https://doi.org/10.2352/EI.2025.37.10.IPAS-347

Automatic Calibration of Multiple Fisheye Cameras Using Recovered Human Body Mesh

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

Abstract

Fisheye cameras providing omnidirectional vision with up to 360° field-of-view (FoV) can cover a given space with fewer cameras for a multi-camera system. The main objective of the paper is to develop fast and accurate algorithms for automatic calibration of multiple fisheye cameras which fully utilize human semantic information without using predetermined calibration patterns or objects. The proposed automatic calibration method detects humans from each fisheye camera in equirectangular or spherical images. For each detected human, the portion of image defined by the bounding box will be converted to an undistorted image patch with normal FoV by a perspective mapping parameterized by the associated view angle. 3D human body meshes are then estimated by 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 image patch. Structure-from-Motion (SfM) algorithm is used to reconstruct 3D shapes from a pair of cameras and estimate the essential matrix. Camera extrinsic parameters can be calculated from the estimated essential matrix and the corresponding perspective mappings. 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 calculated. Fisheye camera pairs spinning different angles are evaluated using (1) 2D projection error and (2) rotation and translation errors as performance metrics. The proposed method is shown to perform more accurate calibration than methods using appearance-based feature extractors, e.g., Scale-Invariant Feature Transform (SIFT), and deep learning-based human joint estimators, e.g., MediaPipe.

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 suffers from occlusion [1] and depth ambiguity [2] problems and may fail for rare or unseen poses due to limited and fixed training data [3]. Fisheye cameras provide omnidirectional vision with up to 360° field-of-view (FoV), much wider than typical perspective vision from traditional cameras, providing advantages in many applications such as telepresence, autonomous driving, cinema, and surveillance. Systems of multiple fisheye cameras with wide baselines can provide more reliable multi-view 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 such systems to accurately cover a wide region using a minimum number of fisheye cameras with overlapping FoVs and properly mitigate the self or mutual occlusion and depth ambiguity problems.

SMPL (Skinned Multi-Person Linear Model) [4] and its extended version SMPL-X (Expressive Body Capture) [5] and upgraded version STAR (Sparse Trained Articulated Human Body Regressor) [6] 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) [7] and its upgraded version HMR 2.0 (Humans in 4D) [1] 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. Therefore, 3D human meshes can be estimated from an image patch defined by a bounding box containing a detected person, without the need to wear any MoCap markers or IMU sensors.

Deep learning-based monocular 3D human pose estimation may fail for rare or unseen poses due to limited and fixed training data [3]. 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 [2]. In worst-case scenarios, incorrect body tilt depends on hand / body stretches for lack of diversified human poses in training data.

Fisheye cameras provide omnidirectional vision with up to 360° FoV and output images typically in equirectangular or spherical format, which cause serious object shape distortion (i.e., deformation), especially for objects captured away from the equator (in equirectangular format) or the optical axis (in spherical format). Applying object detection trained on undistorted perspective images to images from fisheye cameras usually results in lower accuracy due to shape distortion. Applying view-angle dependent perspective mapping before object detection to images from fisheye cameras achieves higher accuracy while introducing excessive computation cost. For accurate and efficient object detection on images from fisheye cameras, new architectures were designed specifically for spherical data. Kernel transformer network (KTN) [8] was proposed to efficiently perform spherical convolution (SphConv) [9] on images in equirectangular format, and can be applied to CNN based object detection, e.g., Faster R-CNN [10], for accurate and efficient object detection on images from fisheye cameras.

Multi-camera systems with wide baselines 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. Even well-trained monocular HPSE methods may suffer from excessive errors due to self or mutual occlusion and out-of-view truncation of human bodies. 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. While fisheye cameras can provide up to 360° FoV to cover a wide region with fewer cameras, there are additional challenges due to shape distortion caused by their distorted output images typically in equirectangular or spherical format. Most deep learning-based monocular HPSE models are trained on images in perspective view with normal FoV (≈ 60°), therefore, perspective mapping is usually a necessary pre-processing step for accurate HPSE. 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 multiple fisheye cameras.

Existing multi-camera automatic calibration methods [11][12] using 2D joints from estimated human skeleton as key points for estimating intrinsic and extrinsic camera parameters are 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 [14]. However, none of the above methods support automatic calibration of multiple fisheye cameras suffering from serious shape distortion.

