ARTIFICIAL INTELLIGENCE AS A LOW-COST SOLUTION FOR MUSEUM VISIT DIGITAL CONTENT ENRICHMENT: THE CASE OF THE FOLKLORE MUSEUM OF XANTHI

George Ioannakis[1,2], Loukas Bampis[1,3], Anestis Koutsoudis[1]*

[1] Athena Research and Innovation Centre, Xanthi’s Division, PO Box 159, Kimmeria Campus, 67100 Xanthi, Greece
[2] Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi 67100, Greece
[3] Department of Production and Management Engineering, Democritus University of Thrace, Xanthi 67100, Greece

Commission II, WG II/8

KEY WORDS: Content Enrichment, Expedition-Center Visits, Mobile Devices, Hybrid Recognition, Machine Learning, Convolutional Neural Networks, Bags of Visual Words

ABSTRACT:

The on-demand content enrichment of an exhibition center visit is an active applied research domain. This work focuses on the exploitation of mobile devices as an efficient medium to deliver information related to an exhibit or an area within the exhibition center by utilizing machine learning approaches. We present YPOPSEI, an integrated system that formulates the information retrieval task as an image recognition mechanism, enabling visitors to simply capture an entity of interest in order to acquire information similar to a tour-guidance experience via their personal mobile devices. This scheme not only minimizes the additional infrastructure requirements, but additionally enhances the versatility in cases of exhibits topology alterations while still providing high accuracy in terms of image content recognition. Two hybrid approaches are developed that set Convolutional Neural Networks (CNNs) and Bags of Visual Words (BOVWs) to operate in a synergistic and cooperative manner. They are evaluated under real-world conditions on a client-server Web architecture system that experimentally operates within the premises of the Folklore Museum of Xanthi, Greece.

1. INTRODUCTION

On-demand delivery of digital content, related to a point-of-interest (POI), is a procedure that is highly correlated to cultural heritage thesaurus interaction and augmented reality applications. It is a fact that museums, as well as other cultural heritage thesaurus exhibitions, compose a demanding domain for applying content recognition and classification methodologies in order to deliver supplemental information about a POI to a visitor. Nowadays, a wide range of recognition approaches based on IoT technologies, RFID tags, QR code plates, WIFI triangulation, Convolutional Neural Networks, etc. have been used to allow the accurate identification of an exhibit (O) or a thematic place (P) within a museum (Kuflik, Dim, 2019), (Sornalatha, Kavitha, 2017), (Kumar et al., 2019), (Koutsoudis et al., 2014), (Seidenari et al., 2019). Although, some of the above approaches can provide highly accurate recognition results (e.g. a unique QR code is usually placed in front of an exhibit), their additional infrastructure requirements combined with a fixed installation topology conclude to an inefficient framework in cases where exhibits need to change place, to be removed or to be replaced by others. This additional infrastructure may also be cost-prohibitive in terms of budget.

Recent advances in machine learning technologies can offer the required, by such a demanding application domain, recognition accuracy. In this work, we present YPOPSEI, an experimental client-server Web system that exploits novel hybrid image content recognition approaches. Those hybrids set methods such as Convolutional Neural Networks (CNNs) (LeCun et al., 1998), (Krizhevsky et al., 2012) and Bags of Visual Words (BOVWs) (Sivic, Zisserman, 2003) to operate under *synergistic* and *cooperative* operational schemata (data fusion) in order to achieve enhanced image content recognition performance. Despite the fact that CNNs are currently considered the state-of-the-art solution for classication tasks, their performance depends on surfaces that retain viewpoint-invariance, and they additionally reject topological information at their architecture’s higher levels (Fei et al., 2016), (Sizikova et al., 2016). On the contrary, the BoVWs model is not influenced by view changes, and thus it is still relevant and therefore composes a viable and relevant solution, especially in systems where a freely moving camera sensor is used. Thus, the combination of the above composes a highly applicable framework that operates under cases where a museum visitor follows a random path and a set of relevant, to a given POI, information is delivered through a minimal num-

Figure 1. An instance of the proposed YPOPSEI system’s deployment. The user is able to retrieve information regarding an exhibit through a minimal number of interactions on his/her personal mobile device.

*Corresponding author (akoutsou@ipet.gr)

ber of interactions (e.g. capture an image)(Fig. 1). The hybrid approaches being proposed (*synergy* and *cooperation* respectively) allowed to merge data at the ranking results of the two approaches (CNNs and BOVWs) that can be defined as Early Fusion (EF) or at the decision level which is considered as Late Fusion (LF). During EF, decisions are based on the complete information available to each approach while in LF each classifier's outcome is treated as an absolute and thus weighted to define a final decision.

