





Article

Using Deep Learning for Collecting Data about Museum Visitor Behavior

Alessio Ferrato ¹, Carla Limongelli ¹, Mauro Mezzini ² and Giuseppe Sansonetti ^{1,*}

¹ Department of Engineering, Roma Tre University, 00146 Rome, Italy; ale.ferrato@stud.uniroma3.it (A.F.); limongel@dia.uniroma3.it (C.L.)

² Department of Education, Roma Tre University, 00185 Rome, Italy; mauro.mezzini@uniroma3.it

* Correspondence: gsansone@dia.uniroma3.it; Tel.: +39-06-5733-3220

Abstract: Nowadays, technology makes it possible to admire objects and artworks exhibited all over the world remotely. We have been able to appreciate this convenience even more in the last period, in which the pandemic has forced us into our homes for a long time. However, visiting art sites in person remains a truly unique experience. Even during on-site visits, technology can help make them much more satisfactory, by assisting visitors during the fruition of cultural and artistic resources. To this aim, it is necessary to monitor the active user for acquiring information about their behavior. We, therefore, need systems able to monitor and analyze visitor behavior. The literature proposes several techniques for the timing and tracking of museum visitors. In this article, we propose a novel approach to indoor tracking that can represent a promising and non-expensive solution for some of the critical issues that remain. In particular, the system we propose relies on low-cost equipment (i.e., simple badges and off-the-shelf RGB cameras) and harnesses one of the most recent deep neural networks (i.e., Faster R-CNN) for detecting specific objects in an image or a video sequence with high accuracy. An experimental evaluation performed in a real scenario, namely, the “Exhibition of Fake Art” at Roma Tre University, allowed us to test our system on site. The collected data has proven to be accurate and helpful for gathering insightful information on visitor behavior.



Citation: Ferrato, A.; Limongelli, C.; Mezzini, M.; Sansonetti, G. Using Deep Learning for Collecting Data about Museum Visitor Behavior. *Appl. Sci.* **2022**, *12*, 533. <https://doi.org/10.3390/app12020533>

Academic Editor: Antonio Fernández-Caballero

Received: 13 October 2021

Accepted: 30 December 2021

Published: 6 January 2022

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: cultural heritage fruition; human factors in artificial intelligence; museum visitors analysis; computer vision; machine learning; deep neural networks

1. Introduction

The fruition modalities of cultural heritage sites can benefit from advanced technologies and methodologies of data analysis that propose solutions aimed at visitor engagement. These proposals must cleverly balance the different needs of a large and diverse set of visitors and the peculiarities of the specific site. It is necessary to understand how visitors use the different spaces within a museum and how their behavior can help identify the strengths and weaknesses of the cultural offerings and, consequently, possible engagement strategies for museum institutions. An in-depth analysis of visitor behavior would stimulate new ways of promoting artworks. It would also serve as a spur for implementing more appropriate measures for museum security and visitor care. Moreover, collecting data about visitor behavior would allow museum curators and staff members to offer stakeholders a better settlement, both in displaying and narrating the artworks and in terms of marketing-related services. There are many studies on audience engagement. Some of them promote the integration of visitor tracking technology with mobile devices that users carry with them [1,2]. Other studies analyze user tracking to examine the flow of visits through complex and expensive tracking systems [3–8].

In this paper, we propose to collect visitor data through accurate, non-intrusive, cheap, and anonymity-preserving tools. The approach is based on computer vision techniques and leverages off-the-shelf RGB cameras and badges such as those provided free to attendees by event and conference organizers. Therefore, the overall cost of the entire instrumentation is

reduced, which is certainly a significant advantage over other state-of-the-art technologies. The methodology is based on deep learning, more specifically, methods and techniques for image detection and classification through Convolutional Neural Networks (CNNs) capable of providing excellent performance in terms of accuracy. Our approach can represent the solution to some of the criticalities shown by the other visitor localization technologies in the museum environment. With the proposed setting, the estimate of the visitor position is extremely accurate (on the order of 10^{-2} m). Moreover, the intrusiveness of the proposed approach is minimal. The user is not required to wear additional devices such as active and passive sensors, Personal Digital Assistants (PDAs), smartphones, or portable Graphics Processing Units (GPUs), but a simple badge. Another advantage of our solution is that the coverage is guaranteed at a low cost, through simple commercial RGB cameras or any video surveillance system present in most national and international museums and exhibitions. Lastly, the collected videos can be processed using a free platform such as the Google Colaboratory environment, thus allowing museum curators and staff to save the cost of hardware and system usage. The contributions of this paper are as follows:

- The analysis of the state of the art to identify the main problems in the timing and tracking of users in indoor environments (i.e., high intrusiveness, low accuracy, high cost, and high consumption);
- The design of a novel, low-cost, and highly accurate system to overcome the aforementioned problems;
- The development of a user timing and tracking system capable of providing data useful both for museum curators and staff (e.g., the possibility to analyze and monitor how visitors enjoy museum objects) and for the visitors themselves (e.g., the possibility to receive personalized suggestions during the visit).

We propose a deep learning-based approach to comprehensively and accurately collect visitor experience data. Specifically, we describe in detail the characteristics of its architecture and the experimental results obtained. We also illustrate a case study in a real environment and finally show how the collected data can be stored and used to provide valuable information relating to the behavior of each visitor.

This paper is structured as follows. In Section 2, we briefly review current technologies used to monitor visitors, focusing on computer vision approaches. In Section 3, we present the proposed system for detecting the exact visitor location anytime. In Section 4, we report the experimental results of the proposed system and the analysis of the data collected in the “Exhibition of Fake Art” at Roma Tre University. In Section 5, we discuss the obtained findings. Finally, in Section 6, we draw our conclusions and identify some of the possible uses of the data collected through the proposed system.

2. State of the Art

Nowadays, technology is increasingly exploited to improve users’ quality of life anywhere and anytime, when they use local transport services [9] or visit points of interest [10]. In particular, the possibility of providing museum curators and staff members with a system to track visitor behavior for improving the service offered is a widely investigated topic. In [11], the authors propose a computer vision algorithm based on Kinect and RGB-D camera. They track groups of visitors at the National Museum of Emerging Science and Innovation (Miraikan) in Tokyo, Japan, to identify the leader and study their dynamics. In [12], the authors present an IoT- (Internet of Things) based system to measure and understand visitor dynamics at the Galleria Borghese museum in Rome, Italy. A similar approach is described in [5], in which the authors report the results of a case study conducted at the CoBrA Museum of Modern Art in Amstelveen, the Netherlands. Tracking can be also used to understand how the flow of visitors inside a museum is oriented. In [13], the authors propose a method based on LIDAR to identify human beings and track their positions, body orientation, and movement trajectories in any public space. The system can accurately track the position of the visitor inside the museum. It has been tested at the Ohara Museum of Modern Arts in the Kurashiki area of Okayama Prefecture, Japan. In [3], Lanir et al.

propose a visual guide for museum curators and staff. Their system can show routes of interest and hotspots, and analyze visitor behavior. It has been tested at the Hecht Museum, University of Haifa, Israel. There are already numerous thorough and exhaustive works that review the main indoor localization technologies (e.g., see [14–19]). On the other hand, fewer works are focused on analyzing the technologies for the localization of museum visitors. The goal of our study is to propose a tracking system easy to use by museum curators and staff. Moreover, it has to capture as much information as possible about visitor behavior in real time. Finally, it should be ready to accommodate several further developments like visitors' micro-expressions recognition when looking at an artwork and subsequent recommendation. For this reason, we now analyze the hardware most commonly used for this aim by examining the strengths and weaknesses of the principal indoor localization technologies.

2.1. Indoor Localization Technologies

In the following, we report the most used technologies for indoor tracking.

- **WiFi.** This technology is extensively used for network connection of various devices in public and private environments. Initially, its maximum coverage was about 100 m, today it has been extended to over 1 km with the IEEE 802.11ah protocol, published in 2017, specifically designed for Internet of Things (IoT) services [20]. The fact that it is supported by almost all the electronic devices on the market makes it one of the most used technologies for indoor localization, without the need for additional infrastructure. However, its characteristics of wide coverage and high throughput yield to a more suitable usage for communication than localization, because of its low accuracy and interference, which make it necessary to use complex processing algorithms.
- **Bluetooth.** This technology is used for wireless connection between mobile and fixed devices within relatively small distances. The latest version, called Bluetooth Low Energy (BLE), provides improved performance in terms of coverage and throughput, with low power consumption [21]. Recently, two BLE-based protocols have been proposed: Eddystone (by Google Inc.) and iBeacons (by Apple Inc.). They are intended more for proximity-based services than localization, due to poor accuracy and high sensitivity to noise.
- **Infrared.** Among others, the IR technology was the first one to be widely used in many projects (e.g., see [8,22]). However, it has several limitations [23]. Firstly, it requires the presence of visible IR emitters and a line of sight between the emitter and receiver. Lastly, the nature of the IR signal requires accurate calibration of the parameters of the IR emitters and the active involvement of visitors in the process of locating their position.
- **Radio-Frequency Identification.** This technology is used to transfer data between a reader and a tag capable of communicating on default radio frequency [24]. There are two types of RFID systems: Active and passive. The first one operates with microwave and Ultra High Frequency (UHF) ranges, and it is characterized by low cost and ease of integration into the objects to be tracked. However, their low accuracy and poor integration in portable devices make them unsuitable for indoor location purposes. Passive RFID systems can operate without a battery but have significant limits in terms of coverage, which makes them unsuitable for indoor location purposes.
- **IEEE 802.15.4.** This technology is mostly used in wireless sensor networks and is characterized by good energy efficiency, low cost, but also by low throughput [25]. This standard is not available on most devices on the market and for this particular reason, it is not suitable for the indoor localization of users.
- **Ultra Wideband.** This technology is mainly used in short-distance communication systems and is characterized by low energy consumption [14]. The main characteristics of the UWB technology are the robustness to interference and the possibility to penetrate various materials. For these reasons, it is extremely suitable for indoor

localization. However, due to its limited implementation in portable devices, it cannot be widely used. The UWB problems have been extensively analyzed by the authors of [26] and in practical scenarios, the Non-Line-of-Sight (NLOS) propagation can be the main issue of this technology.

