Recognizing Personal Locations from Egocentric Videos

Antonino Furnari, Giovanni Maria Farinella, *Senior Member, IEEE,* and Sebastiano Battiato, *Senior Member, IEEE*

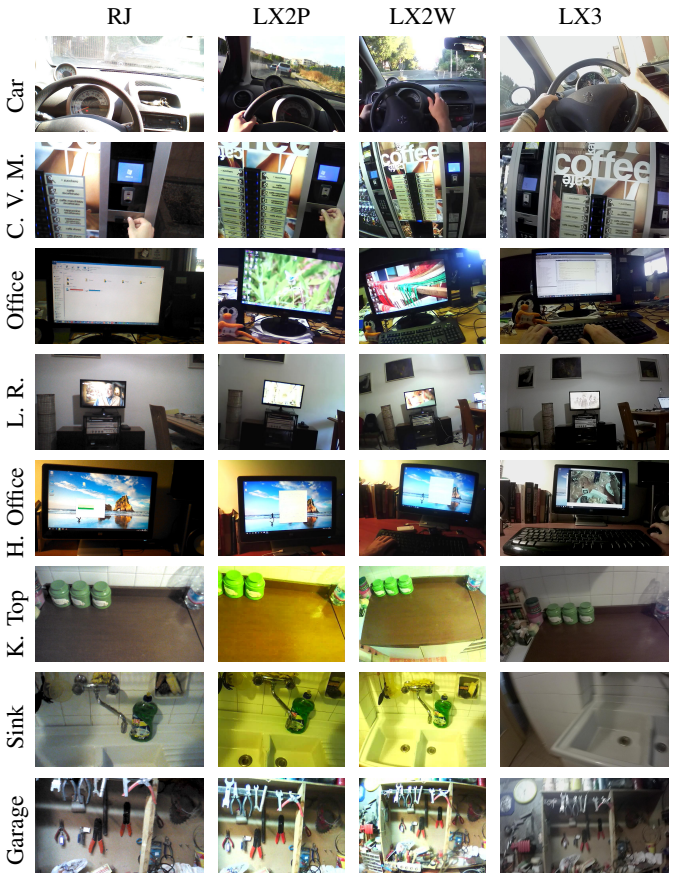
***Abstract*—Contextual awareness in wearable computing allows for construction of intelligent systems which are able to interact with the user in a more natural way. In this paper, we study how personal locations arising from the user's daily activities can be recognized from egocentric videos. We assume that few training samples are available for learning purposes. Considering the diversity of the devices available on the market, we introduce a benchmark dataset containing egocentric videos of 8 personal locations acquired by a user with 4 different wearable cameras. To make our analysis useful in real-world scenarios, we propose a method to reject negative locations, i.e., those not belonging to any of the categories of interest for the end-user. We assess the performances of the main state-of-the-art representations for scene and object classification on the considered task, as well as the influence of device-specific factors such as the Field of View (FOV) and the wearing modality. Concerning the different device-specific factors, experiments revealed that the best results are obtained using a head-mounted, wide-angular device. Our analysis shows the effectiveness of using representations based on Convolutional Neural Networks (CNN), employing basic transfer learning techniques and an entropy-based rejection algorithm.**

***Index Terms*—egocentric vision, first person vision, context-aware computing, egocentric dataset, personal location recognition**

I. Introduction and Motivation

Contextual awareness is a desirable property in mobile and wearable computing [1], [2]. Context-aware systems can leverage the knowledge of the user's context to provide a more natural behavior and a richer human-machine interaction. Although different factors contribute to define the context in which the user operates, two important aspects seem to emerge from past research [2], [3]: 1) context is a dynamic construct and hence it is usually infeasible to enumerate a set of canonical contextual states independently from the user or the application, 2) even if context cannot be simply reduced to location, the latter still plays an important role in the definition and understanding of the user's context. In particular, we argue that being able to recognize the locations in which the user performs his daily activities at the instance level (i.e., recognizing a particular environment such as "my office"), rather than at the category-level, (e.g., "an office"), provides important cues on the user's current objectives and can help improving human-machine interaction.

First person vision systems offer interesting opportunities for understanding the behavior, intent, and environment of a person [4]. Moreover, wearable cameras are becoming more and more used in real-world scenarios. Therefore, we find of particular interest to exploit such systems for contextual sensing and location recognition. It should be noted that, while outdoor location recognition can be trivially addressed using GPS sensors, most daily activities are performed indoor, where GPS devices usually fail. Considering their acquisition modality, egocentric videos introduce some intrinsic challenges [4], [5] which must be taken into account. Such challenges are primarily related to the non-stationary nature of the camera, the non-intentionality of the framing, the presence of occlusions (often by the user's hands), as well as the influence of varying



(a) positive samples

(a) negative samples

Fig. 1. (a) Some sample images of eight personal locations acquired using different wearable cameras. (b) Some negative samples used for testing purposes. The employed devices are discussed in Section IV. The dataset is publicly available at the URL http://iplab.dmi.unict.it/PersonalLocations/. The following abbreviations are used: C.V.M. - Coffee Vending Machine, L. R. - Living Room, H. Office - Home Office, K. Top - Kitchen Top.

The authors are with the Department of Mathematics and Computer Science, University of Catania, Italy e-mails: {furnari, gfarinella, battiato}@dmi.unict.it.

Manuscript received XX XX, XXXX; revised XX XX, XXXX.

lighting conditions, fast camera movements and motion blur. Fig. 1 shows some examples of the typical variability exhibited by egocentric images.

In this paper, we study the problem of recognizing personal locations of interest from egocentric videos. Following the definition in [6], a personal location is considered 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 considered location.*

A simple example of personal location may be the personal office desk, in which the user can perform a number of activities, such as surfing the Web or writing e-mails. It should be noted that, according to the above definition, a personal location is defined (and hence should be recognized) independently from the actions which could be performed by the user. Moreover, a given set of personal locations is meaningful just for a single user and hence user-specific models have to be taken into account when designing this kind of location-aware systems. To clarify the concept of personal location of interest, we consider the scenario of a user wearing an always-on wearable camera. The user can specify a set of personal locations of interest (not known in advance by the system) which he wishes to monitor (e.g., in order to highlight them in the huge quantity of video acquired within a few days or to trigger specific behaviors or alerts). To do so, we suppose that in a real scenario the user wearing the camera records only a short video (≈ 10 seconds) of each of the environments he wants to monitor. Relying on the acquired set of user-specified data, at run time the system should be able to: 1) detect the considered locations and 2) reject negative frames (i.e., frames not depicting any of the locations interesting for the user).

Considering possible real scenarios as above, in addition to the general issues associated with egocentric data, recognizing personal locations of interest involves some unique challenges:

- real-world location detection systems must be able to correctly detect and manage negative samples, i.e., images depicting scenes not belonging to any of the considered personal locations;
- given that an always-on wearable camera is likely to acquire a great variability of different scenes, gathering representative negative samples for modeling purposes is not always possible. In a real scenario, a system able to reject negatives given only user-specific positive samples for learning is hence desirable;
- since personal locations are user-specific, few labeled samples are generally available as it is not feasible to ask the user to collect and annotate huge amounts of data for learning purposes;
- large intra-class variability usually characterizes the appearance of the different views related to a given location of interest;
- personal locations belonging to the same high level category (e.g., two different offices) tend to be characterized by similar appearance, making the discrimination challenging.

Even if previous works already investigated the possibility of recognizing known locations for different purposes [1], [7], [8], [9], [10], a solid investigation on the problem and a benchmark of state-of-the-art representation methods are still missing. We perform a benchmark of different state-of-the-art methods for scene and object classification on the task of recognizing personal locations from egocentric images. To this aim, we built a dataset of egocentric videos containing eight locations acquired by a user over six months. To assess the influence of device-specific factors, such as the wearing modality and the Field Of View (FOV), the data has been acquired multiple times using four different devices. Fig. 1 shows some examples of the acquired data. In order to make the analysis worth in real-world scenarios where personal locations of interest need to be discriminated from negative samples, we propose a classification method which includes a mechanism for the rejection of negative samples. We compare the proposed approach against a baseline method based on the combination of a multi-class SVM classifier and a standard one-class classifier which has been used in [6]. To deal with the problem addressed in this paper, experiments are carried out by training and testing the considered methods on data arising from different combinations of devices and representation techniques.

The remainder of the paper is organized as follows. Section II revises the related work. Section III introduces the proposed method and discusses the reference baseline classification method proposed in [6]. Section IV describes the wearable cameras used to acquire the data and introduces the proposed egocentric dataset of personal locations. Section V summarizes the state-of-the art representation techniques considered in the experimental analysis. Experiments are defined in Section VI, whereas results are discussed in Section VII. Section VIII finally concludes the paper.