The rest of this paper is organized as follows. In section 2 a detailed overview of the proposed recognition framework and the developed hybrid approaches are presented. We discuss computational efficiency and analyze their recognition accuracy on the benchmark image dataset collected from the Folklore Museum of Xanthi. In Section 3 the Web application architecture along with its technologies and recognition integration schema are thoroughly described. Finally, in Section 4 conclusions are drawn and potential future work plans are discussed.

2. METHODOLOGY

In this section, we discuss the two utilized machine learning mechanisms of CNNs and BOVWs, as well as the developed hybrid approaches (Fig. 2) that are used to enhance the recognition accuracy of the system. In addition, the adopted implementation techniques are also detailed, which reduced the computational complexity of our overall architecture resulting in an efficient system.

2.1 Recognition Approaches

The BOVWs approach relies on the description of each image through a set of contained visual words from a pre-trained vocabulary. This vocabulary is formulated through a set of training images recorded from the domain of interest and it is divided into two subsets: one for the description of the contained exhibits *O* and one for the thematic places *P*. More specifically, from each one of the training instances, the most prominent ORB (Rublee et al., 2011) are extracted and used as input into a k-median hierarchical clustering technique with k-means++ seeding (Arthur, Vassilvitskii, 2007). This procedure resulted in two vocabulary trees with $L = 6$ levels and $K = 10$ branches per level, or 10^6 visual words. When a new input image is processed, a Visual Word Vector (VWV) can be produced under the “term frequency - inverse document frequency” (*tf-idf*) (Sivic, Zisserman, 2003) model, allowing to quantify its resemblance with frames of known labels through the *cosine similarity*:

$$C_{sim}\left(\overline{\boldsymbol{v}}_1, \overline{\boldsymbol{v}}_2\right) = \frac{\overline{\boldsymbol{v}}_1 \cdot \overline{\boldsymbol{v}}_2}{\|\overline{\boldsymbol{v}}_1\| \, \|\overline{\boldsymbol{v}}_2\|}, \tag{1}$$

where $\overline{\boldsymbol{v}}_1$ and $\overline{\boldsymbol{v}}_2$ correspond to the VWVs to be compared. Note that the *cosine similarity* produces the same ranking results with the L2-score of unit vectors. Since the user's choice between an *O* or *P* query request is not known, the system formulates two distinct query VWVs, while the actual similarity score is considered as the maximum one. Finally, recognition is achieved by identifying the most encountered label among the k nearest neighbouring ones (k-NN classification).

For the case of CNNs, the standard technique of training a CNN classifier is followed, with each learned class corresponding to a known label. A trade-off equilibrium had to be established between computational speed and recognition accuracy, thus

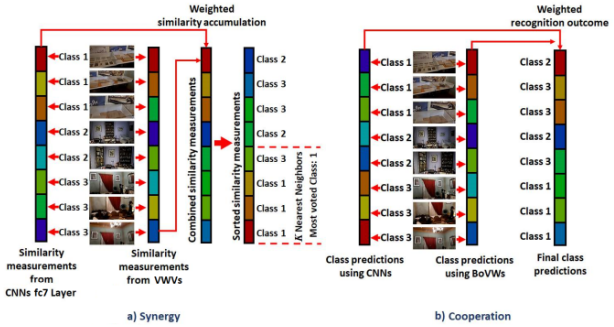

Figure 2. *Synergy* (a) and *cooperation* (b) operational schemata of the hybrid recognition approaches.

the selected CNN architecture is relatively shallow in comparison to the state-of-the-art ones, while still retaining high recognition results. More specifically, the architecture of the proposed CNN consists of five convolutional layers, three fully connected ones and finally a softmax layer. Convolutional layers 1,2 and 5 are followed by a pooling layer, whereas fully connected ones are followed by a dropout layer. Two distinct and identical CNNs are utilized within the scope of this work; a CNN dedicated to the recognition of exhibits *O* and one dedicated to the recognition of thematic places *P*. The intuition behind the decision to treat separately *O*'s and *P*'s respectively relies on the fact that thematic places *P*'s consist of multiple exhibits *O*'s and thus ambiguity could arise at the distinctiveness between them. For both CNN architectures, the Multinomial Logistic/Cross Entropy is used as loss function and are trained using Stochastic Gradient Descent (Bottou, 2010), while the initial weight values derived from training on the ILSVRC12 challenge (Deng et al., 2009). Each image passes through both CNNs, and the final label is defined by the softmax layer output with the highest probability.

