Computer Vision and AI Tools for Enhancing User Experience in the Cultural Heritage Domain

Pavan K. Rachabathuni, Paolo Mazzanti, Filippo Principi, Andrea Ferracani, and Marco Bertini(✉)

Università degli Studi di Firenze - DINFO - MICC, Florence, Italy
{PavanK.Rachabathuni,Paolo.Mazzanti,Filippo.Principi,
Andrea.Ferracani,Marco.Bertini}@unifi.it
http://www.micc.unifi.it

Abstract. To enhance the museum experience for visitors, it's crucial to adopt a people-centered approach. This means engaging with visitors dynamically, providing transformative learning experiences from pre-arrival to post-departure. Using interactive and immersive experiences can evoke emotions, inspire creativity, and facilitate participatory learning. Incorporating advanced technologies like artificial intelligence (A.I.) and computer vision (C.V.) can deepen the connection between artworks and visitors, resulting in a "phygital museum." This helps to reimagine museums with diverse spaces, tools, and interactions for various audiences. AI/CV tools help visitors interact with collections, inspiring insights and enriching data [15].

According to the "contextual mode" of learning proposed by Falk and Dierking [7], the more interactive and participatory the museum experience is, the more it will have a lasting impact on the visitor and thus improve his or her learning on the subject. Visitors are driven by their motivations, emotions, interests, and prior knowledge, and this background has a huge influence on the experience the visitor gets from the stay, including the memory they will have of it and recommendations to other users and friends. Emotions play a motivating role at every stage of the museum experience, from the decision to visit the museum to post-visit enjoyment, and visitors' emotional engagement is a prerequisite for effective and authentic learning. Digital tools can enhance the museum experience by triggering a multisensory experience, enabling physical connections, personalizing the visit to the user's needs, enhancing playful and social interaction, and sharing personal experiences.

In this paper we report on the design and functionalities of a set of CV and AI tools designed to enhance the user experience in cultural heritage institutions like museums. The tools are open source.

Keywords: Computer Vision · User Experience · Gamification · Phygital Museum

P. K. Rachabathuni, P. Mazzanti and F. Principi—Equal contribution.

P. Zaphiris et al. (Eds.): HCII 2024, LNCS 15378, pp. 345–354, 2025.
https://doi.org/10.1007/978-3-031-76815-6_25

1 Introduction

Contemporary literature highlights the importance of engaging audiences in dynamic experiences that enable visitors to enjoy, learn, discover, and interact [7,15]. To achieve this, it is necessary to create new forms of engagement that cater to diverse audiences, with particular attention to younger and local communities. Utilizing a range of digital tools, cultural heritage sites can motivate audiences with an interactive and human-centered approach. Visitors are increasingly familiar with using digital tools and new languages to interact with museum collections, creating and sharing new stories before, during, and after their visit. New technologies, particularly AI/CV tools, hold great potential for motivating and creating a stronger relationship between artworks and visitors. When applied in a playful approach based on gamification and learning-by-doing techniques, these tools can foster interaction and emotional relationships between visitors and artworks, making the content more memorable [4].

Within the ReInHerit project[1], a toolkit has been developed using A.I. and C.V. technologies. The main goal is to allow visitors to interact with collections and exhibitions playfully, using different approaches like gamification, or asking visitors to interact with chat-based systems, or to inspect artworks using virtual magnifying lenses. This encourages visitors to explore artwork content more deeply during and after their visit, creating new narratives and, in some cases, also producing user-generated content for social media sharing, as a memento of the visit. Smaller museums often lack resources for innovative digital services, and only larger institutions can develop applications using A.I. and computer vision tools. Making the toolkit open source aims to bridge this technological gap.

The ReInHerit toolkit has been designed for young visitors who are more likely to use digital tools in a museum setting. The emphasis is on creating mobile-first and web-first applications, prioritizing mobile devices. This approach makes it easier to implement a Bring-Your-Own-Device (BYOD) approach for seamless integration. For this reason, the interfaces and interactions have been designed to be usable both on small and larger screens, to account also for the case in which the organizations prefer to provide the experience through installations, instead of mobile apps.

