Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- We propose a dataset of egocentric videos for visitor behavior understanding.
- Collected by 70 subjects, the data has labels for locations and points of interest.
- We propose four tasks for visitor behavior understanding in cultural heritage.
- Experiments highlight that the proposed dataset can be a valuable benchmark.

Pattern Recognition Letters
journal homepage: www.elsevier.com

EGO-CH: Dataset and Fundamental Tasks for Visitors Behavioral Understanding using Egocentric Vision

Francesco Ragusa[a,b,2], Antonino Furnari[a,2], Sebastiano Battiato[a], Giovanni Signorello[c], Giovanni Maria Farinella[a,c,2,**]

[a] *DMI-IPLab, University of Catania*
[b] *XGD - XENIA s.r.l., Acicastello, Catania, Italy*
[c] *CUTGANA, University of Catania*

ABSTRACT

Equipping visitors of a cultural site with a wearable device allows to easily collect information about their preferences which can be exploited to improve the fruition of cultural goods with augmented reality. Moreover, egocentric video can be processed using computer vision and machine learning to enable an automated analysis of visitors' behavior. The inferred information can be used both online to assist the visitor and offline to support the manager of the site. Despite the positive impact such technologies can have in cultural heritage, the topic is currently understudied due to the limited number of public datasets suitable to study the considered problems. To address this issue, in this paper we propose EGOcentric-Cultural Heritage (EGO-CH), the first dataset of egocentric videos for visitors' behavior understanding in cultural sites. The dataset has been collected in two cultural sites and includes more than 27 hours of video acquired by 70 subjects, with labels for 26 environments and over 200 different Points of Interest. A large subset of the dataset, consisting of 60 videos, is associated with surveys filled out by real visitors. To encourage research on the topic, we propose 4 challenging tasks (room-based localization, point of interest/object recognition, object retrieval and survey prediction) useful to understand visitors' behavior and report baseline results on the dataset.

1. INTRODUCTION

Cultural sites receive many visitors every day. For a cultural site manager, it is hence paramount to 1) provide services able to assist the visitors, and 2) analyze their behavior to measure the performance of the site and understand what can be improved. For example using indicators [1] such as: a) Attraction index: to measure how much a point of interest attracts the visitors, b) Retention index: to measure the average time spent observing information element (e.g., a caption, a video a panel, etc.), c) Sweep Rate Index (SRI): it is used to calculate if visitors move slowly or quickly through the exhibition, d) Diligent Visitor Index (DVI): the percentage of visitors who stopped in front of more than half of the points of interest. Classic approaches addressed the former task through the delivery of printed material (e.g., maps of the museum), the use of audio-guides and the installation of informative panels. Similarly, the analysis of visitors' behavior has generally been performed through the administration of questionnaires. It should be noted that such approaches often require manual intervention and are limited especially when the number of visitors is large. Recent works [2, 3, 4] have highlighted that the use of wearable devices such as smart glasses can provide a convenient platform to tackle the considered tasks in an automated fashion. Using such technology, it is possible to provide to the user services such as automated localization (e.g., to help visitors navigating the site) and recognition of currently observed Ponts Of Interest (POIs)[3] to provide more information on relevant objects and suggest what to see next. Conveniently, localization and POI

**Corresponding author: Tel.: +39 095 7337 219; fax: +39 095 330094;
e-mail: gfarinella@dmi.unict.it (Giovanni Maria Farinella)
[2] These authors are co-first authors and contributed equally to this work.

[3] In this work, we refer to the definition of Point Of Interest (POI) given in [5], as an element which can attract the attention of visitors. Most POIs are objects such as paintings and statues, but architectural elements such as pavements can qualify as POIs, despite not being objects. Therefore, in this paper the notations "Point Of Interest" and "object" are not used interchangeably.

recognition can be used by the manager of the cultural site to obtain information about the visitors and understand their behavior by inferring where they have been, how much time they have spent in a specific environment and what POIs have been liked most.

Despite the aforementioned technologies can have a significant impact on cultural heritage, they are currently underexplored due to the lack of public benchmark datasets. To address this issue, in this paper we propose EGOcentric-Cultural Heritage (EGO-CH), the first large dataset of egocentric videos for visitors behavioral understanding in cultural sites. The dataset has been collected in two cultural sites located in Sicily, Italy: Galleria Regionale di Palazzo Bellomo[4] and Monastero dei Benedettini[5]. The overall dataset contains more than 27 hours of video, including 26 environments, over 200 Points of Interest and 70 visits. We release EGO-CH with a set of annotations useful to tackle fundamental tasks related to visitors behavior understanding in cultural sites, and specifically, temporal labels specifying the location of the visitor as well as the currently observed POI, bounding box annotations around POIs, surveys filled out by visitors at the end of each tour in the cultural site. Figure 1 reports some sample frames from the proposed dataset. The dataset can be publicly accessed upon request to the authors from our webpage *http://iplab.dmi.unict.it/EGO-CH/*.

We propose 4 fundamental tasks for visitors behavioral understanding using egocentric vision: 1) *room-based localization*, consisting in recognizing the environment in which the visitor is located in each frame of the video, 2) *Point of Interest recognition*, which consists in correctly detecting and localizing all objects in the image frames, 3) *object retrieval*, which consists in matching an observed object from the egocentric point of view to a reference image contained in the museum catalogue of all artworks, 4) *survey prediction*, which consists in generating the survey associated to a visit from video. We also provide baseline results for each task on the proposed dataset. The experimental results suggest that the proposed dataset is a challenging benchmark for visitors behavioral understanding using egocentric vision.

