A Cultural Heritage Experience for Visually Impaired People

A Cultural Heritage Experience for Visually Impaired People

Alice Lo Valvo[1], **Domenico Garlisi**[1,3], **Laura Giarré**[2], **Daniele Croce**[1,4], **Fabrizio Giuliano**[1], **Ilenia Tinnirello**[1],

1: DI, Università di Palermo, Viale delle Scienze, ed. 9, 90128 Palermo, Italy; 2: DIEF, Univ. di Modena e Reggio Emilia, Via P. Vivarelli, 10, 41125 Modena, Italy; 3: CNIT Consortium, Viale G.P. Usberti, 181/A - 43124 Parma, Italy; 4: DI, University of Rome La Sapienza, Italy.

Abstract. In recent years, we have assisted to an impressive advance of computer vision algorithms, based on image processing and artificial intelligence. Among the many applications of computer vision, in this paper we investigate on the potential impact for enhancing the cultural and physical accessibility of cultural heritage sites. By using a common smartphone as a mediation instrument with the environment, we demonstrate how convolutional networks can be trained for recognizing monuments in the surroundings of the users, thus enabling the possibility of accessing contents associated to the monument itself, or new forms of fruition for visually impaired people. Moreover, computer vision can also support autonomous mobility of people with visual disabilities, for identifying pre-defined paths in the cultural heritage sites, and reducing the distance between digital and real world.

1. Introduction

Nowadays, thanks to the pervasive availability of connectivity and digital services, we are experiencing new ways of accessing cultural heritage sites and organizing the travel experience, which are often known under the name Tourism 4.0. Although, so far, the most innovative aspect has been the exploitation of mobile connectivity for booking tickets, hotels, online visits and providing feedback, the advent of 5G and emerging technologies for data analysis can represent a real change for the development of a digital cultural heritage integrated and interacting with the real world. Many resources of our cultural heritage, such as archaeological sites, museums, cathedrals, historic buildings, etc., have already been digitized, through the reconstruction of digital maps and models, the collection of photographs, videos and related documents, as well as through the creation of websites that describe the sites and available masterpieces, offering services such as the purchase of tickets or the booking of guided tours. However, we still have a very clear distinction between the experiences of *physical access* to the chosen monument or museum, and *digital access* to the contents available on the website (which represent a digital twin of the good). This distance can be reduced by means of the emerging augmented reality or mixed reality applications, which are becoming possible thanks to the ultra-responsive services of 5G networks.

The real-time association of digital contents to a physical space or monument can combine the advantages of a real fruition experience with those of a personalized navigation, by means of digital mediation tools designed for responding to particular needs of the user (e.g. translation of descriptions, enlargements or different colors for the visually impaired, etc.) and for facilitating the access to the digital contents. This association obviously requires a user localization function (even in indoor spaces), for identifying the user environment, the point of view, as well as the identification of objects and persons for potential interactions in the surroundings.

In this paper we describe the design and implementation of two main components based on computer vision, which in our opinion can enable an innovative experience in the fruition of cultural heritage: i) a localization system, responsible of guiding the user along pre-defined paths; ii) a monument identification system, responsible of facilitating the access to the digital contents associated to specific sites of interest. Although the two components can be customized for many different use cases, we specifically deal with the case of visually impaired users, for which both the digital and physical access to the sites of interest require a special design. The paper is organized as follows: after a brief introduction on the related work in Section 2, Section 3 presents the localization system, called ARIANNA. Section 4 presents the monument identification system based on convolutional neural networks (CNNs). Finally, in Section 5 we draw some conclusions and discuss future evolution of our work.

2. Related work

Autonomous navigation of visually impaired people. Generally, indoor navigation systems for visually impaired people have stricter requirements than general purpose systems, in terms of accuracy and reaction times. In such environments, different solutions have been considered for supporting user orientation, based on the triangulation of RF signals (mainly WiFi), direct sensing of reference points (with RFIDs, ultrasound, bluetooth, Beacons, etc.), These systems, however suffer by a reduced precision and can only provide position information in the range of meters, and they are not useful for a precise visually reduced people guide. Ego-motion estimate using Inertial Measurement Units, i.e., accelerometers, magnetometers, and gyroscopes, are known as *Pedestrian Dead Reckoning*, is accurate in low-medium range, but it suffers from drift due to noise [1] on long distance.

