Pepper4Museum: Towards a Human-like Museum Guide

Giovanna Castellano
Department of Computer Science
University of Bari Aldo Moro
Bari, Italy
giovanna.castellano@uniba.it

Berardina De Carolis
Department of Computer Science
University of Bari Aldo Moro
Bari, Italy
berardina.decarolis@uniba.it

Nicola Macchiarulo
Department of Computer Science
University of Bari Aldo Moro
Bari, Italy
nicola.macchiarulo@uniba.it

Gennaro Vessio
Department of Computer Science
University of Bari Aldo Moro
Bari, Italy
gennaro.vessio@uniba.it

ABSTRACT

With the recent advances in technology, new ways to engage visitors in a museum have been proposed. Relevant examples range from the simple use of mobile apps and interactive displays to virtual and augmented reality settings. Recently social robots have been used as a solution to engage visitors in museum tours, due to their ability to interact with humans naturally and familiarly. In this paper, we present our preliminary work on the use of a social robot, Pepper in this case, as an innovative approach to engaging people during museum visiting tours. To this aim, we endowed Pepper with a vision module that allows it to perceive the visitor and the artwork he is looking at, as well as estimating his age and gender. These data are used to provide the visitor with recommendations about artworks the user might like to see during the visit. We tested the proposed approach in our research lab and preliminary experiments show its feasibility.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**; • **Applied computing** → Fine arts; • **Human-centered computing** → *Interactive systems and tools.*

KEYWORDS

Museum visit; Human-Robot Interaction; Computer Vision.

ACM Reference Format:
Giovanna Castellano, Berardina De Carolis, Nicola Macchiarulo, and Gennaro Vessio. 2020. Pepper4Museum: Towards a Human-like Museum Guide. In *Proceedings of AVI^2CH 2020: Workshop on Advanced Visual Interfaces and Interactions in Cultural Heritage (AVI^2CH 2020).* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

In the last few years, due to technology improvements and drastically declining costs, many innovative Information and Communication Technology (ICT) solutions have been applied to the cultural domain, with the aim of making art more accessible and engaging to a wider population [13]. One of the most promising ICT applications in this domain concerns the provision of personalized services, in which the visitor's specific characteristics and preferences are taken into account [5], and recommendation of personalized museum visiting paths [18]. Monitoring what visitors are looking at, what they like or dislike, etc., can be used to personalize and enhance their experience during the visit [31]. Successful applications range from the simple use of mobile apps and interactive displays to virtual and augmented reality settings.

For instance, in [3] an indoor location-aware architecture, which relies on a wearable device to automatically provide the user with cultural content related to the observed artwork, is proposed. A similar approach was followed in [34], where Zhang et al. proposed the use of a wearable camera equipped with image processing capabilities to solve the task of artwork identification within a museum. A remarkable contribution to the topic of personalized visit experience was provided in [5], where Bartolini et al. proposed a *context-aware* recommendation system aimed at supporting intelligent multimedia services for the users. In [1] and [16], the authors explored the possibility to harness electroencephalograph (EEG) signals captured by off-the-shelf EEG low-cost headsets to understand if an artwork is of interest for a visitor. More recently, Cardoso et al. [7] proposed the use of a mobile application to set museum itineraries where visitors can move at their own pace and, at the same time, have all the complementary information they need about points of interest adapted to the user's needs.

A recent solution to engage visitors in museum tours is to use social robots. Social robots are embodied, autonomous agents that communicate and interact with humans on a social and emotional level. They represent an emerging field of research focused on developing a "social intelligence" that aims to maintain the illusion of dealing with a human being [4]. Thanks to their ability to interact with humans in a natural and familiar way, social robots are spreading more and more often into human life not only for entertainment, but also to assist users in their activities of daily living, or in teaching and educational settings. In particular, they can provide novel, interactive social interfaces in cultural and tourism services [33], thus improving the overall experience of the user [28].

