VEDI: Vision Exploitation for Data Interpretation

G. M. Farinella[1,2](✉), G. Signorello[2], S. Battiato[1], A. Furnari[1], F. Ragusa[1], R. Leonardi[1], E. Ragusa[3], E. Scuderi[3], A. Lopes[3], L. Santo[3], and M. Samarotto[3]

[1] IPLAB - Department of Mathematics and Computer Science,
University of Catania, Catania, Italy
gfarinella@dmi.unict.it
[2] CUTGANA, University of Catania, Catania, Italy
[3] Xenia Gestione Documentale s.r.l. - Xenia Progetti s.r.l., Acicastello, Italy

Abstract. We present VEDI (Vision Exploitation for Data Interpretation), an integrated system to jointly assist the visitors of cultural sites and provide meaningful statistics about the visits to the managers of the sites. To address both goals, VEDI includes a wearable assistant (implemented through a wearable device such as HoloLens) which leverages Computer Vision algorithms to understand where the user is and what they are paying attention to. At the visitor's end, such information is leveraged to augment the visit by displaying additional information on the observed points of interest, helping the visitors to navigate the site and suggesting what to see next. Concurrently, a back-end extracts high-level behavioral information from the captured video content which is used to provide the site manager with meaningful statistics and performance indexes on the cultural site. Experiments show that VEDI achieves good results on both the indoor and outdoor cultural sites considered for the experimentation.

Keywords: First person vision · Egocentric vision · AI for cultural sites

1 Introduction

Cultural sites are visited everyday by many visitors. This foster the interest developing technologies able to assist the visitor by automatically providing information related to the environment (e.g., the visitor's location in the site) or the observed points of interest (e.g., details on the observed points of interest). Also, for site managers, it is important to understand the behavior of visitors (e.g., inferring what they have seen and where they have been) to measure the performance of the cultural site and improve its services. Most cultural sites currently support their visitors through printed material, audio guides, panels and catalogs, whereas behavioral information is collected from visitors through

E. Ricci et al. (Eds.): ICIAP 2019, LNCS 11752, pp. 753–763, 2019.
https://doi.org/10.1007/978-3-030-30645-8_68

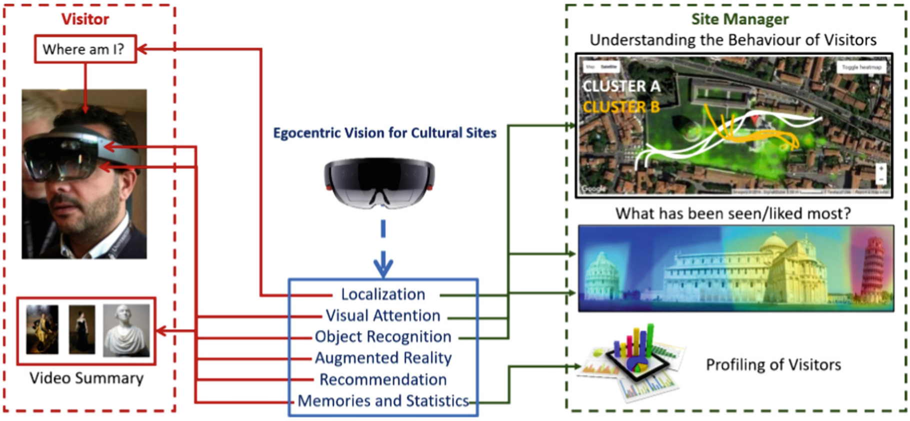

Fig. 1. A scheme of the services provided by VEDI.

surveys. While these classic methodologies are widely employed in cultural sites, they suffer from several limitations. For instance, audio guides and informative panels require the visitor to constantly switch their attention between the cultural site and the supporting media, wheres collecting behavioral information through surveys usually requires time and does not easily scale to large numbers of visitors.