II. Related Work

Mobile and wearable cameras have been widely used in a variety of tasks, such as place and action recognition [1], [7], health and food intake monitoring [11], [12], [13], human-activity recognition and understanding [14], [15], [16], [17], [18], [19], video indexing and summarization [20], [21], [22], as well as assistive-related technologies [23], [24]. The problem of recognizing personal locations from egocentric images, in particular, has already been investigated for different purposes and different methods have been proposed in the literature. The first investigations relevant to the considered problem date back to the late 90s. Starner et al. [1] proposed a context-aware system for assisting the users while playing the "patrol" game. The system proposed in [1] comprises a component able to recognize the room in which the player is operating combining RGB features and a Hidden Markov Model (HMM). Aoki et al. [7] proposed an image matching technique for the recognition of previously visited places. In this case, locations are not represented by a single frame, but rather by an image sequence of the approaching trajectory. Place recognition is implemented by computing the distance between a newly recorded trajectory and a dictionary of trajectories to known places. Torralba et al. [8] proposed

a wearable system able to recognize familiar locations as well as categorize new environments. A low-dimension global representation based on a wavelet image decomposition is proposed in order to include textural properties of the image as well as their spatial layout. Familiar location recognition and new environment categorization are obtained separately training two distinct HMM models. More recently, in the wake of the popularity that always-on wearable cameras have recently gained, Templeman et al. [9] have proposed a system for "blacklisting" sensitive spaces (like bathrooms and bedrooms) to protect the privacy of the user when passively acquiring images of the environment. The system combines contextual information like GPS location and time with an image classifier based on local and global features and a HMM to take advantage of the temporal constraint on human motion. In [10], CNNs and HMM are combined to temporally segment egocentric videos in order to highlight locations important for the user. Images and short-video-based localization strategies have been already investigated in [25], where short videos are used to compute 3D-to-3D correspondences. The authors of [26] propose to model and recognize activity-related locations of interest to facilitate navigation in a visual lifelog. While the discussed approaches generally concentrate on video, some researchers have also investigated the use of low temporal-resolution devices. Such devices generally allow to acquire a few images per minute, but are characterized by a larger autonomy both in terms of memory and battery-life, which makes them particularly suited to acquire large amounts of visual data. In [18], daily activities are recognized from static images within a low temporal-resolution lifelog. In [27], a method for semantic indexing and segmentation of photo streams is proposed. The reader is referred to the work by Bolaños et al. [28] for a review of the advances in egocentric data analysis.

As highlighted in [8], location recognition and place categorization are two related tasks and hence they are likely to share similar features in real-world applications. In this regard, much work has been devoted to designing suitable image representation for place categorization. Torralba and Oliva described a procedure for organizing real world scenes along semantic axes in [29], while in [30] they proposed a computational model for classifying real world scenes. Efficient computational methods for scene categorization have been proposed for mobile and embedded devices by Farinella et al. [31], [32]. More recently, Zhou et al. [33] have successfully applied Convolutional Neural Networks (CNNs) to the problem of scene classification.

Rather than sticking to a specific framework, in this work we aim at systematically studying the performances of the state-of-the-art methods for scene and object representation and classification on the considered task of personal locations recognition. Furthermore, while past literature primarily focused on classification, we pay special attention to the negative-rejection mechanism which is an essential component when building real, robust and effective systems. To make our analysis broader, we assess the influence of device-specific factors such as the wearing modality and the FOV on the performances of the considered methods and provide a dataset of egocentric videos depicting eight different locations to facilitate further research on the topic. Throughout the study, we assume that only visual information is available and that the quantity of training data is limited, according to the assumptions discussed in Section I.

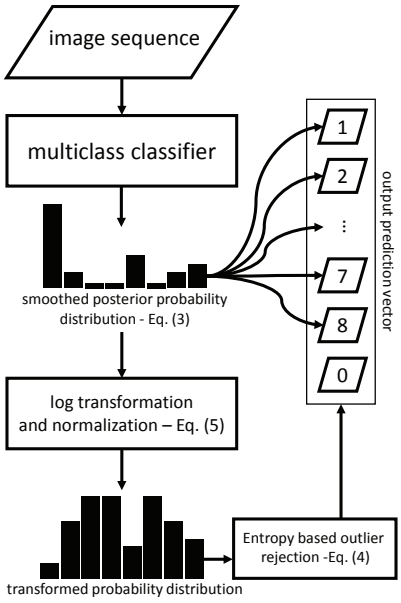


Fig. 2. The proposed classification pipeline combining a multi-class classifier and an entropy-based negative rejection method.

III. Recognizing Personal Locations from Egocentric Videos

A personal location recognition system should be able to: 1) discriminate among different personal locations specified by the user, and 2) reject negative frames, i.e., frames not related to any of the considered locations. Hence, we propose a classification pipeline made up of two main components: 1) a multi-class location classifier, and 2) a mechanism for rejecting negative samples. While the multi-class component can be implemented using standard supervised learning techniques (e.g., an SVM classifier or a fine-tuned Convolutional Neural Network), negative rejection does not always have a straightforward implementation. We propose an entropy-based negative rejection mechanism which leverages the temporal coherence of class predictions within a small temporal window. The input to our system is a small sequence of neighboring frames. For each frame, the multi-class classifier estimates a posterior probability distribution on the considered personal locations. Posterior probabilities are hence smoothed to perform multi-class classification on the input sequence. The input sequence is either classified as a given location or rejected depending on how much the different predictions agree. The proposed method is depicted in Fig. 2 and detailed in the following.

We assume that very close frames in an egocentric video (e.g., less than 0.5 seconds apart) share the same class. This assumption is of course imprecise whenever there is a transition from a given location to another. This phenomenon however mostly affects the accuracy related to the localization of the

exact transition frame between two different locations and it does not impact much (in average) the overall recognition performances. According to this assumption, n subsequent observations $x_1, \ldots, x_n$ share the same class c. This implies the conditional independence between the observations given class c:

$$x_i \perp\!\!\!\perp x_j | c, \ \forall i, j \in \{1, 2, \ldots, n\}. \tag{1}$$

Given the property reported in Eq. (1), the posterior probability $p(c_k|x_1, \ldots, x_n)$ for the generic class c_k, can be expressed as:

$$p(c_k|x_1, \cdots, x_n) = \frac{p(x_1, \cdots, x_n|c_k) \cdot p(c_k)}{p(x_1, \cdots, x_n)} =$$
$$= \prod_{1 \leq i \leq n} p(c_k|x_i) \frac{p(c_k)^{1-n} \prod_{1 \leq i \leq n} p(x_i)}{p(x_1, \cdots, x_n)}. \tag{2}$$

If we assume that all the considered locations of interest have equal probabilities $p(c_k) = \frac{1}{K}, \forall k \in \{1, \cdots, K\}$ (with K being the total number of classes), then Eq. (2) simplifies to:

$$p(c_k|x_1, \cdots, x_n) = \frac{\prod_i p(c_k|x_i)}{\sum_k \prod_i p(c_k|x_i)} \tag{3}$$

where $p(c|x_i)$ denotes the posterior probability distribution on class c estimated by the multi-class classifier, given observation x_i.

Equation (3) is used to smooth the predictions of the multi-class classifier on multiple, contiguous frames of the input sequence for which we assume conditional independence as reported in Eq (1). The predicted class for the input sequence is determined as the one which maximizes the probability reported in Eq. (3). When the samples are positive and hence they belong to a given class, we expect Eq. (3) to produce a resulting posterior distribution which strongly agrees on the identity of the considered samples. On the contrary, when the sequence contains negative samples, we expect the resulting posterior distribution to exhibit a high degree of uncertainty. We propose to measure the uncertainty of the distribution reported in Eq. (3) (i.e., entropy) to quantify the "outlierness" of the considered samples. Given a posterior distribution p, we measure the uncertainty as the entropy:

$$e(p; x_1, \cdots, x_n) = -\sum_k p(c_k|x_1, ..., x_k) log\big(p(c_k|x_1, ..., x_k)\big). \tag{4}$$

The entropy reported in Eq. (4) can be used to discriminate negative sequences (i.e., locations not of interest for the user) from positive ones using a threshold t_e. Sequences are classified as negative if $e(p; x_1, \cdots, x_n) > t_e$, while they are classified as positive if $e(p; x_1, \cdots, x_n) \leq t_e$. The optimal threshold t_e can be selected as the one which of best separates the training set from a small number of negatives used for optimization purposes.

In practice, instead of measuring the uncertainty directly from the distribution reported in Eq. (3), we log-transform the original distribution p as follows:

$$\tilde{p}(c_k|x_1, \cdots, x_k) = \frac{log(p(c_k|x_1, \cdots, x_k))}{\sum_k log(p(c_k|x_1, \cdots, x_k))}. \tag{5}$$

The proposed transformation has the effect of "inverting" the degree of uncertainty carried by the distribution. Therefore, negative samples will be characterized by a high $e(p; x_1, \cdots, x_n)$ value and a low $e(\tilde{p}; x_1, \cdots, x_n)$ value. In Section VI-B1, we show that working with the log-transformed distribution shown in Eq. (5), allows to compute the separation threshold t_e from the training/optimization-negatives set in a more robust way.

