https://doi.org/10.2352/EI.2026.38.9.IPAS-241

Multi-Camera Automatic Calibration Using Human Body Meshes Recovered from Multiple Persons

Chih-Hsien Chou; 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 and occlusion problems. Multi-camera systems with wide baselines can mitigate the problems but accurate and robust multi-camera calibration is a prerequisite. The main objective for the project is to develop fast and accurate algorithms for automatic calibration of multi-camera systems which fully utilize human semantic information from multiple persons in the scene simultaneously seen by multiple cameras, without using predetermined calibration patterns or objects. The proposed method solves the multi-view matching problem by combining geometric consistency (represented by pose and shape from HPSE model) and appearance similarity (represented by feature from Re-ID model) to calculate the affinity scores between human body meshes detected from different views and then calculate the optimal permutation matrix P, which is cycle-consistent across all views for all persons seen by more than one camera. Humans seen by pairs of cameras and identified as the same person are further processed for pairwise camera calibration using Structure-from-Motion (SfM) and RANSAC algorithms to estimate the relative camera pose between the pair of cameras. The proposed method supports multiple persons in the common regions and achieves higher accuracy and faster convergence rate than existing methods using deep learning-based 2D human object detectors or 2D human joint estimators with iterative refinement for multi-person support.

1. Introduction

Human pose and shape estimation 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, where single or multiple persons will present in most scenes. Multi-camera systems provide advantages in many applications such as autonomous driving, cinema, and surveillance, where a plurality of cameras can be deployed to cover the common regions. Multi-camera systems with wide baselines can provide more reliable estimates from less reliable monocular estimates from each individual camera, which are subject to depth ambiguity and self or mutual occlusion problems [1]. However, accurate and robust multi-camera calibration is required for multi-camera systems with overlapping field of view (FoV) to mitigate the depth ambiguity and occlusion problems.

SMPL (Skinned Multi-Person Linear Model) [2] and its extended version SMPL-X (Expressive Body Capture) [3] and upgraded version STAR (Sparse Trained Articulated Human Body Regressor) [4] 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) [5] and its upgraded version HMR 2.0 (Humans in 4D) [6] 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 [7] 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 training data and suffer from non-uniqueness due to depth ambiguity [1]. Among state-of-the-art methods to solve the depth ambiguity problem, multi-camera systems with wide baselines can provide more reliable multi-view estimates from less reliable monocular estimates by each individual camera without redefining a new multi-view 3D HPSE or retraining the existing monocular 3D HPSE, but accurate and robust multi-camera calibration is required. Also, even state-of-the-art monocular HPSE methods may suffer from local rotation and scaling errors due to imperfect loss function during training and from excessive errors due to self or mutual occlusion and out-of-view truncation of human bodies [1]. Therefore, multi-camera systems supporting multi-view image fusion are expected to achieve accuracy with less rotation and scaling errors and 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. It is also highly desirable to support scenes with multiple persons that can be simultaneously seen by multiple cameras. This not only removes the unnecessary restriction that only a single person can be seen by multiple cameras, but it also provides more coverage and diversity during a given time interval in the common regions of a multi-camera system. However, the multi-view matching problem needs to be solved to consistently identify the same person as seen by different cameras at different

view angles without confusion and contradiction. For multi-view matching among multiple persons viewed by multiple cameras, cycle consistency [8] is highly desired where two corresponding humans in two views must be matched to the same human in another view to avoid inconsistent correspondence among multiple views. Also, the total number of people is usually unknown and needs to be accurately estimated.

OpenPose bottom-up network for human skeleton detection was previously used for multi-camera calibration without using calibration patterns, but only a single person in the scene is supported [9]. Re-identification (Re-ID) networks supporting multiple human bounding boxes were also used for multi-camera calibration with multi-person support, but each detected person is represented by only a single 2D point which may affect the accuracy of the calibration [10]. Multiple human skeleton detection together with geometric cross-view matching and iterative refinement for multi-view matching were used for multi-camera calibration with multi-person support [11], where the cross-view matching was based on trial-and-error without using Re-ID techniques. Multiple iterations are required while convergence to optimal results may not always be achieved. YOLOv4 for human detection and HRNet for joint estimation, together with a person Re-ID network for multi-person matching with cycle consistency were used for multi-person 3D pose estimation and temporal tracking [8], where camera calibration is required but automatic calibration is not supported.

