DigitCult | Scientific Journal on Digital Cultures

User Profiling and Context Understanding for Adaptive and Personalised Museum Experiences

Claudio Baecchi, Andrea Ferracani,
Alberto Del Bimbo
MICC
University of Florence
Viale Morgagni 65, Firenze, Italy

Abstract

In this article we present an integrated multimedia system for passive and active profiling of visitors in the Donatello room of the *Bargello Museum* in Florence. The system is composed by two Computer Vision powered modules: 1) the first, **MNEMOSYNE**, based on passive observation of visitors through cameras, builds a list of the artworks of interest for each visitor. These preferred artworks are then used to deliver personalised content and targeted recommendation of other items of interest on an interactive table exploiting user re-identification; 2) the latter, **SeeForMe**, is a wearable embedded system featuring an application which augments the functions of the well-known museum audio guides. The embedded system provides real-time artwork recognition on data obtained through a micro-camera exploiting a Convolutional Neural Network; furthermore it is smart, understanding the context and user behaviours such as walking, talking or being distracted and reacting consequently.

Published December 25th

Correspondence should be addressed to Andrea Ferracani, MICC - University of Florence, Viale Morgagni 65, FI - IT. Email: [name.surname@unifi.it]

DigitCult, Scientific Journal on Digital Cultures is an academic journal of international scope, peer-reviewed and open access, aiming to value international research and to present current debate on digital culture, technological innovation and social change. ISSN: 2531-5994. URL: http://www.digitcult.it

http://dx.doi.org/10.4399/97888255301482
DOI: 10.4399/97888255301482

 doi:10.4399/97888255301482

Introduction

Modern museums have the primary need to address the issue of an easy and targeted access to a massive amount of available information. To solve this problem a lot of researches has focused on providing personalized access to visitors through mobile and hand-held devices (Baber, Bristow, Cheng, Hedley, Kuriyama, Lien, Pollard, and Sorrell 2001; Bruns, Brombach, Zeidler, and Bimber 2007; Kuflik, Stock, Zancanaro, Gorfinkel, Jbara, Kats, Sheidin, and Kashtan 2011; Bay, Fasel, and Van Gool 2006). However, these devices can be intrusive to the museum experience as they require participation from the user and change the way the visitor behaves in the museum and with the artworks. On the other end it can be quite difficult to understand the interests of visitors without asking them to provide some information, both added manually or inferred from actions on specific devices.

User profiling for personalization has been addressed asking the user to input his interests both on the museum website and inside the museum (Wang, Stash, Sambeek, Schuurmans, Aroyo, Schreiber, and Gorgels 2009). Automatic detection of the displacement of the user has been exploited in the museum environment providing localized audio content delivery via a specific audio guide, and not requiring explicit input from the user (Hatala and Wakkary 2005).

Moreover, digital and mobile technologies can now enable visitors to have an enhanced experience, giving them a new mean of interacting with the museum and its contents. This is especially true for audio guides, where the content given to the visitor should be adapted to its interests and needs (Bowen and Filippini-Fantoni 2004). Researchers have also shown (Bowen and Filippini-Fantoni 2004; Karaman, Bagdanov, Landucci, D'Amico, Ferracani, Pezzatini, and Del Bimbo 2016; Wang, Stash, Sambeek, Schuurmans, Aroyo, Schreiber, and Gorgels 2009) that when personalizing on-site or off-site displays of artworks, it is necessary to understand the user behaviour such as what he is looking at and for how long.

In cultural heritage the personalization of the visitor experience may happen in form of virtual information that augments the user experience on some kind of device or in form of physical experience by means of real devices placed in the artwork proximity. In (Wang, Stash, Sambeek, Schuurmans, Aroyo, Schreiber, and Gorgels 2009) they present the Cultural Heritage Information Personalization (CHIP) system, a tour guide that creates a
personalized visit tour through a mobile device with RFID sensors that track the visitor in the museum.