- **Visible Light.** Indoor localization technology based on visible light can be realized using different types of sensors. The most common are Light-Emitting Diodes (LEDs) [27]. The use of LEDs for indoor localization has numerous advantages over other technologies. First of all, emitters and sensors are very popular considering their low cost. They are also resistant to changes in humidity and they have low energy consumption. The main disadvantage of LEDs is that a line of sight between them is required [18]. Another type of sensor used in visible light systems is Light Detection and Ranging Localization (LIDAR). This sensor is able to provide information relating to the contour of surrounding objects. When combined with inertial sensors, LIDAR-based tracking systems can provide accurate results [28]. In order to properly work, the LIDAR-based tracking system needs at least one sensor in each room. Because of that, this particular technology would be extremely expensive for large museums.
- **Acoustic Signal.** This technology can localize the user by capturing acoustic signals emitted from sound sources using a microphone sensor [29]. The acoustic signal localization is accurate only when audible band acoustic signals (i.e., <20 kHz) are used. For these signals, sufficiently low transmission power is required not to cause unwanted noise. This aspect, coupled with the need for additional infrastructure, results in that localization based on acoustic signals is not widely used.
- **Ultrasound.** This technology allows us to compute the distance between a transmitter and a receiver by measuring the time of flight of ultrasonic signals (i.e., >20 kHz) [30]. Indoor localization based on ultrasound is very accurate. However, the measurement process can be heavily influenced by significant changes in temperature and humidity, as well as by ambient noise.

Hence, the solutions above are inaccurate, expensive, or very complex to implement. We, therefore, focus on computer vision, which can represent a non-intrusive solution for the user already accustomed to security cameras.

2.2. Red-Green-Blue (RGB) Video-Based Techniques

Several noteworthy approaches to user timing and tracking rely on RGB video-based models and methods. These techniques are already applied to other fields (e.g., see [31]) such as motion analysis, motion capture, and in general, to most activities related to virtual reality. Here, through RGB cameras, we can collect visual information that can be used to estimate where the visitor is. To achieve this goal, two capture methods can be used. The first one is based on visitor recognition, the second one relies on artwork recognition. The positioning of the camera, therefore, assumes a fundamental role in the implementation of these systems.

Recent work described in [32] takes advantage of computer vision and content-based image retrieval technique to detect visitor behavior. From frames recorded by multiple cameras installed in exhibition chambers, visitors are tracked by an object detector and also modeled with a deep learning technique. The system classifies each person by their appearance, grounded on color similarity as determined by measuring the distances of the distributions. Currently, the system is extremely time-consuming and needs to be enhanced to be applicable.

In [11], the authors propose a computer vision algorithm based on Kinect and RGB-D camera. They track visitor groups in a museum to identify the leader and study its dynamics. They also analyze the body language and the reciprocal position of the group leader to the rest of the group. The final goal of this study is the replacement of the group leader (typically, the guide in a museum) with a robot. They have installed four Kinect V1 sensors in some rooms at the National Museum of Emerging Science and Innovation (Miraikan) in Tokyo, Japan, and for two months they recorded videos of visitor groups

interacting with the museum guides during visits. The motion is detected by computing the difference between bounding boxes of two consecutive frames. The experimentation has been carried out by manually annotating and comparing the motion of the group and the guide with the algorithm results. The main issues with this approach are the inaccuracy of the results when people are too close to the camera and occlusion problems. Moreover, the categorization through bounding boxes has an average accuracy of 70–75%, which can improve with the application of the exponential motion algorithm they propose.

The Kinect sensor is proposed as a tracking solution in [33] as well. The authors use a particular process to estimate the gaze direction from face direction measurements. In their work, they discuss the method for gazed object estimations using face direction measurements and object detection. By measuring the face direction and detecting the object at the same time using a Kinect sensor, they can estimate what the visitors are looking at.

A different solution is SeeForMe [4]. It is a real-time computer vision system that can run on wearable devices to perform object classification and artwork recognition. It uses a video camera on the audio guide to identify artworks. This smart audio guide equipped with a vision system has been tested at the Bargello Museum in Florence. A CNN for object classification and identification runs on an NVIDIA portable GPU. Also, a voice detection module can determine the context (user alone, accompanied, etc.) and stop the guide when, for example, the visitor is talking to someone. Experimental trials were performed with a training set of 300 people and 300 images. Up to 5 m, there is maximum Precision and Recall (with Recall up to 0.8). Through various adjustments to the algorithm, they succeeded in having almost all works recognized, and only 22 of them were not recognized. The System Usability Scale (SUS) questionnaire, filled in by the sample, revealed only the problem of the intrusiveness of the guide during the visit, and the hassle of having to manage the menu. The SUS questionnaire showed good usability. Moreover, the camera must be necessarily placed in a shirt pocket, at chest height, which is a rather limiting constraint.

In [34], an approach in line with the spirit of our proposal is proposed: To collect as much data on user behavior as possible such as itineraries, the number of entries, the flow of visitors, and time spent in front of works. The authors use video cameras with infrared sensors and re-ID (person re-identification). The main difference with our approach is that, while the person re-identification needs a preprocessing phase of the generated videos, in our case the preprocessing is done on the badge before it is given to the user. In this way, we can monitor in real-time the movement of each visitor. This difference is significant because we can imagine using the extracted data also to propose new tools that support both the visitor (e.g., recommender systems [35,36]) and the museum curators and staff (e.g., visitor flow analysis [37]).

3. Proposed Method

Image classification and object identification technologies have become much more successful as a result of recent advancements in the field of deep learning [38]. More specifically, CNN models [39–41] can easily attain accuracy values near to 100% on the training set. In other terms, these models can give the correct prediction, with almost certainty, when they are asked to predict the class of an element of the training set. Thanks to this, it is possible to train such models to recognize an arbitrary, single object, with very high confidence. Based on the above observations, we developed the following idea for tracking museum visitors [42]. A CNN model is trained for recognizing a set of unique and distinct objects. The objects to be recognized are badges, like the ones used in events and conferences (see Figure 1). It should be clear that there will be a fixed number N of distinct badges. Therefore, our model will be trained in order to recognize N different classes: One for each of the N distinct badges. In the research literature, there exist two different types of object detectors [43]: Detectors of specific instances of objects and detectors of broad categories of objects. The former ones aim to detect instances of a particular object

(e.g., the Colosseum, Joe Biden's face, or the neighbor's cat), thus addressing a matching problem. The latter ones aim to detect instances of specific categories of objects (e.g., cars, humans, or cats). In our scenario, badge detection falls into the first type of object detection. Furthermore, the model is also trained for the detection (but not the recognition, for privacy reasons) of visitors' faces (see Figure 2). Therefore, face detection falls into the second type of object detection.



Figure 1. One of the badges used in the experimental trials.

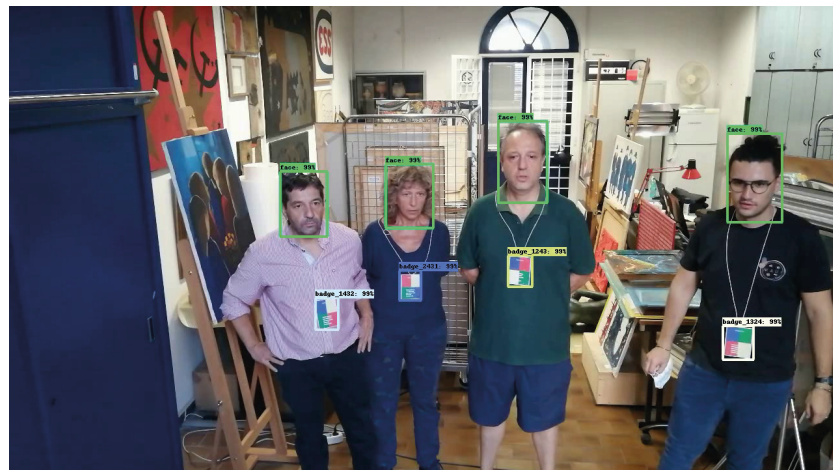


Figure 2. Recognized objects (i.e., badges and faces) in a frame of a 720p video. In this case, the camera is positioned above the artwork of interest, at a height of 2.20 m from the floor.