Please note that the maximum length n of the input sequence in our system should be carefully selected. Indeed, too small values would cause the rejection mechanism to fail for lack of data, while excessively large values would break the assumption reported in Eq. (1) and would greatly affect the localization of the transition frame between two different locations.

IV. Wearable Cameras and Proposed Datasets

The market proposes different wearable cameras, each with its distinctive features. We identify three main factors characterizing such devices: resolution, wearing modality and Field Of View (FOV). The resolution influences the amount of details that a given device is able to capture. While the first generation of wearable devices was characterized by very small resolutions (in the order of 0.1 megapixels), recent devices tend to adhere to the HD and 4K standards. The wearing modality influences the way in which the visual information is actually acquired. In particular, we identify three classes of devices characterized by different wearing modalities: smart glasses, ear mounted cameras and chest mounted cameras. Smart glasses are designed to substitute the user's glasses. Ear mounted cameras are worn similarly to bluetooth earphones and are a little more obtrusive than smart glasses. Both smart glasses and ear mounted devices have the advantage to capture the environment from the user's point of view. Chest mounted cameras are the least obtrusive since they are clipped to the user's clothes rather than mounted on his head (and easily ignored by both the wearer and the people he interacts with). However, the FOV captured by chest mounted cameras does not usually achieve much overlap with the user's FOV. The Field Of View affects the quantity of visual information which is acquired by the device. A larger FOV allows to acquire more information in a similar way to the human visual system at the cost of the introduction of radial distortion, which in some cases requires dedicated processing techniques [34], [35].

In order to assess the influence of the aforementioned device-specific factors for the problem of personal location recognition, we consider four different devices: the smart glasses Recon Jet (RJ)[1], two ear-mounted Looxcie LX2[2], and a wide-angular chest-mounted Looxcie LX3[3]. The Recon Jet and Looxcie LX3 devices produce images at the HD resolution (1280×720 pixels), while the Looxcie LX2 devices have a smaller resolution of 640×480 pixels. The Recon Jet and the Looxcie LX2 devices are characterized by narrow FOVs (70° and $65,5^\circ$ respectively), while the FOV of the Looxcie LX3

[1] http://www.reconinstruments.com/products/jet/
[2] http://www.looxcie.com
[3] http://www.looxcie.com

TABLE I
A SUMMARY OF THE MAIN FEATURES OF THE CONSIDERED DEVICES.

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

is considerably larger ($100°$). One of the two ear-mounted Looxcie LX2 is equipped with a wide-angular converter in order to achieve a large FOV (approximatively $100°$). The wide-angular LX2 camera will be indicated with the acronym LX2W, while the regular (perspective) LX2 camera will be indicated as LX2P. The user is referred to Fig. 1 (a) to assess the differences between similar scenes acquired by the different devices. TABLE I summarizes the main features of the cameras used to acquire the data.

We propose a dataset of egocentric videos acquired by a user in eight different locations using the aforementioned four devices. The dataset has been acquired over six months. The considered personal locations arise from some possible daily activities of a user: Car, Coffee Vending Machine (C. V. M.), Office, Living Room (L. R.), Home Office (H. Office), Kitchen Top (K. Top), Sink, Garage. The considered locations are examples of a possible set of locations which the user may choose. The proposed set of locations has been chosen in order to be challenging and include similar looking locations (e.g., Office vs Home Office) and locations characterized by large intra-class variability (e.g., Garage). Fig. 1 (a) shows some sample frames belonging to the dataset.

Since the considered locations involve static position, we assume that the user is free to turn his head and/or move his body, but he does not change his position in the room. In order to enable fair comparison between the different devices, we built four variants of the dataset. Each variant is an independent, yet compliant, device-specific dataset and comprises its own training and test sets. The training sets include short videos (≈ 10 seconds) of the personal locations of interest. During the acquisition of the training videos, the user turns his head (or chest, in the case of chest-mounted devices) in order to cover the views of the environment he wants to monitor. A single video-shot per location of interest is included in each training set. The test sets contain medium length videos (5 to 10 minutes) acquired by the user in the considered locations while performing regular activities. Each test set comprises 5 videos for each location. In order to gather likely negative samples, we acquired several short videos not representing any of the locations under analysis. The negative videos comprise indoor, outdoor scenes, other desks and other vending machines. The negative videos are divided into two separate sets: test negatives and "optimization" negatives. The role of the latter set of negative samples is to provide an independent set of data useful to optimize the parameters of the negative rejection methods. Some frames from the negative sequences are shown in Fig. 1 (b). The overall dataset amounts to more than 20 hours of video and more than one million frames in total.

In order to facilitate the analysis of such a huge quantity of collected data, we extract each frame in the training videos and temporally subsample the testing videos. To reduce the amount of frames to be processed, for each location in the test sets, we extract 200 subsequences of 15 contiguous frames. This sub-sampling still allows to consider temporal coherence. The starting frames of the subsequences are uniformly sampled from the 5 videos available for each class. The same subsampling strategy is applied to the test negatives. We also extract 300 frames form the optimization negative videos. This amounts to a total of 133770 extracted frames to be used for experimental purposes. The extracted frames are publicly available for download at the URL http://iplab.dmi.unict.it/PersonalLocations/, while the access to full-length videos can be required from the same web page.

V. FEATURE REPRESENTATIONS

To benchmark the proposed method, we consider the main feature representations used for the tasks of scene and object classification. These can be grouped into three categories: holistic, shallow and deep representations.

A. Holistic Feature Representations

Holistic feature representations have been widely used in tasks related to scene understanding [30], [32]. As a popular representative of this class, we consider the GIST descriptor proposed in [30] and use the standard implementation and parameters provided by the authors. According to the standard implementation, all input images are resized to the normalized resolution of 128×128 pixels prior to computing the descriptor. In this configuration, the output GIST descriptors have dimensionality $d = 512$.

B. Shallow Feature Representations

With deep feature representations and Convolutional Neural Networks (CNNs) becoming mainstream in the computer vision literature, classic representation schemes based on the encoding of local features (e.g., Bag of Visual Word models) have been recently referred to as shallow feature representations [36]. The term "shallow" is used to highlight that features are not extracted hierarchically as in deep learning models. Among the different Bag of Visual Word models, we consider Improved Fisher Vectors (IFV) [37] to encode densely-sampled SIFT features.

The IFV features are extracted following the procedures described in [38], [36]. To make computation tractable on a large number of frames, each input image is resized to a normalized height of 300 pixels keeping the original aspect ratio. This produces images of resolutions 400×300 pixels and 533×300 pixels in our dataset. Apart from the standard SIFT descriptors, we also consider the spatially-enhanced local descriptors discussed in [36]. Such descriptors are obtained concatenating the coordinates of the location from which the SIFT descriptor is extracted to the PCA-reduced SIFT features,

obtaining a 82-dimensional vector as detailed in [36]. In our experiments we consider Gaussian Mixture Model (GMM) with $K = 256$ and $K = 512$ centroids. The dimensionality d of IFV descriptors depends on the number of clusters K of the GMM codebook and the number of dimensions D of the local feature descriptors (i.e., SIFT) according to the formula: $d = 2KD$. Using the aforementioned parameters, the number of dimensions of our IFV representations ranges from a minimum of 40960 to a maximum of 83968 components. The VLFeat library [39] has been used to perform all the operations involved in the computation of the IFV representations.

C. Deep Feature Representations

One of the main advantages of CNNs is given by their excellent transfer learning properties. These allow to "reuse" a feature representation learned for a given task in a slightly different one, providing that enough new data is available. In our experiments, we consider two transfer learning approaches: extracting the feature representation contained in the penultimate layer of the network and reusing it in a classifier (e.g., SVM), and fine-tuning the pre-trained network with new data and labels. We consider two popular architectures of convolutional neural networks: AlexNet [40] and VGG16 [41]. Such models have been pre-trained by their authors on the ImageNet dataset [42] to discriminate among 1000 object categories. We also consider two models proposed by Zhou et al. [33], who train the same CNN architectures (AlexNet and VGG16) on the Places205 dataset, which contains images from 205 different place categories. Considering four different models allow us to assess the influence of both the network architectures (AlexNet and VGG16) and the original training data (ImageNet and Places205) in our transfer learning experiments.

1) Reuse of pre-trained CNNs: We obtain the deep feature representations extracting the values contained in the penultimate layer of the network when the input image, appropriately rescaled to the dimensions of the data layer, is propagated into the network. Such feature representation is the one contained in the hidden layer of the multilayer perceptron in the terminal part of the network. For all the considered CNN models, these representations are compact 4096-dimensional vectors.

2) Fine-tuning of pre-trained CNNs: The pre-trained network is fine-tuned using the data contained in the training set. Fine-tuning is performed substituting the last layer of the network (the one carrying the final probabilities) with a new layer containing 8 units (one per each personal location to be recognized) which is initialized with random weights. The training set is divided into two parts: 85% for training and 15% for validation. Optimization of the network is resumed starting from the pre-trained weights. We set a larger learning rate for the randomly initialized layer, and a smaller learning rate for pre-learned layers. The training procedure is stopped when a high validation accuracy is reached or when it is not able to grow any more and the model with maximum validation accuracy is selected. In this case the networks are not used to explicitly extract the representation but directly to predict posterior probabilities.

