

Markerless Facial Reconstruction Motion Capture Using Triangulation Method

Muhammad Alwali

Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
602241013@student.its.ac.id

Sevito Fernanda Pambudi

Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
602232011@student.its.ac.id

Laras Suciningtyas

Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
602241014@student.its.ac.id

Eko Mulyanto Yuniarno

Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ekomulyanto@ee.its.ac.id

Abstract—Motion capture is a popular research topic, with one of its main applications being human face reconstruction. The demand for converting 2D images into 3D reconstructions continues to increase, especially in facial reconstruction, where progress is made in improving the accuracy of facial position prediction. However, there is still a significant gap in developing facial reconstruction technologies that can consistently convert 2D to 3D data with high accuracy, especially in scenarios involving dynamic facial expressions, diverse facial angles, and complex environmental conditions. Therefore, an approach using the triangulation method for 3D face reconstruction in the real world was developed. In the experiments, two cameras were used to obtain two face landmark coordinates so that the triangulation method can be implemented for 3D face reconstruction. This research aims to develop a motion capture approach that is able to accurately and efficiently transform 2D data into 3D face models without the need for complex hardware. The main contribution of this research is the development of a machine learning-based markerless motion capture technique designed to improve the accuracy of face position prediction in 3D face reconstruction from 2D data in realistic environments. This method seeks to bridge the current technology gap by providing a more flexible and reliable solution, expanding the potential applications of motion capture in various fields without dependence on specialized hardware. The results of face reconstruction research using markerless motion capture and triangulation method show RMSE values of 3.560839 for eyes, 1.644749 for nose, and 4.054638 for lips.

Keywords—3D face, motion capture, reconstruction, markerless, triangulation method

I. INTRODUCTION

Motion capture is a popular topic in research, one of which is implemented for human face reconstruction[1]. In recent years, machine learning has grown rapidly with various applications in text, image, and video processing. The demand for generating reconstructions from 2D to 3D dimensions continues to increase, especially for face reconstruction, which is currently progressing in the accuracy of face position prediction. This is due to the challenges arising from face shape variations, different poses, background complexity, to expression detection in multi-person situations[2]. With these various needs, marker-based approaches become inadequate,

so machine learning-based methods are chosen to develop reconstruction technology [3].

Motion capture is an interesting method in the creation of motion for computer animation. This technique relies on recording and sampling the motion of humans, animals, or static objects as data in a three-dimensional format. Today, the use of motion capture for motion detection is growing very rapidly[4]. This technology has several advantages, especially in the production of computer animation and visual effects. Motion capture enables very accurate motion capture. Compared to manual animation, motion capture is much faster because it can directly record complex movements, thus saving time[5]. This technique is also effective for replicating complicated movements that are difficult or time-consuming, resulting in consistent animation from the recording[6]. With all these advantages, motion capture is becoming a popular choice for producing realistic and efficient animations, especially in projects that require high detail in facial reconstruction.

However, in its application, motion capture faces several obstacles that can affect the accuracy of facial position prediction, efficiency, and final results. Motion capture requires sophisticated and expensive hardware and software. Sensors, cameras, and data processing software need to be of high quality to produce accurate results, which is why motion capture is so important resulting in high production costs[7]. Less-than-ideal lighting conditions or busy backgrounds can interfere with motion capture in both marker-based and non-marker-based systems[8]. Face shifting or motion instability can also be a challenge during the recording process[9]. The data generated from motion capture usually requires a cleaning and error correction process, which can be time-consuming and requires additional technical expertise[7]. These constraints pose a challenge in the use of motion capture for facial reconstruction.