In this paper, we present VEDI (Vision Exploitation for Data Interpretation), an integrated system which includes a wearable device capable of supporting the visitors of cultural sites, as well as a back-end to analyze the visual information collected by the wearable system and infer behavioral information useful for the site manager. To achieve the aforementioned goals, VEDI implements algorithms to localize visitors in the cultural site and recognize the points of interest observed during the visits from the visitors' point of view. The inferred information is then used to provide the following services: (1) a "Where am I?" service which informs the visitor on their location in the site during their visit; (2) a service to provide the visitor with additional information on the observed points of interest using Augmented Reality; (3) a service to estimate the visitors' attention during the visits (e.g., what has been seen most, which places have been most visited). The obtained information can be used by the site manager to profile the visitors and gain insights into the quality of the provided services; (4) a recommendation system to suggest visitors what to see next based on their current location and history of observed points of interest; (5) a system to generate a video summary of each visit, which can be given to the visitor as a "digital memory". Figure 1 shows a scheme of the services offered by VEDI to visitors and cultural site mangers.

The proposed system has been tested in two cultural sites: "Monastero dei Benedettini" and "Orto Botanico". The former is an indoor environment in which we have considered 9 different contexts and 57 points of interest. The latter is an outdoor site composed by 9 different areas, each including plants belonging to different families (we consider 16 plants as points of interest for the experimentation). Experiments show that the proposed VEDI system achieves good performance in the tasks of visitor localization and point of interest recognition on the considered cultural sites.

The remainder of the paper is organized as follows. Section 2 reports the related work. Section 3 presents the collected and publicly available datasets. Section 4 describes the architecture of VEDI and discusses the services provided by the system. Section 5 reports the experimental results, whereas Sect. 6 concludes the paper.

2 Related Work

Augmented Cultural Experience with Wearable and Mobile Devices. Different works investigated the use of wearable and mobile devices for augmented cultural experience [3]. Among the most notable works, the authors of [1] exploited gesture recognition to enable interaction between users and artworks. In [17] a system to support the visitors of natural sites through multimodal navigation of multimedia contents was proposed. In [6] it was suggested to analyze georeferenced images through Visual Analytics tools to identify trends, patterns and relationships among images collected from social media. Differently from the aforementioned works, the proposed VEDI has been designed to both support the visitors of cultural sites and provide useful behavioral information to the site manager.

User Localization form Wearable Devices. Beside classic approaches to scene understanding [2], previous works have investigated method to achieve localization from egocentric images. In [18], a system to perform room-based localization and scene recognition from a wearable camera has been introduced. The authors of [7] proposed a system to infer the 6 Degrees of Freedom pose of a camera directly from egocentric images. The authors of [19] presented an approach to perform world-scale photo gelolocation using Convolutional Neural Networks. The authors of [4,5] proposed an approach to perform room-based localization from few training data. The approach has later been applied in the context of a cultural site in [12]. The authors of [9] exploited egocentric images and GPS to address outdoor localization. In [14] it was presented an approach to perform large scale image-based localization based on direct 2D-to-3D matching.

Object Detection/Recognition in Cultural Sites. Our work is related to previous investigations using object detection to estimate the attention of visitors in a cultural site. The authors of [10] investigated the use of Fully Convlutional Networks to perform egocentric image classification and object deteciton in a museum. The authors of [15] presented a system to detect artworks and analyze audio activity to implement a smart audio guide with a smartphone. The proposed VEDI system leverages state of the art object detectors to recognize the points of interest observed by the visitors of a cultural site. Specifically, our system relies on the YOLOv3 object detector [13].

Behavioural Analysis of the Visitors of a Cultural Site. Previous works have investigated approaches for the behavioral analysis of the visitors of a cultural site. In particular, the authors of [16] have proposed empirically grounded models of individual and collective spatial behavior, whereas an accurate analysis of the behaviours of the visitors of an exhibition has been addressed in [8]. The proposed system tackles behavioral analysis by classifying visitors into four different profiles.

3 Experimental Cultural Sites and Datasets

Our system has been tested in two real cultural sites: "Monastero dei Benedettini"[1], which is an indoor environment, and "Orto Botanico"[2], which is a outdoor natural site. Specifically, we have collected and labeled different datasets of egocentric videos useful to drive the design of context-based localization algorithms and point of interest recognition algorithms. The datasets are also useful to and assess the performances of algorithms in both indoor and outdoor environments. The collected datasets, which are described in the following, are publicly available for research purposes at the following URL: http://iplab.dmi.unict.it/VEDI_project.

