Gesture Recognition Using Wearable Vision Sensors to Enhance Visitors' Museum Experiences

Lorenzo Baraldi, Francesco Paci, Giuseppe Serra, Luca Benini, *Fellow, IEEE*, and Rita Cucchiara, *Member, IEEE*

***Abstract*— We introduce a novel approach to cultural heritage experience: by means of ego-vision embedded devices we develop a system, which offers a more natural and entertaining way of accessing museum knowledge. Our method is based on distributed self-gesture and artwork recognition, and does not need fixed cameras nor radio-frequency identifications sensors. We propose the use of dense trajectories sampled around the hand region to perform self-gesture recognition, understanding the way a user naturally interacts with an artwork, and demonstrate that our approach can benefit from distributed training. We test our algorithms on publicly available data sets and we extend our experiments to both virtual and real museum scenarios, where our method shows robustness when challenged with real-world data. Furthermore, we run an extensive performance analysis on our ARM-based wearable device.**

***Index Terms*— Wearable vision, interactive museum, embedded systems, gesture recognition, natural interfaces.**

I. Introduction

IN RECENT years the interest in cultural heritage has reborn, and the cultural market is becoming a cornerstone in many national economic strategies. In the United States, a recent report of the Office of Travel and Tourism Industries claims that 51% of the 40 million Americans traveling abroad visit historical places; almost one third visit cultural heritage sites; and one quarter go to an art gallery or museum [1]. The same interest is found in Europe, where the importance of the cultural sector is widely acknowledged, South Asia and North Africa. The latest annual research from World Travel and Tourism Council shows that travel and tourism's total contribution to total GDP grew by 3.0% in 2013, faster than overall economic growth for the third consecutive year [2].

Manuscript received April 9, 2014; revised September 4, 2014 and February 5, 2015; accepted March 2, 2015. Date of publication March 11, 2015; date of current version March 27, 2015. This work was supported in part by the Seventh Framework Program (FP7) through the PHIDIAS Project under Grant g.a.318013, in part by the FP7 European Research Council through the MULTITHERMAN Project under Grant g.a. 291125, in part by the National Operative Programme for Research and Competitiveness through the DICET-INMOTO Project under Grant Cod.PON04a2_D, and in part by the Vision for Augmented Experiences through the Fondazione CRMO Project. The associate editor coordinating the review of this paper and approving it for publication was Dr. Faisal Qureshi.

L. Baraldi, G. Serra, and R. Cucchiara are with the Dipartimento di Ingegneria Enzo Ferrari, University of Modena and Reggio Emilia, Modena 41100, Italy (e-mail: lorenzo.baraldi@unimore.it; giuseppe.serra@unimore.it; rita.cucchiara@unimore.it).

F. Paci is with the Dipartimento dell'Energia Elettrica e dell'Informazione, University of Bologna, Bologna 40126, Italy (e-mail: f.paci@unibo.it).

L. Benini is with the Dipartimento dell'Energia Elettrica e dell'Informazione, University of Bologna, Bologna 40126, Italy, and also with the Department of Information Technology and Electrical Engineering, ETH Zurich, Zurich 8092, Switzerland (e-mail: luca.benini@unibo.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSEN.2015.2411994

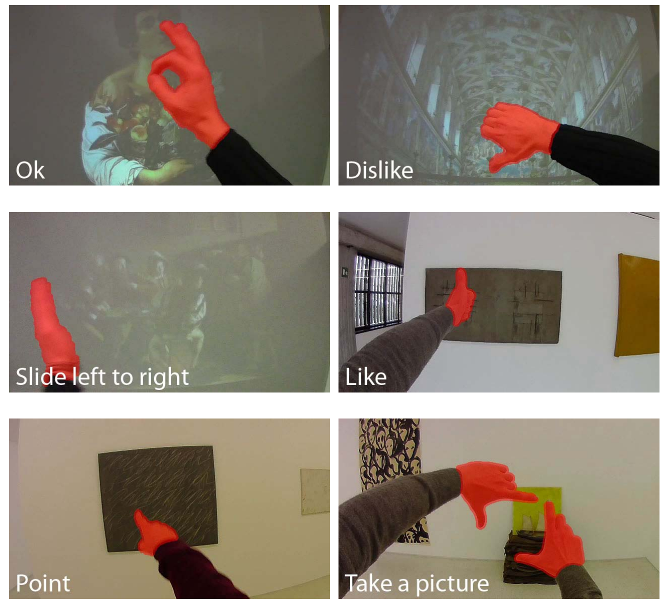

Fig. 1. Natural interaction with artworks: visitors can get specific content or share information about the observed artwork through simple gestures. Hand segmentation results are highlighted in red and detected gestures are reported in the bottom part of each frame.

Consequently, to deal with an increasing percentage of "digital native" tourists, a big effort is underway to propose new interfaces for interacting with the cultural heritage. In this direction goes the solution "SmartMuseum" proposed by Kuusik *et al.* [3]: by the means of PDAs and RFIDs, a visitor can gather information about what the museum displays, building a customized visit based on his or her interests inserted, prior to the visit, on their website. This project brought an interesting novelty when first released, but it has some limitations. First, being tied to RFIDs does not allow reconfiguring the museum without rethinking the entire structure of the exhibition. Furthermore, researches demonstrated how the use of mobile devices on the long term decreases the quality of the visit due to their users paying more attention to the tool rather than to the work of art itself.

In 2007 Kuflik *et al.* [4] proposed a system to customize visitors experiences in museums using software capable of learning their interests based on the answers to a questionnaire that they compiled before the visit. Similarly to SmartMuseum, one of the main shortcomings of this system is the need to stop the visitor and force him into doing something that he/she might not be willing to do. An interesting attempt to user profiling with wearable sensors was the Museum Wearable [5], a wearable computer which orchestrates an audiovisual narration as a function of the visitors' interests gathered from his/her physical path in the museum. However this prototype does not use any computer vision algorithm for understanding the surrounding environment. For instance the estimation of the visitor location is based again on infrared sensors distributed in the museum space.

Museums and cultural sites still lack of an instrument that provides entertainment, instructions and visit customization in an effective natural way. Too often visitors struggle to find the description of the artwork they are looking at and when they finds it, its detail level could be too high or too low for their interests. Moreover, frequently the organization of the exhibition does not reflect the visitors' interests leading them to a pre-ordered path which cultural depth could not be appropriate.

To overcome these limitations, we present a solution to enhance visitors' experiences based on a new emerging technology, namely *ego-vision* [6]. Ego-vision features glass-mounted wearable cameras able to see what the visitor sees and perceiving the surrounding environment as he does. We developed a wearable vision device for museum environments, able to replace the traditional self-service guides and overcoming their limitations and allowing for a more interactive museum experience to all visitors. The aim of our device is to stimulate the visitors to interact with the artwork, reinforcing their real experience, by letting visitors to replicate the gestures (e.g. point out to the part of the painting they're interested in) and behaviors that they would use to ask a guide something about the artwork.