At the beginning of their visit, the visitor is invited to wear one of the badges on which a CNN model is trained. This model, as confirmed by the experimental results reported in Section 4.1, can recognize badges accurately. To this aim, it is required that RGB cameras are installed inside the museum environment. In our case, simple, inexpensive, off-the-shelf RGB cameras are sufficient (in our experimental tests, we used a Logitech HD Webcam CS25 camera and a smartphone Honor View 10 Lite camera). The frame rate of the videos captured by the Logitech camera is eight fps (frames per second) and 30 fps for those captured by the smartphone camera. It should be noted that the value of the frame rate does not affect the accuracy of the model. A higher frame rate increments the amount of collected data but has the drawback to increase the number of computational resources and storage needed. These cameras should be strategically placed inside the museum premises in a way that the badge worn by the active visitor is always visible by at least one camera. A simple assumption is to install one camera in every point of interest of the museum or on each side of every room, at a height that minimizes the possibility that another visitor put herself in front of the active visitor wearing the badge, thus making it not visible from the camera. Since the RGB cameras are inexpensive, the use of more cameras concerning the simple aforementioned assumption should not result in a substantial increase in the installation cost. Another, more sophisticated approach, to optimally position the cameras

inside the museum, is to resort to classical algorithms like those used for solving the Art Gallery Problem [44,45]. Once the recorded video is acquired by cameras, it is given in input to the model to detect the visitor's badge and face inside each video frame. The detection process consists of the following steps. For each video frame and for each object in the video frame, the model provides a score $0 < p \leq 1$, expressing the likelihood of an object being detected, the class c of the object, and a 4-dimensional vector containing the coordinates of the upper left and the lower right vertices of the box inside the video frame where the object c is detected. Therefore, for each video frame, the output of the detection process consists in a set of triples (c_i, p_i, \mathbf{b}_i) . Hereafter, vector \mathbf{b}_i will be referred to as the bounding box of the object of class c_i and we will denote the coordinates of the upper left and lower right corners of the bounding box by $(\mathbf{b}_i(x_1), \mathbf{b}_i(y_1))$ and $(\mathbf{b}_i(x_2), \mathbf{b}_i(y_2))$, respectively. If the value of p for a class c is higher than a prefixed threshold σ (we empirically set $\sigma = 0.8$ in our experimental tests), we assume that the object of class c is detected inside the video frame. The value of σ is a hyperparameter of the system. For high values of the σ parameter, we can have a high number of false negatives, whilst, for low values, we can have a high number of false positives.

3.1. Computation of the Exact Visitor Position

To compute the visitor's spatial position from the bounding box of the detected badge, it is first necessary to calibrate the camera or cameras used. Generally speaking, the procedure of camera calibration consists of the estimation-with acceptable accuracy for the specific application-of the extrinsic (i.e., rotation matrix and translation vector) and intrinsic parameters (i.e., image center, focal length, skew, and lens distortion) of the camera [46]. This process is fundamental for most computer vision applications, especially when metric information related to the scene is required, as is our case. Once the camera has been calibrated, it is possible to determine the angular amplitude α of each pixel of the camera [47]. This can be done with a simple computation consisting in counting the number m of pixels in a video frame (see Figure 3b) of a unit length yardstick put in front of the camera at a unit distance (see Figure 3a). Then, the angular amplitude α of a single pixel can be expressed as follows:

$$\alpha = \frac{2 \arctan(0.5)}{m}. \quad (1)$$

Knowing the angular amplitude of the pixel and the real dimensions of the badge (in our case they are $L = 10.4$ cm and $H = 14.0$ cm), it is straightforward to compute the distance ℓ of the badge from the camera, which can be done as follows (see Figure 4a):

$$\ell = \frac{H}{2 \tan\left(\frac{\alpha m_y}{2}\right)} \quad (2)$$

where $m_y = |\mathbf{b}(y_1) - \mathbf{b}(y_2)|$ is the number of pixels of the height of the badge bounding box in the video frame (see Figure 4b). In Equation (2), we can also replace the term m_y with the term $m_x = |\mathbf{b}(x_1) - \mathbf{b}(x_2)|$ and H with L . As above, we can compute the angle β (respectively, γ) that the badge forms with the vertical (respectively, horizontal) centerline of the video frame. Thus, the triple (ℓ, β, γ) corresponds to the polar coordinates of the badge in the camera reference. The visitor position inside the museum can be obtained by adding the values of the camera coordinates in the museum reference. Knowing the video frame rate, we can also determine the exact time and length of the museum visit and all other temporal information such as how much time a visitor spent in front of a specific artwork and so on.

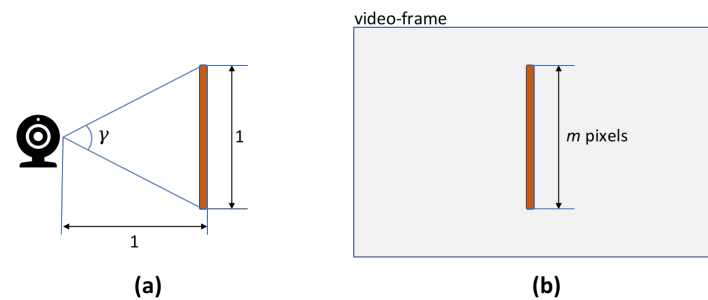


Figure 3. Computation of the angular amplitude of a pixel; (a) a unit length yardstick in front of the camera at a unit distance; (b) the corresponding number m of pixels in a video frame.

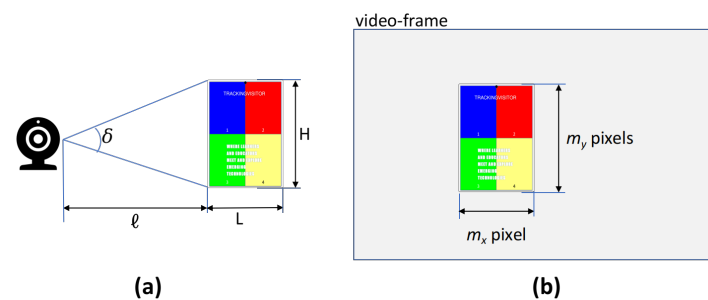


Figure 4. Computation of the distance between the badge and the camera; (a) a badge in front of the camera; (b) the corresponding number $m_x \times m_y$ of pixels in a video frame.

3.2. Experimental Settings

We evaluated several possible designs of the badge during the experimental trials. One requirement of the design is that the badge should be easily and effectively detected by the system. Another requirement is that each visitor who gets a badge should be easily distinguished from all other visitors wearing badges, in order to be able to track back the visit of a single visitor and distinguish their track from the track of any other visitor. This requirement could be easily satisfied when all the badges are distinct. In order to satisfy those requirements, we eventually chose a design in which a rectangular badge is split into four parts that can be in one of eight different colors (see the badge shown in Figure 1). Therefore, the number of different possible badges is equal to the number of dispositions of four colors taken from a set of eight, which is equal to $\frac{8!}{4!} = 1680$. If one splits the badge into six equal-sized parts, then the number of dispositions of six colors taken from a set of eight is $\frac{8!}{2!} = 20,160$. This shows that our design can be easily scaled for ten of thousands of different badges. For the sake of simplicity, we limited the experiments to the design of the badge shown in Figure 1. Furthermore, we trained our model to recognize 12 different badges. Thus, the model can recognize 12 different classes plus the face class. The training set was built first by manually annotating a dataset of about 300 pictures all containing the same badge. The training process with this single badge proved to be particularly efficient. This allowed us to automatically annotate all the other elements in the training set. The images of the training set were extracted from a set of 36 videos (three videos for each badge) from different angles, at 8 fps, 24 of them were about two minutes long, and the other 12, six minutes long, sampling a frame every two. In the first 24 videos, there was only one badge in each frame. In the last 12 videos, there were always two badges in each frame so that all possible pairs of badges were present in one frame. We inspected exhaustively all the automatically annotated images in order to assure the quality of the outcome. However, the accuracy of the model was so high (see Section 4) that very few manual corrections were needed to the automatic annotation process. In other terms, only about one of a thousand images required us to manually insert a missed annotation or delete a false positive annotation. Eventually, we produced a set composed of more

than 30,000 annotated pictures containing 12 different badges or, equivalently, about 2500 pictures for each different badge.

3.3. Model Implementation

For implementing the system, we used the Faster Region-based Convolutional Neural Network (Faster R-CNN) model [48]. The reason was that preliminary studies (e.g., see [38]) showed that the Faster R-CNN model is effective and accurate in relation to other popular deep learning frameworks. The architecture of the proposed system is shown in Figure 5.

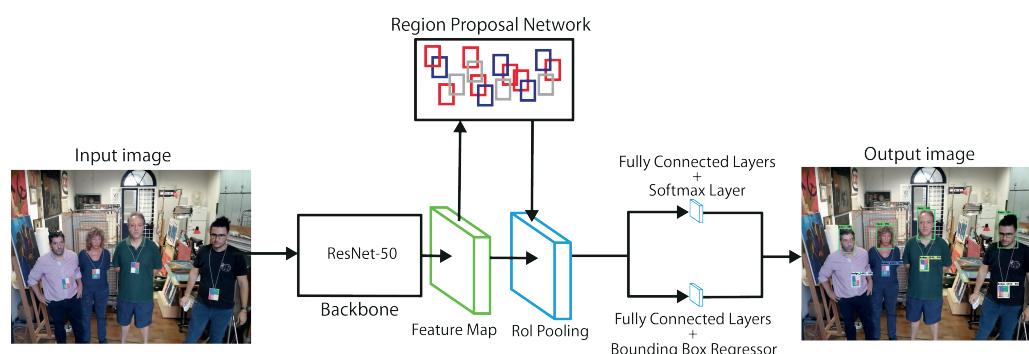


Figure 5. The architecture of the proposed system relies on a Faster Region-based Convolutional Neural Network (Faster R-CNN).