With the above recognition approaches defined, the two produced ranking results are combined under our proposed *synergy* schema through a normalized weighting model (Fig. 2a). Such as the case of BOVWs, the trained CNN architectures are utilized in order to produce two distinct description vectors that better describe the properties of *O* and *P* samples, respectively. More specifically, we utilized the output from the pre-ultimate fully connected (*fc7*) layer as the descriptor from each frame. Thus, through the computation of C_{sim} between the input and all known-labeled frames, a vector containing all the similarity values is produced that can be combined with the equivalent one produced by the BOVWs mechanism. Finally, by combining the two similarity vectors through a normalized weighting schema, the recognition result is once again obtained by identifying the label with the most appearances among the k most similar ones, as in the BOVWs case.

On the contrary, the *cooperation* schema fuses data from BoVWs and CNNs at the decision level. More specifically, the two final decisions from each respective approach conclude into a single one through the assignment of normalized weights on the produced similarity scores. (Fig. 2b). In the case of CNNs, the maximum value from the softmax layer is utilized to define the winning class and the output of the final fully connected layer, that corresponds to this class, is chosen. The output passes through a ReLU layer and is normalized through: $fc8norm(i) = \frac{fc8(i)}{\sum_{k=1}^{B} fc8_k}$, where B is the total number of classes. Finally, through the combination of the BoVWs and CNNs results, a common prob-

ability decision is calculated through a weighting schema: $prob = max\,(b \cdot VWVprob,\ \ (1-b) \cdot CNNprob)$, where b is the weight that is defined through an optimization process during the recognition performance evaluation phase.

2.2 Computational Efficiency

Our algorithms' implementations are carefully designed in order to ensure the computational efficiency of YPOPSEI. This design includes well-established techniques for quickly computing the VWVs of the BOVWs mechanism and measuring similarities between those samples, as well as deploying the CNN architectures, using GPU-acceleration, to classify the input images and extract the *fc7* vectors.

With the view to accelerate the extraction of ORB features from each image, we adopted GPGPU versions (Bampis, Gasteratos, 2019) of the oFAST algorithm[1] (Rublee et al., 2011) and the ORB descriptor. Modern GPUs operate under the “Single Instruction on Multiple Threads” (SIMT) architecture[2] which allows for vast parallelization of computations when the targeted functionality can be segmented into multiple identical operations. Thus, the individual GPU threads are used in order to detect the points of interest in parallel through the rules defined in (Rosten, Drummond, 2006). The descriptor formulation for each one of those keypoints is implemented by 32 parallel threads that fit into one warp, the fundamental execution block of the GPU, sharing their data through local registers (Shuffle Functions (CUDA Library, 2019)). In addition, the choice of a binary local feature descriptor, such as ORB, allows for the utilization of Hamming distance so as to traverse the nodes of the vocabulary tree and identify the most representative visual words.

Furthermore, in order to measure the *cosine similarity* between an input image and the rest of the known instances under an indexing schema, an inverted file structure was implemented (Jegou et al., 2008). This structure takes advantage of the VWVs' high sparsity which is expected when large visual vocabularies are used. Given a training image collection with known labels, the vocabulary's inverted file is formulated as a set of lists, each of which retains the indexes of frames that a given visual word occurs, as well as the corresponding *tf* value. When the input image is processed, the similarity vector is initialized to zero. Then, for each visual word detected in the input frame, the corresponding list is accessed and every contained instance's score is increased by a $tf \times idf$ factor. The result of this procedure corresponds to the dot product in equation 1, and the final *cosine similarities* are computed by dividing the scores with the two L2-norms of the corresponding VWVs.

The cuDNN framework (cuDNN, 2019) enables training and inferring of CNN architectures at high-speed using CUDA. It provides GPU accelerated functionality that includes implementations for standard routines such as forward and backward convolution, pooling, normalization, and activation layers. Thus, within the scope of this work, we utilize GPU-acceleration in order to achieve the highest possible computational speed and efficiency.

[1] oFAST corresponds to the orientation-invariant version of FAST (Rosten, Drummond, 2006) detector used by ORB.

[2] SIMT refers to an extension of the more traditional “Single Instruction on Multiple Data” (*SIMD*) architecture.

