Human-Size Interactive Mirror for Real-Time Movement Feedback: A Technical Framework for Physiotherapy Applications

Ethan Fuchs, Jérôme Kochenburger, Paula Liebrich, Johannes Merschroth, Nina Mugele, Marit Zenker, Angela Dieterich, and Uwe Hahne

Hochschule Furtwangen
firstname.lastname@hs-furtwangen.de

Abstract. We present a modular framework for a life-size interactive mirror based on stereo RGB cameras, markerless human pose estimation, and real-time 3D avatar rendering. The system reconstructs the user's three-dimensional body pose from two camera views, maps the resulting keypoints to a humanoid VRM avatar, and renders the avatar as a perspective-aware virtual reflection on a large display. We evaluate multiple HPE configurations from *MediaPipe* and *RTMLib* with respect to processing performance and robustness under varying subject distance, partial occlusions, additional people, and temporal stability. *MediaPipe Full* provided the best overall balance between pose quality and processing speed and was therefore selected for the prototype. A proof-of-concept interactive application demonstrates real-time movement-based interaction with the virtual avatar. The framework provides a technical foundation for future movement-feedback applications, including physiotherapy, where exercise-specific analysis and visual feedback can be investigated.

Keywords: Human Pose Estimation; Physiotherapy; Interactive Systems

1 Introduction

Physiotherapy is widely used to prevent and treat injuries, restore mobility, and improve balance as well as muscular strength. It typically relies on continuous monitoring by trained professionals to observe, evaluate, and correct patient movements. While this guided feedback is highly effective, it is also resource-intensive and cannot be transferred to self-guided training or home exercise [1]. This creates a need for accessible systems that can provide users with immediate visual feedback on their own movements without requiring specialized equipment or continuous supervision. The presentation of movement feedback can influence how users perceive and interact with their own body. Preston et al. [2] state that mirrors provide a particularly familiar form of self-referential visual feedback: humans learn to interpret the spatial transformation of their reflected body and routinely use mirrors to guide actions. The authors found stronger subjective and physiological evidence of body ownership when a body was viewed through a mirror compared with a non-mirror third-person perspective. Their findings suggest that mirror viewing provides a distinct relationship between the displayed body and the observer's own body representation. Established motion tracking approaches in rehabilitation usually include marker-based optical systems which provide highly accurate three-dimensional reconstructions of the human pose, but require specialized equipment, controlled environments, and wearable markers, limiting their practicality outside of laboratory settings.

Markerless approaches based on neural networks [3] and depth-sensing cameras [4] reduce or eliminate the need for physical markers and can simplify deployment. More recent Human Pose Estimation (HPE) methods can also estimate body keypoints directly from conventional RGB cameras, making markerless 3D motion tracking possible using relatively inexpensive hardware. However, the reliability of these approaches can vary with factors such as occlusion, subject distance, and viewpoint.

To investigate this approach, we present a modular, accessible framework using consumer hardware for a life-size interactive mirror that combines stereo RGB cameras, AI-based HPE, 3D pose reconstruction, and real-time avatar rendering. The estimated body pose is mapped to a humanoid avatar and displayed on a life-size screen, providing a self-referential visual representation of the user's movements. The framework separates pose estimation, 3D reconstruction, avatar retargeting, and visualization, allowing different HPE models and avatar representations to be substituted independently. We demonstrate the technical feasibility of the system as a life-size virtual mirror integrating the tracked avatar with an interactive 3D environment. The main contributions of this work are a modular stereo RGB-based architecture for real-time 3D human pose tracking that includes a model-independent keypoint-to-VRM retargeting pipeline. We further discuss its potential as a foundation for future movement-feedback applications, including physiotherapy.

2 Related Work

HPE is the task of extracting the configuration of human bodies from images or videos, which typically involves the inference of body joint coordinates or *keypoints* [5]. It is commonly used in physiotherapy and rehabilitation to assess movement and treatment outcomes, especially in laboratory settings. For this, Marker-Based Stereophotogrammetry Systems (MBS) are the de-facto standard with several commercial solutions available, most notably *Vicon* [6,7]. MBS usually use infrared illumination and multiple cameras to track the positions of reflective markers placed on the patient's body. Hence, the necessary effort to set up such a system and its application in physiotherapy is very high. Alternatives to MBS include wearable Inertial Measurement Units (IMU), which require one or more devices worn on the relevant body parts. While those system are more affordable, the effort for setting up remains and additional challenges in terms of accuracy, reliability, and ease of use arise. Using depth sensing cameras enables markerless HPE and has therefore been investigated for motor rehabilitation and home-based exercise [6,7]. Markerless HPE from conventional RGB cameras offers a further reduction in hardware requirements and forms the basis of the approach presented in this work.

