learned for a given task in a slightly different one. This is generally done extracting the representation contained in the penultimate layer of the network and reusing it in a classifier (e.g., SVM) or finetuning the pre-trained network with new data and labels. We consider three publicly available networks which have demonstrated state-of-the-art performances in the tasks of object and scene classification, namely AlexNet [27], VGG [26] and Places205 [21]. AlexNet and VGG have different architectures but they have been trained on the same data (the ImageNet dataset). Places205 has the same architecture as AlexNet, but it has been trained to discriminate contexts on a dataset containing 205 different scene categories. The different networks allow us to assess the influence of both network architectures and original training data in our transfer learning settings. To build our deep representations, we extract for each network model the values contained in the penultimate layer when the input image (rescaled to the dimensions of the first layer) is propagated into the network. This consists in a compact 4096-dimensional vector which corresponds to the representation contained in the hidden layer of the final Multilayer Perceptron included in the network.

5. Experimental Settings

The aim of the experiments is to study the performances the state-of-the-art representations discussed in Section 4 on the considered task. For all the experiments, we refer to the benchmark classification pipeline illustrated in Figure 3. The classification into the 6 different classes (the "negative" class, plus the 5 context-related classes) is obtained using a cascade of a one-class SVM (OCSVM) and a regular multi-class SVM (MCSVM). The OCSVM detects the negative samples and assigns them to the negative class. All the other samples are fed to the MCSVM for context discrimination. Following [25, 26], we transform the input feature vectors using the Hellinger's kernel prior to feed them to the linear SVM classifiers. Since the Hellinger's kernel is additive homogeneous, its application can be efficiently implemented as detailed in [25]. Differently from [25, 26], we do not apply the L2 normalization to the feature vectors, but instead we independently scale each component of the vectors in the range $[-1, 1]$ subtracting the minimum and dividing by the difference between the maximum and minimum values. Minima and maxima for each component are computed from the training set and reported on the test set. This overall preprocessing procedure outperforms

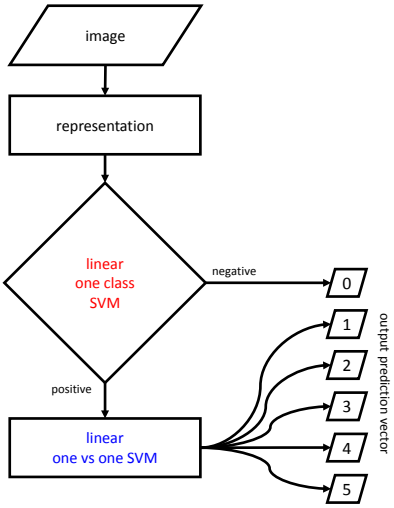


Figure 3. Diagram of the proposed classification pipeline.

or gives similar results to the combination of other kernels (i.e., gaussian, sigmoidal) and normalization schemes (i.e., L1, L2) in preliminary experiments.

For the OCSVM, we consider the method proposed in [28]. Its optimization procedure depends on a single parameter ν which is a lower bound on the fraction of outliers in the training set. In our settings, the training set consists in all the positive samples from the different contexts and hence it does not contain outliers by design. Nevertheless, since the performances of the OCSVM are sensitive to the value of parameter ν, we use the small subset of negative samples available along with the training set, to choose the value of ν which maximizes the accuracy on the training-plus-negatives samples. It should be noted that the negative samples are only used to optimize the value of the ν parameter and they are not used to train the OCSVM.

The multiclass component has been implemented with a multiclass SVM classifier. Its optimization procedure depends only on the cost parameter C. At training time, we choose the value of C which maximizes the accuracy on the training set using cross-validation similarly to what has been done in other works [25, 26].

The outlined training and testing pipeline is applied to different combinations of devices and representations in order to assess the influence of using different devices to acquire the data and different state-of-the-art representations. It should be noted that all the parameters involved in the classification pipeline are computed independently in each experiment in order to yield fair comparisons. We use LibSVM library [29] in all our experiments.