A spherical Faster R-CNN implementation was proposed in [8] where the backbone CNN is replaced by SphConv [9] to transfer the features from planar images to spherical images. The feature map is then projected onto tangent planes before the region proposal network (RPN) is applied on the projected feature maps. Redundant proposals from all tangent planes are removed using a proposed spherical non-maximum suppression (NMS) process [8]. The refined proposals are then fed into the detector network to generate the final object detection outputs. The spherical Faster R-CNN detector can successfully detect objects despite the serious object shape distortion in the spherical images from fisheye cameras.

The main objective of the paper is to verify that automatic calibration of multiple fisheye cameras can be achieved using pretrained models and its accuracy can be improved 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 automatic calibration method for multiple fisheye cameras based on recovered human body meshes. The human detector can be implemented using spherical Faster R-CNN [8] which detects various objects of different classes, scales, and aspect ratios in equirectangular or spherical images from each camera, and outputs a bounding FoV with an object class and a softmax score in [0, 1] for each detected object. In the system, only humans detected with the object class *Person* and a score larger than a preset threshold of 0.6 will trigger the following processes. The human detector outputs an α-degree bounding FoV centered at (θ, φ) associated with each detected human, where θ is the polar angle and φ is the azimuthal angle in the input spherical image I_s from a fisheye camera. For each detected human, a perspective mapping $\mathbb{P}(I_s, \alpha, \theta, \varphi) = I_p$ will be performed to project the associated α-degree FoV from I_s to a $W\times W$ pixels image patch I_p on the tangent plane at view angle (θ, φ).

The view-angle dependent perspective mapping associated with each detected person can be considered to be a virtual camera, with its optical axis pointing at the center of the associated bounding FoV and its view covering the associated bounding FoV. For each virtual camera, the ideal pinhole camera model is assumed to provide an undistorted perspective view to HPSE without suffering from the serious geometric distortion caused by the 360° fisheye camera. It enables HPSE to recover geometrically consistent human body meshes from multiple fisheye cameras with different view angles. Each 360° fisheye camera is typically composed of two back-to-back 180° hemispherical cameras (i.e., front and rear), each captures the incident rays incoming from a hemisphere.

It is assumed that the intrinsic parameters of each 360° fisheye camera are pre-calibrated offline and known, while the extrinsic parameters of each 360° fisheye camera will be estimated by the proposed automatic calibration method. For each virtual camera, it is assumed that its position coincides with the corresponding 360° fisheye camera, while its rotation matrix $\boldsymbol{R_v}$ and intrinsic matrix $\boldsymbol{K_v}$ can be derived directly from the associated bounding FoV: $\boldsymbol{R_v} \approx \mathrm{ROT}(\varphi)\,\mathrm{ROT}(\theta)$ and $\boldsymbol{K_v} = [f_v\ 0\ x_0;\ 0\ f_v\ y_0;\ 0\ 0\ 1]$, where $\mathrm{ROT}(\varphi)$ and $\mathrm{ROT}(\theta)$ are rotation matrices about the Y and X axes, respectively. $f_v = W/2\tan(\alpha/2)$ is the focal length and $x_0 = y_0 = W/2$ are the principal point offsets of the virtual camera. For each virtual camera, the extrinsic camera matrix from the 3D world coordinates to its 2D pixel coordinate is $[\boldsymbol{R_v}|0]\,[\boldsymbol{R_f}|\boldsymbol{t_f}]$, where $\boldsymbol{R_f}$ and $\boldsymbol{t_f}$ are the rotation matrix and translation vector of the corresponding fisheye camera.

The undistorted image patch captured by a virtual camera can be interpolated from the spherical image(s) captured by one or both of the hemispherical cameras inside the corresponding fisheye camera, depending on whether the bounding FoV is seen by one or both. Unified spherical model [15] can be applied for inverse mapping from the targeted perspective image patch I_p with $W\times W$ pixels back to the input spherical image(s) I_s and nearest-neighbor or bilinear interpolation can be used in order to compute perspective mapping efficiently with minor image quality degradation.

