Wearable Vision for Retrieving Architectural Details in Augmented Tourist Experiences

Stefano Alletto, Davide Abati, Giuseppe Serra and Rita Cucchiara
Università degli Studi di Modena e Reggio Emilia
Via Vignolese 905, 41125 Modena - Italy
Email:{name.surname}@unimore.it

***Abstract*—The interest in cultural cities is in constant growth, and so is the demand for new multimedia tools and applications that enrich their fruition. In this paper we propose an egocentric vision system to enhance tourists' cultural heritage experience. Exploiting a wearable board and a glass-mounted camera, the visitor can retrieve architectural details of the historical building he is observing and receive related multimedia contents. To obtain an effective retrieval procedure we propose a visual descriptor based on the covariance of local features. Differently than the common Bag of Words approaches our feature vector does not rely on a generated visual vocabulary, removing the dependence from a specific dataset and obtaining a reduction of the computational cost. 3D modeling is used to achieve a precise visitor's localization that allows browsing visible relevant details that the user may otherwise miss. Experimental results conducted on a publicly available cultural heritage dataset show that the proposed feature descriptor outperforms Bag of Words techniques.**

I. Introduction

Cultural cities and museums are an increasingly common destination for tourists: half of the Americans traveling abroad visits historical places and almost one third of them chooses cultural heritage sites [1]. From the popularity of this kind of tourism, new challenges and possibilities arise. Indeed, many tourists of the 21th century are digital natives and expect mobile and multimedia to follow them during their visits. While the problem of providing immersive and interactive contents have been addressed in the past, e.g. exploiting touch interfaces [5] or creating interactive environments [10], much can still be done. In fact, these systems can only provide an enriched interaction in the environment they are designed for and lack the ability to adapt to changes in the cultural heritage site setting or to follow the user in its cultural visit. For instance, problems arise in unconstrained environments such as outdoor visits in large areas where people move freely around masterpieces, buildings etc.

Following this growing request for multimedia tools and applications for smart tourism, we aim at designing a system that is capable of assisting the visitor in an unconstrained outdoor tour. To be effective, such system requires the ability to see what the user sees in a perspective that resembles his very own. Recently, due to the increased diffusion of wearable devices and head mounted cameras, systems dealing with an egocentric perspective are arousing a growing interest in the research community. Egocentric vision, often referred to as ego-vision, tackles with problems such as activity and gesture recognition [4], social interactions [6] or video summarization [12] exploiting the unique perspective of wearable cameras.

In fact, being tied to its user, the camera follows his path effectively providing a recording of the objects he interacts with, people and events he focuses his attention on and, in short, events and things that are relevant to him. This feature is of extreme interest when designing a system that aims at following a tourist during his visit. Being capable of seeing in real-time what the users sees allows the method to provide him with useful information that is directly relates to his focus of attention, effectively guiding the visit in a natural and intuitive way.

In this paper we propose a system that can retrieve architectural details from images and can provide the tourist with an augmented experience. The main idea behind our framework is that a tourist may not be able to immediately identify all the details in an artwork and may have to rely on a guide to do so. Using a handbook is neither practical not desirable since it often requires time to find the correct building and can distract from the actual artwork. Using a wearable computing board and a glass-mounted camera, the user can ask to the system to provide him with the details of the scene he is looking at. This requires almost no effort from the tourist since the system is already seeing what he is. Using the visitor's smartphone as a screen, the method can then display a view of the captured artwork where the noteworthy details have been highlighted.

Here, we make several contributions: we design a system capable of following a tourist in his visit and provide him with useful information minimizing the amount of input that the user must provide, resulting in a natural and more enjoyable interaction. Furthermore, we propose a retrieval method that relies on a fast descriptor based on the covariance of local features, which is particularly suited for the task at hand and outperforms the descriptors that exploit the Bag of Word paradigm (BOW). In contrast to common BOW techniques [17] our descriptor does not require a pretrained visual vocabulary, resulting in decreased computational costs and independence from a specific context and dataset. We also provide a new and publicly available cultural heritage dataset that features a large set of images of the romanesque dome of Modena, annotated with 10 different possible retrieval queries. Finally, we perform an extensive evaluation in which we compare our descriptor to several competitors showing improved performances.