Using HMR [5] / HMR 2.0 [6] methods with SMPL [2] / SMPL-X [3] / STAR [4] models supports advanced 3D human body mesh representation more realistic than existing joint / skeleton / landmark methods and may provide 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 can be used in animatable human avatar generation. It is expected that automatic multi-camera calibration using 3D human body meshes can be optimized by adaptive sampling of mesh vertices. The main objective of the paper is to develop fast and accurate algorithms for multi-camera automatic calibration to fully utilize dense human semantic information, e.g., 3D human body meshes, from multiple persons, which may be readily available in many human-centric applications. Furthermore, non-iterative algorithms are preferred to achieve cycle consistency for multi-person support.

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 with multi-person support. SMPL (Skinned Multi-Person Linear Model) [2] is a 3D human body model based on skinning and blend shapes. HMR (Human Mesh Recovery) [5] and its upgraded transformer version HMR 2.0 [6] are end-to-end methods for estimating the corresponding SMPL model parameters for reconstructing a full 3D mesh of each human body, even occluded or truncated, from the view of each camera. The proposed automatic calibration method for multi-camera systems takes from each camera the SMPL model parameter and bounding box outputs for each human detected by HMR or HMR 2.0 which are pre-trained with prior knowledge about 3D human body poses and shapes.

A person Re-ID model [12] trained on massive datasets learns a camera-invariant subspace to deal with style variations from different cameras due to lighting and viewpoint changes. A pre-trained Re-ID model extracts discriminative appearance features as descriptor vectors from the cropped image of each bounding box associated with humans detected by HMR or HMR 2.0 in each view from each camera. The output of the Pool-5 layer of the pre-trained Re-ID model is extracted as the cropped image's 2,048-dim feature descriptor κ for each bounding box. The Euclidean distances are computed as the appearance similarity between cropped images containing detected humans. The results are mapped to (0, 1) using a sigmoid function to obtain the *appearance affinity score*.

The output of HMR includes the SMPL model parameters for pose ($\mu \in \mathbb{R}^{24\times3\times3}$) 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. A pre-trained HMR model estimates pose μ and shape β parameters for 3D human body meshes of each human detected in each view. Their differences are used to compute geometric similarity between two 3D human body meshes. The pose μ and shape β differences are mapped to (0, 1) using a sigmoid function to obtain the *geometric affinity score*.

A multi-view matcher establishes the correspondences of the detected body across views using feature κ, pose μ, and shape β parameters associated with each detected human. A cross-view matching selector sends recovered body meshes of the same person from pairs of cameras to a mesh matching unit where the recovered body meshes are projected onto the 2D image plane for each camera. Structure-from-Motion (SfM) algorithm is used to reconstruct 3D shapes from a pair of cameras, using iterative RANSAC algorithm to remove outliers during pairwise camera calibration [1].

Geometric and Appearance Affinity Matrices

Suppose there are N views captured by N cameras in the system and m_i detected human body meshes in view i. Detected 3D human body meshes of the same person are to be matched across multiple views. Appearance affinity (from feature κ) and geometric affinity (from pose μ and shape β) are used to calculate the *combined affinity scores* between all detected human body meshes and bounding boxes from view i and j. For *appearance affinity matrix* block $\boldsymbol{Aa}_{ij}$ (size $m_i \times m_j$), compute the Euclidean distance between the feature κ vectors of each pair of bounding boxes of detected humans from view i and j, then map the distances to values in (0,1) using a sigmoid function. For *geometric affinity matrix* block $\boldsymbol{Ag}_{ij}$ (size $m_i \times m_j$), compute the angle difference between the pose μ vectors (i.e., total angle difference between the two sets of 24 rotation matrices for human joints) and the Euclidean distance between the shape β vectors, respectively, of each pair of recovered 3D human body meshes from view i and j, then map the weighted sum of the results to values in (0,1) using a sigmoid function. For *combined affinity matrix* $\boldsymbol{A} = [\boldsymbol{A}_{ij}]$ (size $\sum_i m_i \times \sum_j m_j$), compute block $\boldsymbol{A}_{ij} = \mathrm{SQRT}(\boldsymbol{Ag}_{ij} \circ \boldsymbol{Aa}_{ij})$ for all views i and view j, where SQRT and $\circ$ denote element-wise square-rooting and multiplication, respectively.

