Development of Multi-Camera System for Evaluating Reprojection Error in 3D Marker Estimation on Simple Motion Capture System

Moses Ananta
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
moses.ananta@gmail.com

Nugraha Priya Utama
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
utama@staff.stei.itb.ac.id

Syihabuddin Yahya Muhammad
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
syihabuddiny@gmail.com

Wisnu Aditya Samiadji
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Cirebon, Indonesia
wisnuaditya14@gmail.com

Jose Galbraith Hasintongan
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bekasi, Indonesia
galbraith.jose02@gmail.com

***Abstract*— In response to the advancements in artificial intelligence (AI) technology, motion capture systems have become more accessible to small content creation industries due to reduced production costs and complexity. However, AI-based motion capture systems that rely on estimating motion without conventional markers may encounter accuracy issues. To address this challenge, we propose a low-cost and straightforward alternative, employing color-based markers and a multi-camera setup to capture actor movements. The primary objective of this study is to develop a reliable multi-camera system capable of estimating 3D coordinates using triangulation techniques. The system's performance is assessed through the reprojection error metric, which serves as an indicator of the accuracy of 3D estimation. Extensive testing involving four distinct camera rotation ranges reveals that reprojection error can be effectively employed as a measure of the reliability of 3D estimation results. Furthermore, our analysis demonstrates that camera rotation has minimal impact on reprojection error, while utilizing three synchronized cameras for triangulation yields the most optimal results with the lowest reprojection error and the best 3D estimation outcomes.**

Keywords*— Motion capture, Multi-camera system, Triangulation, Reprojection error, 3D estimation.

I. INTRODUCTION (*Heading 1*)

The progress in artificial intelligence (AI) technology has significantly enhanced the practicality and affordability of motion capture systems. However, the accuracy of motion capture results obtained from AI-based systems remains questionable. On the other hand, conventional motion capture systems offer precise and clear motion capture results but entail high costs due to the required materials, equipment, and skilled technicians, making them generally inaccessible for small industries. Despite their superior accuracy compared to AI-based systems, the high costs associated with conventional systems prohibit their replication and performance comparison in this study. Thus, to facilitate a fair comparison with AI-based motion capture systems, a non-AI-based motion capture system that can approximate the performance of conventional systems without relying on expensive equipment and materials is essential. This study focuses on discussing the methodology of estimating the 3D positions of markers on actors' bodies using a multi-camera system. However, estimating 3D positions from 2D images is not a straightforward task as valuable 3D information is lost during the image capture process.

II. METHODOLOGY

A. Image Formation in Cameras

To map a 3D point in the world onto a 2D point in a camera, transformations [1] are employed.

$$\boldsymbol{x} = K[R|t]\,\boldsymbol{X} \tag{1}$$

These transformations involve extrinsic matrices consisting of 3x3 rotation matrix and 3x1 translation matrix, which map 3D points from the world coordinate system to 3D points in the camera coordinate system, and intrinsic matrices K, which map 3D points from the camera coordinate system to 2D points in the image coordinate system. Given the 2D points in the image coordinate system and the values of each camera's intrinsic and extrinsic matrices, it is possible to perform the inverse projection, allowing the 2D points in the image coordinate system to be reconstructed as 3D points in the world coordinate system. However, a challenge arises as information regarding the intrinsic and extrinsic matrices of the camera is often unknown or not readily available in the cameras used for this study.

B. Camera Calibration (Single Camera)

Camera calibration, also referred to as single camera calibration in this study, involves the process of determining the intrinsic and extrinsic matrices. The calibration method utilized in this research is the Zhang Calibration Method [2]. This method allows for the estimation of the camera's intrinsic matrix and distortion parameters using multiple calibration pattern images, such as a known-sized chessboard with known numbers of rows and columns and the size of each square on the chessboard pattern. The distortion parameters represent image distortions caused by the camera lens and are crucial for distortion correction. Eliminating distortion is essential to enhance the accuracy of subsequent processes.

