Noname manuscript No.
(will be inserted by the editor)

Interacting with Art: Ego-vision for Enriched Cultural Experience

Lorenzo Baraldi · Stefano Alletto · Giuseppe Serra · Rita Cucchiara

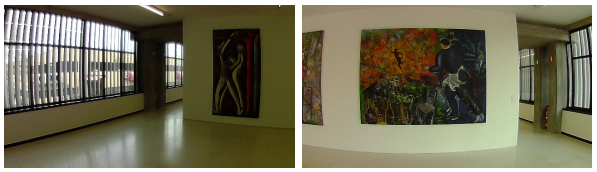
Received: date / Accepted: date

Abstract Everyone has been to a museum once. We all quarrelled with those self-service audio-guides at least once, striving to find a correspondence between what we were looking at and what the guide was trying to explain. Now imagine if we had a device capable of seeing what we saw to which we could have asked for the information we wanted by just pointing with a finger to the artwork. In this paper we provide a novel approach to cultural heritage experience: by the means of an ego-vision embedded platform we develop an approach that aims to a new, more entertaining and multimedia way of accessing museum knowledge. Our proposal deals with two main challenges: gesture recognition and painting recognition.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. We extensively test our approach on two publicly available datasets and we further extend our experiments to both virtual and real museum scenarios where our method shows robustness when challenged with real-world data.

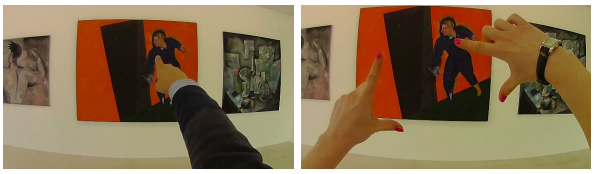
Keywords Wearable vision, interactive museum, embedded systems, gesture recognition, natural interfaces.

1 Introduction

The rebirth of interest in cultural heritage is an unequivocal fact. Thanks to the internet and to social media, the cultural market is becoming a key-point in many national economics strategies: the importance of the cultural sector is traditionally straightforward in Europe, but it is equally high also in other parts of the world. In fact, the Office of Travel and Tourism Industries in USA affirms that half of the Americans travelling abroad visit historical places; almost one-third visit cultural heritage sites; and one-quarter go to an art gallery or museum [1]. To deal with the new behaviours of tourists and art lovers which are more and more digital natives, the cultural places, archaeological sites, museums and exhibitions must deal with the need of new multimedia technologies. There is a big effort in proposing and providing new natural, attractive and immersive interfaces for interacting with the cultural heritage. Visitors need new enhanced experiences in cultural heritage [6]: for example singular and social experiences using mobile and smartphones [32] or in



(a) First-person viewpoint in museum experience



(b) Natural interaction with artwork

Fig. 1: Sample gestures from a real-world contemporary art museum.

L. Baraldi, S. Alletto, G. Serra, R. Cucchiara
Università degli Studi di Modena e Reggio Emilia
Dipartimento di Ingegneria "Enzo Ferrari"
Tel.: +390592056265
Fax: +390592056129
E-mail: {name.surname}@unimore.it

the crowd using crowd sourcing [26]. Further efforts are made in the digitalization of available cultural heritage objects like paintings or sculptures [13,16].

The multimedia research community is largely involved in designing new devices that go beyond standard paper books or audio guides, such as touch interfaces [4], interactive interfaces [20,36] and smartphones with games [5].

In this paper, we present our research in a new emerging technology, namely ego-vision for cultural experience in smart museums. Ego-vision features glass-mounted wearable cameras able to see what the visitor sees and perceiving the surrounding environment as he does. Since small low cost cameras with high processing capabilities are becoming available, these systems will be largely spread in the near future. For this reason, we propose their use in enhanced cultural experiences, making users interact with physical or digital cultural products to improve their fruition of information and to share knowledge on the social community. The technology that we present, being based on scalable wearable devices capable of communicating both with each other and the rest of the world, will be open to further integration with the internet-of-things and sensors such as RFID, GPS and gyroscopes. However, the research in ego-vision is still in its beginning and presents several challenges such as processing video recorded from an unconstrained prospective. In this work, we provide techniques that performs (i) painting recognition in a museum to achieve content-awareness, (ii) self-gesture analysis in order to recognize user interaction with artworks providing an natural and enriched interface to cultural heritage. Figure 1 shows an example of both the setting (Figure 1a) of our experiments and a real-world gesture analysis scenario (Figure 1b).

Our main novelties and contributions are the enabling technologies for multimedia interaction. In particular we propose robust algorithms for gesture analysis based on trajectory and shape information classification and painting recognition, tested both in real-world and virtual museums (see Figure 2. These methods have been compared to the current state of the art techniques over two public datasets showing our superior performance. Moreover, thanks to the very limited training requirements of our method, it can be extended to understand many other gestures not originally included. We also propose a painting recognition technique capable of classifying artworks despite the challenging conditions of real world museums like ever-changing lighting conditions, fast camera motion and partial occlusions.