Historical examples of museum tour guide robots include RHINO [6] and Minerva [32]. RHINO integrates low-level probabilistic reasoning and high-level problem solving, embedded in first order

AVI^2CH 2020, September 29, Island of Ischia, Italy

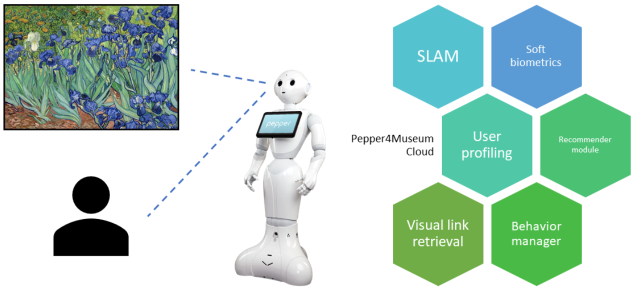

Figure 1: Overview of the system components.

logic, to navigate at high speeds through dense crowds, while reliably avoiding collisions with obstacles. Differently from RHINO, Minerva learns the map from sensor data and presents an improved interaction system with the users. To do this, it adopts a "pervasively probabilistic" approach, which relies on explicit representations of uncertainty in perception and control. More recently, in [17], Germak et al. developed a telepresence robot designed as a tool to explore inaccessible areas of a cultural site. In [2] the humanoid robot Pepper has been used as a tour guide in a museum. Pepper was equipped with several modules, useful for accompanying visitors and interacting with them. Suddrey et al. [30] recently showed that the Pepper's basic functionalities can be improved to enable the robot to provide autonomous and interactive tours.

In this context, the recent advances in Computer Vision are allowing researchers to endow robots with novel and powerful capabilities. In this paper, we present our preliminary work on the use of a social robot, Pepper in this case, as a museum tour guide. In particular, we present a vision-based approach for supporting people during a museum visit. The vision module allows Pepper to perceive the presence of a visitor and localize him in the space, and estimating his age and gender. Moreover, a visual link retrieval module gives Pepper the ability to take the image of the painting observed by the visitor as a visual query to search for visually similar paintings in the museum database. The robot uses these data and other information acquired during the dialog to provide the visitor with recommendations about similar artworks he might like to see in the museum. We tested the proposed approach in our research lab and preliminary experiments show its feasibility.

2 PEPPER4MUSEUM

Designing the behaviors of a social robot acting as a museum guide requires endowing it with different capabilities that would provide visitors with an engaging and effective experience during the visit. These capabilities are meant to allow the robot to detect and localize people in the museum, recognize artworks the visitor is looking at, profile the user during the visit so as to generate suitable recommendations, and finally engage people in the interaction using suitable conversational skills. This is the final aim of the Pepper4Museum project (Fig. 1) which exploits the combination of Computer Vision and Social Robotics.

As robot platform we use Pepper, a semi-humanoid robot developed by SoftBank Robotics. It is an omnidirectional wheeled humanoid robot equipped with several cameras and sensors. In the following, the main modules we are developing for museum visit assistance are briefly described.

2.1 Museum Mapping and Localization

Simultaneous Localization And Mapping (SLAM) is the problem of constructing and updating a map of an unknown environment while keeping track of the robot's location within it. As building maps is one of the fundamental tasks of mobile robots, a lot of researchers focused on this problem. The SLAM algorithms mainly use optical sensor data to reconstruct the map of the environment and determine the orientation and position of the robot. There are two common approaches to SLAM: Visual SLAM, based on data captured from RGB or RGB-D cameras, and LiDAR (Light Detection and Ranging) SLAM, based on data captured from laser sensors. The approach based on LiDAR is typically faster and more accurate than Visual SLAM [15]. As far as concerns the Pepper's capabilities of mapping and navigating in an environment, Pepper is able, using the NaoQI API, to: *i*) map the environment; and *ii*) localize itself and navigate inside the mapped environment in accordance with the SLAM approach. To this aim, Pepper uses its odometry and laser sensors. Thus, as a first step, we developed a module that allows Pepper to map the museum space moving around autonomously. Then, once the mapping has been completed, the resulting map is stored as a 2D image (see Fig. 2a). Successively, the map is annotated with the points of interest close to the artworks' position and each point is tagged with the artwork ID. This ID is then used to retrieve information about the corresponding artwork (i.e., author, description, image, tags).

Besides annotating the points of interest in the space, Pepper has to detect and localize visitors in the mapped space. This is done with the use of a particular deep neural network for object detection. Specifically, we used SSD MobileNetV2 [27], a state-of-the-art deep learning model pre-trained on the MS COCO dataset [22], which is able to detect 80 different objects, including people. Given the frames captured by the Pepper's camera as an input, SSD MobileNetV2 returns a bounding box around all the detected people, as shown in Fig. 2b.