3D human body meshes are then estimated from these image patches by Human Mesh Recovery (HMR [7] or HMR 2.0[1]) followed by SMPL [4] model trained with prior knowledge about 3D human body poses and shapes. The output of HMR model include 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. The 3D human body mesh can be projected onto the 2D image plane of each virtual

camera using a perspective projection with the extrinsic parameters $[R|t]$ estimated from each virtual camera by HMR.

The vertices of each recovered 3D human body mesh are projected onto the image plane of the corresponding virtual camera. Structure-from-Motion (SfM) algorithms in OpenCV library were used to reconstruct 3D shapes from a pair of virtual cameras, using iterative RANSAC algorithm to remove outliers when the essential matrix is being calculated in each iteration. More details about wide-baseline multi-camera automatic calibration using recovered human body mesh are provided in [14]. By assuming one main fisheye camera's pose in the world coordinate is known, the poses (i.e., the camera extrinsic parameters) of all other fisheye cameras in the multi-camera system can be readily calculated from the estimated essential matrix and the view angle parameters for the associated perspective mappings for the corresponding pair of virtual cameras. Figure 2 shows the SfM setup and camera calibration pipeline using human body meshes recovered from a pair of virtual cameras. The calibration results of the overall system can be further optimized using Bundle Adjustment (BA) algorithm [12].

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 projection matrices of the virtual cameras 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.

$$\rho = \left\| \boldsymbol{u}_{\text{estimated}} - \boldsymbol{u}_{\text{reprojected}} \right\|_2 \quad \text{(pixels)}. \tag{1}$$

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

$$\psi = \text{angle}\,(\boldsymbol{R}_{\text{estimated}}, \boldsymbol{R}_{\text{groundtruth}}) \quad \text{(degrees)}. \tag{2}$$

$$\delta = \left\| \boldsymbol{t}_{\text{estimated}} - \boldsymbol{t}_{\text{groundtruth}} \right\|_2 \quad \text{(meters)}. \tag{3}$$

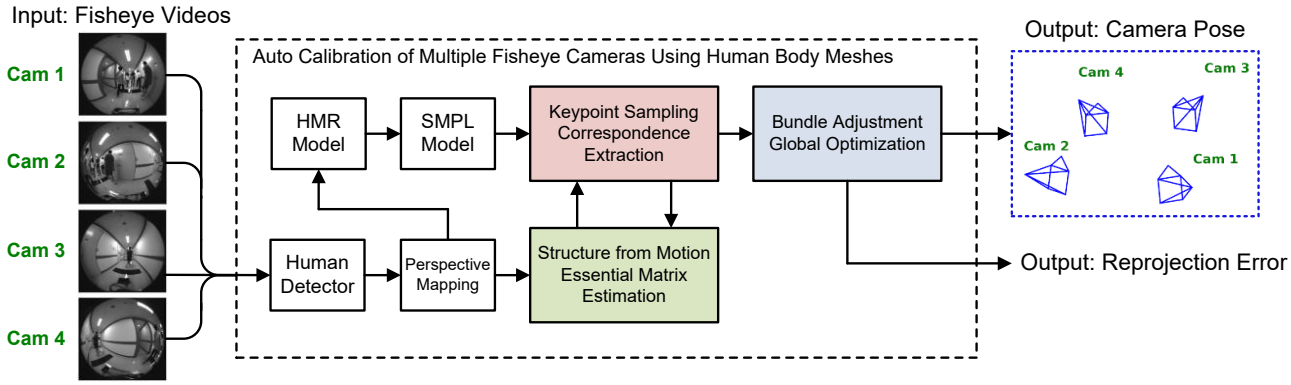


Figure 1. *Top-level block diagram for automatic calibration method for multiple fisheye cameras based on recovered human body meshes.*

