https://doi.org/10.2352/EI.2025.37.14.COIMG-126

Wide-Baseline Multi-Camera Automatic Calibration Using Recovered Human Body Mesh

Chih-Hsien Chou, Lin-Hsi Tsao; Futurewei Technologies, Inc., San Jose, California, USA

Abstract

Human pose and shape estimation (HPSE) is a crucial function for human-centric applications, while the accuracy of deep learning-based monocular 3D HPSE may suffer due to depth ambiguity. Multi-camera systems with wide baselines can mitigate the problem but accurate and robust multi-camera calibration is a prerequisite. The main objective of the paper is to develop fast and accurate algorithms for automatic calibration of multi-camera systems which fully utilize human semantic information without using predetermined calibration patterns or objects. The proposed automatic calibration method for multi-camera systems takes from each camera the 3D human body meshes output from pretrained Human Mesh Recovery (HMR) model, and the vertices of each 3D human body mesh are projected onto the 2D image plane for each corresponding camera. Structure-from-Motion (SfM) algorithm is used to reconstruct 3D shapes from a pair of cameras, using iterative Random Sample Consensus (RANSAC) algorithm to remove outliers when calculating the essential matrix in each iteration. Relative camera extrinsic parameters (i.e., the rotation matrix and translation vector) can be calculated from the estimated essential matrix accordingly. By assuming one main camera's pose in the world coordinate is known, the poses of all other cameras in the multi-camera system can be readily calculated. Using (1) average 2D projection error and (2) average rotation and translation errors as performance metrics, the proposed method is shown to perform calibration more accurate than methods using appearance-based feature extractors, e.g., Scale-Invariant Feature Transform (SIFT), and deep learning-based 2D human joint estimators, e.g., OpenPose.

1. Introduction

Human pose and shape estimation (HPSE) is a crucial function for many human-centric applications in various fields, such as immersive telepresence, interactive conferencing, sports analytics, healthcare monitoring, human motion tracking, avatar and digital human creation, metaverse, AR/VR/MR/XR and entertainment. However, deep learning-based monocular 3D HPSE may fail for rare or unseen poses due to limited and fixed training data. It is challenging due to the fact that 2D-image-to-3D-posture mapping by a monocular 3D human pose estimator is not unique but subject to depth ambiguity. Furthermore, broad diversity in human poses, appearances, and camera viewpoints only make the problem even more difficult and error prone. Among existing methods for solving the problem, multi-camera systems with wide baselines can provide more reliable estimates from less reliable monocular estimates from each individual camera without redefining a new multi-view 3D HPSE or retraining the existing monocular 3D HPSE. However, accurate and robust multi-camera calibration is required for multi-camera systems with overlapping field of view (FoV) to mitigate the self or mutual occlusion and depth ambiguity problems.

SMPL (Skinned Multi-Person Linear Model) [1] and its extended version SMPL-X (Expressive Body Capture) [2] and upgraded version STAR (Sparse Trained Articulated Human Body Regressor) [3] are state-of-the-art 3D human body models based on skinning and blend shapes. They are becoming popular in both industry and academia for human body synthesis by NeRF or 3D Gaussian splatting. HMR (Human Mesh Recovery) [4] and its upgraded version HMR 2.0 (Humans in 4D) [5] are state-of-the-art end-to-end methods for reconstructing a full 3D mesh of a human body, even occluded or truncated, from a single RGB image by estimating its corresponding SMPL model parameters. 3D human meshes are usually estimated within a bounding box containing a detected person in 2D camera coordinates, while existing methods estimating 3D humans in 3D world coordinates run slow and require MoCap markers or IMU sensors.

Bottom-up human pose estimation methods can directly estimate the joints of all people in an input image without running multiple times in multi-person scenarios as the top-down methods, although their accuracy is usually worse than their top-down counterparts. OpenPose [6] is a state-of-the-art bottom-up method for 2D human skeleton detection that can quickly and accurately identify multiple human skeletons and locate associated 2D joints in a single input image, where body parts belonging to the same person are linked, including foot key points. This is achieved by the part affinity fields (PAFs) where a 2D vector in each pixel of every PAF encodes the position and orientation of the limbs.

Deep learning-based monocular 3D human pose estimation may fail for rare or unseen poses due to limited and fixed training data [7]. It is challenging due to depth ambiguity and broad diversity in human poses, appearances, and camera viewpoints. Training 3D pose estimation is severely limited by dataset bias, because collecting accurate 3D pose annotations for 2D images as ground truth for model training is costly and time-consuming and collected training data is usually biased towards specific environment and selected actions. The 2D-image-to-3D-posture mapping by a monocular 3D human pose estimator is not unique subject to depth ambiguity, which may result in, for example, different degrees of body tilt even for common human postures regardless of the camera's shooting angle [8]. In worst-case scenarios, incorrect body tilt depends on hand / body stretches for lack of diversified human poses in training data. State-of-the-art methods to solve the depth ambiguity problem include: (1) synthetic data generation utilizing scarce training data, using only single view; (2) multi-view consistency as supervisory signal when training data is scarce, requiring at least two views; and (3) multi-camera systems with wide baselines to provide more reliable estimates from less reliable estimates by each individual camera.