The strategy used to design the toolkit is based on the following pillar topics:

- Interactivity: to enhance visitor engagement, motivate learning, create a stronger relationship between artworks and visitors;
- BYOD: mobile device applications have been specifically developed as primary targets, simplifying adherence to the Bring-Your-Own-Device (BYOD) approach;
- Artificial Intelligence and Computer Vision: to develop new types of human-computer interactions;

[1] https://www.reinherit.eu.

– Open-Source: to align with a sustainable management perspective, streamlining maintenance and facilitating the reuse of applications by different organizations.

In particular, in this paper we report on the following applications:

– Strike-a-Pose and Face-fit: these two applications are designed to employ gamification and interaction with an artwork to increase the engagement of the visitors of a museum: visitors are asked to replicate an artwork using the pose of their body pose and their facial expression and pose; CV is used to automatically evaluate the similarity to the pose of the artwork. User-generated materials are provided for social media sharing and as mementos of the museum visit, where the user is depicted replicating the poses or painted with his own expression within an artwork.
– Smart Lens: this app transforms a mobile phone in a virtual magnifying lens to observe the details of an artwork and getting the related information; the user is asked to observe more proactively an artwork and think about its parts. Computer vision is used to recognize the relevant artwork details selected by the exhibition curators to provide the associated multimedia information.
– Multimedia Chatbot: this system merges CV and natural language processing capabilities of Multimedia Language Models (MLM) to implement a chatbot usable in web/mobile interfaces; the system also provides speech-to-text functionalities to allow for a more natural interaction during a visit. The goal is to provide descriptions and information on artworks using natural language and interacting in a chat, thus again eliciting a proactive interaction from the user, answering his real interests, and using a type of interaction that has become common especially for younger users.

2 The Toolkit

In the following sections we are going to present the applications developed and published in the Toolkit. Source code and documentation of all the applications is available on the ReInHerit Digital Hub[2]. In Sect. 2.1 we describe two applications implemented following the paradigm of serious gaming (Strike a Pose and Face-Fit). In Sect. 2.2 we face the problem of providing information about details of artworks through object detection on mobile devices. In Sect. 2.3 we present a system (i.e. Viola Chatbot) which engages users on artworks through an AI-powered conversational chat-bot.

2.1 Strike-a-Pose and Face-Fit

Gamification is the process of exploiting strategies and game dynamics into scenarios that are not a game. It has already been proved to be useful to enhance

[2] https://reinherit-hub.eu/.

skills and competences in a variety of domains such as marketing, industry training and entertainment. Certainly also cultural heritage can benefit from a gamification approach which represents an opportunity to engage visitors to museums content through the design of more entertaining, social and challenging digital learning scenarios [3,9,13], to help museums to move from the traditional "look and do not touch" toward a "play and interact" approach. In fact, it has been observed that the availability of tools like gamified e-guides to visitors contributes to the sustainability of museums [2]. The two applications presented in this section use gamification as a strategy to elicit an emotional response from users and to trigger a learning process. More detials regarding the design is provided in [5].

Strike-a-Pose (Fig. 1) is a web-based application that enables users to compare and evaluate their own poses to those present in famous paintings or statues. The system can be accessed via visitors' smartphones using the "Bring Your Own Device" (BYOD) approach or through a dedicated environment that uses a fixed station equipped with a large screen and a camera. The application uses a gamification paradigm that makes the process of learning about works of art fun and engaging for users. After registering, users are challenged to replicate the poses of selected artworks from the museum's collections. The application provides users with skeleton displays of both the artwork and their own bodies to facilitate matching various points and segments. The application presents users with sets of challenges that require replicating poses using different body parts, such as using only the torso to allow wheelchair users to participate. Once all the poses have been matched, the application allows users to generate a video of the interactive experience, which can be saved for social sharing.