In sum, the contributions of this work are: 1) we present EGO-CH, a new challenging dataset of egocentric videos acquired in two cultural sites, 2) the dataset has been labeled to tackle 4 main tasks useful to understand visitors behavior, 3) we report baseline results for each task.

2. RELATED WORK

Visitors Behavioural Understanding and Site Manager Assistance in Cultural Sites Several works investigated the use of wearable systems to augment the fruition in cultural sites [2]. Razavian et al. [6] proposed a method to estimate the attention of the visitors of an exhibition, whereas in [7] a CNN to perform localization and object recognition is introduced in order to develop a context aware audio guide. Raptis et al. [8] studied the design of mobile applications in museum environments and highlighted that context influences interaction. In [3, 4], the problem of localizing the visitors of a museum from egocentric videos is considered. The inferred localization can be used to provide behavioral information to the manager of the site. Past works investigated specific applications, generally relying on data collected on purpose and not publicly released. In this work, we aim at standardizing the fundamental problems of visitors behavioral understanding in cultural sites by proposing a public dataset and a series of tasks.

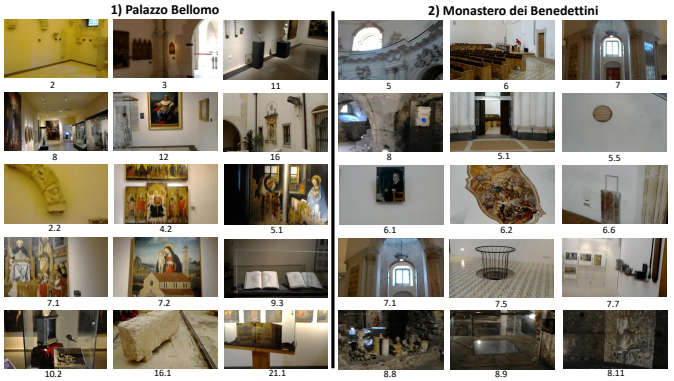


Fig. 1. Sample frames from the two cultural sites belonging to EGO-CH: 1) Palazzo Bellomo, 2) Monastero dei Benedettini. The first two rows show frames extracted from the training videos and related to the environments, whereas the remaining rows show frames of the training videos related to POIs. See Section 3 for more details.

Datasets on Cultural Heritage Few image-based datasets focusing on cultural heritage have been proposed in past works. Koniusz et al. [9] proposed the OpenMIC dataset containing photos captured in ten different exhibition spaces of several museums and explored the problem of artwork identification. DelChiaro et al. [10] proposed NoisyArt, a dataset composed of artwork images collected from Google Images and Flickr correlated by metadata gathered from DBpedia. In contrast with the aforementioned works, we propose the first dataset composed of egocentric videos, and release it publicly. The dataset can be used to address different tasks related to visitors behavioral understanding in cultural sites. A significative part of the proposed dataset has been collected by real visitors (i.e., 60 visits) and hence it is a realistic set of data for benchmarking.

Localization Ahmetovic et al. [11] presented NavCog, a system to navigate with a smartphone in complex indoor and outdoor environments exploiting Bluetooth Low Energy beacons. Kendall et al. [12] proposed to infer the 6 Degrees of Freedom pose of a camera from egocentric images using a CNN. In [3], it has been considered the problem of localizing a visitor in a cultural site from egocentric images to provide behavioral information to the site manager. In this work, we consider the work presented in [3] as a baseline for the localization task.

Point Of Interest/Object Recognition Seidenari et. al [7] and Taverriti et al. [13] proposed to perform object classification and artwork recognition to assist tourists with additional

[4] `http://www.regione.sicilia.it/beniculturali/palazzobellomo/`.

[5] `http://www.monasterodeibenedettini.it/`

Environment	#video	#frame
1 Sala1	1	3721
2 Sala2	4	7968
3 Sala3	4	8285
4 Sala4	5	11497
5 Sala5	5	11461
6 Sala6	1	2630
7 Sala7	4	10613
8 Sala8	2	6910
9 Sala9	4	10505
10 Sala10	3	5830
11 Sala11	3	7343
12 Sala12	1	2463
13 Sala13	1	3040
14 Cortile degli Stemmi	2	5853
15 Sala delle Carrozze	1	3259
16 Cortile Parisio	4	10374
17 Biglietteria	2	4099
18 Portico	2	5701
19 Scala Catalana	2	6399
20 Loggetta	2	4169
21 Box Sala8	2	4661
22 Area Sosta	2	3505
Total	**57**	**140286**

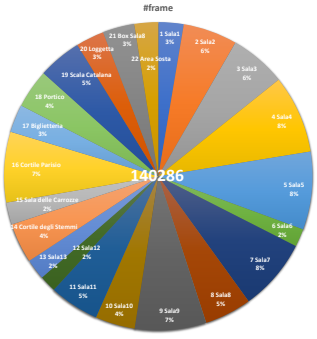


Fig. 2. Number of training videos collected in each environment and corresponding number of frames for the cultural site "Palazzo Bellomo" (left), along with a pie chart representation of the same data (right).

