Enhanced real-time motion transfer to 3D avatars using RGB-based human 3D pose estimation

ILIAS POULIOS, THEODORA PISTOLA, SPYRIDON SYMEONIDIS, SOTIRIS DIPLARIS, KONSTANTINOS IOANNIDIS, STEFANOS VROCHIDIS, and IOANNIS KOMPATSIARIS, Information Technologies Institute - CERTH, Greece

Human motion transfer on 3D avatars has witnessed substantial progress, driven by the advancements of 3D pose estimation using RGB data. This technology analyzes human movements captured through RGB cameras, enabling tracking of 3D body landmarks and leading to the animation of 3D avatars. Utilizing RGB input offers a range of advantages, democratizing avatar creation by eliminating the need for specialized equipment, such as sensors, markers, or specialized studios. Recent years have seen remarkable strides in this field, leveraging deep learning models and sophisticated computer vision algorithms to capture intricate movements and gestures from RGB video footage. This study introduces a novel real-time approach leveraging RGB input to generate realistic 3D animations. It comprises three phases: i) 3D human pose estimation using MediaPipe, ii) correction of MediaPipe's landmarks' inaccuracies, especially regarding depth dimension, and incorporation of bones' rotation information, and, finally, iii) transfer of the motion to the target 3D avatar.

Additional Key Words and Phrases: 3D Pose Estimation, Avatar Animation, MediaPipe, Deep Learning, Computer Vision

1 INTRODUCTION

3D human body pose estimation is the process of determining the three-dimensional (3D) configuration of a person's body from input data such as images or videos. It involves identifying key body joints or landmarks and estimating their positions and orientations in 3D space. The goal is to accurately reconstruct the pose of the human body, including the positions and orientations of body parts such as the head, torso, arms, and legs, in 3D coordinates. Because of its extensive applications across various domains, including motion capture for animation and gaming, human-computer interaction, healthcare, sports biomechanics, robotics, surveillance and security, try-on and fashion, and various VR and AR applications, 3D human pose estimation has garnered significant interest in the computer vision field.

Camera-based 3D human pose estimation holds paramount significance for Virtual Reality (VR) and Augmented Reality (AR) applications as it easily enables the capturing and interpretation of human movements in real-time. In VR environments, where users interact with virtual worlds through immersive experiences, accurate pose estimation is crucial for enabling natural and intuitive interactions. By tracking the movements of users' bodies, VR systems can render realistic avatars and objects that respond dynamically to users' actions, enhancing the sense of presence and immersion. Similarly, in AR applications, where virtual elements are overlaid onto the real world, precise pose estimation enables seamless integration of virtual objects into the user's environment, facilitating interactive experiences such as virtual try-on, gaming, and educational simulations. Moreover, camera-based pose estimation can be leveraged in healthcare, sports training, and motion analysis, offering valuable insights into human movement patterns and facilitating personalized coaching and rehabilitation programs. Overall, the accurate and real-time estimation of 3D human poses using camera-based techniques is pivotal for unlocking the full potential of VR and AR technologies, enabling immersive, interactive, and engaging experiences across various domains.

Authors' address: Ilias Poulios, ipoulios@iti.gr; Theodora Pistola, tpistola@iti.gr; Spyridon Symeonidis, spyridons@iti.gr; Sotiris Diplaris, diplaris@iti.gr; Konstantinos Ioannidis, kioannid@iti.gr; Stefanos Vrochidis, stefanos@iti.gr; Ioannis Kompatsiaris, ikom@iti.gr, Information Technologies Institute - CERTH, Thessaloniki, Greece.

Recent advances in 3D human body pose estimation technology help us move a step forward towards accurate real-time transfer of human motion to an immersive environment using a single camera. Deep learning methods, particularly Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), have significantly enhanced accuracy and robustness in pose estimation tasks. Multi-view fusion techniques have also played a crucial role by integrating information from multiple camera viewpoints to improve performance, especially in challenging scenarios. Attention mechanisms have further refined these models by enabling them to focus on relevant features or body parts, particularly useful in crowded scenes or complex poses. Additionally, self-supervised learning approaches and generative models have expanded the capabilities of pose estimation systems, allowing for better utilization of unlabeled data and generating realistic and diverse 3D human poses. Finally, efforts to optimize algorithms for real-time performance have made these systems more practical for applications in live streaming, Virtual Reality (VR), and Augmented Reality (AR) environments.

Challenges in 3D human body pose estimation persist despite recent advancements. One significant hurdle is the inherent complexity of human body articulation, which includes a wide range of movements and poses. Ambiguities in pose interpretation, such as occlusions and self-occlusions, pose challenges for accurate estimation, particularly in cluttered environments. Limited availability of in-the-wild annotated datasets, especially for diverse populations and activities, hinders the training of robust models, leading to uncertainties in depth perception and unrealistic human poses. Furthermore, issues related to real-time performance, including computational efficiency and latency, remain areas of concern, particularly for applications requiring rapid processing. Additionally, ensuring generalization across various body shapes, sizes, and clothing types presents a persistent challenge for pose estimation algorithms.

In this study, our focus lies on real-time human movement transfer from monocular RGB input to a target 3D avatar within an immersive environment, using 3D human pose estimation technology. More specifically, we want to enable real-time 3D avatar animation in a Unity[1] environment using a single RGB camera with the ultimate aim to develop an eXtended Reality (XR) application for rehabilitation purposes. To achieve our goal, initially, we employ MediaPipe's 3D pose estimation algorithm[2] to extract human body landmarks' coordinates in 3D space. However, inaccuracies in the estimated depth information often degrade the quality of the final result. Moreover, the MediaPipe algorithm's output lacks bone longitudinal rotation insights, occasionally resulting in unrealistic human poses. Consequently, when employing the original MediaPipe output, we may encounter unrealistic human motions within the Unity environment, particularly in scenarios where the viewpoint is suboptimal or the pose is particularly challenging.

To tackle the challenges outlined above, we propose two intermediary steps between the landmark extraction and the final step of motion transfer to the target 3D avatar. These two steps aim to: i) address MediaPipe's output's inaccuracies, particularly for the estimation of the depth dimension, and ii) enrich MediaPipe's output with bones' longitudinal rotation information. Our approach consists of two sequential fully connected deep neural networks: the Landmarks' Correction Network and the Bones' Rotation Information Enrichment Network. The Landmarks' Correction Network receives MediaPipe's output (including the 3D coordinates of body landmarks) as its input and output a corrected version of their 3D coordinates. Subsequently, the Bones' Rotation Information Enrichment Network takes the corrected MediaPipe output as its input and extracts an enhanced version of the corrected MediaPipe's output, which incorporates bones' rotation information. Both deep neural networks were trained using the ground-truth data from the TotalCapture dataset [20], which we augmented with bones' longitudinal rotation information specifically for training the Bones' Rotation Information Enrichment Network. In this manuscript, we begin by outlining the challenges we encountered

[1]https://unity.com/
[2]https://developers.google.com/mediapipe/solutions/vision/pose_landmarker

and then detail all the steps of the development of our proposed system aiming to enhance MediaPipe's 3D human pose estimation output and adapt it to real-time 3D avatar animation needs. We present our methodology, describing the architecture and training configuration of our two deep neural networks, along with the process that we followed to generate an enhanced version of TotalCapture's ground-truth to be used for the training of our Bones' Rotation Information Enrichment Network and the procedure of mapping the final estimations of 3D body landmarks onto the target 3D avatar. Additionally, we present the results of our final evaluation, which encompasses both pose estimation accuracy and inference performance. Our proposed Landmarks' Correction Network effectively reduced the Mean Per Joint Position Error (MPJPE) by 3.5 cm, achieving a notable improvement from 7.4 cm to 3.9 cm compared to the original MediaPipe output.

