MultiView Markerless MoCap - MultiView Performance Capture, 3D Pose Motion Reconstruction and Comparison

Pratik Singh Bisht, Philippe Colantoni, Damien Muselet, Alain Trémeau
Université de Lyon, Université Jean Monnet,
Laboratoire Hubert Curien, UMR 5516
Email: pratik.singh.bisht@etu.univ-st-etienne.fr
(philippe.colantoni,damien.muselet,alain.tremeau)@univ-st-etienne.fr

***Abstract*—This article explores the challenges and advancements in multi-view camera systems, 2D pose estimation, and 3D reconstruction for capturing and reconstructing live performances. It conducts a comparative analysis of methodologies at each stage of the pipeline, identifying strengths, weaknesses, and effective techniques. The study addresses the robustness of existing techniques in diverse scenarios and aims to integrate these fields into a unified framework for high-quality reconstructions. It contributes to the development of advanced multi-camera systems applicable across domains and serves as a valuable resource for future research in the field of performance capture and analysis.**

***Index Terms*—Multiple camera setup, 3D reconstruction, Calibration, 3D Pose Estimation, Triangulation, Mesh fitting**

I. Introduction

Capturing and reconstructing live performances with realism is a complex challenge, requiring advancements in multi-view camera systems, 2D pose estimation, and 3D reconstruction. This study explores these areas, focusing on their application in performance capture. It investigates camera calibration and synchronization, pose estimation using frameworks like OpenPose [1] and AlphaPose [23], 3D skeleton and mesh generation from 2D poses, and data visualization. The research's main aim is to compare and analyze methods across stages, including multi-camera synchronization, pose estimation models, and 3D reconstruction as shown in the framework diagram Fig. 1. It evaluates technique robustness in diverse scenarios to identify those suitable for various conditions, addressing challenges in real-world settings.

This study delves into multi-view camera systems based on GoPro Hero 10 and ZED 2i, investigating advanced methods, research gaps, and applications. Its aim is to create a comprehensive methodology for real-time 3D reconstruction, extending beyond pose estimation.

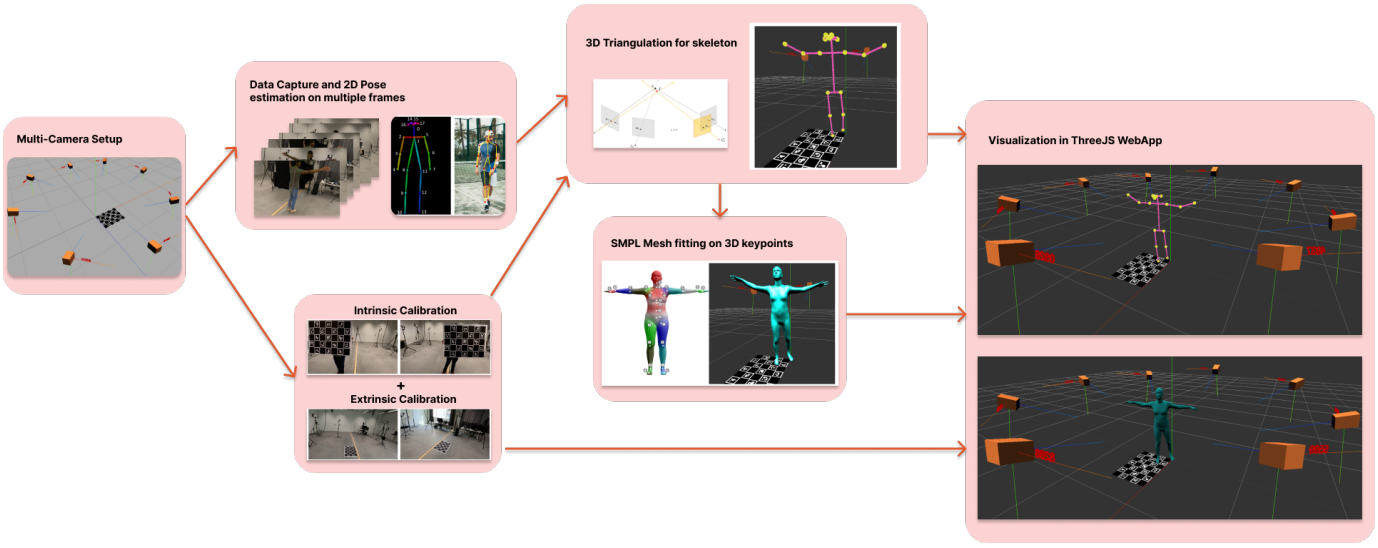

Fig. 1. Overview of the MultiView Markerless MoCap framework

The research proposes a model-agnostic data accumulation approach for versatile multi-camera systems, impacting areas like sports analysis and virtual reality. Additionally, it provides a foundational framework for the PREMIERE project [22],

which aims to digitize performing arts using MoCap (Motion Capture) technology. This research guides future project directions, offering a methodological roadmap. Its comparison methodology benefits fellow researchers, driving progress in the field. In essence, this work advances the field by proposing an integrated framework to bridge gaps and enhance multi-camera systems for live performance reconstructions. We also plan on sharing the pipeline structure, scripts and code through github at a later stage.