6. Experimental Results and Discussion

In order to assess the performances of each component of the classification pipeline depicted in Figure 3, we report the overall accuracy of the system, as well as the performance measures for the one-class and multi-class components working independently. The overall accuracy of the system (ACC) is computed simply counting the fraction of the input images correctly classified by the cascade pipeline into one of the possible six classes (five contexts, plus the "negative" class). The performances of the OCSVM component, are assessed reporting the True Positive Rate (TPR) and the True Negative Rate (TNR). Since the accuracy of the one-class classifier can be biased by the large number of positive samples (about 5000), versus the small number of negatives (about 1000), we report the average between TPR and TNR, which we refer to as One-class Average Rate (OAR). The performances of the MCSVM are assessed bypassing the OCSVM component and running the MCSVM only on the positive samples of the test set. We report the Multi-Class Accuracy (MCA), i.e., the fraction of samples correctly discriminated into the 5 contexts, and the per-class True Positive Rates. Table 2 reports the results of all the experiments. Each row of the table corresponds to a different experiment and is denoted by a unique identifier in brackets (e.g., $[a_1]$). The GMM used for the IFV representations have been trained on all the descriptors extracted from the training set (excluding the negatives) using the settings specified in the table. The table is organized as follows: the first column reports the unique identifier of the experiment and the used representation; the second column reports the device used to acquire the pair of training and test sets; the third column reports the options of the representation, if any; the fourth column reports the dimensionality of the feature vectors; the fifth column reports the overall accuracy of the cascade (one-class and multi-class classifier) depicted in Figure 3 on the six classes; the sixth column reports the One-Class Average Ratio (OAR) of the OCSVM classifier; the seventh and eighth columns report the TPR and TNR values for the OCSVM; the ninth column reports the accuracy of the MCSVM classifier (MCA) working independently from OCSVM on the five contexts classes. The remaining columns report the true positive rates for the five different personal contexts classes. To improve the readability of the table, the per-column maximum performance indicators among the experiments related to a given device are reported as boxed numbers, while the global per-column maxima are reported as underlined numbers.

In the reported results the performance indicators of the MCSVM are in average better than the ones of the OCSVM. This difference is partly due to the fact that one-class classification is usually "harder" than multi-class classification due to the limited availability of representative counterexamples. Furthermore, it can be noted that many of the considered representations yield inconsistent one-class classifiers characterized by large TPR values and very low TNR values. This effect is in general mitigated when deep features are used, which suggests that better performances could be achieved with suitable representations. Moreover, the performances of the one-class classifier have a large influence on the performances of the overall system, even in