Method (3) described above is desirable because it can provide more reliable multi-view estimates from less reliable monocular estimates without redefining a new multi-view 3D HPSE or retraining the existing monocular 3D HPSE, but accurate and robust multi-camera calibration is required. Procrustes transformation (i.e., rigid-body transformation with degrees of freedom in scale and rotation) is usually applied to ignore local rotation and scaling for

loss calculation in training human pose estimator and human body mesh recovery, causing the trained models subject to local rotation and scaling errors. Moreover, top-down methods usually drop global-location information (i.e., human bounding box) which lead to excessive angular error in estimated human pose and shape [9]. Furthermore, 2D or 3D humans are usually estimated in camera coordinates instead of world coordinates, while state-of-the-art methods that estimate human in world coordinates are based on optimization and are usually slow [10]. 2D or 3D human estimations on a per-frame basis usually suffer from temporal jitter problem which is highly undesirable in most applications. Also, even well-trained monocular HPSE methods may suffer from excessive errors due to self or mutual occlusion and out-of-view truncation of human body. Therefore, multi-camera systems supporting multi-camera fusion may achieve accuracy much less susceptible to partially visible human bodies due to occlusion and truncation.

2. Motivation

Fast and accurate automatic multi-camera calibration without using predetermined calibration patterns or objects is highly desirable for various human-centric applications, where human bodies are mostly visible in the scenes, particularly for ad hoc or amateur video capturing. It is especially preferred for systems with wide baselines where using traditional calibration patterns or objects become problematic due to difficulty in correspondence matching among inputs from different cameras. Deep learning-based monocular HPSE may provide relatively reliable key points (at least for the two dimensions other than the depth dimension for each camera) that can be readily utilized as corresponding points for automatic calibration of multi-camera systems, especially when the common field of view (FoV) among cameras are narrow. The main objective of the paper is to develop feasible algorithms to fully utilize human semantic information, e.g., human body meshes, which may be readily available in many human-centric applications, for fast and accurate automatic calibration of multi-camera systems.

The desirable automatic calibration method for multi-camera systems should support reliable image-only 2D-camera-to-3D-world coordinate transformation for 3D human pose and shape estimation, 3D human body reconstruction and tracking which are less affected by self or mutual occlusion and out-of-view truncation without using extra markers or sensors. The desirable multi-camera systems also mitigate 2D reprojection errors and 3D reconstruction errors caused by depth ambiguity, where multiple 3D body configurations result in the same 2D projections, and by scale ambiguity between the size of the person and the camera distance.

Existing methods [11][12] using 2D joints from estimated human skeleton as key points for calibrating intrinsic and extrinsic parameters of a multi-camera system is feasible, but the matching and selection of corresponding key points for camera calibration are limited due to fewer typical number of 2D joints in a 2D human skeleton compared with the typical number of vertices on a 3D human body mesh. For multi-person scenarios, correspondence matching is usually time consuming and error-prone [12]. Re-identification (re-ID) networks may be required to support multi-person within camera FoVs and facilitate human tracking while increasing complexity and reduce accuracy by utilizing human bounding boxes instead of 2D joints [13]. Human semantic features extracted by human body meshes can be used for better camera calibration and multi-person identification. Using HMR / HMR 2.0 methods with SMPL / SMPL-X / STAR models supports advanced 3D human body representation more realistic than prior art joint / skeleton / landmark detections and provides crucial body shape information for more reliable re-identification in multi-person scenarios. 3D human body representation in meshes can readily derive human joint / skeleton / landmark (but not vice versa) and be used in animatable human avatar generation. The main objective of the paper is to verify that multi-camera automatic calibration can be optimized with adaptive sampling of recovered mesh vertices by matching correspondence of key points and checking consistency among selected key points.

3. Main Method

Figure 1 depicts a top-level block diagram for the proposed multi-camera automatic calibration method based on recovered human body meshes. SMPL (Skinned Multi-Person Linear Model) [1] is a 3D human body model based on skinning and blend shapes. HMR (Human Mesh Recovery) [4] and its upgraded transformer version HMR 2.0 [5] are end-to-end methods for reconstructing a full 3D mesh of a human body, even occluded or truncated, from a single RGB image by estimating its corresponding SMPL model parameters. The output of HMR include the SMPL model parameters for pose ($\theta \in \mathbb{R}^{24\times3\times3}$) and shape ($\beta \in \mathbb{R}^{10}$), and extrinsic camera parameters consist of a global orientation matrix $R \in \mathbb{R}^{3\times3}$ and translation vector $t \in \mathbb{R}^3$. Given these parameters estimated by HMR, the SMPL model outputs a 3D human body mesh $M \in \mathbb{R}^{3\times N}$ with $N = 6890$ vertices. The 3D human body mesh can be projected onto the 2D image using a perspective projection with the estimated extrinsic camera parameters. The proposed automatic calibration method for multi-camera systems takes from each camera the 3D human body meshes output from HMR [4] or HMR 2.0 [5] followed by SMPL [1] model trained with prior knowledge about 3D human body poses and shapes.

The vertices of each recovered 3D human body mesh are projected onto the 2D image plane for each corresponding camera. Structure-from-Motion (SfM) algorithm is used to reconstruct 3D shapes from a pair of cameras, using iterative RANSAC algorithm to remove outliers when the essential matrix is being calculated in each iteration. The exit condition can be set according to the zero matrix product rule and the matrix singular property. Relative camera extrinsic parameters (i.e., the rotation and translation matrices) can be calculated from the estimated essential matrix accordingly. Figure 2 shows the SfM setup and camera calibration pipeline using recovered human body meshes.

The main difference of the proposed method from common SfM processes is the use of vertices from detected 3D human body mesh as key points. This approach greatly reduces the complexity of correspondence matching between key points captured by different cameras. Among thousands of vertices of the 3D human body mesh from each camera, usually only a few hundred will be selected as corresponding key points for calculating the most consistent essential matrix between the camera pair. By assuming one main camera's pose in the world coordinate is known, the poses (i.e., the relative camera extrinsic parameters) of all other cameras in the multi-camera system can be readily calculated from the estimated essential matrix using the above mentioned procedures. The calibration results of the overall system can be further optimized using Bundle Adjustment (BA) algorithm.