Fig. 3. Some example bounding box annotations from the cultural site "Monastero dei Benedettini".

information about the observed objects. In general, object detectors (e.g., YOLOv3 [14]) have been used to detect artworks in cultural sites. However, it should be noted that, as pointed out in [5], depending on the cultural site, not all Points Of Interest are objects. For instance, a point of interest can be an architectural element such as a pavement, or even a corridor. In this case, it should be considered that object detectors can be limited. In this work, we consider the YOLOv3 object detector [14] as baseline for Point Of Interest/Object recognition.

Object Retrieval Many previous works investigated approaches to image retrieval. Rubhasy et al. [15] used an ontology-based approach to retrieval in multimedia cultural heritage collections. The goal is to enable the integration of different types of cultural heritage media and to retrieve relevant heritage media given a query. Kwan et al. [16] proposed matrix of visual perspectives to address Content-based Image Retrieval (CBIR) of cultural heritage symbols, whereas Iakovidis et al. [17] perform pattern-based Content-based Image Retrieval. The work of [18] focused on discarding image outliers using Content-based Image Retrieval. Despite the availability of advanced approaches, for generality and ease of comparison, in this paper we consider simple baselines based on image representation and nearest neighbor search to address the object retrieval task.

3. THE EGO-CH DATASET

3.1. Data Collection

The dataset has been acquired using a head-mounted Microsoft HoloLens device in two cultural sites located in Sicily, Italy: 1) Palazzo Bellomo (Table 1), located in Siracusa[6], and 2) Monastero dei Benedettini (Table 2), located in Catania[7].

Palazzo Bellomo This cultural site is composed of 22 environments and contains 191 Points of Interest (e.g., statues, paintings, etc.).[8] Training videos have been collected by operators instructed to walk around in order to capture images of each environment from different points of view. To simplify labeling, each training video contains only frames from a given environment. At least one training video has been collected per environment. In the case of outdoor environments (e.g., courtyards), we collected multiple videos to include different lighting conditions. We have collected a total of 57 training video in this cultural site. Figure 1(left) shows some frames acquired in the considered cultural site, whereas Figure 2 reports the number/percentage of frames acquired in each environment. Ten test videos have been collected separately asking 10 volunteers to visit the cultural site. One of the 10 videos (i.e., "Test 3") was selected randomly and used as validation set, whereas the remaining 9 videos are used for evaluation purposes. No specific instructions on where to go, what to look at and how much time to spend in a specific environment/POI has been provided to the visitors. Most of the subjects had limited confidence with the cultural site. This provided a natural means to collect realistic data of visitors exploring the environments and observing Points of Interest. All the videos have a resolution of 1280×720 pixels and a frame-rate of 29.97 fps. The average duration of test videos is 31.27 *min*, with the longest one being 50.23 *min*. See the supplementary material for more details about training/test videos. We also include 191 reference images related to the considered POIs to be used for one-shot image retrieval. The images are akin to the images generally included in museum catalogs.[9]

Monastero dei Benedettini This dataset is composed of 4 environments and contains 35 Points Of Interest.[10] Differently from "Palazzo Bellomo", the POIs belonging to this cultural site include both objects such as paintings and statues as well as architectural elements, such as pavements, which cannot be easily recognized using object detection techniques as noted in [5]. See Figure 1(right) for some qualitative examples of the considered objects. Training videos have been collected with the same acquisition modality considered for the "Palazzo Bellomo" cultural site. Figure 4 reports the number/percentage of frames acquired in each environment. Training and validation videos have a resolution of 1216×684 pixels and a frame-rate

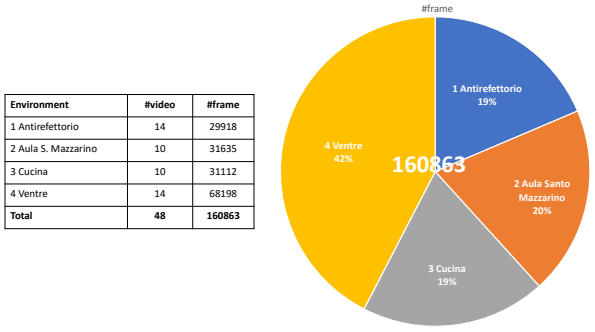
[6] http://www.regione.sicilia.it/beniculturali/palazzobellomo/
[7] http://monasterodeibenedettini.it/
[8] See the supplementary material for the list of environments and POIs.
[9] Examples reference images for both cultural sites are included in the supplementary material.
[10] See the supplementary material for the list of environments and POIs.

Table 1. Details regarding the cultural site "Palazzo Bellomo".

Subset	Resolution	FPS	AVG Time (min)	# POIs	#environments	bbox annotations	temporal segments
Training	1280x720	29.97	1.4	191	22	56686	57
Test	1280x720	29.97	31.27	191	22	13402	340

Table 2. Details regarding the cultural site "Monastero dei Benedettini".