II. Related Work

In the last years several methods and approaches regarding building and visual landmark recognition have been proposed

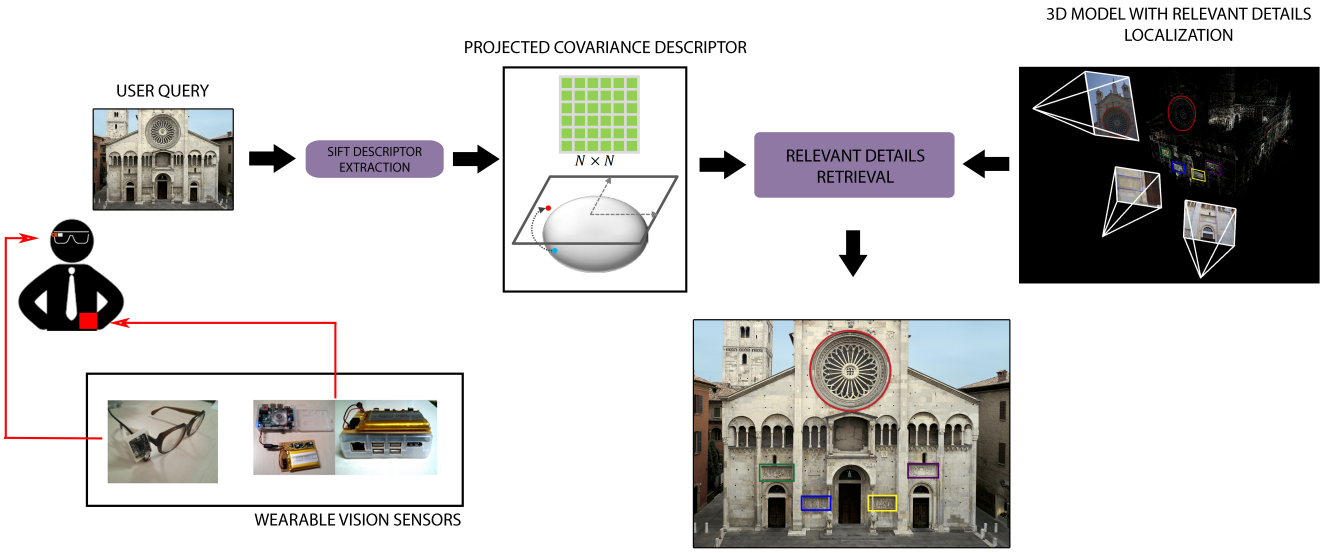


Fig. 1: Schematization of the proposed system.

[13]. They deal with the problem of identifying different buildings in a large-scale dataset and recognizing them in query images, in order to provide existing information about what the user is interested in. These techniques can be roughly divided in methods that are focused on improving recognition performance or reducing computational effort.
In particular, the method described in [7] relies on wide baseline matching (i.e. image matching under significant view-point change), extracting affine invariant vertical segments and describing them with geometrical and color information. Then, the matching score is calculated by Mahalanobis distance between such descriptors and distance between segments. Trinh *et al.* [19] address the problem of recognizing multiple buildings in a dataset by extracting facets of each building exploiting line segments and vanishing point information, and then describing them with color histogram and a list of SIFT features. Nearest-neighbor is then used for matching a query image to its closest model. The problem has also been faced as a ranking one. Relying on the vector space model as image representation, Philbin *et al.* [16] described a system able to propose a ranked list of images in response to a query depicting a building. They also remark how important the spatial verification of such words is, in order to re-rank the top scored retrieved images filtering out false positives. However, these techniques aim at recognizing building landmarks and provide to the user general information of what he/she is looking at, but miss interesting architectural details such as statues, decorated windows or arches etc.

Recently, in [23] the authors analyzed a full pipeline for visual landmark and architectural detail recognition from Internet photo collections, in order to understand if some of the buildings are easier to recognize with respect to other ones, what kind of representation allows the images to be efficiently stored in memory, how reliable user provided semantic information are, and so on. In their analysis, they cluster images into categories to understand how the visual recognition behaves for each of those types. One important remark is the trade off they highlight between the difficulty of the recognition and the reliability of user provided semantic annotation. As they state, landmarks easier to classify due to high cardinality of the cluster (such as building facades) often suffer from noisy semantic information, while well annotated images (such as architectural details) usually belong to less populated categories, and hence are more difficult to recognize.