The following performance metrics for camera calibration can be used to evaluate its performance.

(1) *2D reprojection error* ρ serves as a metric of how well the estimated 3D structure aligns with the observed image data. After reconstructing the 3D points in the world coordinate frame by triangulation, the estimated camera projection matrices are used to reproject these 3D points back into 2D image space. The

2D reprojection error ρ is then computed as the Euclidean distance between the initially observed 2D points and the reprojected 2D points.

$$\rho = \left\| \boldsymbol{u}_{\text{estimated}} - \boldsymbol{u}_{\text{reprojected}} \right\|_2 \quad \text{(pixels)}$$

(2) *3D reconstruction error* ξ serves as a metric of how well the reconstructed 3D structure aligns with the ground truth 3D structure. The 3D points can be reconstructed by applying triangulation, given the projection matrices of two cameras and their corresponding observed 2D points. The 3D reconstruction error ξ is then computed as the Euclidean distance between the reconstructed 3D points and the ground truth 3D points.

$$\xi = \left\| \boldsymbol{p}_{\text{reconstructed}} - \boldsymbol{p}_{\text{groundtruth}} \right\|_2 \quad \text{(meters)}.$$

(3) *Rotation error* φ and *translation error* δ are key metrics used to quantify the discrepancy between estimated and ground truth camera poses. The rotation error φ represents the angular deviation between the estimated camera orientation and the ground true orientation in the world frame, typically measured in degrees. The translation error δ refers to the Euclidean distance between the estimated and ground true position vectors of the camera, expressed in meters within the world frame.

$$\varphi = \text{angle}\left(\boldsymbol{R}_{\text{estimated}}, \boldsymbol{R}_{\text{groundtruth}}\right) \quad \text{(degrees)}.$$

$$\delta = \left\| \boldsymbol{t}_{\text{estimated}} - \boldsymbol{t}_{\text{groundtruth}} \right\|_2 \quad \text{(meters)}.$$

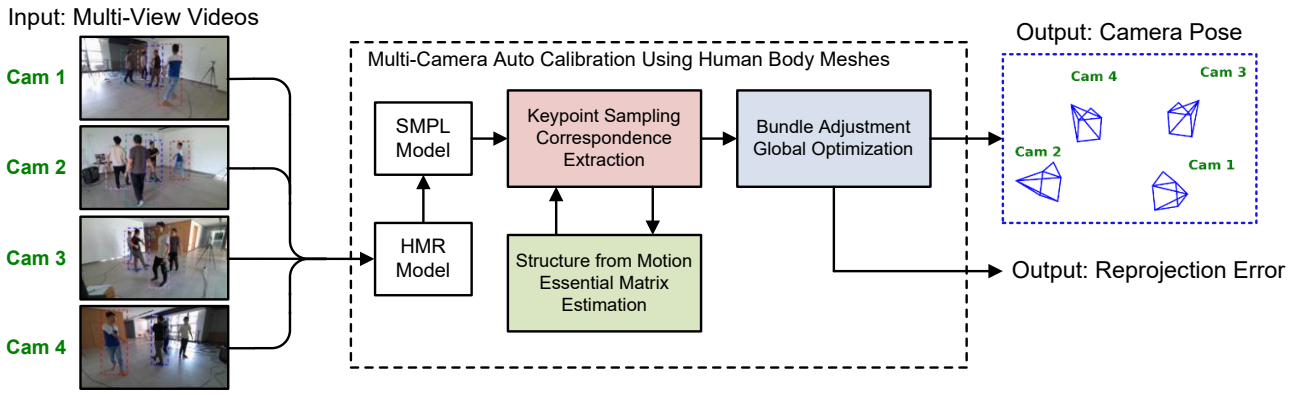


Figure 1. *Top-level block diagram for multi-camera automatic calibration method based on recovered human body meshes.*

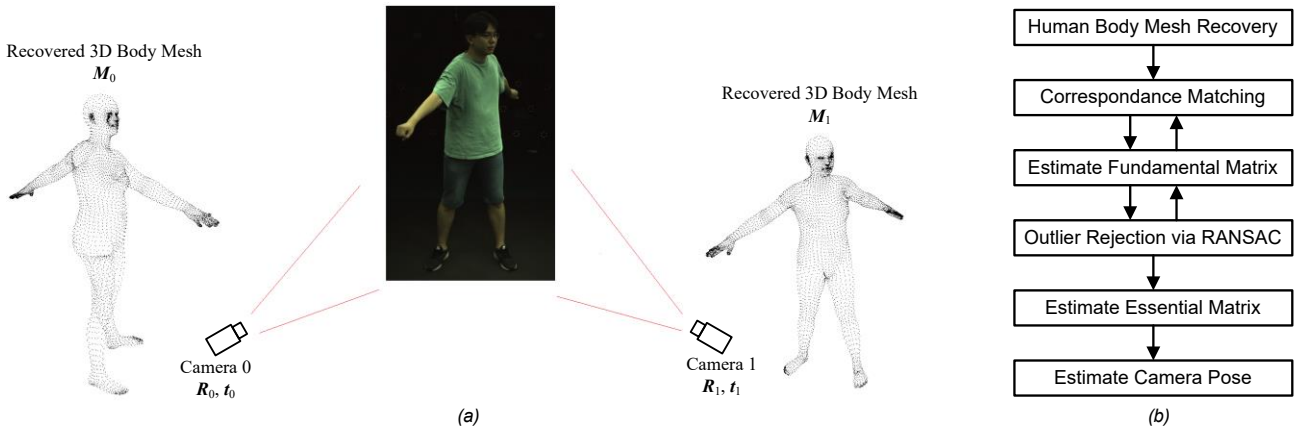


Figure 2. *(a) Structure-from-Motion (SfM) setup using recovered human body meshes. (b) Structure-from-Motion (SfM) camera calibration pipeline.*