Subset	Resolution	FPS	AVG Time (min)	# POIs	#environments	bbox annotations	temporal segments
Training	1216x684	24.00	2.2	35	4	33366	48
Validation	1216x684	24.00	3.5	35	4	2235	20
Test	1408x792	30.03	21	35	4	71310	455



Environment	#video	#frame
1 Antirefettorio	14	29918
2 Aula S. Mazzarino	10	31635
3 Cucina	10	31112
4 Ventre	14	68198
Total	**48**	**160863**

Fig. 4. Number of training videos collected in each environment and corresponding number of frames for the cultural site "Monastero dei Benedettini" (left), along with a pie chart representation of the same data (right).

of 24 fps. Five validation videos have been collected by asking volunteers to visit the cultural site following the same protocol used for "Palazzo Bellomo". Additionally, we collected 60 test videos by asking real visitors inexperienced with both the research project and its goals and the HoloLens device to freely visit the cultural site. No specific instructions have been given to the visitors, who were free to explore the 4 environments and the 35 POIs. This allowed us to obtain realistic data of how a visitor would move in a cultural site. Test videos have been collected over a period of three months. Moreover, at the end of the visit, we administered the visitor a survey, the content of which is described in Section 3.2. The 60 test videos have a resolution of 1408×792 pixels and a frame-rate of 30.03 fps. The average video length is 21 *min*, with the maximum length being 42 *min*. See the supplementary material for more details about training/validation/test videos. Similarly to "Palazzo Bellomo", we include 35 reference images related to the considered POIs for one-shot image retrieval[9]. Please note that this set of data is adapted from and extends significantly the dataset proposed in [3], introducing 60 new labelled videos collected by real visitors. Specifically, the overall dataset presented in this work contains +1600 minutes of video, data from +70 more subjects, +91369 bounding box annotations and an additional cultural site "Palazzo Bellomo" comprising 22 environments and 191 points of interest.

3.2. Annotations

Temporal Labels All test and validation videos have been temporally labeled to indicate in every frame the environment in which the visitor is located and the observed point of interest, if any. If the visitor is not located in one of the considered environment (e.g., a stair), the frame is marked as "negative"[11]. It is worth noting that there are no negative frames in "Palazzo Bellomo" since all environments are part of the museum, whereas negative frames are contained in "Monastero dei Benedettini". This is due to the different nature of the two sites: "Palazzo Bellomo" is a museum, consisting in a limited set of rooms, whereas "Monastero dei Benedettini" is a much more complex environment including many corridors and stairs which have not been labeled as locations of interest for visitors. Similarly, we mark as "negative" all frames in which the visitor is not observing any of the considered POIs. Each location is identified by a number that denotes a specific environment ($1-22$ for "Palazzo Bellomo" and $1-4$ for "Monastero dei Benedettini"). Each point of interest is denoted by a code in the form X.Y (e.g., 3.5) where "X" denotes the environment in which the point of interest is located and "Y" identifies the point of interest. See Figure 1 for some examples.

Bounding Box Annotations A subset of frames from the dataset (sampled at at 1 fps) has been labeled with bounding boxes indicating the presence and locations of all POIs. Specifically, each POI has been labeled with a tuple $(class, x, y, w, h)$ indicating the class of the POI and its bounding box information. It is worth mentioning that, as noted in [5], a POI can be an object (e.g., a painting or a statue) or a different element (e.g., a pavement or a specific location), which cannot be strictly defined as an object. Indeed, the kind of POIs contained in a cultural site depends on the nature of the site itself. In EGO-CH, "Palazzo Bellomo" contains only objects as POIs, whereas "Monastero dei Benedettini" contains both objects and other elements. Nevertheless, all elements are labeled with class type and bounding box annotations. Figure 3 shows examples of labeled frames from the 60 visits of "Monastero dei Benedettini".

Surveys The 60 test videos collected in the "Monastero dei Benedettini" are associated with surveys which have been administered to the visitors at the end of the visits. Specifically, the visitors are asked to rate a subset of 33 out of the 35 Points Of Interest (a picture of each point is shown) or specify if any of them had not been seen it during the visit. The rating is expressed as a number ranging from -7 (not liked) to $+7$ (liked).

The EGO-CH dataset is publicy available at our website: *http://iplab.dmi.unict.it/EGO-CH/*. The reader is referred to the supplementary material for more details about the dataset and the experiments. The dataset can be used only for research purposes and is available upon the acceptance of an agreement.

[11] Examples of "negative" frames are reported in the supplementary material.

4. PROPOSED TASKS AND BASELINES

In this Section, we propose four tasks which can be addressed using the proposed dataset. The tasks are related to problems investigated in previous works on cultural heritage [4, 3, 7, 13]. We believe that solving these tasks can bring useful information about the behavior of the visitors of a cultural site.

4.1. Room-based Localization

Task: The task consists in determining the room in which the visitor of a cultural site is located from egocentric images collected using a wearable device. Localization information can be used both to provide a "where am I" service to the visitor and to collect behavioral information useful for the site manager to understand what paths do visitors prefer and where they spend more time in the cultural site.