II. Background and Related Works

The domains of computer vision and robotics have been enriched by multi-camera capture systems and pose estimation methods, which employ multiple cameras for precise scene representation and object tracking. Advances in machine learning and affordable cameras have driven interest in these technologies, with transformative potential in live performance capture, virtual reality, and human-computer interaction.

A. Multi-Camera Systems: Synchronization, Calibration, and Localization

Utilizing multiple cameras for scene capture involves synchronization, calibration, and localization. Synchronization can be achieved through hardware methods like timecode generators, genlock [3], phase-locked loops or software approaches like cross-correlation [4] and Kalman filtering. Calibration extends single camera calibration to estimate intrinsic parameters and relative positions using automated methods like feature matching and bundle adjustment. Challenges arise from distortion, synchronization, and parameter estimation, driving ongoing research for accurate calibration in complex setups. Aruco markers [5] play a pivotal role in camera calibration and localization by serving as reference points. They aid in calculating intrinsic and extrinsic parameters, enabling real-time 6-degree-of-freedom localization as cameras navigate in the environment. This approach streamlines both calibration and real-time localization processes, offering an efficient solution for multi-camera systems.

B. Pose Estimation and 3D Reconstruction

Pose estimation is a crucial aspect of computer vision that involves detecting and tracking human body parts like joints and limbs in 2D images, 3D images, and videos. It finds applications in robotics, human-computer interaction, and animation. Different methods exist for human pose estimation, including bottom-up and top-down approaches. Bottom-up methods, like OpenPose [1], use convolutional neural networks (CNNs) to detect body parts, with features like Part Affinity Fields (PAF) capturing part relationships. Top-down methods, such as Convolutional Pose Machines (CPMs) [6] and AlphaPose, start with a complete body model and fit it to images, handling challenges like occlusion and varying body scales. AlphaPose (RMPE) [2] employs predefined human body models to identify keypoints in images or videos, estimating body poses. Performance evaluation can be done with datasets like COCO Keypoints [17] and metrics such as average precision (AP) and mean average precision (mAP).

Following pose estimation, 3D reconstruction from multiple 2D poses is the next crucial step. This process involves generating a 3D model of an object or a scene from multiple 2D views, a critical step in applications such as live performance capture and reconstruction.

Two distinct approaches for 3D reconstruction from 2D poses are direct and indirect methods. Direct methods like structure from motion (SfM), multiple view stereo (MVS), and multi-view triangular mesh generation (MVT) create 3D models directly from 2D poses [7] [8]. Indirect methods estimate 3D poses from 2D poses and then reconstruct the scene, often employing human body models like SMPL [9] and MANO [10]. Approaches like HMR [11] and SMPLify-X [12] aim to achieve 3D pose estimation and reconstruction from a single RGB image. They use parametric body models like SMPL and refine poses, joint angles, and scale to generate 3D pose estimates from a monocular image. HybrIK [13] [14] and GLAMR [15] address limitations in depth information due to camera or subject movement. GLAMR enhances results in monocular image/video contexts, achieving consistent global coordinates for dynamic camera scenarios.

C. ZED Fusion

An alternative framework to our pipeline for 3D skeleton is the ZED 2i camera. This powerful stereoscopic camera captures image data and depth maps within a range of 20 centimeters to 20 meters. It uses stereo vision to create depth maps by comparing pixels in two images and calculating the depth of objects. A neural network can be used to correct errors in depth maps for accuracy.

III. Methodology

This section describes step-by-step the process of creating a multi-view camera system for data capture, pose estimation, 3D triangulation and reconstruction. This framework is based on state-of-the-art techniques related to multi-camera calibration, 2D pose estimation, 3D reconstruction and leverages the open-source tools and libraries available for each of these tasks.

A. Multi-Camera Setup

The initial phase consists in configuring the multi-camera system. An optimal arrangement may be found with eight GoPro Hero 10 cameras positioned in a balanced geometric pattern, such as a circle or rectangle, to cover a 360-degree scene. Cameras must be spaced at equiangular intervals to minimize errors in triangulation, a process where accurate distribution impacts precision. The recommended eight-camera setup ensures balanced coverage, enhancing accuracy in pose estimation and 3D reconstruction. However, scene requirements may dictate adjustments. Factors like scene area size, performers, and detail level may affect camera choice. While the eight-camera setup is a solid starting point, scene-specific adjustments might be necessary for optimal outcomes. Further details and comparisons are available in the results section.

B. Data Acquisition

This stage concerns the process of capturing data, mainly to achieve two purposes: camera calibration and pose estimation. The multi-camera setup defined in the previous step first needs to be calibrated (this is further explained in the next sections) and then used to capture actual target subject movements in the scene for 2D pose estimation analysis.

1) Camera Calibration: The calibration process for our multi-camera setup is based on a two-step procedure, focusing on intrinsic and extrinsic parameters.