METHOD	DEV.	OPTIONS	DIM.	ACC	OAR	TPR	TNR	MCA	CAR	C.V.M.	OFFICE	TV	H. OFF.
[a_1] GIST	RJ	—	512	**38,96**	**50,52**	91,54	9,50	**49,85**	43,76	90,84	14,20	76,26	46,78
[b_1] IFV	RJ	KS 256	40960	**42,17**	**46,70**	91,20	2,20	**51,25**	62,28	53,82	34,69	98,69	38,37
[c_1] IFV	RJ	KS 512	81920	**42,16**	**46,61**	90,82	2,40	**51,21**	62,21	53,85	34,55	98,90	38,58
[d_1] IFV	RJ	KS SE 256	41984	**43,24**	**45,42**	85,14	5,70	**53,73**	69,08	50,22	34,65	99,11	46,62
[e_1] IFV	RJ	KS SE 512	83968	**36,06**	**45,68**	89,66	1,70	**44,03**	77,80	46,41	29,65	97,00	21,88
[f_1] IFV	RJ	DS 256	40960	**43,77**	**52,35**	93,50	11,20	**52,63**	65,58	49,50	27,98	91,51	86,92
[g_1] IFV	RJ	DS 512	81920	**47,46**	**48,82**	88,74	8,90	**60,33**	84,34	55,51	37,79	78,09	52,10
[h_1] IFV	RJ	DS SE 256	41984	**47,91**	**49,37**	91,74	7,00	**59,83**	78,92	70,49	40,73	66,96	88,15
[i_1] IFV	RJ	DS SE 512	83968	**49,51**	**45,77**	81,34	10,20	**67,51**	83,80	65,75	41,78	78,73	67,77
[j_1] CNN	RJ	AlexNet	4096	**49,26**	**48,17**	67,03	29,30	**79,50**	93,07	97,10	57,25	94,00	62,10
[k_1] CNN	RJ	Places205	4096	**55,19**	**53,02**	80,14	25,90	**78,02**	97,29	98,43	69,69	96,14	50,86
[l_1] CNN	RJ	VGG	4096	**54,54**	**53,78**	63,35	44,20	**85,26**	94,54	89,83	77,10	90,54	73,27
[a_2] GIST	LX2P	—	512	**48,62**	**61,53**	96,56	26,50	**54,15**	74,15	99,81	30,41	82,68	32,02
[b_2] IFV	LX2P	KS 256	40960	**51,19**	**55,68**	79,26	32,10	**70,93**	60,17	98,40	56,65	98,97	55,16
[c_2] IFV	LX2P	KS 512	81920	**63,83**	**54,97**	95,64	14,30	**76,90**	59,84	97,23	68,39	96,92	72,17
[d_2] IFV	LX2P	KS SE 256	41984	**50,66**	**56,43**	79,16	33,70	**69,75**	58,80	98,29	54,87	98,96	53,10
[e_2] IFV	LX2P	KS SE 512	83968	**59,08**	**50,54**	97,48	3,60	**71,99**	58,29	98,03	60,93	98,44	62,11
[f_2] IFV	LX2P	DS 256	40960	**46,62**	**52,10**	88,10	16,10	**61,73**	71,33	75,65	26,08	62,62	56,10
[g_2] IFV	LX2P	DS 512	81920	**50,59**	**52,20**	90,00	14,40	**65,15**	77,70	68,41	31,21	72,75	66,59
[h_2] IFV	LX2P	DS SE 256	41984	**41,79**	**47,22**	80,64	13,80	**57,61**	74,62	76,88	32,42	71,65	39,86
[i_2] IFV	LX2P	DS SE 512	83968	**56,24**	**55,85**	94,00	17,70	**68,29**	77,34	84,29	37,44	88,29	52,78
[j_2] CNN	LX2P	AlexNet	4096	**48,16**	**51,31**	66,31	36,30	**76,10**	80,54	78,98	50,45	100,0	70,66
[k_2] CNN	LX2P	Places205	4096	**54,84**	**57,30**	60,89	53,70	**87,14**	99,19	92,20	63,38	99,88	96,45
[l_2] CNN	LX2P	VGG	4096	**50,74**	**57,40**	56,39	58,40	**86,02**	98,60	81,04	74,11	99,75	80,21
[a_3] GIST	LX2W	—	512	**61,27**	**60,02**	93,66	26,37	**73,91**	87,51	100,0	80,05	83,84	48,29
[b_3] IFV	LX2W	KS 256	40960	**55,47**	**61,89**	89,92	33,87	**67,27**	55,46	99,30	38,77	98,78	61,73
[c_3] IFV	LX2W	KS 512	81920	**54,82**	**63,41**	88,46	38,36	**66,93**	57,55	99,30	40,58	99,26	57,14
[d_3] IFV	LX2W	KS SE 256	41984	**49,73**	**50,08**	88,38	11,79	**66,53**	63,29	99,69	42,45	99,28	47,94
[e_3] IFV	LX2W	KS SE 512	83968	**55,08**	**54,90**	91,52	18,28	**67,95**	53,43	99,80	46,75	100,0	55,86
[f_3] IFV	LX2W	DS 256	40960	**59,62**	**52,77**	94,36	11,19	**72,81**	87,40	95,28	66,94	97,33	48,22
[g_3] IFV	LX2W	DS 512	81920	**60,50**	**52,77**	95,86	9,69	**73,15**	75,52	90,04	73,72	99,81	53,60
[h_3] IFV	LX2W	DS SE 256	41984	**57,88**	**49,01**	87,84	10,19	**74,33**	82,26	93,71	74,33	99,60	51,99
[i_3] IFV	LX2W	DS SE 512	83968	**62,65**	**54,59**	96,40	12,79	**75,74**	69,61	97,51	79,32	98,85	58,93
[j_3] CNN	LX2W	AlexNet	4096	**71,23**	**70,00**	81,46	58,54	**91,34**	99,70	96,23	90,36	99,03	76,50
[k_3] CNN	LX2W	Places205	4096	**61,63**	**63,77**	66,49	61,04	**94,02**	99,90	99,90	93,90	99,65	80,17
[l_3] CNN	LX2W	VGG	4096	**66,02**	**71,91**	69,29	74,53	**94,42**	100,0	99,60	93,79	99,64	81,91
[a_4] GIST	LX3	—	512	**42,08**	**65,23**	77,86	52,59	**53,07**	65,16	95,24	31,91	58,36	26,55
[b_4] IFV	LX3	KS 256	40960	**40,51**	**49,88**	82,50	17,27	**62,07**	67,21	90,19	46,31	99,15	20,47
[c_4] IFV	LX3	KS 512	81920	**40,21**	**47,23**	83,38	11,08	**62,13**	67,33	90,19	46,37	99,15	20,74
[d_4] IFV	LX3	KS SE 256	41984	**41,48**	**47,61**	85,64	9,58	**61,49**	66,07	89,35	47,04	98,87	16,61
[e_4] IFV	LX3	KS SE 512	83968	**40,49**	**51,34**	81,92	20,76	**61,35**	66,19	89,23	45,58	99,15	19,17
[f_4] IFV	LX3	DS 256	40960	**59,07**	**61,20**	93,46	28,94	**68,81**	78,72	83,11	47,00	92,49	73,08
[g_4] IFV	LX3	DS 512	81920	**63,31**	**50,69**	89,50	11,88	**81,92**	90,82	92,61	59,97	99,81	86,23
[h_4] IFV	LX3	DS SE 256	41984	**67,54**	**58,78**	92,32	25,25	**82,70**	88,61	84,00	66,67	99,29	89,39
[i_4] IFV	LX3	DS SE 512	83968	**66,02**	**57,83**	91,80	23,85	**81,08**	91,42	90,55	58,93	99,84	78,23
[j_4] CNN	LX3	AlexNet	4096	**54,49**	**67,42**	75,16	59,68	**76,32**	99,80	99,90	50,85	97,88	20,23
[k_4] CNN	LX3	Places205	4096	**52,87**	**72,01**	55,19	88,82	**86,28**	95,97	98,21	62,36	97,47	99,12
[l_4] CNN	LX3	VGG	4096	**59,12**	**69,68**	74,59	64,77	**80,74**	99,60	100,0	51,31	99,01	77,63