4. Simulation Results

The overall calibration results can be evaluated using the following performance metrics: (1) average 2D reprojection error and (2) average rotation and translation errors compared with the ground truth camera extrinsic parameters. The first performance metric is universal for almost all use cases as a self-guiding performance metric without using ground truth labelled data, while the second performance metrics are useful in labs or other controlled environments for improving algorithm and fine-tuning hyper-parameters. Average 2D reprojection error can be calculated as the sum of all 2D reprojection errors divided by the total number of selected corresponding key points throughout all the tested frames. Average rotation and translation errors can be calculated as the sum of all estimated rotation matrices and translation vectors per frame divided by the number of all the tested frames, compared with the ground truth rotation matrix and translation vector, respectively.

A multi-camera dataset ZJU-MoCap [14] was used to simulate and evaluate the multi-camera calibration methods. The dataset provides 9 human subjects performing complex motion (e.g., twirling, punching, kicking) and captured by 21 circularly aligned synchronized 30fps cameras. The simulation results for six test

angle ranges are shown in Table 1 to 4 below. The two subjects [CoreView_313] and [CoreView_315] were selected from ZJU-MoCap dataset to test multi-camera calibration methods. Wide ranges of angles between cameras can be selected using different camera pairs among the 21 cameras. Each video has 1470 and 2185 frames from the corresponding camera capturing [CoreView_313] and [CoreView_315], respectively. Performance results are compared for SIFT, Human Joints, and Human Mesh methods. The proposed human mesh method achieves much higher accuracy than methods using appearance-based feature extractors, e.g., Scale-Invariant Feature Transform (SIFT), and somewhat higher accuracy than methods using deep learning-based 2D human joint estimators, e.g., OpenPose, especially for camera pairs spanning larger angles. The proposed human mesh method is also much less susceptible to partially visible human bodies due to self or mutual occlusion and out-of-view truncation.

As a visual inspection for the accuracy of corresponding matching between key points, images of subject [CoreView_313] and [CoreView_315] with color lines connecting the matching key points are shown in Figure 3 and 4, respectively, for each method with camera pairs 0-2 (≈ 30° apart), 0-18 (≈ 90° apart) and 0-12 (≈ 180° apart). It can be seen that many incorrect matches and very few correct ones resulted using the SIFT method, especially when the baseline and the angle of the camera pair become larger. The human joints and human mesh methods performed better, but the latter achieved much more correct matches and resulted in higher accuracy than the former did. These results can be expected because the appearance-based SIFT method has difficulties finding matching key points when the baseline and the angle of the camera pair become larger. Both the deep learning-based human joints and human mesh methods utilize human semantic information, but typical estimated human skeleton only contains tens of joints while typical estimated human meshes contain thousands of vertices. Therefore, the latter provides many more chances for correct correspondence matches and usually results in higher accuracy than the former does.

Table 1: For Subject [CoreView_313]: Average 2D Reprojection Error (Camera 0 is the Reference)

Methods	0° - 30° (Cam ID: 1, 22)	30° - 60° (Cam ID: 2, 3, 21)	60° - 90° (Cam ID: 4, 5, 18)	90° - 120° (Cam ID: 6, 7, 16, 17)	120° - 150° (Cam ID: 8, 9, 14, 15)	150° - 180° (Cam ID: 10, 11, 12, 13)
SIFT	102.532 pixel	684.618 pixel	239.231 pixel	190.069 pixel	1200.252 pixel	657.840 pixel
Human Joints	3.914 pixel	3.621 pixel	4.016 pixel	4.607 pixel	4.487 pixel	5.714 pixel
Human Mesh (Ours)	3.273 pixel	2.623 pixel	2.225 pixel	2.398 pixel	2.547 pixel	2.857 pixel

Table 2: For Subject [CoreView_313]: Average Rotation and Translation (R, T) Error (Camera 0 is the Reference)

Methods	0° - 30° (Cam ID: 1, 22)	30° - 60° (Cam ID: 2, 3, 21)	60° - 90° (Cam ID: 4, 5, 18)	90° - 120° (Cam ID: 6, 7, 16, 17)	120° - 150° (Cam ID: 8, 9, 14, 15)	150° - 180° (Cam ID: 10, 11, 12, 13)
SIFT	9.382°, 0.667 m	19.560°, 2.139 m	34.259°, 5.245 m	27.136°, 5.739 m	28.567°, 6.834 m	28.713°, 6.711 m
Human Joints	8.140°, 0.525 m	12.369°, 0.606 m	17.597°, 1.048 m	17.646°, 1.020 m	17.290°, 1.448 m	21.453°, 2.540 m
Human Mesh (Ours)	0.613°, 0.010 m	1.399°, 0.057 m	0.729°, 0.039 m	0.423°, 0.023 m	0.298°, 0.019 m	0.314°, 0.018 m

Table 3: For Subject [CoreView_315]: Average 2D Reprojection Error (Camera 0 is the Reference)