UNICT-VEDI. This dataset has been originally introduced in [12] to address context-based visitors localization and subsequently extended in [11] to study the problem of point of interest recognition. The dataset consists of several videos acquired using two wearable devices: HoloLens and GoPro Hero 4. The videos have been temporally annotated to indicate in each frame (1) the location of the visitor (the room in which they are located) and (2) the "point of interest" observed by the visitor (e.g., a painting, a statue or an architectural element). The dataset comprises a total of 9 environments and 57 points of interest. We also provide bounding box annotations for about 1000 frames for each of the 57 points of interest. This amounts to a total of 54248 frames labeled with bounding boxes in the whole dataset (see the Fig. 2).

EgoNature. This dataset has been collected in the natural site "Orto Botanico" to test context-based localization. In particular, the dataset contains 9 contexts, which are 9 areas of the site relevant for the visitors. Egocentric videos have been acquired using a Pupil 3D Eye Tracker headset coupled with a smartphone (Honor 9) to collect GPS locations which are later synced to the videos. More details are available in [9].

UNICT-VEDI_Succulente. This dataset has been collected in the natural site "Orto Botanico" to perform point of interest recognition. It includes 16 points of interest representing plants belonging to following families: (1) Apocynaceae, (2) Bombacaceae, (3) Cactaceae, (4) Crassulaceae, (5) Euphorbiaceae, (6) Lamiaceae, (7) Liliaceae. For each frame, we have annotated the plant depicted in the

[1] http://monasterodeibenedettini.it.
[2] http://ortobotanico.unict.it/.

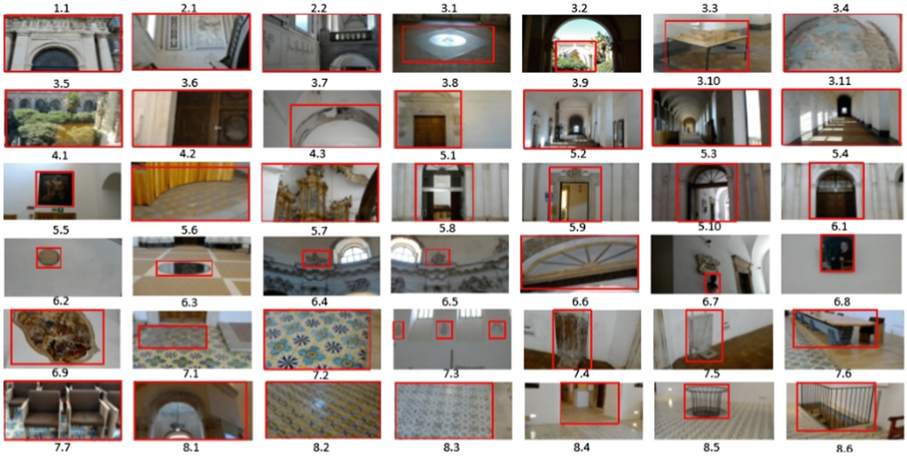

Fig. 2. Some examples from UNICT-VEDI along with bounding box annotations.

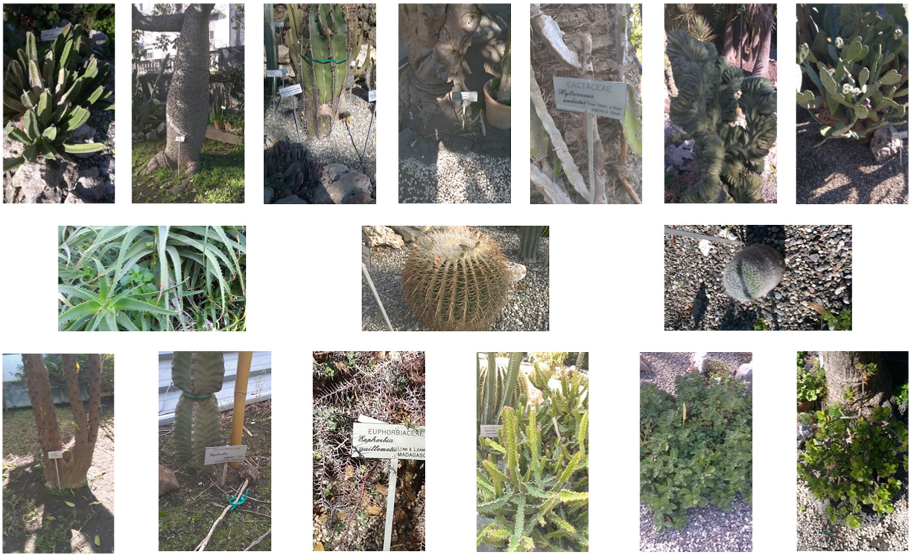

Fig. 3. Examples of the 16 plants belonging to the UNICT-VEDI_Succulente dataset.

image. The dataset contains 36, 728 labeled images. Figure 3 shows some images of the points of interest present in the dataset.