We fused the information about the bounding box of a person in the image with the data captured from the depth camera of the robot in order to compute the coordinates of the visitor in the map previously created with the SLAM algorithm. To determine if a visitor is close to an artwork, we compute the Euclidean distance between the person's point and each point of the artwork. If the distance is less than a threshold, Pepper approaches the visitor.

2.2 Age and Gender Estimation

In order to start gathering information about the target user, a soft biometric module is used [11]. The soft biometric module allows Pepper to automatically infer the *age* and *gender* of the user who is interacting with it. The algorithm follows the approach described in [26], which relies on a fine-tuned version of the VGG16 state-of-the-art deep convolutional neural network [29], using an unconstrained image dataset. The capability of deep neural networks to solve complex perceptual tasks has been shown in several recent works (e.g., [9, 21]). Our approach showed a good performance, as gender recognition reached an accuracy of 85%, while age

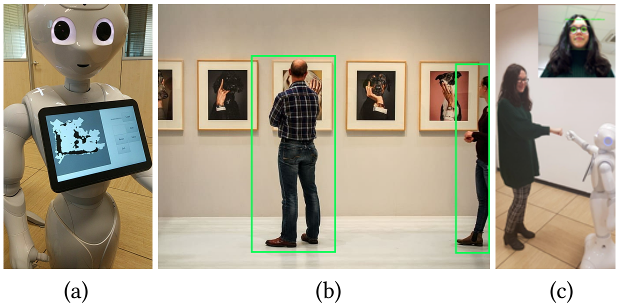

Figure 2: (a) An example of map generated after the exploration of a museum space. (b) People detection. (c) An example of interaction between a young woman and Pepper with real-time gender and age estimation.

estimation reached an accuracy (±1 year) of 84% on the previously mentioned dataset. Soft biometric traits can be used with two main purposes: *i*) improving the recommender module performance by filtering recommendations accordingly; and *ii*) adapting the robot's dialogue to the person it is interacting with. In our museum scenario, we used an approach similar to the one described in [14], in which the robot uses a different level of formality in its dialogue, based on the age and gender of the person being tracked (Fig. 2c).

2.3 User Profiling and Recommender Module

Understanding the user preferences may enable a social robot to adapt its behavior accordingly, hence enhancing the user satisfaction during the interaction [25]. Usually this process is based on explicit feedback, e.g. specific question answering or explicit rating of items. However, this approach, albeit more precise and reliable, is time consuming and requires an effort by the user.

Recent trends use approaches based on implicit feedback that can be inferred by observing and analyzing user's behavior without interrupting the user engagement in the interaction. In the museum visit context, we decided to exploit a hybrid approach that combines observations of the user behavior with explicit questions asked by the robot during the interaction. Then, thanks to the soft biometric analysis, information about user gender and age is used as a feature for triggering an initial stereotypical model for the visitor [24]. Moreover, these data can be used to tailor the dialog and the information presented to the visitor (i.e., descriptions provided to a child will be different from those provided to an adult). In addition, these data, together with information about what is of interest for the visitor (inferred by observing what the visitor is looking at and by answering to specific questions during the visit), can be used to trigger the recommendation about what to see next in the museum. This phase represents an active process of feedback and preference acquisition that allows the robot to acquire new information that can be used for refining subsequent recommendations.

2.4 Visual Link Retrieval

The proposed module for link retrieval assumes that the robot has knowledge about the artworks exhibited in the museum. The goal is to project the raw pixel images into a new, numerical feature space in which to search for similarities among paintings [10]. These similarities can be used to provide semantic links among paintings so as to recommend artworks a visitor may be interested in.

The proposed method is mainly based on "visual attributes" automatically learned by a VGG16-based model. The resulting high dimensional representation is then embedded in a more compact feature space by applying Principal Component Analysis. Finally, similarities among paintings, i.e. visual links, are obtained through a distance measure in a completely unsupervised Nearest Neighbor fashion. The proposed method thus provides the nearest neighbors for each query image, that are those images more similarly linked to the input query. Relying on a completely unsupervised approach makes the proposed method simple and practical, as it excludes the necessity to acquire labels of visual links, which can be unavailable or very difficult to collect.