Recent markerless HPE approaches predominantly employ machine learning models to infer these keypoints directly from visual input [8]. HPE methods differ by the dimensionality of the estimated keypoints or joints (2D or 3D) and whether the pose of a single or multiple persons are estimated [9]. For the proposed system, single-person 3D pose estimation is required to drive a virtual avatar and represent the user's movement in the coordinate system of the physical mirror.

Virtual mirrors have subsequently been explored in rehabilitation and other interactive applications. Gaggioli et al. [10] developed a virtual mirror for post-stroke motor rehabilitation in which movements of the healthy limb were mirrored to support mental practice of movements with the impaired limb. Other systems [11,12] have used mirror-based interfaces for applications including anatomy education and virtual try-on, demonstrating that the mirror paradigm can combine a user's physical presence with digitally

generated visual information. Many existing virtual mirror systems are developed around a specific application and sensing or visualization setup. In contrast, our work focuses on a modular technical framework that combines stereo RGB-based HPE, 3D pose reconstruction, avatar retargeting, and life-size mirror rendering. This separation allows the sensing and avatar components to be adapted independently and provides a foundation for future movement-feedback applications, including physiotherapy.

3 System Architecture and Design

To achieve a modular design, we separate our system in backend and frontend. Figure 1 shows an overview over the proposed system. First, an offline calibration step is needed before running the processing pipeline. During calibration, the intrinsic and extrinsic parameters of the stereo cameras are determined and the camera coordinate systems are related to the coordinate system of the display. During runtime, RGB frames are processed by the HPE module, corresponding keypoints are triangulated into 3D coordinates, and the resulting pose is mapped onto a humanoid avatar. The avatar and virtual environment are subsequently rendered on the life-size display.

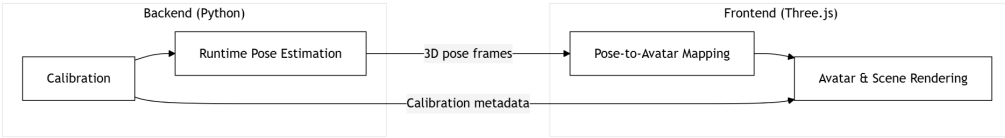


Fig. 1. Overview over the system architecture.

3.1 Hardware and Calibration

The physical setup consists of a large display and two cameras mounted on either side of the display for HPE. We use a vertical 75-inch display to provide a life-size interaction area and two standard webcams capturing HD images at 60 fps. The system is not tied to a specific display or camera model, provided that the cameras provide sufficient image quality for the selected HPE model and sufficient frame rate for the desired system performance.

The intrinsic parameters of all cameras are calibrated beforehand. The extrinsic calibration is then used to determine the position and orientation of the cameras relative to the display. For this purpose, a temporary third camera and a cube equipped with an ArUco [13] marker on each side are used. The temporary camera and cube are positioned such that all three cameras observe at least one of the cube's markers, while the temporary camera also observes the display, which shows a ChArUco board. The positions and orientations of the three cameras relative to the cube are determined from the visible markers, while the pose of the display relative to the cube is determined from the ChArUco board. These transformations enable the calculation of all camera extrinsics relative to the display. The resulting calibration defines a common coordinate frame for the cameras and display. The calibration parameters are transmitted to the frontend during initialization, where they are used to configure the virtual scene and display geometry. They are also required by the backend for triangulating corresponding keypoints from the two camera views.