4 Architecture and Services

In this Section, we first discuss the general architecture of VEDI (Sect. 4.1), then present the services implemented by the system (Sect. 4.2).

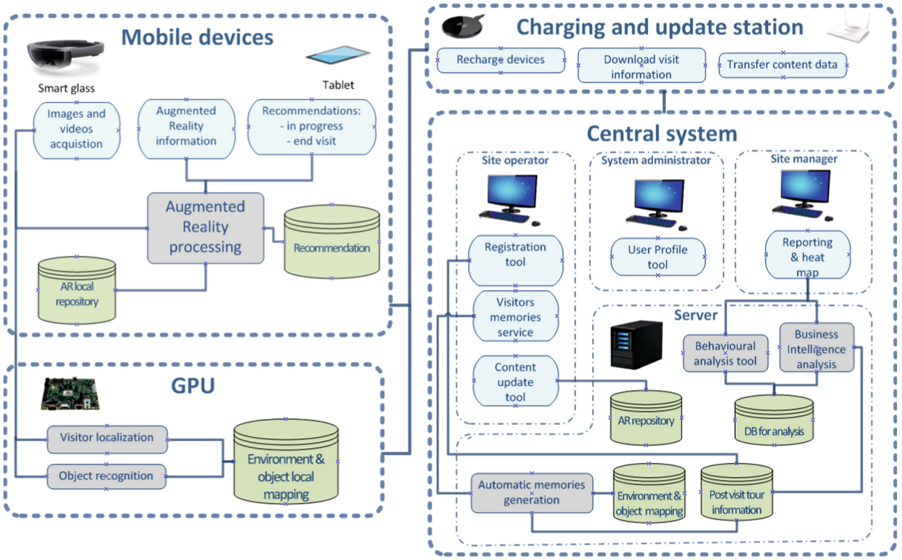
Mobile devices
Smart glass
Tablet
Images and videos acquistion
Augmented Reality information
Recommendations: - in progress - end visit
Augmented Reality processing
Recommendation
AR local repository
GPU
Visitor localization
Object recognition
Environment & object local mapping
Charging and update station
Recharge devices
Download visit information
Transfer content data
Central system
Site operator
System administrator
Site manager
Registration tool
User Profile tool
Reporting & heat map
Visitors memories service
Content update tool
Server
Behavioural analysis tool
Business Intelligence analysis
AR repository
DB for analysis
Automatic memories generation
Environment & object mapping
Post visit tour information

Fig. 4. The VEDI system is made up of 4 components: (1) Mobile devices, (2) GPU, (3) Charging and update station, (4) Central system.

4.1 Architecture

Figure 4 illustrates the high level architecture of the proposed VEDI system, which is made up of the following components:

- **Mobile devices:** mobile devices such as smart glasses and tablets are provided to the visitors of the cultural site. These devices are used to both acquire images and video from the point of view of the visitors, as well as to provide additional information or recommendations to the visitor through Augmented Reality;
- **Graphic Processing Unit (GPU):** directly connected to the wearable device, it is used to provide additional computational power in order to to process egocentric video and address visitor localization and object recognition;
- **Charging and update station:** used at the end of the visit to recharge the wearable devices, transfer the information collected during the visit (e.g., video) to the central system, and update the contents (e.g., 3D models) provided during the visit;
- **Central system:** handles system management, processes and store all data collected by the wearable devices. The central system comprises a *Server*, which includes components to handle the egocentric data collected during the visits and analyze it for behavioral analysis, business intelligence analysis and automatic generation of digital video memories to be provided to the visitors. Moreover, the following actors take part to the central system:

- *System administrator:* can access all system functions, define user profiles (site operator, site manager) and enable/disable specific functions;
- *Site operator:* can access the following functions: (1) “registration tool”, which allows to associate their identity to the assigned mobile device id; (2) “visitors memories service”, which automatically generates a video containing the salient moments of each visit to be sent to the user, postcard or other digital gadgets representing objects observed during the visit; (3) “content update tool” which allows to update the contents stored in the AR repository;
- *Site manager:* can use the “Reporting & Head-Map” tool to visualize performance indexes and statistic indicators generated after normalization, aggregation and management of data, as well as all behavioral information periodically extracted by the system using dedicated algorithms.