The rest of this paper is organized as follows. In section 2, we briefly present the most prominent related works in the field of 3D human body pose estimation divided into specific categories. Section 3 describes in detail the three steps of our proposed approach. Then, in section 4, the training details of our networks, as well as the experimental results are provided. Finally, we conclude our paper with section 5, where we discuss the main outcomes and refer to our future steps to further improve our work.

2 RELATED WORK

Until recently, motion capture technology was the most common method used to transfer the movement of a human to a target 3D avatar. This technology involved recording the movements of actors using special suits equipped with markers or complex sensors, which were then translated onto 3D avatars [2]. While this motion capture technology offers more realistic animations, it is expensive and often requires specialized equipment and facilities. With the recent advances in computer vision and artificial intelligence (AI), RGB-based human 3D pose estimation has significantly advanced animation making by providing cost-effective, flexible, and real-time solutions for capturing human motion. These methods leverage affordable RGB cameras, allowing motion capture in diverse environments without the need for specialized equipment. Moreover, their non-intrusive nature eliminates the requirement for physical markers on subjects, enhancing comfort during capture sessions. Providing this widespread accessibility, RGB-based methods democratize motion capture technology, empowering animators of all skill levels to create realistic animations efficiently. In this section, we present a brief overview of relevant studies in the domain of human 3D pose estimation from RGB video input. The related work is divided into 3D human pose estimation skeleton-only methods and human mesh recovery methods. The differences between the two categories are clarified in the following subsections. In addition, we provide a brief overview of related research works that are focused on refining the output of MediaPipe's 3D pose estimation algorithm, similar to our approach, or enriching the estimated 3D poses with supplementary information to improve the final data for motion transfer to the 3D avatar.

2.1 Skeleton-only methods

3D human pose estimation skeleton-only methods focus solely on estimating the positions of skeletal joints without considering the surface geometry of the human body. These methods typically output a simplified representation of the human body as a skeletal structure.

VNect [11] is a single-person 3D pose estimator that runs in real-time. It consists of a Convolutional Neural Network (CNN) part that regresses 2D and 3D joints on images, and a skeleton fitting part that combines the 2D-3D predictions with the temporal history of the sequence to produce a temporally stable, camera-relative, full 3D skeletal pose. To ensure skeletal stability (e.g. fixed-size bones), during the skeleton fitting process, the regressed 3D joints are converted into

joint angles while keeping only the 3D position of the root joint (hips). XNect [10] is a more advanced version of VNect that supports multi-person 3D pose estimation. Another method is PhysCap [18] that introduces physics constraints in the process of 3D pose estimation. It uses VNect as a backbone to get 3D skeleton joints from images and then passes them through a physics-based pose optimizer in order to output a physically plausible and temporally stable human motion without deformations like floor penetrations or foot skating. To achieve that, it also utilizes a trained neural network to detect foot contact events on images. In contrast to previous methods, Single-Shot Multi-Person Absolute 3D Pose Estimation (SMAP) [24] employs a different approach by initially regressing a series of 2.5D representations of body parts. Subsequently, it reconstructs the 3D absolute poses based on these 2.5D representations using a depth-aware part association algorithm. This single-shot bottom-up strategy enables the system to enhance its understanding and reasoning regarding inter-person depth relationships, thereby enhancing both 3D and 2D pose estimation from a single RGB image. In MotioNet [17] a deep neural network with embedded kinematic priors is used to decompose sequences of 2D joint positions into two separate attributes: a symmetric skeleton and a sequence of 3D joint rotations associated with global root position and foot contact labels. These attributes are fed into a Forward Kinematics (FK) layer that outputs 3D positions. In [21] a single-person 3D pose estimation model named Pose Estimation using TRansformer (PETR) is proposed, which uses a High-Resolution Net (HRNet) [19] to estimate 2D joints on images and then passes those joints through a transformer encoder network combined with a fully connected layer and outputs the final 3D pose. In the same work, a second model, named Pose Estimation on Bone Rotation using Transformer (PEBRT), is presented. PEBRT follows a similar architecture to PETR, but instead of 3D positions, it predicts the rotation matrices for each bone using a 2D pose as input. The rotation matrices are applied to a T-pose skeleton model to get the final 3D pose. No temporal information or receptive fields are required to generate kinematically realistic human poses. In MocapNet [12] they use 2D joints as input, while the output is a BVH (BioVision Hierarchical data) animation file. The 2D joints are first splitted into groups (lower body, upper body and hands) and used to form the enhanced Normalized Signed Rotation Matrices (eNSRM), which encode a relation between each pair of the joints. Both the matrices and the 2D joints are fed into a neural network ensemble and the outputs of the networks are further refined using Hierarchical Coordinate Descent (HCD). Furthermore, in Ray3D [23] they propose a lifting method to map 2D human pose keypoints to 3D. To eliminate the impact of camera parameters variations the 2D keypoints are converted into 3D rays firstly in Camera Coordinate System (CCS) and then in Normalized Coordinate System (NCS). Subsequently, they are fed to pose estimation network and trajectory network to predict the final 3D pose. With unnormalization, the 3D pose under world coordinate system is obtained. By conducting experiments on four benchmarks they show that their method significantly outperforms existing state-of-the-art models.

Google has also developed a 3D human pose estimator that is integrated into its open-source, cross-platform framework, called MediaPipe[3]. MediaPipe enables building on-device machine learning algorithms to analyze arbitrary data such as video, audio and text. Its main advantage is the ready-to-use solutions it offers on many machine learning tasks. These solutions are light enough to run on any device and in real-time (either on CPU or GPU). Concerning the 3D human pose estimation, MediaPipe offers the Pose Landmark Detection solution [4], which is based on the BlazePose [1] method. BlazePose is a lightweight CNN architecture tailored for real-time human pose estimation on mobile devices. During inference, the network generates 33 body keypoints for a single person and runs at over 30 frames per second on a Pixel 2 phone. This rapid performance makes it well-suited for demanding real-time use applications, such as fitness tracking and sign language recognition. Notable contributions of BlazePose include an innovative body pose tracking

[3]https://developers.google.com/mediapipe
[4]https://developers.google.com/mediapipe/solutions/vision/pose_landmarker

solution and a lightweight body pose estimation neural network that combines heatmaps and regression techniques to estimate the keypoints' coordinates while ensuring computational efficiency. The variant of BlazePose model integrated into the MediaPipe framework utilises GHUM [22], a 3D human shape modeling pipeline, to estimate the complete 3D body pose of an individual from images or videos. Additional insights into the MediaPipe 3D human pose estimation algorithm are provided in Subsection 3.2, as we utilize this algorithm within the framework of the proposed method outlined in this paper.