Several research have contributed to advancements in face reconstruction techniques. Boyang et al. proposed a face alignment algorithm to enhance facial landmark accuracy. Their research introduced Weight Reconstruct Alignment with an optimized loss function for 3D face reconstruction, achieving higher accuracy in facial landmark[10]. Weilong et

al. developed an unsupervised 3D face reconstruction method using multiple images in open environments. This research introduced consistency loss, incorporating a novel approach called Inconsistency Elimination to enhance geometric consistency, significantly reducing inconsistencies in 3D facial geometry construction[11]. Cahyo et al. explored the detection and tracking of facial features across various human expressions and poses, accounting for orientation and facial movement. Their method employed Active Appearance Models requiring images with depth information for 3D representation, captured using a stereo camera[12]. Jaeik et al. developed a 3D face reconstruction technique capable of reconstructing a 3D face from a single side-view image, with experimental results demonstrating the method's accuracy when only side-view images are available[7]. Yihao et al. conducted research focused on efficiently generating realistic 3D facial animations, introducing a facial motion capture system that improves both the efficiency and realism of 3D face reconstruction. Their method has applications in 3D animation, gaming, virtual movies, and television[13]. Shi et al. applied facial motion capture techniques in the rehabilitation training of deaf children, aiming to identify optimal facial marker feature points for language rehabilitation. Their research also defined suitable motion capture devices and parameters, showing that correctly placed facial markers significantly enhance training effectiveness[14]. Lastly, Nasser et al. developed an algorithm to construct accurate 3D face models using distance data from 3D scanners. Their approach efficiently generates detailed, smooth 3D face models using only three primary facial feature points, providing an efficient alternative to methods that require numerous reference points[15].

Based on previous literature reviews, various studies related to face reconstruction have focused on improving face position prediction accuracy and efficiency in reconstructing 3D faces from 2D images. The proposed approaches include face alignment algorithms for accurate facial position prediction of facial landmarks, unsupervised methods for geometric consistency, and techniques for tracking and detecting facial features in various expressions and poses. Most studies implement machine learning models and motion capture-based techniques, such as Active Appearance Models[12], and facial motion capture systems[14], to produce more realistic and efficient facial representations. However, there is a significant gap in the development of face reconstruction technologies that can consistently convert 2D to 3D data with high face position prediction accuracy, especially in situations involving dynamic facial expressions, face positions from various angles, and complex environmental conditions. Therefore, further development is still needed in the integration of machine learning techniques and motion capture systems to bring more reliable solutions in 3D face reconstruction in real environments.

Based on previous research, motion capture has been widely applied in various applications, such as body, hand,

and face motion capture. Most of the previous research relied on marker-based motion capture or techniques that require depth information to generate 3D face models. This approach has limitations due to its dependence on specialized hardware and is not always able to handle pose, expression, and lighting variations flexibly. These limitations indicate an urgent need for more adaptive face reconstruction methods that do not depend on specialized markers or cameras.

In the previous research conducted by Feng et. al, a 3D face reconstruction method was obtained from 2D images using a convolutional neural network-based approach and non-linear regression [26]. The dataset used consists of 2000 2D face images from 135 subjects, equipped with 3D face scans as ground truth. Evaluation is done by comparing the reconstruction results against the ground truth using the RMSE (Root Mean Square Error) metric. The results show that this method is able to produce accurate 3D face reconstruction with an average RMSE value of 1.58 mm, showing competitive performance compared to other methods in 3D face reconstruction evaluation.

There is a study conducted by Zhao et. al proposed the Deep Fusion Multi-View Reconstruction (DF-MVR) model to reconstruct 3D faces from weakly supervised multi-viewpoint 2D images [27]. This method combines features from multiple viewpoints using deep fusion to improve reconstruction accuracy. The datasets used in the evaluation include Pixel-Face and Bosphorus. The results show that this method successfully improves the reconstruction accuracy with a decrease in RMSE value of 5.2% on Pixel-Face and 3.0% on Bosphorus compared to other weak supervision methods.

Therefore, this research aims to develop a motion capture approach that is capable of converting 2D data into 3D faces with high accuracy of face position prediction and efficiency without complex hardware. This approach is expected to overcome the challenges of detecting facial expressions in real-time, even under complex conditions such as pose variations, face shapes, viewing angles, less-than-ideal lighting, and multi-person situations.

The main contribution of this research is the development of a markerless motion capture method based on machine learning, designed to improve the accuracy of facial position prediction for 3D facial reconstruction from 2D data in more realistic environments. This method is expected to address the gap in motion capture technology by offering a more flexible and reliable solution, expanding the potential applications of motion capture across various contexts without dependence on specialized hardware.

II. METHODS

This section will describe the methods used in this research. The stages used in this research can be seen in the block diagram in Fig. 1.