4.2 Services

This Section presents the services implemented by VEDI. Demo videos of the different services are available at the following URL: http://iplab.dmi.unict.it/VEDI_project/#video.

Localization and Points of Interest Recognition: Given the different nature of indoor and outdoor contexts, visitors localization and point of interest recognition are carried out using different algorithms. In indoor contexts (e.g., the *UNICT-VEDI* dataset), the system performs context-based localization of visitors by processing the acquired egocentric video with a multi-stage localization algorithm which we describe in details in [12]. The recognition of the points of interest observed by visitors is carried out using an approach based on a Yolov3 object detector [13], which is detailed in [11]. In the *outdoor contexts* (i.e., the *EgoNature* and *UNICT-VEDI_Succulente* datasets), we perform context-based localization by fusing GPS measurements and egocentric images by means of the multi-modal localization algorithm described in [9]. Recognition of points of interest is addressed in *UNICT-VEDI_succulente* as a classification problem, by fine-tuning an AlexNet CNN to discriminate between images belonging to the 16 different points of interest.

Augmented Reality. The AR GUI is triggered when a point of interest is recognized and observed for a significant amount of time. This leaves to the visitor the decision on which “augmented” information they are interested in. To reach this goal, the user interface has been designed according to the following three features:

1. The user interaction panel used to choose the multimedia contents of interest should not remain constantly in front of the visitor;
2. The GUI has been designed relying on the use of transparency to never completely impede the visibility of the external world;
3. The area engaged by the interface is designed to be as small as possible.

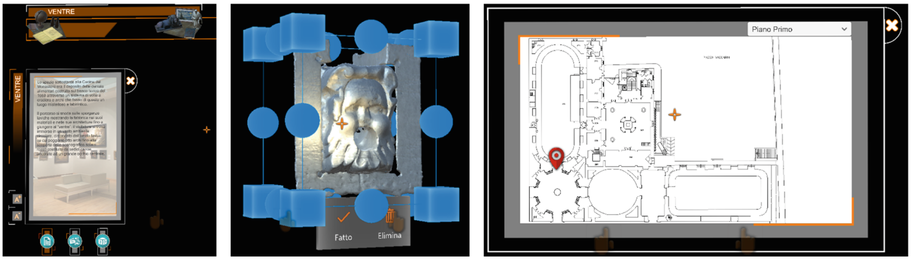

Fig. 5. AR GUI examples. From left to right: additional information on the observed point of interest, a 3D model shown to the visitor, a map showing the position of the visitor in the cultural site.

See Fig. 5 for same examples of the AR GUI.

Behavior Analysis and Visual Analytics. To study the behavior of visitors, we compute the following indicators for each cultural site:

- Attraction index: ratio between the number of visitors observing a given point of interest and total number of visitors;
- Retention index: measuring the average time spent in front of an information-communication element (e.g., a panel, a video, a caption, etc.);
- Usage times: times of use (for the overall visit, for specific sections, for types of users);
- Sweep Rate Index (SRI): the ratio between the total size of the exposure, in square meters, and the average time spent by visitors within the exposure itself;
- Diligent Visitor Index (DVI): the percentage of visitors who stopped in front of more than half of the points of interest of the cultural site.

Data Visualization. The VEDI platform is engineered to provide the managers of cultural sites with utilities and tools to create awareness on the visitors' behavior. Cultural site manager can explore visitors' behavioral data and have insights on the characteristics of each class of visitors (e.g., male-female, young-adult, low-high education, local-alien) through specific data report. This is done relying on the output of the localization and point of interest recognition algorithms discussed in the previous paragraphs. The data visualization tools offer the site manager a way to assess in which areas the visitors spend more time and the most followed routes inside the building (see Fig. 6). Finally, VEDI assists the managers by providing internationally known key performance indexes such as "Attraction Index", "Sweep Rate Index" and "Diligent Visitors Index" to benchmark the performances of the considered cultural site against similar sites.

