A Smart Assistant for Visual Recognition of Painted Scenes

Federico Concone[a,b], Roberto Giaconia[b], Giuseppe Lo Re[a,b] and Marco Morana[a,b]

[a] *University of Palermo, Department of Engineering, Viale delle Scienze, ed. 6, 90128, Palermo, Italy*
[b] *Smart Cities and Communities National Lab CINI - Consorzio Interuniversitario Nazionale per l'Informatica*

Abstract

Nowadays, smart devices allow people to easily interact with the surrounding environment thanks to existing communication infrastructures, i.e., 3G/4G/5G or WiFi. In the context of a smart museum, data shared by visitors can be used to provide innovative services aimed to improve their cultural experience. In this paper, we consider as case study the painted wooden ceiling of the Sala Magna of Palazzo Chiaramonte in Palermo, Italy and we present an intelligent system that visitors can use to automatically get a description of the scenes they are interested in by simply pointing their smartphones to them. As compared to traditional applications, this system completely eliminates the need for indoor positioning technologies, which are unfeasible in many scenarios as they can only be employed when museum items are physically distinguishable. Experimental analysis aimed to evaluate the performance of the system in terms of accuracy of the recognition process, and the obtained results show its effectiveness in a real-world application scenario.

Keywords

Machine Learning, Human-Computer Interaction (HCI), Cultural Heritage

1. Introduction

Smart personal devices, such as smartphones and tablets, have totally changed the way people live. In addition to the traditional calling and messaging capabilities, these devices come with heterogeneous sensors that allow users to collect and share information with the surrounding environment, thus paving the way to a new generation of applications that would not otherwise be possible [1, 2]. For instance, people with visual and hearing impairments may rely on specific services provided by their smartphones to move in public spaces, so helping them to live in a more independent way [3].

In this context, social and cultural inclusion represents a primary goal for many innovative IT systems. A smart museum, for example, is a suitable scenario for this type of solutions because a wide range of services can be provided to users in order to create a more inclusive and informative cultural experience, both physically and virtually. In an ideal smart museum, visitors should be able to get suggestions about the items to visit, as well as personalized descriptions of the works according to their individual knowledge (like a guided tour would do). Artificial Intelligence (AI) methods provide an invaluable help to realize these services, while also being non-invasive and ensuring the visitor's freedom to follow or not the

Joint Proceedings of the ACM IUI 2021 Workshops, April 13-17, 2021, College Station, USA

federico.concone@unipa.it (F. Concone); roberto.giaconia@gmail.com (R. Giaconia); giuseppe.lore@unipa.it (G. Lo Re); marco.morana@unipa.it (M. Morana)

suggestions provided. This last aspect is fundamental to any system, since it certainly makes the exposition more appealing and captivating to visitors.

For these reasons, modern museums have recently started a process of deep transformation by developing new interactive interfaces and public spaces to meet the challenges raised by the technological revolution of the last years. Moreover, it should not be ignored that in cultural sites a broad range of visitors can be found, from the youngest to the oldest. The younger generations are accustomed to the new technologies, while others might feel alienated in a smart environment that encourages them to use their personal devices. Hence, an inclusive smart museum should allow visitors to access all of its services, working towards making them easily accessible to everyone.

In this paper, we address a scenario in which users of a smart museum can exploit ad-hoc intelligent services aimed at improving their visit experience by providing personalized descriptions of the artworks of interest.

The way in which the specific contents for different users are selected [4, 5] is out the scope of this work, which instead focuses on the intelligent system responsible for supporting the visit thanks to the use of smart mobile devices. Our case study is the Sala Magna of Palazzo Chiaramonte (also known as Steri) in Palermo, Italy, which is characterized by a unique wooden ceiling from the 14th century containing a variety of paintings. Visitors can use our system to take a picture of any painting they are interested in and get the corresponding description, completely eliminating the need for indoor positioning technologies, which are unfeasible in our case study as they can only be used when museum items are physically separated.

Our solution exploits visual information of the various illustrated scenes and AI techniques. In particular, a Convolutional Neural Network (CNN) is employed to synthesize a set of features from a given picture in input, and a distance-based classification algorithm is used for the final inference. Such method normally has low accuracy on single images, so the proposed solution relies on the contextual classification of multiple images. This kind of expedient exploits the overall characteristics of the paintings and has not been researched by other works in the literature, since most image recognition algorithms usually try to recognize generic objects inside artworks.

The remainder of the paper is organized as follows: related work is outlined in Section 2. Section 3 introduces the case study, while Section 4 describes the proposed system as well as the algorithms behind the AI recognition modules. The experimental results are shown and discussed in Section 5. Finally, conclusions and future works follow in Section 6.