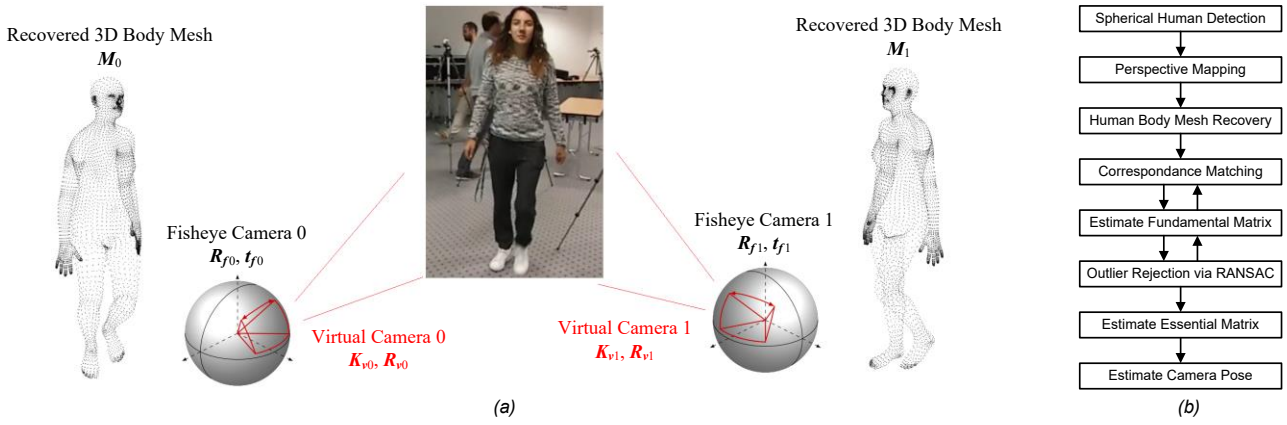


Figure 2. *(a) Structure-from-Motion (SfM) setup using recovered human body meshes. (b) Structure-from-Motion (SfM) camera calibration pipeline.*

4. Simulation Results

The overall calibration results can be evaluated using the following performance metrics: (1) 2D reprojection error and (2) 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.

A multi-view dataset FTV360 [15] was used to simulate and evaluate the calibration methods for multiple fisheye cameras. The dataset provides multiple people performing complex motion (e.g.,

walking, racing, gaming) and captured by 40 synchronized 30fps 360° fisheye cameras arranged in a 5×8 rectangular grid. The spacing between adjacent fisheye cameras is about 1.5m in the rear-front direction and 2.5m in the left-right direction. Each 360° fisheye camera used in capturing the FTV360 dataset is actually composed of two back-to-back 180° hemispherical cameras, which will be considered as two separate cameras in the following discussion. Wide ranges of angles between cameras can be selected using different camera pairs among the 40 fisheye cameras. Without loss of generality, a group of six adjacent fisheye cameras were selected as an example camera configuration. Two walking persons captured in the [indoor-walk] video sequence were selected from FTV360 dataset to evaluate the performance of automatic camera calibration methods for multiple fisheye cameras. Figure 3 depicts two test cases for the selected group of six fisheye cameras from the FTV360 dataset during the [indoor-walk] video sequence at a frame when a walking man or a walking lady can be seen by all six cameras in the group simultaneously.

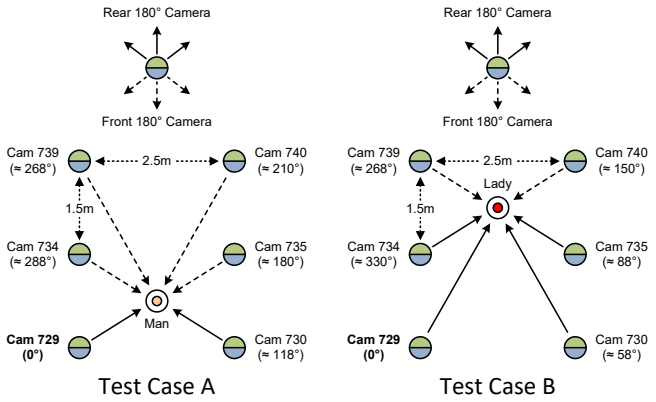


***Figure 3**. Two test cases for a selected group of six adjacent fisheye cameras from the FTV360 dataset during the [indoor-walk] video sequence.*

Figure 4 and 6 show the representative video frames captured by each of the six cameras for the two test cases when the walking man and the walking lady, respectively, can be detected by all six cameras in the group with different view angles. It is obvious to find that the body shapes of the walking persons are seriously distorted in a captured frame if they are away from the optical axis of the capturing hemispherical camera. The simulation results for the two test cases are shown in Table 1 to 4 below. 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., MediaPipe [2] or OpenPose [16], 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 correspondence matching between key points, undistorted image patches of the walking man and the walking lady with color lines connecting the matching key points are shown in Figure 5 and 7, respectively, for each method with camera pairs 729-730, 729-735, 729-740, 729-739, and 729-734. 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.