Memories. This service allows to automatically generate a video summary of a visit by taking into account (1) semantic information about locations and observed points of interest obtained using the localization and point of interest recognition algorithms, (2) meta-data (e.g., photos, descriptions) on the site, contexts and points of interest.

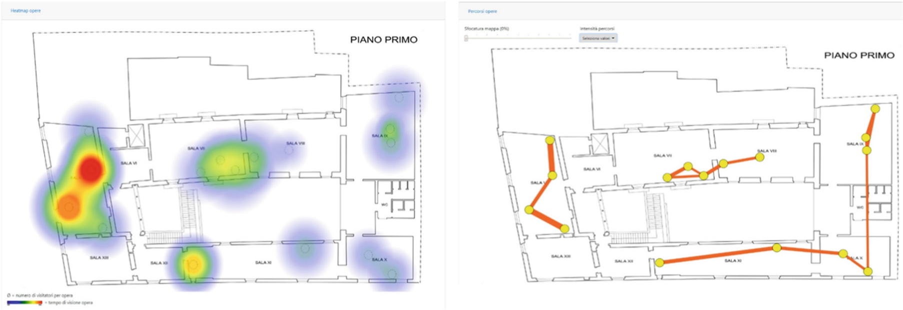

Fig. 6. An example of data visualization, where a heat map is used to demonstrate the behavior of the visitors.

5 Experimental Results

We tested our system to assess the performances of localization and point of interest recognition systems, which are at the core of VEDI. Table 1(a) reports the results of the context-based localization system on UNICT-VEDI. Following [12], we evaluate our system using a frame-based mean F_1 score (mFF_1), which is the mean F_1 scores across classes computed over frames, as well as a segment-based F_1 score ($mASF_1$), which is the mean average F_1 score, when the system is used to retrieve video segments (please see [12] for more details on the evaluation measures). As can be seen, the proposed approach achieves usable results with videos acquired using both HoloLens and GoPro considering all evaluation measures.

Table 1(b) compares the proposed approach for point of interest recognition [11] based on a Yolov3 object detector with respect to the following baseline: (1) 57-POI - the localization method proposed in [11] adapted for point of interest recognition, (2) 57-POI-N, as in 57-POI but training using both positive and negative frames (i.e., frames in which no point of interest is observed), (3) 9-Classifiers, the combination of 9 context-specific 57-POI point of interest recognition methods with the localization system proposed in [11]. It should be noted that $9 - Classifier$ is computational expensive both at training and test time due to the need to train context-specific recongition systems. All results are evaluated using mean frame-based F_1 score. Results confirm that the proposed approach allows to obtain good performances comparable with 9-Classifiers using a context-generic recognition system at a lower computational cost.

Table 1(c) reports the results of different variants of the proposed system for context-based outdoor localization which uses both egocentric images and GPS. Localization results are measured using frame-based accuracy. The time required to process and localize a single image in CPU is reported in milliseconds (ms). All methods use a variant of SqueezeNet to process images and a Decision Tree (DCT) to process GPS. SqueezeNet-n modeles denote a simplified (and hence faster) SqueezeNet architecture which considers only the n convolutional layers.

All methods obtain good results. Considerably faster inference is obtained using SqueezeNet-6 + DCT.

Table 1. The results obtained by VEDI system in the fundamental tasks of localization (a and c) and point of interest recognition (b).

	HoloLens	**GoPro**
mFF_1	0.82	0.81
$mASF_1$	0.71	0.71

(a)

Method	F_1 **score**
57-POI	0.59
57-POI-N	0.62
9-Classifiers	0.66
Proposed	0.68

(b)

Method	**Accuracy**	**Time (ms)**
SqueezeNet-6 + DCT	0.86	4.7
SqueezeNet-9 + DCT	0.86	6.09
SqueezeNet-11 + DCT	0.86	6.60
SqueezeNet + DCT	0.91	22.9

(c)

Regarding to the outdoor point of interest recognition system based on AlexNet, it achieves a mean F_1 score of 89.02% on the UNICT-VEDI_Succulente dataset when discriminating among the 16 considered points of interest.