Table 2. The experimental results. For each experiment we report the corresponding device (DEV.), the options (if any), the dimensionality of the feature vector (DIM.), the accuracy of the overall system (ACC), the One-class Average Rate (OAR), the one-class True Positive Rate (TPR) and True Negative Rate (TNR), the Multi-Class accuracy (MCA) and the per-class true positive rates (last five columns). The following legend holds for the options: SE - Spatially Enhanced [26], KS - keypoint-based SIFT feature extraction, DS - multiscale dense-based SIFT feature extraction. Numbers in the OPTION column indicate the number of clusters of the GMM (either 256 or 512). The following abbreviations are used in the last five columns: C.V.M. - coffee vending machine, H. OFF. - home office. The per-column maximum performance indicators among the experiments related to a given device are reported as boxed numbers, while the global per-column maxima are reported as underlined numbers. Results related to the one-class component are reported in red, while results related to the multi-class component are reported in blue.

the presence of excellent MCA values as in the case of $[j_3]$, $[k_3]$ and $[l_3]$. For example, while the $[l_3]$ method reaches an MCA accuracy equal to $94,42\%$ when only discrimination between the five different contexts is considered, it scores a OAR accuracy as low as $71,91\%$ on the one-class classification problem, which results in the overall system accuracy (ACC) of $66,02\%$.