This two-step calibration process ensures that each camera's perspective is accurately captured and related to a common coordinate system, enabling accurate multi-view capture and 3D reconstruction. The standard measure used for quantitatively analyzing camera calibration is typically referred to as the Reprojection Error (RE). This error assesses the discrepancy caused by various factors when projecting points back onto the scene. It is computed as the average L2 norm of the correspondence between observed feature points on the image plane and the projection of predicted 3D feature points after considering distortion. To convert the Reprojection Error from pixel-wise measurements to real-world coordinates, the following approach can be employed [16]:

$$FOV(rad) = 2 \cdot arctan(\frac{width}{2f_x}) \tag{1}$$

where $width$ is horizontal width and f_x is horizontal focal length.

$$PixelAngle(rad) = \frac{FOV}{width} \tag{2}$$

$$AngularError(rad) = RE \cdot PixelAngle \tag{3}$$

$$LengthError(m) = Distance \cdot AngularError(rad) \tag{4}$$

This results in a Length Error (LE) margin, in meters, for real world coordinate system. This error is only related to camera calibration procedure, there might be further errors introduced by other steps, e.g. at pose estimation and mesh fitting stage.

2) Preprocessing: The Preprocessing stage is a pivotal part of the pipeline, preparing raw data from the multi-camera setup for subsequent analysis. This stage encompasses video capture, synchronization, and relevant section clipping. Initial actions involve recording videos of the subject from multiple viewpoints using the multi-camera setup. Simultaneous recording results in diverse poses and movements. Synchronization can be ensured through an external flash used at the beginning of capture, its duration may match the camera frame rate for precise alignment.

C. Pose Estimation

Computer vision and machine learning are pivotal here. Processed videos go through pose estimation models like AlphaPose or OpenPose. These models, using CNNs, detect and track human body keypoints, estimating pose accurately. Most of models give coordinates for body keypoints in formats like COCO with 17 points. Extra points, like SMPL's 24, are crucial for precise 3D reconstruction and mesh fitting. Precise pose impacts 3D quality, requiring models balancing accuracy, speed, occlusion handling, and appearance variations. Post-processing refines results, categorizing poses and using Kalman filtering for smoothing. This phase also holds potential for multi-pose support. This paper explores various pose estimation models, adaptable to upcoming state-of-the-art models aligned with mesh fitting. While not delving deeply into model intricacies, the focus remains on the approach and output outcomes.

D. Triangulation

Triangulation, a vital step in 3D reconstruction, aims to determine the 3D coordinates of a point in space by analyzing its projections in multiple images. This process relies on epipolar geometry, describing the geometric relationships among different images of the same scene. Input data for triangulation includes calibration information for each camera (intrinsic matrix, distortion coefficients, and extrinsic matrix) and 2D pose estimations. The intrinsic matrix (K) captures internal camera parameters like focal length and principal point. Distortion coefficients (D) address lens distortion causing image curvature. The extrinsic matrix (E) defines the camera's world pose involving rotation and translation.

The 2D pose estimation provides a set of 2D points p_i in each image, which correspond to the projections of the same 3D point P in the world. The goal of triangulation is to find the 3D coordinates of P.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \alpha \begin{bmatrix} p_1 & p_2 & p_3 & p_4 \\ p_5 & p_6 & p_7 & p_8 \\ p_9 & p_{10} & p_{11} & p_{12} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} y\boldsymbol{p}_3^\top - \boldsymbol{p}_2^\top \\ \boldsymbol{p}_1^\top - x\boldsymbol{p}_3^\top \\ y'\boldsymbol{p}_3'^\top - \boldsymbol{p}_2'^\top \\ \boldsymbol{p}_1'^\top - x'\boldsymbol{p}_3'^\top \end{bmatrix} \boldsymbol{X} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{A}\boldsymbol{X} = \mathbf{0}$$

The triangulation process can be formulated as a system of linear equations [18]. For each camera i, the projection of P onto the image plane is given by $p_i = K_i * E_i * P$, where $K_i \; and \; E_i$ are the intrinsic and extrinsic matrices of camera i, respectively. This equation can be rewritten as $p_i = A_i * P, where \; A_i = K_i * E_i$. The system of equations for all cameras can then be written in matrix form as $p = A * P$, where p is a vector of the stacked 2D points p_i, and A is a matrix of the stacked A_i.

This system of equations is typically solved using the Singular Value Decomposition (SVD) method. The solution P corresponds to the eigenvector corresponding to the smallest eigenvalue of the matrix $A^T * A$.

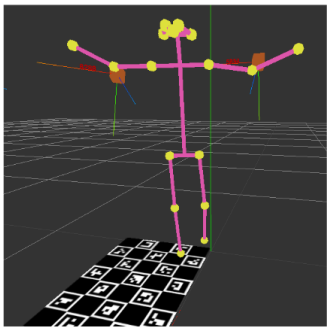

Fig. 2. 3D skeleton after triangulation of 2D joints.