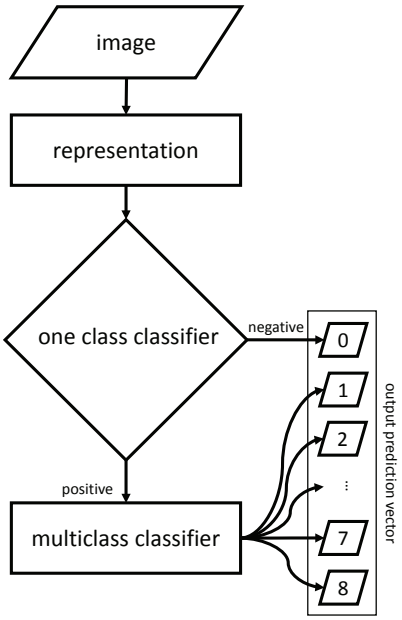


Fig. 3. The baseline classification pipeline proposed in [6] and considered for comparison.

VI. Experiments

The experiments aim at assessing the performances of the state-of-the-art representations discussed in Section V on the considered task of personal location recognition. For all the experiments, we consider the classification pipeline proposed in Section III. We also consider the baseline proposed in [6], where personal location classification is carried on single images. The method in [6] performs negative sample rejection using a one-class classifier learned only on positive samples provided by the user (i.e., the locations of interest). All non-rejected samples are then fed to a multi-class classifier which discriminates among the considered locations. This baseline method is illustrated in Fig. 3.

All experiments are performed considering different device-representation combinations. The considered classification pipelines and all related parameters are independently trained and tested on the training/testing sets related to the different devices. In the following, we discuss the experiments designed to assess the performances of the considered feature representations with respect to 1) the overall location recognition system, 2) the negative rejection mechanism alone, and 3) the multi-class classifier alone.

A. Overall Personal Location Recognition System

The performances of the overall system are assessed considering the two classification pipelines depicted in Fig. 2 and Fig. 3. When the proposed method including the entropy-based negative rejection mechanism is considered (Fig. 2), the short sequences of 15 subsequent frames included in the dataset are used as inputs. Posterior probabilities estimated by the multi-class component for each of the 15 input frames are smoothed using Eq (3). The smoothed posterior probability is used to reject the input sequence or classify it among the different locations.

When the baseline classification pipeline proposed in [6] is considered (Fig. 3), the first image of each sequence is used as input. Input frames are whether rejected by the one-class classifier or discriminated into the positive classes by the multi-class classifier.

B. Rejection of Negative Samples

Rejection of negative samples is known as a hard problem and it can be tackled in different ways. Since all our experiments are performed on unbalanced datasets (the number of positive samples is larger than the number of negative ones – see Section IV), we don't use the accuracy to assess the performances of the methods under analysis. When the number of negative samples is low with respect to the positives one, a method with a high True Positive Rate (TPR) and a low True Negative Rate (TNR) still retains a high accuracy. Therefore, the performances of the proposed methods are assessed using the average between the TPR and the TNR, which we refer to as the True Average Rate (TAR):

$$TAR = \frac{TPR + TNR}{2}. \tag{6}$$

1) Entropy-Based Rejection Option: We apply the proposed entropy-based rejection method to discriminate negative from positive samples. For the experiments, we consider the short sequences of 15 subsequent frames contained in the proposed dataset. It should be noted that, given the standard rate of 30 fps, the length of each sequence is $0.5s$ long and hence the conditional independence assumption reported in Eq. (1) of Section III is satisfied. For each experiment, we choose t_e as the threshold which best separates the training set from the optimization negative samples included in the dataset. All thresholds are computed independently for each experiment (i.e., for each device-representation combination). Since the training set does not comprise 15-frames sequences, no temporal smoothing is performed on the training predictions and entropy is measured on the posterior probabilities predicted for each training sample.

In Section III we proposed to log-transform the smoothed posterior distribution (Eq. (5)) in order to compute the entropy-based score (Eq. (4)) used for negative rejection. To show that the considered log-transformation helps finding threshold t_e more reliably, in Fig. 4 we report the Threshold-TAR curves for some representative experiments. The curves plot thresholds t_e against the True Average Rate (TAR) scores obtained using such thresholds. The depicted curves are used to effectively find the best discrimination threshold t_e (i.e., the x-value corresponding to the curve peak). The figure reports the curves computed on the training sets plus optimization negatives, as well as the ones computed on the test sets. As can be noted, the curves computed using the log-transformation are almost totally overlapped, while there is far less overlap between the curves computed avoiding the log-transformation. To assess the robustness of the estimated thresholds, we also report the True Average Rate (TAR) results for all performed experiments in Fig. 5. The figure compares results obtained using the proposed method (i.e., thresholds t_e are computed from the training/optimization-negatives set) to those obtained with the optimal threshold computed directly on the test set using the ground truth labels. The average absolute difference between obtained and optimal results amounts to 0.06.

2) One Class Classifier: Following [6], we build a one-class SVM classifier for the purpose of rejecting negative samples. The optimization procedure of the one-class SVM classifier depends on a single parameter ν which is a lower bound on the fraction of outliers in the training set. We train the one-class component considering all the positive samples (the entire training set) and use the optimization negatives to choose the value of ν which maximizes the TAR value on the set of training samples plus optimization negatives. It should be noted that the classifier is learned solely from positive data, while the small amount of negatives is only used to optimize the value of the ν hyperparameter.

C. Multiclass Discrimination

To assess the performances of the considered representations with respect to the task of discriminating among the 8 personal locations, we train linear SVM classifiers on the training sets and test them on the corresponding test sets. Similarly to [38], [36], the input feature vectors are transformed using the Hellinger's kernel prior to using them in the linear SVM classifier. Differently from [38], [36], we do not apply L2 normalization to the feature vectors, but instead we independently scale each component of the vectors 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. Using L2 norm and avoiding feature scaling led to inconsistent results in our experiments. Such results are omitted for the sake of brevity. It should be noted that the considered scheme is adopted in order to obtain comparable results considering that very different representations are used. We do not intend to suggest that the employed normalization scheme performs better than others in general. The optimization procedure of the linear SVM classifier depends only on the cost parameter C, which is chosen in order to maximize the accuracy on the training set using cross-validation techniques [38], [36]. It should be noted that, in the case of fine-tuning, Convolutional Neural Networks are jointly used for feature extraction and classification. Therefore, in such cases, we do not rely on a SVM classifier for multi-class classification. When fine-tuned models are employed within the baseline proposed in [6], they are used both to extract features (on top of which the SVM One-Class classifier can be learned) and to directly perform multi-class classification. We would like to emphasize that in our experiments the multi-class classifier is learned using only positive samples.

VII. Results

In this section, we report the performances of the overall system implemented according to the two considered pipelines, as well as detailed performances of the discrimination and negative rejection components individually.