The results related to the MCSVM are more consistent. In particular, the deep features systematically outperform any other representation methods, which suggests that the considered task can take advantage of transfer learning techniques, given the availability of a small amount of labelled data, i.e., we can use models already trained for similar tasks to build the representations. Interestingly, the simple GIST descriptor, gives remarkable performances when used on wide angle images acquired by the LX2W device (i.e., experiment $[a_3]$), where an MCA value of $73,91\%$ is achieved. The different experiments with the IFV-based representations highlight that the keypoint-based extraction scheme (KS) has an advantage over the dense-based (DS) extraction scheme only when the narrow FOV LX2P device is used, while dense-based extraction significantly outperforms the keypoints-based extraction scheme when the field of view is larger, i.e., for the LX2W and LX3 devices. Moreover, when a dense-based extraction scheme is employed, spatially-enhanced descriptors (SE) outperform their non-spatially-enhanced counterparts. The use of larger GMM codebooks (i.e., $K = 512$ clusters) often (but not always, as in the cases of $[e_1]$ vs $[d_1]$ and $[i_4]$ vs $[h_4]$) allows to obtain better performances. However, this come at the cost of dealing with very large representation vectors (in the order of $80K$ vs $40K$ dimensions).

As a general remark, devices characterized by larger FOVs tend to have a significant advantage over the narrow-FOV devices. This is highlighted in Figure 4 which reports the minima, maxima and average ACC values (accuracy of the overall system) for all the experiments related to a given device. These statistics clearly indicate that the LX2W camera is the most appropriate (among the ones we tested) for modelling the personal contexts of the user. The success of such camera is probably due to the combination of the large FOV and the wearing modality, which allows to gather the data from a point of view very alike to the one of the user. Indeed, the LX3 camera, which has a similar FOV, but is worn differently, achieve the top-2 average and maximum results.

We conclude our analysis reporting the confusion matrices (Figure 5) and some success/failure examples (Figure 6) for the best performing methods with respect to the four considered devices. These are: $[k_1]$ CNN Places205 for the RJ device, $[c_2]$ IFV KS 512 for the LX2P device, $[j_3]$ CNN AlexNet for the LX2W device and $[h_4]$ IFV DS SE 256 for the LX3 device. The confusion matrices reported in Figure 5 show that the most part of the error is introduced by the negatives, while there is usually less confusion among

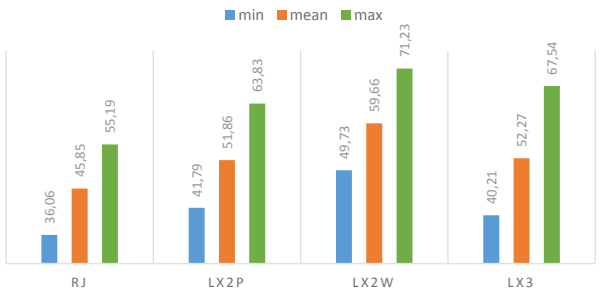
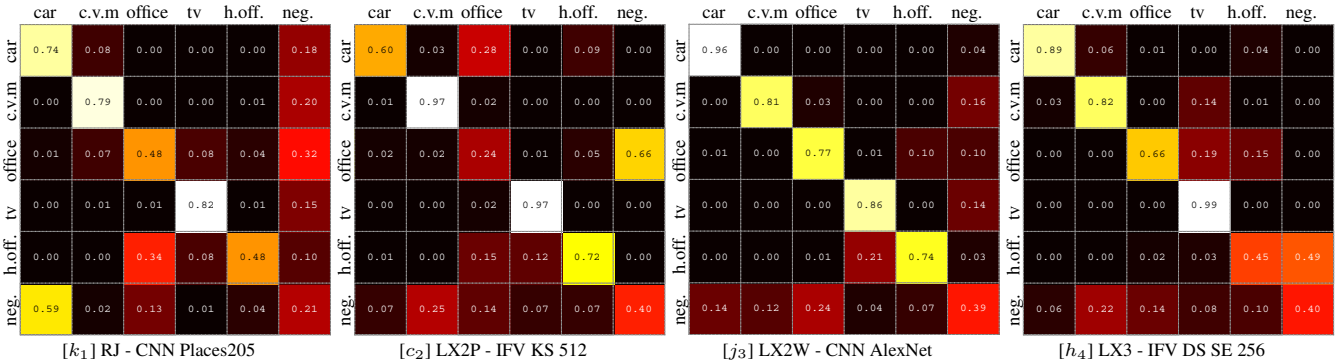