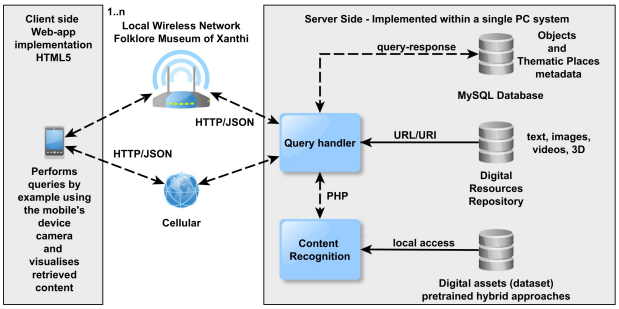

Figure 3. Experimental client-server model system architecture.

Table 1. Total recognition accuracy results over our testing set. Bold table entries highlight the best performing approach.

Method	BoVWs	CNNs	*Synergy*	*Cooperation*
Accuracy (%)	62.24	81.39	**93.17**	82.69

2.3 Recognition Accuracy

Identification of the optimum operational parameters of our image content recognition framework involves the creation of a benchmark image dataset based on exhibits O and thematic places P from the Folklore Museum of Xanthi, Greece. As with all evaluation datasets, this is organized into three parts: training, validation, and testing. There is no content sharing between the three parts in order to achieve objective recognition performance evaluation results and furthermore to setup the system's optimum parameterization schema. The dataset covers 16 places and 92 objects that are exhibited within the museum premises. It should be noted that an object (O) represents a single exhibit (smallest entity) while a place (P) is an area that represents a given theme and may contain many objects. All *Os* and *Ps* have been recorded several times using different hardware (camera sensors, optical lens, etc.). The dataset has a total of $50,054$ images related to *Os* and a total of $161,312$ images related to *Ps*. The ratios between the three parts of the dataset are 60%, 20% and 20%.

Through our recognition performance evaluation experiments (Table 1), we have concluded that although the CNN-based recognition mechanism offers higher performance compared to the BoVW-based one, the two proposed hybrid data fusion approaches do exceed in performance the previously mentioned ones and thus indicate the importance of combining their content recognition properties. Based on the obtained accuracy results, *synergy* outperforms *cooperation*. This difference relies on the fact that in *cooperation* we consider only the final decision from each individual approach; whereas in *synergy*, BoVW and CNN similarities from every instance with known label are considered. Nonetheless, the hybrid schemata achieved higher recognition performance than both CNNs and BoVWs. This is coupled with the fact that the distinct recognition capabilities of the latter two are proven insufficient for the targeted application, proving the importance of the proposed data fusion techniques which lead to an improved estimation of the decisions boundary. The content-based retrieval performance evaluation tasks were performed using an online information retrieval performance evaluation tool, RETRIEVAL (Ioannakis et al., 2018).

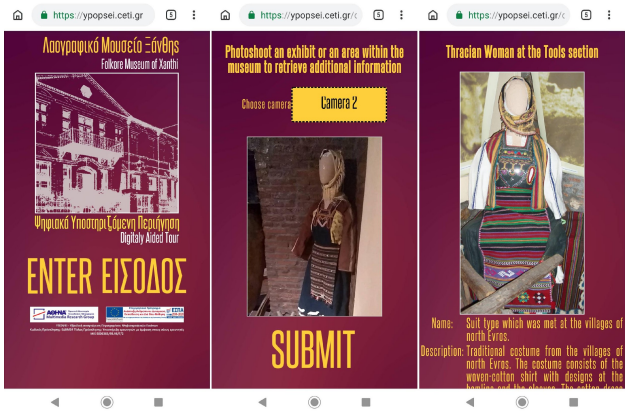

Figure 4. Screenshots from the experimental Web application performing a query-by-example and receiving relevant digital content to a captured exhibit.

3. WEB APPLICATION

The experimental client-server system (Fig. 3) being used within the premises of the museum relies on the exploitation of the best performing *synergy* hybrid approach. The Web application is currently bilingual (Greek and English), and in this section, we describe the design of its architecture.

3.1 The Client Side

The client side is currently implemented as a Web application that exploits HTML5, PhP and Javascript technologies in order to enable the submission of a query image to the server side. The system is based on the idea that a visitor identifies an exhibit (O) or a thematic place (P) in the museum that falls within his/her interests and therefore requests on-demand additional information about it. The visitor aims the camera of a mobile device towards this entity and a low-resolution image-based query-by-example is forwarded to the server's query handler over the local WIFI or Cellular using HTTP/JSON. Figure 4 depicts its current simple graphical user interface that delivers content through a minimum number of necessary interactions.

