Mobile Museum Visitors Guide based on an Indoor Localization System using Deep Learning-based Image Recognition

Bashar Egbariya[1,*], Tsvi Kuflik[1] and Ilan Shimshoni[1]

[1]The Information Systems Department The University of Haifa, Hanamal St. 65, Haifa, 3303220 ,Israel

Abstract

Indoor positioning was and still is a challenge for museum visitors' guide systems aiming at providing context aware and personalized services to their users. There are trivial solutions that require either explicit user actions or the installation of indoors positioning system in the museum. We propose a simple and easy to implement alternatives that is based on image-based positioning.

Keywords

image-based indoor positioning, indoor positioning in cultural heritage, indoor positioning

1. Introduction

Indoor navigation systems are essential for cultural heritage establishments, especially for big and complex ones, including Museums. It is required since visitors need to be able to navigate and find their way in large museum spaces, and also for a museum visitors' guide systems to know the accurate visitor position, as well as orientation in order to deliver relevant information to him/her. Over the years it has been a challenge to navigate big establishments, even if signs and directions were provided. It used to be time consuming and hard to navigate through a building especially for newcomers and visitors. In recent years, with technological advancements, we experienced the use of external sensors, Wi-Fi, Bluetooth and others to assist with localization and navigation. These techniques have proven to be effective but some of them are expensive and/or require constant maintenance, which in turn does not lend itself to widespread adoption [1]. Alternative solutions of object identification using image processing were suggested as well [2], still, having images of all museum's objects while may be technically feasible, does not seem to be practical, given the number of artefacts available in a museum. Hence, we will explore the viability of applying computer vision techniques to better derive the location and orientation of a visitor without using any other external sensors and/or beacons. We formulate the following research question: How can we harness feature extraction models to facilitate image matching within a sustainable images database, with the goal of achieving precise user localization and positioning in indoor environments for context aware information delivery by analyzing a continuous image stream depicting the user's trajectory? Our target application is a mobile visitors' guide system for a museum where the visitors' location is determined seamlessly as the visitor walks freely in the museum. Using machine-vision-based positioning, which yields location and orientation, visitors will be provided proactively with relevant information. For our experiments we will be using the Hecht Museum at the University of Haifa, as our controlled experimental environment.

CEUR-WS.org/Vol-3798/paper5.pdf

Workshop on Advanced Visual Interfaces and Interactions in Cultural Heritage (AVICH 2024), June 4th, 2024, co-located with the 17th ACM Conference on Advanced Visual Interfaces, Arenzano (Genoa), Italy.

*Corresponding author.

e.bashar.t@gmail.com (B. Egbariya); tsvikak@is.haifa.ac.il (T. Kuflik); ishimshoni@is.haifa.ac.il (I. Shimshoni)

https://tsvikak.hevra.haifa.ac.il/ (T. Kuflik); https://sites.google.com/is.haifa.ac.il/ilan-shimshoni/home (I. Shimshoni)

0009-0001-8337-9030 (B. Egbariya); 0000-0003-0096-4240 (T. Kuflik); 0000-0002-5276-0242 (I. Shimshoni)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

2. Planned research

The goal of the research is to explore vision-based positioning techniques as intuitive positioning and orientation detectors for a museum visitors' guide system. Following the POC of vision based positioning demonstrated by the ARDIF system [3], this study aims to investigate, in a realistic setting, the feasibility of leveraging computer vision techniques in conjunction with a smartphone to enhance location determination, eliminating the need for external sensors or beacons, as part of a museum visitors' guide system. Our proposed indoor navigation system comprises two integral components: an Android application and a back-end API. The Android app serves as a user-centric tool, continuously streaming images of the user's real-time path within the museum. Meanwhile, the back-end API functions as a matching service, ingesting images from the Android app and seeking corresponding matches within its database. The amalgamation of these elements aims to provide real-time user's location and orientation within the museum, solely by capturing the user's path through the Android app. The positioning and orientation detection will trigger a museum visitors' guide application that will deliver context aware (location and orientation aware) information to the visitor.

