3D Human Body Model Reconstruction Algorithm Based on Multi-View Synchronized Video Sequences

Yidong Wu[a]
Department of Computing, Harbin Institute of Technology, Harbin, China

Keywords: Computer Vision, Deep Learning, Multi-View Driven, 3D Human Body Model Reconstruction.

Abstract: As an important topic in the field of computer vision, 3D human reconstruction has a wide range of applications in the fields of film and television entertainment, sports and medicine. Traditional 3D human reconstruction methods often require professional equipment and clothing for technical support, and the process is very cumbersome and has great limitations. In recent years, with the development of deep learning, the method of human reconstruction using deep learning has achieved great success. Based on this background, this paper introduces a 3D human reconstruction algorithm based on multi-view synchronized video sequences, which can improve the shortcomings of traditional methods. Specifically, this paper reprojects the key points on the 3D human model back to the 2D plane under multiple perspectives, and uses the key points obtained by 2D human posture detection to optimize the reprojected key points, and finally obtains the body shape and posture parameters of the 3D human model. After comparative experiments, the method of this paper has achieved good accuracy and efficiency.

1 INTRODUCTION

At present, with the rapid development of augmented reality (AR) and virtual reality (VR) technologies, the metaverse (Wang, et al., 2023) has attracted more and more attention. As a digital space based on the virtual world, the metaverse can effectively simulate the physical laws and human activities in the real world. In this artificially created virtual world, users can interact immersively through their own avatars, such as engaging in social activities, work production, etc. At present, one of the key research directions of the metaverse is how to obtain the user's human posture in real time and reconstruct their avatars in the metaverse at the same time. This technology is also called three-dimensional human body reconstruction technology. As a key research topic in computer vision and computer graphics, three-dimensional human body reconstruction technology has been widely used in game modeling, medical imaging, film and television motion capture, identity recognition and other fields.

Traditional three-dimensional human body reconstruction technologies include motion capture methods based on optical marker capture (Siaw, Han, and Wong, 2023) and inertial capture (He, Zheng, Zhu, et al., 2022). The method based on optical marker capture is to capture the marker points attached to different positions of the human body through a multi-view camera, so as to obtain the positions of each joint of the human body and perform human body modeling. The method based on inertial capture is to equip accelerometers, gyroscopes and other measuring instruments at various positions of the human body to capture the speed, acceleration, etc. of different parts of the human body, and finally calculate the human body model. Although these methods are feasible, they require professional equipment, and the process is cumbersome and has great limitations.

The main research purpose of this paper is to propose a multi-perspective jointly driven 3D human body reconstruction technology. This multi-perspective jointly driven 3D human body reconstruction technology no longer relies on professional wearable equipment, can effectively solve the problems existing in traditional methods, and provides a simpler and more efficient 3D human body reconstruction method, which has strong practical significance.

[a] https://orcid.org/0009-0008-6060-2165

Wu and Y.
3D Human Body Model Reconstruction Algorithm Based on Multi-View Synchronized Video Sequences.
DOI: 10.5220/0013512800004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 214-220
ISBN: 978-989-758-754-2
Copyright © 2025 by Paper published under CC license (CC BY-NC-ND 4.0)

2 LITERATURE REVIEW

Two-dimensional human pose detection can be divided into two categories: top-down and bottom-up. Top-down methods, such as the AlphaPose algorithm (Fang, Li, Tang, et al., 2022), first use a human detector (REN, HE, Girshick, et al., 2015)) to detect all the human bodies in the image, and then use a neural network to estimate the human pose for each individual human body. Bottom-up methods, such as the OpenPose algorithm (Wu, Tang, Xiong, et al., 2022), have a network structure divided into many layers. The first layer is used to predict the joint point heat map and the limb association confidence map, and each subsequent layer will gradually optimize the connection between the joints and limbs until all human skeletons are assembled. In addition, many detection methods have improved OpenPose, such as the OpenPifPaf (Kreiss, Bertoni, and Alahi, 2022) multi-person pose estimation method.