2.2 Human mesh recovery methods

Human mesh recovery methods aim to reconstruct the complete 3D surface geometry of the human body, including the skin or clothing. These methods generate a detailed mesh representation that captures the shape and appearance of the human body in three dimensions, often incorporating finer details such as clothing folds and surface textures.

Video Inference for Body Pose and Shape Estimation (VIBE) [6] estimates both body pose and shape using RGB video input. Its output is a sequence of pose and shape parameters in the SMPL [8] body model format. This work introduces a recurrent architecture that propagates information over time. It proposes a discriminative training of motion sequences using the AMASS dataset [9] and a self-attention mechanism in the discriminator so that it learns to focus on the important temporal structure of human motion, as well as a new motion prior (MPoser) from AMASS. The result generated by VIBE can easily be converted into 3D avatar animation files; however, the resulting avatar remains stationary at the axes' origin and lacks dynamic movement within the 3D environment. Human Motion Model for Robust Estimation of temporal pose and shape (HuMoR) [13] introduces an expressive generative model, employing a conditional variational autoencoder, to capture pose changes throughout motion sequences. Additionally, it presents a versatile optimization method, utilizing HuMoR as a motion prior, to effectively estimate feasible pose and shape from uncertain inputs. Extensive assessments confirm the model's ability to generalize across various motions and body types, following training on extensive motion capture data. Moreover, it facilitates motion reconstruction from diverse input modalities, such as 3D keypoints and RGB(-D) videos. This method, though, cannot be used for real-time applications as it requires multiple iterations. Furthermore, [3] introduces Pose2Mesh, which is a system based on graph convolutional neural networks (GraphCNN) that directly estimates the 3D coordinates of human mesh vertices from the 2D human pose. Utilizing the 2D human pose as input offers essential human body articulation information, while maintaining relatively uniform geometric properties between the two domains. Moreover, the proposed system overcomes representation challenges by effectively leveraging mesh topology through a GraphCNN in a coarse-to-fine fashion.

2.3 Refinement methods for 3D pose estimation outputs

As our research revolves around the utilization and refinement of the MediaPipe 3D pose estimation algorithm, it is pertinent to acknowledge prior efforts that have also focused on improving the accuracy and effectiveness of MediaPipe's output.

In contrast to other technical solutions, Google's Mediapipe stands out as a leading framework for 3D human body recognition. Nonetheless, despite its maturity, Mediapipe still exhibits several shortcomings in accurately detecting 3D human posture. In [7], the authors address the issue of inaccurate human pose detection by MediaPipe's algorithm by employing several strategies. Firstly, they enhance the accuracy of 2D human pose detection by applying a speed threshold correction method to each joint. Secondly, to rectify inaccuracies in the depth (Z value) captured by monocular cameras, they statistically correct the Z value of joint points based on human tilt angles. Additionally, they normalize the

simulated proportions of each body limb to accurately correct the Z value of human pose under varying body postures. Finally, to mitigate jitter, lag, and periodic noise in multiple frames caused by changes in joint speed, the authors apply one euro filtering and mean filtering to the joint data. Through extensive testing on individuals of different heights, weights, ages, and genders, their study confirms that the improved Mediapipe achieves a 3D human pose detection accuracy of over 90% in multi-pose recognition tests. However, the aforementioned study does not support real-time applications. Additionally, the authors of [4] endeavor to conduct an initial evaluation of the accuracy and suitability of the MediaPipe library for applications such as physical therapy. Their findings reveal that pose estimation accuracy is strongly influenced by factors such as the camera's viewing angle and the specific exercise being performed. While optimal conditions yield high accuracy, the performance diminishes significantly under less favorable conditions. Moreover, in [16], the authors propose an intensive feature consistency (IFC) network, which refines Mediapipe's 2D landmarks to generate an accurate and stable skeleton. The IFC network utilizes a global body intensity module, which tracks the body position in the frame, and a local joint part adjustment module which ensures the joints' location distribution to be concentrated. The method reduces the impact of body joint movement diversity by interpreting long-term consistent view. The network was tested on two benchmark datasets and has shown an improvement of 99.1% of Percentage of Correct Key-points (PCK) body and 94.7% of PCK torso accuracy under 31 frames per second (FPS) on a CPU. The method is well-suited for fitness applications and mobility activities.

2.4 Longitudinal rotations estimation methods

Although commonly used in human pose estimation methods, minimal stick figure representations, as the one shown in Figure 3, do not provide the needed kinematic information to compute the six degrees-of-freedom (DoF) of each bone in space. Each 'stick' is defined only by two points however, at least three non-collinear points are required to fully define the orientation of the local reference frame of a body segment [14]. In this section, we provide a brief overview of some scientific papers that aim to tackle this problem by enriching the stick figures' representations with extra information to include the longitudinal bones' rotations.

STONES [5] is such an example, that presents an innovative machine learning technique designed to estimate those rotations from a minimal set of body points. This approach utilizes a recurrent deep neural network, which takes 3D joint positions from a simplified stick figure representation, typical of data obtained from conventional depth camera sensors, and accurately predicts longitudinal segment rotations. Validation of this method was conducted in an exergaming context, including activities like lunges, squats, and kicks, which are increasingly prevalent in various healthcare domains. Estimations demonstrated an accuracy exceeding 98% with mean errors of approximately 1 degree. Notably, this deep learning approach outperforms other machine learning strategies and achieves accuracy levels comparable to state-of-the-art motion capture systems while maintaining real-time processing speeds. Moreover, in Motion Envelopes (ME) [14] they use a geometric method that generates a surface based on the line segment defined by two points that compose a body segment, obtained by linear interpolation of the trajectories traced through time. By computing the normal vector to this surface, it is then possible to determine a third vector perpendicular by applying a cross-product between the previous ones and, therefore, to estimate the 3D orientation of the local reference frame of the segment. The proposed ME method enabled the estimation of the segments orientation of a stick figure model during gait movements for lower limbs and upper arms.