2. Related Work

Technologies for smart environments [6, 7], and smart museums in particular, have been deeply researched in recent years. As a result, a variety of solutions have been proposed to address specific challenges. In this paper, we focus on an intelligent IT system capable of supporting computer assisted guides. Early technologies employed in museums usually consisted of recorded descriptions, played by a small sound player through headphones, which required the visitors to follow a predetermined tour, or to manually input a code for every work. More recently, researchers suggested to replace these systems with applications directly usable through the users' smart devices. In this context, in order to make the apps easier to use, several works focused on the specific issue such as activity recognition [8] or indoor positioning, which allow to detect the user's movements and position within the museum so as to provide him/her with ad-hoc information (e.g., a description of the items in the nearby).

Indoor positioning systems typically exploit visitor's smart devices in order to interact with Bluetooth Low Energy (BLE) beacons, sensor networks [9], or WiFi infrastructure. Although these technologies are becoming increasingly accurate, energy efficient and affordable [10], there are some types of exhibitions in which different items are necessarily placed close to each other, so making the positioning system unable to identify the real interests of the user.

In this paper, we address this scenario and present a different approach that can be deployed in a variety of exhibitions, allowing smart touring where indoor positioning is not feasible.

The issue of uniquely referring to items placed close to each other is frequently addressed by means of Quick Response (QR) codes that identify each single work of art. The approach described in [11], for instance, exploits mobile phone apps and QR codes to enable *smart* visits of a museum. Once an item has been identified by means of its QR code, the visitors are provided with an augmented description, including text, images, sounds, and videos. This system, as well as many others in the literature [12, 13], is extremely easy to use, although the users generally prefer to recognize the artwork itself rather than QR codes or numbered codes, as discussed in [14, 15]. To this aim, various image processing algorithms and methods have been proposed for this kind of task, including Scale-Invariant Feature Transform (SIFT) and its faster but less accurate counterpart, Speeded Up Robust Features (SURF) [16]. For example, [17] describes a SIFT-based artwork recognition subsystem that operates in two steps: at first the images captured from the camera are pre-processed to remove blurred frames; then, the SIFT features are extracted and classified. Moreover, authors exploit the visitor's location in order to select the nearby artworks, so greatly reducing the computational effort of the matching process while at the same time increasing the recognition accuracy.

Despite having similar performances to SIFT, Convolutional Neural Networks are more suited for large scale Content Based Image Retrieval (CBIR), and they are generally faster to extract features from an input [18]. An enhanced museum guidance system based on mobile phones and on-device object recognition was proposed in [19]; here, given the limited performance of the devices [20], the recognition was performed by a single layer perceptron neural network. Such a solution can be used together with indoor positioning, but it is outdated, as smart devices are now capable of running larger neural networks, and CNNs should be preferred for large scale CBIR. In [21], the authors rely on the CNN to extract relevant features of a specific artwork, while the classification is treated as a regression problem. CNNs are particularly suitable to synthesize a small set of values (features) from a query image, which can then be compared to a database of examples in order to find images with similar contents. This enables fast image classification within a restricted dataset, even using a pre-trained network for inference.

More recent works are exploring the possibility of using CNNs in combination with other well-known techniques. For example, [22] discusses two novel hybrid recognition systems that combine Bags of Visual Word and CNNs in a cooperative and synergistic way in order to achieve better content recog-

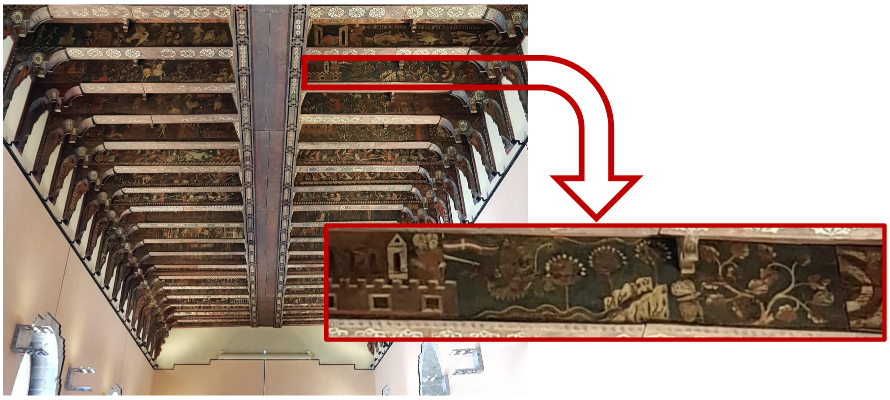


Figure 1: Part of the wooden ceiling of Palazzo Chiaramonte.

nition performances in a real museum.

3. Application Scenario

The system presented here has been designed with the aim of providing a non-invasive solution to enjoy Palazzo Chiaramonte in Palermo, Italy.

One of the worth noting place to visit in Palazzo Chiaramonte is the Sala Magna with its 14th-century wooden ceiling (see Fig. 1), which measures 23 × 8 meters and is composed of 24 beams, parallel to the short sides of the hall, perpendicularly divided by a long *fake* beam. On all sides of the beams and ceiling coffers, various scenes are illustrated, some telling mythological and hunting stories, others with plants and patterns.