In order to simplify the human body model, people have proposed the concept of parametric human body model, that is, using a template and different parameters to generate a variety of human body models, such as the SMPL (Song, Yoon, Cho, et al., 2023) model proposed by Loper in 2015, which is a model based on linear mixed skinning drive. Later, the face parameterized model FLAME (Athar, Shu, and Samaras, 2023) and the hand parameterized model MANO (Potamias, Ploumpis, Moschoglou, et al., 2023) were also proposed one after another. In 2019, Pavlakos et al. improved on SMPL, combined with FLAME and MANO models, and constructed a full-body human model SMPL-X (Pavlakos, et al., 2019). Compared with SMPL, which has simpler parameters, the SMPL-X model can not only adjust the height, weight and posture of the human body, but also make separate adjustments for facial expressions and hand movements.

For single-person human model reconstruction, SMPLify (Hassan, Choutas, Tzionas, et al., 2019) designed a reconstruction method based on joint detection and posture prior for the SMPL model. After the emergence of the SMPL-X model, SMPLify-X (Pavlakos, et al., 2019) also came into being. Compared with SMPLify, it relearned the action prior using variational autoencoders (VAE) and designed new model penetration penalties, thereby achieving better reconstruction results. In 2018, Kanazawa et al. proposed the HMR (Human Mesh Recovery) (Moon, Choi, and Lee, 2022) method, which realized an end-to-end deep learning network, directly regressed a three-dimensional human model from a single image, and the training process did not require supervision of three-dimensional key points of the human body.

Many current studies often sacrifice a certain degree of accuracy in order to improve reconstruction speed, ignoring the error accumulation caused by the two processes of two-dimensional posture detection and three-dimensional reconstruction, resulting in jitter in the reconstructed human body model. The main research content of this paper is to use the key point coordinates obtained by two-dimensional posture detection as the main reference based on multi-view video information, and use the information obtained by triangulation of key points to initialize the human body model, reconstruct the three-dimensional human body model by optimizing the parameters of the human body model, and optimize the jitter problem of the human body model, and finally maximize the efficiency of human body reconstruction without sacrificing accuracy.

3 METHOD

3.1 Camera Model

The camera model used in this paper is the pinhole camera model, which contains four coordinate systems: world coordinate system, camera coordinate system, image coordinate system, and pixel coordinate system. In this model, a point in the real world is transformed into a point in the image through the conversion between these four coordinate systems. Next, this paper will discuss the method of projecting point P from its coordinates (X, Y, Z) in the world coordinate system to its coordinates (u, v) in the pixel coordinate system. The pinhole camera model is shown in Figure 1.

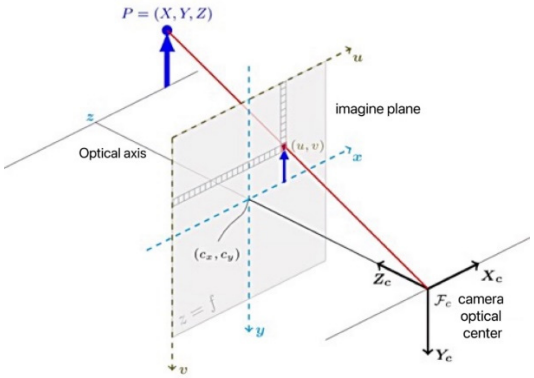

Figure 1: Pinhole camera model. (Photo/Picture credit : Original)

Given the coordinates (X, Y, Z) of point P in the world coordinate system, the formula for projecting this coordinate to the coordinates (u, v) in its pixel coordinate system is:

$$Zc\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}\frac{1}{dx}&0&u_0\\0&\frac{1}{dy}&v_0\\0&0&1\end{bmatrix}\begin{bmatrix}f&0&0&0\\0&f&0&0\\0&0&1&0\end{bmatrix}\begin{bmatrix}R&t\\0_{1\times3}&1\end{bmatrix}\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix}$$
$$=\begin{bmatrix}f_x&0&u_0&0\\0&f_y&v_0&0\\0&0&1&0\end{bmatrix}\begin{bmatrix}R&t\\0_{1\times3}&1\end{bmatrix}\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix} \quad (1)$$

Among them, $\begin{bmatrix}f_x&0&u_0&0\\0&f_y&v_0&0\\0&0&1&0\end{bmatrix}$ is called the intrinsic parameter matrix, and $\begin{bmatrix}R&t\\0_{1\times3}&1\end{bmatrix}$ is called the extrinsic parameter matrix. The intrinsic parameter matrix is the parameter that describes the internal properties of the camera, which includes information such as focal length, principal point coordinates, and distortion coefficients. For a specific camera, these intrinsic parameters are usually fixed and do not change with time or space. The extrinsic parameter matrix is used to describe the pose parameters of the camera in the world coordinate system, which changes with the position and pose of the camera in space.