The paper is structured as follows: in the next section, some related works for both multimedia technologies in cultural experiences and ego-vision systems are reported. In section 3 we present the details of the proposed solutions for painting recognition and self-gesture analysis. In section 4 results are shown and in conclusion some final considerations and proposals for future work are described.

2 Related Work

The present section explores the current state of the art in the two main areas of interest touched by our work: enhanced cultural experience in museums and egocentric vision.

Museums are traditionally spaces that have, by their very nature, an abundance of information available to visitors. In many cases, objects are accompanied by textual descriptions, usually too short or long to be adequate to the cultural interests of all visitors. Because of this, visitor access to museum collections can be often unsatisfactory or not appealing, especially to new generations. Personalization of multimedia museum content is one answer to this problem [27]. Personalization offers visitors a customized presentation of appropriate information related to the visitor's tastes and preferences. For these reasons, many solutions have been recently proposed for interactive user-profile based guides. An example is the work "SmartMuseum" [22]: by means of PDAs and RF-IDs, 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 approach brought an interesting novelty when first released, but it has some very limiting flaws. First of all, being tied to RFIDs does not allow reconfiguring the museum without rethinking the entire structure of the knowledge on which the project is based. Furthermore, researches demonstrated how the use of PDAs 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. Similarly in [21], authors proposed to customize museums experiences with machine learning techniques applied on the answer to questionnaires that the users should compile before the visit. In both proposals the main flaw is the need to invasive interaction, asking the visitors to do something that probably they would not want to do. One of the valuable attempts to user profiling with wearable sensors was the "Museum Wearable" [31], a wearable computer which orchestrates an audio-visual narration as a function of the visitors' interests gathered from his/her physical path in the museum. However this prototype does not use any visual understanding algorithms for understanding the surrounding environment. For instance the estimation of the visitor location is based again on infrared sensors distributed

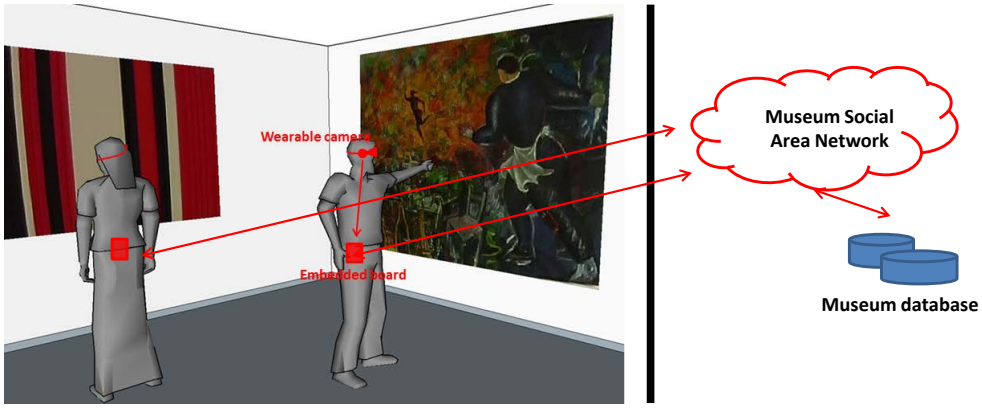


Fig. 2: An example of the interaction architecture of our proposal.

in the museum space. On a different note, the work by Yanulevskaya et al. [35] proposes a framework to automatically classify abstract paintings based on the emotional response they trigger in the visitors. Using statistical analysis and eye tracking devices, they calculate a score of each painting based on the emotions it provokes.

The ego-vision scenario has been addressed only recently by the research community and mainly to understand human activities and to recognize hand regions. Pirsiavash et al. [29] detected activities of daily living using an approach that involves temporal pyramids and object detectors tuned for objects appearance during interactions and spatial reasoning. Sundaram et al. [33] 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. [11] 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 [17] studied color classification for skin segmentation. They pointed out how color-based skin detection has many advantages and potentially high processing speed, and demonstrated that Random Forest is one of the best classifiers for skin segmentation. Fathi et al. [11] proposed a different approach to hand detection, based on the assumption that background is static in the world coordinate frame, thus foreground objects are detected as to be the moving regions respect to the background. This approach is shown to be a robust tool for skin detection and hand segmentation in indoor environments, even if it performs poorly with more unconstrained scenarios. Li et al. [23] 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.

The gesture analysis domain in ego-vision is rather unexplored. Even though not related to ego-vision domain, several approaches to gesture and human action recognition have been proposed. Kim et al. [18] extended Canonical Correlation Analysis to measure video-to-video similarity in order to represent and detect actions in video. Lui et al. [25, 24] used tensors and tangent bundle on Grassmann manifolds to classify human actions and hand gestures. Sanin et al. [30] developed a new and more effective spatio-temporal covariance descriptor to classify gestures in conjunction with a boost classifier. 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, hand presence and background cluttering.

3 Ego-vision for cultural heritage

We present the techniques we propose for improving cultural experience. The first component of our method is the egocentric artwork recognition, which is used to provide knowledge to the user without the need to explicitly input the painting identifier for which the information is desired.