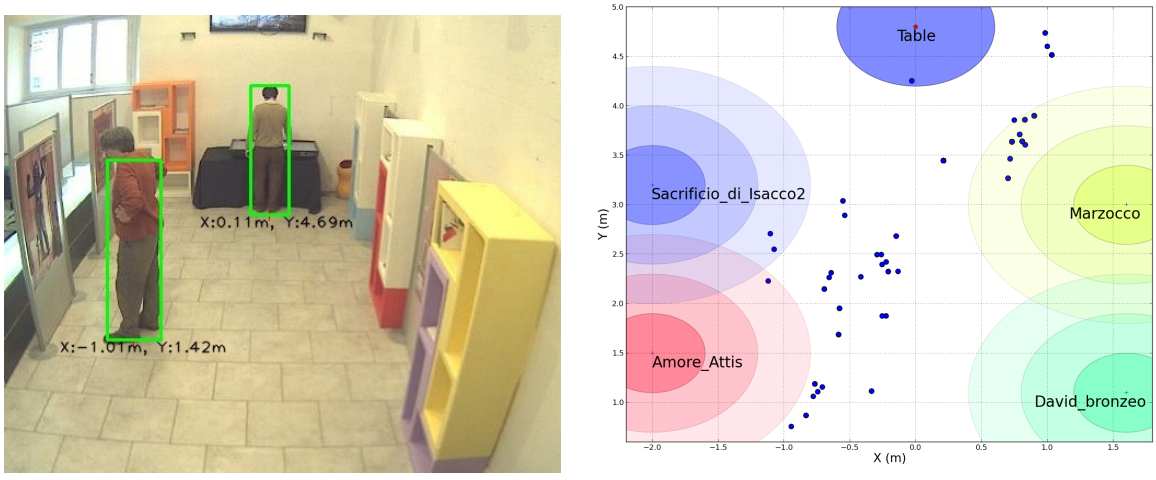
In this article we propose two multimedia systems, both experimented in the *Sala di Donatello* of the *Bargello Museum* in Florence that introduce new ways of exploiting digital technologies. The main goal of the two system is to minimize the cognitive effort and the actions required by the user to get information about artworks and, at the same time, to maximize the level of customization of contents on the basis of his behaviour and profile of interest.

In the two main sections we describe respectively the **MNEMOSYNE** system, whose main features are a passive profiling approach in order to track user behaviour and an interactive table-based interface for showing information, and **SeeForMe**, a prototype of an audio guide which aims to create a personalized on-site museum experience using a non-intrusive computer vision algorithm that can be executed on the web browser of any modern mobile device.

MNEMOSYNE

MNEMOSYNE (Bagdanov, Del Bimbo, Landucci, and Pernici 2012) is a three-year research project where we have studied techniques for passively observing museum visitors (Karaman and Bagdanov 2012) through cameras. Camera tracking is exploited in order to perform user profiling for personalizing multimedia content delivery on an interactive table placed near the museum exit, see Fig. 1. Furthermore, a mobile application allows visitors to download the content of interest of their visit to a smartphone or a tablet.

Figure 1. The tabletop in the *Sala di Donatello* of the *Bargello museum*



(a) (b)

Figure 2. (a) Example of frame with detections (b) Detection map with artwork sphere of influence for one visitor model.

 doi:10.4399/97888255301482

Passive Profiling of Museum Visitors

A passive visual profiling system of visitors moving in the *Bargello Museum* is exploited in order to build a profile of interest inferred from their behaviour. Fixed cameras are used to observe visitors walking inside the *Sala di Donatello* of the museum. A record of what each visitor has observed during his visit is acquired and analysed. In order to make the system work a pre-processing step is required to map the artistic content and the physical features of the museum hall.

Physical Museum Mapping

The Mnemosyne system exploits already-installed cameras in the museum. In order to understand the position of tracked people with respect to artworks given the perspective of the cameras each camera c is calibrated to a common ground plane. A simple tool allows an operator to estimate the homography $\mathcal{H}_c$ from each camera image plane to the ground plane with a few mouse clicks (Hartley and Zisserman 2004).

Given the homography $\mathcal{H}_c$, each artwork of interest can be easily localized by an operator on the ground plane by simply clicking in the camera view where the artwork stands on the ground. Each artwork is associated with a sphere of influence, defined as a bi-dimensional Gaussian of mean equal to the ground position of the artwork and variances in x and y dimensions as defined by the operator. These variances can depend both from the structure of the museum and the artwork dimensions.

Modeling the Visit of Museum Visitors

A pedestrian detector is run on the video stream corresponding to camera $c \in \mathcal{C}$, in order to obtain a set of N person bounding boxes (Bimbo, Lisanti, Masi, and Pernici 2010). The bounding boxes are described by visual, temporal and spatial descriptors.

The descriptor is defined as:

$$d_i = \left\{ \mathbf{d}_i^a, \mathbf{d}_i^s, d_i^t, d_i^c \right\}, \text{ for } i \in \{1, \ldots, N\}, \tag{1}$$

where $\mathbf{d}_i^a$ is an appearance descriptor consisting of RGB and HS color histograms computed on overlapping horizontal stripes and the HoG (Histogram of Oriented Gradients) descriptor (Dalal and Triggs 2005) as proposed for person re-identification in (Karaman, Lisanti, Bagdanov, and Del Bimbo 2013), $\mathbf{d}_i^s = (d_i^x, d_i^y)$ is the absolute position of the person detection on the ground plane, d_i^t is an integer timestamp, and d_i^c indicates the camera source from which the detection comes from c. All video streams are synchronized so that d_i^t and d_j^t can be compared.