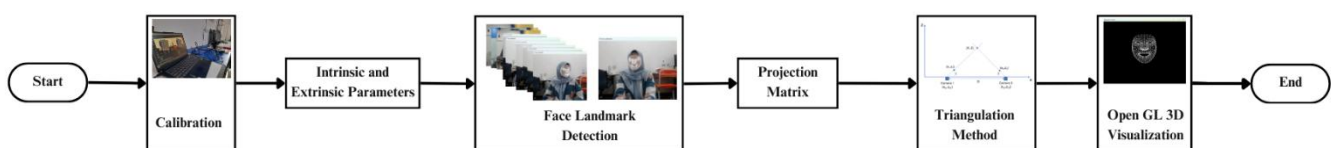


Fig. 1. Research Workflow

A. Camera Calibration

Camera calibration is the process of estimating the parameters of the camera used to take photos. In existing camera calibration methods, the coordinates of the main point are generally treated as the center of distortion. Camera calibration is essential in facial expression detection using motion capture to ensure accurate projection of facial features and minimize image distortion[16]. In camera calibration using multiview cameras, if n cameras are used, then the projection equation for each camera i is.

$$Si \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = K_i [R_i | t_i] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (1)$$

Equation (2) is K_i the intrinsic matrix of each camera (containing the internal parameters of the camera), R_i is the rotation matrix indicating the orientation of the camera, t_i is the translation vector (relative position of the camera), (X, Y, Z) are points in 3D coordinates.

Camera calibration in multiview or stereo systems ensures alignment of images from different viewpoints, enabling accurate 3D position calculation. Recent research emphasizes that by utilizing the intrinsic and extrinsic parameters of the camera, the accuracy of spatial measurements can be significantly improved, making it very important for high-fidelity applications such as facial expression detection in motion capture[17]. In this research, 110 checkerboard were calibrated.

Some related research such as that conducted by Vincent, et al. used camera calibration and focused on improving the accuracy of 3D systems using various methods. Although it provides high face position prediction accuracy, it requires more complicated hardware and increased processing time[18].

Another research conducted by Bisht, et al with the use of markerless motion capture technology to reconstruct 3D poses in a multiview environment. The advantage of this approach is its flexibility, but the challenge lies in calibration accuracy which can be affected by factors such as occlusion and interaction between subjects[19].

B. Face Landmark Estimation using MediaPipe

MediaPipe is a framework developed by Google for various purposes, such as face detection, landmark estimation on the face, hands, and pose, and gesture recognition. Face landmark extraction using MediaPipe is one of the leading solutions for facial expression analysis, utilizing deep learning models to detect important facial points in images or videos. MediaPipe provides 3D face detection with 468 landmark points, allowing for the analysis of facial geometry in various positions and orientations[20]. The process starts with face detection that identifies the presence of faces in the image, and then proceeds with the prediction of facial landmarks that refer to 3D coordinates to accurately depict facial details, even on faces that are distorted due to rotation or position changes[21]. The landmarks obtained from MediaPipe are coordinate points on the X and Y axes of the face.

Related research conducted by Narendra, et al using MediaPipe shows effectiveness in identifying face landmarks in real-time with low latency. MediaPipe excels on low-power

devices, but can struggle in poor lighting conditions or extreme face orientation[22]. Another research conducted by Cha Zhang, et al with a deep learning CNN-based model used for multiview face detection successfully improved accuracy in variable viewing angle conditions. However, this model is more computationally intensive, making it less suitable for devices with limited power[23].

C. 3D Face Pose Reconstruction Estimation

The process of 3D face pose reconstruction estimation is carried out by projecting 2D point data from both multi-view cameras. The approach used for the projection is triangulation. The result of triangulation provides depth information for each landmark point, allowing for the generation of 3D coordinates to create the 3D face pose.