However, the Zhang Calibration Method has a limitation; it does not provide the extrinsic matrix, which maps points from the world coordinate system to the camera coordinate system, hindering the back-projection of 2D image coordinates to 3D world coordinates. Although it is possible to perform the back-projection from the 2D image coordinate system to the 3D camera coordinate system using only the intrinsic matrix. The back-projection of 2D image coordinates to 3D camera coordinates from a single viewpoint may result in unclear 3D projections, particularly concerning the depth information or the Z-axis value of a 3D point. This ambiguity is due to an infinite number of 3D points satisfying the mapping equation from a 3D point to its 2D image point. However, employing more than one view that observes the same 3D point yields at least one 3D point that satisfies the image mapping equation from each view.

C. Triangulation for 3D Point Reconstruction

One of the methods to determine the 3D point using multiple views is triangulation [3]. Triangulation essentially utilizes fundamental linear algebra concepts, seeking the intersection points of two or more lines, where the intersection points in this context represent the sought-after 3D points. The projection function is the result of multiplying the intrinsic and extrinsic matrices.

$$\mathbf{P} = K[R|t] \tag{2}$$

As at least two non-parallel lines are required to find the intersection point, triangulation necessitates a minimum of two views observing the same 3D point. By combining the projection matrices of each camera and the corresponding 2D points of the observed 3D point from each camera, a matrix A is constructed.

$$\boldsymbol{A} = \begin{bmatrix} u_1\,\mathbf{P}^3 - \mathbf{P}^1 \\ v_1\mathbf{P}^3 - \mathbf{P}^2 \\ u_2\,\mathbf{P}^3 - \mathbf{P}^1 \\ v_2\mathbf{P}^3 - \mathbf{P}^2 \end{bmatrix} \tag{3}$$

The solution of Equation II.22 will yield the 3D intersection point of the projection lines from each camera.

$$\boldsymbol{AX} = 0 \tag{1}$$

In real-world scenarios, obtaining the 3D intersection point from the individual camera back-projections is challenging since the back-projection lines of each camera do not typically intersect due to noise introduced from the observed 2D points or the camera projection matrices. Therefore, the problem of finding the 3D intersection points from the camera projection lines is transformed into a problem of searching for the best 3D point that can represent the intersection of the projection lines. To address this issue, the Singular Value Decomposition (SVD) method is commonly applied, performing a decomposition of matrix A to obtain the best-fitting 3D point that approximates the intersection of the camera projection lines.

D. Stereo Calibration for Extrinsics Estimation

One unresolved issue in the previous stages of the process arises from the fact that the projection matrix utilized in triangulation is derived from the multiplication of the intrinsic and extrinsic matrices of each camera. The problem lies in assuming that the extrinsic matrix used in the triangulation process transforms the camera coordinate system to the world coordinate system, while this extrinsic matrix has not been obtained up to this point. An alternative solution is to perform triangulation operations in the context of one camera's coordinate system, rather than transforming the coordinate systems of each camera to the world coordinate system. This approach involves finding a transformation matrix that maps the coordinate system of one camera to the coordinate system of the target camera. This search for the transformation matrix is more feasible compared to obtaining the extrinsic matrices of each camera and is known as stereo calibration.

Similar to the extrinsic matrices that perform rotation and translation from the world coordinate system to the camera coordinate system, stereo calibration seeks a matrix that performs rotation and translation, transforming the coordinate system of one camera to the coordinate system of another camera. Utilizing the stereo calibration technique described in [4], which involves using a calibration pattern such as a chessboard pattern visible to both cameras simultaneously, the required rotation and translation matrices can be obtained.

E. Multi-View Triangulation for 3D Point Estimation

At this stage, the intrinsic matrices and the rotation-translation matrices from one camera to another have been acquired. The information obtained thus far is sufficient to proceed with 3D point estimation using triangulation techniques.