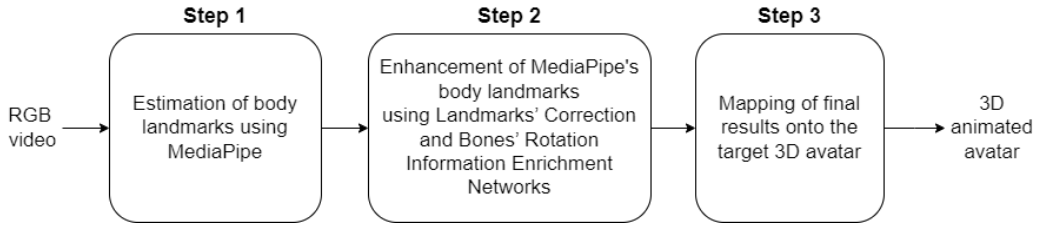

Fig. 1. The pipeline of our proposed approach.

3 THE PROPOSED APPROACH

In this section, we provide a detailed description of our proposed approach for capturing human motion from a conventional RGB video and transferring it onto a target 3D avatar within an immersive environment.

As we mentioned before, our aim is to enable real-time translation of human body motion from RGB input to a 3D avatar. To achieve this goal, we devised a system comprising three principal phases: i) 3D human pose estimation, ii) correction of MediaPipe's landmarks' inaccuracies, especially at the depth dimension, and incorporation of bones' rotation information, and iii) transfer of the motion to the target 3D avatar. In Figure 1 there is a simplified diagram of the proposed system's pipeline. This section begins with an overview of the TotalCapture dataset that we employed throughout our method's implementation. Then we analyze each one of the three phases of the proposed system, providing a comprehensive overview of our methodology in the subsequent subsections.

3.1 The TotalCapture Dataset

Throughout the implementation of our proposed system, we utilized the TotalCapture dataset [20], which employs 8 calibrated, static full HD RGB cameras along with 13 Inertial Measurement Units (IMUs) positioned on different body parts. Captured in an indoors environment within a 4x6 meter volume, the dataset offers synchronized video, IMU data, and Vicon[5] labeling, totaling $\approx$ 1.9 million frames across multiple subjects, activities, and viewpoints. Ground-truth poses provide 21 pixel-accurate 3D joint positions and angles derived from the Vicon marker-based motion capture system, ensuring reliable ground-truth labeling. This dataset comprises 4 male and 1 female subjects, each performing 5 diverse activities repeated 3 times —*Walking, Acting, Running, Freestyle and range of motion sequences (ROM)*— and presents challenging scenarios including yoga, giving directions, bending over, and crawling. For our study, we solely utilized the camera and ground-truth data. The ground-truth is provided in CSV and BVH file formats. The CSV files contain 21 3D joint positions in global coordinates. The BVH files contain 27 3D joint positions in global coordinates and, also, contain the rotation of the joints. In our work, we use the data provided in the BVH files, as the bones' rotation information is important for our final resutls. Figure 2 shows some video frame examples that come from the TotalCapture dataset. The examples depict some poses that subject 1 performs during walking, acting and freestyle activities captured from camera 1. In Figure 3 we can see the 27 3D joints provided in the BVH ground-truth files.

3.2 Step 1: 3D Pose Estimation using MediaPipe

In the context of our work, we employed MediaPipe's 3D body landmark detection algorithm, which extracts human's 3D body joints in real-time given an RGB input. The MediaPipe Pose Landmarker is designed to identify human body

[5]https://www.vicon.com/

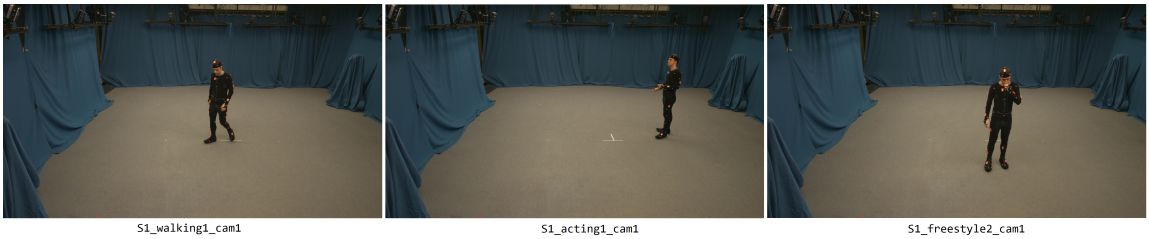

Fig. 2. TotalCapture video frame examples.

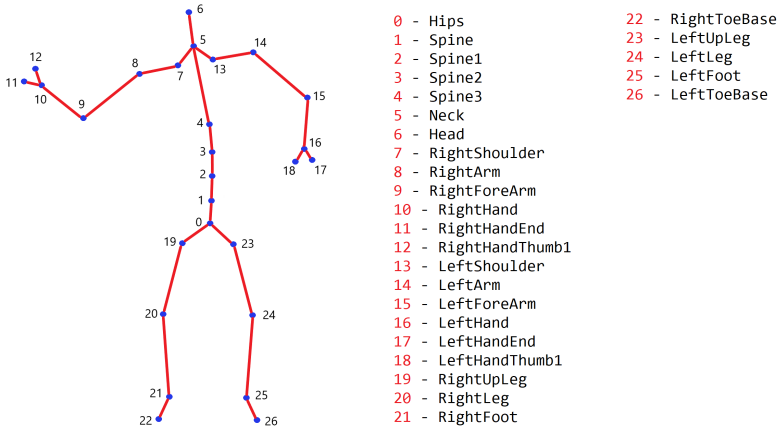

Fig. 3. The skeleton of TotalCapture BVH ground-truth files.

landmarks within images or videos, facilitating the analysis of posture and movement categorization. Leveraging Machine Learning (ML) models, it operates effectively with both single images and video sequences, providing body pose landmarks in both image coordinates and 3D world coordinates. Specifically, this algorithm employs a multi-stage approach to predict pose landmarks. Initially, a pose detection model identifies human bodies within the input image or video frame, while a subsequent pose landmarker model yields estimates for 33 3D body pose landmarks. These 33 body landmarks are shown in Figure 4.

For the pose estimation process Mediapipe uses the BlazePose model. BlazePose is a lightweight Convolutional Neural Network (CNN) which has an architecture similar to MobileNetV2 [15], specifically tailored for real-time on-device fitness applications. BlazePose [1] integrates GHUM [22], a 3D human shape modeling pipeline, facilitating the precise estimation of an individual's complete 3D body pose from images or videos. The 33 output landmarks are represented either as 3D normalized coordinates in image space or 3D world coordinates in camera space. As far as the normalized coordinates, the x and y values range from 0.0 to 1.0 relative to the video frame's width and height, respectively. The z coordinate, denoting the landmark's depth, is measured from the midpoint of the hips which serves as a reference point and the smaller the value the closer the landmark is to the camera. The magnitude of z uses roughly the same scale as x. On the other hand, the x, y and z values of the world coordinates are measured relative to the hips and they are in meters. The *visibility* parameter for each body landmark, ranging from 0.0 to 1.0, indicates the likelihood of the landmark being visible and unobscured in the video frame. For estimating 3D body landmarks, the MediaPipe framework provides three

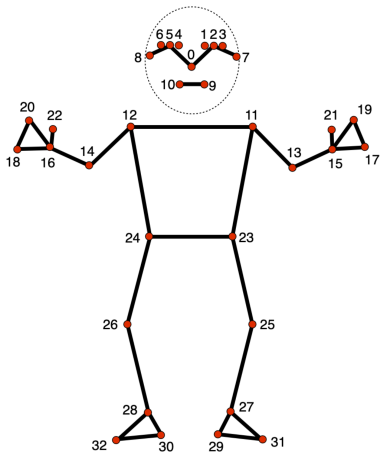

Fig. 4. The 33 body landmarks detected by MediaPipe's 3D Pose Landmarker.