The main objective in passive profiling is to cluster detections $D = \{d_i \mid i = 1 \ldots N\}$ in groups representing individual museum visitors. The adopted algorithm 1 is based on the computation of the distance between a model cluster m_j and the descriptors of a detection d_i with all its appearance and spatio-temporal features. Formally, the distance between a description d_i and the model m_j is computed as:

$$\begin{aligned} \text{dist}(m_j, d_i) &= (1 - \alpha - \beta) \times ||\mathbf{m}_j^a - \mathbf{d}_i^a||_2 \text{ (appearance)} & (2) \\ &+ \alpha \times \text{dist}_w(\mathbf{m}_j^s, \mathbf{d}_i^s, w_s) \quad \text{(spatial)} & (3) \\ &+ \beta \times \text{dist}_w(m_j^t, d_i^t, w_t) \quad \text{(temporal)} & (4) \end{aligned}$$

where $\text{dist}_w(x, y, w)$ is the windowed L2 distance:

$$\text{dist}_w(x, y, w) = \min\left(\frac{||x - y||_2}{w}, 1\right). \tag{5}$$

The parameters w_s and w_t are, respectively, the spatial and temporal window of the observations. The weights α and β defines the contribution of spatial and temporal distances, respectively, with respect to the overall distance calculation and are defined such that $\alpha, \beta \in [0, 1]$ and $\alpha + \beta < 1$. A detection belongs to a model if the distance to it is less than δ. A detection is represented as an accumulation of at least τ detections in a temporary model $\mathbf{m}_j^a$ computed as a running average, while the position and time information are those of the last matched detection. A model is active if it has the last associated detection time.

Data: D, δ, τ
Result: Detection associations
$M_a \leftarrow$ getActiveModels()
$M_{temp} \leftarrow$ getTmpModels()
for $d_i \in D$ **do**
  $\mathbf{dist} \leftarrow \{\mathrm{dist}(d_i, m_j), \forall m_j \in M_a\}$
  if $\min(\mathbf{dist}) \leq \delta$ **and** $M_a \neq \emptyset$ **then**
    $k \leftarrow \mathrm{argmin}(\mathbf{dist})$;
    $m_k.\mathrm{associate}(d_i)$;
  else
    $\mathbf{tmpDist} \leftarrow \{\mathrm{dist}(d_i, m_j), \forall m_j \in M_{temp}\}$
    if $\min(\mathbf{tmpDist}) \leq \delta$ **and** $M_{temp} \neq \emptyset$ **then**
      $k \leftarrow \mathrm{argmin}(\mathbf{tmpDist})$;
      $m_k.\mathrm{associate}(d_i)$;
      if $m_k.\mathrm{AssociationsCount} \geq \tau$ **then**
        $M_a = M_a + \{m_k\}$;
        $M_{temp} = M_{temp} \setminus \{m_k\}$;
      end
    else
      $M_{temp} = M_{temp} + \{d_i\}$;
    end
  end
end

Algorithm 1: Detection and association algorithm

User's Profile of Interest

The profile of interest of each visitor is built when the visitor is inside the interactive table area (see Fig. 2b). The profile consists in a report of which artworks have interested most the visitor. This information is based on the number of detections and the persistence of the visitor in the sphere of influence of each artwork. The level of interest for each artwork as well as contextual information are given on the interactive table.

Re-identification and Multimedia Personalization

The visit of the *Sala di Donatello* of the *Bargello museum* ends in an area near the exit equipped with an interactive tabletop display. When the user enters the area the system tries to re-identify the visitor associating to him the recorded profile of interest on the basis of his visual appearance. The interface is gesture-driven and allows visitors to explore and have in-depth and contextual information about the artworks that interested him most. The system features also a recommendation system, which further personalizes the profile of interest built passively during the visit. Favorite artworks and related multimedia content can be saved to a personal mobile device via a dedicated application.

Recommendation System

The MNEMOSYNE system exploits ontologies and semantic graphs as well as collaborative filtering to provide recommendations and deepenings of content to visitors on the interactive table.

Recommendation using ontologies Suggestions of multimedia and related content are provided using ontologies on a subset of 8 monitored artworks in the *Sala di Donatello* in the *National Museum of Bargello*. A Semantic Search Engine exploits the potential of the Semantic Web through an RDF (Resource Description Framework) ontology. The ontology models the artwork instances, but also related places, events, historical curiosities, other artworks in Florence and all over the world. Six entities are implemented: artist, artwork, category, museum/place and story. Artwork instances provide information about their creator, meaning, materials, techniques and the historical context and are complemented by a variety of multimedia content. Several

 doi:10.4399/97888255301482

(a) (b)

Figure 3. (a) Artwork level: are represented with title, thumbnail and a circular symbol showing the level of interest inferred for the current user during the visit by the passive profiling module. (b) Related resources level: information related thematically to the selected artwork.

thematic links to other artworks and stories are also provided. SPARQL queries are performed to get sub-graphs of data from the ontology, offering different views on the knowledge-base in terms of artwork content, related stories, additional resources related by tags or stories.