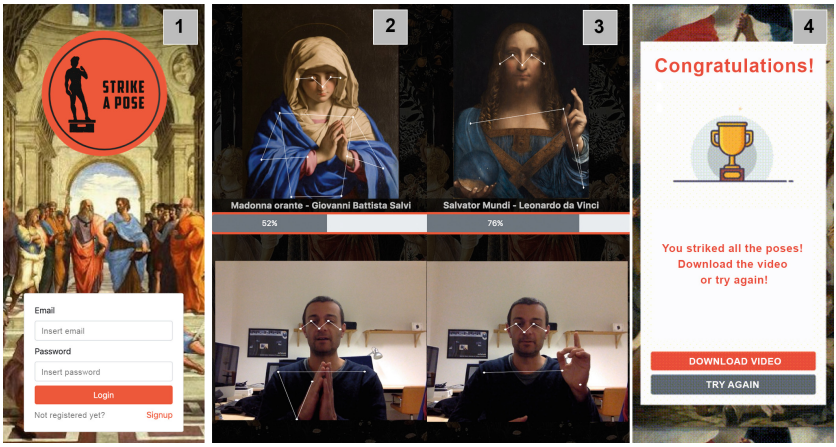

Fig. 1. Strike a Pose App for smartphone. 1) Login. 2–3) The user trying to strike the pose in the painting (playing in "easy" mode, with visible skeletons). 4) Challenge completed: download the video.

The application was developed using JavaScript on the client-side and Python on the server-side. Pose detection on human bodies is achieved using TensorflowJS detection API[3], which exploits the pose detection model, MoveNet. MoveNet is a fast and accurate model that detects 17 key points of a body. The model is used in the variant "Lightning" for latency-critical applications, running faster than real-time on most modern desktops, laptops, and phones. The model runs completely client-side in the browser. Server-side, an SQLLite database is used to store artwork collections, challenges, and artwork metadata and descriptions. Communication between the knowledge-base and the interface is ensured through RESTful APIs developed in Flask[4]. The video is created server-side. The interface designed for smartphones uses a vertical layout, while the one for installations has a horizontal layout, to accommodate different screen aspect ratios.

To ensure that the coordinates of the keypoints were not related to the entire image but only to the pose, both for the user's pose and the pose of the character in the painting, they have been normalized. The algorithm used to test the similarity of the user's pose to the pose of the character in the painting is based on the Euclidean distance between the set of points. The match is verified if the distance is below a predefined threshold for a certain time interval.

Face-Fit (see Fig. 2) is a software application developed in JavaScript and Python that allows users to personalize and gamify famous paintings, particularly portraits. The app is designed for both smartphones and desktop installations in museums. It lets users replicate the head pose and expression of portraits by famous painters and transfer their own face onto the artwork to create a new image. The development of the app was informed by a usability study that followed an iterative design approach with three groups of five people [12].

To use the app, the user stands in front of a camera-equipped smartphone or installation and chooses a painting from a vertical carousel of portraits. The app presents a ghost image of the user's face over the painting, which the user must try to superimpose perfectly. The ghost image was introduced to solve issues related to keeping the user focused on the task while still making the game fun. Initially, visual suggestions were provided to help users find the right pose, but these were found to distract users from the painting and the game.

The app uses the Mediapipe library[5], which employs Tensorflow Lite technology, to detect facial landmarks in real-time, even on mobile devices [8]. The landmarks are used to compute the rotation angles of the user's head with respect to the camera, which are then compared to the pose of the face in the painting. The match is verified when the Euclidean distance of the 3D rotation angles is below a predefined threshold. A second check is then performed to compare the shapes of the three most significant features of the human face - eyes, eyebrows, and mouth - detected in both images. These features convey the seven universal facial expressions as identified by Ekman *et al.* [6]. Image matching and face-

[3] https://www.tensorflow.org/js.
[4] https://flask.palletsprojects.com/.
[5] https://google.github.io/mediapipe/.

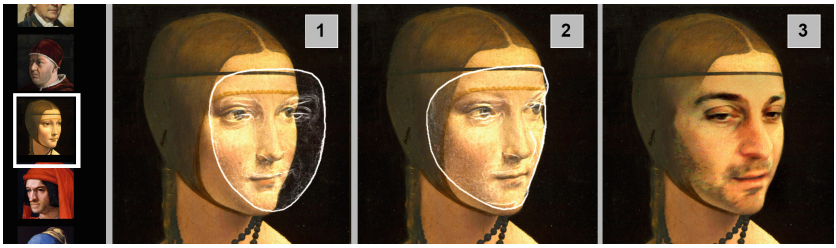

Fig. 2. Face-Fit App for museum installation. 1) trying to match the pose and expression of Leonardo's *Lady with an Ermine*. 2) Face matched. 3) Challenge completed: image generated.

swapping between the user's face and the painting is achieved through affine transformation to corresponding mesh triangles.