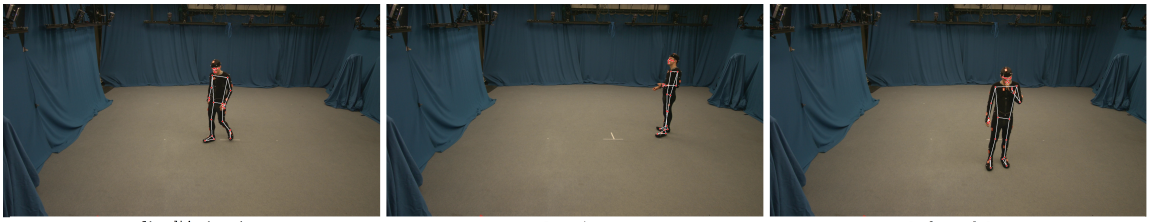

Fig. 5. MediaPipe results on TotalCapture video frame examples.

models with varying complexities (0, 1, or 2), where reduced complexity leads to faster execution but lower estimation accuracy. In our work, we used the model with complexity 1, to keep the balance between estimation accuracy and real-time performance.

3.2.1 Challenges of MediaPipe's 3D Pose Estimator. Though the efficiency gains in inference time offered by MediaPipe's 3D pose estimation algorithm are undeniable, only a few studies have been conducted concerning its qualitative evaluation. As mentioned in [4], MediaPipe's pose estimation is highly dependent on the camera's viewing angle as well as the performed exercise. Through our research, we found that specific limitations persist in accurately determining the depth of each body landmark in certain situations. In Figure 6 we provide an example, where we estimate a pose from a TotalCapture's video frame using MediaPipe's algorithm. The blue dots represent the estimated MediaPipe body landmarks, whereas the red ones are those from the TotalCapture ground-truth data. When examining these 3D points from a frontal perspective, it looks that MediaPipe's estimations closely resemble those of the ground-truth. However, when viewed from the side, it becomes evident that MediaPipe inaccurately estimates depth (depicted by the blue dots). Additionally, MediaPipe lacks any information regarding bones' longitudinal rotation, as it solely outputs 3D points. These inaccuracies and deficiencies pose challenges when transferring the estimated poses to a target 3D avatar, resulting in unrealistic human movements.

Table 1. Correspondance of landmark indices between MediaPipe and TotalCapture skeletons

Mediapipe	11	12	13	14	15	16	17	18	19	20	23	24	25	26	27	28	31	32
TotalCapture BVH	8	9	10	11	12	14	15	16	17	18	19	20	21	22	23	24	25	26

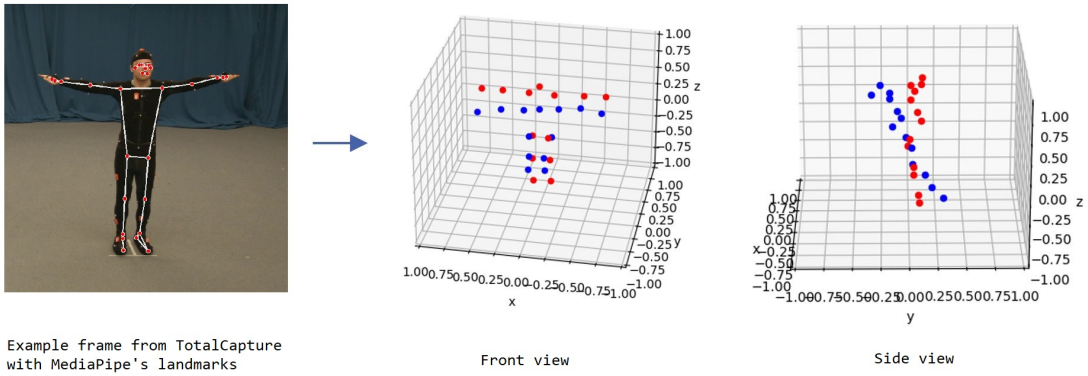

Fig. 6. Example of MediaPipe's inaccurate depth estimation. Blue dots represent the estimated 3D points of Mediapipe, while red dots are the 3D points provided by the TotalCapture ground-truth.

The main focus of this work is to eliminate the depth inaccuracies and enhance the corrected MediaPipe's results with bones' rotation information using Deep Learning (DL).

3.3 Step 2: Enhancement of MediaPipe's Results

As we already mentioned, the purpose of our proposed method is twofold. Firstly, to eliminate MediaPipe's pose estimation errors primarily on depth estimation. Secondly, to enrich MediaPipe's output with extra information so as to include the longitudinal bones' rotations. To achieve both of our goals, we trained two different neural networks, which we describe in the following subsections.

3.3.1 Landmarks' Correction Network. For the automatic correction of the MediaPipe output's inaccuracies, we trained a fully connected neural network that takes the MediaPipe's world landmarks (x, y, z and *visibility*) as input and returns the corrected joint's positions (x, y, z). As ground-truth, we use the TotalCapture data provided in the BVH files. However, as you can see in Figures 4 (MediaPipe skeleton structure) and 10 (TotalCapture's BVH skeleton structure), the TotalCapture's skeleton structure is different from MediaPipe's and for that reason, for the training of our network we used only the joints that are common in both skeletons. In total, we selected 18 joints. Table 1 shows the correspondence between the indices of the MediaPipe 3D joints and the TotalCapture's 3D joints provided in the BVH ground-truth files. The BVH joints are converted in hips-based format (see Section 3.4 for details). The network consists of 4 dense layers that make proper data transformations. In Figure 7 we can see the architecture of our proposed Landmarks' Correction Network, as well as the inaccurate 3D joint points estimated by MediaPipe that we feed the network with and the corrected output. The input layer of the network has 72 nodes to get 18 Mediapipe's landmarks with 4 values (x,

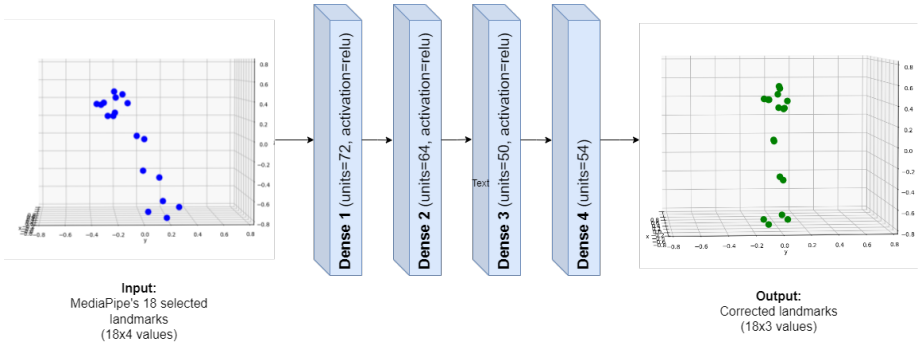

Fig. 7. Landmarks' correction network architecture.