In this work, we provide algorithms that perform gesture analysis to recognize user interaction with artworks and artwork recognition to achieve context awareness (see Fig. 1) The proposed solution is based on scalable and distributed wearable devices capable of communicating with each other and with a central server and hence does not require fixed cameras. In particular the connection with the central server allows our wearable devices to grab gestures of past visitors for improving gesture analysis accuracy, to get information and specific content of the observed artwork through the automatic recognition module, and to share visitor's feelings and photos on social networks.

The main novelties and contributions of this paper are:

- A distributed architecture that improves museum visitors' experience. It is composed by ego-vision wearable devices and a central server, and it is capable of recognizing users' gestures and artworks.
- A gesture recognition approach specifically developed for the ego-vision perspective. Unlike standard gesture recognition techniques, it takes into account camera motion and background cluttering, and does not need markers on hands. It shows superior performance when compared on benchmark dataset, and can achieve good accuracy results even with a few training samples. We further demonstrate that it can benefit from distributed training in which gestures performed by past visitors are exploited.
- A novel hand segmentation approach that considers temporal and spatial consistency, and that is capable of adapting itself to different illumination conditions. It achieves the state-of-the-art results in the ego-vision EDSH dataset. Moreover, we show that when combined with our gesture recognition approach, it can improve the overall system accuracy.
- A performance evaluation of our algorithms on an ARM big.LITTLE heterogeneous platform for embedded devices and an implemented tradeoff between accuracy and performance.

The rest of this article is structured as follows: in the next section we report related works for ego-vision. In Section III we give a detailed description of our system, focusing on self gesture recognition and artwork recognition. In Section IV our algorithms are compared with the state of the art and we present two novel datasets taken in real and virtual museum environments.

II. Related Work

Only recently the ego-vision scenario has been addressed by the research community. The main effort has focused on understanding human activities and detecting hand regions. Pirsiavash and Ramanan [7] detected activities of daily living using temporal pyramids and object detectors tuned for objects appearance during interactions and spatial reasoning. Sundaram and Cuevas [8] proposed instead to use Dynamic Bayesian Networks to recognize activities from low resolution videos, without performing hand detection and preferring computational inexpensive methods. Fathi *et al.* [9] used a bottom-up segmentation approach to extract hand held objects and trained object-level classifier to recognize objects; furthermore they also proposed an activity detection algorithm based on object state changes [10].

Regarding hand detection, Khan *et al.* in [11] studied color classification for skin segmentation. They pointed out how color-based skin detection has many advantages and potentially high processing speed, invariance against rotation, partial occlusion and pose change. The authors tested Bayesian Networks, Multilayers Perceptrons, AdaBoost, Naive Bayes, RBF Networks and Random Forest. They demonstrated that Random Forest classification obtains the highest F-score among all the other techniques. Fathi *et al.* [9] proposed another approach to hand detection, based on the assumption that background is static in the world coordinate frame, thus foreground objects are detected as the moving regions with respect to the background. An initial panorama of the background is required to discriminate between background and foreground regions: this is achieved by fitting a fundamental matrix to dense optical flow vectors. This approach is shown to be a robust tool for skin detection and hand segmentation in limited indoor environments, even if it performs poorly with more unconstrained scenarios.

Li and Kitani [12] provide a historical overview of approaches for detecting hands from moving cameras. They define three categories: local appearance-based detection, global appearance-based detection, where a global template of hand is needed, and motion-based detection, which is based on the hypothesis that hands and background have different motion statistics. Motion-based detection approaches require no supervision nor training. On the other hand,

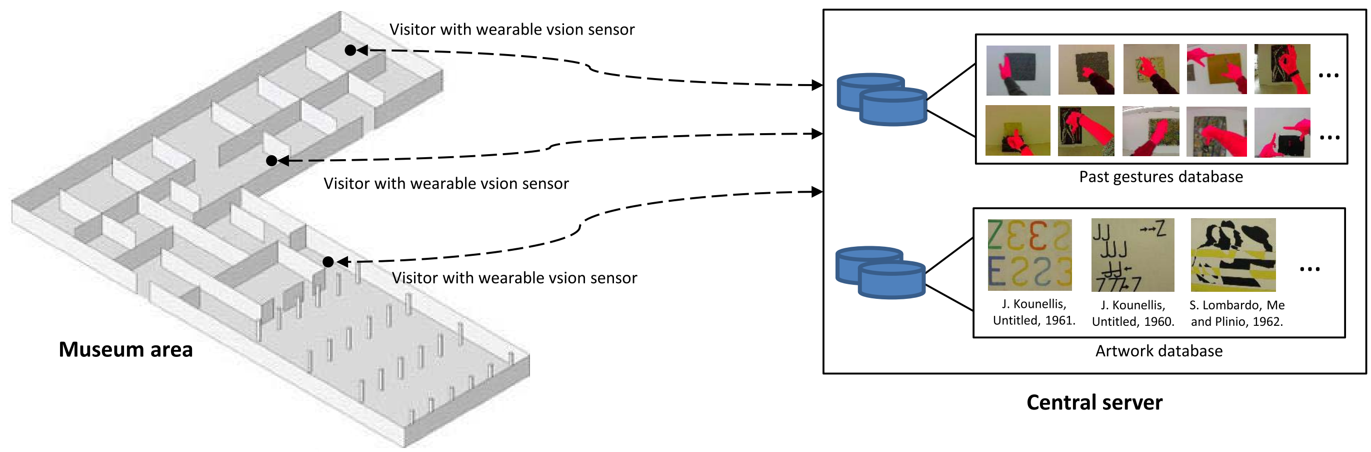

Fig. 2. Schema of the proposed distributed system. Each wearable vision sensor can communicate with a central server to send captured hand gestures and to retrieve gestures from other users and painting templates for artwork recognition. The central server contains two databases: the gesture database, which includes gestures performed by past visitors, and the artwork database, which contains artwork templates.

these approaches may identify as hand an object manipulated by the user, since it moves together with his hands. In addition they proposed a method with sparse feature selection which was shown to be an illumination-dependent strategy. To solve this issue, they trained a set of Random Forests indexed by a global color histogram, each one reflecting a different illumination condition.

