Symotiv

Virtual Insights into the Symphony Orchestra

Michael Zöllner, Markus Bosl, Dirk Widmann, Moritz Krause

The aim of the project Symotiv was to introduce people to classical music through interaction with a virtual orchestra.[1] With the aid of virtual reality (VR) glasses and a controller, it not only enables people to immerse themselves in a three-dimensional virtual concert hall, but also to experience images and sound from the perspective of musicians. This opened up the possibility to experience and understand what is happening from new spatial, but above all sonic perspectives. To do so, it was necessary for us to capture the motions and sound of 50 musicians and the conductor. Whereas a few years ago expensive and complex time-of-flight cameras or sensor-based bodysuits were needed to capture the biomechanical movements of a musician, we used machine learning-based tracking software that relies on ordinary 2D RGB camera images.

Using eight off-the-shelf GoPro[2] cameras distributed throughout the orchestra's venue, the Freiheitshalle in Hof, Germany, we captured all the musicians and their movements during a performance without occlusion. We used opensource software for 2D pose estimation, that is, for extracting the coordinates and angles of all the musicians' joints from the camera image. In a further step we created a three-dimensional biomechanical model of the musicians. This provided the basis for animating the avatars in virtual reality. To process the eight camera tracks with several terabytes of data, from which the movements of the musicians had been extracted and transformed into biomechanical models, we developed an automatic processing pipeline on our graphics workstations.

In the following sections we describe the evaluation and exploration of the current state of single-camera skeleton-tracking hardware that led us to select the approach we implemented as separate prototypes in order to evaluate the quality of the resulting data. We go on to briefly explain the further development into a VR experience. Finally, we discuss the lessons learned during the development and the significance of these technologies for interactions with visitors in museums.

1 http://symotiv.de (all URLs here accessed in June 2023).

2 https://gopro.com.

Figure 1: A violinist's skeleton tracked while playing.

Related Work

Over the past years, we have seen rapid progress in the quality and availability of camera-based pose estimation. An early moment in the accessibility of skeleton tracking for a large audience was the release of the Microsoft Kinect and community-based libraries like OpenKinect[3] and OpenNI (Villaroman/Rowe/Swan 2011).

With the advent of machine learning-based approaches, there was no longer a need for special hardware and we were thus able to use standard cameras. CMU's OpenPose (Cao et al. 2019) and OpenPifpaf (Kreiss/Bertoni/Alahi 2021) are two major developments that produce robust 2D pose estimations of humans based on single RGB images. Even if the quality of PoseNet's (Kendall/Grimes/Cipolla 2015) 2D tracking data is not as advanced as the data of those previously mentioned, its main contribution is its availability in browsers. ml5js[4] facilitates a large developer audience by simplifying the development process even further through the integration of p5js (McCarthy/Reas/Fry 2015). The even simpler Google MediaPipe (Lugaresi et al. 2019) combines a large variety of machine learning applications in one JavaScript framework.

When we started the project, we decided to build our project on top of OpenPose since it gave us the best quality results for robust 2D position data for the skeletons' joints at the time. Unfortunately, at that time, OpenPose was only able to produce 2D position data, while today's version features 3D position data as well. We therefore

3 https://openkinect.org.

4 https://ml5js.org/.

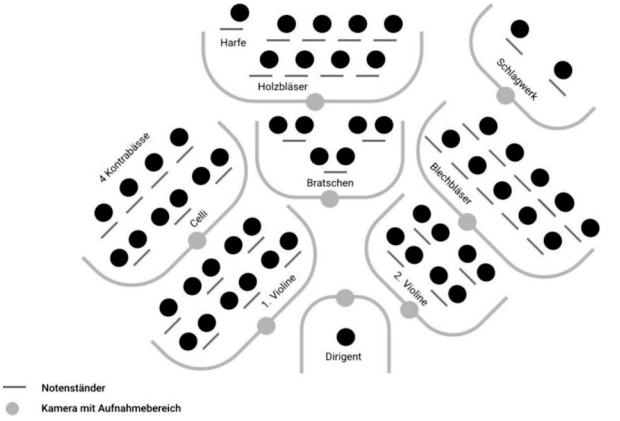
required a second processing step in order to create a 3D biomechanical skeleton model. '3d-pose-baseline' (Martinez et al. 2017) provided this feature and predicted the 3D skeletons. Similar to OpenPose is MeTRAbs' Absolute 3D Human Pose Estimator (Sárándi et al. 2021), which also features not only 2D but also robust 3D position data for the skeletons' joints.

With respect to related applications, we would like to mention several projects in adjoining disciplines. Capturing and analysing dance is a common research field in the field of pose estimation and tracking. With reference to hardware-based approaches, we would like to mention Alexiadis et al.'s 'Evaluating a dancer's performance using Kinect-based skeleton tracking'. An early 3D position estimation approach with a single camera was Kahn et al.'s 'Capturing of contemporary dance for preservation and presentation of choreographies in online scores', in which recurring path patterns in experimental ballet performances are shown.