A few works addressed the problem of directly identifying or retrieving details, instead of relying on semantic annotation. Weyand and Liebe [22] address the problem of discovering popular details given a large collection of images of a specific building. They presented an offline hierarchical technique capable of finding iconic images at various scales, iteratively solving a medoid shift increasing the kernel bandwidth. The technique presented by Mikulik *et al.* [15] proposed a retrieval system based on Bag of Word that returns to the user the images of the details depicted in the query. This is achieved thanks to a distance metric that scores the match between visual words according to the corresponding keypoint scale change, preferring scaling-up rather than scaling down (i.e. zoom-in rather than zoom-out).

III. The Proposed Architecture

Our system consists in a collection of wearable egocentric vision devices and a processing center. The wearable devices embed a glass-mounted camera, an Odroid-XU developer board serving as image processing, GPS module and network communication unit. Our wearable solution is composed by several commercial components in order to have: 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 Odroid-XU developer board [2] embeds the ARM Exynos 5 SoC, that hosts a Quad Big-Little ARM processor (Cortex A15 and A7) [3], and is powered by a battery pack of 3000 mAh to make it portable.

The processing center stores the users' current location in order to collect data for statistical analysis and provides cultural information. In particular, the processing center contains 3D models of cultural heritage buildings generated from a set of uncontrolled images (see below, Section V). A 3D building model consists in a 3D point cloud, where architectural details are annotated, its geographical location and an image collection $C = \{I_1, \dots I_n\}$ used in the reconstruction. For each image I, we store the detected 2D-3D correspondences between 3D and local interest points in the point cloud.

When the user captures an image, the image processing algorithm, that is able to detect the architectural details of the building the user is observing, runs on the wearable board (see Figure 1). In particular, the image is analyzed by extracting local SIFT features and computing, as image global descriptor, the Projected Covariance Descriptor (*pCov*, see Section IV). This descriptor is used to retrieve a ranked list of similar images from the building image collections connecting with the processing center; image collections linked to cultural heritage building far from the user's geographical location are discarded. The search proceeds by calculating the similarity between the query vector and each image feature descriptor in the candidate collections using Euclidean distance.

To select the most similar images in the corpus, we include a spatial verification step on the K top ranked images (in our experiments we fix K=5). The spatial verification estimates a geometrical transformation between the query image and each K top candidates and scores them based on how well local features locations are reprojected by the estimated transformation. Following [16], we use RANSAC algorithm that generates transformation hypotheses using a minimal number of correspondences and then evaluates them based on the number of inliers. We consider the affine transformation as model for generating hypotheses that can cover situations such as zooming or observation of a cultural site from different view points, since it can encode rotation and translation warps. The affine transformation has six degrees of freedoms and can be computed from three point correspondences. Once similar images are obtained we can exploit 2D-3D correspondences to determine the absolute camera pose by solving the perspective-n-point (PnP) problem [8], [9]. Differently from the standard solution that uses three point correspondences (assuming that the intrinsic parameters of the camera are known), we propose to use the recent approach proposed by Kukelova *et al.* [11] that solves the pose problem from cameras with unknown radial distortion and unknown focal length. This technique allows us to obtain a good accuracy even if we are using commercial wearable cameras with large radial distortions.

Currently a Android application allows the user to see his captured photo with interesting architectural details highlighted. As future work, we want to extend this application also to run on a head-mounted display that will enable the visitor to obtain a more natural browsing of the contents.

IV. Projected Covariance Descriptor

Let $F = \{\mathbf{f_1} \dots \mathbf{f_N}\}$ be a set of local SIFT features extracted on an image I, we represent them by a covariance matrix $\mathbf{C}$, that encodes information about the variance of the features and their correlations:

$$\mathbf{C} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{f}_i - \mathbf{m})(\mathbf{f}_i - \mathbf{m})^T, \tag{1}$$

where $\mathbf{m}$ is the mean vector of the set F. Although the space of covariance matrices can be formulated as a differentiable manifold, it does not lie in a vector space (e.g the covariance space is not closed under multiplication with a negative scalar) and Euclidean distance between image descriptors can not be computed. Therefore to use this descriptive feature vector, we need to define a suitable transformation. We exploit a projection from the Riemannian manifold to an Euclidean tangent space, called Log-Euclidean metric as suggested by [20]. The basic idea of the Log-Euclidean metric is to construct an equivalent relationship between the Riemannian manifold and the vector space of the symmetric matrix.