The large number of visitors the Sala Magna receives every year, each of whom surely owns a smart device, and the many frescoes present in the hall make this place a perfect scenario to test intelligent applications aiming to improve the visitors' experience [23] during the tour. For example, information gathered from personal devices may be used to alert the visitors about the influx of crowds for a specific artwork, so allowing them to plan the tours according to their personal needs.

In the scenario addressed in this paper, visitors are interested in knowing the stories painted on the various parts of the ceiling, stories that are very close to each other and hardly distinguishable. Even with the assistance of the tour guides or other tools (such as QR codes), this characteristic makes it very difficult for the visitors to find the scene of interest, unless they manually count the beams. Our idea is to let visitors exploit their smart devices to locate a specific scene, select it, and obtain an augmented description, e.g., by means of 3D images, stories or videos telling of the scene. Moreover, once an item has been recognized, descriptions can be easily transferred to a "totem" located in the Sala Magna, that is a smart touchscreen enabling users to enjoy the scenes of interest in a more comfortable way. This is very important, for instance, to people that are not accustomed to prolonged interactions with small smart devices [14].

Such an alternative is also very useful for visiting groups, as it both allows individual touring on smartphones and tour guides' presentations at the multimedia totems.

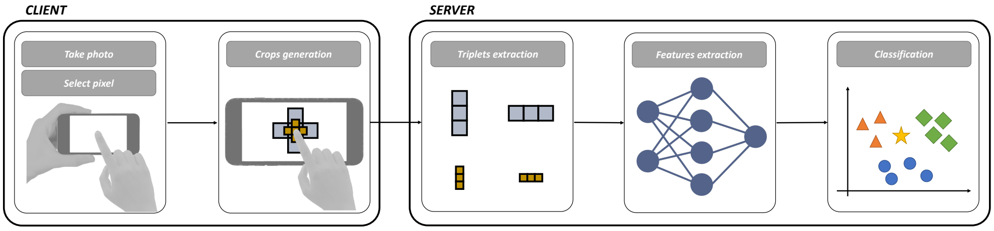


Figure 2: Frescoes classification procedure. The user points its device to a painting, touching the screen to precisely locate it. The client application creates a set of crops and sends it to the server. Crops are then classified.

4. System Overview

The recognition system is based on a client-server architecture in which two different kinds of clients are available. The first is the totem (i.e. the touch-screen monitor), where visitors can browse the artworks and their descriptions in a very clear way. The second client is a mobile app that runs on the visitors' personal smart devices, as well as other smartphones and tablets provided by the museum. In addition to the functionalities provided by the monitor, the mobile software application includes a scene recognition service that can be also used in combination with the other kind of client, i.e., sharing the identifier of the recognized scene with the touch-screen totem. This enables visitors to select an object of interest through the mobile devices and then show the results in the totem, thus enhancing their experience within the museum. This architecture is much lighter than three-level solutions based, for instance, on the fog paradigm [24, 25] and it is adequate for the purposes of the proposed application.

In the following of this section we present the main phases of the recognition procedure, summarized in Fig. 2.

4.1. Crops Generation

Visitors select a scene of interest by pointing a specific region (e.g., object/item) using the touchscreen of the smart device. The picture is then decomposed in two sets of crops of different resolutions, namely 512 x 512 and 336 x 336 pixels. Each set consists of five regions, $S = \{r_C, r_L, r_R, r_T, r_B\}$, the first *centered (C)* on the pixel chosen by the user, the others selected at the *left (L)*, *right (R)*, *top (T)*, and *bottom (B)* of the central one.

While the center crops might seem sufficient to classify the location selected by the visitor, using adjacent crops and different resolutions allows to increase the recognition performances, as it will be discussed in Section 5.2. The obtained crops are sent to the server for the last three steps, i.e. *triplets extraction*, *feature extraction*, and the *classification* tasks.

4.2. Triplets and Feature Extraction

In order to obtain a compact representation of input data, the server groups the crops into two horizontal/vertical triplets $t_1 = (r_L, r_C, r_R)$ and $t_2 = (r_T, r_C, r_B)$, and builds feature vectors by using a convolutional neural network. The adoption of CNN in our system is justified by the intrinsic nature of this category of neural networks, which are specialized in processing data that has a grid-like topology, such as an image. A CNN is typically made up of three different kinds of layers, named *convolutional*, *pooling*, and *fully-connected* layers [26].

The convolutional layer aims to synthesize the spatial relationship between pixels of the input image, without losing features which are critical for getting a good prediction. This layer uses a combination of linear and non-linear operations, i.e., convolution operation and activation function. The convolution is an element-wise product between the input image and a kernel, and its output is a feature map containing different characteristics of the input image. More are the kernels used during the analysis, more feature maps are generated. The feature maps are then evaluated by means of a nonlinear activation function, such as the sigmoid, hyperbolic, or rectified linear unit (ReLU) functions.

The pooling layer performs a downsampling operation with the aim to reduce the spatial dimensionality of the feature maps generated at the previous layer and, at the same time, to extract dominant features invariant in rotation and position. One of the most adopted operation at this stage is the *max pooling*, which applies a filter of size $n \times n$ to the feature maps, and extracts the maximum value for each of them.