The second one is recognizing the gestures of the user, hence the *self-gesture recognition*. In this regard, adapting to personal requests is a key aspect, in fact people in different cultures have very different ways of express through gestures. Our method can indeed learn from a very limited set of examples and it is robust to lighting changes and ego-motion.

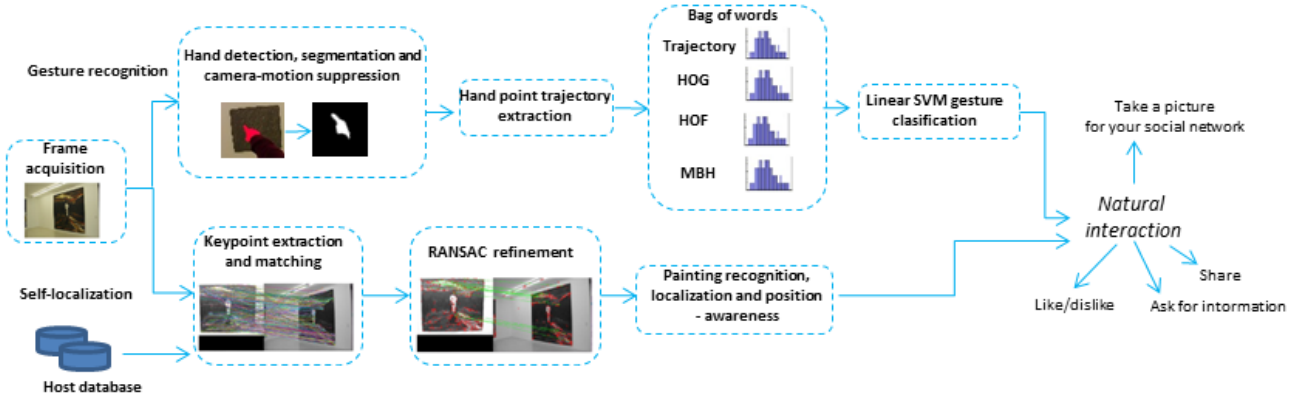


Fig. 3: Schematization of the proposed self-gesture and painting recognition methods.

Figure 3 provides a more detailed schematization of the workflow of the two components of our method, which will be deeper examined in the following of this section.

3.1 Egocentric painting recognition

The ability to understand what painting is the user looking at without the need of directly asking it to him or interact with some early deployed tag is one of the key features of our approach. In fact, this allows for a wide range of applications from user localization to social tagging and sharing. Aiming at the recognition of an artwork in a museum introduces a useful constraint: the number of paintings is finite and know in advance. Furthermore, a database containing painting data and features like descriptors and keypoints can be available, greatly reducing the required processing time. While object-recognition is a field in which some very interesting results are available, the real-world ego-vision setting we deal with makes this task full of challenges. For example, ego-vision presents extremely camera motion and different lighting conditions due to peculiar painting needs. Furthermore, paintings in museums are often protected by reflective glasses or occluded by other visitors, requiring in the necessity of a method capable of dealing with this difficulties too.

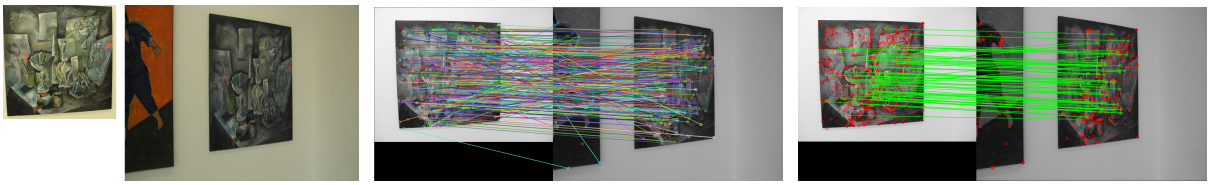
An important step to any recognition attempt in ego-vision is to understand whether the task should be performed or not. In fact, a typical ego-vision characteristic is that the camera wearer can have very fast head motion, e.g. when he is looking around for something. The high head motion can cause a significant blur in the video sequence resulting in an extremely low quality video. If not addressed properly, this situation can degrade the descriptor extraction on which our recognition recognition process is based at the point that its results can loose significance.

To deal with this challenge typical of the ego-vision scenario, previous to the recognition process we evaluate the amount of blurriness and decide whether to proceed with the frame or to skip it. The idea behind our approach is to compute the amount of gradient in the frame and to learn a threshold that discriminates a fast head movement due to the user looking around from the normal blur caused by small and inevitable motions. We define a blur function which recognizes the blur degree in a frame I, according to a threshold θ_B:

$$Blur(I, \theta_B) = \sum_I \sqrt{\nabla S_x^2(I) + \nabla S_y^2(I)}, \quad (1)$$

where $\nabla S_x(I)$ and $\nabla S_y(I)$ are the x and y components of Sobel's gradient in the image and θ_B is the threshold under which the frame is discarded due to excessive motion blurriness, a parameter which can be learned by computing the average amount of gradient in a sequence. This step, that can be done in real-time, effectively allows to remove those frames that could prevent our painting recognition method to work properly.