The first step is the projection of the covariance matrix on an Euclidean space tangent to the Riemannian manifold, on a specific tangency matrix $\mathbf{T}$. The second one is the extraction of the orthonormal coordinates of the projected vector. In the following, matrices (points in the Riemannian manifold) will be denoted by bold uppercase letters, while vectors (points in the Euclidean space) by bold lowercase ones. The projection of $\mathbf{C}$ on the hyperplane tangent to $\mathbf{T}$ becomes:

$$\mathbf{c} = \mathrm{vec}_{\mathbf{I}} \left(\log \left(\mathbf{T}^{-\frac{1}{2}} \mathbf{C} \mathbf{T}^{-\frac{1}{2}} \right) \right), \tag{2}$$

where $\log$ is the matrix logarithm operator and $\mathbf{I}$ is the identity matrix, while the vector operator on the tanget space at identity of a symmetric matrix $\mathbf{Y}$ is defined as:

$$\mathrm{vec}_{\mathbf{I}}(\mathbf{Y}) = \left[y_{1,1} \ \sqrt{2} y_{1,2} \ \sqrt{2} y_{1,3} \dots y_{2,2} \ \sqrt{2} y_{2,3} \dots y_{d,d} \right]. \tag{3}$$

As observed in [14], by computing the sectional curvature of the Riemmanian manifold, the natural generalization of the classical Gaussian curvature for surfaces, it is possible to show that this space is almost flat. This means that the neighborhood relation between the points on the manifold remains unchanged, wherever the projection point $\mathbf{T}$ is located. Therefore, from a computational point of view, the best choice for $\mathbf{T}$ is the identity matrix, which simply translates the mapping into applying the $\mathrm{vec}_{\mathbf{I}}$ operator to the standard matrix logarithm. This also frees us from the problem of optimizing the projection point for the specific data under consideration, leading to a generally applicable descriptor. Since the projected covariance is a symmetric matrix of $d \times d$ values, the image descriptor is a $(d^2 + d)/2$-dimensional feature vector.

V. 3D Model and architectural details localization

In order to generate a 3D model of a cultural heritage building and to locate the relevant architectural details, we

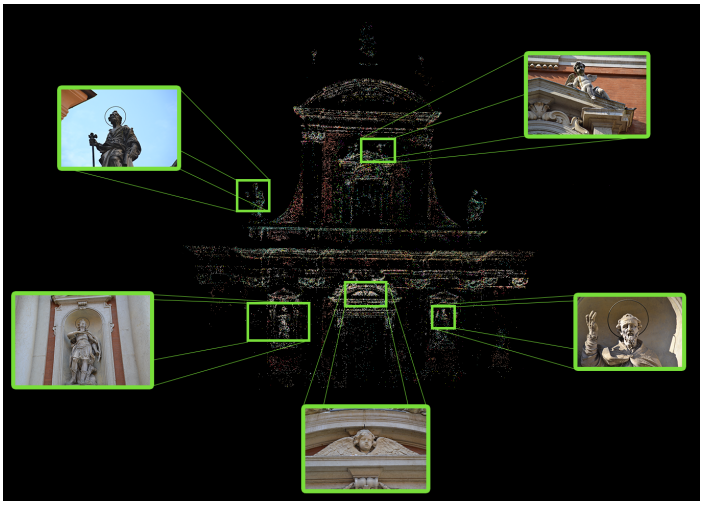


Fig. 2: 3D reconstruction and relevant details of San Giorgio's Church in Modena (Italy).

perform structure from motion (SfM) on a set of images. A SfM system exploits image matching based on local SIFT features to infer information about the scene structure. After finding a set of geometrically consistent matches between each image pair, such matches are organized into tracks. Each track is a set of matching keypoints across multiple images. To recover the set of camera parameters (position, orientation, focal length and radial distortion) and 3D location for each track, a non linear optimization problem is solved, minimizing the reprojection error (the sum of distances between the projections of the track and its corresponding keypoints). This problem can be solved using the Bundle Adjustment technique (BA) [18].
Since non-linear least squares solvers suffer from bad local minima, usually the estimation is evaluated incrementally, starting from a single pair of images and then adding one image at a time. The initial pair is selected as the one with largest number of matches, subject to the condition that such matches cannot be modeled by a single homography. This is done to avoid pairs having a small change in point of view, therefore generating a worst parameter esteem.
The SfM algorithm iteratively adds a single image and solves BA optimization. At each iteration, the image having the largest number of tracks whose 3D locations have already been estimated is added to the evaluation, and its tracks are considered in the optimization.

