BIRD: A Museum Open Dataset Combining Behavior Patterns and Identity Types to Better Model Visitors' Experience

Alexanne Worm, Florian Marchal, Sylvain Castagnos

▶ **To cite this version:**

Alexanne Worm, Florian Marchal, Sylvain Castagnos. BIRD: A Museum Open Dataset Combining Behavior Patterns and Identity Types to Better Model Visitors' Experience. UMAP '25: 33rd ACM Conference on User Modeling, Adaptation and Personalization, Jun 2025, New York City, United States. pp.18-22, ⟨10.1145/3708319.3733686⟩. ⟨hal-05421211⟩

HAL Id: hal-05421211

https://hal.science/hal-05421211v1

Submitted on 22 Dec 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

BIRD: A Museum Open Dataset Combining Behavior Patterns and Identity Types to Better Model Visitors' Experience

Alexanne Worm
alexanne.worm@loria.fr
University of Lorraine (IDMC) - CNRS - LORIA
Vandoeuvre-lès-Nancy, FRANCE

Florian Marchal
florian.marchal.bornert@gmail.com
University of Lorraine - CNRS - LORIA
Vandoeuvre-lès-Nancy, FRANCE

Sylvain Castagnos
sylvain.castagnos@loria.fr
University of Lorraine - CNRS - LORIA
Vandoeuvre-lès-Nancy, FRANCE

ABSTRACT

Lack of data is a recurring problem in Artificial Intelligence, as it is essential for training and validating models. This is particularly true in the field of cultural heritage, where the number of open datasets is relatively limited and where the data collected does not always allow for holistic modeling of visitors' experience due to the fact that data are *ad hoc* (*i.e.* restricted to the sole characteristics required for the evaluation of a specific model). To overcome this lack, we conducted a study between February and March 2019 aimed at obtaining comprehensive and detailed information about visitors, their visit experience and their feedback. We equipped 51 participants with eye-tracking glasses, leaving them free to explore the 3 floors of the museum for an average of 57 minutes, and to discover an exhibition of more than 400 artworks. On this basis, we built an open dataset combining contextual data (demographic data, preferences, visiting habits, motivations, social context…), behavioral data (spatiotemporal trajectories, gaze data) and feedback (satisfaction, fatigue, liked artworks, verbatim…). Our analysis made it possible to re-enact visitor identities combining the majority of characteristics found in the literature [3, 8–10, 16, 19] and to reproduce the Veron and Levasseur profiles [17]. This dataset will ultimately make it possible to improve the quality of recommended paths in museums by personalizing the number of points of interest (POIs), the time spent at these different POIs, and the amount of information to be provided to each visitor based on their level of interest. Dataset URL: https://mbanv2.loria.fr/

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **User models**; • **Applied computing** → *Arts and humanities*; • **Information systems** → Clustering; Location based services.

KEYWORDS

Dataset, Identity-Related Data, Spatiotemporal Data, Gaze Data, Museum Visitors' Behavior, User Modeling, Recommenders

ACM Reference Format:

Alexanne Worm, Florian Marchal, and Sylvain Castagnos. 2025. BIRD: A Museum Open Dataset Combining Behavior Patterns and Identity Types to Better Model Visitors' Experience. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct '25), June 16–19, 2025, New York City, NY, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3708319.3733686

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

UMAP Adjunct '25, June 16–19, 2025, New York City, NY, USA

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1399-6/2025/06
https://doi.org/10.1145/3708319.3733686