3. Challenges

This kind of research, carried out in a realistic setting, faces the following challenges (which we will address only part of them):

- efficient and accurate position and orientation identification: The development of such a navigation system presents several challenges. The Android smartphone serves as a wandering tool, capturing a continuous stream of images representing the user's environment. To prevent overwhelming the system with unnecessary queries, the application must intelligently determine which frames are suitable candidates for location representation. To facilitate this decision-making process, smartphone sensors and specialized techniques will be employed to identify whether a user is stationary or in motion—a valuable feature that demands adaptation. Furthermore, efficient methods for frame-to-image comparison within the database must be devised, as exhaustive comparisons could prove computationally expensive and hinder practical usability. Building and structuring the database present unique challenges as well, requiring us to devise efficient ways to represent objects and locations through images in a manner that optimizes the matching process. The conventional approach of saving numerous images for each object or location is impractical and warrants innovative solutions.
- visitor interest identification: an important aspect for engaging visitors, out of scope for this specific study, but important as future work.
- information delivery technique (e.g. text? audio? multimedia)? Again, out of scope for this specific study, but important as future work.
- proactiveness vs. reactiveness, once more, an interesting and important aspects, but out of scope for this specific study, but important as future work.
- general usability: this will be assessed by a user study that will examine the proposed approach.

This research aims to address some of these challenges and contribute mainly to the advancement of indoor navigation systems in general and museum visitors' guide systems in particular through innovative computer vision and smartphone integration.

4. Tools and methods

For feature extraction, our methodology centers on CLIP [4]. This decision was made after comparing it with alternative tools such as Resnet and OpenCV ORB. Although we attempted to combine these tools, we found that CLIP yielded the most favorable results when used as the sole feature extractor. Specifically, Resnet [5], particularly the ResNet152V2 version, proved to be prohibitively slow, requiring

approximately 10 seconds to extract features for a single image. ORB, on the other hand, exhibited less accuracy, successfully detecting locations in only about 70% of cases. In contrast, CLIP demonstrated superior speed, extracting features within 2-4 seconds, and achieving a 93% success rate in location detection. Consequently, we opted for CLIP as our primary feature extractor. Specifically, we utilized a specialized variant of CLIP known as clip-vit-large-patch14 for image processing and feature extraction (https://huggingface.co/openai/clip-vit-large-patch14). Subsequently, using the CLIP feature vector us our representstion, we employed the Euclidean distance metric to calculate distances between feature vectors representing images.

For storing the locations dataset we utilize a MongoDB database. This dataset encompasses short videos capturing objects from all angles at each location within the database. These videos are subsequently converted into a dataset of images or frames, forming a sequence to represent the video. Each image is then transformed into a vector of features using CLIP. The resultant dataset of feature vectors serves as the foundation for the ARDIF [3] algorithm, aiming to identify the minimal set of images (vectors) necessary to accurately determine specific positions within the museum. These selected images (vectors) are then stored in the database corresponding to each specific location, facilitating precise identification by matching reference images with those captured by visitors' cameras.

For the client side development we used Android SDK that provides the ability to develop apps for mobile devices. For the server side we used Python Flask that is a framework that is used to create server-side applications in Python. We decided to use it since the app requires a connection for a database and a server. Hence, there is a need to handle the API request and potentially run auto matching images and videos.

5. System description

An image-based positioning database was built using the ARIDF algorithm. The system contains the following components: a mobile App. and a web server (see Figure 1).

For museum visitors, the app initiates with a camera interface upon launch, enabling users to explore the museum environment through their device's camera. This real-time feed facilitates the tracking of visitors' movements, serving dual objectives: identifying their positions within the museum and selecting appropriate frames for querying the web server API.

Utilizing a straightforward technique, the application determines standing visitor positions by analyzing differences between consecutive frames, measured through pixel variations. Upon detecting a stationary visitor, the application assumes their presence at a Point of Interest (POI), triggering an upload process. This involves extracting a relevant frame from the stream and constructing a request for the web server. Throughout this process, visitors are kept informed via a brief waiting mode, typically lasting 3-4 seconds. Upon receiving a response, the application provides relevant information about the POI or prompts the user to retry if unsuccessful.

Operating as an Android application developed in Java, the app can be easily installed as an APK on any Android device. On the server side, a standalone Python service manages an API with two distinct entry points, each triggering a different flow of operations. Leveraging a MongoDB database, the service stores image features and general information about museum objects.