Table 2. The axes used for the bones' vertical vectors

BVH Joints	8	9	14	15	19	20	23	24
Axis	x	x	-x	-x	z	z	z	z

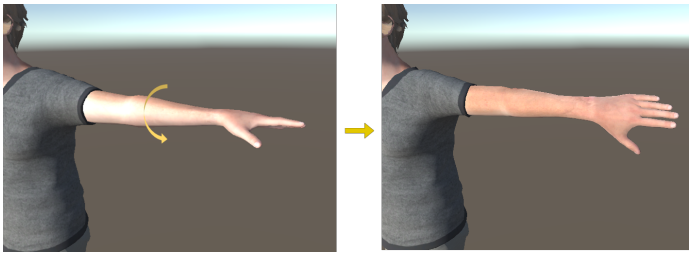

Fig. 8. Example of longitudinal rotation of upper limb in Unity environment.

y, *z* and *visibility*) each (18 x 4 = 72). However, the output size is 54 (18 x 3 = 54) as we only calculate the corrected landmarks' positions.

3.3.2 Bones' Rotation Information Enrichment Network. Given two joint 3D positions, we can calculate the intermediate bone's direction. The longitudinal rotation of the bone can be described either as an angle or a vector orthogonal to the bone's direction (see Figure 8). In our method, we use the vector representation as it was more convenient for the avatar mapping process. We calculated these vectors only for the bones of the upper and lower limbs. For the torso area, the longitudinal rotation of the spine can be calculated using the shoulder and hips joints without the need of extra information. As starting points of the vectors we used the starting points of the corresponding bones. The ending points of the vectors were chosen on an axis vertical to the bones' direction. However, each bone has two axes orthogonal to its direction (see Figure 10), so we choose one of them. In table 2 you can see the axis used for each BVH joint. In total, we calculated 8 ending points, one for each orthogonal vector. Then, we trained a fully connected neural network to estimate those points (the ending points of the bones' orthogonal vectors) using the 18 BVH joints. In Figure 9 we can see the architecture of our proposed Bones' Rotation Information Enrichment Network, which consists of 4 dense layers.

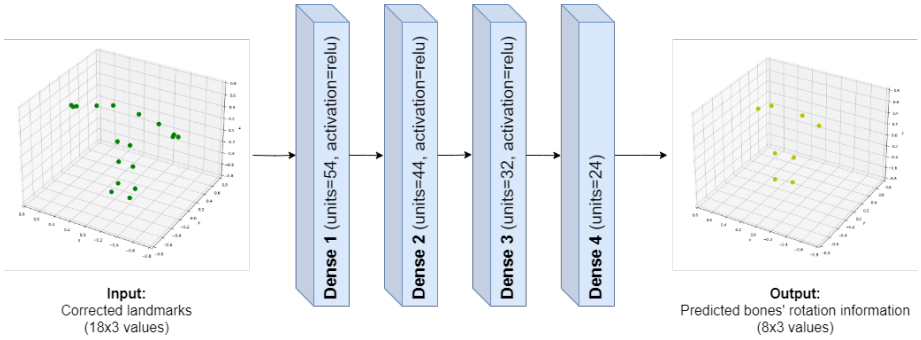

Fig. 9. Bones' rotation information enrichment network architecture.

3.4 Step 3: Transfer of Human Movement to Target 3D Avatar

The main idea on mapping the human pose from a video to an avatar is to determine the avatars' bones orientation based on the detected 3D landmarks. The reason we cannot use the detected landmarks' positions as starting and ending points for the avatar's bones is that the skeleton generated from Mediapipe is not consistent. That means a bone e.g. of the left arm may be larger than the corresponding bone in the right arm or a bone may have different size in consecutive frames of the video. Moreover, the skeleton structure of the avatar may be different from the Mediapipe skeleton (see Figures 4 and 11). Thus, the only way to transfer the movement is to calculate the directional vectors of the bones from the Mediapipe skeleton and apply them to the corresponding bones of the avatar.

Our implementation of this process was based on the DigiHuman project [6]. However, in DigiHuman the creators use the Mediapipe world landmarks without applying any correction method, rather than a scaling and they are not dealing with the longitudinal rotations. That is the reason why, in some cases, the avatars show deformities and the bones are moving unnaturally. For our experiments, we use avatars provided by Mixamo [7], which is an open 3D character animation and rigging platform. As shown in Figure 4, the Mediapipe skeleton has no landmarks in the spine area, whereas the avatar has (see Figure 11). To ease the mapping process we extend the Mediapipe landmarks by adding three more (hips, spine and neck). The extra landmarks are calculated using the shoulders and upper legs positions (Mediapipe landmarks 11, 12, 23 and 24) and they are added after the depth correction process.

Before the mapping process, the first step is to calculate the initial state of the avatar's bones i.e. to define their lengths and their initial orientations. This step is crucial because, as we proceed, the poses from the Mediapipe landmarks will be applied on top of that initial state. In the mapping process, we firstly calculate the torso orientation using the shoulders' landmarks, the upper legs and the 3 extra landmarks we added in the Mediapipe set (hips, spine, neck). Then, we calculate the orientation of each bone in the avatar using the corresponding set of landmarks in the Mediapipe skeleton. The avatar's joints follow a parental structure (e.g. the elbow is the parent joint of the wrist). To find the position of a child joint in the avatar we use the parent's joint position, the directional vector of the intermediate bone and the bone's length.

For the upper and lower limbs, we use the vertical points to calculate the longitudinal rotations. The Unity provides the LookRotation function, inside the Quaternion package, which takes two vectors, orthogonal to each other, as inputs and outputs the orientation they form as a quaternion. In our case, we use the bones' directional and orthogonal vectors

[6]https://github.com/Danial-Kord/DigiHuman
[7]https://www.mixamo.com/

to calculate the bones' orientations. The use of quaternios is preferable in order to eliminate the problem of gimbal lock that is appeared when using euler angles.

As we mentioned earlier, the Mediapipe landmarks are not always stable between consecutive frames. Due to light changes in the video and the fact that Mediapipe follows a per frame approach on pose estimation, there are small changes of the landmarks' positions that lead to poor results on the avatar movement. To reduce this jittering effect, before the mapping process, we pass the landmarks through a low pass filter. This process combines the landmarks' positions from the six last frames using a different weight for each frame in order to smooth the result and reduce unwanted changes.