Finally, the pooling layer output is transformed in a one-dimensional array and mapped by a subset of fully-connected layers to the final outputs of the network. Hence, these layers return class scores for classification and regression purposes, using the same principles of the traditional Multi-Layer Perceptron neural network (MLP).

If such a layer is not included into the neural network, then the CNN can be used to extract a set of features from the input image [27]. As our goal is only to extract the feature, the system leverages on a Convolutional Neural Network in which the last two dense layers and the soft-max function were discarded from the network. The underlying idea is to describe the content and shapes of the graphical content of the crops with a high level of abstraction, so that it is possible to make a comparison by only using the feature vectors.

To be more specific, the network model we adopted (refers to Fig. 3) is made of 13 convolution layers, with 3-by-3 kernel size filters, each one using a ReLU activation function. It also employs 5 pooling layers, specifically 2-by-2 max-pooling, to achieve downsampling. The original network [28] uses 3 dense layers and a final soft-max function, specifically targeted towards the 1000 classes of the ImageNet dataset; our network only uses the first dense layer, so the output of our network is a set of 4096 features.

4.3. Classification

The complete classification process consists in a two steps procedure. At first, the 4096-elements feature vectors obtained from each crop in the triplets are classified according to a minimum distance approach [29]. Generally, given a training set $X = \{x_1, x_2, \dots, x_m\}$ and the corresponding label set $Y = \{y_1, y_2, \dots, y_m\}$, a new point x^* of unknown class y^* is classified to $y_i \in Y$ if the distance measure $d(x^*, x_i)$, $x_i \in X$, is smaller than those to all other points into the training set:

$$y^* = y_i \text{ if } d(x^*, x_i) < d(x^*, x_j), \\ j \neq i, j = 1, \dots, m. \tag{1}$$

Here, each feature vector associated with the crop in t_j is classified as belonging to the class $y_i \in Y$ by using the Frobenius distance [30]; thus, the output of the first classification step

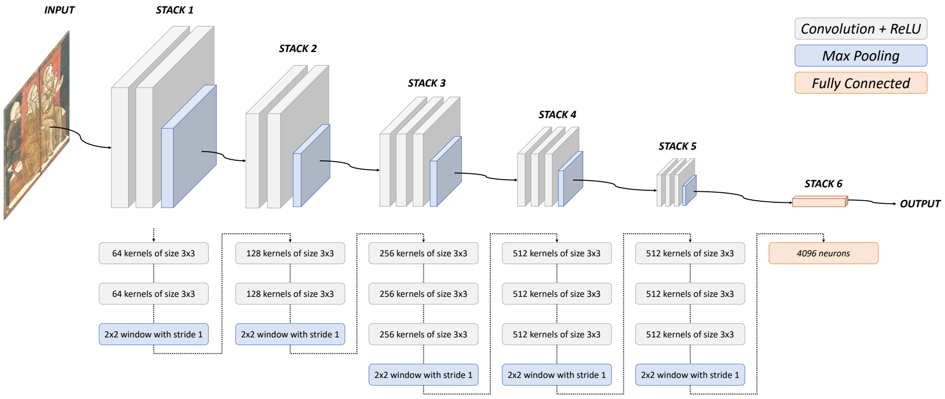


Figure 3: The CNN used for the feature extraction.

Table 1
Devices used for the experiments.

Model	Camera	CPU	RAM
Nokia 7 Plus	12 Mp + 13 Mp	Qualcomm Snapdragon 660	4 GB
Apple iPhone7	12 Mp + 12 Mp	Apple A10 Fusion	3 GB
Samsung Galaxy A50	25 Mp + 8 Mp + 5 Mp	Exynos 9610	6 GB
Samsung Galaxy Tab A10.1	8 Mp	Exynos 7904	3 GB

is represented by two new triplets T_1 and T_2 containing the predicted class for each crop.

The second phase aims to evaluate if crops in a triplet T_j were classified as depicting the same object. To accomplish such a task, we introduced the concepts of *strong confidence* and a *weak confidence* for classification. The first one is achieved when every element of the triplet is associated with the same object; the latter occurs when only two elements are associated with the same class.

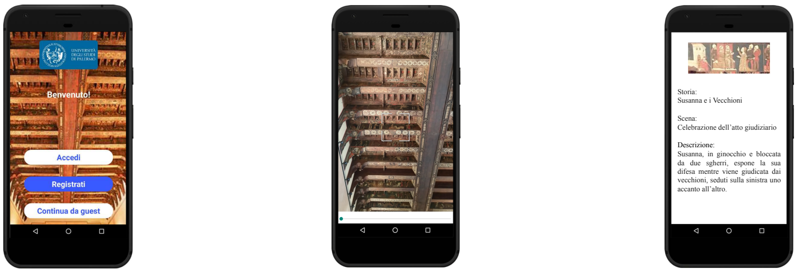
Firstly, the system checks for a *strong* confidence in any of the triplets and, if none is found, it tries for the *weak* confidence. If none of the triplets achieve *strong* or *weak* confidences, the visitor is asked to take a new picture of the artwork. This process is performed in near real-time, therefore it does not slow down the visiting experience, but rather improves the system precision.