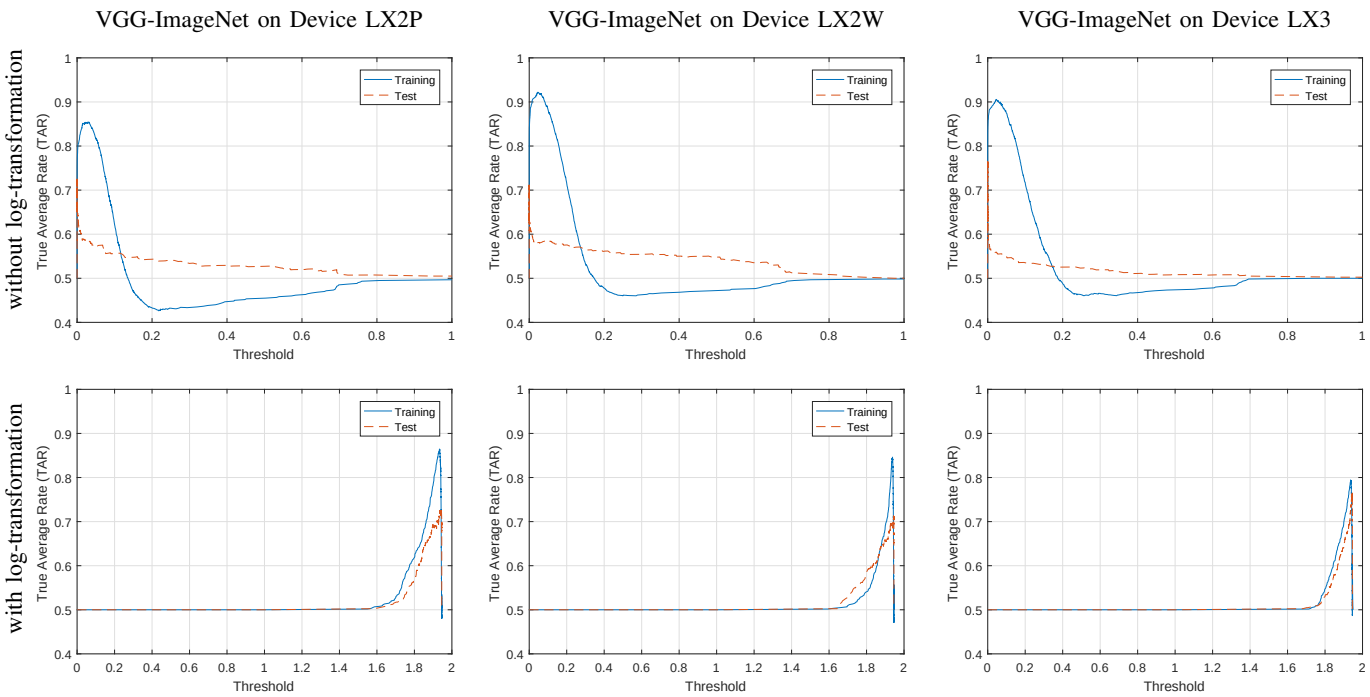


Fig. 4. Threshold-TAR (True Average Rate) curves obtained without (top row) and with (bottom row) log-transformation. All plots are obtained from posterior probabilities estimated by an SVM model trained extracting VGG-ImageNet features from data acquired using three different devices: the LX2P camera (perspective Looxcie LX2), the LX2W camera (wideangular Looxcie LX2), and the LX3 device (chest mounted Looxcie LX3).

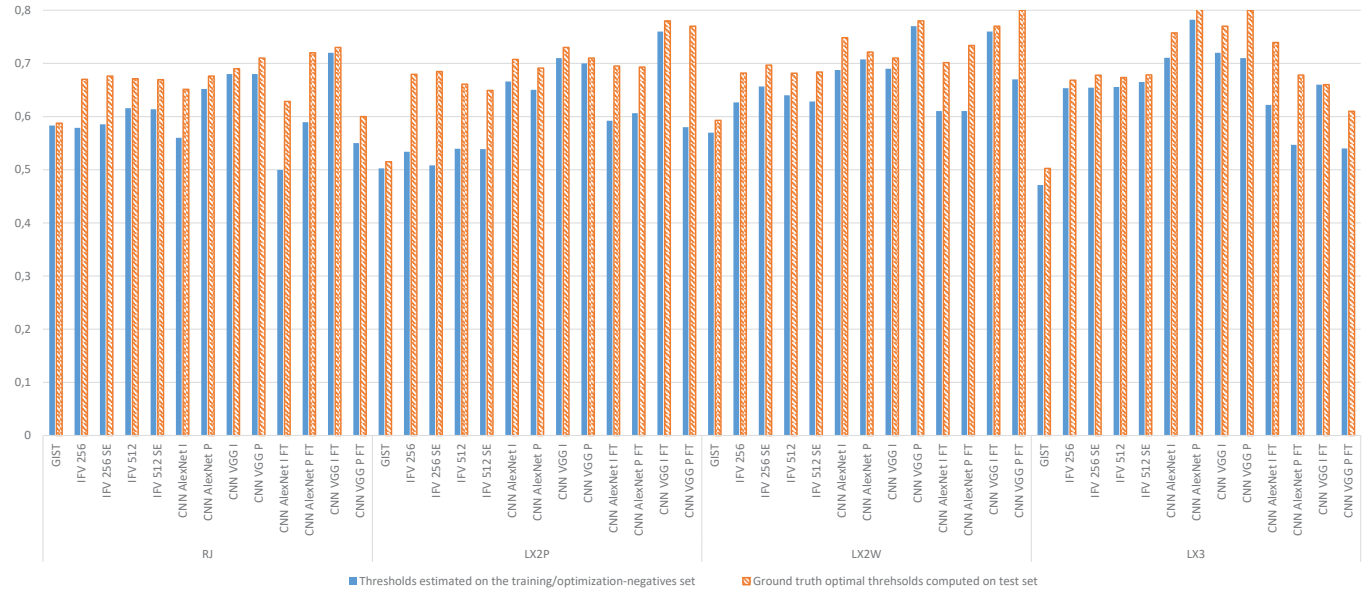


Fig. 5. True Average Rate (TAR) scores obtained on the test sets considering different combinations of devices and representations. The figure reports results obtained using thresholds computed on the training/optimization-negatives sets. Results obtained using the ground truth optimal thresholds computed on the test set are also reported for reference. As can be noted, estimated thresholds often reach close-to-optimal results. The average absolute difference between obtained and optimal results amounts to 0.06.

A. Overall System

TABLE II reports the accuracies of the overall system for the proposed method and the baseline introduced in [6]. Each row of the table corresponds to a different experiment and is denoted by a unique identifier in brackets (e.g., $[a_1]$). The first column (**METHOD**) reports the unique identifier and the representation method used in the experiment. The second column (**DEV.**) reports the device used to acquire the data. The third column (**OPTIONS**) reports the options related to the considered representation method. Specifically, in the case of representations based on the Improved Fisher Vectors (IFV), the numbers 256 or 512 represent the number of centroids used to train the GMMs, while "SE" indicates that the SIFT descriptors have been Spatially Enhanced as discussed in Section V-B.

TABLE II
PERFORMANCES OF THE OVERALL SYSTEM.

METHOD	DEV.	OPTIONS	DIM.	ACCURACY PROPOSED	ACCURACY [6]
$[a_1]$ GIST	RJ	—	512	22,44	25,67
$[b_1]$ IFV	RJ	256	40960	25,11	56,39
$[c_1]$ IFV	RJ	256 SE	41984	26,28	58,56
$[d_1]$ IFV	RJ	512	81920	31,67	55,78
$[e_1]$ IFV	RJ	512 SE	83968	31,33	56,61
$[f_1]$ CNN	RJ	AlexNet I	4096	58,11	58,94
$[g_1]$ CNN	RJ	AlexNet P	4096	67,00	62,33
$[h_1]$ CNN	RJ	VGG16 I	4096	71,61	43,83
$[i_1]$ CNN	RJ	VGG16 P	4096	61,17	60,00
$[j_1]$ CNN	RJ	AlexNet I FT	4096	65,94	60,00
$[k_1]$ CNN	RJ	AlexNet P FT	4096	**76,83**	76,72
$[l_1]$ CNN	RJ	VGG16 I FT	4096	64,11	**76,89**
$[m_1]$ CNN	RJ	VGG16 P FT	4096	75,06	70,78
$[a_2]$ GIST	LX2P	—	512	29,44	22,61
$[b_2]$ IFV	LX2P	256	40960	17,50	51,39
$[c_2]$ IFV	LX2P	256 SE	41984	12,56	55,11
$[d_2]$ IFV	LX2P	512	81920	18,50	48,17
$[e_2]$ IFV	LX2P	512 SE	83968	18,00	48,33
$[f_2]$ CNN	LX2P	AlexNet I	4096	70,06	61,28
$[g_2]$ CNN	LX2P	AlexNet P	4096	64,11	49,89
$[h_2]$ CNN	LX2P	VGG16 I	4096	67,28	52,44
$[i_2]$ CNN	LX2P	VGG16 P	4096	63,33	44,83
$[j_2]$ CNN	LX2P	AlexNet I FT	4096	74,83	63,72
$[k_2]$ CNN	LX2P	AlexNet P FT	4096	69,94	72,00
$[l_2]$ CNN	LX2P	VGG16 I FT	4096	68,28	**75,89**
$[m_2]$ CNN	LX2P	VGG16 P FT	4096	**80,06**	70,50
$[a_3]$ GIST	LX2W	—	512	39,83	23,22
$[b_3]$ IFV	LX2W	256	40960	37,50	59,17
$[c_3]$ IFV	LX2W	256 SE	41984	42,83	58,44
$[d_3]$ IFV	LX2W	512	81920	39,50	52,06
$[e_3]$ IFV	LX2W	512 SE	83968	37,06	51,50
$[f_3]$ CNN	LX2W	AlexNet I	4096	75,22	65,61
$[g_3]$ CNN	LX2W	AlexNet P	4096	73,89	55,06
$[h_3]$ CNN	LX2W	VGG16 I	4096	70,89	54,06
$[i_3]$ CNN	LX2W	VGG16 P	4096	81,67	50,06
$[j_3]$ CNN	LX2W	AlexNet I FT	4096	73,89	65,44
$[k_3]$ CNN	LX2W	AlexNet P FT	4096	76,22	73,78
$[l_3]$ CNN	LX2W	VGG16 I FT	4096	76,78	73,78
$[m_3]$ CNN	LX2W	VGG16 P FT	4096	**87,28**	**80,11**
$[a_4]$ GIST	LX3	—	512	29,50	29,22
$[b_4]$ IFV	LX3	256	40960	39,94	29,11
$[c_4]$ IFV	LX3	256 SE	41984	40,44	37,00
$[d_4]$ IFV	LX3	512	81920	39,50	27,56
$[e_4]$ IFV	LX3	512 SE	83968	39,89	27,28
$[f_4]$ CNN	LX3	AlexNet I	4096	65,39	51,39
$[g_4]$ CNN	LX3	AlexNet P	4096	**76,50**	55,72
$[h_4]$ CNN	LX3	VGG16 I	4096	73,22	34,17
$[i_4]$ CNN	LX3	VGG16 P	4096	76,11	51,94
$[j_4]$ CNN	LX3	AlexNet I FT	4096	73,06	**66,94**
$[k_4]$ CNN	LX3	AlexNet P FT	4096	67,61	56,28
$[l_4]$ CNN	LX3	VGG16 I FT	4096	61,94	60,65
$[m_4]$ CNN	LX3	VGG16 P FT	4096	71,39	44,00