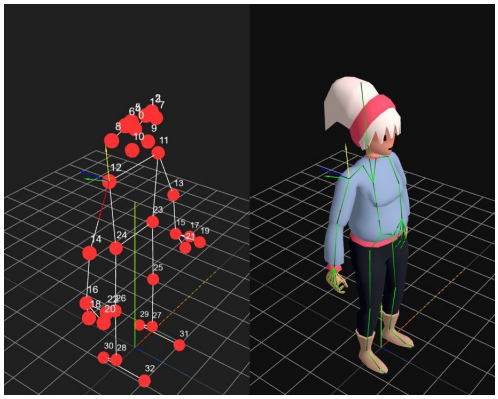
3.2 Human Pose Estimation

During runtime, the backend captures synchronized frames from the two display-mounted cameras. The selected HPE model is applied independently to each camera view to obtain two-dimensional keypoints. Corresponding keypoints from the two views are then triangulated using the calibrated camera parameters, producing three-dimensional keypoints in the common display coordinate system. The reconstructed pose is transmitted to the frontend for avatar retargeting and visualization. Our system currently supports *RTMLib* [14] and *MediaPipe* [15]. *RTMLib* is an open-source lightweight Python library for human and animal pose estimation and can perfom multi-person pose estimation. Its main advantage is its high flexibility as it offers multiple detectors. Depending on the Pose Estimator used, the number of keypoints can range between 17 (body only) and 133 (body including hands, feet, and detailed face keypoints)[14]. *MediaPipe* is a framework for perception pipelines by Google, offering a *Pose Landmarker* for HPE [16]. It uses the *BlazePose* keypoint model with 33 keypoints and provides three model versions: *Lite*, *Full*, and *Heavy* [17,15]. All three variants were evaluated during development. The *Lite* model was unable to reliably detect the complete pose across all frames under the conditions of the prototype. The *Full* and *Heavy* models produced comparable pose quality, while the *Heavy* model introduced higher processing latency. Since the difference in tracking quality between *Full* and *Heavy* was marginal, the *Full* model was selected as the best compromise between pose quality and latency.

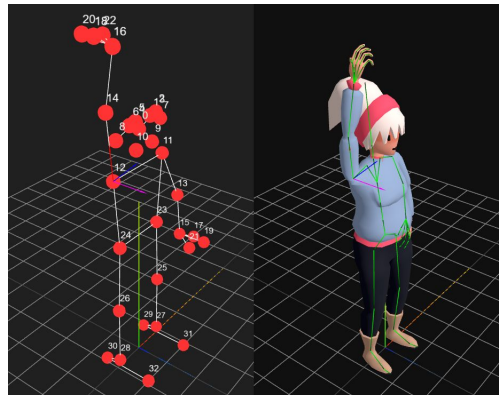
3.3 Pose-to-Avatar Mapping

The system uses the VRM [18] format to represent the virtual avatar. VRM is a runtime-loadable humanoid avatar specification that defines a standardized humanoid skeleton and supports a range of avatar representations. It further enables the creation of realistic avatars from images of a real person [19]. We selected VRM because its standardized bone structure allows the pose-to-avatar mapping to remain independent of a specific avatar model. A VRM avatar is animated by rotating bones defined in a hierarchy of parent-child relationships. Each bone's orientation is defined relative to its parent bone [20]. However, the human pose reconstructed by HPE is represented as a set of 3D keypoints. Therefore, we convert keypoint positions into orientations of the corresponding avatar bones to animate the VRM avatar based on the user's movements. We define a local coordinate system consisting of three orthogonal axes for each VRM bone that needs to be animated (we currently don't animate fingers). The primary axis represents the direction in which the bone points. The secondary axis defines the bone's orientation around the primary axis, and the tertiary one completes the right-handed orthogonal coordinate system. The primary direction can be obtained by the corresponding keypoints. However, the secondary and tertiary axes cannot be directly derived from keypoints and need to be constructed using a helper *up vector* that defines the orientation along the primary axis. This vector is not itself used directly as the final secondary axis. Instead, the tertiary axis is first calculated as the cross product of the primary direction and the configured up direction. The secondary axis is then calculated from the tertiary and primary axes, resulting in an orthogonal local coordinate system. Figure 2(a) shows the local coordinate system for the right upper arm, with the up vector highlighted in yellow.

For certain bone orientations, the configured up direction can become nearly parallel to the primary axis, leading to a degenerate configuration. To handle this case, we define an alternative helper vector that is sufficiently different from the up direction that is used instead in those cases. This approach is inspired by existing MediaPipe-to-avatar retargeting methods [21]. Figure 2(b) shows the local coordinate system for the right



(a) Left: BlazePose keypoints, right: corresponding pose using the VRM avatar (*Shiro* from OpenSourceAvatars).



(b) Left: BlazePose keypoints, right: corresponding pose using the VRM avatar.

Fig. 2. In both images, the coordinate system used for the right upper arm is shown: red is the primary axis, green the secondary, and blue the tertiary. The yellow arrow shows the specified helper up vector.

upper arm in the case where the up vector and the primary axis are nearly parallel, so the fallback helper vector is used instead (highlighted in magenta). The avatar root is positioned using the pelvis position, making the avatar follow the user's movements. In a calibration step, the lengths of its major bones are adjusted to match the corresponding distances between pose keypoints, adapting the avatar to the user's height and individual body proportions.