4 EXPERIMENTS AND RESULTS

In order to train our deep neural network so that it corrects MediaPipe's depth estimation inaccuracies and enriches its output with bones' rotation information, we created an enhanced TotalCapture ground-truth, which contains the bones' rotation information. Below, we describe the process of this enhanced ground-truth, we provide details about the training of our proposed network and, finally, we present our experimental results.

4.1 Enhanced TotalCapture Dataset

The 3D joints provided in the TotalCapture dataset are in global coordinates, while the Mediapipe world landmarks are in hips-based format. Thus, we had to convert the TotalCapture joints into hips-based representation. As hip position, we used the middle point between the upper right and upper left leg joints (joints 19 and 23), instead of the original hip (joint 0) that is provided in the dataset, in order to follow Mediapipe's format. Another issue is that the BVH files contain animated skeletons where each joint's position is described with respect to its parent joint, based on the parental hierarchy of the skeleton. In order to manipulate the BVH data and extract the joints we used the Blender Python API (bpy) [8]. Using the same library we calculated the 3d points to form the orthogonal vectors of the upper and lower limbs (see Sec. 3.3.2 for details). Lastly, we had to process the TotalCapture's videos, with Mediapipe, to extract the world landmarks for each frame. We selected only the frames in which the person is fully visible and their body is entirely inside the frame in order for the Mediapipe's result to be more accurate. We use the MediaPipe model with complexity level of 1, which achieves good results in terms of both performance and accuracy. For our experiments we only use the first camera (cam1), however, we intend to add more cameras in future works. Subjects 1,2,3 and 4 of the TotalCapture dataset were used for training, while subject 5 was used only for the testing of our models.

4.2 Networks' Training Configuration

Some details on the configuration we used for the training of our two neural networks are presented below. Our code was implemented in Python and for the training of our deep neural networks we used Keras with Tensorflow as backend.

Landmarks' correction network: We employed the Adam optimizer and utilized the Mean Per Joint Position Error (MPJPE) as our loss function during training. The network underwent training for 17 epochs. Our dataset, comprising subjects 1 to 4 from the TotalCapture dataset, was divided into 90% for training and 10% for validation purposes.

Bones' orientation information enrichment network: For the training of this network, again, we utilized the Adam optimizer and adopted the Mean Per Joint Position Error (MPJPE) as our loss function. The training was conducted

[8] https://docs.blender.org/api/current/index.html

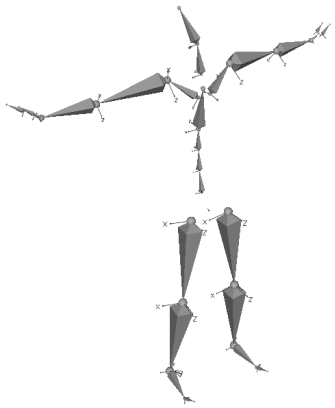

Fig. 10. BVH skeleton with joints' local axis.

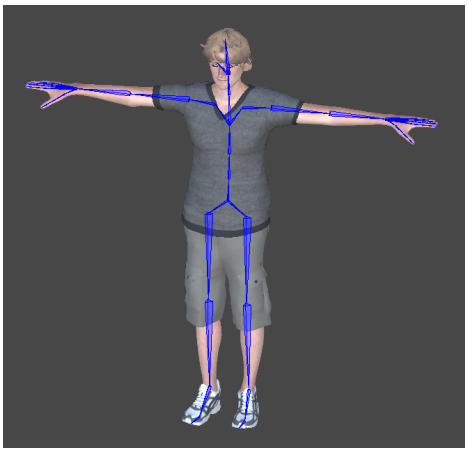

Fig. 11. Mixamo avatar's bone structure.

Table 3. Evaluation Results in MPJPE (cm)

Data	MPJPE (cm)
Original MediaPipe output	7.4
Corrected MediaPipe output	3.9

over 15 epochs. The dataset, which included subjects 1 to 4 from the TotalCapture dataset, was partitioned into 90% for training and 10% for validation.

4.3 Experimental Results

For the evaluation of our results, we used the mean per joint position error (MPJPE) metric in centimeters (cm). This metric calculates the average Euclidean distance between the estimated joint coordinates and the corresponding ground truth coordinates across all joints.

$$MPJPE(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} \|m_{\hat{x}}(i) - m_x(i)\| \tag{1}$$

where N represents the number of processed joints, $m_{\hat{x}}(i)$ is the estimated i-th joint coordinates and $m_x(i)$ is the corresponding ground truth position of the i-th joint.

For the evaluation of our results, we analyzed the camera 1 videos of subject 5 from the TotalCapture dataset, as the rest of the subjects (1-4) were used for the training of our networks. Initially, we computed the Mean Per Joint Position Error (MPJPE) between the original MediaPipe's landmark 3D coordinates and the TotalCapture ground-truth, resulting in an MPJPE of 7.4 cm. Subsequently, employing the corrected MediaPipe results obtained from the depth correction network, we recalculated the MPJPE against the TotalCapture ground-truth. Remarkably, our proposed network successfully reduced the MPJPE to 3.9 cm, underscoring its effectiveness in improving estimation accuracy. Table 3 briefly presents the evaluation results.

Besides the decrease of the Mean Per Joint Position Error (MPJPE), our goal is also to maintain the real-time performance of MediaPipe's algorithm throughout the process of mapping its enhanced output to the target 3D avatar. In order to increase performance and achieve a real-time inference, we optimized our two models using ONNX Runtime [9], which is a powerful machine-learning model accelerator. Our method achieves an inference time of approximately 28-30 frames per second (fps) on CPU.

5 CONCLUSIONS AND FUTURE WORK

In conclusion, in this study we are dealing with the process of real-time human movement transfer from monocular RGB input to a target 3D avatar, leveraging the advancements in 3D human pose estimation technology. Initially, the state-of-the-art MediaPipe pose estimation algorithm serves as a backbone to extract the human body landmarks' 3D coordinates. However, inherent inaccuracies in depth estimation and the absence of bone orientation information pose significant challenges, often resulting in suboptimal outcomes in the avatar mapping process. To handle these difficulties, we introduce an intermediary step between landmark extraction and motion transfer. The proposed intermediary step is designed to correct depth inaccuracies and enrich MediaPipe's output with longitudinal bones' rotation information. To achieve this, we employ a simple yet effective framework comprising two sequential deep neural networks: a depth corrector and a rotation information enricher. The depth corrector was trained on the videos provided by the TotalCapture dataset, using the corresponding BVH skeleton data as ground truth while the rotation information enricher was trained on our enhanced dataset which includes the bones' vertical vectors. The bones' vertical vectors are used in the avatar mapping process to calculate the longitudinal rotations. The Mediapipe's landmarks, firstly, pass through the depth corrector and then the corrected landmarks go to the information enricher network. Finally, both the corrected landmarks and the vertical points are used in the Unity environment to transfer the human movement from video to an avatar. Utilizing the ONNX library we optimize both our networks to reduce the inference time and as a result we achieve real-time performance of around 28-30 frames per second (fps) on CPU. Throughout this manuscript, we describe in detail each phase of our methodology, from the process of generating an enhanced TotalCapture ground-truth to the network architectures and training methodologies that we used. Additionally, we provide detailed insights into the methodology employed for mapping the estimated poses onto the target 3D avatar and provide quantitative and qualitative evaluation of our approach. Notably, our proposed depth corrector network yields remarkable results, substantially reducing the mean per joint position error from 7.4 cm to 3.9 cm when compared to MediaPipe's original output. These findings underscore the efficacy and promise of our approach in enhancing the accuracy and realism of real-time human motion transfer in immersive environments.

