Tracking the Visitor

An Optical Indoor System for Visitor Research in Museums

Franz Koeferl, Matthias Zuerl, Jitin Jami, Jindong Li, Dario Zanca, Bjoern Eskofier[1]

Visitor tracking has become a de facto standard for evaluating the success of exhibitions in museums (Yalowitz/Bronnenkraut 2009, 47). The data collection required for this analysis is, however, usually very labour-intensive or requires a costly setup (ibid.). But such systems usually do not gather, in particular, information about the visitor, like gender, age, and apparent interests. This information is instead painstakingly collected by means of questionnaires (ibid.). We propose a simple and cost-efficient automatic visitor monitoring pipeline to capture not only visitor trajectories, similar to established products, but also personal parameters generally only collected through questionnaires. Furthermore, we evaluate parts of this pipeline using state-of-the-art methods.

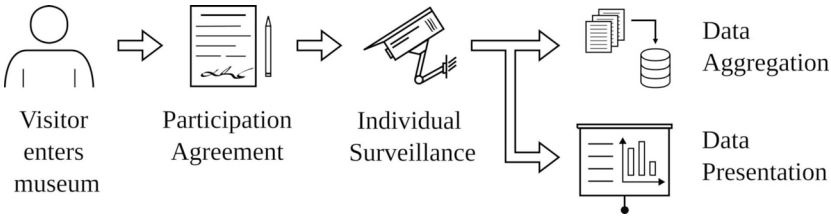
Related Work

Various methods can be used to conduct visitor tracking. The paper-and-pencil method is historically the most common (Yalowitz/Bronnenkraut 2009, 52–53) and requires a researcher to collect visitor data by hand. Modern methods feature radio technologies like Bluetooth, ultra-wide band, or wireless local area networks (De Angelis and Santoni 2022, 8). Older designs, in contrast, rely on infrared light and optical tracking approaches (Kuflik/Lanir/Dim et al. 2011, 375–76). But with recent advances in deep learning and person detection in images and videos, especially in the context of museum research (Bartoli et al. 2015, 19–27), visitor tracking by means of optical methods seems to have become a feasible alternative for visitor studies.

1 We gratefully acknowledge the support of DATEV eG and the Deutsches Museum Nuremberg through the project Tracking in the Deutsche Museum Nuremberg. Bjoern Eskofier gratefully acknowledges the support of the German Research Foundation (DFG) within the framework of the Heisenberg professorship program (grant number ES 434/8-1).

Particularly in the context of large-scale camera networks (Zhang/Scanlon/Yin et al. 2009, 435–56), there are a range of approaches to tracking people with multiple cameras (Ristani/Tomasi 2018, 6037). But most research does not take visitor consent for data capture and processing into account in its pipeline. Our intention is thus to close this research gap by introducing a novel framework.

Figure 1: After a visitor enters the museum, consent is obtained from the visitor at a registration station. The system then tracks this visitor and filters out information from other non-consenting visitors. The information compiled can then be presented to the visitor or used for further analysis.



Tracking System

The use of optical tracking systems in public spaces has generated considerable debate as it is regarded as an invasion of privacy. Most nations and multijurisdictional entities such as the European Union have implemented stringent regulations pertaining to the capture and processing of photographic and video graphic data on individuals (Meints/Biermann/Bromba et al. 2008, 1088). Even countries that were to some extent lax regarding data protection in the past (Weber/Zhang/Wu 2020, 568–70) have been paying more and more attention to data privacy in recent years (Yin/Li/Liu et al. 2022, 1–2). The further processing and, in particular, use of this visual information is heavily regulated. Explicit consent with respect data collection and processing are therefore often legal requirements for such activities. We propose a concept for a consent-based visitor tracking solution that adheres to the laws of countries with stricter regulations, like Germany.

The workflow of our tracking system is depicted in figure 1. A visitor enters the museum and signs a participation agreement at a special edge device called a registration station. Immediately after signing the agreement, the station captures several frontal images of the person, which then constitute the individual reference gallery for the purpose of live re-identification. These images can also be used to determine personal attributes such as age and sex. All the information, including the images, is sent to a tracking system that ensures individual surveillance. The system

thus only tracks people who have signed the agreement, and the data on people who did not give consent are immediately discarded.