Several approaches to gesture and human action recognition have been proposed. Sanin *et al.* [13] developed a new and more effective spatio-temporal covariance descriptor to classify gestures in conjunction with a boost classifier. Lui *et al.* [14], Lui and Beveridge [15] used tensors and tangent bundle on Grassmann manifolds to classify human actions and hand gestures. Kim and Cipolla [16] extended Canonical Correlation Analysis to measure video-to-video similarity to represent and detect actions in video. However, all these approaches are not appropriate for the ego-centric perspective, as they do not take into account any of the specific characteristics of this domain, such as fast camera motion and background cluttering. To our knowledge, the study of gesture recognition in the ego-centric paradigm has been partially addressed by Mistry and Maes [17]. Their work presents a natural interface to interact with the physical world and embeds a projector to show results of that interaction. However they use colored markers on user's fingers to recognize gestures and they require a backpacked laptop as computational unit. Although our work could seem similar to this last approach, we move a step forward with respect to [17]: we proposed a fully automatic gesture recognition approach based on appearance and motion of the hands. Our approach can deal with background cluttering and camera motion and does not require any markers on fingers. In addition we provide an embedded solution that the user can easily wear.

III. Proposed Architecture

Our cultural heritage system consists of a central server and a collection of wearable ego-vision devices, that embed a glass-mounted camera and an Odroid-XU developer board, serving as video-processing and network communication unit (see Fig. 3). There are several benefits in using such a portable device: the commercial availability and low costs for prototypes evaluation, the computational power and energy efficiency of the big.LITTLE architecture, the possibility of peripheral addition to extend connections and input devices. In particular, the developer board [18] we use embeds the ARM Exynos 5 SoC, that hosts a Quad big.LITTLE ARM processor (Cortex A15 and A7) [19]. To make it a portable demo device a battery pack of 3000 mAh has been added (see Figure 4).

This wearable device hosts the two main components of our system. The first one is the software that makes it capable of recognizing the gestures performed by its user and can customize itself, learning the way its user reach out for information. Adapting to personal requests is a key aspect in this process, in fact people in different cultures have very different ways of express through gestures. Our method is robust to lighting changes or ego-motion and can learn from a very limited set of examples gathered during a fast setup phase involving the user. The second component of our architecture is the artwork recognition, which allows not only to understand what the user is observing but also to infer the user's position.

The cooperation of ego-vision devices with the central server is two-fold. First, to increase gesture recognition accuracy, wearable devices receive gesture examples performed by past visitors and then send gestures for future users to augment the training set; second, the server also features a database of all the artworks in the museum, which is used for painting recognition and for obtaining detailed text, audio and video content. A schema of the proposed system is presented in Figure 2.

A. Gesture Recognition

Gestures can be characterized by both static and dynamic hand movements. Therefore, we consider a video sequence captured by a glass mounted camera, in which a gesture may be performed, and describe it as a collection of dense trajectories extracted around hand regions. When the user's

Fig. 3. One user interacting with wearable camera.

Fig. 4. The Odroid-XU board with battery pack.

hands appear, feature points are sampled inside and around the hands and tracked during the gesture; then several descriptors are computed inside a spatio-temporal volume aligned with each trajectory to capture its shape, appearance and movement at each frame. We use the following descriptors, according to [20]: Trajectory descriptor, histograms of oriented gradients (HOG), of optical flow (HOF), and motion boundary histograms (MBH). The first one directly captures trajectory shape, while HOG [21] are based on the orientation of image gradient and thus encode the static appearance of the region surrounding the trajectory. HOF and MBH [22] are based on optical flow and are used to capture motion information enforcing the temporal aspect of our method. These descriptors are coded, using the Bag of Words approach and power normalization, to obtain the final feature vectors, which are then classified using a linear SVM classifier. Figure 5 provides a more detailed outline of the workflow of the proposed gesture analysis module.

1) Camera Motion Removal: To estimate the hand motion, it is first necessary to remove the camera motion, which is, semantically, noise. To do so, the homography transform between two consecutive frames is estimated running the RANSAC [23] algorithm on densely sampled features points: SURF [24] features and sample motion vector are extracted from the Farneback's optical flow [25] to get dense matches between frames. The choice of this particular optical flow algorithm is induced by our preliminary tests, in which Farneback's optical flow showed the best performance when compared to other popular optical flow algorithms, such as TV-L1 [26] and SimpleFlow [27].

In ego-vision, however, it is often the case where camera and hand motions are not consistent, resulting in wrong matches between the frames and degrading the consequent homography estimation. This introduces the need for an additional step based on a totally decoupled feature. We use a hand segmentation mask that allows us to remove the matches belonging to the user's hands, which could have resulted in incorrect trajectories. Computing the homography based only on non-hand keypoints allows to have a motion model consistent with the ego-motion of the camera which can, consequently, be removed.

2) Gesture Description: After the suppression of camera motion, trajectories can be extracted. Using the previously estimated homography, each frame of the sequence is warped and the Farneback's optical flow between each couple of adjacent frames is recomputed to estimate the motion resulting from the hand movement. Feature points around the hand region are sampled and tracked in a way similar to [20]. We build a spatial pyramid with four layers, such that each layer has half the area of the previous one, and at each spatial scale we apply a threshold on the minimal eigenvalue of the covariance matrix of image derivatives to obtain dense keypoints. We also ensure that keypoints are not duplicated among different spatial layers, and that a minimum distance between each couple of points is preserved. Each keypoint $P_t = (x_t, y_t)$ is then tracked by the means of median filtering with kernel M in a dense optical flow field $\omega = (u_t, v_t)$:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\bar{x}_t, \bar{y}_t)} \qquad (1)$$

where $(\bar{x}_t, \bar{y}_t)$ is the rounded position of P_t. Differently from [20], our trajectories are calculated under the constraint that they lie inside and around the user's hand: at each frame the hand mask is dilated and all keypoints still outside are discarded.

A spatio-temporal volume aligned with each trajectory is then build, as a collection of 32×32 patches around the keypoint. Then, Trajectory descriptor, HOG, HOF and MBH are computed inside the volume. We introduce a difference in how to weight the temporal volume of each component of our feature vector: while HOF and MBH are averaged on five consecutive frames, a single HOG descriptor is computed for each frame. This allows us to describe the changes in the hand pose at a finer temporal granularity. This step results in a variable number of descriptors for each video sequence. To obtain a fixed size descriptor, we exploit the Bag of Words approach training four separate codebooks, one for each descriptor. Each codebook contains K visual words (in the experiments we fix $K = 500$) and is obtained running the k-means algorithm on the training data.

Since the histograms obtained from the Bag of Words in our domain tend to be sparse, they are power normalized to unsparsify the representation, while still allowing for linear classification. To perform power-normalization [28], the function:

$$f(h_i) = \text{sign}(h_i) \cdot |h_i|^{\frac{1}{2}} \qquad (2)$$

is applied to each bin h_i in our histograms.

The final descriptor is then obtained by the concatenation of its four power-normalized histograms. Finally, gestures are recognized using a linear SVM 1-vs-1 classifier.