3.2 Database Architecture

The metadata of each O or P are organized in a relational database using MySQL. The database contains metadata concerning the title, description, dimensions, place of production, date of production, external URL (where applicable) and the name of an indicative representation of the exhibit O or thematic place P. Each database entry follows a unique ID encoding that also depicts its relation to the rest of Os and Ps.

3.3 The Server Side - Content Recognition Module

On the server side, the query handler is the first node of a task scheduler that initiates the recognition algorithms of our proposed system. The information flow of our design is depicted in Fig. 5 and it is implemented using the Robotics Operating System (ROS, 2019). The first node of the architecture forwards the input query image to the two machine learning mechanisms which operate in parallel and output the two similarity results between the query and every database instance in the form of two vectors. Then, the *synergy* node receives this data and combines them under the weighting schema described in Section 2.1. The produced recognition result is forwarded back to the query handler in the form of a unique ID that corresponds to the matched O or P. Finally, the handler retrieves the database information regarding the recognized entity, that is subsequently forwarded to the client side on an updated Web-page.

Table 2. Time results of Synergy and its sub-processes.

Process	Time (seconds)
VWV Formulation	0.147
fc7 Vector Formulation	0.370
Cosine Distances Calculation	3.158
Decision Making	0.009

3.4 Timing Results

In this section, an evaluation process is performed in order to define the response speed of the proposed integrated system YPOPSEI. An overview of the sub-processes execution response time results is presented in Table 2. YPOPSEI is installed on a PC system that is equipped with an 8-core Intel i7 processor at 3.60 GHz, 128GB of RAM and a NVidia GeForce GTX 1070 8GB RAM graphics card running Ubuntu 16.04 LTS 64-bit.

The “VWV Formulation” entry corresponds to the time needed for our system to extract ORB local features and convert them to the nearest visual words from the corresponding vocabularies. In the same manner, “*fc7* Vector Formulation” denotes the processing time required for deploying the network architectures and compute the *fc7* description vectors. The “Cosine Distances Calculation” entry refers to the process of computing cosine distances with the whole labeled data using equation 1, and finally, “Decision Making” mechanism is the one responsible for counting the label instances during k-NN classification and highlighting the final recognition decision. In the real-world case scenario additional delays are introduced by other parts of the system (e.g. WIFI/Cellular bandwidth, processing speed of mobile device etc.). In our experiments using a Xiaomi Redmi Note 6 over 4G Cellular gave a response time that was always under eight seconds when a single visitor was using YPOPSEI.

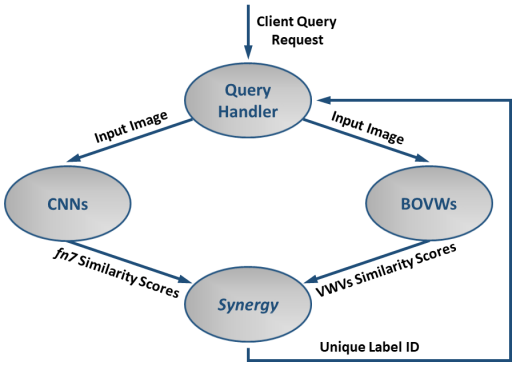

Figure 5. Block diagram of the content recognition module's information flow.

4. CONCLUSIONS AND FUTURE WORK

The overall recognition performance of the hybrid approaches in combination with the subjective feedback of the implemented experimental system indicates that such approaches can be

utilized in real-world applications and be a part of a broader framework that allows the on-demand digital content enrichment of a museum visit. The recognition accuracy achieved after the training of the content recognition subsystem indicates that it can significantly contribute in cases where the need of minimizing additional hardware requirements such as QR code plates, RFID tags or beacons is a prerequisite.

In the near future, we plan to apply the system outdoors (e.g. an archaeological site) without exploiting other localization mechanisms of the mobile device, such as Global Navigation Satellite Systems (e.g. GPS, Glonass, Galileo BeiDou) and compass. Moreover, we are considering to extend our database and adjust it to a simplified version of the CARARE 2.0 (D’Andrea, Fernie, 2013) schema that will be stored in a native XML database (eXist-db, 2019), in case that the information richness provided by the museum demands it. Finally, due to the fact that 3D digital replicas of a number of exhibits were created, we are planning to integrate them in the landing page (depicted on the mobile device using WebGL/X3DOM). The latter will increase the time required to deliver such complex content to the mobile device.

