Development of a Real-Time Multi-Person 3D Keypoint Detection System Using Stereoscopic Cameras and RTMPose

Taufik Hidayat Soesilo

Graduate School of Engineering, University of Miyazaki, 1-1 Gakuen Kibanadai-Nishi, Miyazaki-shi, Miyazaki, 889-2155, Japan

Praveen Nuwantha Gunaratne

Interdisciplinary Graduate School of Agriculture and Engineering, University of Miyazaki, 1-1, Gakuen Kibanadai-Nishi, Miyazaki, 889-2192, Japan

Hiroki Tamura

Faculty of Engineering, University of Miyazaki,1-1 Gakuen Kibanadai-Nishi, Miyazaki-shi, Miyazaki, 889-2155, Japan
E-mail: ti22068@student.miyazaki-u.ac.jp, ti20060@student.miyazaki-u.ac.jp, htamura@cc.miyazaki-u.ac.jp

Abstract

The feature offers a way of tracking the movement patterns of many people at once, and is critical in occupational health, sports performance, and team-based work settings. The selected traditional biomechanical analysis systems, in turn, are largely considered in detecting single person movement, which weakens their applicability to movement analysis that involves interacting with multiple people. In this paper we consider a real time multi-person detection and analysis system using stereoscopic cameras and RTMPose, a novel high real-time pose estimation framework. RTMPose offers real time analysis of 2D key points for the individuals and this data is later augmented with depth data coming from stereoscopic imaging to give 3D skeletal data. The benefit of employing RTMPose is that the system is able to perform accurate and fast multiple persons tracking despite present occlusion scenarios. Consequently, the system overcomes the drawbacks of prior methods, including reliance on wearable devices and unsuitability for out-of-door environments, by employing stereoscopic cameras and RTMPose with low-latency and high-accurate inference. Experimental results demonstrate the system's ability to provide detailed real-time analysis of posture and movement for multiple individuals in diverse scenarios. This research highlights the potential of RTMPose-powered systems to advance multi-person biomechanical analysis for applications ranging from workplace monitoring to sports performance assessment.

Keywords: RTMpose, real-time biomechanical analysis, computer vision, multi-person 3D keypoint detection

1. Introduction

The requirement for the decomposition of multiple individuals at once in real-time has emerged in some fields including occupational health, sports, and ergonomics. Extended conventional frameworks of biomechanical assessment have generally been developed to examine one person at a time, thus being restricted in multi-subject environments. For instance, tasks of analysis in a team environment including construction, healthcare setting or production line work involve capacity to discern group motions for safety, efficiency and order to enhance performance. That is why there is a need to create new systems that would be able to detect and analyze multiple people with increasing complexity of backgrounds [1],[2].

In particular, the progresses in computer vision and stereoscopic imaging technology enable the multi-person detection in real time. While many older approaches only use singe-camera equipment or body-worn sensors, stereoscopic systems are based on two-camera configurations for depth acquisition. This enables the realist rendition of 3D keypoints for the multiple individuals, based on the crowded or overlapping states. These capabilities can be highly employed in cases where all the body movement of the subjects has to be studied at the same time as in sports, ergonomic studies and group work areas [3],[4].

Advanced techniques including deep learning based pose estimation in multi-person detection are currently employed. Some popular benchmarks like OpenPose, AlphaPose, and RTMPose have demonstrated satisfactory accuracy of body-structured 2D human pose estimation of multiple persons [5],[6],[7]. When correlated with depth information gathered from stereoscopic cameras these methods allow for determining three-dimensional corresponding key points for each of participants. This integration solves some of the issues that include occlusion, lighting variation, and objects sharing the same space which tends to complicate the model but produces efficient and accurate tracking applicable in real-life scenarios [8],[9].

Building on these advancements, this study focuses on developing a system for multi-person detection and analysis using stereoscopic cameras. The proposed solution offers detailed 3D key-point data for multiple individuals in real-time, addressing the limitations of single-person systems. By employing state-of-the-art multi-person detection algorithms and robust depth-sensing technology, the system is designed to operate effectively across diverse conditions, making it suitable for a wide range of applications, from workplace monitoring to athletic performance analysis [10].

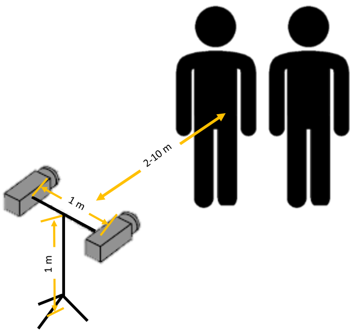


Fig. 1. Camera setup

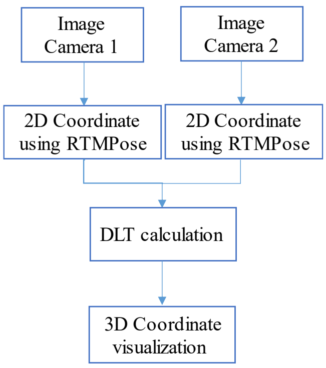


Fig. 2. System Flowchart

2. System Design and Methodology

2.1. System setup

The proposed system also utilizes a stereoscopic camera combination to acquire simultaneous videos that are helpful in the determination of multiple individuals. The setup consists of having two cameras placed in parallel with a defined baseline distance of 1 meter to keep a high level of accuracy of the depth estimation, as depicted in Fig. 1. The cameras are adjusted so as to define the intrinsic and extrinsic parameters that provide the foundations for enriching 3D data.