When the triangulation process incorporates a confidence-based weighting mechanism, it enhances the contribution of high-confidence 2D points for improved triangulation results compared to a naive method. In our experiments, the initial triangulation was refined using the CERES solver [20] [19] (a non-linear least squares solver), but this refinement did not significantly reduce reprojection error. The iterative nature of CERES introduced processing time overhead, particularly for numerous frames. An example of triangulation result is shown in figure Fig. 2

An alternative strategy consists in selectively applying the CERES solver using a threshold logic, refining estimated 3D keypoints only when reprojection error exceeds a set threshold. This approach optimizes the accuracy for challenging scenarios while managing computational resources, similar to OpenPose's built-in multi-view triangulation method. The implementation of this technique demonstrates the significance of intelligent refinement strategies in achieving accurate 3D reconstructions.

E. Human Mesh Fitting process

The Human Mesh Fitting stage is crucial in our pipeline, adapting a parametric human body model to 3D keypoints from the previous triangulation.

We opted for the Skinned Multi-Person Linear (SMPL) model, which accurately represents the human form, aligning well with our project goals.

1) Skinned Multi-Person Linear (SMPL) model: The SMPL model is commonly used to represent the human body, combining realistic deformation and efficient computation. It employs a linear shape space and a linear pose space to depict a human body. SMPL's design includes blend skinning, where each mesh vertex is linked to joints using weights. These weights govern the influence of joints on vertex movement. The linear blend shape model captures shape variations via a low-dimensional linear space, enabling realistic deformations while remaining computationally efficient.

2) Mesh fitting using joints2SMPL: In our implementations, 3D keypoints were converted from JSON to NPY format to enable efficient numerical computations. These keypoints were then processed using the joints2SMPL [21] model to fit the SMPL human mesh with the observed 3D joint coordinates.

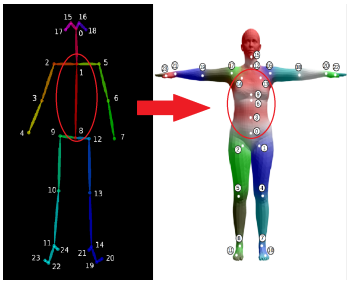

Fig. 3. Comparing the OpenPose-25 keypoints with SMPL 24 keypoints. As highlighted, the spinal and shoulder blade region keypoints cannot be mapped so they were interpolated based on nearby keypoints.

This process minimizes disparities between the SMPL model and observed keypoints and ensures smooth pose transitions. This approach can be adapted to different keypoint formats, such as AlphaPose COCO 17, OpenPose 25, and ZED body38. While ZED body38 adaptation is straightforward, COCO 17 and OpenPose 25 formats require linear interpolation for missing keypoints, particularly in the spine area (as shown in Fig. 3), although this can lead to unusual spinal depiction. The outcome includes PLY (Polygon File Format) models for each frame and a PKL file (Python Pickle File) containing the fitted SMPL parameters. The PLY model represents 3D data, while the PKL file serializes the SMPL parameters for later reconstruction.

3) Merging per frame meshes: The subsequent step consists in combining all PLY models into a single GLB (GL Transmission Format Binary) scene. GLB is a binary file format for 3D models saved in the GL Transmission Format (glTF), providing the advantage of encapsulating all necessary assets in a single file for easier distribution and management. Following this, the gltf-transform tool is utilized to convert the mesh sequence into an animated model.

F. Visualization

The visualization stage is pivotal in the pipeline, providing a visual representation of the 3D human mesh and keypoints generated earlier. This step consists in utilizing camera calibration data to determine accurate camera positions and rotations in the 3D space. The camera projection matrix is crucial for mapping 3D world points to 2D image points. Once cameras are correctly positioned, they are integrated into the 3D scene through translation and rotation, ensuring precise perspectives for observing the 3D keypoints and human mesh. This visualization step aids in comprehending and interpreting the pipeline's outcomes effectively.

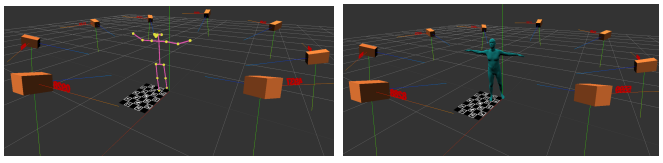

Fig. 4. 3D keypoints skeleton (left) and SMPL mesh after fitting (right) Visualization results for the complete pipeline.

In the subsequent steps, the 3D triangulated keypoints and the animated SMPL GLB file are incorporated into the 3D scene using a rendering library like Three.js. These keypoints denote joint positions, while the SMPL GLB file represents the human mesh. After loading, frames are synchronized by adjusting timing and alignment through frame synchronization algorithms. The final stage consists in visualizing the synchronized skeleton animation, resulting in a vivid 3D representation of human movement captured by the multi-camera setup as shown in Fig. 4.