Meanwhile, for museum workers, the application facilitates the upload of museum objects. This entails submitting a short video of the object along with relevant information. The application then generates a request, which is transmitted to the server for processing. The image processing part, has two parts that are performed at the web server is detailed next (see Figure 2.)

The first part involves enriching the database with object data. Upon receiving a video and object information, the service employs the ARIDF algorithm to minimize the number of representative frames. These frames' features are extracted using the CLIP model and stored in the database alongside the provided information about the objects.

In the second part, the service accepts an image as a request and determines its similarity to any stored position representation (images) in the database. By extracting features representing the input

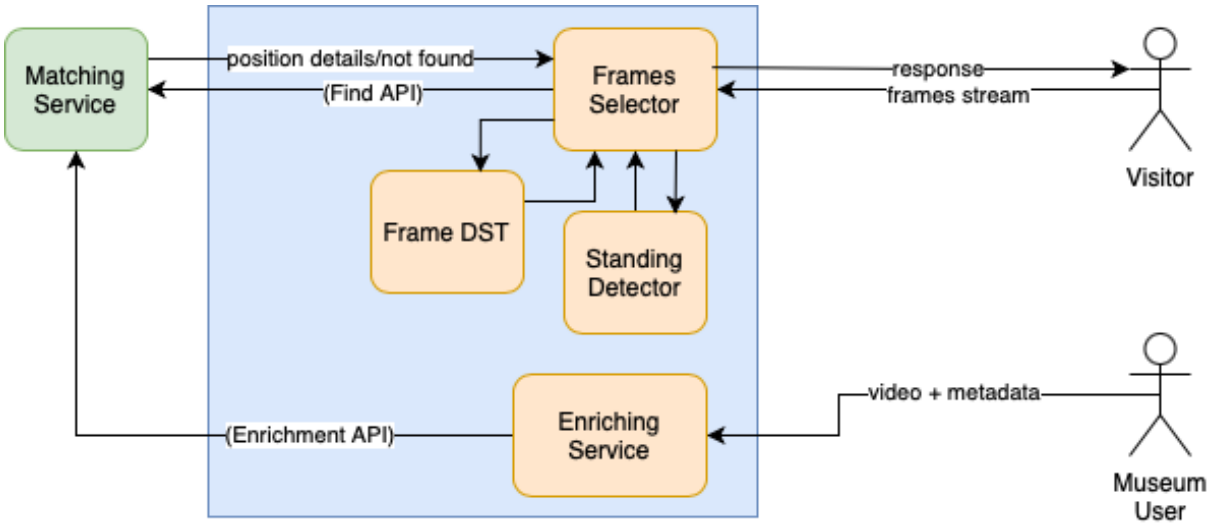

Figure 1: System architecture with two modes of operation: Mode 1. A museum staff member can upload position video and description. Mode 2. A visitor is walking in the museum while images are being continuously analysed. When the visitor is identified to stop at a certain position, a search is performed and information is delivered to the visitor according to the visitor's position and orientation.