To adapt the style of the user's image to that of the painting, the app uses color correction based on statistical analysis by Reinhard *et al.* [14]. Once the user completes a challenge, they can obtain the generated images for sharing on social media and unlock descriptions of the matched artworks.

2.2 Smart Lens

Smart Lens (see Fig. 3) is a mobile application with the purpose to automatically recognize details of artworks. The user can employ his mobile phone (or tablet) as if it was a lens to inspect and analyze details of interest, that are recognized using a variety of computer vision techniques such as Content-based Image Retrieval (CBIR), classification, or object detection. As details are recognized, the app provides multimedia information about them. The app can be used as a smart guide to a collection or for a set of artworks that are of particular relevance and require thorough analysis. An additional benefit is that it is possible to collect anonymized information on the details and artworks of interest, providing curators with statistics and information on how the exhibitions are experienced by visitors.

The application allows the recognition of artwork details using three alternative methods: Content-based Image Retrieval (CBIR), classification, and object detection.

Content-based Image Retrieval (CBIR): the CBIR mechanism allows the recognition of the works and their details through the comparison between the visual features extracted from a dataset of images (the collection of details stored in a database, i.e., in the guide) and those obtained from the frame produced by the user's camera. To allow the recognition of the details of the works, the features have been extracted not only from the entire work but also from some of its parts obtained through a rigid subdivision.

Classification: this approach allows the recognition of the framed detail/work through a classification network on which a fine-tuning operation

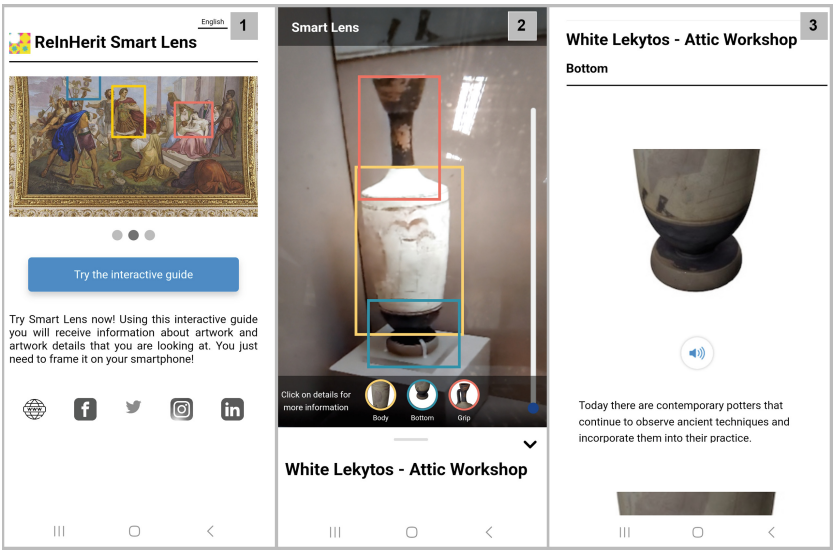

Fig. 3. Smart-Lens App for smartphone. 1) Home. 2) The user frame the artwork with the device camera. Bounding boxes of the details and image details previews are shown. 3) After the selection of a detail the user can hace insights about it and listen to the audioguide.

has been performed on the paintings and their details present in the dataset. This requires retraining of the neural network, typically achievable even using CPUs. In this version, the classification of the entire frame is performed, and it is identified with the most probable class only if the degree of confidence is greater than a specific threshold for each image.

Object Detection: in this version, the recognition is entrusted to a network for object detection on which fine-tuning has been performed on the details of the work. This retraining is more expensive than the previous one and requires a GPU. In this case, only the details within the work are identified by taking the most probable bounding-box (i.e., a rectangle that most probably contains the object of interest) for each class provided that its degree of confidence exceeds a specific threshold for each detail.