Recommendation based on Collaborative Filtering The Recommendation Engine (RE) uses two types of recommendation algorithms that give the system the ability to suggest to users items that they are more likely to find interesting. The RE implements metrics based on both user-similarity and item-similarity. The data model is preference-based. Preferences are stored as triples in a database table containing the following fields: the user ID, the item ID, and a value of preference. The preference is computed by the MNEMOSYNE passive visual profiling module . This value expresses the strength of the user preference for the item inferred by the system on the basis of the detections of the visitor in the hall and the built profile of interest. From this information we compute which users or items are more similar. Both similarity metrics, user-based and item-based, exploit the same components: the data model, a similarity metric, a notion of proximity (i.e. a neighborhood of users or items). The algorithm predicts values of preference according to the similarity metric. Euclidean distance is used for the computation of the similarity: a greater distance indicates a lower similarity.

The Natural Interface

When a visitor approaches the table, he is re-identified through a dedicated camera. The system triggers the interface to download the profile of interest which had been built by the passive profiling module during the visit.

The interface is touch driven and offers an augmented perspective of the museum through the exploitation of the profile of interest of the visitor, modeled by the computer vision system, and of the recommendation engine. Playing with the interface visitors can get deepenings and discover new, interesting information and resources, that can then be collected on their mobile phone for future inspection.

The main user interface consists of two levels of navigation: the *artworks* and the *related resources*.

The artworks level shows the artworks of the museum for which the visitor has shown the highest grade of interest (see Fig. 3a). This is done on the basis of the profile created by the passive profiling module. Each icon representing the artwork can be selected in order to get information about it.

The related resources level follows the selection of an artwork and allows the fruition of additional multimedia content organized in different domains: (*stories*, *secrets*, *recommendations*, *insights*) according to the relations to the artwork computed by the recommendation system or

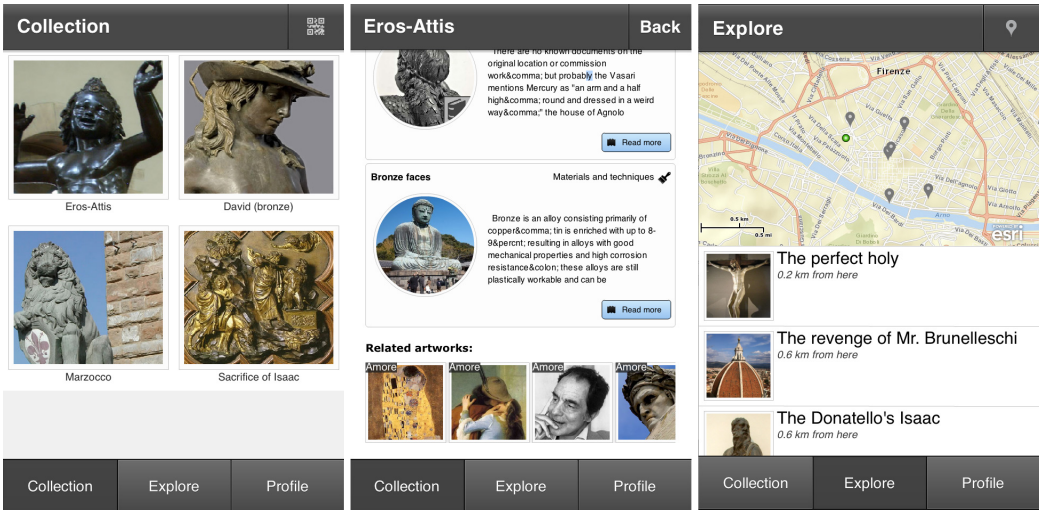


Figure 4. The mobile application. Left: user's favorite artworks; Center: in-depth information on the artwork; Right: map of suggested points of interest.

inferred through semantic expansion:

- **stories:** stories are directly related to the artwork in the ontology;
- **secrets:** resources related to the artwork and its related stories in the ontology according to the knowledge-based reasoning;
- **recommendations:** similar artworks suggested by the recommendation system using collaborative filtering;
- **insights:** resources related by tags or stories to the artwork according to the semantic knowledge-base.

The user can navigate these domains using a visual horizontal sub-menu at the bottom of the interface. Different layouts are used in each section: the diversity in background color, shape and arrangement of the interactive items maximize the difference between the visualized data and provide the user a consistent navigation experience. As an example, the *insights* section is illustrated in Fig. 3b.

The Mobile Application

The MNEMOSYNE mobile application, shown in Fig. 4, allows the user to collect the personalized digital content provided by the interactive tabletop interface. The mobile app is intended to be used at the end of the visit and not as an interactive device throughout the museum tour. The app can be installed on devices running iOS (iPhone, iPad) or Android.

A QR code present on the interface of the interactive table can be scanned using the app. The scanning allows the transfer of the identifier of the user to the mobile device. The application queries the MNEMOSYNE database to retrieve the user's profile of interest and the suggested resources, generated both through the passive profiling module and from the user interaction on the interface.