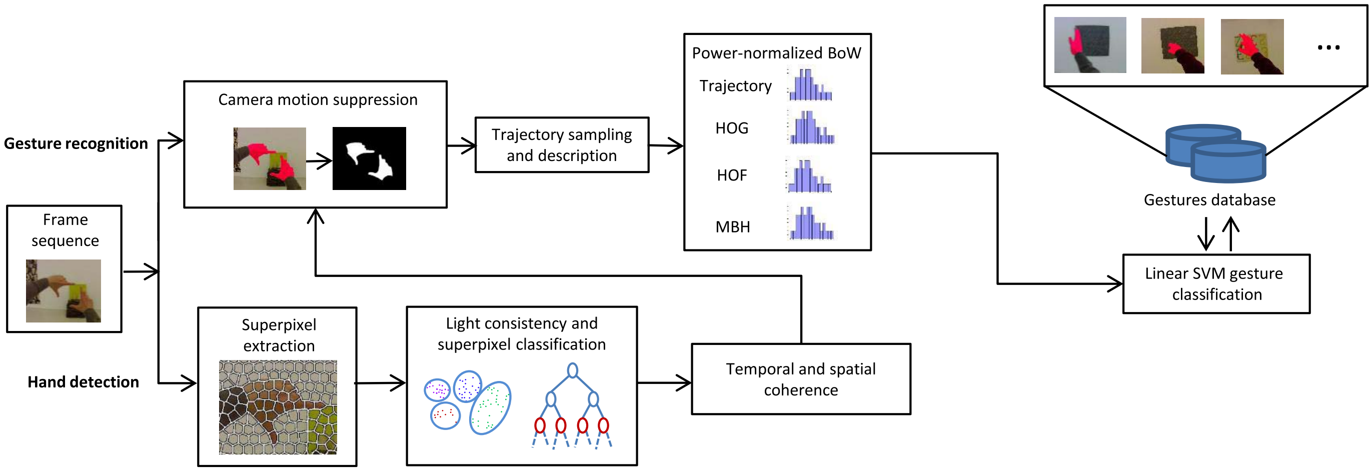

Fig. 5. An outline of the proposed gesture recognition module. It is roughly composed by three steps: the first step consists of hand segmentation and feature extraction, the second step performs BoW coding, the third step is the classification enhanced by past visitors' gestures.

B. Hand Segmentation

As stated before, a hand segmentation mask is used to distinguish between camera and hand motions, and to prune away all the trajectories that do not belong to the user's hand. In this way, our descriptor captures hands movement and shape as if the camera was fixed, and disregards the noise coming from other moving regions that could be in the scene.

At each frame we extract superpixels using the SLIC algorithm [29], that performs a k-means-based local clustering of pixels in a 5-dimensional space, where color and pixel coordinates are used. Superpixels are then represented with several features: histograms in the HSV and LAB color spaces (that have been proven to be good features for skin representation [11]), Gabor filters and a simple histogram of gradients, to discriminate between objects with a similar color distribution.

1) Illumination Invariance: To deal with different illumination conditions, we cluster the training images running the k-means algorithm on a global HSV histogram. Hence, we train a Random Forest classifier for each cluster. By using a histogram over all three channels of the HSV color space, each scene cluster encodes both the appearance of the scene and its illumination. Intuitively, this models the fact that hands viewed under similar global appearance will share a similar distribution in the feature space. Given a feature vector $\mathbf{l}$ of a superpixel $\mathbf{s}$ and a global appearance feature $\mathbf{g}$, the posterior distribution of $\mathbf{s}$ is computed by marginalizing over different clusters c:

$$P(\mathbf{s}|\mathbf{l}, \mathbf{g}) = \sum_{c=1}^{k} P(\mathbf{s}|\mathbf{l}, c)P(c|\mathbf{g}) \tag{3}$$

where k is the number of clusters, $P(\mathbf{s}|\mathbf{l}, c)$ is the output of the cluster-specific classifier and $P(c|\mathbf{g})$ is a conditional distribution of a cluster c given a global appearance feature $\mathbf{g}$. In test phase, the conditional $P(c|\mathbf{g})$ is approximated using an uniform distribution over the five nearest clusters. It is important to highlight that the optimal number of classifiers depends on the characteristics of the dataset: a training dataset with several different illumination conditions, taken both inside and outside, will need an higher number of classifiers than one taken indoor. In addition, we model the hand appearance not only considering illumination variations, but also including semantic coherence in time and space.

2) Temporal Coherence: To improve the foreground prediction of a pixel in a frame, we replace it with a weighted combination of its previous frames, since past frames should affect the prediction for the current frame.

We define a smoothing filter for a pixel x_t^i from frame t as:

$$\begin{aligned} P(x_t^i = 1) = \sum_{k=0}^{\min(t,d)} & w_k (P(x_t^i = 1|x_{t-k}^i = 1) \\ & \cdot P(x_{t-k}^i = 1|\mathbf{l}_{t-k}, \mathbf{g}_{t-k}) + P(x_t^i = 1|x_{t-k}^i = 0) \\ & \cdot P(x_{t-k}^i = 0|\mathbf{l}_{t-k}, \mathbf{g}_{t-k})) \end{aligned} \tag{4}$$

where d is the number of past frames used, and $P(x_{t-k}^i = 1|\mathbf{l}_{t-k}, \mathbf{g}_{t-k})$ is the probability that a pixel in frame $t-k$ is marked as hand part, equal to $P(\mathbf{s}|\mathbf{l}_{t-k}, \mathbf{g}_{t-k})$, being x_t^i part of $\mathbf{s}$. In the same way, $P(x_{t-k}^i = 0|\mathbf{l}_{t-k}, \mathbf{g}_{t-k})$ is defined as $1 - P(\mathbf{s}|\mathbf{l}_{t-k}, \mathbf{g}_{t-k})$. Last, $P(x_t^i = 1|x_{t-k}^i = 1)$ and $P(x_t^i = 1|x_{t-k}^i = 0)$ are prior probabilities estimated from the training set as follows:

$$\begin{aligned} P(x_t^i = 1|x_{t-k}^i = 1) &= \frac{\#(x_t^i = 1, x_{t-k}^i = 1)}{\#(x_{t-k}^i = 1)} \\ P(x_t^i = 1|x_{t-k}^i = 0) &= \frac{\#(x_t^i = 1, x_{t-k}^i = 0)}{\#(x_{t-k}^i = 0)} \end{aligned} \tag{5}$$

where $\#(x_{t-k}^i = 1)$ and $\#(x_{t-k}^i = 0)$ are the number of times in which x_{t-k}^i belongs or not to a hand region, respectively; $\#(x_t^i = 1, x_{t-k}^i = 1)$ is the number of times that two pixels at the same location in frame t and $t-k$ belong to a hand part; similarly $\#(x_t^i = 1, x_{t-k}^i = 0)$ is the number of times that a pixel in frame t belongs to a hand part and the pixel in the same position in frame $t-k$ does not belong to a hand region. Based on our preliminary experiments we set d equal to three.