Baseline: As a baseline for this task, we consider the approach proposed in [3, 19]. This approach is selected as a baseline due to the limited work on room-based localization in the cultural heritage domain [3] and due to the state-of-the-art performance of the approach shown in [19]. Given a set of locations, the considered approach allows to segment a given video into video shots related to the specified locations. If a given shot is not related to any of the locations, the algorithm automatically labels it as a "negative segment" through a "negative rejection" stage. The method is composed by three steps, as illustrated in Figure 5. For each cultural site, we trained a VGG-19 CNN to discriminate between locations ("Discrimination" stage). The "Negative Rejection" step has been considered only for the data of "Monastero dei Benedettini", since "Palazzo Bellomo" does not contain negative locations. The "Sequential Modeling" stage allows to obtain a temporal segmentation of the input video where each segment is associated to one of the considered environments. This algorithm is chosen as it achieves state-of-the-art performance in the task of location-based egocentric video segmentation [19, 3]. Two hyper-parameters are involved in the algorithm: K, related to the "negative rejection" stage and ϵ, which regulates the amount of temporal smoothing applied to the predictions. The reader is referred to [3] for more details.

Implementation Details and Evaluation Measures: We evaluated our method following [3] using FF_1 score and ASF_1 score. Specifically, the FF_1 score is the F_1 score applied to individual frames and, as such, it does not evaluate the ability of the methods to produce a temporally coherent segmentation. ASF_1 is the F_1 score applied to temporal segments rather than frames and measures the ability to detect video segments coherent with the ground truth. Both scores are normalized between 0 and 1. The hyper-parameters of the algorithm K and ϵ are tuned on the validation sets of the proposed dataset. Specifically, $\epsilon = 10^{-273}$ is found by optimizing the validation ASF_1 score with a grid search in the range $[10^{-1} : 10^{-299}]$ on "Palazzo Bellomo" (see [3] for details). Since no negative locations are contained in "Palazzo Bellomo", the "negative rejection" stage is not performed and hence the parameter K is not optimized. Similarly, we find $\epsilon = 10^{-89}$ and $K = 100$ on "Monastero dei Benedettini".[12]

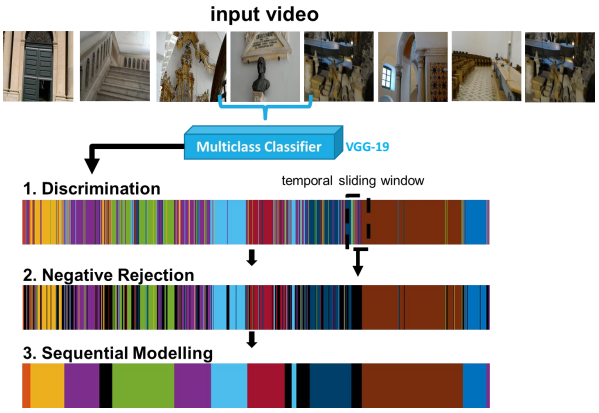


Fig. 5. The method used to perform room-based localization. The method is composed by three steps: 1) Discrimination, 2) Negative Rejection, 3) Sequential Modeling. See [19, 3] for more details.

Results: Table 3 reports the results obtained by the baseline in the two cultural sites[13]. On "Palazzo Bellomo", the baseline achieves good FF_1 scores for most rooms, obtaining an average value of 0.81. Much lower results are observed when the ASF_1 score is considered. In this case, an average value of 0.59 is reached. Lower results equal to 0.68 and 0.40 are obtained in the "Monastero dei Bendettini". This is partly due to the presence of negatives, which are not included in "Palazzo Bellomo" and to the more challenging nature of the test set of "Monastero dei Benedettini", which contains 60 videos collected by real visitors within 3 months with different lighting condition and blur as shown in Figure 6. The overall results highlight that addressing the considered task on the proposed dataset is challenging. In particular, issues such as varying lighting conditions and the presence of negatives need to be addressed in task-specific investigations.

4.2. Point of Interest/Object Recognition

Task: This tasks consists in recognizing the points of interest which the user is looking at. This can be useful to understand the visitor's behavior and answer questions as "What are the most viewed points of interest?" and "How long have they been observed?". Moreover, a system able to recognize points of interest could suggest the visitor what to see next, as well as provide information with Augmented Reality. The dataset could be used to perform standard object detection task.

Baseline: Due to its real-time performance and to its popularity in the cultural heritage domain [5, 7, 13], we consider a YOLOv3 [14] object detector as a baseline for the task. The detector has been trained on the training sets of "Palazzo Bellomo" and "Monastero dei Benedettini".

Implementation Details and Evaluation Metrics: We trained YOLOv3 using the standard anchors provided by the authors for the COCO dataset. We use mean Average Precision

[12] The supplementary material reports more implementation details.

[13] Extended tables, qualitative results and confusion matrix are included in the supplementary material.

Table 3. Room-based localization results. For each cultural site, the last row reports the Average (AVG) of the FF_1 and ASF_1 scores.