1 INTRODUCTION AND RELATED WORK

Museums are often faced with the question of whether they should impose a narrative to visitors, for example through appropriate signage or audio tours, or whether they should leave them free to explore their environment. Recommender systems offer an interesting alternative between these two approaches, by proposing personalized routes both based on the individual expectations and the scenography of the museum [13]. To improve the performance of such systems, it is necessary to effectively exploit data of a very varied nature to model and understand visitors' behaviors in detail [2]. Collecting visitor data in the museum context is a widely studied subject, and many technologies are dedicated to this task [4, 11, 15]. Nevertheless, there are few publicly available datasets that can be exploited for recommendation purposes. Several types of data can be collected for in-depth analysis of visitor identities. Furka *et al.* [5] consider demographic characteristics and artwork ratings given by visitors in order to predict the future ones. Packer and Roy [14] and Falk [3] are particularly interested in visitor identities, including their motivations and their learning experience in a museum. Data is collected via questionnaires. Zancanaro *et al.* [19] collected data on the visited artworks, including order, time spent and percentage observed. A particular focus is also made on the overall visitors' behavior based on their trajectories. Sparacino [16] manually retrieves visitor trajectories, annotating the visit with the number and duration of stops, as well as the items observed (12 in total). As an extension of this work, Hatala and Wakkary [9] also collect interaction history (discrete time-space points of locations and selection of objects), user type according to Sparacino's nomenclature and user interests. Yoshimura *et al.* [18] propose an automated approach: bluetooth technology is used with sensors. The trajectories consist of a sequence of sensors that detected visitors. Girolami *et al.* [6] also use Bluetooth beacons to obtain approximate trajectories for a visit around 10 works with 32 visitors. Similar techniques are employed by Lanir *et al.* [12] to get global trajectories of visitors, with radio frequency based positioning system. Another possibility to access and understand visitor behavior in an indoor environment consist to use data from another context, such as a conference [20] or a shopping mall [1]. However, such domain transpositions do not capture the specific characteristics of a museum, in particular the cultural surroundings that are introduced and not found in other spaces. At last, Grazioso *et al.* [8] investigate how mobile eye

trackers allow a deeper understanding of visitors' behavior while they observe artworks. They collected gaze data from participants from two sessions, before and after a course in art history.

As we can observe in the literature, the analyzed datasets focus on only a few aspects of each visitor. Moreover, these datasets are rarely made public, which restricts their use. In this paper, we propose a dataset with a variety of visitor information to capture their complete profile. The data collected includes the complete trajectories of visitors, the works viewed and appreciated, and data on their identity (motivation, art knowledge...). The museum journeys of 51 visitors are made available in our dataset called BIRD so far, together with museum information (floor plans, artwork descriptions, daily statistics). BIRD is the acronym for Behavioral and Identity-Related Dataset. We plan to progressively increase the number of participants.

As far as we know, no dataset to exploit the different characteristics of visitors for recommendation and identity analysis in an indoor environment has been produced and made public. The remainder of this paper is organized as follows. In Section 2, we provide the procedure of data collection and the pre-processing implemented to obtain a correctly formatted dataset. We then overview the dataset and their suitability for visitor identities in Section 3. Finally, Section 4 is dedicated to the dataset benefits, conditions of use and future work.

2 EXPERIMENT PROCEDURE

2.1 Origin and landmarks

The creation of this dataset is part of the MBANv2 project[1] in collaboration with the Nancy Museum of Fine Arts[2], which boasts a varied collection of artworks, and welcomes around 300,000 visitors every year. This project has two major objectives: firstly, to create a dataset that is as complete and faithful as possible to capture the visitors' experience, and secondly, to create a simulation and test environment (reproduction of a real museum in Unity). These two objectives enable researchers to develop and compare different machine learning models: crowd and trajectory simulation, recommender systems and virtual guides. User studies could also be carried out in the virtual environment to gather information about them and analyze their behavior. This article is dedicated to the first objective of the project.

2.2 Data acquisition

Our study involved 51 visitors and more than 400 pieces of artwork from 16th to 21th centuries. Participants were free to explore the 3 floors of the Nancy Museum of Fine Arts as they wished, without any influence such as a guide or a recommender system. Each visitor was approached at the beginning of the visit with the agreement of the museum. After being informed on the purpose of the study and data collected, they were asked to complete a consent form and were invited to visit the museum as they would under normal conditions. The experimental protocol was declared and authorized by the Data Protection Officer of our university to guarantee its legal compliance.

[1]https://mbanv2.loria.fr/
[2]https://musee-des-beaux-arts.nancy.fr/en/museum

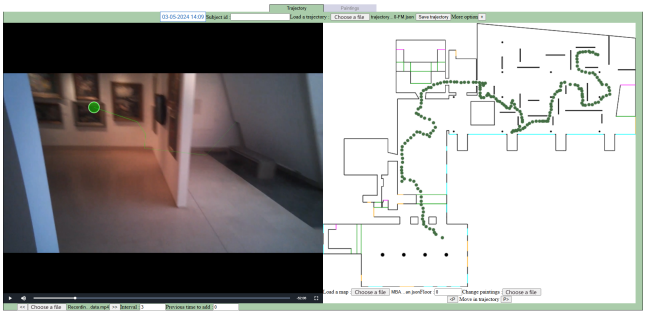
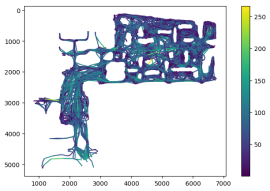


Figure 1: Illustration of the platform used to obtain trajectories.