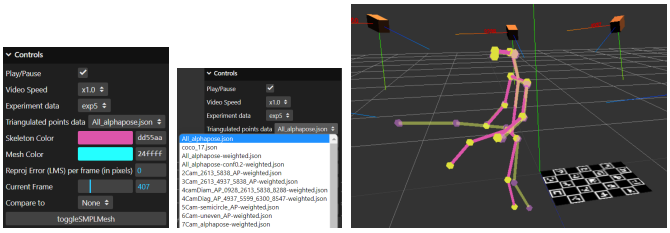

Fig. 5. From left to right, List of Menu options, Different datasets generated from same experiment, Visualization of dataset comparison option.

The web application, designed using ThreeJS, can offer users a seamless interface to navigate through experiments and datasets, including various camera setups like 8-cam, 6-cam, and 4-cam. Users can switch between skeleton and mesh views and compare datasets, as shown in Fig. 5. While mesh comparison is unavailable, it allows detailed examination. This aids in understanding differences due to camera variations. Valuable for motion analysis, 3D animations, and virtual reality development, this visualization tool improves results interpretation of 3D scenes captured from a multi-camera capture system.

IV. Results and Discussion

A. Impact of Cameras - calibration, number and placement

The initial goal of our experiments was to evaluate how the number, calibration, and positioning of cameras impact pose estimation and reconstruction quality. We began with an eight-camera 360-degree setup for comprehensive coverage, but real-world scenarios may not allow this. Thus, we aimed to assess effects of camera reduction and repositioning. Here are some important calibration-related observations to consider.

- The cameras should adequately cover the center of the scene where the ChArUco board is placed for common calibration, as this location is considered as the origin of the scene.
- The size of the board must be proportional to its distance from the camera. While ChArUco or OpenCV do not provide a specific metric for such calibration board, we can establish a correlation between the size of the board and its distance from any camera using the principles of projective geometry. This correlation is possible if we know the camera's focal length (in pixels) and the size of the board in the image (in pixels), as given by the equation:

$$distance = f \cdot (\frac{size_{real}}{size_{image}}) \tag{5}$$

- In our practical experiments, we discovered that an A0 size (841 x 1189 mm) ChArUco chart is effective for common calibration up to a maximum distance of 4 meters from the camera. The aforementioned equation provides a relative distance assuming the board is parallel to the focal plane.

The Table I presents the calibration precision for our procedure, grounded on the principle of reprojection error and the real-world length error, as derived from equations (2) to (4) discussed in preceding sections.

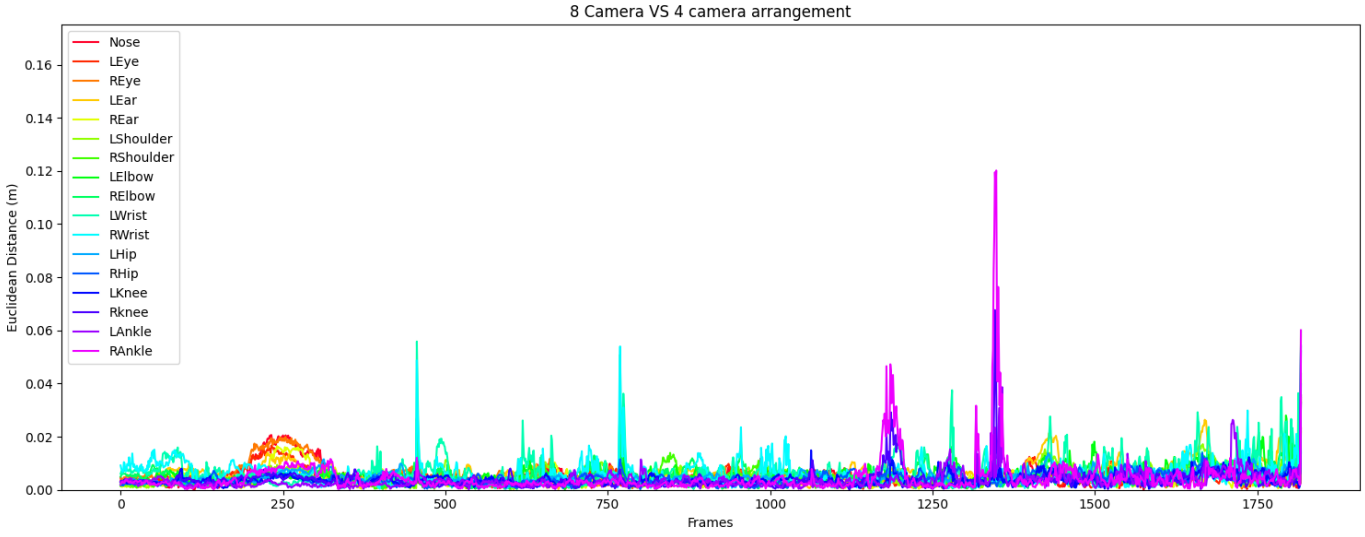

Fig. 6. Comparison of Euclidean distance for each keypoint, for 8 camera setup vs 4 camera. At each keypoint corresponds a false color.

These first results show length errors due to calibration inaccuracies. However, it's important to note that errors can

also emerge during triangulation and reconstruction, due to inaccuracies in pose estimation models and SMPL fitting. As illustration, data collection involved a predefined sequence of poses and movements, recording 60 seconds of video at 60 frames per second, totaling 3000 - 3500 frames. Videos were split into Easy, Medium, and Hard segments (shown in Fig. 7), with pose estimation models facing challenges for the latter two.