6 Conclusion

We have presented VEDI, an integrated wearable system to assist the visitors of cultural sites by providing additional information on the observed points of view during their visits, as well as the site managers by automatically inferring useful performance indicators and behavioral information on the cultural sites. Experiments on two cultural sites highlight the good performance achieved by the proposed approach in the fundamental tasks of localizing visitors and recognizing points of interest.

Acknowledgment. This research is supported by PON MISE - Horizon 2020, Project VEDI - Vision Exploitation for Data Interpretation, Prog. n. F/050457/02/X32 - CUP: B68I17000800008 - COR: 128032, and Piano della Ricerca 2016-2018 linea di Intervento 2 of DMI of the University of Catania.

References

1. Baraldi, L., Paci, F., Serra, G., Benini, L., Cucchiara, R.: Gesture recognition using wearable vision sensors to enhance visitors' museum experiences. IEEE Sens. J. **15**, 2705–2714 (2015)
2. Battiato, S., Farinella, G.M., Gallo, G., Ravì, D.: Scene categorization using bag of textons on spatial hierarchy. In: International Conference on Image Processing, pp. 2536–2539. IEEE (2008)
3. Cucchiara, R., Del Bimbo, A.: Visions for augmented cultural heritage experience. IEEE Multimedia **21**(1), 74–82 (2014)
4. Furnari, A., Battiato, S., Farinella, G.M.: Personal-location-based temporal segmentation of egocentric video for lifelogging applications. J. Vis. Commun. Image Represent. **52**, 1–12 (2018)

5. Furnari, A., Farinella, G.M., Battiato, S.: Recognizing personal locations from egocentric videos. IEEE Transact. Hum. Mach. Syst. **47**, 6–18 (2017)
6. Gallo, G., Signorello, G., Farinella, G., Torrisi, A.: Exploiting social images to understand tourist behaviour. In: ICIAP, pp. 707–717 (2017)
7. Kendall, A., Grimes, M., Cipolla, R.: Posenet: A convolutional network for real-time 6-dof camera relocalization. In: ICCV, pp. 2938–2946 (2015)
8. Levasseur, M., Veron, E.: Ethnographie d'une exposition. Histoires d'expo, Peuple et culture, pp. 29–32 (1983)
9. Milotta, F.L.M., Furnari, A., Battiato, S., Salvo, M.D., Signorello, G., Farinella, G.M.: Visitors localization in natural sites exploiting egoVision and GPS. In: International Conference on Computer Vision Theory and Applications (2019)
10. Portaz, M., Kohl, M., Quénot, G., Chevallet, J.P.: Fully convolutional network and region proposal for instance identification with egocentric vision. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2383–2391 (2017)
11. Ragusa, F., Furnari, A., Battiato, S., Signorello, G., Farinella, G.M.: Egocentric point of interest recognition in cultural sites. In: VISAPP (2019)
12. Ragusa, F., Furnari, A., Battiato, S., Signorello, G., Farinella, G.M.: Egocentric visitors localization in cultural sites. ACM J. Comput. Cult. Heritage **12**, 11 (2019)
13. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. CoRR abs/1804.02767, http://arxiv.org/abs/1804.02767 (2018)
14. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & effective prioritized matching for large-scale image-based localization. PAMI **39**, 1744–1756 (2017)
15. Seidenari, L., Baecchi, C., Uricchio, T., Ferracani, A., Bertini, M., Bimbo, A.D.: Deep artwork detection and retrieval for automatic context-aware audio guides. TOMM **13**(3s), 35 (2017)
16. Shell, D.A., et al.: Spatial behavior of individuals and groups: preliminary findings from a museum scenario. In: IROS 2007 Workshops (2007)
17. Signorello, G., Farinella, G.M., Gallo, G., Santo, L., Lopes, A., Scuderi, E.: Exploring protected nature through multimodal navigation of multimedia contents. In: ACIVS, pp. 841–852 (2015)
18. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: Proceedings of Ninth IEEE International Conference on Computer Vision, 2003, pp. 273–280. IEEE (2003)
19. Weyand, T., Kostrikov, I., Philbin, J.: PlaNet - photo geolocation with convolutional neural networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 37–55. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_3