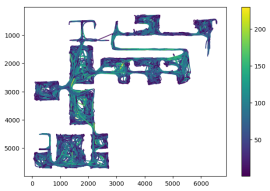
The museum contains a large variety of artworks (paintings, sculptures…). The scenography is thought out chronologically, with each floor dedicated to a specific period, and each room to a specific art movement (impressionism, post-impressionism, cubism…). The spatial topology also differs from one floor to another. The first floor gives more leeway to visitors through big rooms and multiple routes to reach the same painting. The second and third floors constrain the path with rooms only accessible by few entries. The layout of the museum and the position of the artworks were mapped out to determine visitor trajectories. The map is available in PDF and JSON formats. The dataset of artworks has also been created, with information on each item (period, theme, description, etc.). Information was collected via museum panels, the Nancy Museum of Fine Arts database, and specialized websites.

Tobii Glasses 2 (100Hz) were employed to anonymously acknowledge the items observed and the path taken by the visitors, respectively from its eye-tracking cameras and its scene camera. We developed a web platform to get the trajectories and list of items from the videos, by manually pointing the visitor's position on the map in JSON format (NMFA_3floors_plan.json). At the end of the process, a JSON file named after his anonymous unique ID is generated for each visitor. An illustration of the process is visible in Figure 1. Another window is dedicated to the acquisition of items. In this situation, all the walls containing items are clickable. When a visitor notices an item, it can be added to the list by clicking on the wall and select the correct image. Only paintings have been included in the list of visible artworks. It is possible that the visitor stops at a place to observe an item of another kind (sculpture…) but these items have not been taken into account in the analysis. The lists of artworks observed were divided into two groups, in order to obtain the duration of viewing an artwork: one group corresponds to the beginning of viewing a work, while the other corresponds to the end. To make it easier to use the trajectories and list of items with multiple languages, the JSON files obtained from the platform have been converted into CSV files (items_idVisitor.csv, items_idVisitor_end.csv, trajectory_idVisitor.csv).

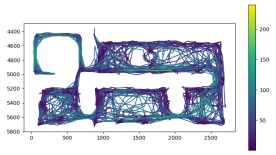
As well as collecting trajectories and list of items, pre- and post-questionnaires were drawn up to supplement the information we were able to obtain about visitors. They were completed at the beginning and end of the visit. Visitors could choose whether or not to complete the questionnaires. All decided to fill in both documents.



(a) Normalized trajectories of the first floor.



(b) Normalized trajectories of the second floor.



(c) Normalized trajectories of the third floor.

Figure 2: Normalized trajectories. The color corresponds to speed, in pixel/sec (unit from the platform). Axes also have pixel units and each point corresponds to a visitor's position at a specific timestamp. All trajectories are represented in these figures. 100 pixels correspond to 1 meter.

The questionnaires were designed to obtain data on different aspects of the visitor's identity: demographic (age, gender, diploma…), physiological (physical fatigue…), psychological (motivations, reasons to come, crowd tolerance…) and group factor (accompanied or not). Details of information acquired from the questionnaires are shown in the subsection Format and Key features (see Section 2.4).

At the end of the visit, after filling in the questionnaires, visitors were invited to use a platform to select items they liked on a mosaic of all the photographs of the artworks in the museum, randomly distributed by floor. Visitors were free to choose the artworks they liked, regardless of whether they had seen them during the visit or not. This step allows us to elicit the artistic preferences of visitors, but also to measure the coverage rate of their visit relative to their preferences (items likely to interest them and which they did not notice during their exploration). Each list of items selected by a visitor was then retrieved and placed in a CSV file (explicit_feedback_visitors.csv).

2.3 Data pre-processing

Raw trajectories may contain noise, as they are produced manually and can sometimes be too precise, preventing the detection of global patterns. To remove this noise, we normalize the trajectories. We relied on the MovingPandas library v0.20.0 [7] to normalize the trajectories with an interval of two seconds between two different positions. We obtain consistent trajectories for each floor as observed in 2.

Each cleaned trajectory is available along the raw trajectories in the dataset. An important added value of our dataset compared to the state-of-the-art is that we know precisely the direction and speed of movement of visitors in real time, whereas the datasets mentioned in Introduction only have the positions of participants at different timestamps.