TABLE I
REPROJECTION ERROR, ANGULAR ERROR AND LENGTH ERROR FOR EACH OF THE 8 GOPRO HERO 10 CAMERAS USED FOR EXPERIMENTATION OVER A MAX DISTANCE COVERAGE OF 4 METERS.

Length Error over a distance of 4 meters			
GoPro Camera ID	Reprojection Error (pixels)	Angular Error (degrees)	Length Error (mm)
0928	2.8533	0.0485	3.391
2613	3.9920	0.0688	4.809
4937	4.6171	0.0811	5.664
5599	2.7551	0.0464	3.241
5838	3.6880	0.0627	4.378
6300	2.7853	0.0473	3.305
8288	3.3410	0.0578	4.041
8547	3.4729	0.0592	4.136

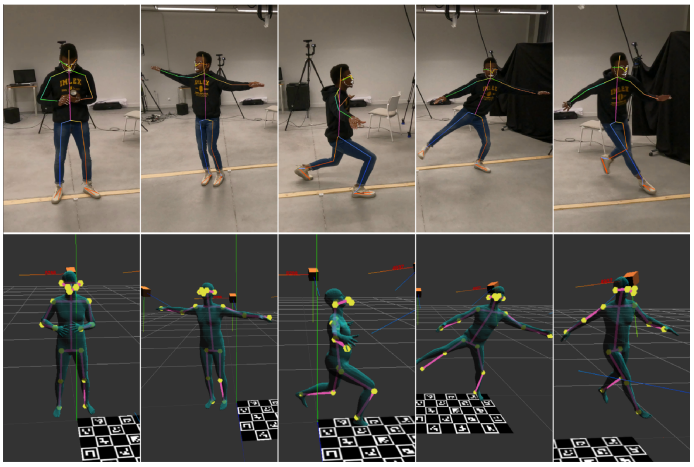

Fig. 7. Different poses from left: standing, T-pose, knee bend, leg raise, running (with motion blur). The reconstruction shows a semi-transparent mesh with triangulated skeleton showing the offset between 3D triangulation and SMPL mesh fitting.

The next section focuses on reprojection error analysis based on camera quantities and positions. The Table II presents reprojection errors for different camera numbers and placements, considering inherent camera-specific error and pose estimation model error. This analysis is averaged across keypoints and frames in various configurations.

The illustrations shown in Fig. 6 and Fig. 8) give a comprehensive analysis of the Euclidean distance, measured in meters, for the standard 17 COCO keypoints. These illustrations allow for a thorough comparison of the variation in triangulation accuracy across different camera setups. This analysis emphasizes the crucial role of the camera setup in

TABLE II
REPROJECTION ERROR, ANGULAR ERROR AND LENGTH ERROR FOR DIFFERENT GOPRO HERO 10 CAMERA SETUPS USED FOR EXPERIMENTATION.

RE and LE comparison for different camera arrangements			
Setup Info	Reprojection Error (pixels)	Angular Error (degrees)	Length Error (mm)
8 Cameras	34.53	0.5917	41.31
7 Cameras	39.15	0.6713	46.86
5 Cameras (semi-circular)	32.82	0.5622	39.24
4 Cameras (diagonal)	38.97	0.6684	46.66
3 Cameras	37.30	0.6444	44.98

TABLE III
REPROJECTION ERROR, ANGULAR ERROR AND LENGTH ERROR FOR 3D POSE USING ALPHAPOSE VS OPENPOSE ON THE SAME SETUP AND ACQUISITION.

RE and LE comparison for AlphaPose vs OpenPose			
Model	Reprojection Error (pixels)	Angular Error (degrees)	Length Error (mm)
AlphaPose	34.53	0.5917	41.31
OpenPose	58.61	1.0043	70.11

achieving reliable triangulation and pose estimation results. Such comparison enable to demonstrate the impact of different camera quantities and placements, rather than providing absolute accuracy assessment. Without available ground truth data, comprehensive accuracy evaluation is not possible. Yet, this comparison sheds light on how camera setup affects the overall pose estimation process.

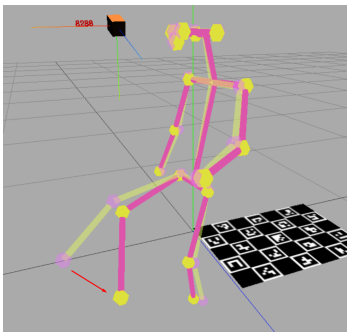

Fig. 8. 8 cameras VS 4 cameras: The peak observed around the frame 1300 is caused by the leg raising action. The 8 camera pose is illustrated with translucent green segments meanwhile the 4 camera pose is illustrated with pink segments.