Methods	0° - 30° (Cam ID: 1, 22)	30° - 60° (Cam ID: 2, 3, 21)	60° - 90° (Cam ID: 4, 5, 18)	90° - 120° (Cam ID: 6, 7, 16, 17)	120° - 150° (Cam ID: 8, 9, 14, 15)	150° - 180° (Cam ID: 10, 11, 12, 13)
SIFT	146.678 pixel	635.920 pixel	403.411 pixel	363.926 pixel	1640.536 pixel	814.577 pixel
Human Joints	3.954 pixel	3.624 pixel	3.912 pixel	4.440 pixel	4.881 pixel	5.172 pixel
Human Mesh (Ours)	2.852 pixel	2.233 pixel	1.861 pixel	1.937 pixel	2.158 pixel	2.596 pixel

Table 4: For Subject [CoreView_315]: Average Rotation and Translation (R, T) Error (Camera 0 is the Reference)

Methods	0° - 30° (Cam ID: 1, 22)	30° - 60° (Cam ID: 2, 3, 21)	60° - 90° (Cam ID: 4, 5, 18)	90° - 120° (Cam ID: 6, 7, 16, 17)	120° - 150° (Cam ID: 8, 9, 14, 15)	150° - 180° (Cam ID: 10, 11, 12, 13)
SIFT	6.514°, 0.662 m	18.407°, 2.260 m	32.973°, 5.056 m	26.035°, 5.701 m	28.083°, 6.759 m	29.771°, 6.328 m
Human Joints	7.793°, 0.556 m	14.537°, 0.784 m	18.718°, 1.125 m	20.190°, 1.590 m	25.495°, 2.617 m	25.106°, 3.456 m
Human Mesh (Ours)	0.730°, 0.011 m	1.190°, 0.030 m	1.192°, 0.046 m	0.488°, 0.026 m	0.426°, 0.027 m	0.437°, 0.025 m

As a visualization for the accuracy of the camera calibration methods, the estimated versus the ground truth camera poses resulted from images of subject [CoreView_313] are shown in Figure 5 for each method with camera pairs 0-2 (≈ 30° apart), 0-18 (≈ 90° apart) and 0-12 (≈ 180° apart). It can be seen that the proposed human mesh method achieves much higher accuracy (i.e., the estimated and the ground truth camera poses are better aligned) than the SIFT method, and somewhat higher accuracy than human joints method, especially for camera pairs spanning larger angles.

5. Conclusion

The proposed automatic calibration method for multi-camera systems support reliable 3D reconstruction for human pose and shape estimation and human body tracking with reduced occlusion and truncation without extra MoCap markers or IMU sensors. Without using any calibration patterns, the proposed method uses pretrained models for recovering 3D-native human body meshes as the basis for multi-camera calibration with more reliable correspondence matching while focusing on their 2D projections onto the camera input images to avoid the depth ambiguity issues while performing automatic calibration. The scale ambiguity between the size of the person and the camera distance can also be solved by specifying a single reference length value. The proposed method uses an iterative refinement method to remove outliers in the 2D-projected human body meshes and choose the most consistent inliers for automatic binocular camera calibration without depending on confidence estimates. The proposed method uses 2D reprojection errors as self-guiding performance metric for binocular camera calibration without relying on predetermined camera calibration results such as extrinsic parameters as ground truth. The proposed method is also much less susceptible to partially visible human bodies due to self or mutual occlusion and out-of-view truncation, compared with methods using SIFT and human joints as key points.

6. Future Works

A possible extension of the proposed method is to support wide-angle or fisheye cameras (e.g., spherical or hemispherical) with wider FoVs to cover a region with less cameras but suffer from lens distortion, which can be handled by additional processing such as perspective mapping [15]. The proposed method can also be enhanced by integrating with re-identification (re-ID) network [13] to support multi-person auto calibration and joint optimization [11] for system-level camera calibration. The fully calibrated multi-camera systems are expected to substantially improve 3D reconstruction accuracy degraded by depth ambiguity, where multiple 3D body poses result in the same 2D projection. For humans only seen by one single camera in the multi-camera system, pretrained monocular 3D human pose and shape estimation (e.g., HMR+SMPL [4][5]) can be applied for 3D reconstruction. For humans seen by multiple cameras within their overlapping FoVs, multi-view 3D human pose estimation (e.g., Mvpose [16]) can be applied for more accurate 3D reconstruction which is expected to be less susceptible to self or mutual occlusion and out-of-view truncation.

References

[1] Matthew Loper, Naureen Mahmood, et al., "SMPL: A Skinned Multi-Person Linear Model," in ACM Transactions on Graphics (TOG), Vol. 34, No. 6, 2015.

[2] Georgios Pavlakos, Vasileios Choutas, et al., "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image," in Proceedings of IEEE/CVF CVPR, 2019.

[3] Ahmed A. A. Osman, Timo Bolkart, et al., "STAR: A Sparse Trained Articulated Human Body Regressor," in Proceedings of ECCV, 2020.

[4] Angjoo Kanazawa, Michael J. Black, et al., "End-to-end Recovery of Human Shape and Pose," in Proceedings of IEEE/CVF CVPR, 2018.

[5] Shubham Goel, Georgios Pavlakos, et al., "Humans in 4D: Reconstructing and Tracking Humans with Transformers," in Proceedings of IEEE/CVF ICCV, 2023.

[6] Zhe Cao, Gines Hidalgo, et al., "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 43, No. 1, January 2021.

[7] Shichao Li, Lei Ke, et al., "Cascaded Deep Monocular 3D Human Pose Estimation with Evolutionary Training Data," in Proceedings of IEEE/CVF CVPR, 2020.