Moreover, visually impaired people are used to continuous reference signals, which guide the users along the path to the destination. Today, these signals are currently implemented using tactile pavings that the blind person is sensing trough his/her cane.

Recent approaches of assistive technology for blind navigation have tried to implement sensory substitution, such as the one based on LIDARs [2], or vibrotactile stimulation [3] applied to the palms of both hands and directions were coded through the temporal structure of the stimuli. To achieve safe navigation, time-of-flight distance sensors [4] or cameras [5] have been exploited, as well as a vibrating belt giving navigation information. Concerning indoor navigation, both radio and visual landmarks have been largely investigated. In [6], specific landmarks are deployed on the points of interest in the environment; in [7] a vision system is developed.

Monument recognition for visually impaired people. Deep Convolutional Neural Networks (DCNN) are largely exploited in the context of image or object recognition. However, their utilization in the context of monument or environment recognition is a relatively new topic of research. These neural networks are based on very interesting structures, called Convolution Neural Networks (CNNs), able to automatically extract the features characterizing different images or objects. In the context of digital services of Tourism 4.0, current research projects have been mainly considered the problem of monument classification, such as the problem of identifying a special building architecture as a church, a palace, a bridge or tower. This is not a trivial task, especially because of the wide ranges of architectural solutions that have been proposed worldwide for implementing a special class of monuments [8]. Moreover, images of buildings in a real exploration can be partially captured by users, under varying environmental conditions (light, background, overlapping objects and people) which may complicate the identification process [9] and [10]. A project similar to ours, devised to identify a special site rather than a whole class of monuments is presented in [11], by still working on image analysis.

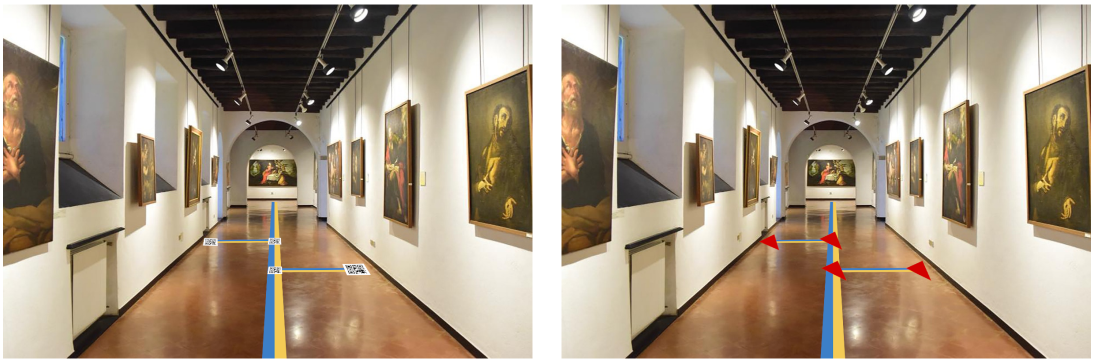

Figure 1. A possible installation of the ARIANNA system in a museum, with QRcode (a) and visual landmarks (b).

Differently from previous approaches, we consider the application of object detection algorithms to the problem of monument recognition. Object detection allows not only to localize the precise monument position within the image by means of a bounding box, but also to recognize small objects within a general background. This is very suitable to our scenario of visually impaired users, whose smartphone cannot be oriented easily towards the most relevant part of the monument. Since training an object recognition model from scratch takes a long time, we based our training mechanism on transfer learning, a popular method in the machine learning field, according to which a model developed for a specific problem is reused as the start point for a new model.