3.4 Virtual Scene and Perspective

To create the mirror effect, we render a virtual scene to be displayed on the screen consisting of the VRM avatar and a background environment. We render this scene from the user's head position and use off-axis projection constrained by the display dimensions. The resulting rendering produces a realistic perspective and parallax effect consistent with the user's viewpoint. To create a background that matches the physical environment, we use *Polycam* [22] to reconstruct the surroundings of the display, achieving a greater degree of immersion.

4 Evaluation and Results

The prototype was evaluated with respect to the robustness and practical behavior of the pose tracking pipeline under conditions relevant to the intended mirror environment. In particular, we compared the supported HPE frameworks with respect to subject distance, partial occlusions, additional people in the camera view, and temporal stability. We additionally examined the behavior of the resulting pose when retargeted to the virtual avatar and used the interactive prototype to assess the feasibility of the complete system.

4.1 HPE Performance

The HPE models were quantitatively evaluated using a standalone test script processing a fixed 30-second sequence. The benchmark measured the average time spent in HPE

inference as well as the total processing time per frame, including frame acquisition and visualization. Since the benchmark was performed independently of the complete system, these measurements characterize the HPE component rather than the end-to-end latency of the proposed system. We reached about 30 fps with the fastest HPE model needing only 5.05 ms per frame.

The benchmark was complemented by a qualitative comparison under conditions relevant to the intended mirror environment. The same camera setup was used to examine the behavior of the HPE frameworks with respect to subject distance, partial occlusions, temporal stability, and additional people in the camera view. Here, *MediaPipe Full* provided more consistent tracking across the tested conditions. In particular, it maintained a recognizable pose at larger subject distances and was less affected by partial occlusions. It also provided more stable tracking when additional people entered the camera view and exhibited less visible frame-to-frame jitter. The *RTMLib* configurations showed greater sensitivity to these conditions, with detections becoming increasingly unstable as the subject moved farther from the cameras and with more frequent failures under partial occlusion.

Based on the combined quantitative and qualitative observations, *MediaPipe Full* was selected as the primary HPE framework for the prototype. While the fastest *RTMLib* configurations achieved a similar processing rate in the standalone benchmark, *MediaPipe Full* exhibited more consistent tracking under the tested conditions and provided a favorable balance between processing speed and pose quality.

4.2 Interactive Prototype



(a) Physiotherapy step application with avatar.

(b) Yoga application with skeleton feedback.

(c) Interactive demo game.

Fig. 3. Interactive application prototypes using our display in different settings.

To demonstrate the application potential of the system, we consider lower-limb physiotherapy exercises focusing on knee joint stabilization, with the mirror indicating alignment deviations from an ideal leg axis. We developed one interactive prototype (see

Fig. 3) that illustrates the application in a physiotherapy (a) resp. yoga (b) setting providing visual feedback either with an avatar (a) or a skeleton model (b). In particular, the system can support the estimation of clinically relevant parameters such as the dynamic knee valgus (DKV) angle to assess whether movement remains within a physiologically acceptable range.

Another interactive prototype game (see Fig. 3 (a)) was presented in a university setting and demonstrated the feasibility of real-time interaction with the virtual environment. Most participants intuitively understood the objective of the game and were able to interact with the virtual objects through their avatar. One participant even turned around and interacted without looking at the screen, suggesting that the spatial relationship between the user's movements and the virtual scene can become sufficiently intuitive to support interaction without continuously observing the virtual objects. We further observed that the HPE was very stable even when users wear dark and loose-fitting clothing. These observations demonstrate the feasibility of using the proposed system for real-time movement-based interaction, and allow a formal evaluation of embodiment or usability in controlled studies in the future.

5 Conclusions

The results demonstrate the technical feasibility of the proposed interactive mirror as a real-time movement tracking and visualization system. The complete pipeline combines stereo RGB cameras, markerless HPE, 3D keypoint reconstruction, avatar retargeting, and perspective-aware rendering into a single interactive environment. A major advantage of the proposed approach is its use of conventional RGB cameras rather than specialized motion-capture hardware. Compared with marker-based systems, the setup does not require reflective markers, dedicated infrared cameras, or extensive preparation of the user. The system calibration has to be performed only once unless the physical placement of the display or cameras changes, and can be completed in under five minutes. This reduces the hardware and setup requirements and makes the approach more suitable for interactive environments outside specialized laboratories. The modular architecture further supports scalability, since the HPE component, keypoint mapping, and avatar representation can be adapted independently. However, the use of commodity cameras also introduces limitations in tracking robustness and accuracy compared with specialized motion-capture systems. The current evaluation therefore demonstrates practical feasibility rather than equivalent measurement accuracy. We plan an updated setup that includes the computation unit as well as hardware-synchronized cameras optimized for low-latency.