2.5 Behavior Manager

A behavior is a program that combines and coordinates the utterances, gestures, expressions, touch-screen interactive elements, and locomotion based on the current robot perceptions. In the context of museum visiting, the behavior manager module can trigger a particular behavior according to two approaches: *reactive* and *proactive*.

In the reactive case, the triggered behavior is an answer to the recognized user's intent. In particular, user input can be provided via voice or through a touch screen. In the first case, the input is processed by the automated speech recognition module that is already encoded in the programming environment of the robot. In the second case, the user can interact with the tablet of the robot, which shows the available choices. It is worth stating that the touch screen is needed, since the overall quality of the speech recognition module encoded in Pepper is typically low. At the current stage of implementation of the system, five families of intents are captured: *greet, small_talk*, *current_painting information*, *suggestions*, and *tour*. Clearly, each intent invokes a different, more or less complex, behavior as an answer. In the proactive case, a rule-based system, in which the current state of the perception is periodically matched with the preconditions of rules, has been adopted. Then, the behavior associated to the selected rule is executed. If no rule is selected, the robot executes the idle behavior in which it moves a bit around, randomly displaying on its tablet artworks present in the museum exhibition with the invitation to ask about them. Each behavior may require the fulfillment of a service execution in the cloud of Pepper4Museum as in the case of the recommendation generation.

3 PRELIMINARY STUDY

As a proof of concept, we preliminarly investigated the effectiveness of the different modules embedded in Pepper4Museum.

About the gender and age estimation, the soft biometric module in the wild was able to recognize with 87.5% accuracy the gender of the visitors and with 62.5% accuracy the age of the visitors (more details in [8, 14]). The classification of the age was lower than we expected because the interaction in the wild did not guarantee a static and frontal position of the user with respect to the camera. Also the variation of lighting conditions influenced the analysis of the face for age estimation.

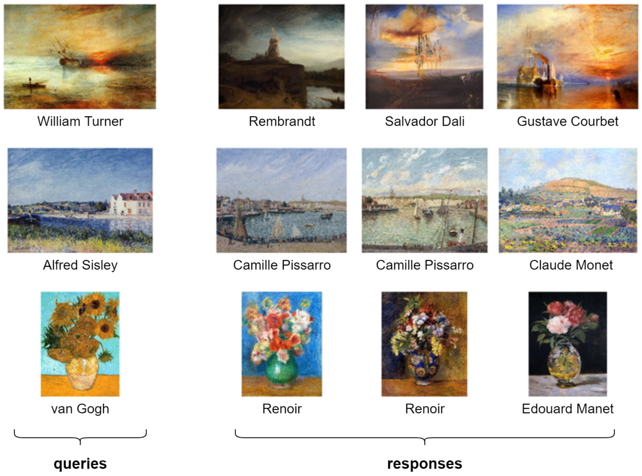

Figure 3: Sample artwork queries and corresponding visually linked paintings provided by the system.

Then, we tested the visual link retrieval method on a database collecting paintings of 50 very popular painters. We used data provided by the Kaggle platform,[1] scraped from an art challenge Website.[2] Artists belong to very different epochs and painting schools, ranging from Giotto di Bondone and Renaissance painters such as Leonardo da Vinci and Michelangelo, to Modern Art exponents, including Pablo Picasso, Salvador Dalí, and so on. Once the reduced features representing paintings were obtained, we applied the Nearest Neighbor matching mechanism to derive, for each query image, the top k matching images ($k = 3$ in this case). To give an illustrative example of the behavior of our system, in Fig. 3 we provide three sample image queries, together with the corresponding top visually linked artworks retrieved by the system. For each query, a brief description of the results is given below:

- Q1 The first image query is the Romanticist “Fort Vimieux” by William Turner, depicting a classic red sunset of the author. It can be seen that the system was able to retrieve paintings similar both in content and color distribution.
- Q2 The second query is the Impressionist “Confluence of the Seine and the Loing” by Alfred Sisley. It can be noticed that the three neighbors, i.e. two artworks by Camille Pissarro and a work by Claude Monet, share the same painting style, characterized by the typical color vibration.
- Q3 Finally, we considered as query a version of the “Sunflowers” series by Vincent van Gogh. As expected, the 3-top images retrieved by the system represent still lifes, two of them by Renoir, the other one by Edouard Manet.