3. The ARIANNA system

ARIANNA is a system for indoor and outdoor localization and navigation, based on the joint utilization of dead-reckoning and computer vision techniques on a common smartphone [12, 13, 14, 15, 16, 17]. The system is explicitly designed for visually impaired people, but it can be easily generalized to other users, and it is built under the assumption that landmarks, such as colored tapes, painted lines, or tactile paving, are deployed in the environment for guiding visually impaired users along pre-defined paths.

A key component of ARIANNA is the computer vision algorithm which recognizes the painted line on the floor, under varying environmental light conditions. There are many different computer vision functions that can be combined for the identification of the painted line, taking into account the constraints of our problem: i) the path identification has to be prompt and reliable, without perceivable latency for the users, which could correspond to discontinuous signals; ii) the lifetime of the smartphone battery has to be compatible with the timing required, in order to guarantee the practical usage of the system. These constraints correspond to the identification of robust solutions, with limited complexity, able to work in real-time. Two main features can be exploited for detecting the paths: the geometry of the tapes (which in the end are given by piecewise lines), and the colors of the tapes (which combine two different colors for representing a direction without ambiguity).

ARIANNA identifies the painted lines in the acquired images, quantifies the slope of the lines seen by the camera, and converts this slope in an absolute orientation of the user, on the basis of a rough positioning of the user on the map. Indeed, the map of the paths are given by a sequence of segments, i.e., a piece-wise line, with different (absolute) orientations: it is enough to know at which segment the user can be located for converting her/his relative orientation to the line in an absolute heading measurement. To identify the slope of the line seen by the camera, we implemented three different steps: i) filtering the image, for reducing the noise and the details of the image background; ii) applying the Canny algorithm, for detecting the edges of the objects in the image; iii) identifying the sub-set of edges which can be considered as a line using the Hough transform.

Along the path, the system also permits to find some points of interests, by detecting landmarks (such as QRcodes or iBeacons) and retrieve location-based information. Figure 1 shows two possible installations of ARIANNA in a museum, where the colored paths are represented by two colors, which make the identification of the lines more robust to other similar colors present in the environment, and allow the application to immediately retrieve a direction information. Landmarks can be based on QR codes (as in figure 3) or other visual landmarks (3), because they are generally more precise in terms of positioning, although they are more visible and potentially impact the aesthetic of the installation. In order to make our navigation system more independent from a synthetic visual pattern, the iBeacons solution can be applied. In addition, iBeacons can be hidden and can represent a preferable choice in cultural environments such as museums. A data network connection is optional but if available can be used to download additional information such as the ambient map (e.g. via a WiFi or cellular network), as well as contents associated to the points of interest. The user interface adopts tactile stimuli (vibration of the smartphone) to provide feedback. It has been shown that the current consumption of typical vibration motors has a limited impact on the battery life of commercial smartphones[18] and that the energy savings coming from switching off the screen are higher than the costs introduced by vibrational cues [19].

4. Cultural heritage recognition

A monument is a building erected in order to commemorate people, events or become important for remembering historic times or cultural heritage. For these reasons, monuments are tourist destinations in any country, also for blind people. In order to support the blind people experience, we decided to search a solution for an automatic monument recognition. The idea is facilitating the access on the digital contents (voice description, sounds/musics, different representations such as images with very bright colors, etc.) associated to the physical monument.

4.1. Neural Network Design

We propose to address the problem of monument identification in an image captured by a smartphone as an object detection problem. Our idea is that the user, following the painted line by means of the ARIANNA system in both indoor and outdoor spaces, from time to time or upon the explicit signal of a point-of-interest can change the orientation of the smartphone camera from the floor to the front space. This operation could be somehow easy for low-vision people (which can identify a building, without perceiving the details), but could also be possible for blind people, assisted by means of vocal messages suggesting the right orientation. In any case, the acquired images are in general of different qualities, with the monument located at any possible position within the image and, in some cases, only partially captured by the image.