5. Experimental Evaluation

The effectiveness of the proposed solution was evaluated through several experiments focused on the case study described in Section 3.

5.1. Experimental Setup

The experiments were carried out using three different models of smartphones, and one tablet (see Table 1) provided with the client software application. The mobile app (supporting both Android and iOS systems) supports visitors during all the recognition phases described so far, i.e., it allows to observe and select a scene in the wooden



Storia:
Susanna e i Vecchioni

Scena:
Celebrazione dell'atto giudiziario

Descrizione:
Susanna, in ginocchio e bloccata da due sgherri, espone la sua difesa mentre viene giudicata dai vecchioni, seduti sulla sinistra uno accanto all'altro.

Figure 4: Interface of the smartphone-side application. From left to right: the log-in screen, the camera screen, and an example page describing a scene.

ceiling, automatically extracts the crops from it and sends them to the classification server. Moreover, the software application is also able to manage information received by the server, thus enabling visitors to read the description of the scene in the device itself or on the touch-screen monitor.

Fig. 4 shows three examples of the smartphone-side application. The leftmost image represents the interface visitors can use to login to the service. The image in the center is the main interface of the application that allows visitors to zoom in/click on a detail of the scene of interest and stars the remote recognition process. Finally, the rightmost image shows information provided to users for the recognized element; in particular, in addition to the title and the description of the scene, the visitors are provided with high resolution pictures of the details of interest that are difficult to distinguish by standing at a great distance from the ceiling.

While the CNN is pre-trained on ImageNet, the dataset to train the classification algorithm and perform the experiments was captured using the devices listed in Table 1. Each class in the dataset corresponds to a part of the ceiling, captured in three pictures taken from different positions (Fig. 5), for each of which five regions of interest have been manually selected (Fig. 6). We considered 100 different relevant locations within the ceiling, thus the number of images obtained from each device is 1500.

The number of locations is calculated by taking into account the specific structure of our case study. The ceiling is made of 24 beams, each of which is divided in two parts by a central beam; on each side, we defined two locations: one for the side of the beam facing East, and one for that facing West (i.e. 4 locations for each beam). In addition to these 96 classes, the East and West walls of the hall also have paintings similar to the sides of the beams, so 4 further locations were considered.

Early experiments were conducted by randomly splitting such a dataset into training and testing sets; we will refer to this case as *mixed* dataset. Then, other tests were performed by dividing the dataset so that the training and testing sets contained images acquired from different cameras. This *separate* dataset is closer to the application scenario because every visitor will use devices with camera settings and characteristics that might differ from the ones used to train the system.

5.2. Recognition Results

The first set of experiments aimed to assess the system performance when considering

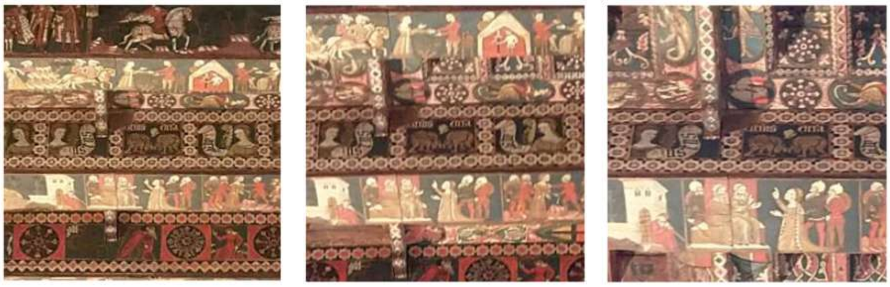


Figure 5: Images of the same scene at different distances.

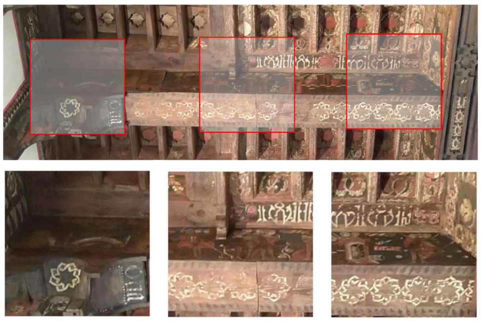


Figure 6: Example of manual cropping of the photo to create the dataset.