Permutation Matrix for Multi-View Matching

The combined affinity matrix block $\boldsymbol{A}_{ij}$ shows the likelihood that each pair of recovered 3D human body meshes from view i and j belongs to the same person. A multi-view matching algorithm should optimize a *permutation matrix* $\boldsymbol{P} = [\boldsymbol{P}_{ij}]$ (size $\sum_i m_i \times \sum_j m_j$) to establish the correspondences (i.e., 0 or 1) of 3D human body meshes across all views, while the total number of people K in the scene is unknown. The *optimal permutation matrix* $\boldsymbol{P}_{\mathrm{opt}}$ maximizes affinities and is cycle-consistent across all views, i.e., any two corresponding human body meshes in two views should correspond to the human body mesh (from the same person) in another view. With proper relaxations for discrete rank operator and integer matrix, Alternating Direction Method of Multipliers (ADMM) [8],

[13] is used to solve the convex optimization problem to obtain $\boldsymbol{P}_{\text{opt}}$, whose rank($\boldsymbol{P}_{\text{opt}}$) will be the estimated total number of people K. It can also be decomposed as $\boldsymbol{P}_{\text{opt}} = \boldsymbol{Y}\boldsymbol{Y}^T$, where the *correspondence matrix* $\boldsymbol{Y}$ (size $\sum_i m_i \times K$) provides cycle-consistent correspondences of human body meshes across all views to each identified person in the scene. Figure 2 shows a processing pipeline for affinity matrices calculation and multi-view matching among 3D human body meshes recovered from all views. Figure 3 provides an illustrative example showing optimization of permutation matrix for multi-view matching among $K = 6$ humans viewed by all $N = 4$ cameras.

Figure 4 provides an illustrative example showing procedures of the proposed multi-camera automatic calibration with multi-person support. Input multi-view videos captured by the four cameras are sent to HMR + Re-ID models where the feature κ, pose μ, and shape β parameters associated with each detected human are estimated. Multi-view matcher + SMPL model then solves the multi-view matching problem and reconstructs 3D body mesh for each identified person. Note that 2D skeletons instead of 3D body meshes are marked on each detected human only for illustration purposes. Cross-view matching selector selects pairs of body meshes identified as the same person from two different views for pairwise camera calibration to be performed by mesh matching unit. With a simultaneous view of the identified 3D human body mesh by a pair of cameras, the relative pose of one camera to the other can be expressed by the resulted essential matrix.

Pose accumulation to accumulate the essential matrix values for the same pair of cameras is then performed after each pairwise camera calibration is performed. 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 averaged essential matrix values after pose accumulation. The pose accumulation process is iterated for more input video frames until a stopping criterion is met, where the condition can be set as each camera has been covered by pairwise camera calibration for at least L (as a hyperparameter) times. To solve the overall scale ambiguity for the multi-camera system, a single known length value should be assigned, usually by assuming an average person height or by measuring the distance between cameras. The calibration results of the overall system can then be globally optimized using bundle adjustment or similar algorithms.

Those human body meshes recovered from only one view with no matches in other views are regarded as useless and skipped during multi-view matching. Wrong matchings of different people as the same person will be rejected as outliers by SfM + RANSAC in pairwise camera calibration. The application specific knowledge about the multi-camera configuration can be obtained through user input or automatic detection to reduce the camera pairs to be matched. Automatic multi-camera calibration can be performed in either record-then-compute mode, or on-the-fly mode when only one out of every M (as a hyperparameter) frames will be captured by each camera during calibration and then processed. Pairwise camera calibration can be performed in a sequential, concurrent, or hybrid scheme depending on system resources, speed, calibration time, and accuracy requirements. For privacy-preserving use cases, each camera only sends feature κ, pose μ, and shape β parameters associated with detected humans to central or cloud servers for performing multi-camera calibration.

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

(1) *Average 2D reprojection error* $\bar{\rho}$ 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. $\bar{\rho}$ is then computed as the average Euclidean distance (in pixels) between the initially observed 2D points and the reprojected 2D points, where N_1 is the total number of reprojected points.

$$\bar{\rho} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(\left\|\boldsymbol{u}_i^{\boldsymbol{estimated}} - \boldsymbol{u}_i^{\boldsymbol{reprojected}}\right\|_2\right). \quad (1)$$

(2) *Average 3D reconstruction error* $\bar{\xi}$ 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. $\bar{\xi}$ is then computed as the average Euclidean distance (in meters) between the reconstructed 3D points and the ground truth 3D points, where N_2 is the total number of reconstructed points.