[8] Yiqiao Lin, Xueyan Jiao, and Lei Zhao, "Detection of 3D Human Posture Based on Improved Mediapipe," in Journal of Computer and Communications, Vol. 11, 2023.

[9] Zhihao Li, Jianzhuang Liu, et al., "CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation," in Proceedings of ECCV, 2022.

[10] Soyong Shin, Juyong Kim, et al., "WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion," in Proceedings of IEEE/CVF CVPR, 2024.

[11] Kang Liu, Lingling Chen, et al., "Auto calibration of multi-camera system for human pose estimation," in IET Computer Vision, Vol.16, No.7, 2022.

[12] S. Dehaeck, C. Domken, et al., "Wide-baseline multi-camera calibration from a room filled with people," in Machine Vision and Applications, Vol. 34, April 2023.

[13] Yan Xu, Yu-Jhe Li, et al., "Wide-Baseline Multi-Camera Calibration using Person Re-Identification," in Proceedings of IEEE/CVF CVPR, 2021.

[14] Sida Peng, Yuanqing Zhang, et al., "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in Proceedings of IEEE/CVF CVPR, 2021.

[15] Chih-Hsien Chou and Lin-Hsi Tsao, "Automatic Calibration of Multiple Fisheye Cameras Using Recovered Human Body Mesh," in Electronic Imaging 2025, Feb. 2025.

[16] Junting Dong, Qi Fang, et al., "Fast and Robust Multi-Person 3D Pose Estimation and Tracking from Multiple Views," in IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Vol. 44, No. 10, October 2022.

Author Biography

Chih-Hsien Chou is currently a Principal Engineer in Futurewei Technologies, Inc. He has been working in R&D of real-time video / image processing algorithms for chip products since 2013. He developed WDR, NR, color correction / enhancement, video stabilization, and autofocusing algorithms. Currently his research focuses on multimodal sensing, processing, and computer vision for ARVR applications. He is the inventor or co-inventor of 20+ patents. He has a B.S. degree from Tatung University, Taiwan, a M.S. and a Ph.D. degree from University of Maryland, College Park, all in Electrical Engineering.

Lin-Hsi Tsao received his B.S. degree in Electrical Engineering from National Taiwan University and M.S. degree in Electrical and Computer Engineering from University of California, San Diego with focus on intelligent systems, robotics, and control. He has developed a deep learning-based gaze and head redirection model for photo-realistic character pose manipulation and a robust camera calibration and pose estimation method for a multi-camera system during his 2023 and 2024 internship with Futurewei Technologies, Inc.

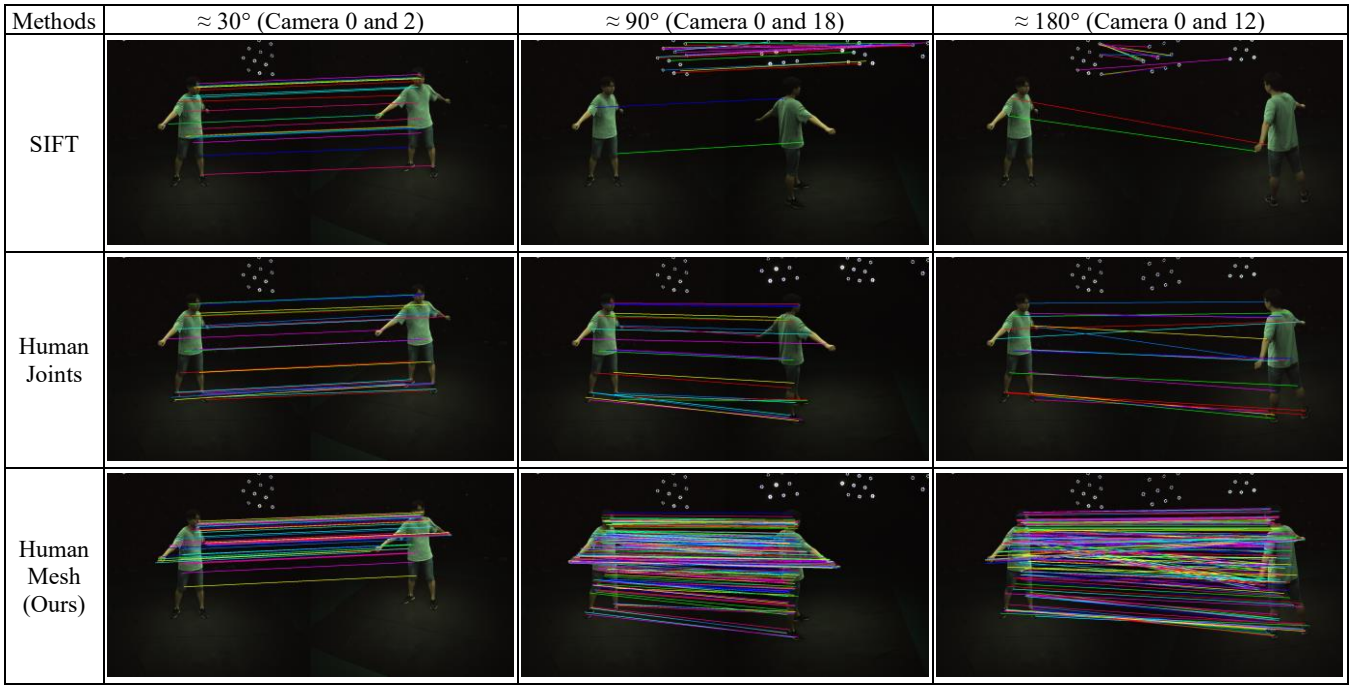


***Figure 3**. Subject [CoreView_313] correspondence matching comparison for SIFT (row 1), human joints (row 2), and human mesh (row 3) methods with narrow-baseline camera pair 0-2 (≈ 30° apart, column 1) and wide-baseline camera pairs 0-18 (≈ 90° apart, column 2) and 0-12 (≈ 180° apart, column 3).*