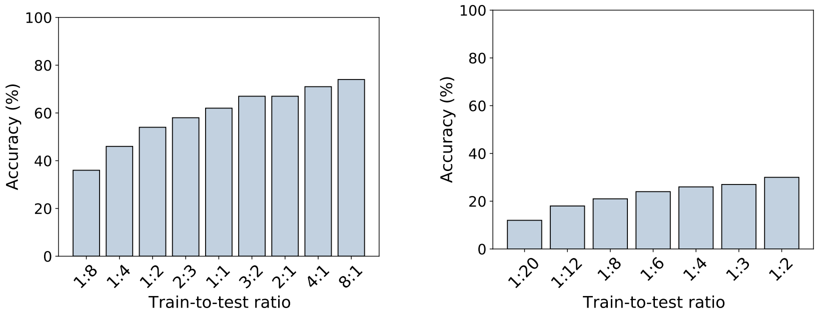
the recognition of a single (central) crop from the *mixed* dataset. The mean accuracy achieved in this case is shown in Fig 7-a, where each bar represents a different ratio of training and testing samples. Since our classifier has to be able to distinguish between 100 classes (scenes), results indicate that a larger number of training samples are required in order to obtain satisfactory accuracy values. In this case, images for both the training and the testing sets were captured by using the same devices, which is not representative of a real scenario that involves hundreds of testing devices equipped with different cameras.

For this reason, we also evaluated the single crop classification by using the *separate* dataset. Results from Fig 7-b show a similar trend as the train-to-test ratio increases, but also highlight a significantly lower mean accuracy, thus demonstrating the inadequacy of a single crop to drive the classification process.

The next set of experiments concerns the evaluation of the proposed three crops classification system, in which two different classification confidence settings are introduced, namely *weak* and *strong*.

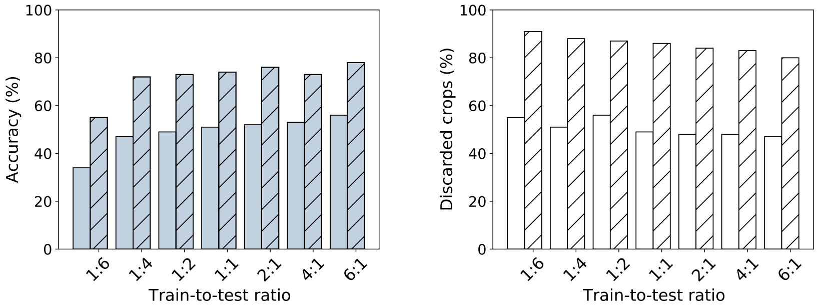
Performances were evaluated both in terms of accuracy and percentage of crop discarded because they have not reached a *weak* or *strong* classification confidence. It is worth noting that discarded images imply that visitors would be asked to take the photos again; thus, the lower this value, the higher the usability of the system.

Fig. 8 shows the result obtained on the *separate* dataset. By observing the mean accuracy values (Fig. 8-a) we can notice a significant improvement as compared to the single crop classification (Fig 7-b). Unfortunately, Fig. 8-b indicates that as the number of samples in the training set varies, the number of discarded crops is stably high. This is mainly due to not having enough images in the dataset. For this reason, the next set of experiments aimed to evaluate the impact of *data augmentation*. This technique



(a) (b)

Figure 7: Single crop classification accuracy on (a) mixed and (b) separate datasets.



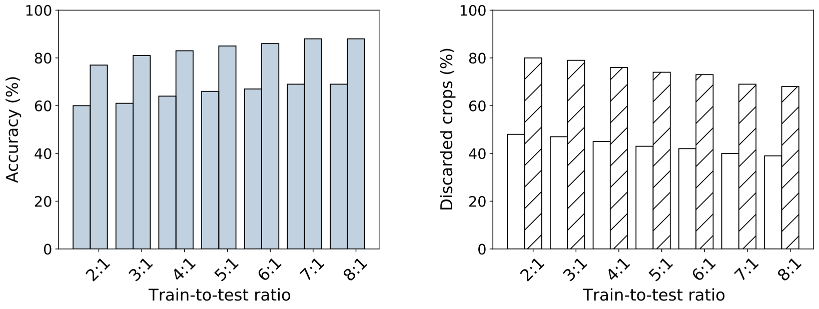
(a) (b)

Figure 8: (a) Accuracy and (b) percentage of discarded crops for weak (solid bars) and strong (hatched bars) classifications when using the *separate* dataset.

is used to artificially increase the number of samples in the training set in order to extract additional information [31]. In our system, *data augmentation* is performed by creating crops of different resolutions of the original locations of interest so as to obtain new samples.

Results in Fig. 9 show that data augmentation causes an increase in accuracy and a decrease in the number of discarded crops. Weak classification best results improved from 56% accuracy and 47% discarded queries, to 69% accuracy and 41% discarded queries respectively. The strong classification improved less, but still in a noticeable way: from 78% accuracy and 80% discarded queries, to 88% accuracy and 68% discarded queries. This confirms that data augmentation enhances the performances of the classifier without requiring the involvement of new capturing devices.

The last set of experiments was aimed to assess the classification procedure that will be actually performed by the smart museum application: instead of using only one triplet of crops, users' devices will send to the classification server all ten crops introduced in Section 4.1, divided in 4 horizontal/vertical



(a) (b)

Figure 9: (a) Accuracy and (b) percentage of discarded crops for weak (solid bars) and strong (hatched bars) classifications when using the data augmentation on the *separate* dataset.