Figure 4. Minimum, average and maximum accuracies per device. As can be noted, all the statistics are higher for the LX2W-related experiments. This suggests that the task of recognizing personal contexts is easier for images acquired using such device.

the 5 contexts, especially in the case of $[j_3]$. This confirms our earlier considerations on the influence on the whole system of the low performances of the one-class component used for the rejection of contexts not of interest for the user. It should be noted that a rejection mechanism (implemented in our case by the one-class component) is crucial for building effective systems, not only able to discriminate among a small set of known contexts, but also able to reject outliers and that building such component can usually rely only on a small number of positive samples with few or no representative negative examples. Moreover, there is usually some degree of confusion between the office, home office and TV classes. This is not surprising, since all these classes are characterized by the presence of similar objects (e.g., a screen) and by similar user-context interaction paradigms. Such considerations suggest that discrimination among similar contexts should be considered as a fine-grade problem and that the considered task could probably benefit from coarse-to-fine classification paradigms. All the considerations above are more evident looking at the samples reported in Figure 6.

7. Conclusion and Further Works

We have studied the problem of recognizing personal contexts from egocentric images. To this aim, we have acquired a dataset of five personalized contexts using four different devices. We have proposed a benchmark evaluation pipeline and we have assessed the performances of many state-of-the-art representations on the considered task with respect to the different devices used to acquire the data. The results show that, while the discrimination among a limited number of personal contexts is an easier task, detecting the negative samples still requires some efforts. The best results have been achieved considering deep representations and a wide angular, ear mounted wearable camera. This suggests that the considered task can effectively take advantage of the transfer learning properties of CNNs and that wide FOV, head mounted cameras are the most appropriate to model the user's personal contexts. Moreover, despite the good performances of the discriminative component, there is still some degree of confusion among personal contexts



[k_1] RJ - CNN Places205 [c_2] LX2P - IFV KS 512 [j_3] LX2W - CNN AlexNet [h_4] LX3 - IFV DS SE 256

Figure 5. Confusion matrices of the four the best performing methods on the considered devices. Columns represent the ground truth classes, while rows represent the predicted labels. The original confusion matrices have been row-normalized (i.e., each value has been divided by the sum of all the values in the same row) so that each element on the diagonal represents the per-class True Positive Rate. Each matrix is related to the row of Table 2 specified by the identifier in brackets. The following abbreviations are used: c.v.m - coffee vending machine, h.off - home office, neg. - negatives.