3) Spatial Consistency: Given pixels elaborated by the previous steps, we want to exploit spatial consistency to prune away small and isolated pixel groups that are unlikely to be part of hand regions and also aggregate bigger connected pixel groups. For every pixel x, we extract its posterior probability $P(x_i^t)$ and use it as input for the GrabCut algorithm [30]. Each pixel with $P(x_i^t) \geq 0.5$ is marked as foreground, otherwise it's considered as part of background. After the segmentation step, we discard all the small isolated regions that have an area of less than 5% of the frame and we keep only the three largest connected components.

C. Artwork Recognition

The second component of our system is artwork recognition: a matching is established between the framed artwork and its counterpart on the system database. The real-world ego-vision setting we are dealing with makes this task full of challenges: paintings in a museum are often protected by reflective glasses or occluded by other visitors and even by user's hands, requiring a method capable of dealing with these difficulties too.

For this reason, we follow common approaches of object recognition based on interest points and local descriptors [31], [32], that have been proved to be able to capture sufficiently discriminative local elements and are robust to large occlusions.

First of all, SIFT keypoints are extracted from the whole image. The need to proceed with this approach instead of sampling from a detected area derives from the difficulties that arise when trying to detect paintings from a first person perspective. Detection based on shape resulted in high false positive rate, hence we rely on sampling over the whole image. To improve the match quality, we process the matched keypoints using the RANSAC algorithm. The ratio between the remaining matches and the total number of keypoints is then thresholded, allowing to recognize if the two images refer to the same artwork even in presence of partial occlusions. In addition, to avoid occlusions with user's hands we perform artwork recognition on the frames captured before the recognized gesture using a temporary buffer.

IV. Experimental Evaluation

To evaluate the performance of our gesture recognition and hand segmentation algorithms we first compare them with existing approaches. In particular we test our gesture module on the Cambridge-Gesture database [33], which includes nine hand gesture types performed on a table, under different illumination conditions. Whereas to evaluate the hand segmentation approach, we test it on the publicly available CMU EDSH dataset [12] which consists of three ego-centric videos with indoor and outdoor scenes and large variations of illuminations.

Furthermore, to investigate the effectiveness of the proposed approach in videos taken from the ego-centric perspective and in a museum setting, we also propose and release publicly two realistic and challenging datasets recorded in in an interactive exhibition room, which functions as a virtual museum, and a real museum of Modern Art. Finally, we perform a performance evaluation of the proposed algorithms on one of our wearable devices.

Fig. 6. Sample images from the Cambridge hand gesture dataset.

TABLE I
RECOGNITION RATES ON THE CAMBRIDGE DATASET

Method	Set1	Set2	Set3	Set4	Overall
TCCA [16]	0.81	0.81	0.78	0.86	0.82
PM [14]	0.89	0.86	0.89	0.87	0.88
TB [15]	**0.93**	0.88	0.90	0.91	0.91
Cov3D [13]	0.92	**0.94**	0.94	0.93	0.93
Our method	0.92	0.93	**0.97**	**0.95**	**0.94**

A. Cambridge Hand Gesture Dataset

The Cambridge Hand Gesture dataset contains 900 sequences of nine hand gesture classes. Although this dataset does not contain ego-vision videos it is useful to compare our results with recent gesture recognition techniques. In particular, each sequence is recorded with a fixed camera, placed over one hand, and hands perform leftward and rightward movements on a table, with different poses (see Figure 6). The whole dataset is divided in five sets, each of them containing image sequences taken under different illumination conditions. The common test protocol, proposed in [33], requires to use the set with normal illumination for training and the remaining sets for testing, thus we use the sequences taken in normal illumination to generate the BoW codebooks and to train the SVM classifier. Then, we perform the test using the remaining sequences.

Table I shows the recognition rates obtained with our gesture recognition approach, compared with the ones of tensor canonical correlation analysis (TCCA) [16], product manifolds (PM) [14], tangent bundles (TB) [15] and spatio-temporal covariance descriptors (Cov3D) [13]. Results show that the proposed method is effective in recognizing hand gestures, and that it outperforms the existing state-of-the-art approaches.

B. EDSH Hand Segmentation Dataset

The CMU EDSH dataset consists of three ego-centric videos (EDSH1, EDSH2, EDSHK) containing indoor and outdoor scenes where hands are purposefully extended outwards to capture the change in skin color. As this dataset does not contain any gesture annotation, we use it to evaluate only the hand segmentation part.

We validate the techniques that we have proposed for temporal and spatial consistency. In Table II we compare the performance of the hand segmentation algorithm in terms of F1-measure, firstly using a single Random Forest classifier, and then incrementally adding illumination invariance,

TABLE II

PERFORMANCE COMPARISON CONSIDERING ILLUMINATION INVARIANCE (II), TEMPORAL COHERENCE (TC) AND SPATIAL CONSISTENCY (SC)

Features	EDSH2	EDSHK
Single RF classifier	0.761	0.829
II	0.789	0.831
II + TC	0.791	0.834
II + TC + SC	**0.852**	**0.901**

TABLE III

HAND SEGMENTATION COMPARISON WITH THE STATE-OF-THE-ART

Method	EDSH2	EDSHK
Hayman and Eklundh [34]	0.211	0.213
Jones and Rehg [35]	0.708	0.787
Li and Kitani [12]	0.835	0.840
Our method	**0.852**	**0.901**

the temporal smoothing filter and the spatial consistency technique via the GrabCut algorithm application. Results shows that there is a significant improvement in performance when all three techniques are used together: illumination invariance increases the performance with respect to the results obtained using only a single Random Forest classifier, while temporal smoothing and spatial consistency correct incongruities between adjacent frames, prune away small and isolated pixel groups and merge spatially nearby regions, increasing the overall performance.

Then, in Table III we compare our segmentation method with different techniques: a video stabilization approach based on background modeling [34], a single-pixel color method inspired by [35] and the approach proposed in [12] by Li *et al.*, based on a collection of Random Forest classifiers. As can be seen, the single-pixel approach, which basically uses a random regressor trained only using the single pixel LAB values, is still quite effective, even if conceptually simple. Moreover, we observe that the video stabilization approach performs poorly on this dataset, probably because of the large ego-motions these video present. The method proposed by Li *et al.* is the most similar to our approach, nevertheless exploiting temporal and spatial coherence we are able to outperform their results.