$$\bar{\xi} = \frac{1}{N_2}\sum_{i=1}^{N_2}\left(\left\|\boldsymbol{p}_i^{\boldsymbol{reconstructed}} - \boldsymbol{p}_i^{\boldsymbol{groundtruth}}\right\|_2\right). \quad (2)$$

(3) *Average rotation error* $\bar{\varphi}$ and *translation error* $\bar{\delta}$ are key metrics used to quantify the discrepancy between the estimated and ground truth camera poses. $\bar{\varphi}$ represents the angular deviation (in degrees) between the estimated and the ground true camera orientation in the world frame. $\bar{\delta}$ refers to the Euclidean distance (in meters) between the estimated and the ground true position vectors of the camera within the world frame, where N_3 is the total number of pairwise camera calibrations.

$$\bar{\varphi} = \frac{1}{N_3}\sum_{i=1}^{N_3}\text{angle}\left(\boldsymbol{R}_i^{\boldsymbol{estimated}}, \boldsymbol{R}_i^{\boldsymbol{groundtruth}}\right). \quad (3)$$

$$\bar{\delta} = \frac{1}{N_3}\sum_{i=1}^{N_3}\left(\left\|\boldsymbol{t}_i^{\boldsymbol{estimated}} - \boldsymbol{t}_i^{\boldsymbol{groundtruth}}\right\|_2\right). \quad (4)$$

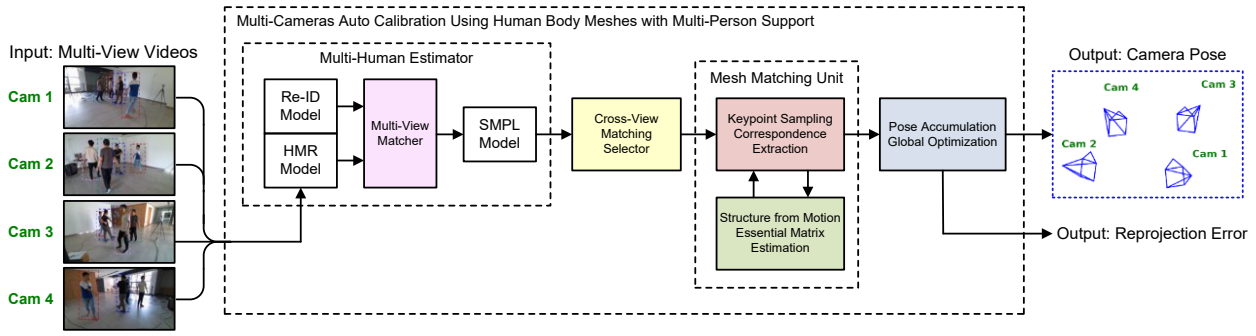


Figure 1. *Top-level block diagram for multi-camera automatic calibration method based on recovered human body meshes with multi-person support.*

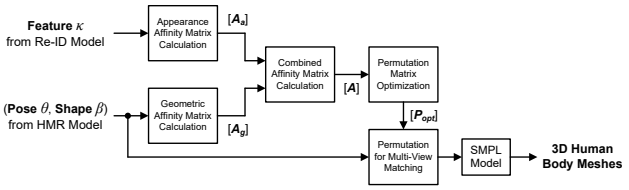


Figure 2. *Processing pipeline for affinity matrices calculation and multi-view matching among 3D human body meshes recovered from all views.*

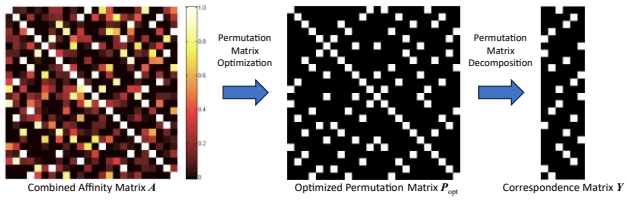


Figure 3. *An illustrative example showing optimization of permutation matrix for multi-view matching among all humans detected from all cameras.*