The mobile app extends the personalized user experience of the visit from an indoor to an outdoor scenario. In fact, suggested resources area also geolocalized and are visualized by the mobile interface on a map which provides routing from the user position. In this way the user can find on his phone other geo-localized resources, linked to his preferred artworks in the museum, that let him to continue the visit even once outside the museum.

 doi:10.4399/97888255301482

Testing MNEMOSYNE

The MNEMOSYNE system has been installed and beta-tested in the *National Museum of Bargello*. The installation has used four cameras, passively observing the visitor behaviour with respect to the artworks selected in the *Sala di Donatello*. The profiling system operated following the procedure described in the *Passive profiling of Museum visitors* section. Identity modeling has been performed performed independently in each camera stream using the *Detection and Association algorithm*. Profiles of interest are obtained by merging local profiles.

Results of the testing showed a computational bottleneck of the profiling task due to the heaviness of the detection process. Hence, we introduced a detector that learns only with weak supervision (the output of a rather slow pedestrian detector) where and at which scale detections usually appear (Bartoli, Lisanti, Karaman, Bagdanov, and Bimbo 2014). In this way candidate detection windows are evaluated only at scales and positions that are relevant for the image framed by each camera.

Usability Study

In the context of the installation of the MNEMOSYNE in the *Bargello Museum* we conducted a usability test in order to measure the effectiveness of the system. Twenty-four visitors were invited to use the system and then asked to complete a questionnaire. We used a standard questionnaire based on the System Usability Scale (SUS) (Bangor, Kortum, and Miller 2008). With the SUS the participants score the following ten questions with one of five responses ranging from “Strongly Agree" to “Strongly disagree":

1. I think that I would like to use this system frequently;
2. I found the system unnecessarily complex;
3. I thought the system was easy to use;
4. I think that I would need the support of a technical person to be able to use this;
5. I found the various functions in this system were well integrated;
6. I thought there was too much inconsistency in this system;
7. I would imagine that most people would learn to use this system very quickly;
8. I found the system very cumbersome to use;
9. I felt very confident using the system;
10. I needed to learn a lot of things before I could get going with this system.

The SUS scale gives an overall subjective assessment of the usability of a system. The questionnaire is given after the user has used the system in an unsupervised way, before any debriefing or discussion takes place. SUS scores have a range of 0 to 100.

Users’ age ranged between 20 and 55 years. 60% of users were male.

The usability evaluation results for the MNEMOSYNE system were encouraging. The average SUS score was around 73.5 which corresponds to a letter-grade of B-. A SUS score above a 68 is considered above average. The worst score was for question #5: users found functions to be poorly integrated. This is probably due to the fact that the users were not aware of being observed by cameras and of the existence of the passive profiling module. Best results were obtained for question #8, with users finding the interface intuitive.

SeeForMe: Wearable Audio Guide

Here we describe an web-based smart audio guide that adapts its behaviour to the actions and interests of the user while he walks through a museum (see Fig. 5). The application is able to understand the environment, that is what happens around the user, and what the visitor is looking

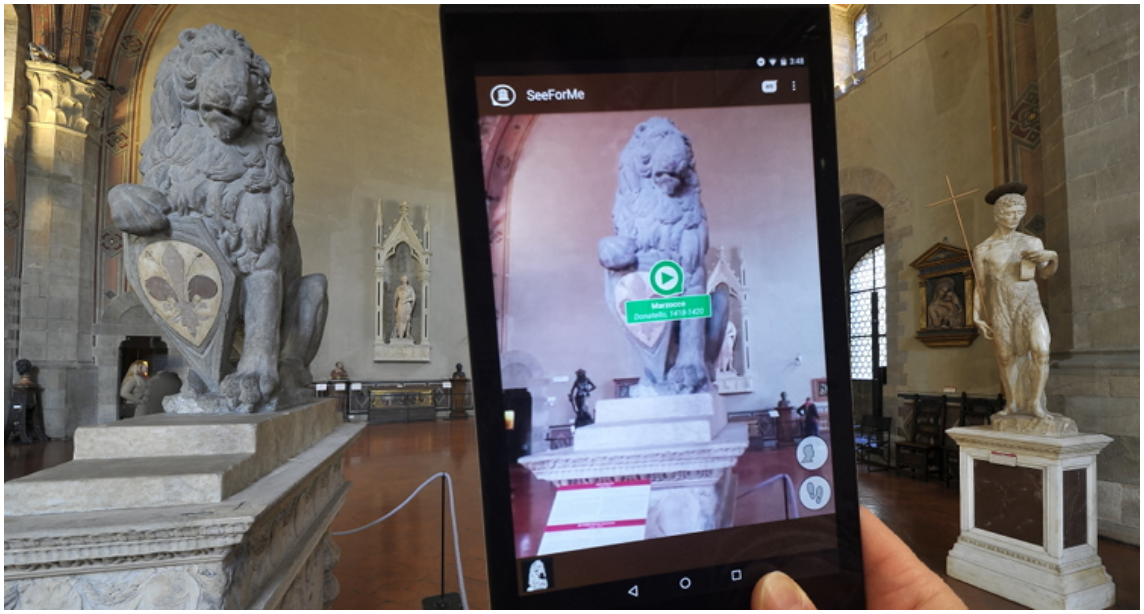


Figure 5. The SeeForMe app. The user frames one of the artwork inside the *Sala di Donatello* inside the *Bargello Museum*