Table 1: Data description of the dataset with information on trajectories.

Name	Type
Trajectory id	Integer
Duration	Float (in seconds)
Speed	pixels/sec (100 pixels = 1 meter)
Nb_items	Float
Nb_stops	Integer
Length	Float (in pixels, 100 pixels = 1 meter)

From each trajectory, we can extract a set of features to further analyze the visitor's behavior by relying on MovingPandas. In our study, we chose to select a few characteristics to check its suitability for some museum identities, such as those of Veron and Levasseur [17]. We therefore collected the duration of each trajectory, the average speed, the number of stops, the length of the trajectory and the number of items observed. This processing is also available in a CSV file (semantic_info_entire_trajectories.csv).

2.4 Format and key features

Based on the data collected, we have created a dataset with various files providing access to specific visitor information.

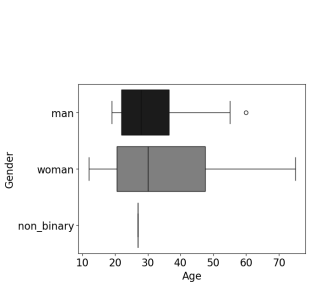
The trajectories of the N visitors are presented in a sequence of tuples containing for each timestamp the position of the visitor v: $T_v = \{(t, fl, x, y) | t = 0, ..., t_{end}; fl \in (0, 1, 2); (x, y) \in R^2\}, \forall v \in [1, ..., N]$, t_{end} the timestamp of the last point of the trajectory. The coordinates (x and y) of the visitor for each floor are in pixels, corresponding to the web platform unit (100 pixels = 1 meter). fl corresponds to the floor.

The list of items seen (items_idVisitor.csv, items_idVisitor_end.csv) are also represented by a sequence of tuples for each visitor v: $I_v = \{(t, fl, paintingID) | t = 0, ..., t_{end}; fl \in (0, 1, 2)\}$, $paintingID$ a string corresponding to the ID of the artwork image. It should be noted that lists with the beginning of items seen may contain more items than end lists. These items which are not present in the end list, have been seen for a too short duration to be included in the end list.

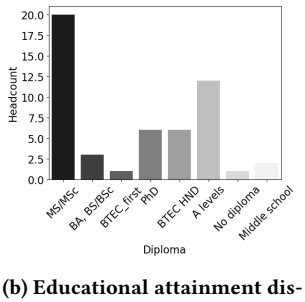
The dataset giving the global characteristics of the trajectories is composed as in Table 1.

Concerning the questionnaires, 23 questions were asked in the pre-questionnaire, which was available in French and in English. The focus is on demographic data (age, qualifications, gender), visit preferences, visit frequency, motivation, objectives, physiological and psychological data (fatigue, distance tolerance, crowd tolerance, etc.). The items related to the objectives were defined on the basis of Falk's identities ([3]). For the post-questionnaire, also in French and English, 29 questions were elaborated. This time, the main topics were art expertise, the benefits of new technologies (applications, recommender systems), the physical and psychological states at the end of the visit, the information about artworks memorized at the end of the visit, and other comments on the visitor's experience.

Finally, the artworks appreciated by visitors correspond of a file of tuples of this form: {visitor_id, id_item, timeOfSelection}.



(a) Age distribution of visitors.



(b) Educational attainment distribution of visitors.

Figure 3: Statistics of the visitor identities.

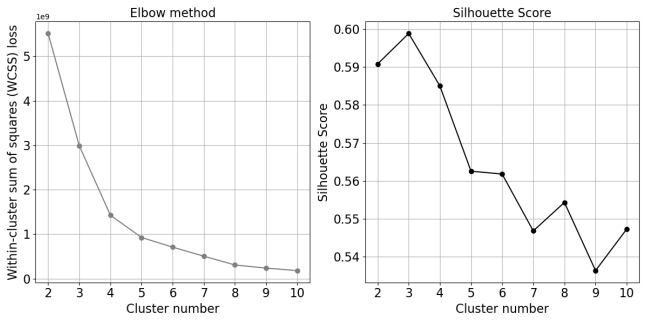


Figure 4: Results obtained from Kmeans clustering on the trajectories information (length, number of items seen, number of stops, speed and duration).

The project presentation and access to the dataset are available at the following address: https://mbanv2.loria.fr/

3 DATASET OVERVIEW

3.1 Statistics

The dataset shows a diversity of visitor identities. 51 visitor identities are currently available. Visitor statistics can be seen in Figure 3. The experiment included 27 males, 23 females and 1 non-binary person. The mean age is 33 years, going from 12 to 75 years old.