Bundle adjustment algorithm is a non-linear optimization used to refine the model structure and parameters. As stated, this optimization can be formulated as a non-linear least squares problem, in which the error function is based on the difference between the observed 2D corresponding locations and the projections of the corresponding 3D point on the image plane of the camera. More formally, let p be the parameter vector and $f(p)$ be the residual reprojection errors for a 3D reconstruction, the optimization can be defined as:

$$\hat{p} = \underset{p}{\operatorname{argmin}} \|f(p)\|^2 . \tag{4}$$

A solution to this problem is obtained by using the Levenberg-Marquardt (LM) algorithm that computes a series of regularized linear approximations to the non-linear problem. Let J be the Jacobian of $f(p)$, the LM at each iteration solves the linear least squares problem defined as:

$$(J^T J + \lambda D^T D)\delta = -J^T f, \tag{5}$$

where D is a non-negative diagonal matrix and λ is used as regularization term. Then the update of p is computed by:

$$p \leftarrow p + \delta \qquad if \quad \|f(p+\delta)\| < \|f(x)\| . \tag{6}$$

The Matrix $H_\lambda = J^T J + \lambda D^T D$ is known as the augmented Hessian matrix.

To solve this problem with a large photo collection of a cultural heritage building, we propose to use the multicore bundle adjustment proposed in [24]. This approach shows that inexact step Levenberg-Marquardt can be implemented without storing any Hessian or Jacobian matrices into memory. This allows us to exploit hardware parrallelism and to obtain good balance between speed and accuracy.

Once the sparse 3D model is reconstructed, the most interesting architectural details are manually identified by cultural heritage experts. Figure 2 shows an example of a 3D model and some of the selected details (San Giorgio's Church in Modena - Italy). Since each image of the collection has associated its set of camera parameters and the 3D location where the photo is taken (estimated with the structure from motion step), we can use them to obtain the absolute pose of query images and highlight relevant architectural details.

VI. Experimental Results

To test our system we evaluate its performance on two different problems: comparing the proposed image descriptor based on covariance of local features with a large variety of visual descriptors based on BoW and evaluating the user experience in real scenarios. To evaluate the core functionality of the retrieval algorithm, we acquire and publicly release a new and challenging dataset that revolves around the romanic cathedral of Modena. It features 743 high quality images capturing different views and different architectural details of the cathedral, fully annotated with 10 different possible queries taking into account the whole structure or individual details. The dataset also contains 20 sample query images taken from Google that can be used to reproduce the described results (see examples in Figure 3).

The first component of our method we experimentally evaluate is the retrieval algorithm. To show its superior performance, we compare our descriptor to several recent visual descriptors extracted by the implementation proposed by [21]: Color Moments, generalized color moments up to the second order, giving a 27-dimensional shift-invariant descriptor; RGB Histograms, a combination of three histograms based on the R, G, and B channels; RG Histograms; Hue Histograms; Opponent Histograms; Transformed Color Histograms, RGB histograms obtained by normalizing the pixel value distributions, that achieve scale-invariance and shift-invariance with

Fig. 3: Random samples from the Modena Cathedral dataset.

TABLE I: Comparison between different descriptors employed in our evaluation.