The physiotherapy application ultimately requires systematic evaluation with representative users and exercises. We will further investigate not only the technical accuracy of the pose estimation and movement measurements, but also whether the provided visual feedback is understandable, useful, and appropriate for therapeutic exercise. A future study will compare the proposed system with conventional exercise instruction or therapist-guided feedback and investigate whether the system can support independent exercise while maintaining sufficient movement quality.

References

1. Abdullah, N., Al-Kazzaz, S.: Evaluation of physiotherapy exercise by motion capturing based on artificial intelligence: A review. Al-Rafidain Engineering Journal (AREJ) **28** (2023)

2. Preston, C., Kuper-Smith, B.J., Ehrsson, H.H.: Owning the body in the mirror: The effect of visual perspective and mirror view on the full-body illusion. Scientific Reports **5**(1) (2015)
3. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 7122–7131
4. Zimmermann, C., Welschehold, T., Dornhege, C., Burgard, W., Brox, T.: 3d human pose estimation in rgbd images for robotic task learning. In: IEEE International Conference on Robotics and Automation (ICRA). (2018)
5. Desmarais, Y., Mottet, D., Slangen, P., Montesinos, P.: A review of 3d human pose estimation algorithms for markerless motion capture. Computer Vision and Image Understanding **212** (2021)
6. Milosevic, B., Leardini, A., Farella, E.: Kinect and wearable inertial sensors for motor rehabilitation programs at home: state of the art and an experimental comparison. Biomedical engineering online **19**(1) (2020)
7. Roggio, F., Ravalli, S., Maugeri, G., Bianco, A., Palma, A., Di Rosa, M., Musumeci, G.: Technological advancements in the analysis of human motion and posture management through digital devices. World journal of orthopedics **12**(7) (2021)
8. Roggio, F., Trovato, B., Sortino, M., Musumeci, G.: A comprehensive analysis of the machine learning pose estimation models used in human movement and posture analyses: A narrative review. Heliyon **10**(21) (2024)
9. Ben Gamra, M., Akhloufi, M.A.: A review of deep learning techniques for 2d and 3d human pose estimation. Image and Vision Computing **114** (2021)
10. Gaggioli, A., Morganti, F., Meneghini, A., Alcañiz, M., Lozano, J.A., Montesa, J., Martínez Sáez, J.M., Walker, R., Lorusso, I., Riva, G.: The virtual reality mirror: Mental practice with augmented reality for post-stroke rehabilitation. In: Annual Review of CyberTherapy and Telemedicine: Interactive Media in Training and Therapeutic Interventions. Volume 3. Interactive Media Institute, San Diego, CA, USA (2005)
11. Blum, T., Kleeberger, V., Bichlmeier, C., Navab, N.: mirracle: An augmented reality magic mirror system for anatomy education. In: 2012 IEEE Virtual Reality Workshops (VRW). (2012)
12. Huang, Z., Zhao, W., Guo, T., Huang, J., Li, P., Sheng, B.: Magicmirror: A 3-d real-time virtual try-on system through cloth simulation. In: Advances in Computer Graphics, Cham, Springer Nature Switzerland (2024)
13. Monkman, G.G., Hyder, K., Kaiser, M.J., Vidal, F.P.: Using machine vision to estimate fish length from images using regional convolutional neural networks. Methods in Ecology and Evolution **10**(12) (2019)
14. Jiang, T.: rtmlib (2023) Accessed September 24, 2026.
15. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for building perception pipelines (2019)
16. Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., Grundmann, M.: Blazepose: On-device real-time body pose tracking (2020)
17. Google: Pose landmarker (2026) Accessed August 15, 2026.
18. Iwaki, S., Kurata, T.: VRM: Concept and Implementation of an Application-Interoperable Avatar Format . In: 2025 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Los Alamitos, CA, USA, IEEE Computer Society (2025)
19. Kondo, N., Asano, Y., Ochiai, Y.: Instant skinned gaussian avatars for web, mobile and vr applications (2025)
20. VRM Consortium: VRMC_vrm.humanoid: VRM 1.0 Humanoid Specification (2026) Accessed August 13, 2026.
21. SunnyView Technology Co., L.: Animating 3d characters with mediapipe (Apr 2026) Accessed June 30, 2026.
22. Polycam: Polycam: Ai 3d scanning, floor plans & reality capture (2026) Accessed August 15, 2026.