The purpose of the projection is to obtain the world coordinates X, Y , and Z for each point, which is a combination of the landmark coordinates obtained from both cameras. The 3D projection is performed on each camera individually. Therefore, we need matrices K_1 and K_2 which are intrinsic matrices for camera 1 and camera 2. In addition, a 4x4 identity matrix is also needed. The projection point of camera one can be calculated using the following equation. S_1 is the transpose matrix of the face landmark coordinates.

$$\begin{aligned} P_1 &= I^{-1} K_1^{-1} S_1 \\ P_2 &= R | t^{-1} K_2^{-1} S_2 \end{aligned} \quad (2)$$

Equation (2) is projection point for camera two is added to the extrinsic matrix containing the rotation and translation matrices for camera 2 with respect to camera 1. The projection point for camera 2 can be calculated using the following equation

Based on Equation (3) is the Z value can be calculated using the triangulation method. Point (x_{c1}, z_{c1}) is the coordinate of the first camera, while point (x_{c2}, z_{c1}) is the coordinate of the second camera. Point (x_1, z_1) is the focal length coordinate of the first camera and (x_2, z_2) is the focal length coordinate of the second camera. The coordinates of point A are obtained from the intersection between the first camera line and the second camera. This intersection point can be calculated with the following equation.

$$\frac{X - x_{c1}}{x_1 - x_{c1}} = \frac{Z - z_{c1}}{z_1 - z_{c1}} \quad (3)$$

$$\frac{X - x_{c2}}{x_2 - x_{c2}} = \frac{Z - z_{c2}}{z_2 - z_{c2}} \quad (4)$$

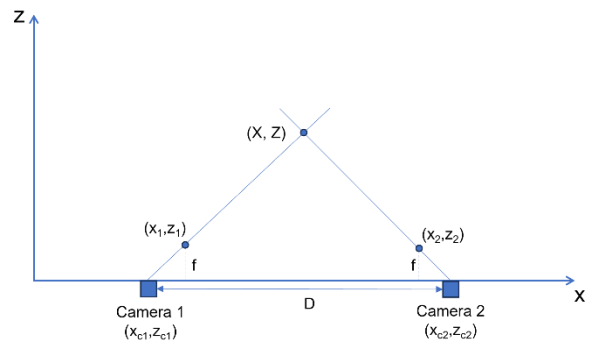


Fig. 2. Triangulation Method

Fig. 2. above represents the triangulation method used in MediaPipe refers to a technique for estimating the three-dimensional position and distance of a detected object (e.g., body parts, face, or hands) using data from two-dimensional image or video. Triangulation essentially involves utilizing data from two viewpoints or two reference points to calculate the 3D location of an object. If a single camera is used, mediapipe uses the relative size or positional shift of key points to estimate depth[24].

After obtaining the Z coordinate using equations (3) and (4), the X and Y coordinates are calculated using the equations (5) and (6). After obtaining the Z coordinate using equations (3) and (4), the X and Y coordinates are calculated using the following equations. Then the X, Y, and Z world coordinates for each point on the face of the two cameras are obtained.

$$X = \frac{xZ}{f} \quad (5)$$

$$Y = \frac{yZ}{f} \quad (6)$$

D. RMSE

Root Mean Squared Error (RMSE) is a metric used to assess how close a model's predictions are to its actual values. RMSE measures the average squared error between the prediction and the actual observation. The smaller the RMSE value, the better the model is at predicting data because it means that the predicted value is closer to the actual value[25]. Mathematically, the RMSE formula can be expressed as.

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_i')^2} \quad (7)$$






Equation (7) is y_i is the actual value at the data, y_i' is the predicted value on the data, n is the number of data.




III. RESULTS AND DISCUSSION

A. Result Face Landmark Using MediaPipe

Mediapipe is a framework for processing multimedia data (mainly videos and images) using machine learning. Mediapipe provides various pre-built solutions for visual processing, such as face detection, hand tracking, body pose detection, and more. This research focuses on face detection using Mediapipe.

TABLE 1. DETECTING FACIAL

Data	Person
Person 1	
1 st data	
2 nd data	
3 rd data	
4 th data	
Person 2	
5 th data	

6 th data	
7 th data	
8 th data	

Based on Table 1. shows the results of facial landmark detection using two cameras, the right image is the result of facial landmark detection using camera A and the left image is the result of facial landmark detection using camera B.

Table 1. shows the result of detecting facial landmark points using MediaPipe. Mediapipe detects various landmark points on the face, such as the eyes, nose, mouth, and facial contours, which help identify the position and orientation of a person's facial features. In each photo, MediaPipe assigns landmarks to areas of the face. These landmark points are used to map the position of key facial features, which can be useful for applications such as facial recognition, expression analysis, or facial motion tracking.