The statistical study provides some information on the 51 developed trajectories. The average visit duration is 57.6 minutes, with a minimum duration of 11 minutes and a maximum duration of 1 hour and 55 minutes. 144 items per visitor are seen on average, with a speed of around 0.26 m/s. This speed is low in view of the environment, which causes many stops (approximately 54 stops per visit). The average length of a trajectory is 838 meters, with a minimum of 254 meters and a maximum of 1361 meters. Each visitor spends an average of 29 seconds in front of selected artworks.

3.2 Suitability for museum identity analysis

To verify the suitability of our dataset for the study of visitor identities and as a proof of concept, we performed clustering to detect Veron and Levasseur profiles [17]. To do this, we used data from the file containing global information on trajectories (semantic_info_entire_trajectories). The clustering technique employed is Kmeans. We used the WCSS loss with the elbow method and Silhouette score, as well as the Davies-Bouldin Index and Calinski-Harabasz Index to select the most relevant number of clusters. As shown in Figure 4, and considering the index values, 4 seems to be the most appropriate number of clusters. A closer look at the centroids obtained reveals the nature of each group: Grasshopper, Ant, Fish and Butterfly.

An interesting future work would be to add information on the order of visit (chronological or not), in order to check whether this could have an impact on the classification.

4 LICENSE, BENEFITS AND PERSPECTIVES

This dataset provides access to detailed visitor identities with information that can be used in combination or separately. It can be employed for a wide range of applications, including the study of visitor identities and behaviors during visits, trajectory prediction, crowd simulation, use in Natural Language Processing (study of visitor comments...). Recommender systems are particularly well-suited to this type of data. Several types of systems can be used (Content-Based Filtering, Collaborative Filtering, Trajectory recommendation...) and compared thanks to this BIRD dataset. In other words, this dataset has been built independently of any research hypothesis or model to be evaluated, in order to be as generic as possible and useful to the research community, as MovieLens used to be in another application domain. Nevertheless, it is important to note that these data were collected in a specific context and over a limited period of time. The dataset seems to be in line with a certain part of the literature (Veron and Levasseur profiles [17]), and one of our future works will reside in verifying whether this dataset can be generalized to other situations.

Other future works will consist in enriching the dataset with more visitors, while giving access to gaze data. Our dataset will also be enriched with more behavioral and identity-related data (isovists, artwork complexity...) so as to train a sequence-based recommender system and test its capacity to predict visiting styles over time. We believe that this information will enable the system to tailor recommendations according to the visible items and the cognitive load that the artworks can bring. To enable this dataset to be used in Deep Learning systems, a data augmentation technique will be used to enlarge it while trying to maintain the coherence of each visitor, in particular their behaviors and trajectories. The complete dataset will be made available, along with the code used to obtain trajectories, information and analysis results under CC-BY-NC-SA 4.0 license. Any use of this dataset for research purposes must be accompanied by a citation of this paper.

ACKNOWLEDGMENTS

This research was supported by the non-economic valuation project called MBANv2. It was the subject of an agreement signed by the University of Lorraine, the metropolitan area of Nancy (Ville de Nancy) and the Nancy Museum of Fine Arts. We would like to thank Sophie Mouton, Sophie Toulouze, Charles Villeneuve de Janti, Michèle Leinen, and Jean-Paul Darada for providing information on the artworks and for authorizing this study to be conducted within the museum.

REFERENCES

[1] Dražen Brščić, Takayuki Kanda, Tetsushi Ikeda, and Takahiro Miyashita. 2013. Person tracking in large public spaces using 3-D range sensors. *IEEE Transactions on Human-Machine Systems* 43, 6 (2013), 522–534.

[2] Sofia Ceccarelli, Amedeo Cesta, Gabriella Cortellessa, Riccardo De Benedictis, Francesca Fracasso, Laura Leopardi, Luca Ligios, Ernesto Lombardi, Malatesta Saverio Giulio, Angelo Oddi, Alfonsina Pagano, Augusto Palombini, Gianmauro Romagna, Marta Sanzari, and Marco Schaerf. 2024. Evaluating visitors' experience in museum: comparing artificial intelligence and multi-partitioned analysis. *Digital Applications in Archaeology and Cultural Heritage* 33 (2024). https://doi.org/10.1016/j.daach.2024.e00340

[3] John H. Falk. 2016. *Identity and the museum visitor experience.* Routledge.