Similarly, the stereoscopic camera setup is positioned to obtain a rich view and capture all regions of the detected scene to identify different postures and orientations of the individuals. The videos captured are then processed in real time in order to obtain 2D key points for each person by using RTMPose. It maintains a high degree of visibility and small-object discernibility because of the proper positioning of the devices.

Key Hardware Details:

- **Camera Model:** Elgato Facecam Full HD
- **Resolution:** 1920x1080 pixels
- **Frame Rate:** 60 fps
- **Baseline Distance:** 1 meter
- **Calibration Tools:** OpenCV stereo calibration (to compute intrinsic/extrinsic parameters and distortion coefficients)

2.2. *Pose Estimation Framework*

The identification of multiple people and their skeletons is done using RTMPose, an efficient real-time pose estimation tool. RTMPose takes the synchronized video feeds from the stereoscopic cameras, and identifies the 2D key points of all people in each frame. The RTMPose model uses a deep learning model for real time object detection with high speed and accuracy the model for RTMPose is ResNet or MobileNet. This framework extracts 17 specific points (according to the COCO dataset) for every individual: shoulders, elbows, knees, ankles as well as other parts. These 2D key points are moreover linked with unique IDs corresponding to each person to afford stable correspondence from frame to frame. The system flowchart can be seen in the Fig 2

Fig. 3. Multi-Person and detected keypoints in 2D for applications of skeleton overlays.

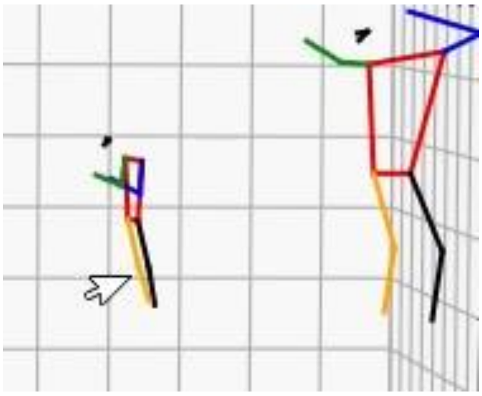


Fig. 4. The 3D coordinates for the ribs of both individuals were reconstructed.

2.3. *3D Keypoint Reconstruction*

The 3D key points for each subject are reconstructed using the 2D key point data obtained from RTMPose and the depth obtained from the stereoscopic cameras. This process involves extracting the 2D key point coordinates from the left and right camera; projected into 3D space.

The reconstruction is performed using the Direct Linear Transformation (DLT) technique that estimates the coordinates of all the important points with respect to the coordinate frame of the camera [11]. Extrinsic and intrinsic calibration data is used to compute the coefficients for lens distortion and thereby facilitates the calculation of depth.

3. *Results and Discussion*

3.1. *Results*

The proposed system can display real time multi-person detection, and 3D keypoints reconstruction with stereoscopic camera and the RTMPose. As it can be seen from the outcomes that the developed system successfully recognizes several people concurrently literally in a real conditions as well and yields quite high accuracy rates in terms of the Keypoints tracking in addition to the real life possibilities to generate the 3D bone models.

Key Observations:

1. 2D Keypoint Detection:
   - The system successfully detected key points for all bodies like some primary joints of shoulder, elbow, knee, ankles of more than one person. Congruent outlines drawn around each of the individuals established sharp and clear segmentation and identification of subjects as shown in Fig 3.
2. 3D Keypoint Reconstruction:
   - With data obtained from the stereoscopic camera setup the system constructed 3D keypoints for each person. These are well oriented with the detected individuals in the 2D images, in order to have accurate spatial correspondences in the 3D environment as the skeletal models can be seen in the Fig 4.
3. Performance Metrics:
   - Accuracy: The computed 3D keypoints of the reconstructed ego-vehicle appeared to match the position of the individuals regardless of their execution of different movements and partially obscured body parts.
   - Processing Speed: The system was able to obtain real time performance with average frame rate of the order of 15 fps.
4. Multi-Person Capability:
   - Several people were effectively detected and tracked within the field of view, in the 2D view demonstrated by the bounding boxes and in the 3D view through the skeleton overlays.

3.2. Discussion

The results validate the effectiveness of the proposed system in handling real-time multi-person detection and 3D keypoint reconstruction. Several key aspects are

Challenges and Limitations:

1. Occlusion Handling: In most cases, the system provided accurate detection of keypoints and the estimation of 3D poses, but full occlusions of the body parts led to small errors. Scanner input occlusion handling could be the subject of future work, with temporal data or limit algorithms used.
2. Complex Environments: The current testing environment was relatively controlled. Performance with signals in more difficult conditions such as in a noisy environment with many subjects and with greatly fluctuating light conditions requires assessment.
3. Depth Accuracy: Disparity based depth estimation technique may contain noises at some time period when the camera is moving fast or during large change in lighting. However, this problem could be reduced by improved calibration on the camera setup or using better filtering methods.