ACKNOWLEDGEMENTS

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme “Human Resources Development, Education and Lifelong Learning 20142020” in the context of the project “Ypopsei: Hybrid content recognition from bitmap images”, MIS code:5006383. We would also like to thank the personnel from Folklore Museum of Xanthi for their collaboration and the admission to install and evaluate our system performance in their premises, as well as the provision of metadata related to its content. Finally, we gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research work.

REFERENCES

Arthur, D., Vassilvitskii, S., 2007. k-means++: The advantages of careful seeding. *Proceedings of the ACM-SIAM symposium on Discrete algorithms*, 1027–1035.

Bampis, L., Gasteratos, A., 2019. Revisiting the Bag-of-Visual-Words Model: A Hierarchical Localization Architecture for Mobile Systems. *Robotics and Autonomous Systems*, 113, 104–119.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. *in COMPSTAT*.

CUDA Library, 2019. Compute Unified Device Architecture. http://docs.nvidia.com/cuda/cuda-c-programming-guide.

cuDNN, 2019. The NVIDIA CUDA Deep Neural Network library. https://developer.nvidia.com/cudnn.

D’Andrea, A., Fernie, K., 2013. CARARE 2.0: A Metadata Schema for 3D Cultural Objects. *In Proceedings IEEE Digital Heritage International Congress*, 2, 137–143.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A Large-Scale Hierarchical Image Database. *CVPR09*.

eXist-db, 2019. Open Source Software Project for NoSQL Databases. http://exist-db.org/.

Fei, X., Tsotsos, K., Soatto, S., 2016. A Simple Hierarchical Pooling Data Structure for Loop Closure. *In Proceedings of the European Conference on Computer Vision Compyter Vision*, 321–337.

Ioannakis, G., Koutsoudis, A., Pratikakis, I., Chamzas, C., 2018. RETRIEVALAn Online Performance Evaluation Tool for Information Retrieval Methods. *IEEE Transactions on Multimedia*, 20(1), 119-127.

Jegou, H., Douze, M., Schmid, C., 2008. Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search. *European Conference on Computer Vision*, 304–317.

Koutsoudis, A., Arnaoutoglou, F., Pavlidis, G., 2014. Passive Markers as a Low-Cost Method of Enriching Cultural Visits on Users Demand. *Journal of Advanced Computer Science & Technology*, 3(1), 12–17.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 1097–1105.

Kuflik, T., Dim, E., 2019. Integrating signals for reasoning about visitors’ behavior in cultural heritage. *Multimodal Behavior Analysis in the Wild*, 159–169.

Kumar, V. A., Saranya, G., Elangovan, D., Chiranjeevi, V. R., Kumar, V. A., 2019. IOT-Based Smart Museum Using Wearable Device. *In Proceedings of the International Conference on Innovative Computing and Communications*, 33–42.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. e. a., 1998. Gradient-based learning applied to document recognition. 86Number 11, 2278–2324.

ROS, 2019. Robotics Operating System. https://www.ros.org.

Rosten, E., Drummond, T., 2006. Machine learning for high-speed corner detection. *Proceedings of the European Conference on Computer Vision*, 430–443.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An Efficient Alternative to SIFT or SURF. *In Proceedings of the IEEE International Conference on Computer Vision*, 2564–2571.

Seidenari, L., Baecchi, C., Uricchio, T., Ferracani, A., Bertini, M., Bimbo, A. D., 2019. Chapter 9 - wearable systems for improving tourist experience. X. Alameda-Pineda, E. Ricci, N. Sebe (eds), *Multimodal Behavior Analysis in the Wild*, Computer Vision and Pattern Recognition, Academic Press, 171 – 197.

Sivic, J., Zisserman, A., 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. *In Proceedings of the IEEE International Conference on Computer Vision*, 1470–1477.

Sizikova, E., Singh, V. K., Georgescu, B., Halber, M., Ma, K., Chen, T., 2016. Enhancing Place Recognition Using Joint Intensity-Depth Analysis and Synthetic Data. *In Proceedings of the European Conference on Computer Vision, Workshop*, 901–908.

Sornalatha, K., Kavitha, V. R., 2017. Iot based smart museum using bluetooth low energy. *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 520–523.