The interface is entirely made in HTML5 and CSS and has a fluid behavior adapting to the size of any screen (tablet or smartphone). It is dynamically modified through JavaScript scripts that allow you to show the results of the matches with the images framed by the camera in real-time. In the camera view, the work or detail is framed, and the recognition operation is carried out in real-time. Tensorflow[6] networks were used for Retrieval and classification, and SSD/MobileNetV3 for Detection. For each work or detail, a textual description, the relative image, and any audio and video content are provided. If an audio

[6] https://www.tensorflow.org/.

file is not available, the application will allow the text-to-speech of the textual information shown on the screen, allowing, in any case, the use of the application as an audio guide.

The application can be used to perform artwork recognition, without considering the details. To use the application, a server is needed to host the backend and the web app itself. A QR code can be used to avoid typing the URL of the web app; this approach has been used during a set of exhibitions organized by the Museum of Cycladic Art in Athens, The Graz Museum and the Museum of the Bank of Cyprus Cultural Foundation, that used Smart lens to let visitors of the exhibitions to learn more in depth about details of selected artworks.

2.3 Multimedia Chatbot: Viola

Viola is a new mobile and web application that has been developed to provide a unique form of interaction through a chatbot on cultural heritage subjects. The web app allows users to ask questions about the content and context of artworks using natural language, either by typing or using voice-to-text input. This type of interaction has gained popularity recently with the emergence of chat systems like ChatGPT. This task can be considered as a form of Visual-Question Answering (VQA).

The chatbot has been designed to use two different types of backends: a version is based on GPT -based LLMs, while a second newer version employs Multimodal Language-and-Image Models (MLLMs). In particular we have developed a MLLM that extends a LLaVa model [11] with GLIP [10], and uses Retrieval Augmented Generation (RAG) to reduce hallucinations in the answer, so to produce scientifically accurate answers using curated artwork information, providing broader historical and cultural context.

To train the MLLM we have developed a new dataset, based on the VISCOUNTH [1] dataset, that provides approximately 500K images of Italian cultural assets, including paintings, statues, finds, prints, and churches, along with 6.5M question-answer pairs. The issue with this dataset is that these questions are based on predefined templates applicable to different cultural assets, thus limiting the types of questions, and the answers often tend to repeat the same words used in the questions. Therefore training a MLLM on this dataset limits the model's ability to learn the context of the cultural assets. To address these limitations, we have thus created a new dataset developed from VISCOUNTH. In this novel dataset we have created about 55,000 question-answer pairs related to 2,890 images of cultural assets images. The question-answer pairs have been created using the capabilities of GPT-4.

An example of the app is shown in the figure below (Fig. 4). The app is web-based and consists of a backend system where the visual question answering system is executed to enable chat-based functionalities, and an HTML5-based frontend interface for end users.

The system can be easily reused by modifying the database information, including images of the gallery and textual information related to the visual and contextual aspects of the artworks. A backend functionality allows museum

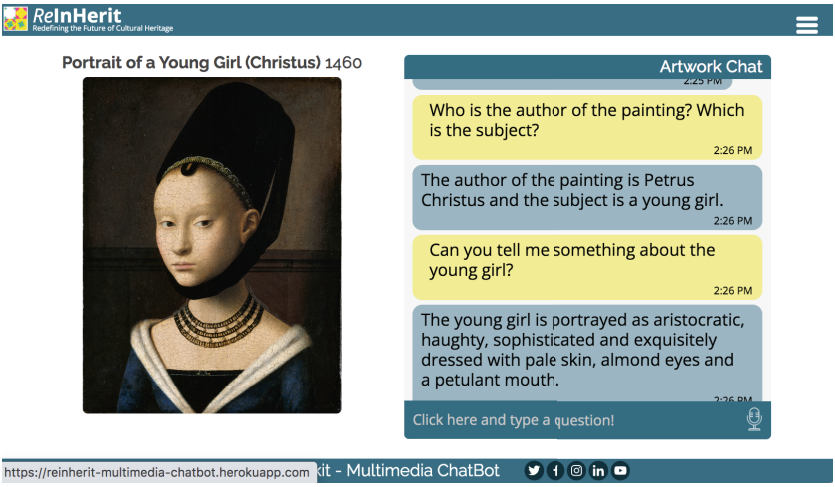

Fig. 4. Multimedia Chatbot Viola answering some questions about a portrait

curators to manage collections aand to add the descriptions used in the RAG functionality of the chatbot to ensure the accuracy o the answers provided by it. Furthermore, VIOLA can serve as a foundation for more complex systems, such as chat-based retrieval systems, serious games, or other interactions based on natural language processing.