at. Being a web application, the user does not need to install a specific application into the phone. In this work we aim at implementing a real-time computer vision system that runs inside a browser on a mobile phone, to perform artwork recognition. Image classification, together with motion sensors and audio voice activity detection, helps to understand what the visitor is looking at, if he is moving and whether he is talking with someone. This lets the application give a personalized experience to the user, but only when he is really interested in an artwork and he is not engaged in other activities. Moreover, leveraging which artwork the user has visited and for how long, the application can build his personalized interest profile. The app makes use of the camera device. Though it works and has been tested as a mobile app the app can also be used as a wearable system with the device put inside the pocket of a shirt and the camera facing forward.

The System

The system we propose comprises several components that work together to enable a smart experience. Fig. 6 shows the architectural diagram, which illustrates the main sub-modules of the system. Here we can identify two main sub-systems, one that recognized the artwork the user is looking at by matching the camera against a database, and the other that models the user status. These two modules provide appropriate input to the control module that in turn is responsible to automatically start and stop the audio description of the recognized artwork. Here we leverage three different sensors: the device camera, the microphone and the motion sensors. All three sensors are accessed through the browser by leveraging recently introduced multimedia APIs.

The camera output is used to understand what the user is looking at. The images are fed into a computer vision system that classifies (Image Classifier) each image into one of the artwork in the database if one of them is recognized (Artwork Recognition). This last module is comprised of two sub-modules: one that retrieves the artwork from the database (Artwork Retrieval) and the other that smooths out the predicted output at each video frame (Temporal Smoothing) by analysing how often and for how long an artwork persists in front of the visitor. Finally, the application (Context Modeling) leverages the audio taken from the microphone (Voice Activity Detection) and the data from the motion sensors (Walking Detection) to build up the user behaviour (User Status). This last output is used to trigger audio descriptions (Playback Control).

 doi:10.4399/97888255301482

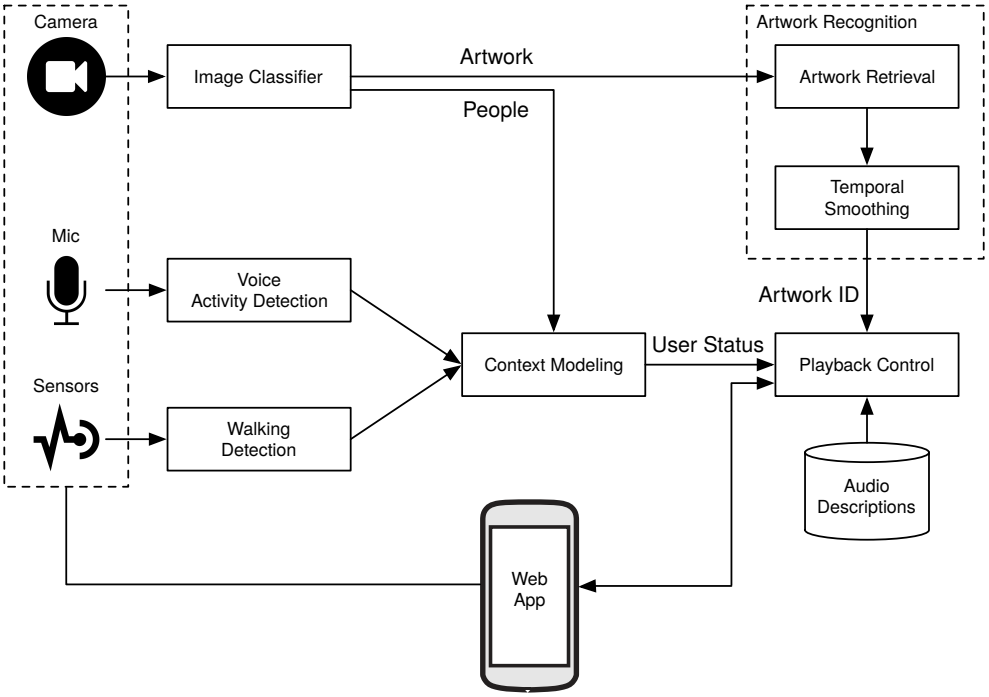


Figure 6. The overall system architecture.