3.2 Parametric Human Body Model

In the SMPL model, the mesh model of a 3D human body model consists of $N = 6890$ vertices, including $K = 23$ human joints, and the driving relationship of each joint to each mesh vertex is described by the Blend Weight matrix $\omega(6890 \times 24)$. The author uses PCA (Principal Component Analysis) to reduce the dimension of a large number of parameters and extracts some parameters that have the greatest impact on the body shape and movement of the human body model, ultimately achieving the purpose of simplifying the number of parameters. For body shape, the parameter $\beta \in \mathbb{R}^{10}$ is defined to describe height, weight, and thinness. For posture, the parameter $\theta \in \mathbb{R}^{3(K+1)}$ is defined, and the axis angle is used to represent the rotation of the joint relative to the initial posture. The first three parameters control the rotation posture of the root orientation, and the remaining three parameters each describe the rotation axis and angle of a joint. After obtaining a set of determined body shape parameters β and posture parameters θ, the human body model can be driven using linear blend skinning (LBS) (Khamis, Taylor, Shotton, et al., 2015). A schematic diagram of the SMPL human body model is shown in Figure 2.

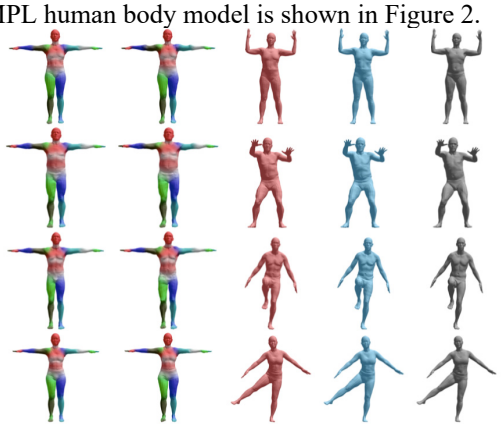

Figure 2: SMPL human body model (Loper, 2024).

The model used in this paper is the human body model SMPL-X (SMPL eXpressive) (Hassan, Choutas, Tzionas, et al., 2019), which was improved by Pavlakos et al. in 2019 based on the SMPL model. It combines the original SMPL model of the body, the MANO (hand Model with Articulated and Non-rigid deformations) (Potamias, Ploumpis, Moschoglou, 2023) model of the hand, and the FLAME (Faces Learned with an Articulated Model and Expressions) (Athar, Shu, and Samaras, 2023) model of the head to describe the human body with a unified model. A schematic diagram of the SMPL-X human body model is shown in Figure 3.

Figure 3: SMPL-X human body model (Pavlakos, et al., 2019).

The SMPL-X model uses a standard LBS-driven human body model. The mesh model of a 3D human body model consists of $N = 10475$ vertices and contains $K = 54$ key points (including joints, chin, fingers, eyes, etc.). The SMPL-X model can be expressed as the following function:

$$M(\theta,\beta,\psi):\mathbb{R}^{|\theta|\times|\beta|\times|\psi|}\rightarrow\mathbb{R}^{3N} \quad (2)$$

Among them, $\theta \in \mathbb{R}^{3(K+1)}$ is the posture parameter, which includes three types of parameters:

chin parameters, finger parameters, and parameters of other body joints. And $\beta \in \mathbb{R}^{10}$ is the body shape parameter as in the SMPL model. The new $\psi \in \mathbb{R}^{10}$ is the facial expression parameter, which is used to represent the human expression after dimensionality reduction. After PCA dimensionality reduction, the author reduced the total number of model parameters of SMPL-X to 119, of which 10 are body shape parameters, 10 are facial expression parameters, 24 are hand parameters used to represent finger movements, and the remaining 75 are body posture parameters used to represent the rotation of the remaining key points (body, chin, eyes, etc.).