In data 1 and data 5 are images of faces towards the front, in data 2 and data 6 are images of faces to the right side, in data images 3 and 7 are images of faces to the left side, in data 4 and data 8 are images of facial expressions. Seen from the object using glasses, MediaPipe is generally quite robust in detecting facial landmarks even when the object uses glasses, but this condition can still affect the level of accuracy of predicting the position of the detected face around the eyes or eyebrows. In the 5th data point, MediaPipe showed a good ability to recognize facial landmarks despite variations in clothing (such as the use of headscarves) and facial expressions (such as smiles). This is important in applications that require face detection independent of the specific appearance of the object.

B. Camera Calibration Result 5, 6, and 7

The result of this calibration contains matrices that describe the geometric relationship between two cameras or between a camera and an object in three-dimensional space. This calibration is very important in applications such as computer vision, where we need to understand the orientation of the camera with respect to the observed object.

TABLE 2. MATRIX CALIBRATION

Rotation matrix: $\begin{bmatrix} 0.99071751 & 0.02043446 & -0.13439216 \\ -0.02004232 & 0.99979001 & 0.00427031 \\ 0.1344512 & -0.00153714 & 0.99091902 \end{bmatrix}$ Translation vector: $\begin{bmatrix} 4.31944805 \\ -0.24238817 \\ -0.28442107 \end{bmatrix}$ K1: $\begin{bmatrix} 623.23167429, & 0. & , \\ 322.64932086, & & , \\ & 0. & , 625.29959212, \\ 238.94256166, & & , \\ & 0. & , 0. & , 1. \end{bmatrix}$ K2: $\begin{bmatrix} 629.35648288, & 0. & , \\ 338.29296983, & & , \\ & 0. & , 633.02074654, 240.1059726 \\ , & & , \\ & 0. & , 0. & , 1. \end{bmatrix}$

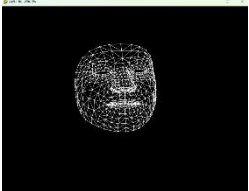
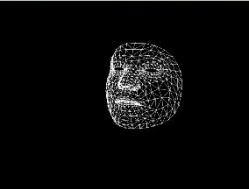

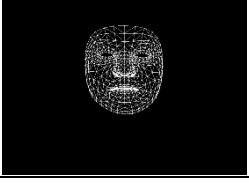
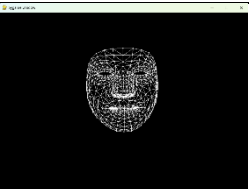
Based on Table 2. matrix, there is a rotation matrix that shows the relative rotation between the two cameras. This 3x3 matrix determines how a point in three-dimensional space rotates from one orientation to another. The rotation matrix indicates that there is little rotation between the two frames or cameras. Values close to 1 on the diagonal (for example, 0.99071751, 0.99979001, 0.99091902) indicate that this rotation is relatively small.

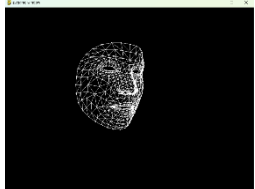
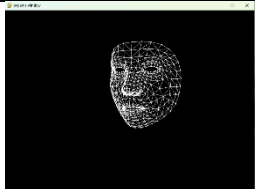
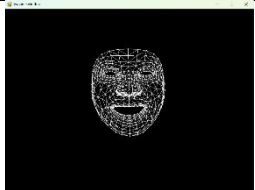
Translation vector shows the distance and direction of displacement between two frames or cameras in 3D space. In this matrix there is a displacement along the X-axis of 4.319, slightly downwards on the Y-axis (-0.242), and slightly towards the Z-axis (-0.284). This vector provides information about the relative position of the second camera to the first camera. Intrinsic Matrix is a 3x3 matrix that contains the internal parameters of each camera, such as focal length and optical center. This matrix serves to map 3D points on the 2D image recorded by the camera.

C. 3D Face Pose Result

The 3D reconstruction results are visualized using the Open Graphics Library (OpenGL), which is used as a tool to display 3D data due to its ability to display 3D objects quickly and efficiently, making it very suitable for data visualization such as 3D facial poses.