Figure 2: The overall tracking system is structured into several edge devices, each of them connected to at least one camera. The image information is processed and send to a central server for information aggregation, mapping, and trajectory estimation.

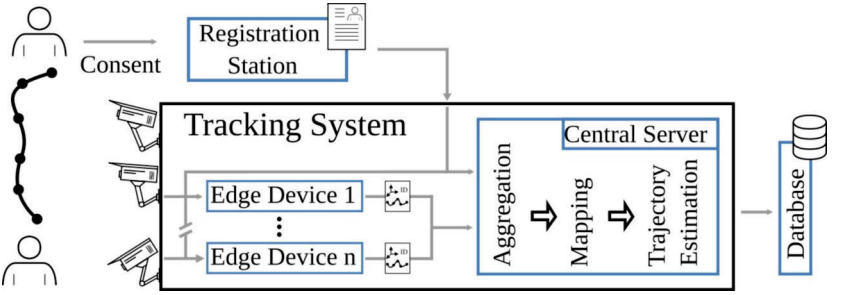
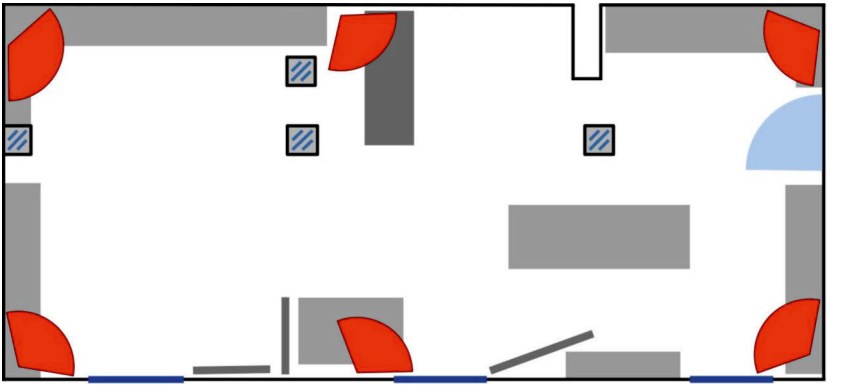


Figure 3: The tracking system is installed in an indoor laboratory setting. The room plan shows the installation location from a top-view perspective. The cameras are illustrated in red, the windows and the entry door in blue, and tables, shelves, or columns in the room in grey.



The tracking data resulting from this surveillance can be used for a live presentation of the data to the consenting visitor or for data aggregation, hence allowing further analysis. Note that the data from the tracking system intended for analysis no longer contains image information, but solely anonymized trajectories, visit duration, and generic personal information.

The tracking system consists of a central server and several edge devices linked to at least one camera. The edge devices are responsible for image processing in the pipeline, which processes the video stream captured by the cameras on a per-image basis. The central server, in contrast, aggregates the information from the various edge devices and creates the complete trajectories of consenting visitors. Note that this adheres to the typical design for large-scale tracking (Zhang/Scanlon/Yin, Weihong et al. 2009, 435–56). The overall structure of the system is visible in figure 2.

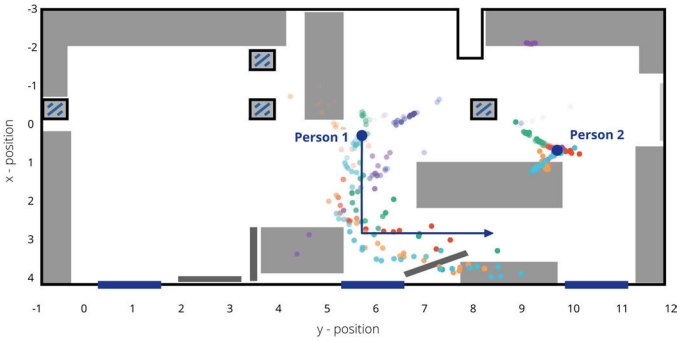
In what follows, we focus on the image processing aspect of the tracking system, in particular, on the processing of the data provided by the edge devices. After an image is captured, we apply a pre-trained DetectNet model (Tao/Barker/Sarathy 2016) to detect the visitor. The resulting bounding boxes are used to further localize the person. We estimate the 3D position by projecting the 2D image coordinates, specifically the lower middle point of the bounding box, onto the 3D ground plane using the camera's intrinsic and extrinsic matrix. We infer the intrinsic matrices using a checkerboard pattern (Zhang 1998) and the extrinsic matrices using an iterate PnP solving scheme (Eade 2013) for defined 3D points, both using the implementation in the OpenCV library (Bradski 2000). For the extrinsic calibration, we placed seven coloured balls on a one-meter-spaced grid and labelled them manually based on image coordinates to achieve the 2D–3D correspondences.