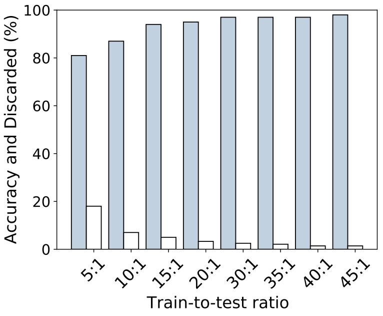


Figure 10: Accuracy (blue) and percentage of discarded crops (white) of the final classification procedure.

triplets. These are evaluated in terms of *strong* and *weak* confidence, and the input is discarded if neither is obtained. Results in Fig. 10 indicate that by applying this criterion, the number of discarded samples is significantly lower than in previous experiments; in particular, considering train-to-test ratios in the same range as Fig. 9, we can notice that only 10% – 20% are rejected while achieving an accuracy of 80% – 90%.

6. Conclusions

In this paper, we presented an intelligent system that exploits personal smart devices to support visitors during their tours within a museum. Differently from other works in the literature, the system completely eliminates the need for indoor positioning technologies, which might not be suitable for some kinds of expositions, such as our case study. In fact, the stories represented in the ceiling of the Sala Magna of Palazzo Steri are composed of a single artwork containing multiple scenes that cannot be physically distinguished, thus making indoor technologies unfeasible.

Thanks to the camera of their smart device, visitors are able to get the description for anything they are interested in by simply pointing the device and selecting a region of interest. Features are extracted and sent to a remote server in order to recognize the specific scene, and send back the related information. Moreover, the mobile app running on the client is capable of communicating with a touch-screen totem, so visitors can read and listen artworks' information in the way they prefer basing on their technology background.

As regards the recognition process, the

adoption of CNNs allows to extract features from 10 different regions of the photo taken by a visitor, taking advantage of the shape of the items in our specific scenario. Experimental results showed the performance of the recognition system in terms of accuracy and percentage of discarded crops, thus proving its effectiveness in a real-world application scenario.

The system will be soon deployed to support the visitors of Palazzo Chiaramonte. This will enable us to collect a greater number of query examples (captured from a wide range of devices), which could also be exploited to further refine the model.

7. Acknowledgments

This research is partially funded by the Project VASARI of Italian MIUR (PNR 2015-2020, DD MIUR n. 2511).

References

[1] V. Agate, F. Concone, P. Ferraro, Wip: Smart services for an augmented campus, in: 2018 IEEE International Conference on Smart Computing (SMARTCOMP), IEEE, Piscataway, NJ, USA, 2018, pp. 276–278. doi:10.1109/SMARTCOMP.2018.00056.

[2] S. Gaglio, G. Lo Re, M. Morana, C. Ruocco, Smart assistance for students and people living in a campus, in: 2019 IEEE International Conference on Smart Computing (SMARTCOMP), Piscataway, NJ, USA, 2019, pp. 132–137. doi:10.1109/SMARTCOMP.2019.00042.

[3] I. Apostolopoulos, N. Fallah, E. Folmer, K. E. Bekris, Integrated online localization and navigation for people with visual impairments using smart phones, ACM Trans. Interact. Intell. Syst. 3 (2014). doi:10.1145/2499669.

[4] V. Agate, P. Ferraro, S. Gaglio, G. Lo Re, M. Morana, Vasari project: a recommendation system for cultural heritage, Proc. of the 5th Italian Conference on ICT for Smart Cities And Communities (I-CiTies 2019) (2019) 1–3.

[5] A. De Paola, A. Giammanco, G. Lo Re, M. Morana, Vasari project: Blended recommendation for cultural heritage, Proc. of the 6th Italian Conference on ICT for Smart Cities And Communities (I-CiTies 2020) (2019) 1–3.

[6] A. De Paola, P. Ferraro, S. Gaglio, G. Lo Re, M. Morana, M. Ortolani, D. Peri, An ambient intelligence system for assisted living, in: 2017 AEIT International Annual Conference, volume 2017-January, 2017, pp. 1–6. doi:10.23919/AEIT.2017.8240559.

[7] A. De Paola, P. Ferraro, G. Lo Re, M. Morana, M. Ortolani, A fog-based hybrid intelligent system for energy saving in smart buildings, Journal of Ambient Intelligence and Humanized Computing 11 (2020) 2793–2807. doi:10.1007/s12652-019-01375-2.

[8] F. Concone, S. Gaglio, G. Lo Re, M. Morana, Smartphone data analysis for human activity recognition, Lecture Notes in Computer Science 10640 LNAI (2017) 58–71. doi:10.1007/978-3-319-70169-1_5.

[9] A. De Paola, A. Farruggia, S. Gaglio, G. Lo Re, M. Ortolani, Exploiting the human factor in a wsn-based system for ambient intelligence, in: Complex, Intelligent and Software Intensive Systems, 2009. CISIS '09. International Conference on, 2009, pp. 748–753. doi:10.1109/CISIS.2009.48.