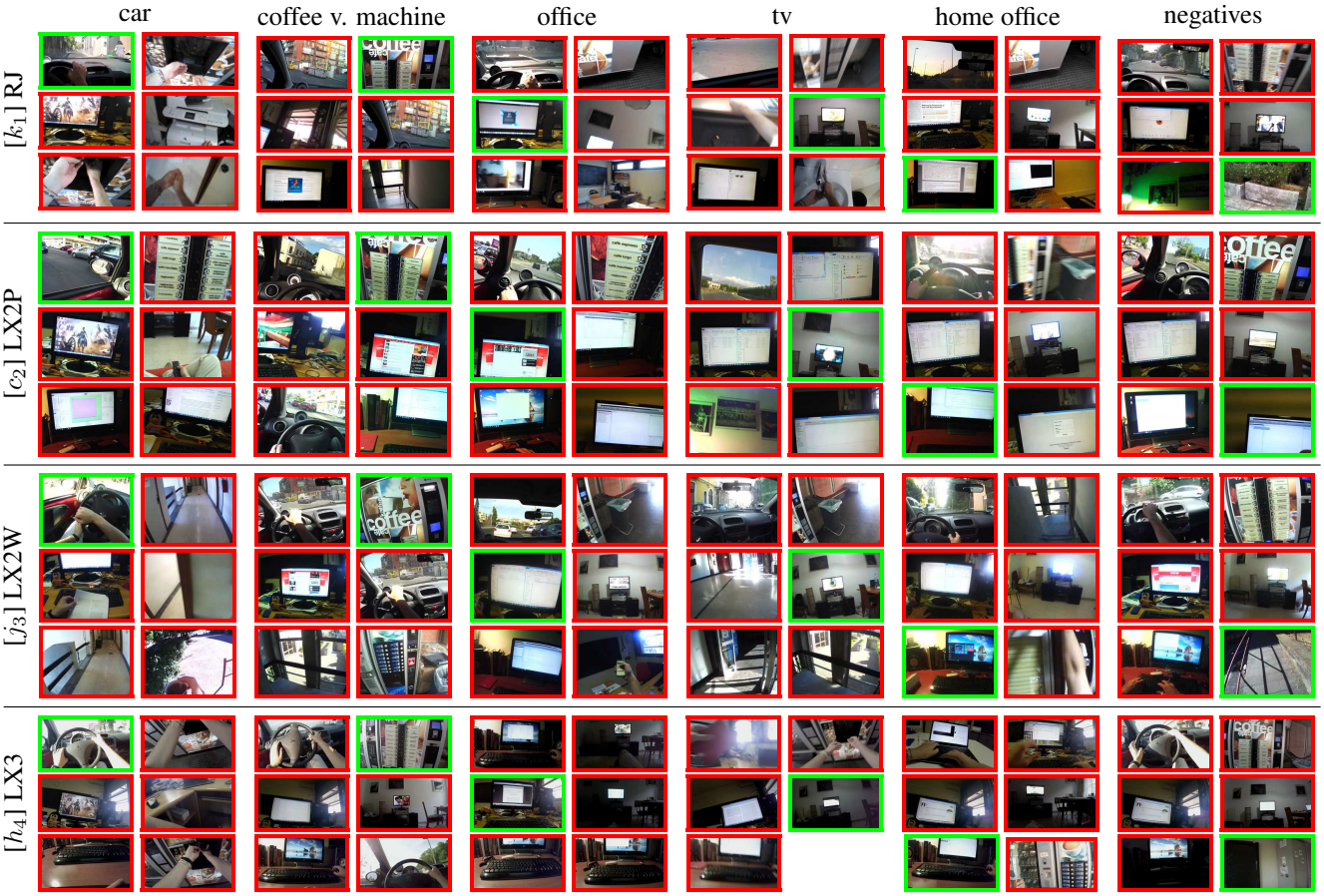


Figure 6. Some success (green) and failure (red) examples and according to the best performing methods on the four considered devices. Samples belonging to the same class are grouped by columns, while samples related to the same method are grouped by rows.

belonging to the same, or similar categories (e.g., office, home office, tv). This suggests that better performances could be achieved fine-tuning the CNN-based representation to the required instance-level granularity. Future works will be devoted to overcome the limitations of the present study by providing larger datasets also acquired by multiple users and better exploring deep representations. Moreover, the spatio-temporally coherence between neighbouring frames could be leveraged to provide meaningful representations (e.g., by exploiting the 3D structure of the scene) and to improve the classification results by disambiguating the predictions for neighbouring frames. Finally, more attention should be devoted to outlier-rejection mechanisms in order build effective and robust systems.

References

[1] A. Betancourt, P. Morerio, C.S. Regazzoni, and M. Rauterberg. The evolution of first person vision methods: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(5):744–760, 2015.