Efficient Web-Based Image Classification

The smart audio guide we developed is based on an efficient computer vision pipeline that simultaneously performs artwork detection and recognition. The pipeline analyses the input camera frame by frame to detect whether an artwork is present or not, and in the former case it is able to tell which of the artworks in the database is the best candidate. Since we are dealing with a sequence of frames, it is also necessary to maintain some kind of temporal coherence between consecutive frames to provide a stable output. The system makes use of a small low-power device oriented convolutional neural network: MobileNet V2 (Sandler, Howard, Zhu, Zhmoginov, and Chen 2018). The network has been re-trained on a small set of artwork images that we wanted the system to recognize, so that each output class of the network corresponded to one of the artwork. To make it run in real-time on the browser we leveraged the Tensorflow JS framework [1] which is a browser port of the machine learning framework TensorFlow (Abadi, Agarwal, Barham, Brevdo, Chen, Citro, Corrado, Davis, Dean, Devin, Ghemawat, Goodfellow, Harp, Irving, Isard, Jia, Jozefowicz, Kaiser, Kudlur, Levenberg, Mané, Monga, Moore, Murray, Olah, Schuster, Shlens, Steiner, Sutskever, Talwar, Tucker, Vanhoucke, Vasudevan, Viégas, Vinyals, Warden, Wattenberg, Wicke, Yu, and Zheng 2015). The framework is able to convert and run models in the web browser making use of the GPU acceleration of the device, speeding up considerably the computation.

Temporal Smoothing

The idea behind the temporal smoothing module is that if there exists a continuity in the recognition output, then the predicted artwork is probably the one the user is looking at. For this reason we consider an artwork recognized only if it persists for at least M frames. It could also happen that the vision system is not always able to give the correct output for every frame. To provide a stable output we implemented a smoothing algorithm that keeps track of how many time an artwork is recognized. We keep a counter p for the most frequent output and we increment it every time the same artwork is recognized in two consecutive frames. If a different output is given we decrement p. We identify the artwork as correct only if $p > P > M$. This technique greatly reduces the number of false recognitions. Following our experiments we set the parameters as M = 15 and P = 20.

[1] https://www.tensorflow.org/js

Context Modeling

We wanted out audio guide to understand when it is the right time to provide information to the user and when it is not the case to disturb his visit. To accomplish this, it is necessary to first understand what happens in the environment surrounding the user. So on top of analysing the frames from the camera, trying to understand if and what artwork he is looking at, we also focus on understanding is moving and whether he is engaged in a conversation with someone else. If any of these is true, we choose not to disturb the user, keeping the application silent.

Audio Analysis for Speech Detection Our audio guide needs to understand whether the user is speaking with someone to convey the best possible experience. In such case, it is in fact desirable not to disturb the user with any information, not providing any audio or interrupting the current one if any of the audio description is playing. By doing this we let the user carry out the conversation, resuming the audio when its focus goes back to the artwork only. To achieve this, we exploit the audio coming from the device microphone and perform Voice Activity Detection (VAD) to understand if the input audio contains speech or not. This task output must of course take precedence over the visual artwork detection output so that the Playback Control Module can interrupt any currently playing audio even if the user is still facing an artwork. Voice Activity Detection is necessary since simple noise detection cannot be considered accurate. In fact, even if a museum is usually a quite place, it is possible that there exist situation where music is present or other people are talking nearby. We thus apply VAD to the audio stream, continuously listening to the environment for high volume voices. To do this we leveraged the work of (Eyben, Weninger, Squartini, and Schuller 2013), that is a state-of-the-art method based on a Long Short Term Memory recurrent neural network. The system models long range dependencies between the audio inputs and is also computationally trivial with respect to the vision module. The system has been integrated as a native web assembly module for the browser. The computational complexity for evaluating the networks is linear with respect to the number of input frames. Only a constant number of operations needs to be performed for every audio frame. The web assembly is obtained by recompiling the open source implementation available in the OpenSMILE framework 2 (Eyben, Wöllmer, and Schuller 2010). The tool is used to evaluate a whole second of audio with granularity of 0.01 before giving a prediction. The final value is then computed as the mean of the single outputs.

Walking Detection As for Speech Detection, we do not want to degrade the user experience by giving information in the wrong moment. There is a significant difference if the user walks or if he is standing still while looking at an artwork. We make the following reasoning: *i.* if the user is moving fast then he probably does not want to receive information about any of the artwork he may face. So even if the vision system detects an artwork, no audio description should be played; *ii.* if the user is listening to an audio description while standing still in front of an artwork, the audio should not be stop even if the vision system stops detecting the artwork. In fact, it can happen that some other visitor can occlude the visual of the user at any moment. If the user did not move, it means that he is still facing the same artwork and thus the audio should not be interrupted. We then perform walking detection by leveraging the device accelerometer by querying the browser APIs. We consider each peak in the accelerometer data as a step. We then take into account a sliding window of 1 second, and consider the subject walking if positive output is given for at least 5 seconds. We also leverage gyroscope data through the browser APIs to detect whether a person has changed or not his facing direction. To this extent we average the orientation vector over the same 1 second interval. The final facing direction is considered changed if the current orientation differs from the average for at least 45°.

Testing SeeForMe

The SeeForMe audioguide has been tested in the *National Museum of Bargello* in order to evaluate the user experience. To this end we used a questionnaire based on the System Usability Scale (SUS) as we did for the MNEMOSYNE system. 22 participants were enrolled in the test. Users' age ranged between 22 and 47 years. 45% of users were male. We asked them to carry out the task to obtain information of at least two artworks in the *Sala di Donatello*.