Figure 4a shows an example of a room view captured by a first person perspective: the painting on the wall is heavily distorted by perspective preventing template matching techniques. In order to find a match between the framed artwork and its counterpart in the museum database, 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 fist-person perspective. Is it a painting or is that the window? Detection based on shape resulted in a false positive rate that prevented any further evaluation, hence we rely on sampling over the



(a) The painting template stored in the museum database (left) and the scene captured from the first person camera view (right).

(b) SIFT matching between the two artworks.

(c) Remaining keypoints after processing the matches with the RANSAC algorithm.

Fig. 4: Painting recognition in our approach.

whole image. Figure 4b shows the SIFT matching between the current frame and the template painting from the database. In order to improve the match quality, we process the matched keypoints using the RANSAC algorithm (Figure 4c). To further improve the matching results, a first thresholding step is performed: using a threshold over the distances between SIFT descriptors θ_S, we remove the matches which have a distance greater than θ_S. This allows to tune our method to different situations and primarily influences the recognition performance.

In order to understand whether the current frame really contains a painting or the computed matches refer to different elements (e.g. windows or architectonic details) a second thresholding step is applied. Using the threshold θ_d over the ratio between matches that survived the previous pruning steps (RANSAC and θ_S) and the original amount of keypoints in the current frame, it can be decided whether the current frame contains an artwork or not. Adjusting this threshold can render the method more robust to noise and clutter situations or increase its detection range. A detailed analysis of the impact of these thresholds on the proposed method is presented in experimental section. A summarization of the painting recognition method is presented in Algorithm 1.

Algorithm 1: Painting recognition

input : Current frame, template database
output: Painting identifier

Compute current frame keypoints and local descriptors;;
for *each painting template* **do**
  read SIFT descriptors;
  calculate matching keypoints;
  apply RANSAC algorithm to discard outliers;
  remove matches with distance greater than θ_S;
  compute ratio between remaining matches and total keypoints of the current frame;
Extract the painting with the highest $\frac{matches}{keypoints}$ ratio;
if $\frac{matches_{max}}{keypoints_{max}} > \theta_d$ **then**
  return Recognized painting identifier;
else
  return No painting detected

3.2 Gesture Recognition

Gesture recognition systems should recognize both static and dynamic hand movements. Therefore, we propose to describe each gesture as a collection of dense trajectories extracted around hand regions. Feature points are sampled inside and around the user's hands and tracked during the gesture; then several descriptors are computed inside a spatio-temporal volume aligned with each trajectory, in order to capture its shape, appearance and movement at each frame. These descriptors are coded, using the Bag of Words approach and power normalization, in order to obtain the final feature vectors which are then classified using a linear SVM classifier.

To describe information of shape, appearance and movement of the hand trajectory we rely on the following descriptors according to [34]: Trajectory descriptor, histograms of oriented gradients (HOG), of optical flow, and motion boundary histograms. The first one directly captures trajectory shape, while HOG [7] are a spacial descriptor representing the orientation of image grandients and thus encode the static appearance of the region surronding the trajectory. HOF and MBH [8] are based on optical flow and are used to capture motion information enforcing the temporal aspect of our method.

3.2.1 Camera motion removal

In order to estimate 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 [12]

algorithm on densely sampled features points. SURF [3] features and sample motion vector are extracted from Farneback's optical flow [9] to get dense matches between frames.

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 hand, which would have resulted in incorrect trajectories. Computing the homography based only of non-hand keypoints allow to have a motion model consistent with the ego-motion of the camera which can, consequently, be removed.

3.2.2 Gesture Description

After the suppression of camera-motion, hand trajectories can be extracted. Using the previously estimated homography, the second frame is warped and the optical flow between the two frames is then recomputed in order to estimate the motion resulting from the hand movement. Feature points around the hand region are sampled and tracked in a way similar to what [34] does for human action recognition. 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. Each resulting 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 $\omega = (u_t, v_t)$:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (M * \omega)|_{(\bar{x_t}, \bar{y_t})}, \qquad (2)$$

where $(\bar{x_t}, \bar{y_t})$ is the rounded position of P_t.

Differently to [34], 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 it are discarded.

A spatio-temporal volume aligned with each trajectory is then considered, 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 trajectory descriptors for each gesture. In order to obtain a fixed size descriptor, we exploit the Bag of Words approach training four separate codebooks, one for each descriptor. Each codebook contains 500 visual words and is obtained running the k-means algorithm in the feature space.

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:

$$f(h_i) = \text{sign}(h_i) \cdot |h_i|^{\frac{1}{2}} \qquad (3)$$

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. Eventually, gestures are recognized using a linear SVM 1-vs-1 classifier.

3.2.3 Hand Segmentation

A key step in our gesture recognition process is the hand segmentation: this allows to distinguis between camera and hand motions, effectively pruning the trajectories that do not regard the user hand. Our method disregards all the semantic noise resulting from other motions in the scene, obtaining a descriptor that captures hand movements and shape as if the video sequence were captured by a fixed camera.