TABLE 3. 3D FACE

Data	Person
Person 1	
1 st data	
2 nd data	
3 rd data	
4 th data	
Person 2	
5 th data	

6 th data	
7 th data	
8 th data	

Based on Table 3. is the result of 3D reconstruction of facial landmarks based on triangulation of the landmark coordinates of camera A and camera B. The resulting 3D face pose shows the reconstruction of the face in three-dimensional form visualized using OpenGL. The figure shows a 3D view of the face, generated from the face pose data. This 3D reconstruction enables a more realistic and detailed view of the face. Each point or line on the face in the figure represents important landmarks or reference points on the facial structure, such as the nose, mouth, eyes, and facial contours. 3D Face Pose refers to the orientation or position of the face in three-dimensional space, including rotation in the X, Y, and Z axes. From the eight-figure displayed, it appears that the face is positioned at different angles, which may represent a change in the viewpoint or orientation of the face (e.g., frontal view, slight tilt, or other views).

Based on the experiments conducted, the following is Table 4. of the experimental results.

TABLE 4. EXPERIMENTAL RESULT

	Ground Truth (CM)			Testing Model (CM)		
	Eyes	Nose	Mouth	Eyes	Nose	Mouth
Person 1						
1 st data	14.3	5.00	6.00	14.34	1.04	8.26
2 nd data	14.3	5.00	6.00	17.96	3.89	9.23
3 rd data	14.3	5.00	6.00	17.93	3.99	10.06
4 th data	14.3	5.00	6.00	18.82	4.76	10.81
Person 2						
5 th data	17.8	4.00	6.00	17.88	3.9	10.07
6 th data	17.8	4.00	6.00	12.96	2.25	7.14
7 th data	17.8	4.00	6.00	12.9	3.59	8.08
8 th data	17.8	4.00	6.00	17.81	4.64	13.36
RMSE				3.560839	1.644749	4.054638

Based on Table 4. using the triangular method obtained the RMSE value in the eye section has an RMSE value of 3.560839, the nose section is 1.644749, the lips section is 4.054638.

D. Discussion

Based on experiments that have been carried out based on the table, it can be seen that the ground truth of the eyes and measurements using the triangular method are 14.3 and 17.96, it appears to have a significant difference so that the RMSE value is 3.560839. In the nose section the ground truth is 5.00 and the triangular method is 3.89, it looks to have a difference that is not too significant with an RMSE value of 1.64474, and finally in the lip section the ground truth is 6.00 and the triangular method is 9.23, it looks to have a difference that is not too significant with an RMSE value of 4.054638.

In this research, ground truth refers to the reference data used to evaluate the accuracy of 3D face reconstruction generated by the triangulation-based markerless motion capture method. The ground truth is obtained from manual measurements of facial dimensions, including distances between key facial features such as eyes, nose, and lips, which are then compared with the 3D reconstruction results to calculate the Root Mean Square Error (RMSE).

The analysis of the RMSE level of facial landmark detection using mediapipe, with the influence of accessories such as the use of glasses and headscarves. This shows that mediapipe is quite resistant to visual variations on the face, but also requires further testing on objects with more complex accessories. Furthermore, the effect of facial expression shows a good ability to recognize facial landmarks even when the facial expression (e.g., smile) changes. This is important in real-world applications, where facial expressions change frequently.

In terms of camera calibration analysis, based on the rotation matrix and translation vectors, it shows that there is little rotation between the two cameras, which indicates that the camera setups are relatively aligned or have slightly different viewing angles. The translation vectors provide important information about the relative distance between the cameras in 3d space, which is important for computer vision applications that involve merging data from two different sources to make the output into 2 parts. The given camera initialization matrices (K_1 and K_2) can be interpreted as a representation of the intrinsic parameters of the cameras indicating the focal length and optical center of each camera. The small variation between these two matrices indicates that the two cameras have similar configurations, but with small differences in their optical parameters.