Concerning our future steps, we aim to further enhance the capabilities of our system by exploring different deep neural network architectures to address the specific challenges of our real-time motion transfer framework. For example, Long Short-Term Memory (LSTM) networks will be explored to incorporate temporal information into our system. By using temporal information, we aim to capture the dynamic nature of human motion more effectively, thereby achieving smoother and more natural motion transfer outcomes. Another idea is to use RGB input from two stereo cameras. In this way, we will have more information, as the subject will be simultaneously captured from two different viewing angles, and, consequently, this will lead to reduced inaccuracies in 3D pose estimation. Moreover, we will examine different training methodologies to optimize the performance of our deep neural networks like advanced optimization algorithms, regularization techniques, and data augmentation strategies to refine the learning process

[9] https://onnxruntime.ai/

and enhance the generalization capabilities of our models. Finally, as we recognize the important role of data in the development and evaluation of our system, we plan to expand our dataset selection, both for the training and testing phases. By incorporating diverse and representative datasets, we aim to ensure that our system is capable of accurately capturing and transferring a wide range of human motions across different scenarios and contexts.

ACKNOWLEDGMENTS

This work has received financial support by the Horizon Europe Research & Innovation Programme under Grant agreement N. 101092612 (Social and hUman ceNtered XR - SUN project).

REFERENCES

[1] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Lixuan Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. *ArXiv* abs/2006.10204 (2020), 11. https://api.semanticscholar.org/CorpusID:219793039

[2] Polona Caserman, Augusto Garcia-Agundez, and Stefan Göbel. 2019. A survey of full-body motion reconstruction in immersive virtual reality applications. *IEEE transactions on visualization and computer graphics* 26, 10 (2019), 3089–3108.

[3] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. 2020. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 769–787. https://doi.org/10.1007/978-3-030-58571-6_45

[4] Sebastian Dill, Maurice Rohr, Gökhan Güney, Christoph Hoog Antink, Andreas Rösch, Luisa De Witte, and Elias Schwartz. 2023. Accuracy Evaluation of 3D Pose Estimation with MediaPipe Pose for Physical Exercises. *Current Directions in Biomedical Engineering* 9 (2023), 563 – 566. https://api.semanticscholar.org/CorpusID:262087385

[5] Francisco Fernandes, Ivo Roupa, Sérgio B Gonçalves, Gonçalo Moita, Miguel Tavares da Silva, João Pereira, Joaquim Jorge, Richard R Neptune, and Daniel Simões Lopes. 2023. Sticks and STONES may build my bones: Deep learning reconstruction of limb rotations in stick figures. *Pattern Recognition Letters* 165 (2023), 138–145.

[6] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, Seattle, WA, USA, 5252–5262. https://doi.org/10.1109/CVPR42600.2020.00530

[7] Yiqiao Lin, Xueyan Jiao, and Lei Zhao. 2023. Detection of 3d human posture based on improved mediapipe. *Journal of Computer and Communications* 11, 2 (2023), 102–121.

[8] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[9] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black. 2019. AMASS: Archive of Motion Capture As Surface Shapes. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, Los Alamitos, CA, USA, 5441–5450. https://doi.org/10.1109/ICCV.2019.00554

[10] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. 2020. XNect: Real-time multi-person 3D motion capture with a single RGB camera. *Acm Transactions On Graphics (TOG)* 39, 4 (2020), 82–1.

[11] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *Acm transactions on graphics (tog)* 36, 4 (2017), 1–14.

[12] Ammar Qammaz and Antonis A Argyros. 2021. Towards Holistic Real-time Human 3D Pose Estimation using MocapNETs. In *British Machine Vision Conference (BMVC 2021)*. BMVA, online, 418.

[13] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. 2021. HuMoR: 3D Human Motion Model for Robust Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, virtual event, 11488–11499.

[14] Ivo Roupa, Sérgio Barroso Gonçalves, Miguel Tavares da Silva, Richard R. Neptune, and Daniel Simões Lopes. 2021. Motion envelopes: unfolding longitudinal rotation data from walking stick-figures. *Computer Methods in Biomechanics and Biomedical Engineering* 25 (2021), 1459 – 1470. https://api.semanticscholar.org/CorpusID:245262754

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Salt Lake City, UT, USA, 4510–4520.

[16] Ming-Hwa Sheu, S. M. Salahuddin Morsalin, Chung-Chian Hsu, Shin-Chi Lai, Szu-Hong Wang, and Chuan-Yu Chang. 2023. Improvement of Human Pose Estimation and Processing With the Intensive Feature Consistency Network. *IEEE Access* 11 (2023), 28045–28059. https://doi.org/10.1109/ACCESS.2023.3258417

[17] Mingyi Shi, Kfir Aberman, Andreas Aristidou, Taku Komura, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. *ACM Transactions on Graphics (TOG)* 40, 1 (2020), 1–15.

[18] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. 2020. Physcap: Physically plausible monocular 3d motion capture in real time. *ACM Transactions on Graphics (ToG)* 39, 6 (2020), 1–16.

[19] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. CVPR, Long Beach, CA, USA, 5693–5703.

[20] Matthew Trumble, Andrew Gilbert, Charles Malleson, Adrian Hilton, and John P. Collomosse. 2017. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In *British Machine Vision Conference*. BMVA, London, UK, 1–13. https://api.semanticscholar.org/CorpusID:52271809

[21] Yen-Lin Wu. 2021. *One Pose Fits All: A novel kinematic approach to 3D human pose estimation.* Master's thesis. Delft University of Technology.

[22] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. 2020. GHUM & GHUML: Generative 3D Human Shape and Articulated Pose Models. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 6183–6192. https://doi.org/10.1109/CVPR42600.2020.00622

[23] Y. Zhan, F. Li, R. Weng, and W. Choi. 2022. Ray3D: ray-based 3D human pose estimation for monocular absolute 3D localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 13106–13115. https://doi.org/10.1109/CVPR52688.2022.01277

[24] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. 2020. SMAP: Single-Shot Multi-Person Absolute 3D Pose Estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XV 16*. Springer, ECCV, Glasgow, UK, 550–566.