In order to compute the aforementioned segmentation mask, we extract superpixels at each frame using the SLIC algorithm [2]. It performs a k-means based local clustering of pixel in a space generated by $(labxy)$ where (lab) are the coordinates of the LAB color space and (xy) are the spatial coordinates on the image. Superpixels are represented with several features: histograms in the HSV and LAB color spaces (that have been proved to be good features for skin representation [17]), Gabor filters and a simple histogram of gradients, to discriminate between objects with a similar color distribution.

In order to achieve robustness to varying illumination conditions, instead of using a single classifier we train a collection of Random Forest classifiers indexed by a global HSV histogram. This results in distributing the training images among the classifiers using the k-means clustering on the feature space. At test time, the predictions from the five nearest classifier are averaged to make the final prediction.

Furthermore, semantic coherence over space and time is taken into account. Since past frames should affect the prediction for the current frame, a smoothing filter is applied: the prediction for each frame is replaced with a combination of the classifier results from past frames.

Then, in order to remove small and isolated pixel groups and also to aggregate bigger connected pixel groups, the GrabCut algorithm is applied to exploit spatial consistency.

4 Experimental Results

To investigate the performance of our method we record and publicly release two datasets: Interactive Museum and Maramotti Collection. The Interactive Museum is a dataset taken from the ego-centric perspective in a virtual environment where users can interact with digital artworks using gestures[1]. The Maramotti Collection dataset is a completely unconstrained real-world museum dataset recorded at the Maramotti Museum of contemporary art[2]. This dataset features different lighting conditions due to different museum rooms having different illumination requirements and partial artwork occlusions due to the presence of the user hands performing gestures. This dataset challenges both our painting and gesture recognition algorithms on a real-world scenario.

To further compare the performance of the proposed gesture recognition algorithm with existing approaches, we test it on the Cambridge-Gesture database [19], which includes nine hand gesture types performed on a table, under different illumination conditions. To evaluate the hand segmentation approach, we test it on the publicly available CMU EDSH dataset [23] which consists of three ego-centric videos with indoor and outdoor scenes and large variations of illuminations.

4.1 Painting Recognition

We test our painting recognition method on a subset of the real world Maramotti Collection museum dataset, which subset contains more than 13000 frames at 960×540 resolution annotated with the current visible painting. This scenario challenges our method in several ways: being a real museum lighting conditions can vary greatly from one room to another due to different artworks needs. Furthermore, being the same dataset used for gesture recognition, artworks can be partially occluded by the user hand during the recognition process. We evaluate our results in terms of detection precision and recall and classification accuracy: this allows to loosely decouple the detection performance from the classification phase, effectively reflecting the fact that our method is based on two different steps. Figure 5 shows the results of our experiments under varying detection and distance thresholds. It can be seen how deep can be the impact of different threshold values on the method performance: using a detection threshold $\theta_d = 0$ immediately produces a 100% recall due to, de facto, not performing any detection but accepting everything as a painting. This produces good classification accuracy performance if the distance threshold is small, but quickly degrades increasing the number of keypoints that are considered good matches (Fig. 5a). On the other hand, a detection threshold too high ($\theta_d = 0.15$) effectively produces a high amount of false negatives resulting in a high precision but very low recall and subsequently low classification accuracy due to the lack of detections (Fig. 5d). In this scenario, increasing the distance threshold θ_S increases the number of available keypoints to use in the detection increasing both accuracy and recall, with an accuracy upper bound that is significantly lower that the recall one. With these results in mind, our method fixes $\theta_S = 150$ and $\theta_d = 0.025$.

4.2 Gesture Recognition

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 to 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. 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 [19], 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 1 shows the recognition rates obtained with our gesture recognition approach, compared with the ones of tensor canonical correlation analysis (TCCA) [18], product manifolds (PM) [25], tangent bundles (TB) [24] and spatio-temporal covariance descriptors (Cov3D) [30]. Results show that proposed method is able to overcome the existing state-of-the-art approaches.

We then present experiments on the Interactive Museum dataset: it consists of 700 video sequences, all shot with a wearable camera, in an interactive exhibition room, in which paintings and artworks are projected over a wall, in a virtual museum fashion. The camera is