We analyzed various methods for object detection including Region-based Convolutional Neural Networks (R-CNN), Faster-RCNN[20], Single Shot Detector (SSD)[21] and You Only Look Once (YOLO). R-CNN and Faster R-CNN are composed by two stage detectors: the first stage identifies a subset of regions of interests in an image that might contain an object, while the second stage is used for object classification and bounding-box regression. The first stage in R-CNN is a slow object detection algorithm called Selective Search, instead Faster R-CNN uses a very small convolutional network called Region Proposal Network to generate regions of interests. Conversely, SSD and YOLO are methods that consider detection as a regression problem and they are composed by only one stage detector. YOLO has the best performance in terms of training time, but not the best in term of accuracy.

Thus, we compare the performance of different structures, based on the combination of multi stages: namely, Faster R-CNN and SSD in combination with Inception v2[22] and ResNet [23].

More specifically, we evaluate the detection performance of different models for the recognition of monuments which are part of the UNESCO Arab-Norman itinerary in Palermo, Italy, wonderful Mediterranean city in the island of Sicily. We compare three different object detection structures (or models):

- Model 1: Faster R-CNN, Inception v2;
- Model 2: Faster R-CNN, Inception v2, ResNet;
- Model 3: SSD, Inception v2.

As mentioned above, all the CNNs considered in our design are composed by different feature extractors: Faster R-CNN, Inception v2, ResNet and SSD. The main layers in a CNN are: convolution layer, pooling layer and fully-connected layer. The primary purpose of a convolution layer is to extract features from the input image through a set of independent filters. The function of a pooling layer is to progressively reduce the spatial size of the representation, in order to reduce the amount of parameters and computation in the network. Finally a fully connected layer takes the end result of the convolution/pooling process and reaches a classification. An example of the composition of a CNN is shown in Figure 2. As previously mentioned, we adopted a Transfer Learning approach for training. In other words, our neural networks are pre-trained by means of the Common Objects in Context (COCO) data set, a large-scale data set containing 1.5 million object instances and more than 200,000 labeled images.

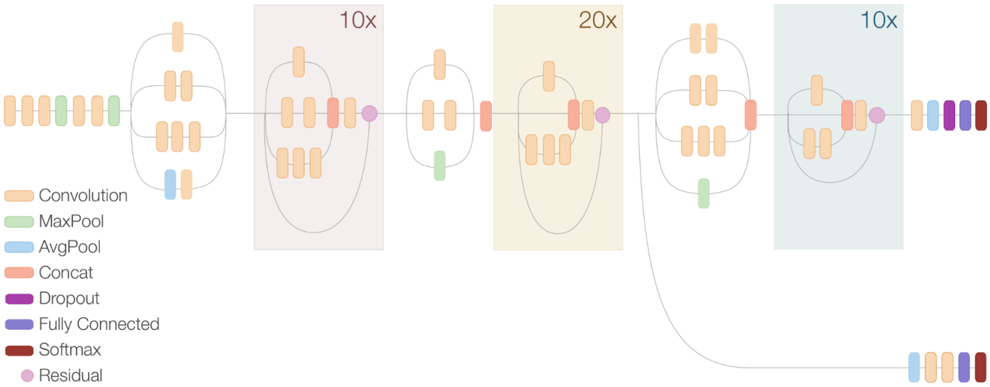

Figure 2. Faster R-CNN Inception ResNet v2 neural network layers.

Models have been trained to recognize 10 monuments (i.e. theaters and churches) of different size and/or structures. Figures 3-12 show the complete set of monuments used as point-of-interest to be recognized. They include very popular sites in Palermo, as well as monuments of the same class (e.g. churches) with very evident similarities. The input dataset consists of 1572 photos (more than 100 pictures per monument), which have been explicitly taken for this purpose. We chose different poses, weather and light conditions and with the presence of object or people in background. Finally, all the pictures were down-sampled in order to satisfied the training system requirements. For each image, we selected the exact location of the monument within the image, through LabelImg, a graphical image annotation tool to label images for bounding box object detection, and we associated the exact label. TensorFlow Object Detection API, an open source framework developed by Google, was used for training our three pre-trained neural networks. More specifically, we used the 80% of the images for the training and the remaining 20% for the evaluation.