Potential Applications:

1. Workplace Monitoring: They include real-time tracking of multiple workers to help monitor safety and even workers' ergonomics.
2. Healthcare: To maintain correct movements for a number of patients at once, group rehabilitation sessions are used.

4. Conclusion

The proposed system is capable of real-time multi-person detection and 3D keypoint reconstruction, while using stereoscopic cameras and RTMPose. It shows high precision, low delay, and multiple person tracking ability, hence, it can be used for workplace monitoring, sports performance analysis, and group physical therapy.

The performance of the system is excellent, however, it is essential to consider the following factors: Occasion where one of the cameras is occluded by the other, noisy depth estimation, and the existence of other objects within a scene. The future work will therefore aim at enhancing these aspects as well as incorporating other features such as action recognition to the system. This system demonstrates better advance of multi-person biomechanical analysis and will have an enhanced foundation for related purposes.

References

1. Z. Cao, G. Hidalgo, T. Simon, et al., "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, 2019. doi:10.1109/TPAMI.2019.2929257.
2. H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, and Y. Xiu, "AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 6, pp. 7157–7173, Jun. 2023. doi:10.1109/TPAMI.2022.3222784.
3. S. Dill, A. Ahmadi, M. Grimmer, D. Haufe, M. Rohr, Y. Zhao, M. Sharbafi, and C. H. Antink, "Accuracy Evaluation of 3D Pose Reconstruction Algorithms Through Stereo Camera Information Fusion for Physical Exercises with MediaPipe Pose," Sensors, vol. 24, no. 23, p. 7772, 2024. doi:10.3390/s24237772.
4. F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking," arXiv, Jun. 2020. doi:10.48550/arXiv.2006.10214.
5. J.-K. Tsai, C.-C. Hsu, W.-Y. Wang, and S.-K. Huang, "Deep Learning-Based Real-Time Multiple-Person Action Recognition System," Sensors, vol. 20, no. 17, p. 4758, Aug. 2020. doi:10.3390/s20174758.
6. A. Mathis, P. Mamidanna, K. M. Cury, et al., "DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning," Nature Neuroscience, vol. 21, pp. 1281–1289, Sep. 2018. doi:10.1038/s41593-018-0209-y.
7. T. Jiang, P. Lu, L. Zhang, N. Ma, R. Han, C. Lyu, Y. Li, and K. Chen, "RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose," arXiv, Jul. 2023. doi:10.48550/arXiv.2303.07399.
8. T. Golda, T. Kalb, A. Schumann, and J. Beyerer, "Human Pose Estimation for Real-World Crowded Scenarios," in 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, Sep. 2019, pp. 1–8. doi:10.1109/AVSS.2019.8909823.
9. N. Sarafianos, B. Boteanu, B. Ionescu, and I. A. Kakadiaris, "3D Human Pose Estimation: A Review of the Literature and Analysis of Covariates," Computer Vision and Image Understanding, vol. 152, pp. 1–20, Nov. 2016. doi:10.1016/j.cviu.2016.09.002.
10. D. Rodriguez-Criado, P. Bachiller-Burgos, G. Vogiatzis, et al., "Multi-person 3D pose estimation from unlabeled data," Machine Vision and Applications, vol. 35, no. 46, Apr. 2024. doi:10.1007/s00138-024-01530-6.
11. T. Batpurev, "Real time 3D body pose estimation using MediaPipe," GitHub. [Online]. Available: https://github.com/TemugeB/bodypose3d. [Accessed: Nov. 14, 2022].

Authors Introduction

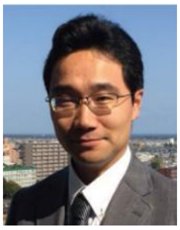
Mr. Taufik Hidayat Soesilo

He received his Bachelor's degree in Engineering in 2021 from the Faculty of Engineering, University of Brawijaya in Indonesia. He is currently a doctoral student in University of Miyazaki, Japan

Mr. Praveen Nuwantha Gunaratne

He received his bachelor's degree in engineering in 2018 from the Faculty of Engineering, University of Moratuwa, Sri Lanka. He is currently a doctoral student in University of Miyazaki, Japan.

Prof. Hiroki Tamura



He received the B.E. and M.E. degree from Miyazaki University in 1998 and 2000, respectively. From 2000 to 2001, he was an Engineer in Asahi Kasei Corporation, Japan. In 2001, he joined Toyama University, Toyama, Japan, where he was a Technical Official in the Department of Intellectual Information Systems. In 2006, he joined Miyazaki University, Miyazaki, Japan, where he was an Assistant Professor in the Department of Electrical and Electronic Engineering. Since 2015, he is currently a Professor in the Department of Environmental Robotics. His main research interests are Neural Networks and Optimization Problems. In recent years, he has had interest in Biomedical Signal Processing using Soft Computing.