Specifically, the image taken by the RGB camera is given as an input to a backbone network, that is, a typical pre-trained convolutional network, which returns a feature map. We chose a 50-layer Residual Network (ResNet-50) as the backbone of our architecture because residual networks can usually achieve better performance than most other backbones [49]. The features are then sent in parallel to two different components of the Faster R-CNN architecture:

- A Region Proposal Network (RPN) that is used to determine the position of the image in which a potential object could be (i.e., at this stage we do not yet know what the object is, but only that there may be an object in a certain position of the image);
- A Region of Interest (RoI) pooling layer that is used to extract fixed size windows from the feature map before giving the RoI input to the fully connected layers. This component makes use of max pooling to convert the features within any valid RoI into a small feature map with a fixed spatial extension of height $H \times$ width W .

The output is then given as an input to two fully connected layers: One for the classification of the object and one for the prediction of the bounding box coordinates to obtain the final locations. The most important hyperparameters (e.g., the batch size and other optimization parameters) were left to the values suggested in [48]. Through grid search, we selected the best values for the learning rate and the scheduler that reduces the learning rate at a specific number of epochs. We chose Stochastic Gradient Descent (SGD) as an optimization algorithm. Another hyperparameter that we changed from the suggested value was the one that specifies the minimum dimension in pixels of the input image. Based on this parameter, the input image is resized in a way that at least one of its dimensions is equal to the parameter, before being fed to the CNN for the forward pass. The suggested value of the parameter was 800 pixels, but we increased it to 960 pixels. The reason was that when a visitor is far from the camera, the spatial dimension of the badge in the frame could be very little, making the detection difficult. In an ad hoc experiment, we analyzed all the dimensions of all the annotated boxes and we found that in no case was the dimension of any bounding box lower than 32×32 . The minimum dimension detected (of the order of 40×40 pixels) has been encountered for the badges when the visitor was approximately 3.5 m far from the camera. In our opinion, this allows for training a model in order to detect badges that are at five or more meters of distance from the camera. We later augmented the dataset by randomly shrinking each picture

by a factor between 0.3 and 0.5 (chosen randomly). This data augmentation enabled the model to detect badges located up to 6 m far from the camera, thus avoiding the need of adding other images to the training set. As a backbone, we employed a ResNet-50 network (where 50 is the number of convolutional layers in the network) that had already undergone two previous pretraining: One with the ImageNet [50] dataset and a second one, with the COCO [51] dataset containing about 90 classes. The authors of [52] strongly recommend the pretraining of the backbone on both datasets because empirical evidence shows that a network that had only been pretrained with the ImageNet dataset was much less accurate. Therefore, in order to verify if the ResNet-34 network (with only 34 convolutional layers) was faster but at the same time maintained, the same performance in terms of Accuracy and Precision, it was necessary to pretrain the ResNet-34 network with the COCO data set. Using the ResNet-34 network confirmed the boost in speed while maintaining an almost equal level of accuracy. Note that if Accuracy and Precision are the most important system performance metrics (instead of detection speed), the use of ResNet-100, or even ResNet-150, could improve system Precision. The pretrained model, as well as the pdf file with the trained badges, are available online (<https://colab.research.google.com/drive/1-Kr0c6dOuMUdoShJjbLhqaVtM9b-gwc6?usp=sharing> (accessed on 13 October 2021)).

4. Experimental Results

4.1. Performance Analysis

In order to assess the performance of the proposed system, we employed the detection evaluation metrics used in the most popular competitions, such as the COCO Detection Challenge (<https://competitions.codalab.org/competitions/20794> (accessed on 13 October 2021)). Before illustrating these metrics, however, it is necessary to introduce some fundamental concepts. The goal of an object detector is to predict the position of objects of a certain class in an image or a video with a high degree of confidence. For this purpose, object detectors place bounding boxes in the image to identify the positions of the detected objects. A detection can, therefore, be represented by three features: The class of the object, the bounding box that contains it, and the confidence score. The Confidence Score is defined as the probability that a bounding box contains an object. It is, hence, usually a value between 0 and 1 that expresses how confident the model is about the prediction [53]. Another fundamental concept is that of Intersection over Union (IoU), which is defined as the ratio between the area of the intersection between a predicted bounding box (B_p) and a ground-truth bounding box (B_{gt}) and the area of their union:

$$IoU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}. \quad (3)$$

We have a perfect match when $IoU = 1$, while if the bounding boxes do not overlap at all, we have $IoU = 0$. Therefore, IoU values near to 1 are significantly better. Confidence Score and Intersection over Union are used to evaluate a detection. Specifically, there is a True Positive (TP) when:

1. The confidence score is higher than a given threshold value;
2. The predicted class is the same as that of the ground-truth;
3. The predicted bounding box has an IoU higher than a threshold value (e.g., 0.75).

On the other hand, there is a False Positive (FP) if one of the last two conditions is not valid. In the event that multiple predictions match the same ground-truth, the one with the highest confidence score is considered a TP, whilst all the others are considered as false positives. We have a False Negative (FN) when the Confidence Score of a detection of a supposed ground-truth is lower than the threshold value, whilst we have a true negative (TN) when the Confidence Score of a detection of anything is lower than the threshold value. True negatives, however, are usually not taken into account in evaluating object detection algorithms. Based on the previous definitions, it is possible to define the *Precision* as follows:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

and *Recall* as follows:

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

By setting the threshold for the Confidence Score at different values, we can obtain different pairs of Precision-Recall, which can be plotted on a graph in the form of *Precision-Recall curves*. It is possible to summarize the shape of these curves through a single numerical value, known as *Average Precision (AP)* [54]. This value is defined as the Precision averaged over a set of eleven Recall values equally spaced $[0, 0.1, 0.2, \dots, 1]$:

$$AP = \frac{1}{11} \sum_{r \in \{0, 0.1, \dots, 1\}} p_{interp}(r) \quad (6)$$

The Precision value for each Recall level is interpolated considering the maximum Precision calculated for a system for which the corresponding Recall exceeds r :

$$p_{interp}(r) = \max_{\tilde{r}: \tilde{r} \geq r} p(\tilde{r}) \quad (7)$$

where $p(\tilde{r})$ is the Precision measured at Recall \tilde{r} . The purpose of the interpolation is to reduce the impact of wiggles in the Precision-Recall curves due to small variations in the classification of the retrieved objects. For a system to obtain a high value of Average Precision, it must therefore have a high Precision value at all levels of Recall. This penalizes systems capable of achieving high Precision only in retrieving a subset of objects. Normally, this particular curve is used to compare one system to another, but when it comes to performance analysis, it shows how a system is performing when its parameters are changed. As mentioned above, there is another type of curve known as *Recall-IoU curves*, which are the basis of another metric used to evaluate the performance of a detector, namely, the *Average Recall (AR)* [55]. Such curves are obtained by plotting the Recall values corresponding to the *IoU* values $\in [0.5, 1.0]$. The Average Recall is defined as the Recall averaged over all *IoU* values $\in [0.5, 1.0]$. It can be calculated as twice the area under the *Recall-IoU* curve:

$$AR = 2 \int_{0.5}^1 Recall(o) do \quad (8)$$

where o is *IoU* and $Recall(o)$ is the corresponding value of Recall. There exist several variants of the metrics above. Among the others,

- $AP@IoU=0.50:.5:.95$ is the *AP* value averaged over 10 different *IoU* threshold values (i.e., 0.50, 0.55, 0.60, ..., 0.95).

Furthermore, there is also Average Precision calculated for different object scales. So, we have:

- $AP@ \text{area} = \text{small}$, which represents *AP* for objects that cover an area less than 32^2 pixels;
- $AP@ \text{area} = \text{medium}$, which represents *AP* for objects that cover an area higher than 32^2 pixels but lower than 96^2 pixels;
- $AP@ \text{area} = \text{large}$, which represents *AP* for objects that cover an area higher than 96^2 pixels;
- $AP@ \text{area} = \text{all}$, which represents *AP* for objects of any size.

The area is given by the number of pixels present in the segmentation mask. Finally, we have *AP* calculated for different detection numbers per image, defined as follows:

- $AP@ \text{maxDets} = 1$, which represents *AP* given 1 detection per image;
- $AP@ \text{maxDets} = 10$, which represents *AP* given 10 detections per image;

- $AP@ \text{maxDets} = 100$, which represents AP given 100 detections per image.

The same variants of the Average Precision metric also apply to the Average Recall. Before presenting the experimental results, it is necessary to describe the test set used. To evaluate the performance of our system using the metrics introduced above, we randomly selected 300 images from 10 videos containing a total of 13 object classes (12 specific badges + the face object). Figure 6 shows the values of the Average Precision metric on the test set for our object detector as the number of epochs increases.

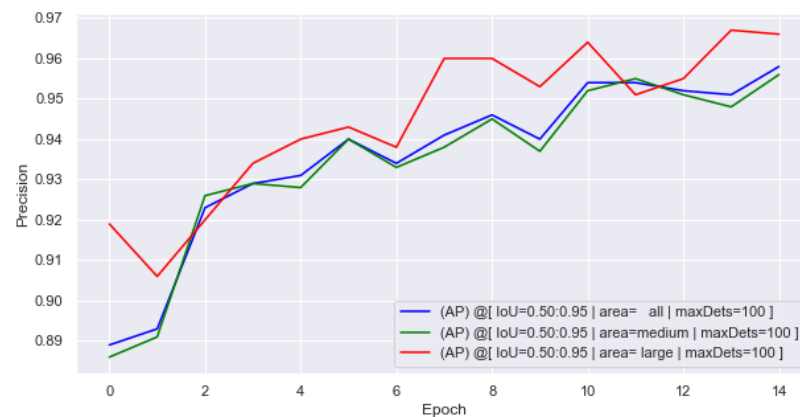


Figure 6. Average Precision of the proposed system on the test set.