Although our focus is on visualization rather than data analysis, we would also like to mention RunwayML's most recent motion tracking tool for video editing[5] and Najeeb Tarazi's application of RunwayML's rotoscoping techniques in the impressive *One More Try* experimental skating video.[6]

Implementation and Evaluation

Figure 2: Setup of eight GoPro cameras positioned around the musicians.



5 https://runwayml.com/.

6 https://vimeo.com/717945664.

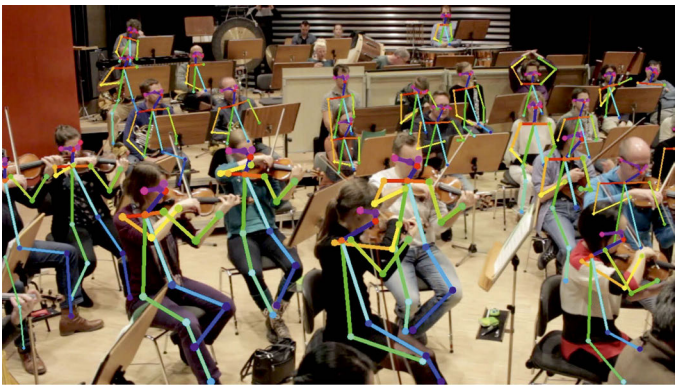
Hardware Setup

To capture the poses of the 50 musicians, we used a reproducible setup consisting of eight GoPro cameras (see fig. 2). Since we required an affordable solution, we used off-the-shelf GoPro Hero 5 Black Cameras attached to long tripods with a 3D-printed mount. A major challenge was positioning the cameras to capture all the musicians without occluding other people, instruments, or chairs. We resolved this by calculating, defining, and fixing the positions of the chairs and the camera tripods for the recordings of the rehearsals. A GoPro Remote served as the control for the cameras to capture a synchronized recording of the various views. Along with the GoPro app on a smartphone, it was also the touchpoint for the setup of the image detail and verification of the recordings.

Tracking and Data

The resulting video sequences amounted to several terabytes of data. They were transferred from the cameras to our graphics workstation and processed by an OpenPose script into CSV files containing the 2D skeleton coordinates of the persons recognized per time frame. The sets of coordinates represent a frontal projection of the joint position that matches the video image and its pixel space (see fig. 3). Each set consists of the same number of frames as the video clips. It contains all the individuals detected along with their skeleton joint positions with their x and y coordinates.

Figure 3: Early visualization of the 2D skeleton data for the musicians.



In the next step the '3d-pose-baseline' script scanned all the OpenPose 2D CSV files and predicted the skeleton data of the musicians in 3D space. The output was

also saved in single CSV files with x, y, and z coordinates. To evaluate the resulting datasets, we developed a visualization prototype, which will be described in the next section.

For a subsequent visualization in Blender[7] and Unity3D,[8] we wrote a custom script to convert the data into the Biovision Hierarchy (BVH) character animation file format (Meredith/Maddock et al. 2001). Based on the fixed position of the cameras and the musicians, we were able to combine the 3D datasets from the different cameras and instrument groups into the real 3D positions of the whole orchestra in the hall.

Data Visualization Prototype

To evaluate the data, we developed a web-based visualization that renders the skeleton data on top of the corresponding image in the recorded scenario. We used p5js (McCarthy/Reas/Fry 2015), a Javascript variant of the Processing language, for the visualization. Each frame is parsed via CSV into p5js. All of a person's joints per frame are then rendered as lines and ellipses. The colour coding enhances the recognition of the individual joints and bones.

Figure 4: Real-time prototype for visualizing the skeleton joint positions of a musicians superimposed on top of an image of the real persons.

The result was a visualization of the dataset on top of the corresponding video frame. This thus enabled us to recognize the persons in the dataset, including their positions and movements, over time. In the example in figure 4, we are observing a

7 https://www.blender.org/.

8 https://unity.com/.

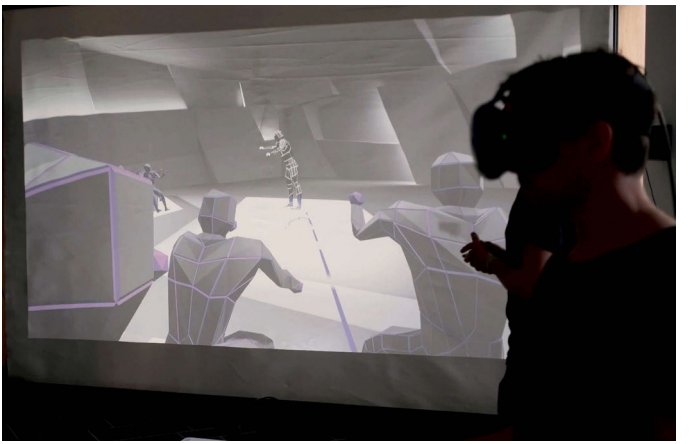
violin player playing a short sequence. The person's movements are overlayed with the drawing of the reconstructed skeleton. Based on the components described, we were able to evaluate the quality of the reconstructed skeleton data in a first integrated prototype in real-time based on previously recorded video segments subsequently analysed in p5js.