Based on a qualitative evaluation of the retrieval results, we can conclude that, overall, the proposed system is able to find visual links that are not in contrast with the human perception. The visual links discovered by the system are sufficiently justifiable by a human observer and in most cases resemble the intrinsic criteria humans adopt to link visual arts. These criteria combine visual elements, such as colors and shapes, and conceptual elements, such as subject matter and meaning of the painted scene.

The mapping and localization process could not be tested in a real museum due to the COVID-19 emergency. We tested this module in the “Museum of History of Computers” located in our department and we observed that its performance was overall acceptable. Some delay was registered when Pepper found an unexpected obstacle on its planned path (e.g., people crossing). The other modules need to be tested in the wild as soon as it will be possible.

4 CONCLUSION AND FUTURE WORK

In this paper, we have presented our preliminary work towards the development of Pepper4Museum: a human-like museum guide. Promising results in our research lab have been obtained. As future work, we plan to test and refine all the behaviors we implemented in this domain. Then, in order to run an experiment in a real museum context, a test on the integration of the described components is needed. A test in a museum will allow for measuring the visitor experience and evaluating the impact of this technology in this context. Finally, it is worth remarking that the data collected by Pepper represent a valuable source of information that can be profitably used to better understand and predict the visitors’ behavior [18, 19, 23]. Such an analysis could be carried out by means of graph theory, e.g. [20], or process mining techniques [12].

ACKNOWLEDGMENTS

Funding for this work was partially provided by Fondazione Puglia that supported the Italian project “Programmazione Avanzata di Robot Sociali Intelligenti”. Gennaro Vessio acknowledges funding support from the Italian Ministry of Education, University and Research through the PON AIM 1852414 project.

REFERENCES

[1] F. Abbattista, V. Carofiglio, and B. De Carolis. 2018. BrainArt: a BCI-based Assessment of User’s Interests in a Museum Visit.. In *AVI*CH*.

[2] D. Allegra, F. Alessandro, C. Santoro, and F. Stanco. 2018. Experiences in Using the Pepper Robotic Platform for Museum Assistance Applications. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 1033–1037.

[3] S. Alletto et al. 2015. An indoor location-aware system for an IoT-based smart museum. *IEEE Internet of Things Journal* 3, 2 (2015), 244–253.

[4] M. Alqaderi and A. Rad. 2018. A Multi-Modal Person Recognition System for Social Robots. *Applied Sciences* 8 (03 2018), 387. https://doi.org/10.3390/app8030387

[5] I. Bartolini et al. 2016. Recommending multimedia visiting paths in cultural heritage applications. *Multimedia Tools and Applications* 75, 7 (2016), 3813–3842.

[6] W. Burgard et al. 1998. The interactive museum tour-guide robot. In *AAAI/IAAI*. 11–18.

[7] P.J.S. Cardoso et al. 2019. Cultural heritage visits supported on visitors’ preferences and mobile devices. *Universal Access in the Information Society* (2019), 1–15.

[8] B. De Carolis, N. Macchiarulo, and G. Palestra. 2018. A Comparative Study on Soft Biometric Approaches to Be Used in Retail Stores. In *Foundations of Intelligent Systems - 24th International Symposium, (ISMIS2018) (Lecture Notes in Computer Science)*, M. Ceci, N. Japkowicz, J. Liu, G.A. Papadopoulos, and Z.W. Ras (Eds.), Vol. 11177. Springer, 120–129.

[9] G. Castellano, C. Castiello, C. Mencar, and G. Vessio. 2020. Crowd Detection for Drone Safe Landing Through Fully-Convolutional Neural Networks. In *International Conference on Current Trends in Theory and Practice of Informatics*. Springer, 301–312.

[10] G. Castellano and G. Vessio. 2020. Towards a Tool for Visual Link Retrieval and Knowledge Discovery in Painting Datasets. In *Italian Research Conference on Digital Libraries*. Springer, 105–110.

[11] A. Dantcheva, C. Velardo, A. D’angelo, and J.-L. Dugelay. 2011. Bag of soft biometrics for person identification. *Multimedia Tools and Applications* 51, 2 (2011), 739–777.