[2] T. Starner, B. Schiele, and A. Pentland. Visual contextual awareness in wearable computing. In *International Symposium on Wearable Computing*, pages 50–57, 1998.

[3] U. Blanke and B. Schiele. Daily routine recognition through activity spotting. In *Location and Context Awareness*, pages 192–206. 2009.

[4] D. Damen, O. Haines, T. Leelasawassuk, A. Calway, and W. Mayol-Cuevas. Multi-user egocentric online system for unsupervised assistance on object usage. In *Workshop on Assistive Computer Vision and Robotics, in Conjunction with ECCV*, pages 481–492, 2014.

[5] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa. Predicting daily activities from egocentric images using deep learning. *International Symposium on Wearable Computing*, 2015.

[6] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1371–1375, 1998.

[7] H. Aoki, B. Schiele, and A. Pentland. Recognizing personal location from video. In *Workshop on Perceptual User Interfaces*, pages 79–82, 1998.

[8] R. Templeman, M. Korayem, D. Crandall, and K. Apu. PlaceAvoider: Steering First-Person Cameras away from Sensitive Spaces. In *Annual Network and Distributed System Security Symposium*, pages 23–26, 2014.

[9] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *Computer Vision and Pattern Recognition*, pages 2537–2544, 2014.

[10] Y. Lee and K. Grauman. Predicting important objects for egocentric video summarization. *International Journal of Computer Vision*, 114(1):38–55, 2015.

[11] A. Ortis, G. M. Farinella, V. D'Amico, L. Addesso, G. Torrisi, and S. Battiato. RECfusion: Automatic video curation driven by visual content popularity. In *ACM Multimedia*, 2015.

[12] Jindong L., E. Johns, L. Atallah, C. Pettitt, B. Lo, G. Frost, and Guang-Zhong Y. An intelligent food-intake monitoring system using wearable sensors. In *Wearable and Implantable Body Sensor Networks*, pages 154–160, 2012.

[13] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, volume 7572, pages 314–327, 2012.

[14] J. Hernandez, Yin L., J. M. Rehg, and R. W. Picard. Bioglass: Physiological parameter estimation using a head-mounted wearable device. In *Wireless Mobile Communication and Healthcare*, 2014.

[15] M. L. Lee and A. K. Dey. Capture & Access Lifelogging Assistive Technology for People with Episodic Memory Impairment Non-technical Solutions. In *Workshop on Intelligent Systems for Assisted Cognition*, pages 1–9, 2007.

[16] P. Wu, H. Peng, J. Zhu, and Y. Zhang. Senscare: Semi-automatic activity summarization system for elderly care. In *Mobile Computing, Applications, and Services*, pages 1–19. 2012.

[17] A. Torralba and A. Oliva. Semantic organization of scenes using discriminant structural templates. *International Conference on Computer Vision*, 2:1253–1258, 1999.

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[19] G. M. Farinella and S. Battiato. Scene classification in compressed and constrained domain. *IET Computer Vision*, (5):320–334, 2011.

[20] G. M. Farinella, D. Ravì, V. Tomaselli, M. Guarnera, and S. Battiato. Representing scenes for real–time context classification on mobile devices. *Pattern Recognition*, 48(4):1086–1100, 2015.

[21] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

[22] N. Jojic, A. Perina, and V. Murino. Structural epitome: a way to summarize one's visual experience. In *Advances in neural information processing systems*, pages 1027–1035, 2010.

[23] A. Furnari, G. M. Farinella, G. Puglisi, A. R. Bruna, and S. Battiato. Affine region detectors on the fisheye domain. In *International Conference on Image Processing*, pages 5681–5685, 2014.

[24] A. Furnari, G. M. Farinella, A. R. Bruna, and S. Battiato. Generalized sobel filters for gradient estimation of distorted images. In *International Conference on Image Processing (ICIP)*, 2015.

[25] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, volume 2, page 8, 2011.

[26] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[28] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[29] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.