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

[2] `http://imagelab.ing.unimore.it/files/ego_maramotti.zip`

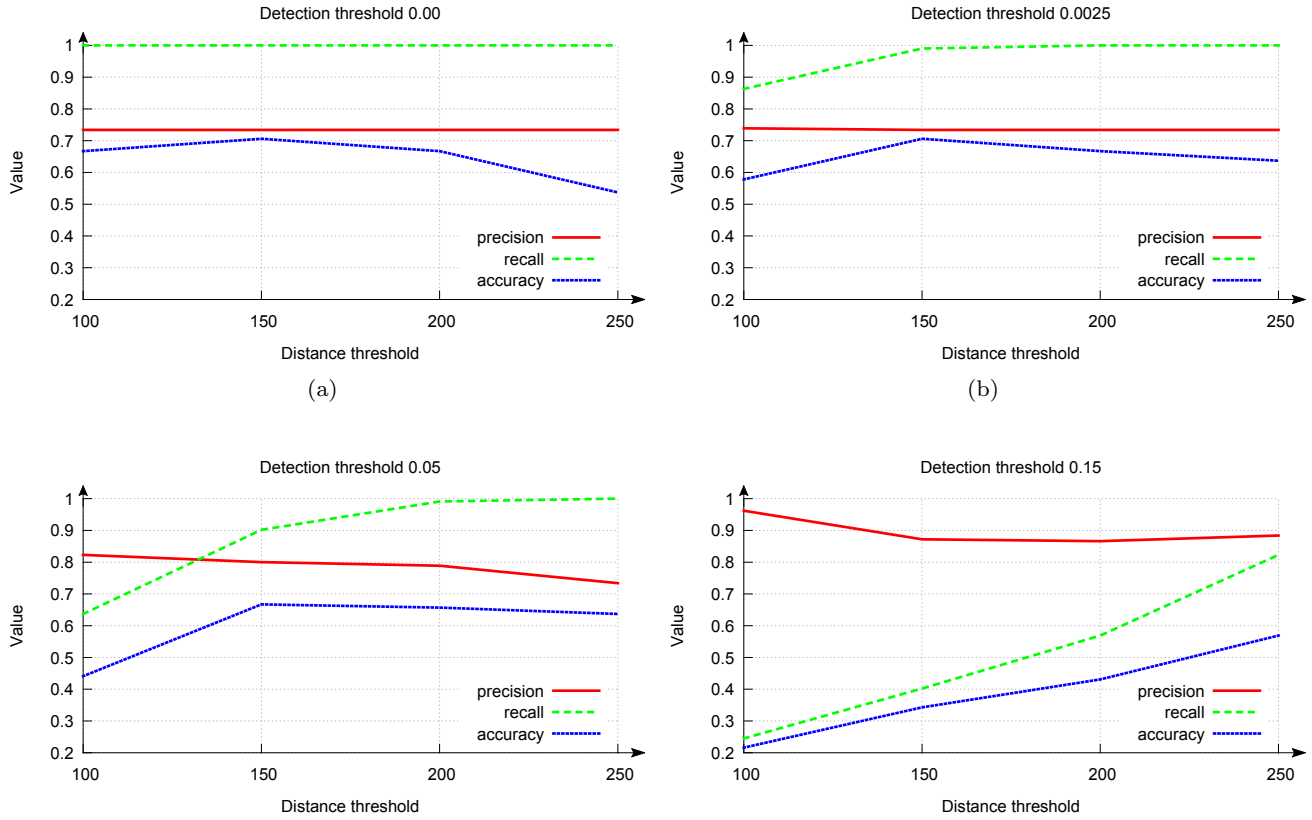


Fig. 5: Results of our painting recognition evaluation in terms of precision and recall of the detection step and accuracy of the classification phase.

Method	Set1	Set2	Set3	Set4	Overall
TCCA [18]	0.81	0.81	0.78	0.86	0.82
PM [25]	0.89	0.86	0.89	0.87	0.88
TB [24]	**0.93**	0.88	0.90	0.91	0.91
Cov3D [30]	0.92	**0.94**	0.94	0.93	0.93
Our method	0.92	0.93	**0.97**	**0.95**	**0.94**

Table 1: Recognition rates on the Cambridge dataset.

placed on the user's head and captures a 800 × 450, 25 frames per second 24-bit RGB image sequence. In this setting, 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. This dataset is very challenging since there is fast camera motion and users have not been trained before recording their gestures, so that each user performs the gestures in a slightly different way, as would happen in a realistic context.

Furthermore, we show an evaluation on our second more challenging dataset recorded in the Maramotti Collection contemporary art museum. It features more than 27000 frames annotated with gesture labels of the 7 gestures previously described. This dataset presents a more challenging environment due to its setting a real-world museum, featuring the inclusion of random visitors in the scene or different lighting conditions.

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. The reported results are the average over 100 independent runs.

In Table 2 we show the gesture recognition accuracy for each of the five subjects, and we also compare with the ones obtained without the use of the hand segmentation mask for camera motion removal and trajectories pruning. Results show that our approach is well suited to recognize hand gestures in the ego-centric do-

User	Virtual Room
Subject 1	0.95
Subject 2	0.87
Subject 3	0.95
Subject 4	0.94
Subject 5	0.96
Average	0.93

Table 2: Gesture recognition accuracy on the Interactive Museum dataset with and without hand segmentation.

User	Maramotti
Subject 1	0.94
Subject 4	0.52
Subject 5	0.68
Subject 6	0.56
Subject 7	0.54
Average	0.53

Table 3: Gesture recognition accuracy on the Maramotti Collection museum.

main, even using only two positive samples per gesture, and that the use of the segmentation mask can improve recognition accuracy. Furthermore, table 3 shows the accuracy results of our gesture recognition method applied to the Maramotti Collection museum, a real-world scenario where the resulting accuracy reflects all the challenges of a testing a method on an unconstrained environment.