3.3 2D Human Posture Detection

Before reconstructing and optimizing the three-dimensional model, it is first necessary to perform two-dimensional human posture detection on each frame image from different perspectives in the video and extract the two-dimensional human posture information at each perspective. The two-dimensional detection method used in this paper is the MediaPipe Holistic method (Huu, Hong, Dang, et al., 2023) released by Google engineer Grishchenko in 2020. A new Pipeline is proposed in MediaPipe Holistic, which successfully optimizes the detection components of the face, hands, and body, and completes the semantic level combination between the three components, so that it can complete the synchronous detection of human face, hands, and posture in real time. The detection results output by MediaPipe Holistic includes 33 human posture key points, 21 hand movement key points for each hand, and 468 facial expression key points, which can provide comprehensive and sufficient data support for the successful execution of subsequent research and analysis work. Figure 4 shows an example of MediaPipe Holistic human pose detection.

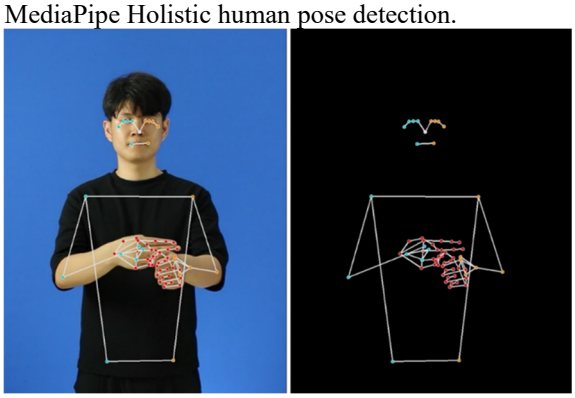

Figure 4: Example of MediaPipe Holistic human pose detection (Kim, Baek, 2023).

After using MediaPipe Holistic to extract the 2D human posture key points of each frame at each viewpoint, in order to use it to reconstruct the SMPL-X model, it is also necessary to find the correspondence between the SMPL-X key points and the key points output by MediaPipe Holistic to complete the mapping between the two models. Figure 5 shows key point annotation of body and hands in MediaPipe Holistic.

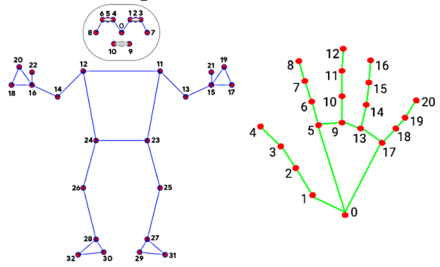

Figure 5: Key point annotation of body and hands in MediaPipe Holistic (Kim, Baek, 2023).

3.4 Residual Function

3.4.1 Residual Term For Body

At a specific viewing angle, the residual of the body part is calculated by the previously obtained reprojected key point 2D coordinates P_s and the key point 2D coordinates P_t obtained by detection. The L2 norm is calculated for $(P_s - P_t)$, and different weights ω_{body} are designed for different key points to optimize the reconstruction effect.

$$dis_{body} = \omega_{\text{body}} \times \text{L2.norm}(P_s - P_t) \quad (3)$$

Because at a specific viewing angle, the confidence C_{p_i} (the probability that the error between the calibrated key point and its corresponding real key point on the human body is small) of each key point p_i obtained by 2D human posture detection is not the same: some key points have a higher confidence, and these key points can be used for fitting optimization; while some key points have a lower confidence, that is, these key points are likely to be inaccurate and have a lower reference value, so it is not accurate to use dis_{body} as the value of the residual function. In a gesture to avoid this problem, this paper only selects key points with $C_p > 0.5$ (C_p is the confidence) for residual function calculation. In summary, the residual function is as follows (where M is the number of viewing angles):

$$\mathcal{L}_{\text{body}} = \frac{1}{\text{M}}\sum_{\text{j=1}}^{\text{M}} dis^{j}_{body_{for\ all\ p_i\ with\ C_{p_i}>0.5}} \quad (4)$$