When considering the structure of the 3D point estimations obtained through triangulation, employing more than two camera views can significantly improve the quality of the resulting 3D point estimates [5]. The use of multiple views enhances the robustness of the 3D point estimation process against noise present in the utilized parameters, leading to reduced errors in the estimation. Using multiple views, matrix A from equation (3), that is used for triangulation, can be expanded into following form.

$$\boldsymbol{A} = \begin{bmatrix} u_1\,\mathbf{P}^3 - \mathbf{P}^1 \\ v_1\mathbf{P}^3 - \mathbf{P}^2 \\ u_2\,\mathbf{P}^3 - \mathbf{P}^1 \\ v_2\mathbf{P}^3 - \mathbf{P}^2 \\ \vdots \\ u_i\,\mathbf{P}^3 - \mathbf{P}^1 \\ v_i\mathbf{P}^3 - \mathbf{P}^2 \end{bmatrix} \tag{2}$$

F. *Bundle Adjustment for 3D Point Optimization*

The final stage, which is optional but considered a crucial aspect and a standard practice in 3D point estimation, involves the optimization of results obtained from the triangulation process. To achieve this, a commonly used optimization method called bundle adjustment [1], [6] is applied. The main objective of this optimization is to minimize errors in the back-projection of 3D points to their corresponding 2D points from each camera used during the triangulation process. By applying the bundle adjustment process, the reprojection errors are minimized, resulting in more accurate 3D point estimation.

III. EXPERIMENT

A. *Camera Calibration and Distortion Correction*

In this experiment, the evaluation of both the camera calibration and distortion correction processes is. The camera calibration process begins with capturing calibration pattern images for each camera. The number of calibration pattern images, using a chessboard pattern, ranges from 10 to 20 images as per Zhang's guidelines. After detecting the edges within the calibration patterns, the 2D detection points and their corresponding 3D positions in arbitrary coordinate system, are fed into the algorithm, resulting in the intrinsic camera matrices and distortion parameters.

One consistent finding during image capture is that the initial images, taken before applying distortion correction, do not exhibit noticeable distortion. This observation leads to the assumption that the inherent distortion in the cameras, if present, is not perceivable to the naked eye. This can be attributed to the use of lenses with relatively minor curvature, resulting in insignificant distortion. Consequently, a new assumption is drawn that the distortion correction process should not substantially alter the original images due to the small magnitude of initial distortion.

However, in the early stages of the experiment, this assumption was not met. The distortion correction process not only failed to eliminate minor distortion but also introduced significant distortion, as illustrated in Fig. 1 below.

Fig. 1. Initial distortion correction result. (a) Original image before distortion correction. (b) Image after distortion correction.

After conducting several iterations of the experiment, two factors contributing to this issue were identified:

- The calibration pattern images were not uniformly distributed across the camera's field of view, primarily concentrated in the center. This caused the calibration algorithm to model distortion parameters primarily in the central region, neglecting other areas.
- The calibration algorithm utilized a low degree of distortion coefficients. According to OpenCV documentation, the default degree of distortion coefficients is set to 3 out of a total of 6 available degrees. During the experiments, it was evident that the use of only three degrees was inadequate in accurately modeling the distortion.

Upon correcting these errors by capturing calibration pattern images uniformly across the camera's view and increasing the distortion coefficient degree from 3 to 6, the distortion correction produced results similar to the original images, as shown in Fig. 2 below.

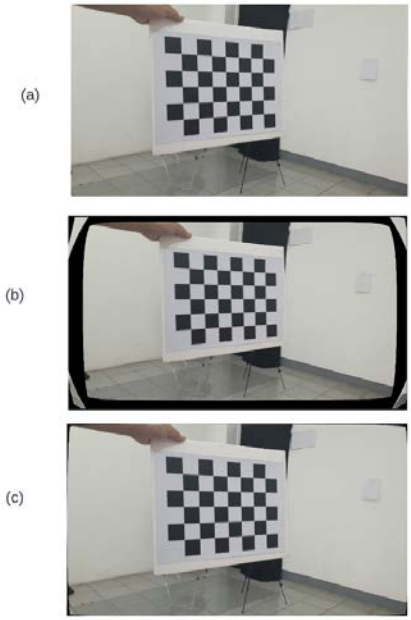

Fig. 2. (a) Original image before distortion correction. (b) Image after distortion correction with remaining errors. (c) Image after distortion correction with significant improvements. The black region in (c) indicates a smaller distortion correction area compared to (b), resulting in (c) being more similar to (a) than (b).