1) Palazzo Bellomo

Room	FF_1 score	ASF_1 score
Sala1	0.71	0.48
Sala2	0.92	0.79
Sala3	0.84	0.50
Sala4	0.92	0.59
Sala5	0.94	0.64
Sala6	0.77	0.52
Sala7	0.94	0.61
Sala8	0.89	0.64
Sala9	0.91	0.47
Sala10	0.84	0.69
Sala11	0.84	0.58
Sala12	0.80	0.66
Sala13	0.80	0.66
Cortile degli Stemmi	0.85	0.64
Sala Carrozze	0.91	0.67
Cortile Parisio	0.75	0.50
Biglietteria	0.65	0.44
Portico	0.69	0.51
Scala Catalana	0.76	0.63
Loggetta	0.71	0.51
Box Sala8	0.94	0.79
Area Sosta	0.43	0.47
AVG	**0.81**	**0.59**

2) Monastero dei Benedettini

Class	FF_1 score	ASF_1 score
Antirefettorio	0.75	0.54
Aula S. Mazzarino	0.33	0.12
Cucina	0.79	0.34
Ventre	0.97	0.60
Negative	0.54	0.33
AVG	**0.68**	**0.40**

(mAP) with threshold on IoU equal to 0.5 for the evaluations. In order to use YOLOv3 to detect artworks, a detection threshold is specified to discard detections with low confidence scores. For each cultural site, we tuned this threshold on the validation sets by choosing the value which maximizes mAP in the range $[5^{-4}; 1^{-3}; 5^{-3}; 1^{-2}; 3^{-2}; 5^{-2}; 0.10; 0.15; 0.2; 0.25; 0.3; 0.35; 0.40]$. To train the detector on "Palazzo Bellomo", we set the initial learning rate to 0.001 and the detection threshold to 0.01. On "Monastero dei Benedettini", we set the initial learning rate to 0.01 and the detection threshold to 0.001.

Results: Table 4 reports the results obtained in the two cultural sites. The results obtained on "Palazzo Bellomo" are much lower than the ones obtained on "Monastero dei Benedettini" mainly because of the larger set of POIs contained in the former site (191) versus the lower number of POIs contained in the latter (35). In both cases, the results are in general very low, which highlights the challenging nature of the proposed dataset and tasks. Among the challenges of the dataset, as previously discussed, it should be considered that some of the points of interest represent architectural elements such as corridors or pavements, which might be challenging to detect with a simple object detector, as pointed out in [5]. Moreover, differently from other object detection tasks, POIs here need to be recognized at the instance level. For instance, the dataset contains multiple paintings which should be recognized as separate objects. We leave the investigation of more specific approaches to future investigations.

Fig. 6. Some sample frames from different visits acquired within 3 months. Each row represents similar positions in the same environment with different lighting conditions.

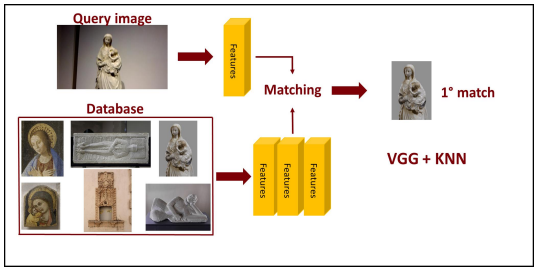


Fig. 7. Diagram of the baseline for the object retrieval task.

4.3. Object Retrieval

Task: Given a query image containing an object, the task consists in retrieving an image of the same object from a database. This task can be useful to perform automatic recognition of artworks when detection can be bypassed, i.e., when the user places the artwork in the center of the field of view using a wearable or mobile device. Moreover, the task is particularly of interest especially considering that artwork detection is a hard task, as highlighted in the previous section. We obtain a set of query images by extracting image patches from the bounding boxes annotated in the test set and consider two variants of the task. This accounts to *23727* image patches for "Palazzo Bellomo" and *44978* image patches for "Monastero dei Benedettini".[14] We consider two variants of this task. In the first variant, object retrieval is framed as a one-shot retrieval problem. In this case, the database contains only the reference images associated to each POI, whereas the whole set of image patches is used as the test set, i.e., only a single labeled sample is assumed to be available for each object. In the second variant, we split the set of image patches into a training set (70% - used as DB) and a test set (30%). It should be noted

[14]The supplementary material reports examples of extracted image patches.

Table 4. Object detection results. The reported mean Average Precision (mAP) is averaged over all test videos. Per-class Average Precision (AP) values are reported in the supplementary material.

Cultural Site	mAP
1) Palazzo Bellomo	10.59%
2) Monastero dei Benedettini	15.45%

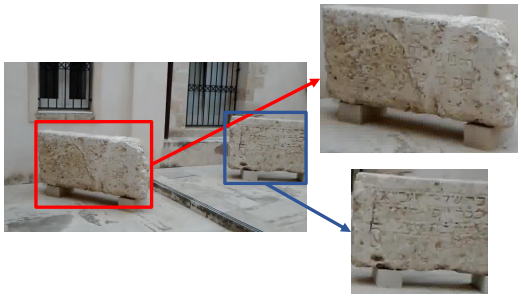


Fig. 8. Example of patches extracted using bounding boxes annotations.

that the first variant of the task is much more challenging both due to the presence of few labeled samples and to the domain shift which affects the two sets of images: reference images for the POIs and image patches cropped from egocentric images. Fig. 8 shows an example of image patches cropped from the egocentric images using bounding box annotations.

Baseline: Given the lack of investigation of approaches for retrieval in the scenario of First-Person vision in the cultural heritage domain, we consider a simple image-retrieval pipeline for both variants of the task. The pipeline uses VGG19 CNN pre-trained on ImageNet to represent image patches, while matching is perfor and matched using a K-NN. A scheme of the considered baseline is shown in Figure 7.

Implementation Details and Evaluation Measures: We have extracted all features from the FC7 layer of the VGG19 network pre-trained on ImageNet. When the second variant of the task is considered, we perform K-NN using $K = [1; 3; 5]$. We evaluated the performance of our baseline using standard metrics for image-retrival: precision, recall and F_1 score.