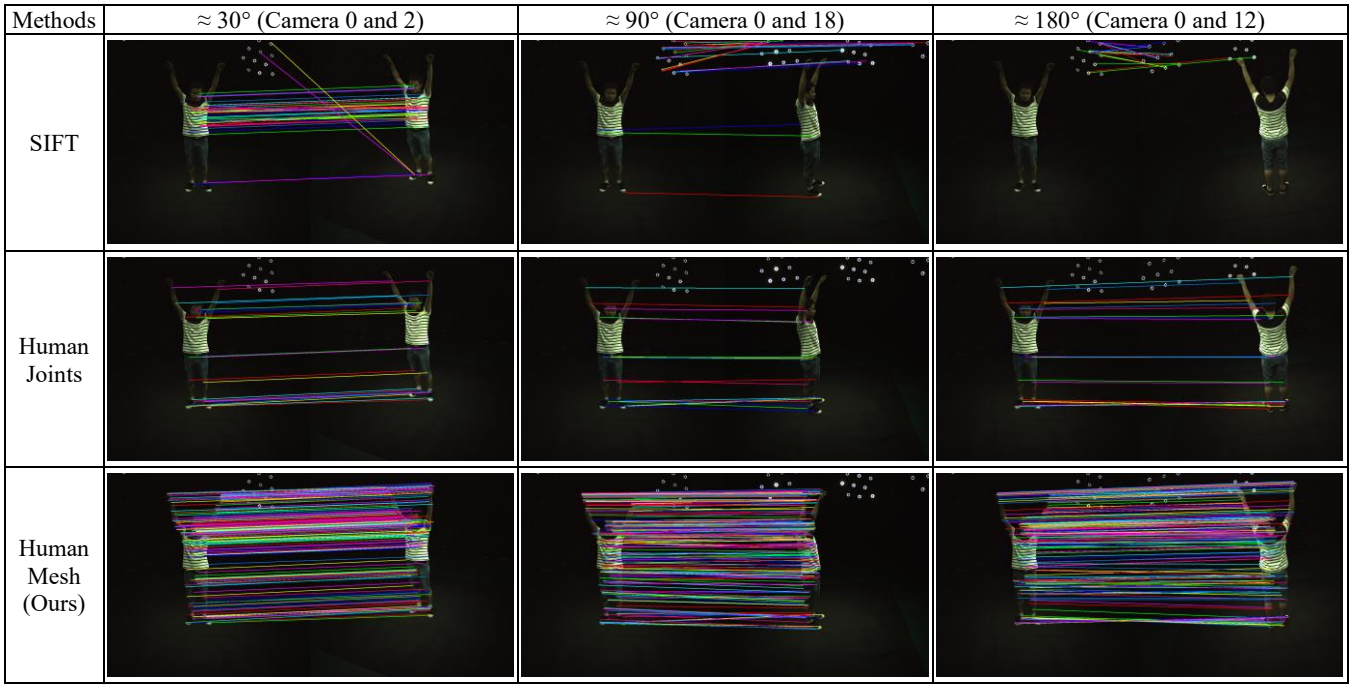


***Figure 4**. Subject [CoreView_315] correspondence matching comparison for SIFT (row 1), human joints (row 2), and human mesh (row 3) methods with narrow-baseline camera pair 0-2 (≈ 30° apart, column 1) and wide-baseline camera pairs 0-18 (≈ 90° apart, column 2) and 0-12 (≈ 180° apart, column 3).*