The networks have been fine-tuned running 7500 consecutive iterations. During the training, we generated a log file for listing the labels and three new different models with the final weights. We trained all the three proposed structures with a common laptop (i.e. i7 processor, 16 GB RAM and without using any powerful GPUs) and the training times were, respectively, about 11 hours for the second structure and 3 hours for both the first and the third structure.

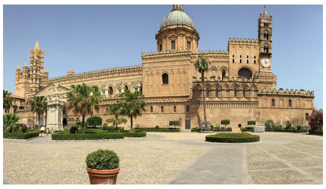

Figure 3. Cathedral.

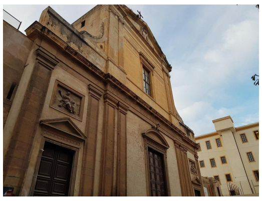

Figure 4. St. Cita Church.

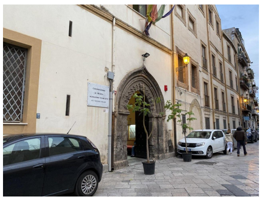

Figure 5. Conservatory of Palermo.

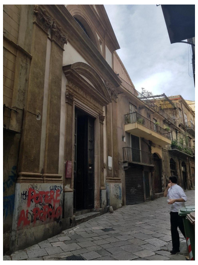

Figure 6. St. Domenico Church.

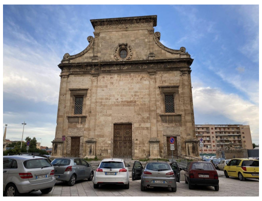

Figure 7. St. Giorgio Church.

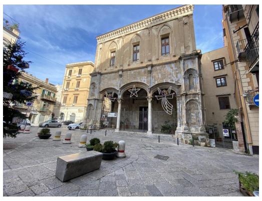

Figure 8. St. Maria La Nova Church.

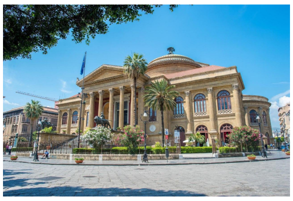

Figure 9. Massimo Theater.

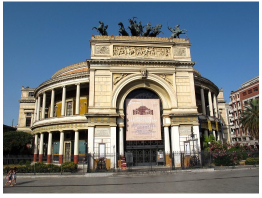

Figure 10. Politeama Theater.

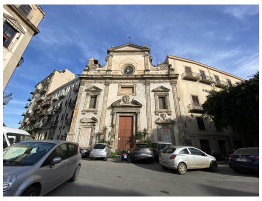

Figure 11. St. Sebastiano Church.

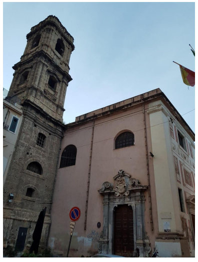

Figure 12. St. Maria Valverde Church.

4.2. Experimental Results

Once the neural networks have been trained, we evaluated the results taking into account confusion matrices for each model. Precisely, the horizontal rows represent what the ground truth, i.e. the correct monument label of the image, while the vertical columns represent the percentage of predictions corresponding to each possible monument of the set. In the last row, we also specify the percentage of times in which the monument has been identified as different from any other monuments of the set. Figures 13-15 show three confusion matrices corresponding to the three different structures or models considered in our work. From the results, it is evident that the last structure, namely the one corresponding to the usage of SSD and Inception v2 (model 3), despite of its simplicity and prompt training times, provides the best results. In most cases, the monument is correctly identified in more than 90% of the cases.

There are other two important metrics that can be evaluated: precision and recall. Precision refers to the percentage of results which are relevant, recall refers to the percentage of total relevant results correctly classified by your algorithm. For example, Figure 16 shows the Precision vs Recall curve of St. Maria Valverde Church. This further demonstrates the good performance of model 3.