It can be noted that there are high AP values already with a low number of epochs. We have only reported the value of $AP@ \text{maxDets} = 100$, as the values for $\text{maxDets} = 1$ and $\text{maxDets} = 10$ are the same as above. Furthermore, we have not reported the value of $AP@ \text{area} = \text{small}$, because we excluded a priori the detection of badges that are too small, that is, worn by visitors at such a distance from the point of interest that they cannot be considered in its surroundings. Figure 7 shows the trend of the Average Recall values on the test set as the number of epochs increases.

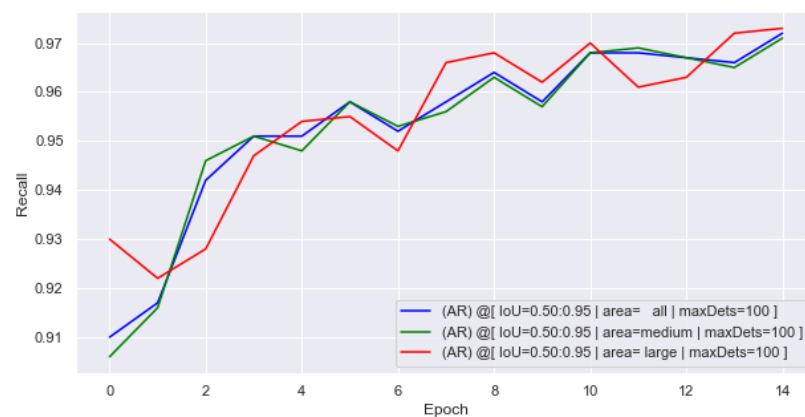


Figure 7. Average Recall of the proposed system on the test set.

Also in this case, the values are already high after a few epochs. The experimental analysis was performed on an NVIDIA QUADRO P2000 GPU capable of analyzing 4 frames per second. This system can, therefore, be used to perform a real-time analysis with fine tuning and optimization of the parameters.

4.2. Data Analysis

In this section, we report some of the analyses that can be carried out on the unfiltered data collected through the proposed object detection system. For this purpose,

we use the data collected in a real scenario, namely, the “Exhibition of Fake Art” (<https://www.facebook.com/indifesaadellabellenza/> (accessed on 13 October 2021)) of Roma Tre University. For each frame captured by the camera, the system provides the following information in output:

- The coordinates in pixels of the four corners of the bounding box that contains the object;
- The class of the recognized object;
- The confidence score of the detection.

From this data, the system can derive the center in pixels of the badge and its distance in meters from the camera (see Section 3.1). Obviously, it is possible to map the data in pixels to geometric coordinates and vice versa, only after camera calibration. Graphing this data not only makes its analysis more effective but also facilitates the use of information by the museum staff and all the operators in the field interested in making the museum data-driven. The system, therefore, not only allows information on the individual user or groups of users to be obtained but also provides the information needed to better manage the visitor flow in the various rooms [56]. The following graphs are taken from a video in which four visitors are present in the room. Specifically, the visitors are in front of the artwork and the camera is positioned above it at a height of 2.20 m (Figure 2 shows a frame of the video).

One of the possible analyses can be performed on the trajectories followed by visitors in the room. For example, in the scatterplot shown in Figure 8 it can be observed how the behavior followed by the green visitor (badge_3) differs from the other three, as the visitor tends to remain in the same position.

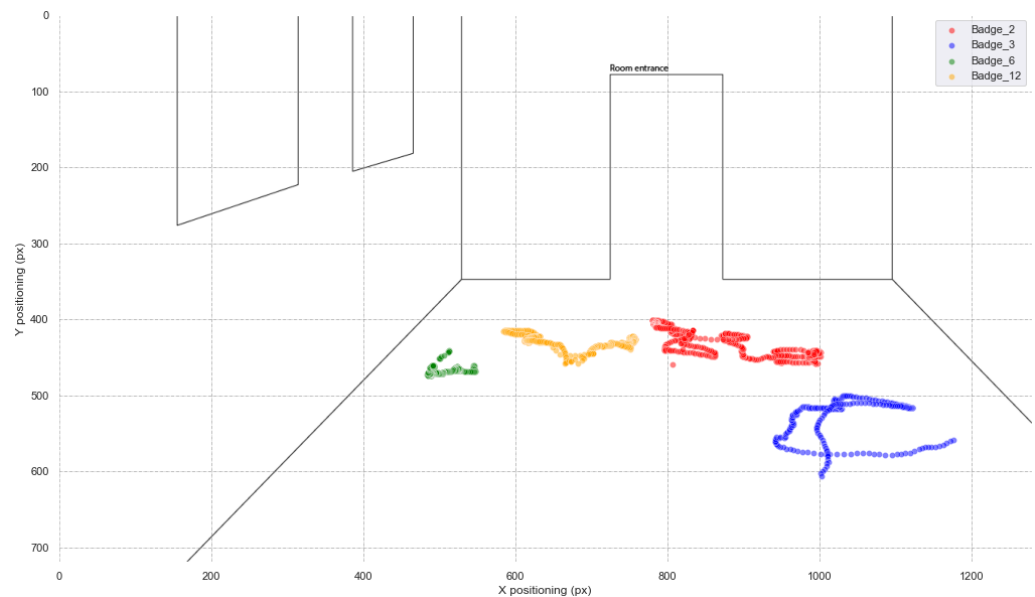


Figure 8. 2D scatterplot of four visitors in the sketched environment.

Figure 9 reports the processing output on the initial video frame.

The data from the monitoring of different environments could be easily integrated with each other to provide heatmaps. This analysis could also be useful for the museum staff to identify any problems in the fruition of the artworks, due, for example, to their arrangement or lighting.



Figure 9. 2D scatterplot of Figure 8 reported on one of the video frames.

From the video analysis, it is possible to easily obtain temporal information by knowing the number of frames per second captured by the camera. For example, from the graph shown in Figure 10, it is possible to obtain accurate and complete information relating both to the time spent by the visitor in front of a specific artwork and to their distance from it. The data collected confirm the differences in the behavior of the four monitored visitors. In particular, the green visitor slightly changes their position and remains in front of the artwork throughout the video, while the blue visitor is detected only from a certain instant of time and tends to change position more often to finally exit the framing of the camera.

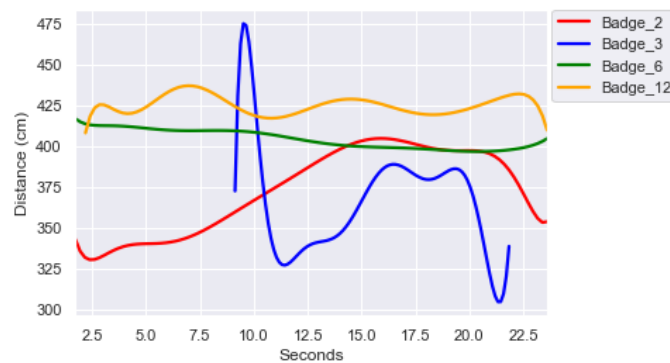


Figure 10. Badge-camera distance as a function of the time (related to four visitors) obtained through polynomial regression of order five.

The information collected can be further integrated with each other to generate 3D scatterplots like the one shown in Figure 11. The accuracy and completeness of the data are such that it can be supplied as input to graphic libraries such as Plotly’s Python graphing library or advanced tools like Blender to generate particularly expressive and informative 3D heatmaps.

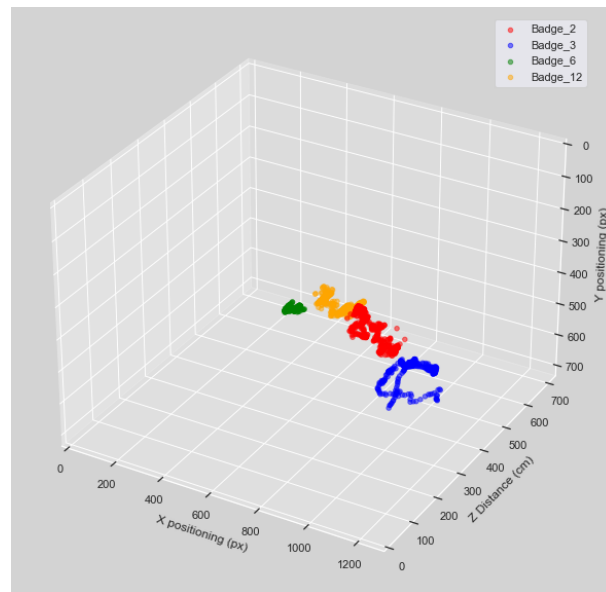


Figure 11. 3D scatterplot of four visitors.