3.4.2 Residual Term For Hand

By taking the L2 norm between the reprojected key points and the key points obtained by detection, the residual term of the hand can also be obtained. However, unlike the calculation of the residual function for the body, the calculation of the residual function for the hand does not need to consider the confidence $\text{C}_{\text{p}_\text{i}}$. This is because MediaPipe Holistic detection will directly discard the key points of the hand image with insufficient confidence. Therefore, all hand key points are high confidence. In summary, the residual function of all views of the two hands is averaged to obtain the residual function:

$$\mathcal{L}_{\text{hand}} = \frac{1}{\text{2M}}\sum_{\text{j=1}}^{\text{2M}} dis^{j}_{hand} \quad (5)$$

3.4.3 Residual Term For Face

This paper uses 478 specific vertices of the human body mesh model for fitting optimization. Since all key points obtained by face detection are also high-confidence key points, there is no need to consider the confidence $\text{C}_{\text{p}_\text{i}}$. The residual function of all views is averaged to obtain the facial residual term:

$$\mathcal{L}_{\text{face}} = \frac{1}{\text{M}}\sum_{\text{j=1}}^{\text{M}} dis^{j}_{face} \quad (6)$$

3.4.4 Regularization Term

Since direct training may lead to the problem of excessively large parameters, making the model susceptible to noise and ultimately leading to inaccurate fitting results, this paper uses a combination of L1 regularization and L2 regularization to constrain and penalize the parameter size.

Combining the above three residual terms, the final residual function expression is as follows:

$$\mathcal{L}_{total} = \omega_{body_{2d}} \times \mathcal{L}_{body} + \omega_{hand_{2d}} \times \mathcal{L}_{hand} + \omega_{face\_2d} \times \mathcal{L}_{body} + \omega_{reg} \times \mathcal{L}_{reg} \quad (7)$$

4 EXPERIMENT

4.1 Dataet

The CMU Panoptic Dataset is a series of datasets released between August 2016 and April 2019. The dataset contains 65 video sequences (5.5 hours) and 1.5 million 3D skeletons (Joo, Liu, Tan, et al., 2015). This paper selects the video sequences of the single-person part as the training dataset.

The parameters for each set of video sequences are as follows:

(1) 480 VGA cameras with a resolution of 640 × 480, capturing at 25 fps, all synchronized using a hardware clock (Joo, Liu, Tan, et al., 2015).

(2) 31 HD cameras with a resolution of 1920 × 1080, capturing at 30 fps, also synchronized using a hardware clock and time-aligned with the VGA cameras (Joo, Liu, Tan, et al., 2015).

(3) 10 Kinect II sensors, providing 1920 × 1080 (RGB) and 512 × 424 (depth) resolutions, capturing at 30 fps, synchronized both among themselves and with the other sensors (Joo, Liu, Tan, et al., 2015).

(4) 5 DLP projectors, synchronized with the HD cameras (Joo, Liu, Tan, et al., 2015).

The experiment uses the videos of 8 viewpoints in the above dataset as the final training dataset input. To ensure the consistency of the training data, all videos of 8 viewpoints are synchronized. To prevent overfitting, the upper limit of the training rounds of each frame model is set to 15 rounds.

4.2 Evaluation Indicators

This paper selects the values of mean joint position error (MPJPE, Mean Per Joint Position Error) and reconstruction error (PA-MPJPE, Procrustes Aligned MPJPE) as the evaluation indicators of the experiment. The accuracy of the algorithm is judged by calculating the mean joint position error (MPJPE, Mean Per Joint Position Error) and reconstruction error (PA-MPJPE, Procrustes Aligned MPJPE) between the key point coordinates of the fitted 3D human model and the real coordinates of the key points given in the data set, and comparing them with the full-body motion capture method published by Zhang (Zhang, Li, An, et al., 2021).

On this basis, this experiment also statistically analyzes the 3D human reconstruction speed of the algorithm and evaluates the efficiency of the algorithm based on this indicator.