In terms of 3D facial pose reconstruction analysis, it provides many benefits, especially in facial recognition and facial expression analysis applications. The ability to analyze the orientation and position of a face in three-dimensional space enables a variety of advanced applications. However, to achieve optimal face position prediction accuracy, image quality, lighting conditions, and input data quality are very important factors. Improvements in landmark detection and the use of high-quality cameras, as well as the fusion of data from multiple angles, will help improve the prediction accuracy of 3D reconstructed face positions.

Potential applications and implementations of this technology include facial recognition for security systems, facial expression tracking for emotion analysis, or even use in Augmented Reality (AR). In addition, potential uses in the industrial or health world, such as for emotion detection or user interaction. Further development can be done to improve

detection results, such as overcoming detection problems in masks or improving the accuracy of 3D face pose in various lighting conditions.

IV. CONCLUSION

Based on the experiments that have been carried out, the conclusion of this experiment is obtained:

1. In this research, the Triangular method is applied to measure the prediction error of several parts of the face. The results show that the RMSE values for the eyes are 3.560839, nose 1.644749, and lips 4.054638. These values indicate that the lips have the highest prediction error rate, followed by the eyes, while the nose has the lowest prediction error. The smaller the RMSE value, the better the performance of the model in predicting a value close to the actual value of a particular part.
2. 3D facial pose reconstruction using technologies such as OpenGL provides significant advantages in providing a more realistic and detailed representation of the face's position and orientation in three-dimensional space. The ability to track and analyze facial landmarks in 3D enables better recognition and tracking of facial expressions.
3. This research applies MediaPipe for facial landmark detection and does not explicitly estimate the influence of lighting, camera quality, or other technical parameters on the results of detection. Consequently, though reconstruction accuracy may be influenced by such factors, this study did not present direct experiments concerning changes in lighting or camera specifications. Future experiments should be designed in order to quantitatively assess the influence of such external conditions.

ACKNOWLEDGMENT

This research is supported by Multimedia and Internet of Things (M-IOT) Laboratory, Institut Teknologi Sepuluh Nopember, Surabaya, East Java, Indonesia.

REFERENCES

- [1] Q. Zhuang, Z. Kehua, J. Wang, and Q. Chen, "Driver fatigue detection method based on eye states with pupil and iris segmentation," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3025818.
- [2] H. O. Shahreza and S. Marcel, "Comprehensive Vulnerability Evaluation of Face Recognition Systems to Template Inversion Attacks via 3D Face Reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, 2023, doi: 10.1109/TPAMI.2023.3312123.
- [3] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, vol. 07-12-June-2015, doi: 10.1109/CVPR.2015.7298682.
- [4] U. Solichah, M. H. Purnomo, and E. M. Yuniarno, "Marker-less Motion Capture Based on Openpose Model Using Triangulation," 2020, doi: 10.1109/ISITIA49792.2020.9163662.
- [5] X. Zhu, C. Yu, D. Huang, Z. Lei, H. Wang, and S. Z. Li, "Beyond 3DMM: Learning to Capture High-Fidelity 3D Face Shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, 2023, doi: 10.1109/TPAMI.2022.3164131.
- [6] G. Tao, S. Sun, S. Huang, Z. Huang, and J. Wu, "Human modeling and real-time motion reconstruction for micro-sensor motion capture," 2011, doi: 10.1109/VECIMS.2011.6052193.
- [7] J. Jo, Y. J. Jung, and J. Kim, "3D face reconstruction from one side-view face images," 2014, doi: 10.1109/ELINFOCOM.2014.6914359.
- [8] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh,