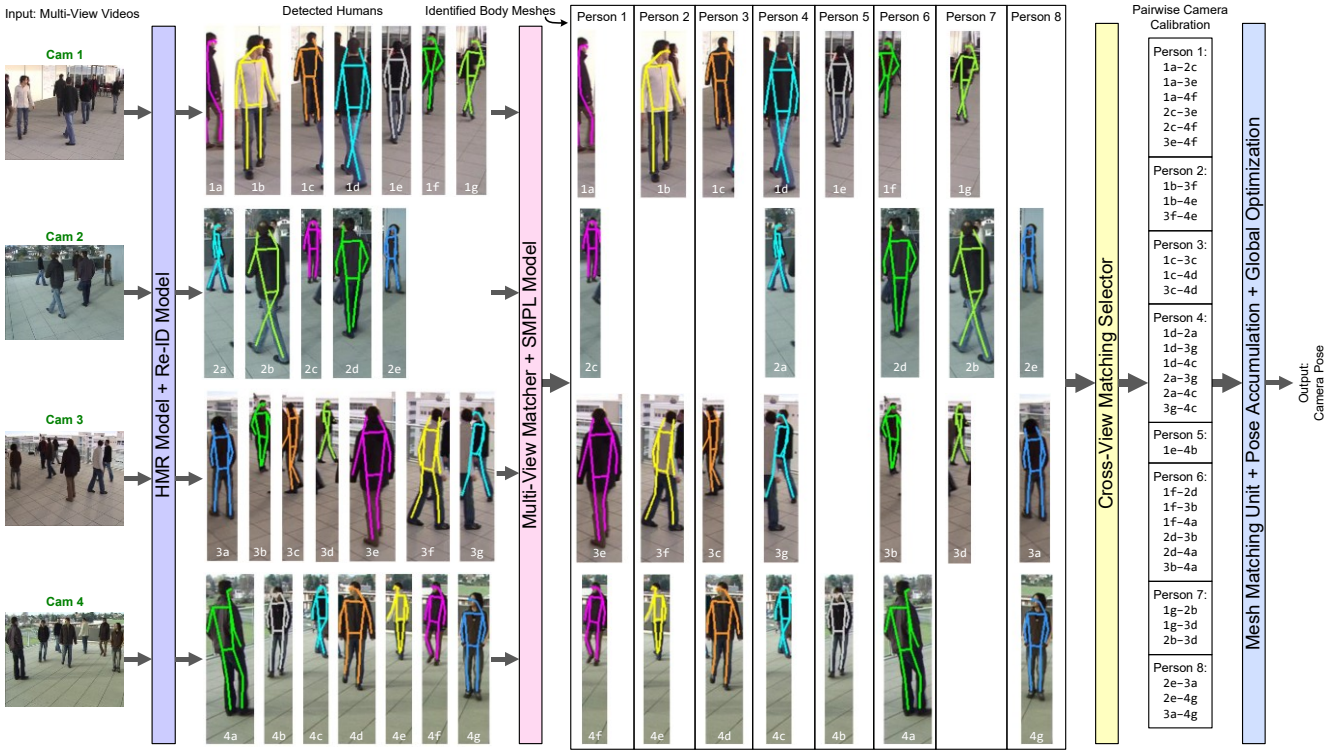


Figure 4. *An illustrative example showing procedures of the proposed multi-camera automatic calibration using recovered multi-person 3D human body meshes.*

4. Simulation Results

A single-person multi-camera ZJU-MoCap dataset [14] was used to simulate and evaluate the pairwise 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 of pairwise camera calibration are compared among SIFT, Human Joints, and Human Mesh methods for six test angle ranges [1], as shown in Figure 5. The proposed human mesh method for pairwise camera calibration achieves much higher accuracy (i.e., the estimated and the ground truth camera poses are better aligned) than methods using appearance-based feature extractors, e.g., Scale-Invariant Feature Transform (SIFT) [15], and somewhat higher accuracy than methods using deep learning-based 2D human joint estimators, e.g., OpenPose, especially for camera pairs spanning larger angles. 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.

A multi-person multi-camera EPFL dataset [16] of pedestrian videos was used to simulate and evaluate the multi-camera calibration methods with multi-person support. In the *Terrace* sequence, there are 8 people walking slowly in a 7m × 11m outdoor area in front of 4 cameras. In the *Basketball* sequence, there are up to 14 people playing basketball in a 17.5m × 22m outdoor area in front of 4 cameras. For both sequences, all 4 cameras capture videos in CIF (352×288) resolution @25 fps for 5000 frames with ground truth camera calibration data provided in the dataset.