4.3 Result

According to the method in the above experiment, the 3D human body reconstruction was performed on several single-person action video sequences in the CMU panoramic data set. The final 3D human body reconstruction model has a very small visual jitter that is almost imperceptible to the naked eye. After the model training is completed, it is tested using a test set, which consists of multi-view synchronized video sequences collected under similar conditions. The 3D human body reconstruction effect of the test set is shown in Figure 6 below.

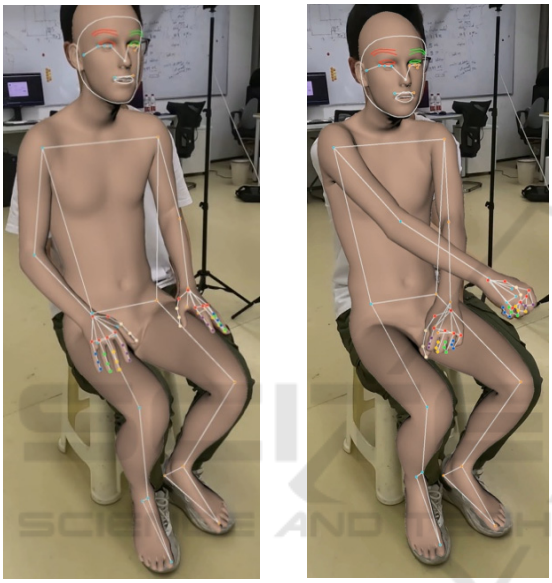

Figure 6: 3D reconstruction effect of the test set. (Photo/Picture credit: Original)

From a data perspective, the MPJPE, PA-MPJPE, and reconstruction speed collected by the algorithm described in this paper are compared with the corresponding indicators in the full-body motion capture method published by Zhang (Zhang, Li, An, et al., 2021). The final experimental results are shown in Table 1:

Table 1 experimental results.

method	MPJPE（mm）	PA-MPJPE（mm）	Recon-struction speed (minutes/f-rame)
This paper	24.10	19.83	0.59
Zhang	24.38	20.06	0.95

Comparing the experimental results, the MPJPE of this paper's method is reduced by about 1.15% and the PA-MPJPE is reduced by about 1.12% compared with Zhang's full-body motion capture method. Therefore, the human body model reconstructed by this paper's method is slightly more accurate than Zhang's method. In addition, this paper's method has a huge improvement in the reconstruction speed of the model, which is increased by about 60%.

In summary, the 3D human body reconstruction algorithm proposed in this paper has achieved improvements in both accuracy and efficiency.

5 CONCLUSION

This paper studies the pinhole camera model, 2D human pose detection, uses the MediaPipe Holistic method, adopts the SMPL-X model, and finally obtains an algorithm for reconstructing a 3D human model from synchronized 2D human pose video sequences from various angles through deep learning. In addition, this paper adds restrictions to the parameters of the human model, reduces error accumulation by adjusting the relationship between the residual function and the parameter optimization, and effectively suppresses the instability of the model. The experimental results verify that these technical improvements have greatly improved the stability and reconstruction effect of the model, and have achieved good results in both accuracy and efficiency.

From a visual perspective, although the jitter observable to the naked eye is extremely subtle and almost imperceptible, these jitters may still need to be further optimized in extreme application scenarios. However, under the existing experimental conditions and test sets, the method in this paper has been able to achieve high visual stability and reconstruction accuracy.

In future research, it is possible to consider further reducing possible jitters by improving the details of the model and improving the adaptability of the model in more complex scenarios. In addition, the versatility and extensibility of the algorithm can also be further verified by introducing more different types of test sets to evaluate its performance in different application scenarios.

REFERENCES

Athar, S., Shu, Z. and Samaras, D., 2023. FLAME-in-NeRF: Neural control of Radiance Fields for Free View Face Animation. 2023 IEEE 17th International Conference on Automatic Face and Gesture

Recognition (FG), Waikoloa Beach, HI, USA, pp. 1-8, doi: 10.1109/FG57933.2023.10042553.

Fang, H. S., Li, J., Tang, H., et al., 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence: 7157-7173.