Results: Table 5 shows the results of the baseline on the image retrieval variants. In both cultural sites, one-shot retrieval does not achieve good results. This is probably due to the fact that one-shot retrieval relies on a limited number of training samples, which are drawn from a different distribution as compared to test samples. This suggests that dedicated methodologies should be considered to tackle one-shot retrieval and the domain shift problem. Better results are obtained on both sites in the second variant of the task, when the effect of one-shot retrieval and domain shift is reduced. Best results are obtained in "Palazzo Bellomo" for $K = 1$ (F_1 score of 0.67) and in "Monastero dei Benedettini" for $K = 5$ (F_1 score of 0.88).

4.4. Survey Prediction

Task: Each test video of the "Monastero dei Benedettini" is associated to a survey collected from visitors at the end of the visit. We define this task as predicting the content of a survey from the analysis of the related egocentric video. We deem this to be possible as the egocentric video contains information

Table 5. Object retrieval results for the two variant of the task.
Points of Interest Retrieval

1) Palazzo Bellomo

Variant	K	Precision	Recall	F_1 score
1 - One Shot	1	0.004	0.007	0.001
2 - Many Shots	1	0.69	0.66	**0.67**
	3	0.69	0.62	0.62
	5	0.69	0.62	0.62
	7	0.68	0.62	0.62
	9	0.67	0.61	0.62
	11	0.67	0.61	0.61

2) Monastero dei Benedettini

Variant	K	Precision	Recall	F_1 score
1 - One shot	1	0.29	0.07	0.08
2 - Many Shots	1	0.87	0.87	0.87
	3	0.88	0.87	0.87
	5	0.88	0.88	**0.88**
	7	0.88	0.87	0.87
	9	0.87	0.87	0.87
	11	0.87	0.86	0.86

on what the visitor has seen during the visit. In particular, the task consists in predicting for each POI 1) if the POI has been remembered by the visitor and 2) how the POI would be rated by the visitor in a $[-7, 7]$ scale. This task investigates automatic algorithms for automatically "filling in" surveys from videos.

Baseline: Since the proposed task is novel and very challenging, as a proof of concept, we propose a baseline which takes as input the temporal annotations annotations indicating the objects observed by the visitors in the 60 visits. To obtain fixed-length descriptors for each video, we accumulate the number of frames in which a given POI has been observed in a Bag Of Word representation. In such representation, each component of the fixed-length vector indicates the total time in which a specific point of interest has been observed by the visitor. The vector is hence sum-normalized to reduce the influence of videos with different lengths. The whole training set is normalized with z-scoring and classification is performed using K-NN. We consider two baselines. The first one simply performs a binary classification to predict whether a POI has been remembered by the visitor or not. The second one predicts both if the POI has been seen and what score has been assigned to it. This is tackled as a 15-class classification problem, where class -8 indicates that the POI has not been remembered, whereas the other 14 classes represent the scores from -7 to 7 assigned by the visitors to POIs. We would like to note that we treat the problem as a classification task, as the scores assigned by the visitors are discrete integer numbers. Also, the dataset contains a limited set of data-points, which would prevent the algorithm from generalizing beyond the discrete set of labels available at training time.

Implementation Details and Evaluation Measures: We perform our experiments using a leave-one-out strategy. We tested different values for k ranging from 1 to 9 and chose $K = 9$ which resulted to be optimal in our experiments. We evaluate results with weighted precision, recall and F_1 score.

Results: Table 6 reports the results obtained in the case of binary classification (remembered vs not remembered)[15]. The

[15] See the supplementary material for the extended tables.

Table 6. Survey prediction results - binary classification task.

Class	Precision	Recall	F_1 score	support
Not Remembered	0,43	0,2	0,27	561
Remembered	0,74	0,89	0,81	1419
AVG	0,65	0,7	**0,66**	1980

Table 7. Survey prediction results - multi-class classification. "Weighted AVG" reports the average scores weighted by the number of samples in each class.

Class	Precision	Recall	F_1 score	Support
Not Remem.	0,32	0,63	0,43	561
-7	0,52	0,24	0,33	49
-6	0	0	0	8
-5	0	0	0	8
-4	0	0	0	5
-3	0	0	0	5
-2	0,09	0,08	0,08	13
-1	0	0	0	10
0	0,18	0,15	0,17	104
1	0	0	0	36
2	0,02	0,02	0,02	65
3	0,12	0,02	0,04	91
4	0,1	0,04	0,06	181
5	0,13	0,07	0,09	213
6	0,14	0,09	0,11	248
7	0,33	0,29	0,31	383
weighted AVG	0,23	0,27	**0,23**	1980

number of instances belonging to each class is reported in the last column. The results suggest that this task is very challenging. Indeed, even if a POI appears in some frames, this does not imply that the visitor remembers it. Table 7 shows that the multi-class task[15] is even more challenging, with classes containing fewer examples (e.g., $-6, -5, -4, -3$) hard to recognize. As a final remark, it is worth noting that the results suggest that the task can be addressed to some degree. We expect that more complex approaches leveraging the analysis of the semantics of the input videos and the estimation of the attention of the visitor can achieve much better performance.