Virtual Reality Reexperience

The venue and the space where an orchestra plays are of great importance for its performance. Because of this significance, for our visualization of the motion data we worked with virtual reality, because it facilitates a recreation of the room that can be entered with complete freedom. VR was thus used to visualize and clarify the data collected.

Unlike in a real concert, in our application the users have the possibility to intervene interactively in what is happening (see fig. 5). They can move freely around the concert hall, stand on the conductor's podium, and turn groups of instruments on and off, or stand among the musicians, observe their movements and sounds, and thus expand and understand the spatial and sonic experience of a concert. The immersive nature of this medium offers scope for new, inclusive information delivery. Various visualizations here reach the user on multiple levels. Like the concert event itself, VR offers an independent form of experience.

Figure 5: Virtual reality reexperience of the orchestra rehearsal captured.



Current Developments in Interaction with Visitors

In section one, we described the state of the art and the exciting developments in pose estimation over the past years. Today, we now have high-quality solutions for 3D human pose estimation like OpenPose and MeTRAbs for processing video recordings into skeleton data. Though it is not as precise but much simpler to use due to its web-based approach, Google MediaPipe facilitates 2D human pose estimation, hand and/or gesture tracking, object recognition, and many more features in its browser. Anyone with web-development skills can thus use these machine learning technologies for their applications.

We think this will also enable museums and exhibitions that use these technologies to prompt their visitors to interact with exhibits. The interactive exhibit therefore knows whether and how many people are standing in front of it. It is able to recognize activating gestures like pointing or waving. And it can even see tangible objects like artifacts that may play an interactive role in storytelling. And it does all of this with a simple camera in a PC, tablet, or smartphone and a few lines of JavaScript.

Conclusion and Future Work

We have presented the technology and development of a contemporary digital process for capturing all the motions and sounds of a symphony orchestra using commercially available cameras and AI-based 3D pose estimation software. As a result, the cultural institution of the Hof Symphony was able to communicate its own work to a larger audience by means of new technologies and to improve the internal processes of orchestra rehearsals and the training of musicians.

Based on these developments and current technological progress, we have also described and proposed the use of pose estimation, gesture tracking, and object detection in museum exhibits for interacting with and activating visitors.

We are currently teaching these skills to our students of communication design, the next generation of designers. These skills have already been applied in the exhibition projects *Walderlebniszentrum Mehlmeisel* and *Maximilian von Welsch*, and in the workshop series Co-Learning Lab.

References

Cao, Zhe/Hidalgo, Gines/Simon, Tomas et al. (2019). OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1812.08008. https://doi.org/10.48550/arXiv.1812.08008 (all URLs here accessed in August 2023).

Kendall, Alex/Grimes, Matthew/Cipolla, Roberto (2015). PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. Proceedings of the IEEE International Conference on Computer Vision, 2938–46. Available online at https://openaccess.thecvf.com/content_iccv_2015/html/Kendall_PoseNet_A_Convolutional_ICCV_2015_paper.html.

Kreiss, Sven/Bertoni, Lorenzo/Alahi, Alexandre (2021). Openpifpaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. arXiv:2103.02440. https://doi.org/10.48550/arXiv.2103.02440.

Lugaresi, Camillo/ Tang, Jiuqiang/Nash, Hadon et al. (2019). Mediapipe: A Framework for Building Perception Pipelines. arXiv:1906.08172. https://doi.org/10.48550/arXiv.1906.08172.

Martinez, Julieta/Hossain, Rayat/Romero, Javier et al. (2017). A Simple Yet Effective Baseline for 3d Human Pose Estimation. arXiv:1705.03098. https://doi.org/10.48550/arXiv.1705.03098.

McCarthy, Lauren/Reas, Casey/Fry, Ben (2015). Getting Started with p5.js: Making Interactive Graphics in JavaScript and Processing. San Francisco, Maker Media. Available online at http://people.uncw.edu/tompkinsj/112/JavaScript/GettingStartedwithP5js.pdf.

Meredith, Michael/Maddock, Steve (2001). Motion Capture File Formats Explained. Department of Computer Science, University of Sheffield. Available online at https://staffwww.dcs.shef.ac.uk/people/S.Maddock/publications/Motion%20Capture%20File%20Formats%20Explained.pdf.

Sárándi, István/Linder, Timm/Arras, Kai Oliver (2021). MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation. IEEE Transactions on Biometrics, Behavior, and Identity Science 3 (1), 16–30. https://doi.org/10.1109/TBIOM.2020.3037257.

Villaroman, Norman/Rowe, Dale/Swan, Bret (2011). Teaching Natural User Interaction Using OpenNI and the Microsoft Kinect Sensor. Proceedings of the 2011 Conference on Information Technology Education, 227–32. https://doi.org/10.1145/2047594.2047654.