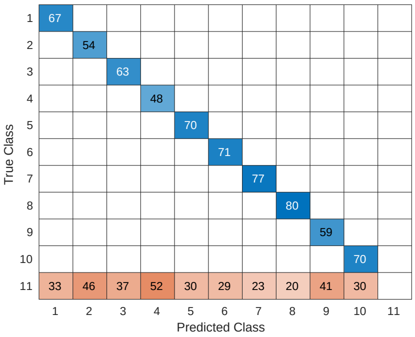

Figure 13. Confusion matrix of model 1.

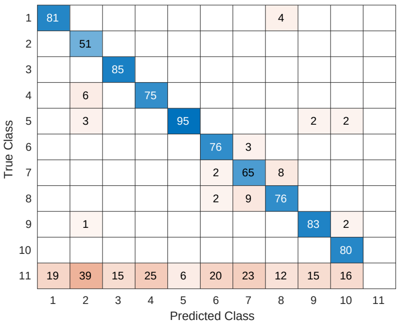

Figure 14. Confusion matrix of model 2.

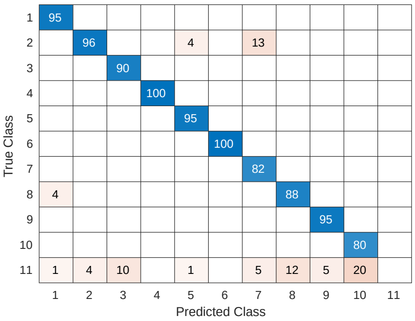

Figure 15. Confusion matrix of model 3.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}; \quad Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

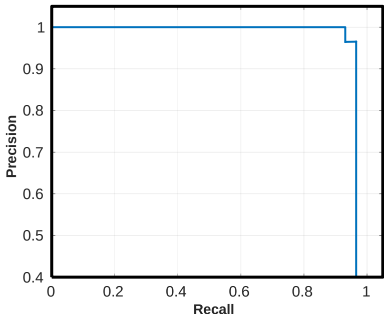

Figure 16. Precision vs Recall curve of St. Maria Valverde Church.

We can conclude that SSD Inception v2 for our point of interests in the Arab-Norman itinerary is the best solution.

5. Conclusion

In this paper we have presented two main components for improving the accessibility of tourist sites, by means of computer vision algorithms, in order to assist the autonomous navigation and understanding of the environment for visually impaired people or for other users with low vision and reduced mobility (including the elderly). We want to demonstrate that it is possible to facilitate the access of digital contents associated to real monuments, by advertising their availability near their physical locations and making virtual navigation possible even with normal smartphones. Thanks to the possibility of recognizing the points of interest and guiding the users along a path, we can envision the real-time creation of customized service instances (voice description of the places, simplification of contents for kids, enlargement of targets for elder people, etc.), responding to particular user needs, which may significant extend the touristic experience.

Acknowledgments

This work was supported in part by the European Union (EU) Founded Research Projects INTERREG V-A Italia-Malta 2014–2020 ”I-ACCESS Implementing the Accessibility to Urban Historic Center’s Use and Knowledge”. The project is carried out by the DARCH department at UNIPA with the code project C1-1.1-41.

References

[1] A.R. Jimènez, F. Seco, J.C. Prieto, J. Guevara Rosas, Indoor pedestrian navigation using an INS/EKF framework for yaw drift reduction and a foot-mounted IMU, Workshop on Positioning Navigation and Communication, Dresden, pp. 135–143, 2010.

[2] C. Ton et al., ”LIDAR Assist Spatial Sensing for the Visually Impaired and Performance Analysis,” in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 9, pp. 1727-1734, Sept. 2018. doi: 10.1109/TNSRE.2018.2859800

[3] R. Kessler, M. Bach and S. P. Heinrich, ”Two-Tactor Vibrotactile Navigation Information for the Blind: Directional Resolution and Intuitive Interpretation,” in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 25, no. 3, pp. 279-286, March 2017. doi: 10.1109/TNSRE.2016.2569258

[4] R. K. Katzschmann, B. Araki and D. Rus, ”Safe Local Navigation for Visually Impaired Users With a Time-of-Flight and Haptic Feedback Device,” in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 26, no. 3, pp. 583-593, March 2018. doi: 10.1109/TNSRE.2018.2800665

[5] H.-C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarré, D. Rus, ”Enabling independent navigation for visually impaired people through a wearable vision-based feedback system”, Proc. IEEE Int. Conf. Robot. Autom. (ICRA), pp. 6533-6540, May 2017.