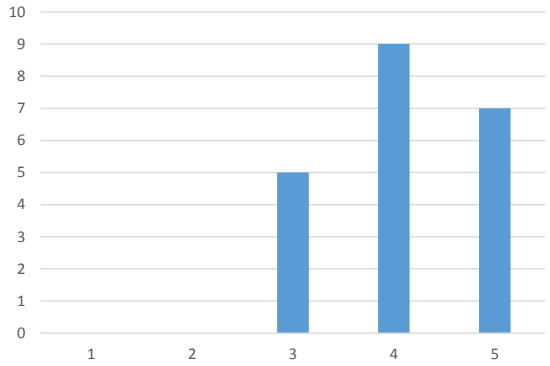
Descriptor	Precision@1	Precision@3	Precision@5	MAP
Our Descriptor pCov-SIFT	**0.800**	**0.717**	**0.600**	**0.362**
RGB SIFT	0.750	0.650	0.570	0.281
Opponent SIFT	0.750	0.600	0.500	0.266
SIFT	0.750	0.617	0.550	0.268
RG SIFT	0.650	0.483	0.420	0.235
C-SIFT	0.600	0.567	0.430	0.236
HSV-SIFT	0.560	0.483	0.440	0.203
Transformed Color Histograms	0.500	0.283	0.300	0.187
Hue SIFT	0.450	0.350	0.260	0.129
Opponent Histograms	0.300	0.233	0.210	0.109
RGB Histograms	0.300	0.217	0.190	0.112
Color Moments	0.200	0.150	0.170	0.115
RG Histograms	0.150	0.133	0.140	0.098
Hue Histograms	0.100	0.100	0.100	0.088

respect to light intensity; SIFT; RGB-SIFT; RG-SIFT; HSV-SIFT; Hue-SIFT; Opponent-SIFT and C-SIFT, obtained using a C-invariant color space which eliminates the remaining intensity information from the opponent channels.

In all cases, to sparsely sample the keypoints in the images we adopt the *Harris-Laplace* keypoint detector. Since all these are local descriptors, a Bag of Words approach is employed to obtain a global description of the image in order to perform the retrieval phase. Using this approach, a 4096-dimensional codebook is built and used to compute the codes of both dataset images and the query ones. The requirement of a codebook is indeed a liability in this context, since is has two major drawbacks: the first one is the time required to compute it, which in the worst case has been of more than 260000 seconds (more than 3 days) on an Intel i7 3.0 GHz CPU. This preprocessing step must be taken for each descriptor type and due to the $O(ndki)$ complexity of the clustering step, which does not depend on the number of images, cannot be shortened without decreasing the precision of the procedure. Our method does not require the computation of a codebook prior to being able to employ the descriptor, effectively saving significant time. On the other hand, being tied to a codebook computed on a particular dataset inevitably leads to being dependent from the training data. Hence, our algorithm is better suited to generalizing than any of the different descriptors it is compared to.

Evaluating our method, we show its performance and we compare it against the aforementioned techniques in terms of Precision@1, Precision@3, Precision@5 and Mean Average Precision. These metrics respectively show the precision of the first k, with $k \in \{1, 3, 5\}$, top ranked images and the overall precision. As the results in Table I show, the usage of the covariance of SIFT descriptors leads to the best results. While other SIFT-based approaches can achieve good performance, it clearly emerges how relying only on color information is not sufficient to perform an effective retrieval. This is mainly due to the complexity of the setting we are dealing with, where most of the architectural details share the same color patterns. Indeed this is the reason why most of the descriptors based on color histograms show very poor performances. Computing the gradient information using the SIFT descriptor on different color spaces can achieve slightly better results than relying only on grayscale information. Nonetheless, using our pCov descriptor, which does not rely on color information, can achieve better results than any of the SIFT-based BoW descriptors, whether they use color or not.

On a second note, since our system is innovative, it requires



(a) "Express how much did you enjoy using the tool"

Fig. 4: The results of the user experience evaluation of the proposed system.

a usability validation step. This evaluation should aim at establishing how the users respond to this new kind of technology, in terms of how they enjoy it, how natural does the interaction feel and how effective the tourists deem the application to be. With these objectives in mind, we staged a preliminary evaluation that involved 5 people of different sex and ages (20-40 years old). Each one of them has been provided with a prototype of our system and has been accompanied in a small tour of Modena, focusing on the cathedral. Here, the tourist had the occasion to test our system. After the visit we asked them to respond to the question "Express how much did you enjoy using the tool" using a Likert scale with scores from 1 (lower) to 5 (higher). The results of this interrogation, that can be seen in Figure 4, validate the proposed system. In fact, 80% of the users evaluated the system as enjoyable (score of 4 or more).