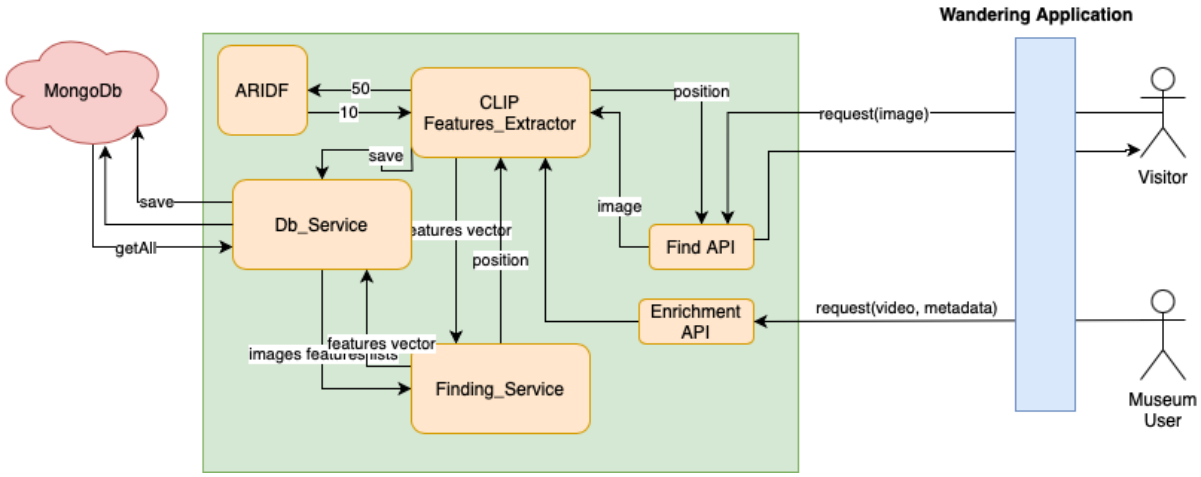

Figure 2: Image processing: Here again, there are two modes of operation. In the first mode, a museum staff member uploads information to the database, including position video and description. The video is processed for selecting the best representing images and their features are stored in the cloud-based database together with the description. In the second mode, when the visitor is stationary, features extracted from the image at the current position are used to identify a position. Once (if) identified, relevant information is delivered to the visitor.

image and comparing them with stored object features using Euclidean distance, the service identifies potential matches. If the similarity meets a specified threshold, the stored information for the object is returned as a result; otherwise, a general response is provided, indicating that a correlated object cannot be found.

6. Planned evaluation

For evaluation and as a case study we use the Hecht Museum, located at the university of Haifa. It is a challenging environment as it has a complicated structure and it is full of objects a visitor may be interested in, hence accurate visitor position is an important precondition for information delivery. For our experiments we built a small scale museum visitors guide system that covers one area at the