4.3. Database with Collected Data

In order to make the analysis of data collection easy and at the same time effective, we propose the following database implementation and give some sample queries that could cover the most basic and useful needs when a museum staff member wants to extract useful information about visitor behavior from the database. The data collected through the proposed system can be stored in a data structure that supports spatial and temporal analyses of visitor behavior, such as those seen in Section 4.2. Let us suppose, for example, that we have m cameras and n badges. Each camera detects, at a generic timestamp, a badge at certain coordinates from the camera. We can store all those detections in a database composed of two tables. The first table, called *position*, has attributes (TIMESTAMP, CAMERA_ID, BADGE_ID, X, Y, Z) and the second table, called *camera*, has attributes (CAMERA_ID, CT, X, Y, Z). A single tuple (t, c_id, b_id, x, y, z) of *position* represents a detection at timestamp t from the camera c_id of the badge b_id at coordinates x, y, z with respect to camera c_id . A single tuple (c_id, ct, x, y, z) of *camera* represents the coordinates x, y, z of the camera c_id in relation to the museum. The value ct is the time period of a frame. If f is the frame rate of the camera, then we have $ct = 1/f$. For the sake of simplicity, hereafter, we suppose that ct assumes the same value for all cameras (i.e., $1/24$ s), but all the discussion can be extended with simple and minimal modifications to the general case, in which cameras can have different frame rates. We note that, whilst the table *position* is fed by the detections of the model, the table *camera* is determined and created in advance by the system supervisor. For instance, it can be convenient to create the view *dist_positions* using the SQL Listing A1, shown in Appendix A.

In order to build the track of a visitor wearing the badge b_id in the time lapses between timestamp t_0 and timestamp t_1 , that is, the ordered timestamp sequence of the visitor positions, we may execute the SQL Listing A2.

We also add to the database another table called *grid* with attributes (GRID_ID, X, Y, Z), in which for each tuple (g_id, x, y, z) , x and y represent the coordinates of the lower-left corner of a square inside the museum and z is the height of the floor (with respect to the museum) to which the square is referred. The width w of each square of the grid can be set, for example, to 0.5 m. Furthermore, we suppose that the badge is located somewhere between the ground floor, whose height is the coordinate z of the square and $z = 2.7$ m. In order to build a heatmap, that is, a visual indication that shows where the visitors spend more or less of the time in a given grid square inside the museum, we associate with each square element g of the grid a value that represents the sum of the number of seconds any visitor was present inside the square g in the time between t_0 and t_1 . Listing A3 returns

such values for all elements of the grid. Moreover, through Listing A4 we can detect how much time a person, identified by the badge b_id , stationed or passed in front of an artwork of the museum. We assume that the constants AX, AY, AW, and AH are given as parameters of the query and they represent what we consider as the space in front of the artwork and AZ the height of the floor (with respect to the museum) of this rectangle.

5. Discussion

In this paper, we reviewed some of the most authoritative and recent works proposed in the literature for indoor localization, focusing on those deployable in museum environments. As we saw in Section 2, each technology has pros and cons. Consequently, we have proposed a solution that requires simple badges and off-the-shelf RGB cameras and relies on deep learning techniques to monitor visitors and their behavior. The source code of the proposed system is available online (see Section 3). The main advantages of such a solution consist in the low cost of the instrumentation and the accuracy ensured by the detection and classification procedures based on the latest generation of Convolutional Neural Networks. As for the first point, the system leverages inexpensive badges and off-the-shelf cameras, which makes it economically viable. As for the second aspect, in Section 3, we have seen that the accuracy of our approach in estimating the visitor position can be pushed on the order of 10^{-2} m. In this regard, it should be noted that the operation of the Faster R-CNN, on which our system relies, does not depend on the number of objects to be recognized within the image. Therefore, the model accuracy is not affected if, in an image, there is only one badge or there are one hundred badges to be recognized. In our experimental trials, we limited ourselves to 12 badges because the SARS-CoV-2 restrictions did not allow us to test our system with more users. Anyway, the performance in terms of Average Precision and Average Recall remained unchanged when there were 12 badges to be recognized within the frame or when there was only one. What could instead be affected is system efficiency, if the number of region proposals in output from the Region Proposal Network significantly increases. We performed our experimental evaluation using an NVIDIA QUADRO P2000 GPU, which allowed us to process four frames per second even when the badges to be recognized were 12. However, it is reasonable to expect that if the badges to be recognized within an image become hundreds, more performing hardware solutions are needed (e.g., based on the use of several GPUs in parallel) if we want to preserve the real-time nature of the process.

However, the possible advantages are not limited to those mentioned above. First of all, the intrusiveness of the proposed approach is minimal. It is sufficient for the visitor to wear a simple badge like those provided free of charge by the organizers of events and conferences to be identified and tracked by the proposed technology. Therefore, no active involvement of the visitor is needed, nor are they required to bring additional devices such as active and passive sensors, PDAs, smartphones, or portable GPUs. As a result, our technology is not affected by power consumption issues. Another positive aspect of our solution is its coverage. It is sufficient that in each point of interest there is a commercial camera to make recognition possible. Moreover, the visitor timing and tracking system could also exploit visual data from any video surveillance systems present in most national and international museums and exhibitions. A further benefit of our solution is the possibility of integrating additional functionalities into it. As seen in Section 4.1, our system can efficiently capture other visitor aspects in addition to the badge worn. More specifically, the system can associate the visitor's face with their badge through a simple correlation (see Figure 12).

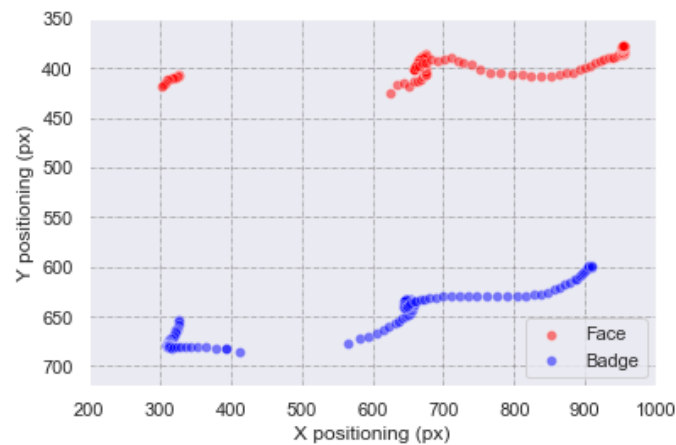


Figure 12. Correlation between face and badge positioning.

We emphasize once again that our system does not make the recognition of the visitor's face, but only its detection, for privacy reasons. In other terms, the mapping occurs only between face and badge, and not between face and visitor. The detected face can be analyzed to derive further information. It has been shown in the research literature (e.g., see [57,58]) how it is possible to analyze the user's micro facial expressions to infer information relating to emotions during the visit to predict their valence, arousal, and engagement. This information can be used to suggest objects [59] and personalized itineraries [60] based on these factors. For example, the exhibition could be organized by providing at its beginning the display of objects and artworks specially selected to derive the visitor's tastes without having to administer ad hoc questionnaires. We tested the system in a real scenario, that is, at the "Exhibition of Fake Art" at Roma Tre University. However, our experimental trials have been carried out with a low number of visitors due to SARS-CoV-2 restrictions. Generally speaking, occlusions can occur in overcrowded environments. Some noteworthy solutions have been proposed in the literature (e.g., see [61]). In our case, this problem can be mitigated by using several RGB cameras positioned in strategic positions, as shown in Figure 13, where the RGB camera is located at 4.20 m from the floor.

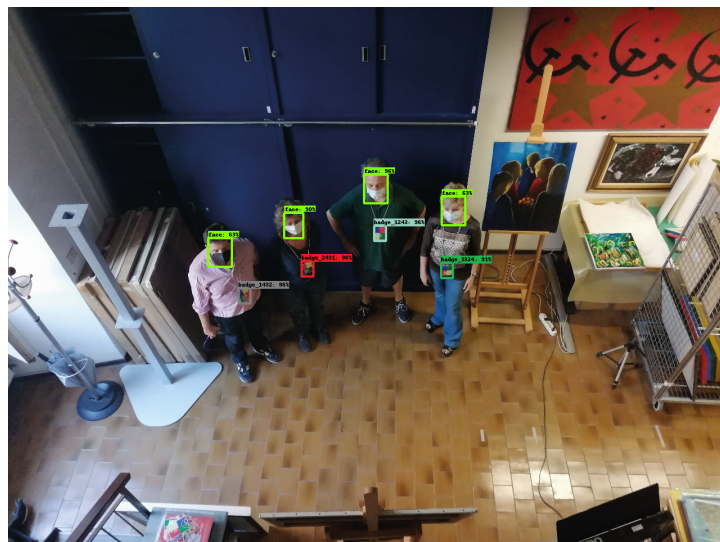


Figure 13. Another frame with the objects recognized by the model. It should be noted that, in this case, the camera is positioned higher than in the scene shown in Figure 2. It is now positioned at 4.20 m from the floor, but this does not affect the object detection and classification process.

Our system makes use of low-cost cameras, so the use of a large number of visual sensors would not involve a significant increase in costs.

6. Conclusions and Future Directions

Visiting museums and exhibitions around the world can indeed be an unforgettable experience. For over a century, studies have been published that show the possible relevance of their role in modern society and analyze the visitor behavior (e.g., see [62–66]). Current technology can make a decisive contribution in further improving the visitor experience, customizing it based on users' tastes and interests [2,67]. To achieve this goal, the first step is to automatically acquire information about the active user. This information can then be used for various purposes, among which:

- Provide visitors with personalized services such as recommendations of points of interest and additional textual and multimedia content [68];
- Analyze the individual and social behavior of visitors;
- Improve artwork arrangement;
- Optimize visitors' flow.