Experimental Setup

We installed our proposed system in an indoor laboratory setting resembling a museum exhibition. The experimental environment has a floor area of 72 square meters, with the dimensions of 6 meters in length and 12 meters in width. It was furnished with various objects, thus resulting in occlusion. We mounted six cameras from multiple angles (see fig. 3), each connected via a universal serial bus (USB) to an edge device for image processing in each top corner of the room. We used the DFK 37BUX178 for all our cameras and the Boxer 8251AI, containing the Nvidia Jetson Xavier NX GPU, as our edge device.

In this setting, we collected data for two participants using the following procedure: In the first scenario, one person remained in the same position for the entire duration, and the other moved to a different fixed position and remained there for several seconds. This second person repeated this step for various positions with varying distances from person one. This procedure makes it possible to evaluate the localization performance. In the second scenario, both participants moved around in the room following specific trajectories at specific times, a scenario intended to evaluate the detection performance.

Figure 4: The localization performance suffices to dtermine the trajectory of person one and the position of person two in the room. Note that the colouring of the position estimate is determined by the camera that performs the estimation.



Evaluation

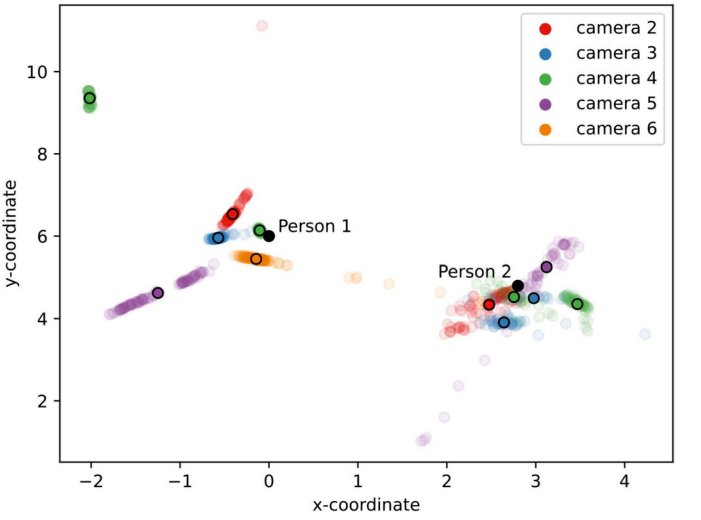
The performance of the tracking system as a whole relies to a great extent on the performance of the algorithms on the edge devices. Within the scope of this work, we therefore present the preliminary results of the detection and localization method used for the tracking system, thus showing that the methods achieve sufficient accuracy for further server-side tracking.

The most important aspect of the system is the person detector. Typical metrics for evaluating detector performance are average precision (AP) and average recall (AR). AP measures false positives, with one indicating none and zero indicating that all predicted detections are incorrect. AR measures true positives, with one indicating that all individuals are detected accurately and zero indicating that none are detected. The system, without any fine-tuning, achieves a total-camera-averaged AR of 0.19 with a standard deviation of 0.07 and an average AP of 0.2 with a standard deviation of 0.11. This means that the system detects a person just 20 per cent of the time, and there are about 20 per cent anomaly detections. This may appear low, but if we include the qualitative results in figure 4, which show the 3D positions of all detections, we are still able to determine the trajectories of persons. This is due to the number of frames taken per second. Note that filtering techniques are able to remove 30 per cent of anomaly detections.