VII. Conclusion

In this paper we presented a system that provides the user with a new way to interact with cultural heritage sites. Benefiting by egocentric paradigm our solution can propose to the user a detailed view of an historical building and allow the visitor to browse through architectural details. The system consists of two main components: the retrieval of similar images and the user's absolute localization. To deal with the retrieval task in this unconstrained scenario, we propose a fast visual descriptor based on the covariance of local SIFT features extracted from an image. Our *pCov* descriptor achieves better performance than its Bag of Words competitors, without the need of a precomputed visual vocabulary. This increases its generality and lowers the computational requirements, making it well suited for wearable and embedded applications. To locate the visitor with regard to the building he is looking at, a 3D model, automatically generated using a structure from motion algorithm, combined with the results of the retrieval phase is employed. This allows the proposal of a set of relevant details visible from the user's current position.

Acknowledgment

This work was partially supported by the Fondazione Cassa di Risparmio di Modena project: "'Vision for Augmented Experience" and the PON R&C project DICET-INMOTO (Cod. PON04a2 D).

References

[1] How the americans will travel 2015. tech rep. http://tourism-intelligence.com/.

[2] Odroid-XU dev board by Hardkernel. http://www.hardkernel.com.

[3] Samsung Exynos5 5410 ARM CPU. http://www.samsung.com/global/business/semiconductor/minisite/Exynos/products5octa_5410.html.

[4] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *Proc. of IEEE Embedded Vision Workshop*, 2014.

[5] Magdalena Blöckner, Svetlana Danti, Jennifer Forrai, Gregor Broll, and Alexander De Luca. Please touch the exhibits!: Using nfc-based interaction for exploring a museum. In *Proc. of MobileHCI*, 2009.

[6] A. Fathi, J.K. Hodgins, and J.M. Rehg. Social interactions: A first-person perspective. In *Proc. of CVPR*, 2012.

[7] Toon Goedemé, Tinne Tuytelaars, and Luc Van Gool. Fast wide baseline matching for visual navigation. In *Proc. of CVPR*, 2004.

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[9] Arnold Irschara, Christopher Zach, Jan michael Frahm, and Horst Bischof. From structure-from-motion point clouds to fast location recognition. In *Proc. of CVPR*, 2009.

[10] Karen Johanne Kortbek and Kaj Grønbæk. Interactive spatial multimedia for communication of art in the physical museum space. In *Proc. of ACM Multimedia*, 2008.

[11] Z. Kukelova, M. Bujnak, and T. Pajdla. Real-time solution to the absolute pose problem with unknown radial distortion and focal length. In *Proc. of ICCV*, 2013.

[12] Yong Jae Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proc. of CVPR*, 2012.

[13] Jing Li, Wei Huang, Ling Shao, and Nigel Allinson. Building recognition in urban environments: A survey of state-of-the-art and future challenges. *Information Sciences*, 277:406–420, 2014.

[14] S. Martelli, D. Tosato, M. Farenzena, M. Cristani, and V. Murino. An FPGA-based Classification Architecture on Riemannian Manifolds. In *Proc. of DEXA Workshops*, 2010.

[15] Andrej Mikulík, Filip Radenović, Ondřej Chum, and Jiří Matas. Efficient image detail mining. In *Proc. of ACCV*, 2014.

[16] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. of CVPR*, 2007.

[17] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proc. of CVPR*, 2003.

[18] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM transactions on graphics (TOG)*, volume 25, pages 835–846. ACM, 2006.

[19] Hoang-Hon Trinh, Dae-Nyeon Kim, and Kang-Hyun Jo. Facet-based multiple building analysis for robot intelligence. *Applied Mathematics and Computation*, 205(2):537–549, 2008.

[20] Oncel Tuzel, Fatih Porikli, and Peter Meer. Pedestrian Detection via Classification on Riemannian Manifolds. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, 2008.

[21] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[22] T. Weyand and B. Leibe. Discovering details and scene structure with hierarchical iconoid shift. In *Proc. of ICCV*, 2013.

[23] Tobias Weyand and Bastian Leibe. Visual landmark recognition from internet photo collections: A large-scale evaluation. *arXiv preprint arXiv:1409.5400*, 2014.

[24] Changchang Wu, S. Agarwal, B. Curless, and S.M. Seitz. Multicore bundle adjustment. In *Proc. of CVPR*, 2011.