The simulation results in comparison with two baseline methods and three deep learning-based methods are shown in Table 1. For the first two baseline methods SIFT + BFM and SuperPoint + BFM, keypoints are detected from images using SIFT [15] or SuperPoint network [17], respectively, and matched across cameras using Brute-Force Matching (BFM) [18]. For the following three methods using person Re-ID models [12], manually annotated point correspondences (Manual-pts), manually associated bounding boxes (Manual-bbox), and Re-ID network selected bounding boxes (ReID-bbox) [10] are used, respectively. The proposed method

using multi-view matching to select 3D human body meshes performs marginally better than the method using ReID network to select bounding boxes [10]. Improvement in the calibration accuracy is expected after performing global optimization and fine-tuning the parameter values, such as the computation of appearance and geometric affinity scores. In the simulation of the proposed algorithm, only one out of every M (set to 10) frames will be captured by each camera during calibration, and the stopping criterion will be met when each camera has been covered by pairwise camera calibration for at least L (set to 200) times. When a human body is simultaneously viewed by n cameras in the captured frames during calibration, it will trigger $n(n-1)/2$ pairwise camera calibration processes. Therefore, the running time for the proposed multi-camera calibration algorithm will depend on the number of cameras N, total number of people K in captured frames, and the values of hyperparameter M and L. In densely populated areas that can be covered by multi-camera systems, the typical running time will be in seconds to tens of seconds. To further verify the accuracy of the proposed algorithm in various camera configurations and human behaviors, synthetic datasets may be used where the ground truth extrinsic calibration parameters are precisely known, as they have been used in the definition of the camera locations for rendering the synthetic videos.

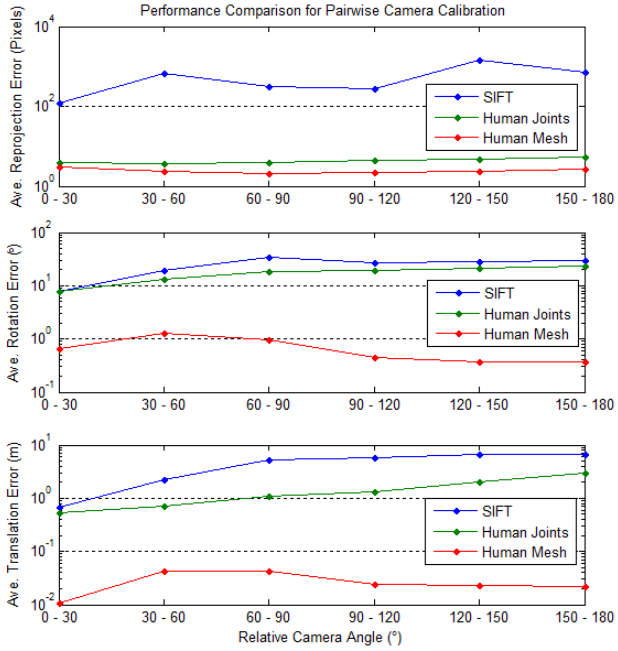


***Figure 5**. Performance of pairwise camera calibration compared among SIFT, Human Joints, and Human Mesh methods for six test angle ranges.*

Table 1: Performance comparison for Multi-Camera Calibration

Performance Metrics	**Average 2D Projection Error** (pixels)		**Average Rotation and Translation** Errors (°, m)	
Methods	*Terrace*	*Basketball*	*Terrace*	*Basketball*
SIFT [15] + BFM [18]	254.40	533.15	55.03°, 4.599 m	65.45°, 6.140 m
SuperPoint [17] + BFM [18]	53.07	9.50	54.68°, 0.358 m	53.97°, 20.065 m
Manual-pts [10]	0.45	0.51	1.18°, 0.390 m	0.66°, 0.358 m
Manual-bbox [10]	2.30	0.88	0.52°, 0.308 m	0.85°, 0.490 m
ReID-bbox [10]	2.30	1.01	0.52°, 0.308 m	0.88°, 0.410 m
Proposed	**2.15**	**0.97**	**0.43°, 0.236 m**	**0.74°, 0.363 m**

5. Conclusion

The proposed automatic calibration method for multi-camera systems supports reliable 3D reconstruction and tracking of multiple human bodies without using extra MoCap markers or IMU sensors. Without using any calibration patterns or objects, the proposed method uses pre-trained models to extract human feature descriptors and recover 3D human body meshes for robust multi-person support and reliable correspondence matching to avoid inconsistency and depth ambiguity issues during automatic calibration. The proposed method uses a non-iterative algorithm for multi-view matching among people. The proposed method is much less susceptible to partially visible human bodies due to self or mutual occlusion and out-of-view truncation, compared with using SIFT and human joints. For privacy-preserving applications, cameras can only send estimated human feature, pose, and shape parameters to central server to perform the multi-camera automatic calibration.