Table 1: For walking man: 2D Reprojection Error (Camera 729 is the Reference)

Methods	≈ 118° Cam 729-730	≈ 180° Cam 729-735	≈ 210° Cam 729-740	≈ 268° Cam 729-739	≈ 288° Cam 729-734
SIFT	137.924 pixel	857.716 pixel	369.217 pixel	237.804 pixel	1571.382 pixel
Human Joints	6.481 pixel	7.293 pixel	9.614 pixel	11.743 pixel	12.576 pixel
Human Mesh (Ours)	5.346 pixel	4.721 pixel	4.096 pixel	4.796 pixel	4.843 pixel

Table 2: For walking man: Rotation and Translation (R, T) Errors (Camera 729 is the Reference)

Methods	≈ 118° Cam 729-730	≈ 180° Cam 729-735	≈ 210° Cam 729-740	≈ 268° Cam 729-739	≈ 288° Cam 729-734
SIFT	12.635°, 0.863 m	24.454°, 3.833 m	37.792°, 5.587 m	29.463°, 6.329 m	30.625°, 6.748 m
Human Joints	11.285°, 0.586 m	13.693°, 0.647 m	19.715°, 1.124 m	20.415°, 1.308 m	18.965°, 1.672 m
Human Mesh (Ours)	1.236°, 0.027 m	2.696°, 0.118 m	1.426°, 0.076 m	0.834°, 0.048 m	0.592°, 0.039 m

Table 3: For walking lady: 2D Reprojection Error (Camera 729 is the Reference)

Methods	≈ 58° Cam 729-730	≈ 88° Cam 729-735	≈ 150° Cam 729-740	≈ 268° Cam 729-739	≈ 330° Cam 729-734
SIFT	217.523 pixel	948.673 pixel	389.734 pixel	347.907 pixel	1372.253 pixel
Human Joints	7.671 pixel	Fail	9.156 pixel	13.328 pixel	10.597 pixel
Human Mesh (Ours)	6.845 pixel	5.384 pixel	5.607 pixel	5.875 pixel	6.087 pixel

Table 4: For walking lady: Rotation and Translation (R, T) Errors (Camera 729 is the Reference)

Methods	≈ 58° Cam 729-730	≈ 88° Cam 729-735	≈ 150° Cam 729-740	≈ 268° Cam 729-739	≈ 330° Cam 729-734
SIFT	9.753°, 0.873 m	19.884°, 3.255 m	33.328°, 5.318 m	27.541°, 6.131 m	29.203°, 7.179 m
Human Joints	8.317°, 0.665 m	Fail	19.617°, 1.276 m	21.279°, 1.782 m	23.375°, 3.134 m
Human Mesh (Ours)	1.227°, 0.023 m	1.792°, 0.061 m	1.894°, 0.094 m	0.786°, 0.053 m	0.831°, 0.056 m

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. Note that the human joints method (using MediaPipe [2]) failed to detect human skeleton of the walking lady in the image patch from Cam 735 rear.

5. Conclusion

The proposed automatic calibration method for multiple fisheye cameras supports reliable 3D reconstruction for human pose and shape estimation and human body tracking in a setting with wide baseline among cameras. Without using any calibration patterns, the proposed method uses pretrained models to detect humans in equirectangular or spherical images and to recover 3D-native human body meshes for more reliable correspondence matching while focusing on their 2D projections onto the corresponding perspective images to avoid the depth ambiguity issues during automation calibration. The proposed method applies spherical objection detection and perspective mapping to support wide-angle or fisheye cameras (e.g., spherical or hemispherical) with wider FoVs for covering a region with less cameras but suffering from lens distortion. 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

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. For multi-camera systems in larger or more complicated environments, not all cameras have highly overlapping FoVs among them. For a larger number of cameras in a multi-camera system, camera pairs can first be locally calibrated within different connected sub-systems. The application specific knowledge about the multi-camera configuration can be obtained through user input or automatic detection to reduce the number of camera pairs where cross-view correspondence matching should be performed. A set of correct correspondences between all cameras can be extracted for a global calibration of the entire system if there are common cameras shared by these sub-systems. The fully calibrated multi-camera systems are expected to substantially improve 3D reconstruction accuracy degraded by depth ambiguity, which causes multiple 3D body poses to result in the same 2D projection.