The calibration matrices and distortion parameters for each camera can be found in Table I and below.

TABLE I. SINGLE CAMERA CALIBRATION RESULT

Camera	**Intrinsic Matrix**			**Distortion Parameter** [a]
Left Camera	1436.854 0 0	0 1437.605 0	951.71 548.9 1	6.37 −46.576 0.0022 −0.0011 138.793 6.389 −47.004 139.1384
Center Camera	1473.503 0 0	0 1471.862 0	950.21 545.757 1	−2.194 79.834 0.0003 0.0013 −115.973 −2.485 8 −116.852

Camera	Intrinsic Matrix			Distortion Parameter [a]
Right Camera	1346.381	0	983.342	−17.635 −26.373 0.0036 0.0008 1353.113 −17.941 −18.649 1301.006
	0	1346.383	530.83	
	0	0	1	

[a.] Arranged in the following order $k_1, k_2, p_1, p_2, k_3, k_4, k_5, k_6$. Where k is the radial distortion coefficient and p is the tangential distortion coefficient.

B. Stereo Calibration for Multi-Camera Setup

In this experiment, the process involves capturing calibration pattern images for the three cameras simultaneously, if feasible. Alternatively, the images are taken for one camera pair first and then for the other pair if simultaneous capture is not possible. This is due to the dependency on camera positions and rotations, which might hinder all three cameras from capturing calibration patterns concurrently.

We conducted tests for four rotation ranges between the cameras: 0°-10°, 25°-35°, 40°-50°, and 51°-70°. The four rotation ranges will be used on the 3D estimation experiment. The obtained rotation and translation results from the stereo calibration process for each test scenario are presented in Table II.

TABLE II. STEREO CAMERA CALIBRATION RESULT

Test	Camera	Axis	Camera Rotation (degree)	Camera Translation (cm)
Rotation Test 1 (0°-10°)	Left	x	2.350	−8.004
		y	−5.857	71.802
		z	3.035	10.348
	Right	x	−2.046	−8.004
		y	3.396	71.802
		z	−0.184	10.348
Rotation Test 2 (25°-35°)	Left	x	0.138	142.416
		y	−25.507	2.622
		z	−3.863	24.636
	Right	x	−4.4	−8.004
		y	28.068	71.802
		z	2.922	10.348
Rotation Test 3 (40°-50°)	Left	x	−5.107	130.589
		y	−43.253	−2.923
		z	−7.118	33.548
	Right	x	−5.13	−152.713
		y	46.452	18.2
		z	5.7732	67.064
Rotation Test 4 (51°-70°)	Left	x	2.375	234.541
		y	−54.465	4.299
		z	−8.408	108.937
	Right	x	−8.004	−170.365
		y	71.802	−16.114
		z	10.348	76.514

The rotations and translations of the cameras relative to the main camera, i.e., the center camera.

C. Evaluation of 3D Estimation Results: Triangulation and Bundle Adjustment

In this experiment, we compare the performance of 3D estimation results obtained through triangulation and those optimized with bundle adjustment. To achieve the best triangulation results, we conducted tests for four rotation ranges between the cameras: 0°-10°, 25°-35°, 40°-50°, and 51°-70°. For each test range, we identified the test number and the camera pair that resulted in the lowest reprojection error. Data on the 2D marker locations during the actor's upright standing pose were provided for each test, as shown in Fig. 3.