6. Future Works

One limitation of the proposed method may result in failure cases when some or all people in the scene have (almost) same uniforms, similar body shapes, and identical poses, e.g., band parade, group dance, or team exercise, etc. Although the conditions for stopping criterion in pose accumulation may be modified to reject such incorrect results, more robust algorithms may need to be developed to achieve correct results under such challenging cases. 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 spherical human detection (to find bounding boxes of humans from distorted input images) and perspective mapping (to map distorted image patches to an ideal virtual camera) [19]. Multi-person multi-view datasets, e.g., FTV360 [20], for multiple fisheye cameras can be used to simulate and evaluate the calibration methods. The proposed method may support use cases with larger (e.g., cars with key points) or smaller (e.g., hands with meshes or faces with landmarks) scales, but the object detection, mesh recovery, and Re-ID models need to be replaced or retrained to fully utilize semantic information of the targeted objects in the scene.

References

[1] Chih-Hsien Chou and Lin-Hsi Tsao, “Wide-Baseline Muli-Camera Automatic Calibration Using Recovered Human Body Mesh,” in Electronic Imaging 2025, Feb. 2025.

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

[3] 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.

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

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

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

[7] Zhe Cao, Gines Hidalgo, et al., "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), Vol. 43, No. 1, January 2021.

[8] Junting Dong, Qi Fang, et al., "Fast and Robust Multi-Person 3D Pose Estimation and Tracking from Multiple Views," in IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), Vol. 44, No. 10, October 2022.

[9] 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.

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

[11] 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.

[12] Zhun Zhong, Liang Zheng, et al., "Camera Style Adaptation for Person Re-Identification," in Proceedings of IEEE/CVF CVPR, 2018.

[13] Stephen Boyd, Neal Parikh, et al., "Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers," in Found. Trends Mach. Learn., Vol. 3, No. 1, 2011.

[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] David G Lowe, "Distinctive image features from scale-invariant keypoints," in International Journal of Computer Vision, Vol. 60, No.2, 2004.

[16] François Fleuret, Jérôme Berclaz, et al., "Multicamera People Tracking with a Probabilistic Occupancy Map," in IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI), Vol. 30, No. 2, February 2008.

[17] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich, "Superpoint: Self-supervised interest point detection and description," in Proceedings of the IEEE/CVF CVPR Workshops, 2018.

[18] Amila Jakubović and Jasmin Velagić, "Image feature matching and object detection using brute-force matchers," in 2018 International Symposium ELMAR, 2018.

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

[20] Thomas Maugey, Laurent Guillo, and Cédric Le Cam, "FTV360: a Multiview 360° Video Dataset with Calibration Parameters," in Proceedings of ACM MMSys, June 2019.

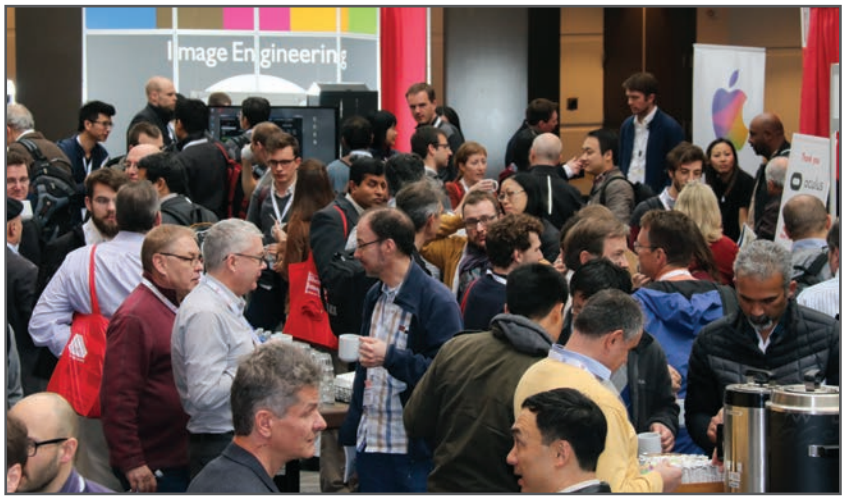
Author Biography

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.

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