References

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

[2] 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.

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

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

[5] 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.

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

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

[8] Yu-Chuan Su and Kristen Grauman, "Learning Spherical Convolution for 360° Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 44, No. 11, November 2022.

[9] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling, "Spherical CNNs," in Proceedings of 6th International Conference on Learning Representations (ICLR), May 2018.

[10] Shaoqing Ren, Kaiming He, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 39, No. 6, June 2017.

[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] Chih-Hsien Chou and Lin-Hsi Tsao, "Wide-Baseline Multi-Camera Automatic Calibration using Recovered Human Body Mesh," in Electronic Imaging 2025, Feb. 2025.

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

[16] 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.

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.

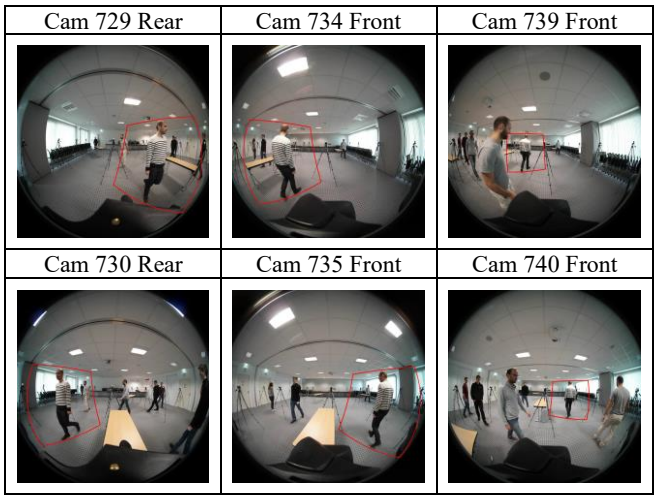


Figure 4. *A walking man detected by the selected group of six adjacent fisheye cameras from the FTV360 dataset during the [indoor-walk] video sequence.*