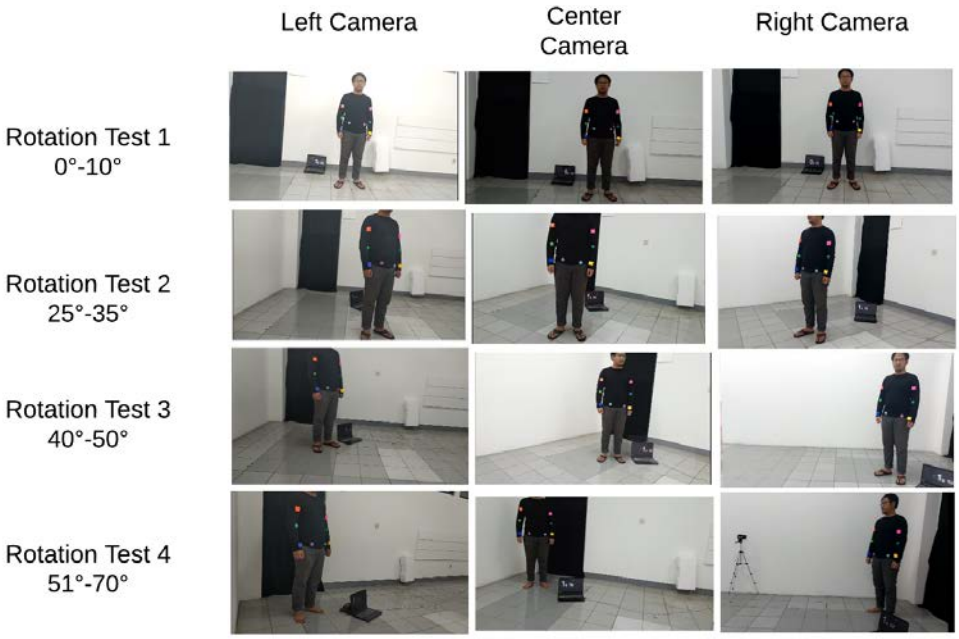

Fig. 3. Different views and rotations of the actor.

The reprojection error results for each test are presented in Table III.

TABLE III. 3D ESTIMATION RESULT

Test	**Triangulation Camera Pair**	**Reprojection Error (pixel)**	
		Triangulation	***Bundle Adjustment***
Rotation Test 1 (0°-10°)	left-center	2.634	0.563
	center-right	3.176	0.855
	left-right	2.845	0.487
	all camera	2.516	0.557
Rotation Test 2 (25°-35°)	left-center	2.423	0.831
	center-right	3.789	1.757
	left-right	2.316	0.811
	all camera	2.299	0.719
Rotation Test 3 (40°-50°)	left-center	2.84	0.757
	center-right	2.8	0.626
	left-right	2.124	0.661
	all camera	2.114	0.69
Rotation Test 4 (51°-70°)	left-center	2.076	0.839
	center-right	2.878	0.715
	left-right	1.9	1.072
	all camera	1.866	0.777

Based on the findings from Table III, the following observations were made:

- There is no significant impact of camera rotation changes on the reprojection error. This confirms the accuracy of the calibration process for both intrinsic and extrinsic camera matrices as well as the usage of 2D data.
- The application of bundle adjustment significantly influences the reduction of mean reprojection error. This confirms the successful and beneficial application of bundle adjustment for optimizing 3D estimation.
- Prior to bundle adjustment, triangulation using all three cameras simultaneously resulted in the lowest mean reprojection error. This suggests that the use of more than two cameras or multi-camera setups can reduce reprojection error.
- After bundle adjustment, triangulation using all three cameras simultaneously did not always result in the lowest mean reprojection error. This could be due to the better initial estimates for the triangulation pairs, making it easier for the optimization algorithm to find the optimal solution. The optimization algorithm relies on derivative principles and is sensitive to the quality of initial estimates, leading to varied results.

Further evaluations of the 3D quality are conducted, focusing on the overall 3D structure rather than the reprojection error, which is qualitative rather than quantitative. Three motion tests—pose-A (raising and lowering the actor beside the body), left-right movement, and forward-backward movement—are performed to assess the system's ability to estimate simple to complex movements. These tests are conducted for the predetermined four test ranges.

Key findings from the experiments are as follows:

- Overall, most movements are accurately captured. However, some 3D structure results show discrepancies or inaccuracies in depth estimation. For instance, Fig. 4 shows a "floating" 3D structure, indicating depth estimation errors. The analysis revealed that asynchrony between the camera frames caused erroneous pairings of 2D points, leading to depth estimation errors. This demonstrates that the use of three cameras does not always guarantee the best results, as each camera might detect different points.