The code of our baselines is public available. See our webpage for the details: https://iplab.dmi.unict.it/EGO-CH/#code.

5. CONCLUSION

We presented EGO-CH, a dataset for visitors behavioral understanding using egocentric vision. The dataset includes more than 27 hours of video, 70 visits acquired by real visitors, 26 environments and over 200 different points of interest related to two different cultural sites. We publicly release the dataset along with temporal labels for locations and observed points of interest, bounding box annotations for objects, and surveys associated to 60 visits. Baseline results on the challenging tasks of Room-based Localization, Point of Interest/Object Recognition, Object Retrieval and Survey Prediction show the potential of the dataset for visitors behavioral understanding. We believe that EGO-CH can be a valuable benchmark to tackle the proposed tasks, as well as others not investigated in this paper.

Acknowledgments

This research is part of the project VALUE - Visual Analysis for Localization and Understanding of Environments (N. 08CT6209090207, CUP G69J18001060007) supported by PO FESR 2014/2020 - Azione 1.1.5. - "Sostegno allavanzamento tecnologico delle imprese attraverso il finanziamento di linee pilota e azioni di validazione precoce dei prodotti e di dimostrazioni su larga scala" del PO FESR Sicilia 2014/2020, and Piano della Ricerca 2016-2018 linea di Intervento 2 of DMI, University of Catania. The authors would like to thank Regione Siciliana Assessorato dei Beni Culturali dell'Identit Siciliana - Dipartimento dei Beni Culturali e dell'Identit Siciliana and Polo regionale di Siracusa per i siti culturali - Galleria Regionale di Palazzo Bellomo.

References

[1] A. Bollo, L. Pozzolo, Analysis of visitor behaviour inside the museum: An empirical study, in: Arts Cultural Management, 2005.

[2] R. Cucchiara, A. Del Bimbo, Visions for augmented cultural heritage experience, IEEE MultiMedia 21 (1) (2014) 74–82.

[3] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella, Egocentric visitors localization in cultural sites, J. Comput. Cult. Herit. 12 (2) (2019) 11:1–11:19.

[4] F. Ragusa, L. Guarnera, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella, Localization of visitors for cultural sites management, in: International Joint Conference on e-Business and Telecommunications - Volume 2: ICETE, 2018, pp. 407–413.

[5] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella, Egocentric point of interest recognition in cultural sites, in: VISAPP, 2019.

[6] A. S. Razavian, O. Aghazadeh, J. Sullivan, S. Carlsson, Estimating attention in exhibitions using wearable cameras, ICPR (2014) 2691–2696.

[7] L. Seidenari, C. Baecchi, T. Uricchio, A. Ferracani, M. Bertini, A. D. Bimbo, Deep artwork detection and retrieval for automatic context-aware audio guides, TOMM 13 (3s) (2017) 35.

[8] D. Raptis, N. K. Tselios, N. M. Avouris, Context-based design of mobile applications for museums: a survey of existing practices, in: Mobile HCI, 2005.

[9] P. Koniusz, Y. Tas, H. Zhang, M. Harandi, F. Porikli, R. Zhang, Museum exhibit identification challenge for domain adaptation and beyond (2018). arXiv:1802.01093.

[10] R. D. Chiaro, A. Bagdanov, A. D. Bimbo, {NoisyArt}: A Dataset for Webly-supervised Artwork Recognition, in: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2019.

[11] D. Ahmetovic, C. Gleason, K. M. Kitani, H. Takagi, C. Asakawa, Navcog: Turn-by-turn smartphone navigation assistant for people with visual impairments or blindness, in: Web for All Conference, W4A '16, 2016, pp. 9:1–9:2.

[12] A. Kendall, M. Grimes, R. Cipolla, Posenet: A convolutional network for real-time 6-dof camera relocalization, in: ICCV, 2015, pp. 2938–2946.

[13] G. Taverriti, S. Lombini, L. Seidenari, M. Bertini, A. Del Bimbo, Real-time wearable computer vision system for improved museum experience, in: ACM Multimedia, 2016, pp. 703–704.

[14] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, CoRR abs/1804.02767. arXiv:1804.02767.

[15] A. Rubhasy, A. A. G. Y. Paramartha, I. Budi, Z. A. Hasibuan, Management and retrieval of cultural heritage multimedia collection using ontology, International Conference on Information Technology, Computer, and Electrical Engineering (2014) 255–259.

[16] P. Kwan, K. Kameyama, J. Gao, K. Toraichi, Content-based image retrieval of cultural heritage symbols by interaction of visual perspectives., IJPRAI 25 (2011) 643–673.

[17] D. K. Iakovidis, E. E. Kotsifakos, N. Pelekis, H. Karanikas, I. Kopanakis, T. Mavroudakis, Y. Theodoridis, Pattern-based retrieval of cultural heritage images, 2007.

[18] K. Makantasis, A. Doulamis, N. Doulamis, M. Ioannides, In the wild image retrieval and clustering for 3d cultural heritage landmarks reconstruction, Multimedia Tools and Applications 75 (7) (2016) 3593–3629.

[19] A. Furnari, S. Battiato, G. M. Farinella, Personal-location-based temporal segmentation of egocentric videos for lifelogging applications, Journal of Visual Communication and Image Representation 52 (2018) 1 – 12.