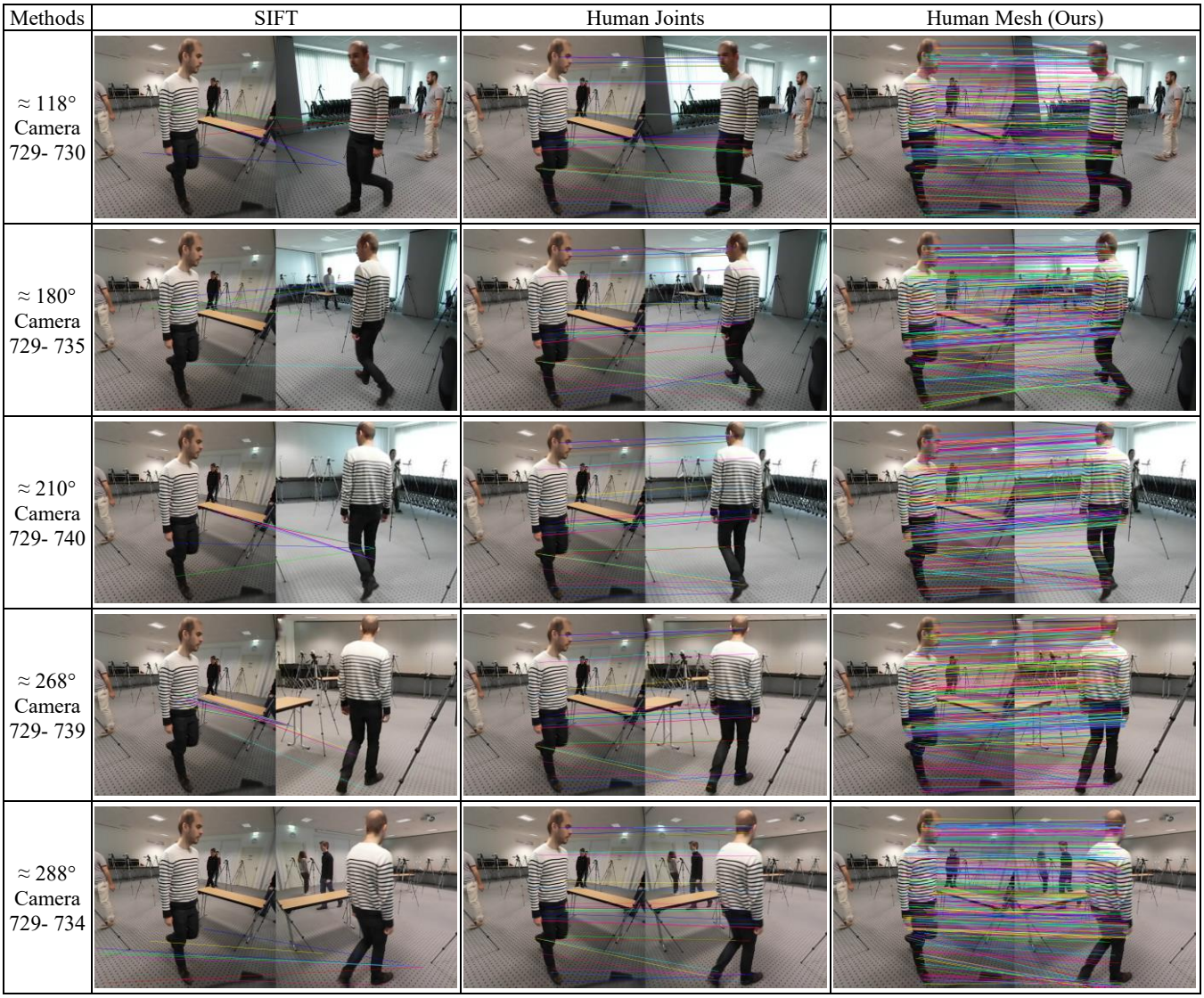


Figure 5. *Walking man correspondence matching comparison on undistorted image patches for SIFT (column 1), human joints (column 2), and human mesh (column 3) methods with camera pairs 729-730, 729-735, 729-740, 729-739, and 729-734.*

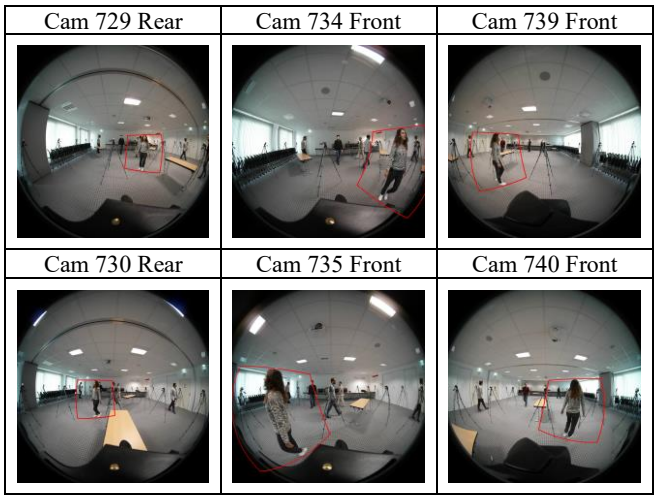


Figure 6. *A walking lady detected by the selected group of six adjacent fisheye cameras from the FTV360 dataset during the [indoor-walk] video sequence.*