[1] https://www.kaggle.com/ikarus777/best-artworks-of-all-time
[2] http://artchallenge.ru

[12] B. De Carolis, S. Ferilli, and D. Redavid. 2015. Incremental Learning of Daily Routines as Workflows in a Smart Home Environment. *ACM Trans. Interact. Intell. Syst.* 4, 4, Article 20 (Jan. 2015), 23 pages.

[13] B. de Carolis, C. Gena, T. Kuflik, and J. Lanir. 2018. Special issue on advanced interfaces for cultural heritage. *International Journal of Human-Computer Studies* 114 (03 2018).

[14] B. De Carolis, N. Macchiarulo, and G. Palestra. 2019. Soft Biometrics for Social Adaptive Robots. In *Int. Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 687–699.

[15] M. Filipenko and I. Afanasyev. 2018. Comparison of Various SLAM Systems for Mobile Robot in an Indoor Environment. In *2018 International Conference on Intelligent Systems (IS)*. 400–407.

[16] C. Gena, C. Mattutino, S. Pirani, and B. De Carolis. 2019. Do BCIs Detect User's Engagement? The Results of an Empirical Experiment with Emotional Artworks. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP'19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 387–391.

[17] C. Germak, M.L. Lupetti, L. Giuliano, and M.E.K. Ng. 2015. Robots and cultural heritage: New museum experiences. *Journal of Science and Technology of the Arts* 7, 2 (2015), 47–57.

[18] T. Kuflik, Z. Boger, and M. Zancanaro. 2012. Analysis and Prediction of Museum Visitors' Behavioral Pattern Types. *Cognitive Technologies* (04 2012).

[19] J. Lanir, T. Kuflik, J. Sheidin, N. Yavin, K. Leiderman, and M. Segal. 2016. Visualizing museum visitors' behavior: Where do they go and what do they do there? *Personal and Ubiquitous Computing* (11 2016).

[20] E. Lella and E. Estrada. 2020. Communicability distance reveals hidden patterns of Alzheimer disease. *Network Neuroscience* Just Accepted (2020), 1–38.

[21] E. Lella and G. Vessio. 2020. Ensembling complex network 'perspectives' for mild cognitive impairment detection with artificial neural networks. *Pattern Recognition Letters* (2020).

[22] T.-Y. Lin et al. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[23] C. Martella et al. 2017. Visualizing, clustering, and predicting the behavior of museum visitors. *Pervasive and Mobile Computing* 38 (2017), 430–443.

[24] E. Rich. 1998. *User Modeling via Stereotypes.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 329–342.

[25] S. Rossi, F. Ferland, and A. Tapus. 2017. User Profiling and Behavioral Adaptation for HRI. *Pattern Recogn. Lett.* 99, C (Nov. 2017), 3–12.

[26] R. Rothe, R. Timofte, and L. Van Gool. 2015. Dex: Deep expectation of apparent age from a single image. In *Proc. of the IEEE international conference on computer vision workshops*. 10–15.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4510–4520.

[28] J. et al. Santos. 2018. A Personal Robot as an Improvement to the Customers' In-Store Experience. *Service Robots* (2018), 1.

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

[30] G. Suddrey, A. Jacobson, and B. Ward. 2018. Enabling a pepper robot to provide automated and interactive tours of a robotics laboratory. *arXiv preprint arXiv:1804.03288* (2018).

[31] A. Tavčar, A. Csaba, and E. V. Butila. 2016. Recommender system for virtual assistant supported museum tours. *Informatica* 40, 3 (2016).

[32] S. Thrun et al. 2000. Probabilistic algorithms and the interactive museum tour-guide robot minerva. *The International Journal of Robotics Research* 19, 11 (2000), 972–999.

[33] V. Tung and R. Law. 2017. The potential for tourism and hospitality experience research in human–robot interactions. *International Journal of Contemporary Hospitality Management* 29 (08 2017), 00–00. https://doi.org/10.1108/IJCHM-09-2016-0520

[34] R. Zhang, Y. Tas, and P. Koniusz. 2018. Artwork identification from wearable camera images for enhancing experience of museum audiences. *arXiv preprint arXiv:1806.09084* (2018).