[4] Alessio Ferrato, Carla Limongelli, Mauro Mezzini, and Giuseppe Sansonetti. 2022. Using deep learning for collecting data about museum visitor behavior. *Applied Sciences* 12, 2 (2022), 533.

[5] Bc Marek Furka. 2022. *Exhibit rating prediction and vistior path prediction in a museum setting.* Ph. D. Dissertation. Wien.

[6] Michele Girolami, Davide La Rosa, and Paolo Barsocchi. 2024. Bluetooth dataset for proximity detection in indoor environments collected with smartphones. *Data in Brief* 53 (2024), 110215.

[7] Anita Graser. 2019. MovingPandas: Efficient Structures for Movement Data in Python. *GI_Forum – Journal of Geographic Information Science* 7, 1 (2019), 54–68. https://doi.org/10.1553/giscience2019_01_s54

[8] Marco Grazioso, Roberto Esposito, Emma Maayan-Fanar, Tsvi Kuflik, and Francesco Cutugno. 2020. Using Eye Tracking Data to Understand Visitors' Behaviour. In *AVI2CH Workshop on Advanced Visual Interfaces and Interactions in Cultural Heritage.* Island of Ischia, Italy.

[9] Marek Hatala and Ron Wakkary. 2005. Ontology-Based User Modeling in an Augmented Audio Reality System for Museums. *User Modeling and User-Adapted Interaction* 15, 3-4 (2005), 339–380.

[10] K. Kontiza, O. Loboda, L. Deladiennee, S. Castagnos, and Y. Naudet. 2018. A Museum App to Trigger Users' Reflection. In *2nd International Workshop on Mobile Access to Cultural Heritage (MobileCH).*

[11] La-or Kovavisaruch, Virach Sornleardlumvanich, Thatsanee Chalernporn, Pobsit Kamolvej, and Nitirat Iamrahong. 2012. Evaluating and collecting museum visitor behavior via RFID. In *2012 Proceedings of PICMET'12: Technology Management for Emerging Technologies.* IEEE, 1099–1101.

[12] Joel Lanir, Tsvi Kuflik, Julia Sheidin, Nisan Yavin, Kate Leiderman, and Michael Segal. 2017. Visualizing museum visitors' behavior: Where do they go and what do they do there? *Personal and Ubiquitous Computing* 21 (2017), 313–326.

[13] Lukas Najbrt and Jana Kapounová. 2016. Categorization of Museum Visitors as Part of System for Personalized Museum Tour. *International Journal of Information and Communication Technologies in Education* 3, 1 (2016), 17–27.

[14] Jan Packer and Roy Ballantyne. 2002. Motivational factors and the visitor experience: A comparison of three sites. *Curator: The Museum Journal* 45, 3 (2002), 183–198.

[15] Lorenzo Seidenari, Claudio Baecchi, Tiberio Uricchio, Andrea Ferracani, Marco Bertini, and Alberto Del Bimbo. 2017. Deep artwork detection and retrieval for automatic context-aware audio guides. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13, 3s (2017), 1–21.

[16] Flavia Sparacino. 2002. The Museum Wearable: Real-Time Sensor-Driven Understanding of Visitors' Interests for Personalized Visually-Augmented Museum Experiences. In *Museums and the Web.* Boston, MA.

[17] Eliseo Verón and Martine Levasseur. 1989. *Ethnographie de l'exposition: l'espace, le corps et le sens.* Centre Georges Pompidou, Bibliothèque publique d'information.

[18] Yuji Yoshimura, Fabien Girardin, Juan Pablo Carrascal, Carlo Ratti, and Josep Blat. 2012. New tools for studying visitor behaviours in museums: a case study at the Louvre. In *Information and Communication Technologies in Tourism 2012.* Springer, 391–402.

[19] Massimo Zancanaro, Tsvi Kuflik, Zvi Boger, Dina Goren-Bar, and Dan Goldwasser. 2007. Analyzing Museum Visitors' Behavior Patterns. In *User Modeling*. Springer Berlin Heidelberg, Berlin, Heidelberg, 238–246.

[20] Ying Zhao, Xin Zhao, Siming Chen, Zhuo Zhang, and Xin Huang. 2021. An indoor crowd movement trajectory benchmark dataset. *IEEE Transactions on Reliability* 70, 4 (2021), 1368–1380.