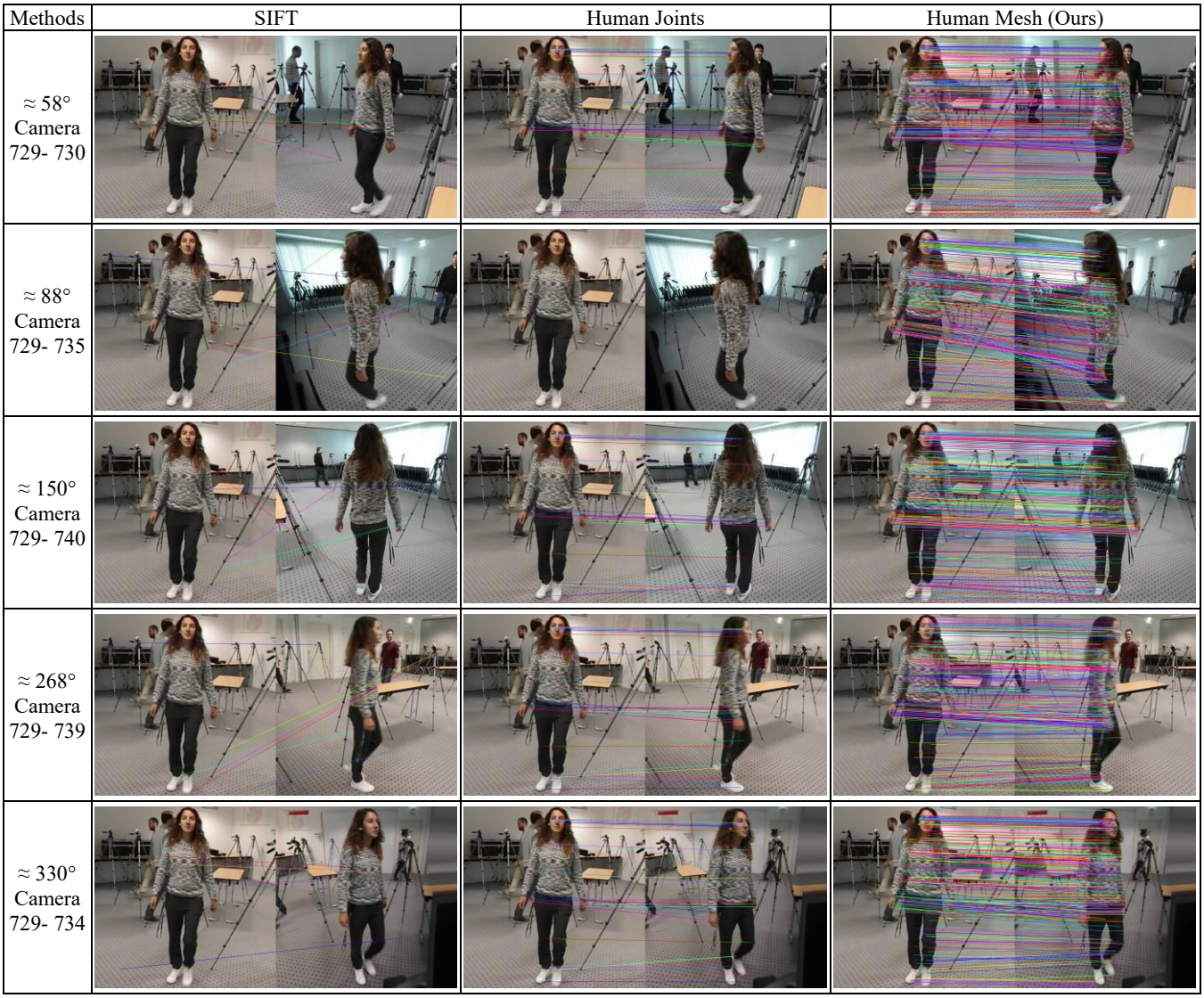
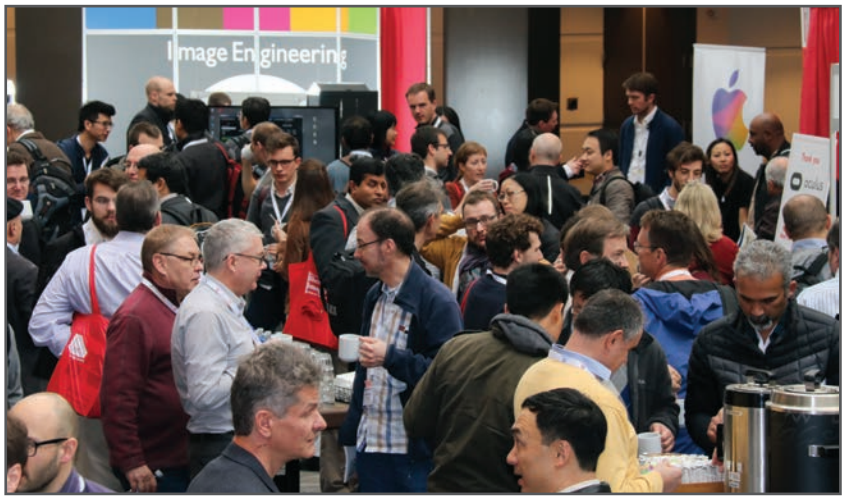


Figure 7. *Walking lady correspondence matching comparison on undistorted image patches for SIFT (column 1), human joints (column 2), and human mesh (column 3) methods with camera pairs 729-730, 729-735, 729-740, 729-739, and 729-734.*

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