In the CNN-related experiments, "I" denotes that the considered model has been pre-trained on the ImageNet dataset, "P" denotes that the considered model has been pre-trained on the Places205 dataset, "FT" indicates that the network has been fine-tuned, while, when no "FT" tag is reported, the pre-trained network is only used to extract the representation vectors. The fourth column (**DIM.**) reports the dimensionality of the feature vectors. The fifth and sixth columns report the accuracies of the model according to the two compared methods. To improve readability, for each method, the maximum accuracies among the experiments related to a given device are reported in **bold numbers**, while the global maximum accuracy is reported in **boxed bold numbers**.

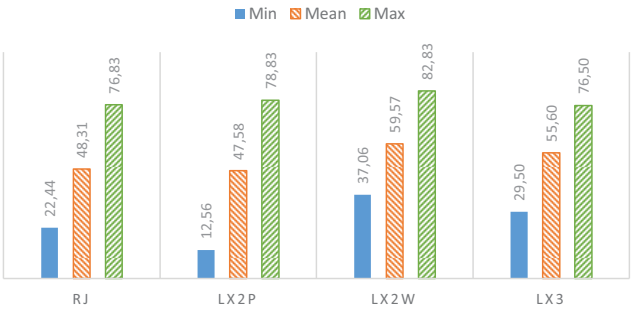


Fig. 6. Minimum, average and maximum accuracies of the overall system with the different representations per device. All the statistics are higher for the LX2W-related experiments. This suggests that the task of recognizing personal locations is easier on images acquired using a head mounted, wide-FOV device.

The proposed entropy-based negative rejection method generally allows to obtain better results with respect to the baseline method [6] when deep representations are used. Comparable or worse performances are generally obtained when using other representations. The holistic GIST representation is usually unable to model the personal locations with the appropriate level of detail (compare methods $[a_1]$, $[a_2]$, $[a_3]$ and $[a_4]$ to others). Improved Fisher Vectors (IFV) generally work better than GIST, but provide inconsistent results in some cases (e.g., $[b_1]$ to $[e_1]$ and $[b_2]$ to $[e_2]$). Using larger codebooks allows to obtain better results in some cases (e.g., when smart glasses Recon Jet (RJ) and narrow-angle ear-mounted LX2P camera are used) at the cost of a significantly larger representation ($80k$ vs $40k$ dimensions). The Spatially Enhancement option (SE) does not in general result in significant improvements. The best performances are given by deep representations. Fine-tuning the model often, but not always (e.g., compare $[h_1]$ to $[l_1]$, $[f_3]$ to $[j_3]$ and $[h_4]$ to $[l_4]$) results in a significant performance improvement.

One important fact emerging from the analysis of the results in TABLE II consists in the superior performances obtained on the data acquired using the LX2W device. This observation is supported by Fig. 6, which reports the minimum, maximum and average accuracies of the overall system for all the experiments related to a given device when the proposed method is considered. All three indicators are higher in the case of the LX2W camera, which suggest that, among the ones being tested, such device is the most appropriate for modelling the user's personal location. This result is probably due to the combination of the large FOV which allows to capture more information and the wearing modality, which enables the acquisition of the data from the user's point of view.

In Fig. 7 and Fig. 8, we report confusion matrices and some success/failure examples (true/false positive) for the best performing methods and each device. All confusion matrices point out how the most errors are due to the need to handle negative samples. In fact, most false positives are due to the misclassification of negative samples as shown in Fig. 8. Moreover, there is usually confusion between pairs of similar

Ground Truth classes \ Predicted Class	Car	C.V.M.	Office	L.R.	H. Office	K. Top	Sink	Garage	Negatives
Car	**91. 50**	0. 00	1. 00	0. 00	0. 00	0. 00	0. 50	0. 50	6. 50
C.V.M.	0. 00	**90. 00**	0. 00	0. 00	0. 00	1. 00	0. 00	3. 00	6. 00
Office	0. 00	0. 50	**69. 50**	1. 00	1. 00	3. 50	1. 00	0. 50	23. 00
L.R.	0. 00	0. 00	1. 00	**88. 00**	0. 50	1. 50	0. 00	0. 00	9. 00
H. Office	0. 00	1. 50	6. 50	1. 00	**54. 00**	6. 00	0. 00	0. 00	31. 00
K. Top	0. 00	0. 00	0. 00	0. 00	0. 00	**91. 50**	2. 50	0. 00	6. 00
Sink	0. 50	0. 00	0. 00	0. 00	0. 00	5. 00	**79. 00**	1. 00	14. 50
Garage	0. 00	0. 00	0. 00	0. 00	0. 00	4. 00	0. 00	**87. 00**	9. 00
Negatives	14. 50	0. 50	16. 50	0. 50	0. 50	21. 50	2. 00	13. 00	**31. 00**

(a) AlexNet-Places-FT/RJ ($[k_1]$)

Ground Truth classes \ Predicted Class	Car	C.V.M.	Office	L.R.	H. Office	K. Top	Sink	Garage	Negatives
Car	**98. 00**	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00	1. 00	1. 00
C.V.M.	0. 00	**98. 00**	0. 00	0. 00	0. 00	0. 00	0. 00	2. 00	0. 00
Office	0. 00	13. 00	**66. 50**	0. 00	7. 50	4. 50	0. 00	0. 00	8. 50
L.R.	0. 00	0. 00	0. 50	**97. 50**	0. 50	0. 00	0. 00	0. 00	1. 50
H. Office	0. 00	0. 00	15. 00	0. 00	**73. 00**	3. 50	0. 00	0. 00	8. 50
K. Top	0. 50	0. 00	0. 00	0. 00	0. 50	**95. 00**	1. 50	0. 50	2. 00
Sink	0. 50	0. 00	0. 00	0. 00	0. 50	18. 50	**74. 50**	1. 00	5. 00
Garage	1. 00	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00	**98. 50**	0. 50
Negatives	12. 00	14. 50	5. 00	0. 00	8. 50	12. 50	17. 50	10. 50	**19. 50**

(b) VGG-Places-FT/LX2P ($[m_2]$)

Ground Truth classes \ Predicted Class	Car	C.V.M.	Office	L.R.	H. Office	K. Top	Sink	Garage	Negatives
Car	**100. 00**	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00
C.V.M.	0. 00	**98. 50**	0. 50	0. 00	0. 00	0. 50	0. 00	0. 50	0. 00
Office	0. 00	0. 00	**92. 00**	0. 00	5. 00	0. 00	0. 00	0. 00	3. 00
L.R.	0. 00	0. 00	0. 00	**86. 00**	6. 00	0. 00	0. 00	0. 00	8. 00
H. Office	0. 00	0. 00	14. 00	0. 00	**79. 00**	0. 00	0. 00	0. 00	7. 00
K. Top	0. 00	0. 00	0. 00	0. 00	0. 50	**99. 50**	0. 00	0. 00	0. 00
Sink	0. 00	0. 00	0. 00	0. 00	0. 00	2. 50	**94. 50**	0. 00	3. 00
Garage	0. 50	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00	**99. 00**	0. 50
Negatives	9. 50	18. 50	9. 50	0. 00	3. 00	6. 50	10. 50	5. 50	**37. 00**

(c) VGG-Places-FT/LX2W ($[m_3]$)

Ground Truth classes \ Predicted Class	Car	C.V.M.	Office	L.R.	H. Office	K. Top	Sink	Garage	Negatives
Car	**96. 50**	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00	3. 50
C.V.M.	0. 00	**84. 00**	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00	16. 00
Office	0. 00	0. 00	**97. 00**	0. 00	0. 50	0. 00	0. 00	0. 00	2. 50
L.R.	0. 00	0. 00	0. 00	**92. 00**	0. 50	0. 00	0. 00	0. 00	7. 50
H. Office	0. 00	0. 00	27. 00	0. 00	**1. 50**	0. 00	0. 00	0. 00	71. 50
K. Top	0. 00	0. 00	0. 00	0. 00	0. 00	**94. 00**	0. 50	0. 00	5. 50
Sink	0. 00	0. 00	0. 00	0. 00	0. 00	5. 00	**52. 50**	0. 50	42. 00
Garage	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00	0. 00	**92. 00**	8. 00
Negatives	6. 00	0. 50	0. 00	0. 00	1. 50	6. 00	1. 50	8. 50	**76. 00**

(d) AlexNet-Places/LX3 ($[g_4]$)

Fig. 7. Confusion matrices related to the best performing methods on each of the considered devices. Rows represent ground truth classes, while columns represent the predicted labels. Each element of the confusion matrix is normalized by the sum of the elements in the corresponding row. Hence, values along the principal diagonal are class-related true positive rates. Confusion matrices are related to the following methods: (a) AlexNet Convolutional Neural Network pre-trained on the Places205 dataset and fine-tuned on data acquired using the Recon Jet (RJ) smart glasses, (b) VGG16 Convolutional Neural Network pre-trained on the Places205 dataset and fine-tuned on data acquired using the ear-mounted perspective Looxcie LX2 camera (LX2P), (c) VGG16 Convolutional Neural Network pre-trained on the Places205 dataset and fine-tuned on data acquired using the ear-mounted wideangular Looxcie LX2 camera (LX2W), (d) SVM classifier trained on AlexNet Convolutional Neural Network pre-trained on the Places205 dataset with data acquired using the chest mounted Looxcie LX3 camera. The reader is referred to TABLE II, TABLE III and TABLE IV for detailed results of all experiments.