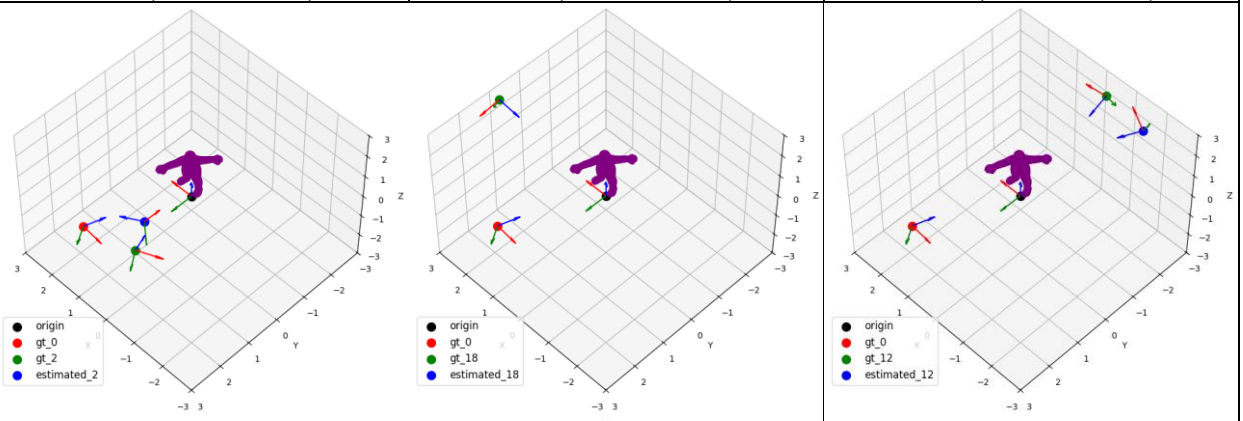
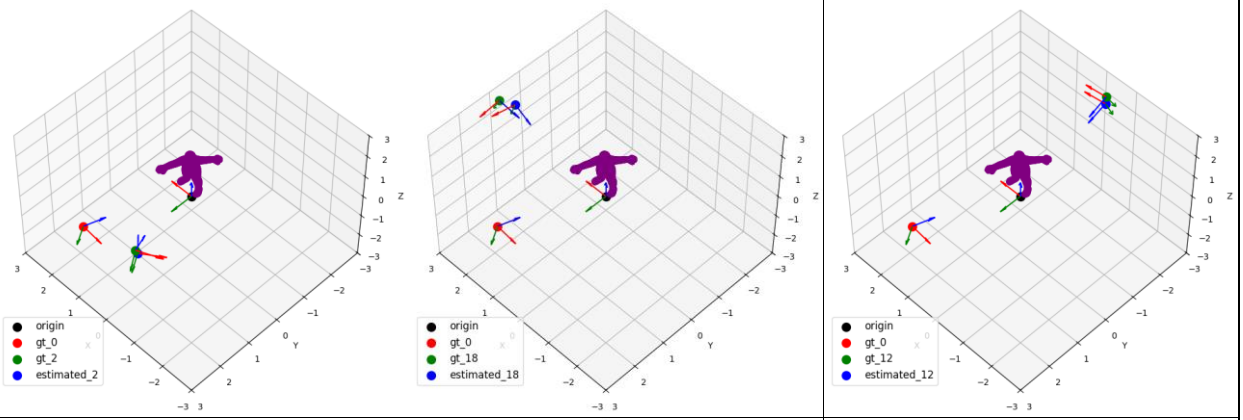
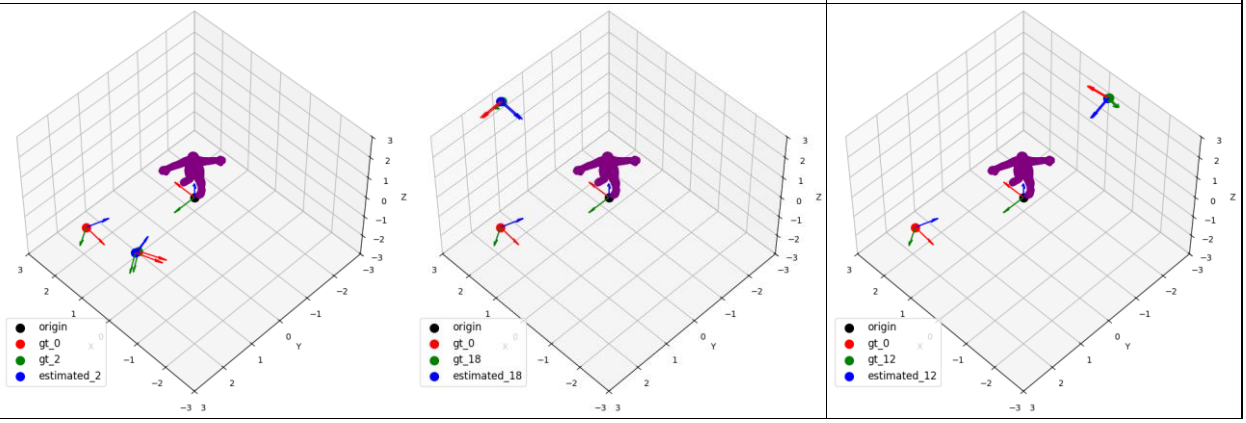
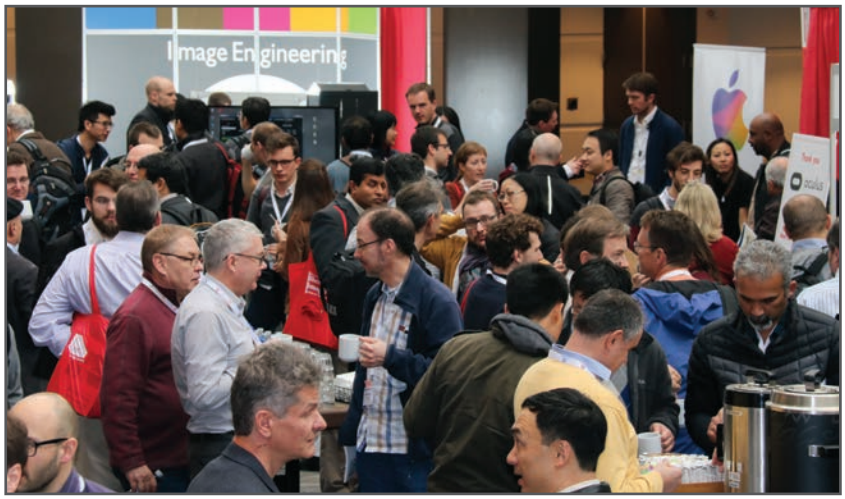
Methods	≈ 30° (Camera 0 and 2)	≈ 90° (Camera 0 and 18)	≈ 180° (Camera 0 and 12)
SIFT			
Human Joints			
Human Mesh (Ours)			

Figure 5. *Subject [CoreView_315] estimated versus ground truth camera poses for SIFT (row 1), human joints (row 2), and human mesh (row 3) methods with narrow-baseline camera pair 0-2 (≈ 30° apart, column 1) and wide-baseline camera pairs 0-18 (≈ 90° apart, column 2) and 0-12 (≈ 180° apart, column 3).*

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org