4.3 Hand Segmentation

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 4 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, 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 the 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.

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

Table 4: Performance comparison considering Illumination Invariance (II), Temporal Smoothing (TS) and Spatial Consistency (SC).

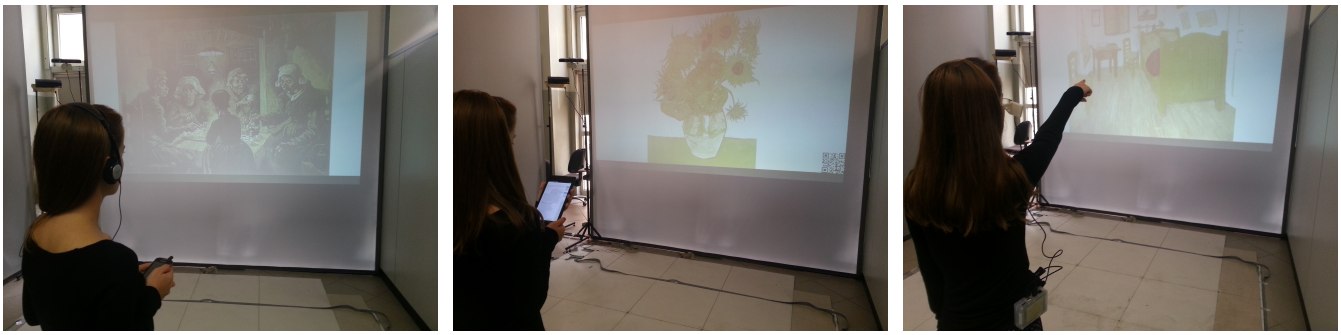
Method	EDSH2	EDSHK
Hayman and Eklundh [14]	0.211	0.213
Jones and Rehg [15]	0.708	0.787
Li and Kitani [23]	0.835	0.840
Our method	**0.852**	**0.901**

Table 5: Hand segmentation comparison with the state-of-the-art.

Then, in Table 5 we compare our segmentation method with different techniques: a video stabilization approach based on background modeling [14], a single-pixel color method inspired by [15] and the approach proposed in [23] 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.

4.4 User experience evaluation

The system we provide allows for a novel kind of interaction with the artwork and hence requires some sort of user experience evaluation. Exploiting the setting of our virtual museum, we compare our method to two of the most common interfaces to museum knowledge: audio guides and QR-CODE based guides. The first one requires the user to input a number displayed next to the artwork and provides an audio description of the painting (Fig. 6a). Similarly, the second one requires



(a) (b) (c)

Fig. 6: Examples of evaluated interfaces in our virtual-museum environment.

Interface	Score
Audio-Guide	0.389
QR-CODE	0.401
Our Method	0.422

Table 6: Results of the user experience evaluation of different museum interaction interfaces.

the user to frame a QR-CODE sited close to the artwork with a tablet. A written description of the artwork is then displayed on the device's screen (Fig. 6b). To be compared with the aforementioned tools, we made our method recognize the current painting and provide an audio description of it at the detection of the "point" gesture (Fig. 6c). Using a Likert scale, we asked a set of 20 test subjects to answer the question "How natural did the interaction feel?" in our virtual museum setting with a score between 1 and 5, where 1 was "Unnatural" and 5 "Extremely natural". Most of the test subjects agree that our solution improves the fruition since it does not require the user to divert his attention from the painting, as showed by Table 6.

5 Conclusion

We described a novel approach to cultural heritage fruition. With the devices and the algorithms we proposed we overcame some of the limitations of self-service museum guides. We proposed a new technique of hand gesture recognition in ego-centric videos: our model can deal in real-time with static and dynamic gestures and can achieve high accuracy results even when trained with a few positive samples, which allows for an easy personalization to the different ways each user performs the same gestures. We also showed how our gesture recognition and hand segmentation results outperform the state-of-the-art approaches on Cambridge Hand Gesture and CMU EDSH datasets. In addition, we proposed a technique for automatic painting recognition that does not require any user interaction nor predetermined museum hardware such RF-IDs. While we use this recognition technique to provide the user content based on what he is really looking at, we recognize how the painting recognizion can be a key step for many further applications. For example, automatic painting recognition can be used to locate the user inside the museum providing services like customized visiting paths. We evaluated our new museum interface providing evidence that test subjects prefer to use an ego-vision gesture-based approach instead of the more common audio guides or QR-CODE based applications, showing great promise for future works.