Fig. 8. True positive (green) and false positive (red) samples related to the best performing methods on the four considered devices. Rows represent the ground truth labels, while the predicted label is shown in yellow in case of a failure. The shown samples are related to the the same methods considered in Fig. 7: (a) AlexNet Convolutional Neural Network pre-trained on the Places205 dataset and fine-tuned on data acquired using the Recon Jet (RJ) smart glasses, (b) VGG16 Convolutional Neural Network pre-trained on the Places205 dataset and fine-tuned on data acquired using the ear-mounted perspective Looxcie LX2 camera (LX2P), (c) VGG16 Convolutional Neural Network pre-trained on the Places205 dataset and fine-tuned on data acquired using the ear-mounted wideangular Looxcie LX2 camera (LX2W), (d) SVM classifier trained on AlexNet Convolutional Neural Network pre-trained on the Places205 dataset with data acquired using the chest mounted Looxcie LX3 camera.

TABLE III
RESULTS RELATED TO THE NEGATIVE REJECTION METHODS.

			EB			OCSVM [6]		
METHOD	DEV.	OPTIONS	TAR	TPR	TNR	TAR	TPR	TNR
$[a_1]$ GIST	RJ	—	58,31	37,63	79,00	53,72	55,44	52,00
$[c_1]$ IFV	RJ	256 SE	58,53	17,06	100,00	**54,00**	72,00	36,00
$[h_1]$ CNN	RJ	VGG16 I	67,59	73,69	61,50	48,06	46,63	49,50
$[l_1]$ CNN	RJ	VGG16 I FT	**72,31**	62,13	82,50	48,59	96,19	1,00
$[a_2]$ GIST	LX2P	—	50,25	63,00	37,50	**59,16**	34,81	83,50
$[d_2]$ IFV	LX2P	512	53,94	08,38	99,50	41,94	75,38	08,50
$[h_2]$ CNN	LX2P	VGG16 I	71,44	67,88	75,00	54,69	59,38	50,00
$[l_2]$ CNN	LX2P	VGG16 I FT	**76,34**	68,69	84.00	52,50	96,00	9,00
$[a_3]$ GIST	LX2W	—	56,97	66,44	47,50	**64,25**	50,50	78,00
$[c_3]$ IFV	LX2W	256 SE	65,66	36,31	95,00	51,63	79,25	24,00
$[i_3]$ CNN	LX2W	VGG16 P	**76,97**	84,44	69,50	59,41	50,31	68,50
$[m_3]$ CNN	LX2W	VGG16 P FT	67,16	97,31	37,00	59,03	91,06	27,00
$[a_4]$ GIST	LX3	—	47,13	50,75	43,50	67,16	44,81	89,50
$[e_4]$ IFV	LX3	512 SE	66,50	34,00	99,00	30,44	41,38	19,50
$[g_4]$ CNN	LX3	AlexNet P	**78,22**	80,44	76,00	**70,06**	57,13	83,00
$[k_4]$ CNN	LX3	AlexNet P FT	54,69	92,88	16,50	52,53	72,06	33,00

looking locations, e.g., Office - Home Office, Sink - Kitchen Top, Living Room - Home Office (see Fig. 8 for some examples). The confusion matrices shown in Fig. 7 (b) and (c) use similar models (a fine-tuned VGG16 network pre-trained on the ImageNet dataset) trained on data acquired using similar devices, differing mainly in their Field Of View (FOV): a narrow-angle Looxcie LX2 (LX2P) and a wide-angle Looxcie LX2 (LX2W). This allows to make direct considerations on the influence of the Field Of View (FOV) in the task of detecting locations of interest. In particular, the use of a wide-angle camera (Fig. 7 (b)) allows to acquire a larger portion of the Field Of View, which is useful to reduce the confusion between similar locations (e.g., Sink vs Kitchen Top).

B. Negative Samples Rejection

TABLE III reports the results related to the two rejection methods considered in our analysis: the proposed Entropy Based method (EB) and the One-Class SVM method proposed in [6] (OCSVM).[4] The table is organized similarly to TABLE II, except for the performance indicators used in this case. Columns 4 to 6 are related to the proposed Entropy-Based method (EB), while columns 7 to 9 are related to the baseline One-Class SVM component (OCSVM). Columns 4 and 7 report the True Average Rate (TAR). Columns $\{5, 8\}$ and $\{7, 9\}$ report respectively the True Positive Rate (TPR) and True Negative Rate (TNR) scores related to the considered methods. The proposed entropy-based method systematically outperforms the one-class SVM baseline, with the exception of the GIST-related methods $[a_2]$, $[a_3]$, $[a_4]$. Consistently with the observations made earlier, the best performing methods are generally the ones related to deep representations.

C. Multiclass Discrimination

TABLE IV reports the results related to the multi-class discrimination component. It should be noted that, in these experiments, negative rejection is not considered and methods are evaluated ignoring negative samples. The structure of TABLE IV follows the one of TABLE II, with the following differences: column 5 reports the accuracy of the multi-class discrimination component when negative samples are removed from the test sets, columns 6 to 13 report the True Positive Rates related to each of the considered classes. It should be noted that the reported results are related to the proposed method and hence they have been obtained using the smoothed posterior probabilities computed as defined in Eq. (3). As noted for TABLE II, the holistic GIST representations are unable to model the personal locations with the appropriate level of detail. Even if the accuracy values related to the GIST representations are always low, in some cases they are still able to model some classes like for instance Coffee Vending Machine (e.g., $[a_1]$, $[a_2]$, $[a_3]$ and $[a_4]$), Living Room (e.g., $[a_3]$) and Sink (e.g., $[a_4]$) which are characterized by distinctive spatial layouts. Interestingly, the shallow representations, albeit consistently outperformed by CNN, give remarkable performances in some cases (e.g., $[c_1]$). The best performances (bold numbers) are given again by the deep representations. While in the reported results, fine-tuned models significantly outperform their pre-trained counterparts, please note that this is not true for all experiments (see the supplementary material for more details).

D. Discussion

The experimental results presented in the previous sections show how the considered problem is a challenging one. As discussed earlier, the performances of all the considered methods are dominated by the limits of the negative rejection module, while the multi-class discrimination remains an "easier" subtask. This suggests that more efforts should be devoted to the design of efficient and robust negative rejection methods. The systematic emergence of deep representations as the best performing ones, not only indicates the higher representational power of such methods, but also suggests that the considered problem can take great advantage of transfer learning techniques. All CNN-based representations have been obtained using models pre-trained on a large number of images, which compensates for the scarce quantity of training data assumed in this study. Nevertheless, such a small number of frames can also limit the considered transfer learning techniques, especially when fine-tuning existing models. We believe that such problem could be mitigated by a more application-aware data augmentation technique. In particular, considering that the training frames belong to a given environment, they could be used to infer a 3D model using structure from motion techniques. Synthetic, yet realistic, samples from different points of view could be then extracted in order to augment the number of training samples.

As already pointed out in Section VII, the LX2W device is the one obtaining the best performances. This suggests that head-mounted wide-angular cameras are probably the best option when modelling the user's location. This is not surprising since such configuration allows to better replicate the user's point of view and provides a FOV similar to the

[4]For sake of brevity, we report only some representative results for each device-representation combination. For the full table containing all the results please refer to the supplementary material available at the url http://iplab.dmi.unict.it/PersonalLocations/thms_supplementary.pdf.

TABLE IV
RESULTS RELATED TO THE MULTI-CLASS COMPONENT.