Therefore, in addition to testing our system in museums and exhibitions with a high number of visitors, we plan to concentrate our next research efforts on the design and the realization of tools that can derive the maximum benefit from the data collected through the system proposed herein.

To conclude, in this paper, we presented a deep learning-based approach to collect data regarding the visitor's experience in an accurate and comprehensive way. The solution we propose makes use of low-cost equipment (i.e., off-the-shelf RGB cameras) and requires the visitor to wear a simple badge, thus being non-intrusive. We do hope that our research efforts will contribute to making the museum visiting experience even more enjoyable, thus persuading more and more people to leave the comfort of their homes and experience cultural heritage on site.

Author Contributions: Conceptualization, A.F., C.L., M.M. and G.S.; investigation, A.F., C.L., M.M. and G.S.; methodology, A.F., C.L., M.M. and G.S.; software, A.F., C.L., M.M. and G.S.; validation, A.F., C.L., M.M. and G.S.; writing—original draft, A.F., C.L., M.M. and G.S.; writing—review and editing, A.F., C.L., M.M. and G.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to sincerely thank Professor Giuliana Calcani and her colleagues from the Department of Humanities of Roma Tre University for allowing us to experience the system proposed herein in a real scenario, namely, the "Exhibition of Fake Art" at Roma Tre University.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. SQL Queries

This appendix contains all the queries described in Section 4.3.

Listing A1: View creation.

```

1 CREATE VIEW dist_positions AS
2 SELECT DISTINCT P.TIMESTAMP, P.BADGE_ID, C.CT
3 /* Changing the reference system */
4 P.X + C.X AS X,
5 P.Y + C.Y AS Y,
6 P.Z + C.Z AS Z
7 FROM positions P, camera C
8 WHERE P.CAMERA_ID = C.CAMERA_ID

```


Listing A2: Badge positions tracking.

```

1 SELECT TIMESTMP, BADGE_ID, X, Y, Z
2 FROM dist_positions
3 WHERE BADGE_ID = bid AND
4 P.TIMESTMP BETWEEN t_0 AND t_1
5 ORDER BY TIMESTMP

```

Listing A3: Heatmap.

```

1 SELECT G.GRID_ID, SUM(P.CT)
2 FROM grid G LEFT JOIN dist_positions P ON
3 WHERE P.TIMESTMP BETWEEN t_0 AND t_1 AND
4 P.X BETWEEN G.X AND G.X + 0.499 AND
5 P.Y BETWEEN G.Y AND G.Y + 0.499 AND
6 P.Z BETWEEN G.Z AND G.Z + 2.70
7 GROUP BY G.GRID_ID

```

Listing A4: Badge time tracking.

```

1 SELECT SUM(CT)
2 FROM dist_positions
3 WHERE TIMESTMP BETWEEN t_0 AND t_1 AND
4 X BETWEEN AX AND AX + AW AND
5 Y BETWEEN AY AND AY + AH AND
6 Z BETWEEN AZ AND AZ + 2.70 AND
7 BADGE_ID = bid

```

References

- Mokatren, M.; Kuflik, T.; Shimshoni, I. Listen to What You Look at: Combining an Audio Guide with a Mobile Eye Tracker on the Go. In *Proceedings of the 10th International Workshop on Artificial Intelligence for Cultural Heritage Co-Located with the 15th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2016), CEUR Workshop Proceedings*; Bordoni, L., Mele, F., Sorgente, A., Eds.; CEUR-WS.org: Aachen, Germany, 2016; Volume 1772, pp. 2–9.
- Ardissono, L.; Kuflik, T.; Petrelli, D. Personalization in Cultural Heritage: The Road Travelled and the One Ahead. *User Model. User-Adapt. Interact.* **2012**, *22*, 73–99. [[CrossRef](#)]
- Lanir, J.; Kuflik, T.; Sheidin, J.; Yavin, N.; Leiderman, K.; Segal, M. Visualizing Museum Visitors' Behavior: Where Do They Go and What Do They Do There? *Pers. Ubiquitous Comput.* **2017**, *21*, 313–326. [[CrossRef](#)]
- Seidenari, L.; Baecchi, C.; Uricchio, T.; Ferracani, A.; Bertini, M.; Del Bimbo, A. Deep Artwork Detection and Retrieval for Automatic Context-Aware Audio Guides. *ACM Trans. Multimed. Comput. Commun. Appl.* **2017**, *13*, 35:1–35:21. [[CrossRef](#)]
- Martella, C.; Miraglia, A.; Frost, J.; Cattani, M.; Steen, M.V. Visualizing, clustering, and predicting the behavior of museum visitors. *Pervasive Mob. Comput.* **2017**, *38*, 430–443. [[CrossRef](#)]
- Stepień, J.; Kołodziej, J.; Machowski, W. Mobile user tracking system with ZigBee. *Microprocess. Microsyst.* **2016**, *44*, 47–55. [[CrossRef](#)]
- Rocchi, C.; Stock, O.; Zancanaro, M.; Kruppa, M.; Krüger, A. The museum visit: Generating seamless personalized presentations on multiple devices. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*; ACM: New York, NY, USA, 2004; pp. 316–318.
- Oppermann, R.; Specht, M. A Nomadic Information System for Adaptive Exhibition Guidance. *Arch. Mus. Inform.* **1999**, *13*, 127–138. [[CrossRef](#)]
- D'Aniello, G.; Gaeta, M.; Orciuoli, F.; Sansonetti, G.; Sorgente, F. Knowledge-based smart city service system. *Electronics* **2020**, *9*, 965. [[CrossRef](#)]
- Sansonetti, G.; Gasparetti, F.; Micarelli, A.; Cena, F.; Gena, C. Enhancing Cultural Recommendations through Social and Linked Open Data. *User Model. User-Adapt. Interact.* **2019**, *29*, 121–159. [[CrossRef](#)]
- Trejo, K.; Angulo, C.; i Satoh, S.; Bono, M. Towards robots reasoning about group behavior of museum visitors: Leader detection and group tracking. *J. Ambient Intell. Smart Environ.* **2018**, *10*, 3–19. [[CrossRef](#)]
- Centorrino, P.; Corbetta, A.; Cristiani, E.; Onofri, E. Measurement and analysis of visitors' trajectories in crowded museums. In *Proceedings of the IMEKO TC4 International Conference on Metrology for Archaeology and Cultural Heritage (MetroArchaeo 2019)*, Florence, Italy, 4–6 December 2019; pp. 423–428.
- Rashed, M.G.; Suzuki, R.; Yonezawa, T.; Lam, A.; Kobayashi, Y.; Kuno, Y. Tracking visitors in a real museum for behavioral analysis. In *Proceedings of the 2016 Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS)*, Sapporo, Japan, 25–28 August 2016; pp. 80–85.
- Liu, H.; Darabi, H.; Banerjee, P.; Liu, J. Survey of Wireless Indoor Positioning Techniques and Systems. *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)* **2007**, *37*, 1067–1080. [[CrossRef](#)]