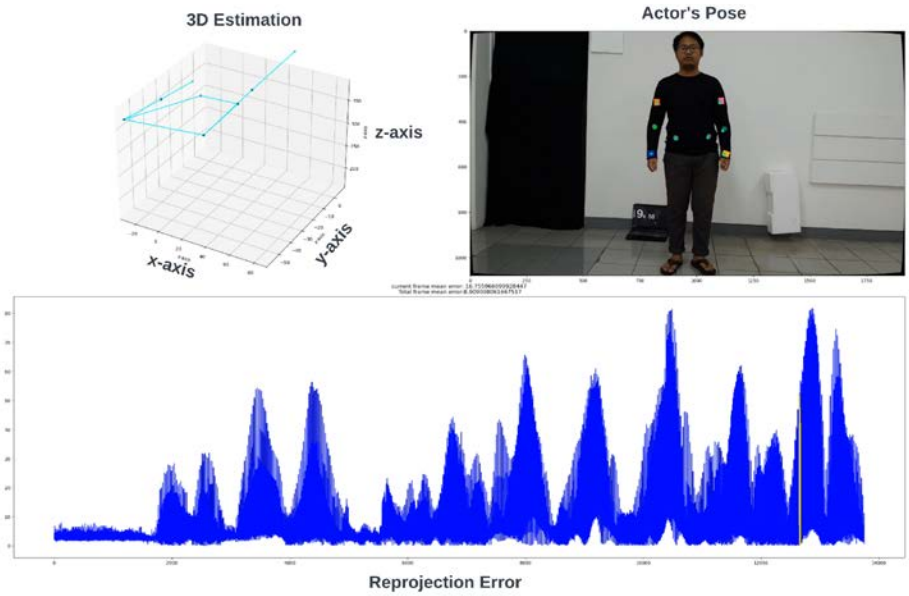

Fig. 4. Visualization of left-right movement. The 3D structure appears to "float" upwards.Top left panel, image showing the 3D estimation result. The x-axis represents the right-left direction. The y-axis represents the up-down direction. The z-axis represents the front-back direction. Top right panel, actor's pose image. Bottom panel, visualization of the reprojection error for each frame, with the parts highlighted in yellow indicating the reprojection error for the current frame.

- Reprojection error serves as an indication of inaccurate 3D estimation. As shown in Fig. 4, when the 3D results "float," the reprojection error increases, as indicated by the yellow-colored areas in the bottom panel. This correlation between reprojection error and 3D results highlights the significance of reprojection error as an indicator of the quality of the obtained 3D structure, where incorrect 3D estimations fail to align with their corresponding 2D data points.

IV. Conclusion

Key findings from this study highlight the significance of reprojection error as a reliable metric for assessing 3D estimation quality and identifying inaccuracies. Continuous monitoring and improvement of reprojection errors are essential to ensure system reliability. Surprisingly, camera rotations have minimal impact on reprojection error, indicating the system's robustness under varying angles and positions. Additionally, using three cameras simultaneously for 3D estimation resulted in the lowest reprojection error and most accurate results, emphasizing the advantages of multi-camera setups, especially when precision is crucial. Synchronization of image capture among cameras also plays a critical role in achieving optimal outcomes.

References

[1] R. Hartley and A. Zisserman, Multiple view geometry in computer vision. Cambridge University Press, 2003.

[2] Z. Zhang, "A flexible new technique for camera calibration," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 11, pp. 1330-1334, Nov. 2000, doi: 10.1109/34.888718.

[3] R. Hartley and P. Sturm, “Triangulation,” *Computer Vision and Image Understanding*, vol. 68, no. 2, pp. 146–157, Nov. 1997, doi: 10.1006/cviu.1997.0547.

[4] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. “O’Reilly Media, Inc.,” 2008.

[5] N. Snavely and Z. Li, “Multi-View Stereo,” *www.cs.cornell.edu*. https://www.cs.cornell.edu/courses/cs5670/2018sp/lectures/lec16_mvs.pdf

[6] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, ‘Bundle Adjustment - A Modern Synthesis’, in Workshop on Vision Algorithms, 1999.