Hassan, M., Choutas, V., Tzionas, D. and Black, M., 2019. Resolving 3D Human Pose Ambiguities With 3D Scene Constraints. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 2282-2292, doi: 10.1109/ICCV.2019. 00237.

He, Q., Zheng, Z., Zhu, X., Zhang, H., Su, Y. and Xu, X., 2022. Design and Implementation of Low-Cost Inertial Sensor-Based Human Motion Capture System. 2022 International Conference on Cyber-Physical Social Intelligence (ICCSI), Nanjing, China, pp. 664-669, doi: 10.1109/ICCSI55536.2022.9970563.

Huu, P. N., Hong, P. D. L., Dang, D. D., Quoc, B. V., Bao, C. N. L. and Minh, Q. T., 2023. Proposing Hand Gesture Recognition System Using MediaPipe Holistic and LSTM. 2023 International Conference on Advanced Technologies for Communications (ATC), Da Nang, Vietnam, pp. 433-438, doi: 10.1109/ATC58710.2023.10318885.

Joo, H., Liu, H., Tan, L., et al., 2015. Panoptic studio: A massively multiview system for social motion capture. Proceedings of the IEEE International Conference on Computer Vision. 3334-3342.

Khamis, S., Taylor, J., Shotton, J., Keskin, C., Izadi, S. and Fitzgibbon, A., 2015. Learning an efficient model of hand shape variation from depth images. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 2540-2548, doi: 10.1109/CVPR.2015.7298869.

Kim, H., Baek, S. W., 2023. Implementation of wearable glove for sign language expression based on deep learning. Microsystem Technologies. 29. 1-17. 10.1007/s00542-023-05454-5.

Kreiss, S., Bertoni, L. and Alahi, A., 2022. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 13498-13511, doi: 10.1109/TITS.2021.3124981.

Loper, M., 2024. SMPL: A Skinned Multi-Person Linear Model. SMPL Project Homepage. Available: https://smpl.is.tue.mpg.de/. Accessed September 16, 2024.

Moon, G., Choi, H. and Lee, K. M., 2022. NeuralAnnot: Neural Annotator for 3D Human Mesh Training Sets. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, pp. 2298-2306, doi: 10.1109/CVPRW56347.2022.00256.

Pavlakos, G., et al., 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 10967-10977, doi: 10.1109/CVPR.2019.01123.

Potamias, R. A., Ploumpis, S., Moschoglou, S., Triantafyllou, V. and Zafeiriou, S., 2023. Handy: Towards a High Fidelity 3D Hand Shape and Appearance Model. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, pp. 4670-4680, doi: 10.1109/CVPR52729.2023.00453.

Ren, S., He, K., Girshick, R. B., et al., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. Computing Research Repository, abs/1506.01497. http://arxiv.org/abs/1506.01497

Siaw, T. U., Han, Y. C. and Wong, K. I., 2023. A Low-Cost Marker-Based Optical Motion Capture System to Validate Inertial Measurement Units. in IEEE Sensors Letters, vol. 7, no. 2, pp. 1-4, Art no. 5500604, doi: 10.1109/LSENS.2023.3239360.

Song, H., Yoon, B., Cho, W. and Woo, W., 2023. RC-SMPL: Real-time Cumulative SMPL-based Avatar Body Generation. 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Sydney, Australia, pp. 89-98, doi: 10.1109/ISMAR59233.2023.00023.

Wang, Y. et al., 2023. A Survey on Metaverse: Fundamentals, Security, and Privacy. in IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 319-352, doi: 10.1109/COMST.2022.3202047.

Wu, E. Q., Tang, Z. -R., Xiong, P., Wei, C. -F., Song, A. and Zhu, L. -M., 2022. ROpenPose: A Rapider OpenPose Model for Astronaut Operation Attitude Detection. in IEEE Transactions on Industrial Electronics, vol. 69, no. 1, pp. 1043-1052, doi: 10.1109/TIE.2020.3048285.

Zhang, Y., Li, Z., An, L., et al., 2021. Lightweight multi-person total motion capture using sparse multi-view cameras. Proceedings of the IEEE/CVF International Conference on Computer Vision. 5560-5569.