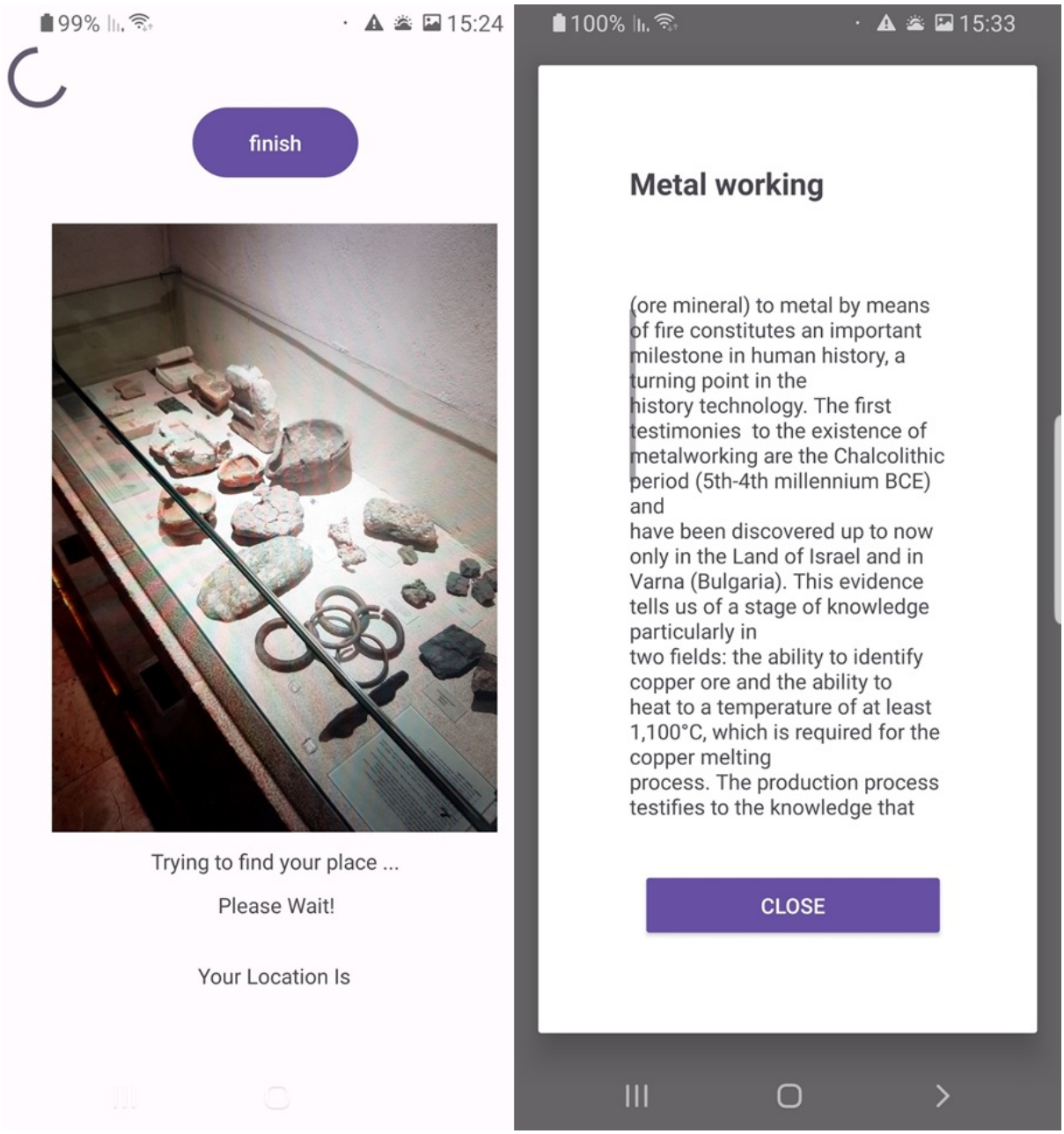

Figure 3: Application screenshots: Searching for location (left) and Information delivery (right).

Hecht Museum and contains 28 points of interest (POIs). The visitor is required to activate the App and walks around when the mobile device faces forward. When the visitor stops at a POI, for more than X seconds without changing orientation a search for the position starts and the system alerts the visitor about that. Once the position is identified, a relevant textual description is delivered to the visitor (see Figure 3). The planned evaluation is by a user study where visitors will be asked to walk around at the exhibition freely and visit at least 15 POIs. The visitors will be asked to fill the SUS questionnaire and will be interviewed about their experience and especially about the positioning function - whether it was easy to use and accurate.

7. Discussion and summary

Getting accurate indoor positioning of visitors in museums was and is still a major challenge. In most cases any technology needs to be invisible, so not to interfere with the museum architecture, a requirements that poses a major challenge for most positioning technologies that require power, communication, line of sight etc. Positioning technology that does not rely on the museum infrastructure is thus preferred. ARIDF offers such a solution, based solely on a dataset of images of POIs and the local computation of the visitor's device. However, there is a difference between a POC and an application of the idea in a realistic setting. When considering in door positioning, it should work efficiently enough to support the application. While in museums, visitors usually walk in a slow pace, stopping occasionally to look at objects. This behavior may guide the positioning system, as image matching is a time consuming process that may have negative effect on the ability to provide continuous accurate position. Still, accurate position is required when the visitor is interested in an artefact - when the visitor stops and looks at an object. In this case, there is no change in the visitor's position and orientation. From practical reasons, there is no change in the image that a front looking camera captures. The practical challenge is to identify when the visitor stops and looks at an artefact. Identifying this situation may trigger the process of position identification and information delivery that will follows. Being able to do that in a "reasonable" amount of time is the focus of our current study. As the focus of the study is image based-positioning and how intuitive is it to visitors to use it, we did not invest in a high quality content at the moment, hoping that the lack of it will not have a negative effect on the study.

References

[1] E. J. Alqahtani, F. H. Alshamrani, H. F. Syed, F. A. Alhaidari, Survey on algorithms and techniques for indoor navigation systems, in: 2018 21st Saudi Computer Society National Computer Conference (NCC), IEEE, 2018, pp. 1–9.

[2] H. Bay, B. Fasel, L. Van Gool, Interactive museum guide: Fast and robust recognition of museum objects, in: Proceedings of the first international workshop on mobile vision, 2006.

[3] M. Mokatren, T. Kuflik, I. Shimshoni, Aridf: Automatic representative image dataset finder for image based localization, in: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization, 2022, pp. 383–390.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[5] S. Targ, D. Almeida, K. Lyman, Resnet in resnet: Generalizing residual architectures, arXiv preprint arXiv:1603.08029 (2016).