ROBUST SELF-OCCLUSION MITIGATION FOR ENHANCED ACCURACY IN MULTI-VIEW 3D MARKERLESS MOTION CAPTURE SYSTEMS

Bo-Hung Chen[1], Cheng-Hung Tsai[2], Liang-Wei Huang[1], Mu-Hua Wang[1], Chiun-Sheng Hsu[2], You-Yin Chen[1]
[1] Department of Biomedical Engineering, National Yang Ming Chiao Tung University, Taipei, Taiwan.
[2]III Software Technology Institute, Institute for Information Industry, Taipei, Taiwan.
Email: youyin.chen@nycu.edu.tw

INTRODUCTION

Occlusion is the major factor affecting the accuracy of 3D triangulated keypoints of a person in the markless motion capture system, alongside baseline noise and camera calibration quality [2]. In our system, occlusion can be dealt with by identifying the occluded camera and excluding the 2D prediction data in the triangulation step. We employ a method with low computational power and an improved loss function by selecting the combination that minimizes the mean of the residuals from the singular value decomposition (SVD) within the direct linear transformation (DLT) process. After identifying the optimal combination, we triangulate the 2D coordinate into 3D using weighed direct linear transformation (wDLT) , where the confidence score is used as the weight.

METHODS

In the triangulation step, DLT [1] converts 2D coordinates from multiple cameras into 3D coordinates using the equation $\lambda q = PQ$, where $Q = (X, Y, Z, 1)$ represents the 3D point, $q = (u, v, 1)$ is single 2D image prediction, and P is the camera's projection matrix. With λ as an unknow scale factor, the equations become $(P_1^T - u\, P_3^T)Q = 0$ and $(P_2^T - v\, P_3^T)Q = 0$ for each camera. With $N = 4$ cameras in this study, this yields $2N$ equations, simplified to $AQ = 0$.

Using SVD to solve this equation by minimizing AQ gives us the best outcome from the equation. The residuals from SVD of AQ are able to indicate how well the triangulation is performed, which the residuals are significantly greater when inaccurate predicted 2D keypoints are included. Therefore, our method finds the combination of cameras which has the minimum value of AQ. It indicates the best quality of triangulation and eventually excludes the 2D prediction of cameras with occlusion issues. When the combination is determined, we perform wDLT [1]:

$$c \times (P_1^T - u\, P_3^T)Q = 0\, , c \times (P_2^T - v\, P_3^T)Q = 0 \quad (1)$$

where 2D coordinate with higher confidence scores have more effect on the triangulation, which enhances the trianlguated result.

RESULTS AND DISCUSSION

We find a significant difference of distribution in SVD residuals (p<0.0001) when using 3 cameras and 4 cameras. This distribution indicates that the index for identifying occlusion is sensitive enough to define when an occlusion occurs rather than relying on an experimental value [1]. Based on the threshold we set, **Figure 1**. compares different approaches on identifying occlusion after the triangulated result using SVD residuals and reprojection error [1]. The left panel successfully rules out the right elbow, which is apparently occluded, and the right panel fails to do so using reprojection error. This result shows in certain circumstances, SVD residuals have a better ability to identify occlusion compared to the conventional reprojection error method.

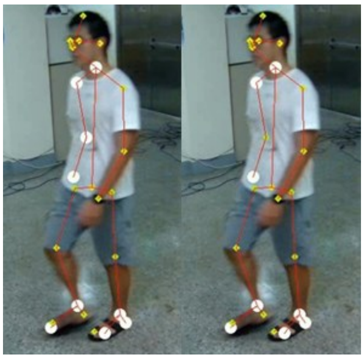

Figure 1. Result Comparison using SVD residuals and conventional reprojection error.

We also test our algorithm and the reprojection method against the VICON marker-based system. In a walking gait cycle, both correlation of coefficient (CC) and root-mean-square error (RMSE) for the right ankle are better comparing with reprojection error method shown in **Table1**.

Table 1. Result Comparison from VICON

Right Ankle	*CC*	*RMSE (degree)*
SVD residual	0.7027	11.8174
Reprojection error	0.6434	12.4597

CONCLUSION

There are many ways to deal with occlusion in motion capture systems, such as volumetric methods which require immense computational power [2], and the reprojection error index which cannot directly indicate the quality of the result from triangulation. Our approach of finding the minimum value of residuals improve the accuracy of final triangulated coordinate, due to directly reflect the triangulation quality, which reprojection error method fails, and thereby excluding the correct occluded camera.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the funding and support provided by the National Science and Technology Council, Republic of China, and the Administration for Digital Industries, Ministry of Digital Affairs, Republic of China. Thanks for their generous contributions that made this project possible.

REFERENCES

1. Pagnon, D., M. Domalain, and L. Reveret, Pose2Sim: An End-to-End Workflow for 3D Markerless Sports Kinematics-Part 1: Robustness. Sensors, 2021. 21(19).
2. Wang, J. B., et al. (2021). "Deep 3D human pose estimation: A review." Computer Vision and Image Understanding 210.