C. Virtual and Real Museum Environments

We propose two new gesture recognition datasets taken from the ego-centric perspective in virtual and real museum environments. The Interactive Museum dataset consists of 700 video sequences, all shot with a wearable camera, taken in a interactive exhibition room, in which paintings and artworks are projected over a wall in a virtual museum fashion (see Figure 7). The camera is placed on the user's head and captures a 800 × 450, 25 frames per second 24-bit RGB image sequence. Five different users perform seven hand gestures: *like*, *dislike*, *point*, *ok*, *slide left to right*, *slide right to left* and *take a picture*. Some of them (like the *point*, *ok*, *like* and *dislike* gestures) are statical, others (like the two *slide* gestures) are dynamical. We have publicly released the dataset.[1]

[1] http://imagelab.ing.unimore.it/files/ego_virtualmuseum.zip

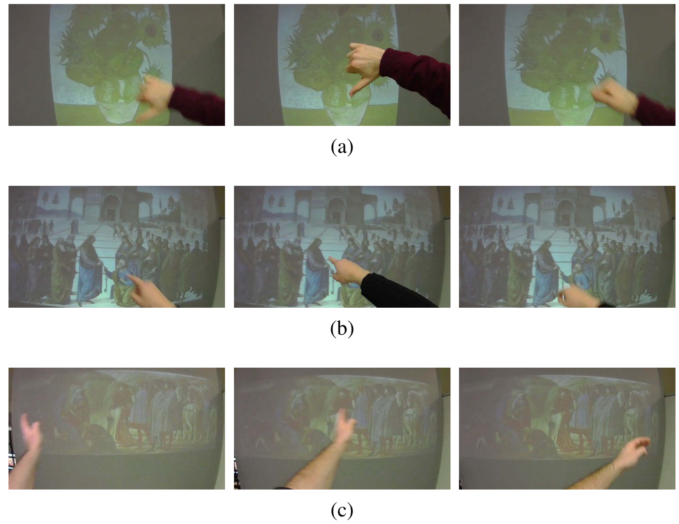

Fig. 7. Gestures from the Interactive Museum dataset. (a) *Dislike* gesture. (b) *Point* gesture. (c) *Slide left to right* gesture.

TABLE IV

GESTURE RECOGNITION ACCURACY ON THE INTERACTIVE MUSEUM DATASET WITH AND WITHOUT HAND SEGMENTATION

User	No segmentation	With segmentation
User A	0.91	**0.95**
User B	**0.96**	0.94
User C	0.91	**0.96**
User D	0.87	**0.87**
User E	0.92	**0.95**
Average	0.91	**0.93**

Since ego-vision applications are highly interactive, their setup step must be fast (i.e. few positive examples can be acquired). Therefore, to evaluate the proposed gesture recognition approach, we train a 1-vs-1 linear classifier for each user using only two randomly chosen gestures per class as training set.

In Table IV we show the gesture recognition accuracy for each of the five subjects of the Interactive Museum dataset. To validate the proposed technique, that combines gesture recognition and hand segmentation, we also show the results obtained without the use of the hand segmentation mask. As can be seen, our approach is well suited to recognize hand gestures in the ego-centric domain, even using only two positive samples per gesture, and the use of the segmentation mask for camera motion removal and trajectories pruning can improve recognition accuracy. The reported results are the average over 100 independent runs.

On a different note, to test our approach in a real setting, we created a dataset with videos taken in the Maramotti modern art museum, in which paintings, sculptures and *objets d'art* are exposed. As in the previous dataset, the camera is placed on the user's head and captures a 800 × 450, 25 frames per second image sequence. The Maramotti dataset contains 700 video sequences, recorded by five different persons (some are the same of the Interactive Museum dataset), each performing the same gestures as before in front of different artworks.

(a) (b)

(c) (d)

Fig. 8. Gestures from the Maramotti dataset. (a) *Like* gesture. (b) *Ok* gesture, in low light. (c) *Slide right to left* gesture, while another visitor walks in. (d) *Take a picture* gesture.

TABLE V

GESTURE RECOGNITION ACCURACY ON THE MARAMOTTI DATASET

User	Single user's Gestures	Augmented
User A	0.54	**0.65**
User B	0.52	**0.72**
User C	0.68	**0.68**
User F	0.56	**0.79**
User G	0.53	**0.72**
Average	0.57	**0.71**

We are currently waiting for the permission to release this dataset from the Maramotti museum.

Figures 7 and 8 show some examples of gestures performed in the two datasets. In the Interactive Museum dataset, users perform gestures in front of a wall over which the works of art are projected. This setting is quite controlled: the illumination is constant, the art works are in low light, while hands are well illuminated. On the other hand, in the Maramotti dataset, users perform gestures in front of real artworks inside a museum. This is a realistic and very challenging environment: the illumination changes, other visitors are present and sometimes walk in. In both cases there is significant camera motion, because the camera moves as the users move their heads or arms. It is also important to underline that users have not been trained before recording their gestures, so each user performs the gestures in a slightly different way, as would happen in a realistic context.

In Table V we show the results of our gesture recognition approach on the Maramotti dataset. As can be seen, in this case the challenging and real environment causes a drop in accuracy. This is mainly due to the illumination changes, to the presence of other visitors, and to the fact that often the artworks are better illuminated than hands. Since our wearable vision devices is fully connected to a central server, we show how the use of other visitors' gestures can improve the recognition accuracy. In our scenario each visitor coming to the museum performs, in the initial setup phase, two training gestures for each class. These training gestures from past visitors, manually checked, are used to augment the training set, so no erroneous data is accumulated into the model. In particular, in our test "Augmented" (Table V) each ego-vision wearable device uses two randomly chosen gestures performed by its user as training, plus gestures performed by the remaining four users supplied by their devices to the central server. Results show that this distributed approach is effective and leads to a significant improvement in accuracy.

D. Performance Evaluation

In this section we present our gesture recognition approach performance and optimizations. They are evaluated on the Hardkernel Odroid-XU board, already introduced in Section III. The tests we further present are performed on the Maramotti dataset. To evaluate the performance of our gesture recognition application, we split our algorithm in five main sub-modules (already deeply explained in the previous sections): Hand Segmentation, Camera motion removal, Trajectory extraction, Trajectory description, Power-normalized BoW and SVM-based Classification. To reach good performance on the Odroid-XU embedded device we applied different optimization techniques. Firstly compiler optimization has been used to speed-up code execution adding -O3 to compilation flags. Then we used Neon optimized instructions, by including neon library in source code and using these flags at compile time: -mfpu=neon-vfpv4 -mfloat-abi=hard -mtune=cortex-a15 -marm. Several low level "for cycles" have been balanced on different processors using OpenMP parallel regions. In Figure 9 we show the impact of each sub-module, separately, to elaborate 38 frames, that is the average gesture length within the Maramotti dataset. On the bottom part of each column we report the number of times each sub-module is called.