B. Comparison of different pose estimation models

This section focuses on comparing the impact of pose estimation models on triangulation and 3D reconstruction. AlphaPose (top-down) and OpenPose (bottom-up) were analyzed using two methods: swapping models while keeping the pipeline consistent, and using OpenPose's multi-view triangulation. Reprojection and Length errors were compared

(refer Table III), revealing significant differences. OpenPose's bottom-up approach leads to missing occluded keypoints, affecting triangulation due to limited viewpoints (shown in Fig. 9). AlphaPose's top-down approach estimates keypoints within bounding boxes, enhancing accuracy. However, OpenPose's limitations extend to SMPL mesh fitting, requiring all 25 keypoints. An error in fitting can cascade, compromising pose accuracy and triangulation quality. This underscores the interdependency of pipeline stages and error impacts.

C. Comparison with ZED fusion

The study compares the ZED fusion method with the proposed approach, assessing their integration feasibility and mesh fitting results. An initial setup with 4 ZED2i cameras encountered frame drops, prompting a configuration change to 2 cameras at HD720 and 15 fps due to GPU limitations. A comparative experiment captured a scene with both methods, evaluating SMPL mesh fitting and comparing fitting loss (Fig. 10). Results indicate that the proposed approach maintains stable fitting loss across frames, while ZED achieves lower loss for about 3/5th of frames. ZED's results exhibit instability and occasional glitches, while GoPro Hero 10 data shows smoother transitions despite slightly less accurate fitting, highlighting a trade-off between fitting accuracy and animation stability.

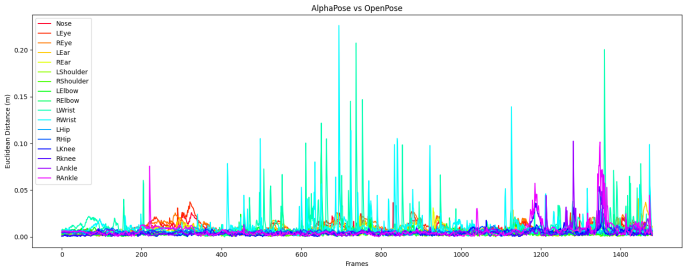

Fig. 9. Difference in Euclidean distance between AlphaPose and OpenPose 3D joints triangulation.

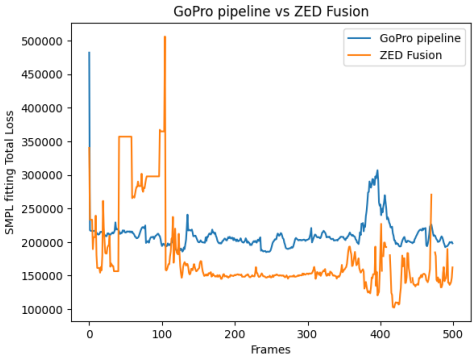

Fig. 10. SMPL fitting loss for GoPro pipeline vs ZED Fusion.

V. Limitations and Future Works

A. Calibration Accuracy

The calibration process is crucial for the pipeline's accuracy, relying on the ChArUco marker technique which can be time-consuming and marker-size dependent. While a marker chart of A0 size supports small setups, larger environments like stages pose challenges due to marker size limitations. To overcome this, alternative calibration methods like the ZED 360 auto calibration offer more efficient solutions. These methods enhance efficiency and scalability while maintaining accuracy, addressing limitations related to marker size and setup scale.

B. Pose Estimation Model

Although not direct limitations of the pipeline or tools, it is important to highlight some observations regarding the state-of-the-art (SOTA) pose estimation models and their accuracy in real-world scenarios. There are two main points to consider:

- **Ghost Poses:** Ghost poses can occur when background objects resemble human poses, caused by factors like lighting, occlusion, and more.
- **Challenges with horizontal or parallel poses:** When the subject is lying horizontally or parallel to the floor, the entire body may not be clearly visible as the subject is in a perspective that is perpendicular to the camera. Pose estimation models do not perform well in such scenarios.

C. Multiple Poses and Reconstruction

While our current pipeline lacks multi-pose detection and reconstruction capabilities, it can be easily incorporated in the future. The observed challenge lies in the reliability of pose IDs assigned by the models, especially in cases of occlusion or overlap. Robust pose tracking is essential for consistent identification of individuals, aiding accurate 2D point retention and reliable triangulation. Another challenge involves matching poses across different camera angles to prevent the mixing of data and maintain accurate triangulation outcomes. This highlights the importance of maintaining pose consistency across frames and angles for successful multi-pose handling.

D. Impact of Interpolated Points

Converting pose estimation keypoints to SMPL format involves challenges, requiring linear interpolation for specific points like the spine and shoulder blades. A comparison shown in Fig. 11 between ZED body38 SMPL fitting and AlphaPose COCO 17 keypoints, interpolated to 24, revealed ZED's natural spinal and upper body shapes but not universal superiority due to overall loss. Standardizing keypoint formats and corresponding model training could establish a consistent protocol for accurate reconstruction. Alternatively, exploring methods based on real spinal dynamics and vertebral structure might enhance missing keypoint accuracy beyond linear interpolation.