The results of the evaluation were good. The average SUS score was around 76 which

 doi:10.4399/97888255301482

corresponds to a letter-grade of B+. A SUS score above a 68 is considered above average. The worst score was for question #1: users didn't think users would use this system frequently. This is probably due to the fact that we asked the users to test the app holding the device with the hands and framing the artworks to perform recognition. Keeping hands up for a long time is tiresome for users. Best results were obtained for question #4, with users finding the app easy to use without requiring the help of technical staff.

Conclusions

In this paper we presented two system which have been tested in the *Sala di Donatello* of the *Bargello Museum* in Florence. The two systems share the intention to offer to the visitor non-invasive digital tools that nevertheless manage to provide a high degree of customization and personalization of the information. **MNEMOSYNE** track visitors in the hall and build a profile of interest of the artwork liked by the user on the basis of the time spent in the artwork area. Then it provides personalized content on an interactive table. **SeeForMe** is a new generation audio-guide: it is wearable but it can also provide visual augmented information on the camera; it recognizes artworks; it understands user behaviour and what's happening around him, adapting itself to the situation. The usability of both systems have been tested through SUS questionnaires with good results.

Acknowledgements

The **MNEMOSYNE** project was partially supported by Thales Italia and the MNEMOSYNE project (POR-FSE 2007-2013, A.IV-OB.2)

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Baber, C., H. Bristow, S.-L. Cheng, A. Hedley, Y. Kuriyama, M. Lien, J. Pollard, and P. Sorrell (2001). Augmenting museums and art galleries. In *Human-Computer Interaction. In: INTERACT*, Volume 1, pp. 439–447.

Bagdanov, A. D., A. Del Bimbo, L. Landucci, and F. Pernici (2012). Mnemosyne: Enhancing the museum experience through interactive media and visual profiling. In *Multimedia for Cultural Heritage*, pp. 39–50. Springer.

Bangor, A., P. T. Kortum, and J. T. Miller (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction 24*(6), 574–594.

Bartoli, F., G. Lisanti, S. Karaman, A. D. Bagdanov, and A. D. Bimbo (2014). Unsupervised scene adaptation for faster multi-scale pedestrian detection. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR 2014)*.

Bay, H., B. Fasel, and L. Van Gool (2006, May). Interactive museum guide: Fast and robust recognition of museum objects. In *Proceedings of the first international workshop on mobile vision*.

Bimbo, A. D., G. Lisanti, I. Masi, and F. Pernici (2010). Person detection using temporal and geometric context with a pan tilt zoom camera. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pp. 3886–3889. IEEE.

Bowen, J. P. and S. Filippini-Fantoni (2004). Personalization and the web from a museum perspective. In *Proc. of Museums and the Web (MW)*.

Bruns, E., B. Brombach, T. Zeidler, and O. Bimber (2007). Enabling mobile phones to support large-scale museum guidance. *MultiMedia, IEEE 14*(2), 16–25.

Dalal, N. and B. Triggs (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Volume 1, pp. 886–893. IEEE.

Eyben, F., F. Weninger, S. Squartini, and B. Schuller (2013). Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 483–487. IEEE.

Eyben, F., M. Wöllmer, and B. Schuller (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462. ACM.

Hartley, R. I. and A. Zisserman (2004). *Multiple View Geometry in Computer Vision* (Second ed.). Cambridge University Press, ISBN: 0521540518.

Hatala, M. and R. Wakkary (2005). User modeling and semantic technologies in support of a tangible interface. *Journal of User Modeling and User Adapted Interaction 15*(3-4), 339–380.

Karaman, S. and A. D. Bagdanov (2012). Identity inference: generalizing person re-identification scenarios. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pp. 443–452. Springer.

Karaman, S., A. D. Bagdanov, L. Landucci, G. D'Amico, A. Ferracani, D. Pezzatini, and A. Del Bimbo (2016). Personalized multimedia content delivery on an interactive table by passive observation of museum visitors. *Multimedia Tools and Applications 75*(7), 3787–3811.

 doi:10.4399/97888255301482

Karaman, S., G. Lisanti, A. D. Bagdanov, and A. Del Bimbo (2013). From re-identification to identity inference: labelling consistency by local similarity constraints. In *Person Re-identification*, Advances in Computer Vision and Pattern Recognition, pp. 287–307. Springer.

Kuflik, T., O. Stock, M. Zancanaro, A. Gorfinkel, S. Jbara, S. Kats, J. Sheidin, and N. Kashtan (2011). A visitor's guide in an active museum: Presentations, communications, and reflection. *Journal on Computing and Cultural Heritage (JOCCH) 3*(3), 11.

Sandler, M., A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520.

Wang, Y., N. Stash, R. Sambeek, Y. Schuurmans, L. Aroyo, G. Schreiber, and P. Gorgels (2009). Cultivating personalized museum tours online and on-site. *Interdisciplinary Science Reviews 34*(2-3), 2–3.