[6] W. C. S. S. Simões and V. F. de Lucena, ”Hybrid Indoor Navigation as sistant for visually impaired people based on fusion of proximity method and pattern recognition algorithm,” 2016 IEEE 6th International Conference on Consumer Electronics - Berlin (ICCE-Berlin), Berlin, 2016, pp. 108-111. doi: 10.1109/ICCE-Berlin.2016.7684732

[7] R. Tapu, B. Mocanu and T. Zaharia, ”A computer vision system that ensure the autonomous navigation of blind people,” 2013 E-Health and Bioengineering Conference (EHB), Iasi, 2013, pp. 1-4. doi: 10.1109/EHB.2013.6707267

[8] A. Saini, T. Gupta, R. Kumar, A. K. Gupta, M. Panwar and A. Mittal, ”Image based Indian monument recognition using convoluted neural networks,” 2017 International Conference on Big Data, IoT and Data Science (BID), Pune, 2017, pp. 138-142.

[9] Amato, G., Falchi, F., Gennaro, C. ”Fast image classification for monument recognition.” J. Comput. Cult. Herit. (JOCCH). 2015, 8, 18.

[10] S. Gada, V. Mehta, K. Kanchan, C. Jain and P. Raut, ”Monument Recognition Using Deep Neural Networks,” 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, 2017, pp. 1-6.

[11] Palma, V., ”Towards deep learning for architecture: a monument recognition mobile app.” Int. Arch. Photogrammetry, Remote Sens. Spat. Inf. Sci. XLII-2/W9, 551–556 (2019)

[12] D. Croce, P. Gallo, D. Garlisi, L. Giarré, S. Mangione, I. Tinnirello, ARIANNA: A smartphone-based navigation system with human in the loop 22nd Mediterranean Conference of Control and Automation (MED), pp. 8–13, 2014.

[13] D. Croce, L. Giarré, F.G. La Rosa, E. Montana, I. Tinnirello, Enhancing tracking performance in a smartphone-based navigation system for visually impaired people, 24th Mediterranean Conference of Control and Automation (MED), 2016.

[14] Italian Patent N. BG2014A000054, Sistema di navigazione per non vedenti, presented 2015, patented 2017.

[15] G. Galioto, I. Tinnirello, D. Croce, F. Inderst, F. Pascucci, L. Giarré, Demo: Sensor Fusion Localization and Navigation for Visually Impaired People, MobiCom 2017.

[16] G. Galioto, I. Tinnirello, D. Croce, F. Inderst, F. Pascucci, L. Giarré, Sensor Fusion Localization and Navigation for Visually Impaired People, ECC 2018.

[17] D. Croce et al., ”An indoor and outdoor navigation system for visually impaired people,” in IEEE Access.doi: 10.1109/ACCESS.2019.2955046, 2019.

[18] M. Pielot, and R. de Oliveira. Peripheral Vibro-Tactile Displays, ACM Press, 2013. Mobile HCI 2013 - tactile user interfaces. doi:10.1145/2493190.2493197.

[19] M. Pielot, How the Phone's Vibration Alarm Can Help to Save Battery Accessed May 12, 2014. http://pielot.org/2012/12/11/how-the-phones-vibration-alarm-can-help-to-save-battery/.

[20] S. Ren, K. He, R. Girshick and J. Sun, ”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[21] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, ”SSD: Single shot multibox detector”, European Conference on Computer Vision, pp. 21-37, 2016.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, ”Rethinking the inception architecture for computer vision,” arXiv, no. 1512.00567, 2015.

[23] He K., Zhang X., Ren S., Sun J. ”Deep residual learning for image recognition”, Computer Vision and Pattern Recognition (CVPR) (2016), pp. 770-778