Our experimental pipeline has illuminated pose estimation intricacies and motion reconstruction. Despite limitations, our commitment to improvement remains strong. By addressing challenges and exploring innovative solutions, we can boost accuracy and efficiency, expanding applications in motion analysis and computer vision. Continuous refinement drives progress in this field.

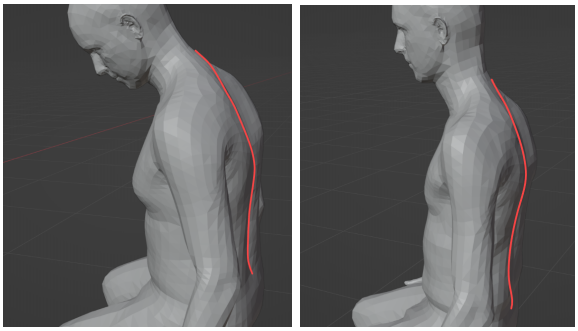

Fig. 11. Comparison of ZED fitting (left) and interpolation fitting on AlphaPose COCO-17 (right)

VI. Conclusion

The article extensively explored multi-camera systems for 3D reconstruction and live performance recording. It introduced a universal data collection approach, compared pose estimation methods and camera setups, and advanced multi-camera system development. Challenges persist, but the study anticipates pushing boundaries in human motion analysis and computer vision applications. Future opportunities include refining calibration, pose estimation, and supporting face and hand analysis for enhanced accuracy. In conclusion, the study aims to foster progress in the field and contribute to creating state-of-the-art multi-camera systems for high-quality performance reconstructions.

References

[1] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, 'OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 1, pp. 172–186, 2021.

[2] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, 'RMPE: Regional Multi-person Pose Estimation', arXiv preprint arXiv:1612. 00137, 2016.

[3] C. Musunuri, 'EP2272247A1 - Genlock synchronizing of remote video sources', 2008. [Online]. Available: https://patents.google.com/patent/EP2272247A1/en.

[4] N. J. Bryan, P. Smaragdis, and G. J. Mysore, 'Clustering and synchronizing multi-camera video via landmark cross-correlation', in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 2381–2384.

[5] F. J. Romero-Ramirez, R. Muñoz-Salinas, and R. Medina-Carnicer, 'Speeded up detection of squared fiducial markers', Image and Vision Computing, vol. 76, pp. 38–47, 2018.

[6] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, 'Convolutional Pose Machines', arXiv preprint arXiv:1602. 00134, 2016.

[7] R. Hartley and P. Sturm, 'Triangulation', Computer Vision and Image Understanding, vol. 68, no. 2, pp. 146–157, 1997.

[8] S. Henry and J. A. Christian, 'Absolute Triangulation Algorithms for Space Exploration', Journal of Guidance Control and Dynamics, vol. 46, no. 1, pp. 21–46, 2022.

[9] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, 'SMPL: A Skinned Multi-Person Linear Model', ACM Transactions on Graphics (TOG), vol. 34, no. 6, pp. 1–16, 2015.

[10] J. Romero, D. Tzionas, and M. J. Black, 'Embodied hands: Modeling and capturing hands and bodies together', ACM Transactions on Graphics (TOG), vol. 36, no. 6, pp. 1–17, 2017.

[11] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, 'End-to-end Recovery of Human Shape and Pose', in Computer Vision and Pattern Recognition (CVPR), 2018.

[12] G. Pavlakos et al., 'Expressive Body Capture: 3D Hands, Face, and Body from a Single Image', in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.

[13] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, 'Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3383–3393.

[14] J. Li, S. Bian, C. Xu, Z. Chen, L. Yang, and C. Lu, 'HybrIK-X: Hybrid Analytical-Neural Inverse Kinematics for Whole-body Mesh Recovery', arXiv preprint arXiv:2304. 05690, 2023.

[15] Y. Yuan, U. Iqbal, P. Molchanov, K. Kitani, and J. Kautz, 'GLAMR: Global Occlusion-Aware Human Mesh Recovery with Dynamic Cameras', in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[16] Gl, 'The Reprojection Error?' - https://www.camcalib.io/post/what-is-the-reprojection-error 9 2022.

[17] T.-Y. Lin et al., 'Microsoft COCO: Common Objects in Context', arXiv [cs.CV]. 2015.

[18] Kris Kitani, 'Triangulation'. 2017.

[19] S. Agarwal, N. Snavely, S. M. Seitz, and R. Szeliski, 'Bundle Adjustment in the Large', in Computer Vision – ECCV 2010, 2010, pp. 29–42.

[20] K. Madsen, H. Nielsen, and O. Tingleff, 'Methods for Non-Linear Least Squares Problems (2nd ed.)', p. 60, 01 2004.

[21] X. Zuo et al., 'Sparsefusion: Dynamic human avatar modeling from sparse rgbd images', IEEE Transactions on Multimedia, vol. 23, pp. 1617–1629, 2021.

[22] 'Premiere - Performing arts in a new era', 18-Oct-2022. [Online]. Available: https://premiere-project.eu.

[23] H.-S. Fang et al., 'AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time', arXiv [cs.CV]. 2022.