As can be seen, the Hand Segmentation is by far the most time consuming sub-module compared to the others. This is also due to the number of times each of sub-module is called: while Classification and Power-normalized BoW are executed just one time per gesture, the others are called one time per frame.

Therefore, we studied the performance-accuracy tradeoff of hand segmentation introducing a frame step between subsequent elaborations. The idea is to benefit of the hand segmentation not on each frame, but to introduce a gap between segmentation processing of the video stream and see how this impact on the gesture recognition accuracy. In this case, the hand segmentation mask is computed

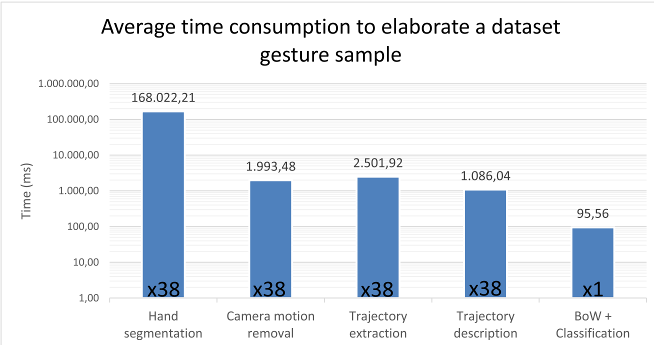

Fig. 9. Average time consumption of each submodule to elaborate a gesture sample from the Maramotti dataset.

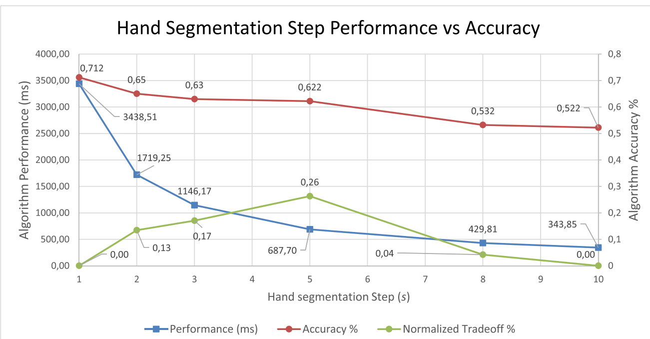

Fig. 10. Performance-accuracy trade-off of the proposed gesture recognition approach with different Hand Segmentation frame steps.

every s frames. Trajectories and descriptors are still computed using all frames, but new keypoints are sampled only when the hand segmentation mask is available. The choice of higher values of s results in an increasing loss of accuracy, since the gesture is described with less trajectories, but in a decreasing computational power needed to process the input.

Figure 10 summarizes the whole gesture recognition algorithm performance and accuracy, applying different hand segmentation frame steps. We evaluated it as an average of the five Maramotti subjects, and the execution step of the Hand Segmentation is evaluated on the average length of the dataset samples (38 frames).

Three lines are shown in the graph: accuracy, performance and the normalized tradeoff. This last line has been computed as plain multiplication of normalized accuracy by normalized performance. The best normalized tradeoff is given by a step size of 5 frames. The average hands segmentation accuracy decreases of 9% (from 71.2% to 62.2%) in a tradeoff with a speed-up of 5x. This is a good result for performance, because paying a 9% accuracy loss we reduce the execution time from *3438.51 ms* to *687.70 ms*.

In Table VI we show a summary of the performances obtained with different step sizes. As can be seen, the best computational performance on Odroid-XU platform is reached when using a step size of 10, and paying an accuracy loss of about 19% (from 71.2% to 52.2%). Based on this analysis, we can state that our gesture recognition with hand

TABLE VI
GESTURE RECOGNITION PERFORMANCE WITH DIFFERENT STEP SIZES

Step size	ms per frame	Frame/second	Accuracy
s = 1	3438.51	0.29	71,2 %
s = 5	687.70	1.45	62,2 %
s = 10	343.85	2.91	52,2 %

segmentation is sufficiently accurate for real-life deployment and runs with an acceptable computation performance on ARM-based embedded devices.

V. CONCLUSION

We described a novel approach to cultural heritage fruition based on ego-centric vision devices. Our work is motivated by the increasing interest in ego-centric vision and by the growth of the cultural market, which encourages the development of new interfaces to interact with the cultural heritage. We presented a gesture and painting recognition model that can deal with static and dynamic gestures and can benefit from a distributed training. Our gesture recognition and hand segmentation results outperform the state-of-the-art approaches on Cambridge Hand Gesture and CMU EDSH datasets. Finally, we ran an extensive performance analysis of our system on a wearable board.

ACKNOWLEDGMENT

The authors would like to thank Collezione Maramotti for granting the use of their space in order to test our system in a realistic scenario.

REFERENCES

[1] *How the Americans Will Travel 2015*. [Online]. Available: http://tourism-intelligence.com, accessed Apr. 7, 2014.
[2] *Economic Impact of Travel & Tourism 2014*. World Travel and Tourism Council, London, U.K., 2014.
[3] A. Kuusik, S. Roche, and F. Weis, "SMARTMUSEUM: Cultural content recommendation system for mobile users," in *Proc. 4th Int. Conf. Comput. Sci. Converg. Inf. Technol. (ICCIT)*, Nov. 2009, pp. 477–482.
[4] T. Kuflik *et al.*, "A visitor's guide in an active museum: Presentations, communications, and reflection," *J. Comput. Cultural Heritage*, vol. 3, no. 3, p. 11, Mar. 2011.
[5] F. Sparacino, "The museum wearable: Real-time sensor-driven understanding of visitors' interests for personalized visually-augmented museum experiences," in *Proc. Museums Web*, 2002, pp. 17–20.
[6] T. Kanade and M. Hebert, "First-person vision," *Proc. IEEE*, vol. 100, no. 8, pp. 2442–2453, Aug. 2012.
[7] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *Proc. CVPR*, Jun. 2012, pp. 2847–2854.
[8] S. Sundaram and W. W. M. Cuevas, "High level activity recognition using low resolution wearable vision," in *Proc. CVPR*, Jun. 2009, pp. 25–32.
[9] A. Fathi, X. Ren, and J. M. Rehg, "Learning to recognize objects in egocentric activities," in *Proc. CVPR*, Jun. 2011, pp. 3281–3288.
[10] A. Fathi and J. M. Rehg, "Modeling actions through state changes," in *Proc. CVPR*, Jun. 2013, pp. 2579–2586.
[11] R. Khan, A. Hanbury, and J. Stoettinger, "Skin detection: A random forest approach," in *Proc. 17th IEEE ICIP*, Sep. 2010, pp. 4613–4616.
[12] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *Proc. CVPR*, Jun. 2013, pp. 3570–3577.
[13] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell, "Spatio-temporal covariance descriptors for action and gesture recognition," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2013, pp. 103–110.
[14] Y. M. Lui, J. R. Beveridge, and M. Kirby, "Action classification on product manifolds," in *Proc. CVPR*, Jun. 2010, pp. 833–839.