METHOD	DEV.	OPTIONS	DIM.	ACC	CAR	C. V. M.	OFFICE	L.R.	H. OFFICE	K. TOP	SINK	GARAGE
$[a_1]$ GIST	RJ	—	512	37,56	62,86	98,59	48,65	0,00	32,79	48,72	25,00	29,06
$[c_1]$ IFV	RJ	256 SE	41984	82,94	95,50	88,83	89,23	100,00	97,86	68,42	95,83	59,70
$[h_1]$ CNN	RJ	VGG16 I	4096	**93,50**	100,00	99,49	97,37	100,00	81,48	84,00	97,24	94,03
$[l_1]$ CNN	RJ	VGG16 I FT	4096	90,00	74,19	76,92	100,00	98,45	97,50	87,00	95,24	96,08
$[a_2]$ GIST	LX2P	—	512	42,69	42,25	99,29	17,18	65,28	44,81	37,41	83,33	24,22
$[c_2]$ IFV	LX2P	256 SE	41984	74,44	64,47	100,00	36,28	100,00	60,14	72,97	100,00	86,15
$[h_2]$ CNN	LX2P	VGG16 I	4096	**90,00**	100,00	96,14	72,17	100,00	83,03	77,65	99,34	98,98
$[l_2]$ CNN	LX2P	VGG16 I FT	4096	88,06	96,80	77,27	100,00	92,23	98,02	62,50	93,88	100,00
$[a_3]$ GIST	LX2W	—	512	51,75	57,97	94,74	48,36	93,48	30,32	33,95	92,50	31,48
$[c_3]$ IFV	LX2W	256 SE	41984	74,06	53,76	100,00	97,50	100,00	83,78	53,04	100,00	80,97
$[i_3]$ CNN	LX2W	VGG16 P	4096	**95,44**	99,49	99,00	85,97	100,00	96,05	89,45	100,00	95,67
$[m_3]$ CNN	LX2W	VGG16 P FT	4096	94,88	83,76	83,48	100,00	100,00	99,50	99,49	95,22	99,50
$[a_4]$ GIST	LX3	—	512	46,31	42,41	84,07	35,56	45,60	33,33	56,05	83,52	23,26
$[c_4]$ IFV	LX3	256 SE	41984	68,31	100,00	100,00	81,52	100,00	100,00	31,24	97,69	80,24
$[i_4]$ CNN	LX3	VGG16 P	4096	**87,63**	100,00	99,48	62,26	91,28	96,77	88,21	92,93	92,02
$[m_4]$ CNN	LX3	VGG16 P FT	4096	81,81	60,00	49,62	99,44	97,55	93,75	100,00	76,89	97,04

one characterizing the human visual system. Moreover, head mounted devices are particularly helpful for understanding the user's activities [15], which plays an important role in modeling the user's context.

VIII. CONCLUSION AND FUTURE WORKS

In this paper, we have studied the problem of recognizing personal locations from egocentric videos. We have proposed a dataset containing more than 20 hours of videos acquired by a user in 8 different locations using four different wearable devices. We have analyzed the performances of the main state-of-the-art representation techniques for scene and object classification on the considered task, emphasizing the role of a negative rejection mechanism for building effective location detection systems. A negative rejection option has been proposed and compared with respect to a baseline based on a one-class SVM classifier. The results highlight that deep representations systematically outperform the competitors and that the best results are achieved using the LX2W device, which suggests that head-mounted, wide-angular devices are the most suited to recognize the user's personal locations.

Future works will focus on investigating contextual sensing using action-related video features such as the ones proposed in [43], [44], as well as motion-related features such as the ones proposed in [21], [45], [46]. Moreover, this study could be extended to the multi-user case to assess the generalization ability of the investigated methods. In particular, it would be interesting to investigate how data acquired by different subjects can be leveraged to mitigate the lack of training data and improve the recognition system. This is reasonable since personal locations selected by different subjects are likely to present some degree of overlap, e.g., different users are likely to select similar locations, such as "Kitchen" and "Office". In order to make the system more valuable in real and complex scenarios, methods to enforce temporal coherence between neighboring predictions will be investigated. Finally, applications related to the robotic domain will be considered in future works.

ACKNOWLEDGEMENTS

This work has been performed in the project FIR2014-UNICT-DFA17D.

REFERENCES

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

[2] A. K. Dey, G. D. Abowd, and D. Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Compututer Interaction*, 16(2):97–166, 2001.

[3] S. Greenberg. Context as a dynamic construct. *Human-Compututer Interaction*, 16(2), 2001.

[4] T. Kanade and M. Hebert. First-person vision. *Proceedings of the IEEE*, 100(8):2442–2453, 2012.

[5] 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.

[6] A. Furnari, G. M. Farinella, and S. Battiato. Recognizing personal contexts from egocentric images. In *Workshop on Assistive Computer Vision and Robotics (ACVR) in conjunction with ICCV*, 2015.

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

[8] A. Torralba, K. P Murphy, W. T. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In *International Conference on Computer Vision*, pages 273–280, 2003.

[9] 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.

[10] A. Furnari, G. M. Farinella, and S. Battiato. Temporal segmentation of egocentric videos to highlight personal locations of interest. In *International Workshop on Egocentric Perception, Interaction and Computing (EPIC) in conjunction with ECCV*, 2016.

[11] E. Thomaz, A. Parnami, I. Essa, and G. D. Abowd. Feasibility of identifying eating moments from first-person images leveraging human computation. In *SenseCam and Pervasive Imaging Conference*, pages 26–33, 2013.

[12] 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.

[13] D. Ravì, B. Lo, and G. Yang. Real-time food intake classification and energy expenditure estimation on a mobile device. *Body Sensor Network, MIT, Boston, MA, USA*, 2015.

[14] 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.

[15] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. *Proceedings of the IEEE International Conference on Computer Vision*, pages 407–414, 2011.
[16] 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.
[17] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *British Machine Vision Conference*, 2014.
[18] 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.
[19] Y. Yan, E. Ricci, G. Liu, and N. Sebe. Egocentric daily activity recognition via multitask clustering. *Transactions on Image Processing*, 24(10):2984–2995, 2015.
[20] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *Computer Vision and Pattern Recognition*, pages 2714–2721, 2013.
[21] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *Computer Vision and Pattern Recognition*, pages 2537–2544, 2014.
[22] 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.
[23] 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.
[24] P. Wu, H.-K. Peng, J. Zhu, and Y. Zhang. Senscare: Semi-automatic activity summarization system for elderly care. In *International Conference on Mobile Computing, Applications, and Services*, pages 1–19, 2011.
[25] G. Lu, Y. Yan, L. Ren, J. Song, N. Sebe, and C. Kambhamettu. Localize me anywhere, anytime: A multi-task point-retrieval approach. In *International Conference on Computer Vision*, 2015.
[26] H. Wannous, V. Dovgalecs, R. Mégret, and M. Daoudi. Place recognition via 3d modeling for personal activity lifelog using wearable camera. In *International Conference on Multimedia Modeling*, pages 244–254, 2012.
[27] M. Dimiccoli, M. Bolaños, E. Talavera, M. Aghaei, S. G. Nikolov, and P. Radeva. Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation. *arXiv preprint arXiv:1512.07143*, 2015.
[28] M. Bolaños, M. Dimiccoli, and P. Radeva. Towards storytelling from visual lifelogging: An overview. *Transactions on Human-Machine Systems*, 2015.
[29] A. Torralba and A. Oliva. Semantic organization of scenes using discriminant structural templates. *International Conference on Computer Vision*, 2:1253–1258, 1999.
[30] 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.
[31] G. M. Farinella and S. Battiato. Scene classification in compressed and constrained domain. *IET computer vision*, 5(5):320–334, 2011.
[32] 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.
[33] 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.
[34] 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.
[35] 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*, 2015.
[36] 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.
[37] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *International Conference on Computer Vision*, pages 143–156, 2010.
[38] 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.
[39] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.
[40] 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.
[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.
[42] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
[43] Heng W. and C. Schmid. Action recognition with improved trajectories. In *IEEE conference on Computer Vision*, pages 3551–3558, 2013.
[44] Yin L., Zhefan Y., and J.M. Rehg. Delving into egocentric actions. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 287–295, 2015.
[45] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.
[46] M. S. Ryoo, B. Rothrock, and L. Matthies. Pooled motion features for first-person videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

Antonino Furnari received his bachelor degree and his master degree from the University of Catania in 2010 and 2013 respectively. Since 2013 he is a Computer Science PhD student at the University of Catania under the supervision of Prof. Sebastiano Battiato and Dr. Giovanni Maria Farinella. His research interests concern Pattern Recognition and Computer Vision, with focus on First Person Vision. More information at http://www.dmi.unict.it/~furnari/.

Giovanni Maria Farinella (M'11–SM'16) 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. His research interests concern Computer Vision, Pattern Recognition and Machine Learning. He founded (in 2006) and currently directs the International Computer Vision Summer School - ICVSS. More information at http://www.dmi.unict.it/farinella/.

Sebastiano Battiato (M'04–SM'06) 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 as assistant professor in 2004 and became associate professor in the same department in 2011. His research interests include image enhancement and processing, image coding, camera imaging technology and multimedia forensics. More information at http://www.dmi.unict.it/~battiato/.