Acknowledgements This work was partially supported by the PON R&C project DICET-INMOTO (Cod. PON04a2_D) and the CRMO project "Vision for Augmented Experiences". 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. tech rep. http://tourism-intelligence.com/
2. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(11), 2274–2282 (2012)
3. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Proc. of ECCV. Springer (2006)
4. Blöckner, M., Danti, S., Forrai, J., Broll, G., De Luca, A.: Please touch the exhibits!: Using nfc-based interaction for exploring a museum. In: Proc. of MobileHCI, pp. 71:1–71:2 (2009)
5. Coenen, T., Mostmans, L., Naessens, K.: Museus: Case study of a pervasive cultural heritage serious game. J. Comput. Cult. Herit. **6**(2), 8:1–8:19 (2013)
6. Cucchiara, R., Bimbo, A.D.: Visions for augmented cultural heritage experience. IEEE Multimedia **21**(1), 74–82 (2014)

7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proc. of CVPR, vol. 1, pp. 886–893. IEEE (2005)
8. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Proc. of ECCV. Springer (2006)
9. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Image Analysis, pp. 363–370. Springer (2003)
10. Fathi, A., Rehg, J.M.: Modeling actions through state changes. In: Proc. of CVPR (2013)
11. Fathi, A., Ren, X., Rehg, J.M.: Learning to recognize objects in egocentric activities. In: Proc. of CVPR (2011)
12. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
13. Guidi, G., Frischer, B., Russo, M., Spinetti, A., Carosso, L., Micoli, L.: Three-dimensional acquisition of large and detailed cultural heritage objects. Machine Vision and Applications **17**(6), 349–360 (2006)
14. Hayman, E., Eklundh, J.O.: Statistical background subtraction for a mobile observer. In: Proc. of ICCV (2003)
15. Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection (1999)
16. Khan, F., Beigpour, S., van de Weijer, J., Felsberg, M.: Painting-91: a large scale database for computational painting categorization. Machine Vision and Applications **25**(6) (2014)
17. Khan, R., Hanbury, A., Stoettinger, J.: Skin detection: A random forest approach. In: Proc. of ICIP (2010)
18. Kim, T.K., Cipolla, R.: Canonical correlation analysis of video volume tensors for action categorization and detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31**(8), 1415–1428 (2009)
19. Kim, T.K., Wong, K.Y.K., Cipolla, R.: Tensor canonical correlation analysis for action classification. In: Proc. of CVPR (2007)
20. Kortbek, K.J., Grønbæk, K.: Interactive spatial multimedia for communication of art in the physical museum space. In: Proc. of ACM Multimedia (2008)
21. Kuflik, T., Stock, O., Zancanaro, M., Gorfinkel, A., Jbara, S., Kats, S., Sheidin, J., Kashtan, N.: A visitor's guide in an active museum: Presentations, communications, and reflection. Journal on Computing and Cultural Heritage (JOCCH) **3**(3), 11 (2011)
22. Kuusik, A., Roche, S., Weis, F., et al.: Smartmuseum: Cultural content recommendation system for mobile users. In: Proc. of ICCIT (2009)
23. Li, C., Kitani, K.M.: Pixel-level hand detection in egocentric videos. In: Proc. of CVPR (2013)
24. Lui, Y.M., Beveridge, J.R.: Tangent bundle for human action recognition. In: In proc. of Automatic Face and Gesture Recognition and Workshops (2011)
25. Lui, Y.M., Beveridge, J.R., Kirby, M.: Action classification on product manifolds. In: Proc. of CVPR (2010)
26. Oomen, J., Aroyo, L., Marchand-Maillet, S., Douglass, J.: Personalized access to cultural heritage: Multimedia by the crowd, for the crowd. In: Proc. of ACM Multimedia (2012)
27. Pechenizkiy, M., Calders, T.: A framework for guiding the museum tours personalization. In: Proc. of PATCH'07 (2007)
28. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Proc. of ECCV (2010)
29. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: Proc. of CVPR (2012)
30. Sanin, A., Sanderson, C., Harandi, M.T., Lovell, B.C.: Spatio-temporal covariance descriptors for action and gesture recognition. In: Proc. of Workshop on Applications of Computer Vision (2013)
31. Sparacino, F.: The museum wearable: Real-time sensor-driven understanding of visitors' interests for personalized visually-augmented museum experiences. (2002)
32. Suh, Y., Shin, C., Woo, W., Dow, S., Macintyre, B.: Enhancing and evaluating users' social experience with a mobile phone guide applied to cultural heritage. Personal Ubiquitous Compututing **15**(6), 649–665 (2011)
33. Sundaram, S., Cuevas, W.W.M.: High level activity recognition using low resolution wearable vision. In: Proc. of CVPR (2009)
34. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action Recognition by Dense Trajectories. In: Proc. of CVPR (2011)
35. Yanulevskaya, V., Uijlings, J., Bruni, E., Sartori, A., Zamboni, E., Bacci, F., Melcher, D., Sebe, N.: In the eye of the beholder: Employing statistical analysis and eye tracking for analyzing abstract paintings. In: Proc. of ACM Multimedia, MM '12 (2012)
36. Zabulis, X., Grammenos, D., Sarmis, T., Tzevanidis, K., Padeleris, P., Koutlemanis, P., Argyros, A.: Multicamera human detection and tracking supporting natural interaction with large-scale displays. Machine Vision and Applications **24**(2) (2013)