- "Computationally efficient face spoofing detection with motion magnification," 2013, doi: 10.1109/CVPRW.2013.23.
- [9] Y. W. Cha *et al.*, "Towards fully mobile 3D face, body, and environment capture using only head-worn cameras," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 11, 2018, doi: 10.1109/TVCG.2018.2868527.
- [10] B. Hu *et al.*, "Face Landmark Calibration Based on 3D Reconstruction and Deep Learning," 2022, doi: 10.1109/DDCLS55054.2022.9858436.
- [11] W. Peng *et al.*, "IE-aware Consistency Losses for Detailed 3D Face Reconstruction from Multiple Images in the Wild," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6, doi: 10.1109/ICME57554.2024.10687861.
- [12] C. Darujati and M. Hariadi, "Facial motion capture with 3D active appearance models," 2013, doi: 10.1109/ICICI-BME.2013.6698465.
- [13] Y. Zhang, X. He, Y. Hu, J. Zeng, H. Yang, and S. Zhou, "Face animation making method based on facial motion capture," 2021, doi: 10.1109/ICESIT53460.2021.9696547.
- [14] L. Shi, Z. An, J. Zhao, L. Wang, and Q. Du, "A study of face motion capture and its data processing technique applied to the speech training of hearing-impaired children," 2012, doi: 10.1109/ICINIS.2012.63.
- [15] A. N. Ansari, M. Abdel-Mottaleb, and M. H. Mahoor, "3D face mesh modeling from range images for 3D face recognition," in *Proceedings - International Conference on Image Processing, ICIP*, 2007, vol. 4, doi: 10.1109/ICIP.2007.4380066.
- [16] G. Wen, Y. Yang, Z. Yang, J. Yue, and S. Yan, "A High-accuracy Camera Calibration Method Based on Special Circular Target," in *2022 International Conference on Industrial Automation, Robotics and Control Engineering (IARCE)*, 2022, pp. 24–28, doi: 10.1109/IARCE57187.2022.00015.
- [17] S. Bi, Y. Gu, Z. Zhang, H. Liu, C. Zhai, and M. Gong, "Multi-camera stereo vision based on weights," 2020, doi: 10.1109/I2MTC43012.2020.9128927.
- [18] C. Y. Vincent and T. Tjahjadi, "Multiview camera-calibration framework for nonparametric distortions removal," *IEEE Trans. Robot.*, vol. 21, no. 5, 2005, doi: 10.1109/TRO.2005.851383.
- [19] D. M. and A. T. P. S. Bisht, P. Colantoni, "MultiView Markerless MoCap - MultiView Performance Capture, 3D Pose Motion Reconstruction and Comparison," pp. 333–340, 2023, doi: 10.1109/SITIS61268.2023.00061.
- [20] M. Meghana, M. Vasavi, and D. D. Shrivani, "FACIAL LANDMARK DETECTION WITH MEDIAPIPE & CREATING ANIMATED SNAPCHAT FILTERS," *Int. J. Innov. Eng. Manag. Res.*, 2022, doi: 10.48047/ijiemr/v11/i06/10.
- [21] A. M. Al-Nuimi and G. J. Mohammed, "Face Direction Estimation based on Mediapipe Landmarks," 2021, doi: 10.1109/ICCITM53167.2021.9677878.
- [22] N. Kumar Rao B, N. Panini Challa, E. S. P. Krishna, and S. S. Chakravarthi, "Facial Landmarks Detection System with OpenCV Mediapipe and Python using Optical Flow (Active) Approach," 2023, doi: 10.1109/ICACITE57410.2023.10182585.
- [23] C. Zhang and Z. Zhang, "Improving multiview face detection with multi-task deep convolutional neural networks," 2014, doi: 10.1109/WACV.2014.6835990.
- [24] S. V. F. Barreto, R. E. Sant'Anna, and M. A. F. Feitosa, "A method for image processing and distance measuring based on laser distance triangulation," 2013, doi: 10.1109/ICECS.2013.6815509.
- [25] J. Lin, M. Zhao, G. Yin, H. Zhou, T. Hudoyberdi, and B. Jiang, "A Method for Depth Camera Calibration Based on Motion Capture System," 2023, doi: 10.1109/ICCD59681.2023.10420758.
- [26] Z.-H. Feng, P. Huber, J. Kittler, P. J. B. Hancock, X.-J. Wu, Q. Zhao, P. Koppen, and M. Räscher, "Evaluation of Dense 3D Reconstruction from 2D Face Images in the Wild," *arXiv preprint arXiv:1803.05536*, 2018. [Online]. Available: <https://arxiv.org/abs/1803.05536>.
- [27] W. Zhao, C. Yang, J. Ye, R. Zhang, Y. Yan, X. Yang, B. Dong, A. Hussain, and K. Huang, "From 2D Images to 3D Model: Weakly Supervised Multi-View Face Reconstruction with Deep Fusion," *arXiv preprint arXiv:2204.03842*, 2024. [Online]. Available: <https://arxiv.org/abs/2204.03842>.