[15] Y. M. Lui and J. R. Beveridge, “Tangent bundle for human action recognition,” in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 97–102.
[16] T.-K. Kim and R. Cipolla, “Canonical correlation analysis of video volume tensors for action categorization and detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 8, pp. 1415–1428, Aug. 2009.
[17] P. Mistry and P. Maes, “SixthSense: A wearable gestural interface,” in *Proc. ACM SIGGRAPH ASIA Sketches*, 2009, p. 11.
[18] *Odroid-XU Development Board by Hardkernel*. [Online]. Available: http://www.hardkernel.com, accessed Apr. 7, 2014.
[19] *Samsung Exynos5 5410 ARM CPU*. [Online]. Available: http://www.samsung.com/global/business/semiconductor/minisite/Exynos/products-5octa_5410.html, accessed Apr. 7, 2014.
[20] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, “Action recognition by dense trajectories,” in *Proc. CVPR*, Jun. 2011, pp. 3169–3176.
[21] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 886–893.
[22] N. Dalal, B. Triggs, and C. Schmid, “Human detection using oriented histograms of flow and appearance,” in *Computer Vision*. Berlin, Germany: Springer-Verlag, 2006, pp. 428–441.
[23] M. A. Fischler and R. C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981.
[24] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features,” in *Proc. ECCV*, 2006, pp. 404–417.
[25] G. Farnebäck, “Two-frame motion estimation based on polynomial expansion,” in *Image Analysis*. Berlin, Germany: Springer-Verlag, 2003, pp. 363–370.
[26] C. Zach, T. Pock, and H. Bischof, “A duality based approach for realtime TV-L^1 optical flow,” in *Pattern Recognition*. Berlin, Germany: Springer-Verlag, 2007, pp. 214–223.
[27] M. Tao, J. Bai, P. Kohli, and S. Paris, “SimpleFlow: A non-iterative, sublinear optical flow algorithm,” *Comput. Graph. Forum*, vol. 31, no. 2, pp. 345–353, 2012.
[28] F. Perronnin, J. Sánchez, and T. Mensink, “Improving the Fisher kernel for large-scale image classification,” in *Proc. ECCV*, 2010, pp. 143–156.
[29] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, “SLIC superpixels compared to state-of-the-art superpixel methods,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[30] C. Rother, V. Kolmogorov, and A. Blake, “‘GrabCut’: Interactive foreground extraction using iterated graph cuts,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004,
[31] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” in *Proc. ICCV*, Oct. 2003, pp. 1470–1477.
[32] A. D. Bagdanov, L. Ballan, M. Bertini, and A. D. Bimbo, “Trademark matching and retrieval in sports video databases,” in *Proc. ACM Int. Workshop Multimedia Inf. Retr. (MIR)*, 2007, pp. 79–86.
[33] T.-K. Kim, K.-Y. K. Wong, and R. Cipolla, “Tensor canonical correlation analysis for action classification,” in *Proc. CVPR*, Jun. 2007, pp. 1–8.
[34] E. Hayman and J.-O. Eklundh, “Statistical background subtraction for a mobile observer,” in *Proc. ICCV*, Oct. 2003, pp. 67–74.
[35] M. J. Jones and J. M. Rehg, “Statistical color models with application to skin detection,” *Int. J. Comput. Vis.*, vol. 46, no. 1, pp. 81–96, 2002.

Lorenzo Baraldi received the B.Sc. and M.Sc. degrees in computer engineering from the University of Modena and Reggio Emilia, in 2011 and 2014, respectively, where he is currently pursuing the Ph.D. degree at the Imagelab Laboratory. His research interests include egocentric vision, video understanding, and deep learning.

Francesco Paci received the B.Sc. and M.Sc. degrees in computer engineering from the University of Bologna, where he is currently pursuing the Ph.D. degree in electronics, telecommunications, and information technologies. He was a Visiting Student with Trinity College Dublin, a Visiting Researcher with Movidius, Dublin, in 2014, and STMicroelectronics, Grenoble, in 2012. His research areas of interest are computer vision algorithms developed for embedded devices, graphical processing unit and many-core software acceleration, and thermal/power-aware scheduling.

Giuseppe Serra received the Laurea degree in computer engineering and the Ph.D. degree in computer engineering, multimedia, and telecommunication from the University of Florence, in 2006 and 2010, respectively. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, and Telecom ParisTech/ENST, Paris, in 2006 and 2010, respectively. He is currently an Assistant Professor with the University of Modena and Reggio Emilia, Italy. He has authored over 40 publications in scientific journals and international conferences. His research interests include egocentric vision, image and video analysis, multimedia ontologies, image forensics, and multiple view geometry. He was a recipient of the Best Paper Award from the ACM-SIGMM Workshop on Social Media in 2010. He co-organized the First International Workshop on Wearable and Ego-Vision Systems for Augmented Experience in conjunction with ICME 2015.

Luca Benini (F’07) is currently a Full Professor with the University of Bologna and the Chair of Digital Circuits and Systems at ETHZ. He has served as the Chief Architect for the Platform 2012/STHORM project in STmicroelectronics, Grenoble, from 2009 to 2013. He has held visiting and consulting researcher positions at EPFL, IMEC, HewlettPackard Laboratories, and Stanford University. His research interests are in energy efficient system design and MultiCore SoC design. He is also active in the area of energy efficient smart sensors and sensor networks for biomedical and ambient intelligence applications. He has authored over 700 papers in peer reviewed international journals and conferences, four books and several book chapters. He is a member of the Academia Europaea.

Rita Cucchiara (M’15) received the M.S. degree in electronics engineering, in 1989, and the Ph.D. degree in computer engineering from the University of Bologna, in 1992. Since 2011, she has been scientific responsible for the ICT platform in the High Technology TV Network of Emilia Romagna Region. She has been a Coordinator of several projects in video surveillance mainly for people detection, tracking, and behavior analysis. She is currently a Full Professor of Computer Engineering with the University of Modena and Reggio Emilia. She is the Director of the Research Center with Softech-ICT, and heads the Imagelab Laboratory. In the field of multimedia, she works on annotation, retrieval, and human-centered searching in images and video big data for cultural heritage. She has authored over 300 papers in journals and international proceedings, and is a reviewer for several international journals. Her research interests include pattern recognition, computer vision, and multimedia. She is a member of the IEEE Computer Society, the Association for Computing Machinery, the Gruppo Italiano Ricercatori in Pattern Recognition, and the Italian Association for Artificial Intelligence. Since 2006, she has been a fellow of the International Association for Pattern Recognition.