[10] M. Majd, R. Safabakhsh, Impact of machine learning on improvement of

user experience in museums, in: 2017 Artificial Intelligence and Signal Processing Conference (AISP), IEEE, Piscataway, NJ, USA, 2017, pp. 195–200. doi:10.1109/AISP.2017.8324080.

[11] T. Octavia, A. Handojo, W. T. KUSUMA, T. C. YUNANTO, R. L. THIOSDOR, et al., Museum interactive edutainment using mobile phone and qr code, volume 15-17 June, 2019, pp. 815–819.

[12] M. S. Patil, M. S. Limbekar, M. A. Mane, M. N. Potnis, Smart guide–an approach to the smart museum using android, International Research Journal of Engineering and Technology 5 (2018).

[13] S. Ali, B. Koleva, B. Bedwell, S. Benford, Deepening visitor engagement with museum exhibits through handcrafted visual markers, in: Proceedings of the 2018 Designing Interactive Systems Conference, DIS '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 523–534. doi:10.1145/3196709.3196786.

[14] L. Wein, Visual recognition in museum guide apps: Do visitors want it?, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 635–638. doi:10.1145/2556288.2557270.

[15] M. K. Schultz, A case study on the appropriateness of using quick response (qr) codes in libraries and museums, Library & Information Science Research 35 (2013) 207 – 215. doi:https://doi.org/10.1016/j.lisr.2013.03.002.

[16] B. Ruf, E. Kokiopoulou, M. Detyniecki, Mobile museum guide based on fast sift recognition, in: M. Detyniecki, U. Leiner, A. Nürnberger (Eds.), Adaptive Multimedia Retrieval. Identifying, Summarizing, and Recommending Image and Music, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 170–183.

[17] S. Alletto, R. Cucchiara, G. Del Fiore, L. Mainetti, V. Mighali, L. Patrono, G. Serra, An indoor location-aware system for an iot-based smart museum, IEEE Internet of Things Journal 3 (2016) 244–253. doi:10.1109/JIOT.2015.2506258.

[18] V. D. Sachdeva, J. Baber, M. Bakhtyar, I. Ullah, W. Noor, A. Basit, Performance evaluation of sift and convolutional neural network for image retrieval, Performance Evaluation 8 (2017).

[19] P. Föckler, T. Zeidler, B. Brombach, E. Bruns, O. Bimber, Phoneguide: Museum guidance supported by on-device object recognition on mobile phones, in: Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia, MUM '05, Association for Computing Machinery, New York, NY, USA, 2005, p. 3–10. doi:10.1145/1149488.1149490.

[20] S. Gaglio, G. Lo Re, G. Martorella, D. Peri, Dc4cd: A platform for distributed computing on constrained devices, ACM Transactions on Embedded Computing Systems 17 (2017). doi:10.1145/3105923.

[21] G. Taverriti, S. Lombini, L. Seidenari, M. Bertini, A. Del Bimbo, Real-time wearable computer vision system for improved museum experience, in: Proceedings of the 24th ACM International Conference on Multimedia, MM '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 703–704. doi:10.1145/2964284.2973813.

[22] G. Ioannakis, L. Bampis, A. Koutsoudis, Exploiting artificial intelligence for digitally enriched museum visits, Journal of Cultural Heritage 42 (2020)

171 – 180. doi:https://doi.org/10.1016/j.culher.2019.07.019.

[23] G. Lo Re, M. Morana, M. Ortolani, Improving user experience via motion sensors in an ambient intelligence scenario, 2013, pp. 29–34.

[24] F. Concone, G. Lo Re, M. Morana, A fog-based application for human activity recognition using personal smart devices, ACM Transactions on Internet Technology 19 (2019). doi:10.1145/3266142.

[25] F. Concone, G. Lo Re, M. Morana, Smcp: a secure mobile crowdsensing protocol for fog-based applications, Human-centric Computing and Information Sciences 10 (2020). doi:10.1186/s13673-020-00232-y.

[26] R. Yamashita, M. Nishio, R. K. G. Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Insights into imaging 9 (2018) 611–629.

[27] T. Bluche, H. Ney, C. Kermorvant, Feature extraction with convolutional neural networks for handwritten word recognition, in: 2013 12th International Conference on Document Analysis and Recognition, IEEE, IEEE, Piscataway, NJ, USA, 2013, pp. 285–289.

[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[29] P. Kamavisdar, S. Saluja, S. Agrawal, A survey on image classification approaches and techniques, International Journal of Advanced Research in Computer and Communication Engineering 2 (2013) 1005–1009.

[30] G. H. Golub, et al., Cf vanloan, matrix computations, The Johns Hopkins (1996).

[31] A. Mikołajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 International Interdisciplinary PhD Workshop (IIPhDW), IEEE, Piscataway, NJ, USA, 2018, pp. 117–122. doi:10.1109/IIPHDW.2018.8388338.