Based on these detections, we apply a localization method to the sub-scenario of our sequence, where both participants remain in the same position for several seconds. The results are depicted in figure 5. The figure shows the localizations for each camera cluster closely and spread solely along the camera's viewing direction.

Moreover, each cluster is quite close to the actual ground truth position vis-à-vis the position of the other person. In this scenario, a separation of persons can be achieved, and tracking ensured in a later stage. On average, the localization error is 0.64 meters, with a minimum error of 0.17 meters and a maximum error of 1.86 meters.

Figure 5: The ground truth positions of each person are highlighted in black. The positions estimated by each camera are highlighted in different colours.



Discussion

The detection and localization performance, though acceptable in principle, requires greater accuracy to be applied on a large scale. Especially the localization may be problematic for information aggregation on the server side when multiple people are in close proximity, which is usually the case when one visits a museum with friends or family. The experimental setup should thus reflect this scenario with multiple participants in a museum-like setting. This scenario must, of course, also be repeated multiple times in order to ensure high statistical power.

Conclusion

This work proposes and partially evaluates a simple, mobile, and cost-efficient automatic visitor monitoring pipeline. We show that the proof of concept works sufficiently on edge devices, but still has room for improvement. Future work will include integrating re-identification capabilities into the edge device pipeline and further improving the established algorithms by transfer learning and image coordinate refinement. With this work, we are one step closer to an optical tracking system applicable for visitor research in museums.

References

Bradski, Gary (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools, 122–25. Available online at https://www.proquest.com/trade-journals/opencv-library/docview/202684726/se-2 (all URLs here accessed in August 2023).

De Angelis, Alessio/Francesco, Santoni (2022). Advanced Sensors and Sensing Technologies for Indoor Localization. Applied Sciences 12 (8), 3786. https://doi.org/10.3390/app12083786.

Eade, Ethan (2013). Gauss-Newton / Levenberg-Marquardt Optimization. https://ethaneade.com/optimization.pdf.

Kuflik, Tsvi/Lanir, Joel/Dim, Eyal et al. (2011). Indoor Positioning: Challenges and Solutions for Indoor Cultural Heritage Sites. International Conference on Intelligent User Interfaces, Proceedings IUI, 375–78. https://doi.org/10.1145/1943403.1943469.

Meints, Martin/Biermann, Heinz/Bromba, Manfred et al. (2008). Biometric Systems and Data Protection Legislation in Germany. 2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, 1088–93. https://doi.org/10.1109/iih-msp.2008.314.

Ristani, Ergys/Tomasi, Carlo (2018). Features for Multi-Target Multi-Camera Tracking and Re-Identification. Proceedings of the IEEE conference on computer vision and pattern recognition, 6036–46. arXiv:1803.10859. https://doi.org/10.48550/arXiv.1803.10859.

Tao, Andrew/Barker, Jon/Sarathy, Sriya (2016). DetectNet: Deep Neural Network for Object Detection in DIGITS. Available online at https://developer.nvidia.com/blog/detectnet-deep-neural-network-object-detection-digits.

Weber, Philip Andreas/Zhang, Nan/Wu, Haiming (2020). A Comparative Analysis of Personal Data Protection Regulations between the EU and China. Electronic Commerce Research 20 (3), 565–87. https://doi.org/10.1007/s10660-020-09422-3.

Yalowitz, Steven/Bronnenkant, Kerry (2009). Timing and Tracking: Unlocking Visitor Behavior. Visitor Studies 12 (1), 47–64. https://doi.org/10.1080/10645570902769134.

Yin, Daoxin/Li, Xiaojie/Liu, Ruishuang et al. (2022). China's Personal Information Protection Law. BMJ 379, e072619. https://doi.org/10.1136/bmj-2022-072619.

Zhang, Zhengyou (1998). A Flexible New Technique for Camera Calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (11), 1330–34. https://doi.org/10.1109/34.888718.

Zhang, Zhong/Scanlon, Andrew/Yin, Weihong et al. (2009). Video Surveillance Using a Multi-Camera Tracking and Fusion System. In: Hamid Aghajan/Andrea Cavallaro (Eds.). Multi-Camera Networks. Oxford, Academic Press, 435–56. https://doi.org/10.1016/b978-0-12-374633-7.00020-3.