15. Zafari, F.; Gkelias, A.; Leung, K.K. A Survey of Indoor Localization Systems and Technologies. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 2568–2599. [[CrossRef](#)]
16. Fiorucci, M.; Khoroshiltseva, M.; Pontil, M.; Traviglia, A.; Del Bue, A.; James, S. Machine Learning for Cultural Heritage: A Survey. *Pattern Recognit. Lett.* **2020**, *133*, 102–108. [[CrossRef](#)]
17. Augello, A.; Infantino, I.; Pilato, G.; Vitale, G. Site Experience Enhancement and Perspective in Cultural Heritage Fruition—A Survey on New Technologies and Methodologies Based on a “Four-Pillars” Approach. *Future Internet* **2021**, *13*, 92. [[CrossRef](#)]
18. Obeidat, H.; Shuaieb, W.; Obeidat, O.; Abd-Alhameed, R. A Review of Indoor Localization Techniques and Wireless Technologies. *Wirel. Pers. Commun.* **2021**, *119*, 289–327. [[CrossRef](#)]
19. Roy, P.; Chowdhury, C. A Survey of Machine Learning Techniques for Indoor Localization and Navigation Systems. *J. Intell. Robot. Syst.* **2021**, *101*, 63. [[CrossRef](#)]
20. Centenaro, M.; Vangelista, L.; Zanella, A.; Zorzi, M. Long-range communications in unlicensed bands: The rising stars in the IoT and smart city scenarios. *IEEE Wirel. Commun.* **2016**, *23*, 60–67. [[CrossRef](#)]
21. Zafari, F.; Papapanagioutou, I.; Christidis, K. Microlocation for Internet-of-Things-Equipped Smart Buildings. *IEEE Internet Things J.* **2016**, *3*, 96–112. [[CrossRef](#)]
22. Stock, O.; Zancanaro, M.; Busetta, P.; Callaway, C.; Krüger, A.; Kruppa, M.; Kuflik, T.; Not, E.; Rocchi, C. Adaptive, intelligent presentation of information for the museum visitor in PEACH. *User Model. User-Adapt. Interact.* **2007**, *17*, 257–304. [[CrossRef](#)]
23. Kuflik, T.; Lanir, J.; Dim, E.; Wecker, A.; Corra', M.; Zancanaro, M.; Stock, O. Indoor positioning: Challenges and solutions for indoor cultural heritage sites. In Proceedings of the 16th international conference on Intelligent user interfaces, Palo Alto, CA, USA, 13–16 February 2011; pp. 375–378.
24. Holm, S. Hybrid ultrasound-RFID indoor positioning: Combining the best of both worlds. In Proceedings of the 2009 IEEE International Conference on RFID, Orlando, FL, USA, 27–28 April 2009; pp. 155–162.
25. Baronti, P.; Pillai, P.; Chook, V.W.C.; Chessa, S.; Gotta, A.; Hu, Y.F. Wireless Sensor Networks: A Survey on the State of the Art and the 802.15.4 and ZigBee Standards. *Comput. Commun.* **2007**, *30*, 1655–1695. [[CrossRef](#)]
26. Maranò, S.; Gifford, W.M.; Wymeersch, H.; Win, M.Z. NLOS identification and mitigation for localization based on UWB experimental data. *IEEE J. Sel. Areas Commun.* **2010**, *28*, 1026–1035. [[CrossRef](#)]
27. Armstrong, J.; Sekercioglu, Y.A.; Neild, A. Visible light positioning: A roadmap for international standardization. *IEEE Commun. Mag.* **2013**, *51*, 68–73. [[CrossRef](#)]
28. Xiao, Y.; Ou, Y.; Feng, W. Localization of indoor robot based on particle filter with EKF proposal distribution. In Proceedings of the 2017 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), Ningbo, China, 19–21 November 2017; pp. 568–571.
29. Huang, W.; Xiong, Y.; Li, X.Y.; Lin, H.; Mao, X.; Yang, P.; Liu, Y.; Wang, X. Swadloon: Direction Finding and Indoor Localization Using Acoustic Signal by Shaking Smartphones. *IEEE Trans. Mob. Comput.* **2015**, *14*, 2145–2157. [[CrossRef](#)]
30. Hazas, M.; Hopper, A. Broadband ultrasonic location systems for improved indoor positioning. *IEEE Trans. Mob. Comput.* **2006**, *5*, 536–547. [[CrossRef](#)]
31. Desmarais, Y.; Mottet, D.; Slangen, P.; Montesinos, P. A review of 3D human pose estimation algorithms for markerless motion capture. *Comput. Vis. Image Underst.* **2021**, *212*, 103275. [[CrossRef](#)]
32. Hong, S.; Yi, T.; Yum, J.; Lee, J.H. Visitor-artwork network analysis using object detection with image-retrieval technique. *Adv. Eng. Inform.* **2021**, *48*, 101307. [[CrossRef](#)]
33. Saito, N.; Kusunoki, F.; Inagaki, S.; Mizoguchi, H. Novel application of an RGB-D camera for face-direction measurements and object detection: Towards understanding museum visitors' experiences. In Proceeding of the 13th International Conference on Sensing Technology (ICST 2019), Sydney, NSW, Australia, 2–4 December 2019.
34. Angeloni, R.; Pierdicca, R.; Mancini, A.; Paolanti, M.; Tonelli, A. Measuring and evaluating visitors' behaviors inside museums: The Co. ME. project. *SCIRES-IT-Sci. Res. Inf. Technol.* **2021**, *11*, 167–178.
35. Caldarelli, S.; Gurini, D.F.; Micarelli, A.; Sansonetti, G. A Signal-Based Approach to News Recommendation. In *CEUR Workshop Proceedings*; CEUR-WS.org: Aachen, Germany, 2016; Volume 1618.
36. Hassan, H.A.M.; Sansonetti, G.; Gasparetti, F.; Micarelli, A.; Beel, J. BERT, ELMO, USE and InferSent Sentence Encoders: The Panacea for Research-Paper Recommendation? In *RecSys 2019 Late-Breaking Results*; Tkalcic, M., Pera, S., Eds.; CEUR-WS.org: Aachen, Germany, 2019; Volume 2431, pp. 6–10.
37. Centorrino, P.; Corbetta, A.; Cristiani, E.; Onofri, E. Managing crowded museums: Visitors flow measurement, analysis, modeling, and optimization. *J. Comput. Sci.* **2021**, *53*, 101357. [[CrossRef](#)]
38. Zhao, Z.Q.; Zheng, P.; Xu, S.T.; Wu, X. Object Detection With Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [[CrossRef](#)]
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity Mappings in Deep Residual Networks. In *European Conference on Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 630–645.
41. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]

42. Mezzini, M.; Limongelli, C.; Sansonetti, G.; De Medio, C. Tracking Museum Visitors through Convolutional Object Detectors. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization; UMAP'20 Adjunct*; ACM: New York, NY, USA, 2020; pp. 352–355. [[CrossRef](#)]
43. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep learning for generic object detection: A survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [[CrossRef](#)]
44. Mezzini, M. Polynomial time algorithm for computing a minimum geodetic set in outerplanar graphs. *Theor. Comput. Sci.* **2018**, *745*, 63–74. [[CrossRef](#)]
45. O'Rourke, J. *Art Gallery Theorems and Algorithms*; Oxford University Press, Inc.: New York, NY, USA, 1987.
46. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: New York, NY, USA, 2003.
47. Hsu, C.C.; Lu, M.C.; Wang, W.Y.; Lu, Y.Y. Distance measurement based on pixel variation of CCD images. *ISA Trans.* **2009**, *48*, 389–395. [[CrossRef](#)]
48. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15)—Volume 1*; MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
49. Zhang, H.; Deng, Q. Deep Learning Based Fossil-Fuel Power Plant Monitoring in High Resolution Remote Sensing Images: A Comparative Study. *Remote Sens.* **2019**, *11*, 1117. [[CrossRef](#)]
50. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis. (IJCV)* **2015**, *115*, 211–252. [[CrossRef](#)]
51. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755.
52. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 22–29 October 2017; pp. 2980–2988. [[CrossRef](#)]
53. Wenkel, S.; Alhazmi, K.; Liiv, T.; Alrshoud, S.; Simon, M. Confidence Score: The Forgotten Dimension of Object Detection Performance Evaluation. *Sensors* **2021**, *21*, 4350. [[CrossRef](#)] [[PubMed](#)]
54. Salton, G.; McGill, M.J. *Introduction to Modern Information Retrieval*; McGraw-Hill, Inc.: New York, NY, USA, 1986.
55. Hosang, J.; Benenson, R.; Dollár, P.; Schiele, B. What Makes for Effective Detection Proposals? *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 814–830. [[CrossRef](#)]
56. Yoshimura, Y.; Krebs, A.; Ratti, C. Noninvasive Bluetooth Monitoring of Visitors' Length of Stay at the Louvre. *IEEE Pervasive Comput.* **2017**, *16*, 26–34. [[CrossRef](#)]
57. Ferrato, A.; Limongelli, C.; Mezzini, M.; Sansonetti, G. Exploiting Micro Facial Expressions for More Inclusive User Interfaces. In *Joint Proceedings of the ACM IUI 2021 Workshops Co-Located with 26th ACM Conference on Intelligent User Interfaces (ACM IUI 2021)*; Glowacka, D., Krishnamurthy, V.R., Eds.; CEUR-WS.org: Aachen, Germany, 2021; Volume 2903.
58. McDuff, D.; Mahmoud, A.; Mavadati, M.; Amr, M.; Turcot, J.; Kaliouby, R.e. AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*, San Jose, CA, USA, 7–12 May 2016; pp. 3723–3726.
59. Sansonetti, G. Point of Interest Recommendation Based on Social and Linked Open Data. *Pers. Ubiquitous Comput.* **2019**, *23*, 199–214. [[CrossRef](#)]
60. Fogli, A.; Sansonetti, G. Exploiting Semantics for Context-Aware Itinerary Recommendation. *Pers. Ubiquitous Comput.* **2019**, *23*, 215–231. [[CrossRef](#)]
61. Li, Y.; Zhang, X.; Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1091–1100.
62. Robinson, E.S.; Sherman, I.C.; Curry, L.E.; Jayne, H.H.F. The behavior of the museum visitor. *Publ. Am. Assoc. Mus.* **1928**, *1*, 72.
63. Melton, A.W. Visitor Behavior in Museums: Some Early Research in Environmental Design. *Hum. Factors* **1972**, *14*, 393–403. [[CrossRef](#)]
64. Falk, J.H. Assessing the Impact of Exhibit Arrangement on Visitor Behavior and Learning. *Curator Mus. J.* **1993**, *36*, 133–146. [[CrossRef](#)]
65. Serrell, B. *Paying Attention: Visitors and Museum Exhibitions*; G-Reference, Information and Interdisciplinary Subjects Series; American Association of Museums: Washington, DC, USA, 1998.
66. Agrusti, F.; Gasparetti, F.; Gena, C.; Sansonetti, G.; Tkalcic, M. SOcial and Cultural IntegrAtion with PersonalIZED Interfaces (SOCIALIZE). In *IUI '21: 26th International Conference on Intelligent User Interfaces*; Hammond, T., Verbert, K., Parra, D., Eds.; ACM: New York, NY, USA, 2021; pp. 9–11. [[CrossRef](#)]
67. Pavlidis, G. Recommender systems, cultural heritage applications, and the way forward. *J. Cult. Herit.* **2019**, *35*, 183–196. [[CrossRef](#)]
68. Sansonetti, G.; Gasparetti, F.; Micarelli, A. Cross-Domain Recommendation for Enhancing Cultural Heritage Experience. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization; UMAP'19 Adjunct*; ACM: New York, NY, USA, 2019; pp. 413–415.