3 Conclusions

In conclusion, the ReinHerit project and its Toolkit model confirm and reinforce the concept of an innovative interactive museum that focuses on visitors and the use of digital tools throughout the entire visitor journey. The museum is considered as a generator of unique and informal educational experiences that attract diverse audiences and stimulate their learning. Museum studies emphasized the potential of Digital Learning to offer museums new educational approaches, employing various digital technologies for content mediation, storytelling, and engagement with different visitor types and styles.

The ReinHerit Toolkit model focuses on visitor-centered interactive experiences facilitated by innovative digital tools, such as AI and computer vision. It emphasizes the promotion of emotional and playful digital learning and encourages visitors to use their own smartphones (BYOD) through web applications to enhance their interaction throughout the visitor experience.

Acknowledgments.. This work was partially supported by the European Commission under European Horizon 2020 Programme, grant number 101004545 - ReInHerit.

Disclosure of Interests. The authors have no competing interests.

References

1. Becattini, F., et al.: VISCOUNTH: a large-scale multilingual visual question answering dataset for cultural heritage. ACM Trans. Multimedia Comput. Commun. Appl. **19**(6), 1–20 (2023)
2. Bieszk-Stolorz, B., et al.: Impact of the availability of gamified e-guides on museum visit intention. Procedia Comput. Sci. **192**, 4358–4366 (2021)
3. Bonacini, E., Giaccone, S.C.: Gamification and cultural institutions in cultural heritage promotion: a successful example from Italy. Cultural Trends **31**(1), 3–22 (2022). https://doi.org/10.1080/09548963.2021.1910490
4. Del Bimbo, A.: Emotions in digital. In: of European Museum Organisations, N. (ed.) Emotions and Learning in Museums, pp. 24–30. Network of European Museum Organisations, online (2021)
5. Donadio, M.G., Principi, F., Ferracani, A., Bertini, M., Del Bimbo, A.: Engaging museum visitors with gamification of body and facial expressions. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 7000–7002. MM '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3503161.3547744
6. Ekman, P., et al.: Universals and cultural differences in the judgments of facial expressions of emotion. J. Pers. Soc. Psychol. **53**(4), 712 (1987)
7. Falk, J.H., Dierking, L.D.: The Museum Experience Revisited. Left Coast Press Inc. (2012)
8. Kartynnik, Y., Ablavatski, A., Grishchenko, I., Grundmann, M.: Real-time facial surface geometry from monocular video on mobile GPUS. In: CVPR Workshop on Computer Vision for Augmented and Virtual Reality 2019. Long Beach, CA (2019). https://arxiv.org/abs/1907.06724
9. Khan, I., Melro, A., Amaro, A.C., Oliveira, L.: Systematic review on gamification and cultural heritage dissemination. J. Digit. Media Interact. **3**(8), 19–41 (2020)
10. Li*, L.H., et al.: Grounded language-image pre-training. In: CVPR (2022)
11. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
12. Nielsen, J.: Iterative user-interface design. Computer **26**(11), 32–41 (1993)
13. Paliokas, I., et al.: A gamified augmented reality application for digital heritage and tourism. Appl. Sci. **10**(21) (2020). https://doi.org/10.3390/app10217868, https://www.mdpi.com/2076-3417/10/21/7868
14. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Comput. Graphics Appl. **21**(5), 34–41 (2001)
15. Villaespesa, E., Murphy, O.: This is not an apple! Benefits and challenges of applying computer vision to museum collections. Museum Manag. Curatorship **36**(4), 362–383 (2021). https://doi.org/10.1080/09647775.2021.1873827