Comparing the Quality of Human Pose Estimation with BlazePose or OpenPose

Sarah Mroz
Department of Mechanical Engineering
University of Ottawa
Ottawa, Canada
smroz040@uottawa.ca

Natalie Baddour
Department of Mechanical Engineering
University of Ottawa
Ottawa, Canada
nbaddour@uottawa.ca

Connor McGuirk
Department of Mechanical Engineering
University of Ottawa
Ottawa, Canada
cmcgu019@uottawa.ca

Pascale Juneau
Department of Mechanical Engineering
University of Ottawa
Ottawa, Canada
pjune022@uottawa.ca

Albert Tu
Department of Surgery
Children's Hospital of Eastern Ontario
University of Ottawa
Ottawa, Canada
atu@cheo.on.ca

Kevin Cheung
Department of Surgery
Children's Hospital of Eastern Ontario
University of Ottawa
Ottawa, Canada
kcheung@cheo.on.ca

Edward Lemaire
Faculty of Medicine
Ottawa Hospital Research Institute
University of Ottawa
Ottawa, Canada
elemaire@ohri.ca

2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART) | 978-1-6654-0810-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIOSMART54244.2021.9677850

***Abstract*— Human pose estimation is a computer vision task that predicts the position of person's body landmarks within a given image or video. This technology could help provide virtual motion assessments by analyzing videos captured when the patient is outside a clinical setting. In this study, a newer pose estimation model that can run on a smartphone (BlazePose) was compared to a well-accepted solution (OpenPose) to determine if these models can provide clinically viable body keypoints for virtual motion assessment. Using ten videos of clinically relevant movements (recorded by physicians), keypoint coordinates were generated from each model. Using OpenPose as a baseline, Pearson correlation and root mean square error were calculated between the BlazePose and OpenPose keypoint trajectories. BlazePose had more instances where keypoints deviated from anatomical joint centres, compared to OpenPose, indicating the BlazePose was not yet viable for clinically relevant assessments. However, BlazePose runtime was much faster than OpenPose and returned metrics that could be incorporated into a smartphone solution. Future designs of a smartphone-based system for conducting virtual motion assessments should utilize OpenPose for pose estimation; however, BlazePose could be used for other design aspects such as movement pre-screening or activity classification.**

Keywords*— *artificial intelligence, human pose estimation, remote assessment, motion analysis

I. Introduction

Artificial intelligence (AI) powered systems are increasingly changing conventional medical processes by implementing technology to enhance elements of patient care [1],[2]. The demand for virtual delivery of medical services was heightened by the COVID-19 global pandemic. Efforts have expanded to develop applications incorporating AI algorithms to execute tasks typically completed in clinical settings [3]. This concept extends to the need for virtual movement assessments. Virtual assessments often rely on a caregiver or family member capturing relevant images or sensor data. Therefore, clinically relevant data needs to be collected without the use of specialized equipment, ideally though an on-device smartphone solution.

Quantitative motion analysis typically involves a dedicated motion laboratory, requiring specialized equipment to collect clinically relevant data [4]. A multidisciplinary team of experienced personnel is needed conduct the assessment, process the data, and interpret the results upon which to base clinical decisions [5]. While motion laboratories can provide very accurate and valuable data, barriers to accessing this service include available infrastructure, geographical accessibility for patients, equipment requirements, and clinician resources. These barriers could be reduced by implementing virtual motion assessments.

AI models that estimate body keypoints to describe body position have become a potentially useful tool for virtual motion assessments. Human pose estimation typically uses Convolution Neural Networks (CNN) to predict the position of a person by performing inference on input images or videos [6]. Determining accurate pixel coordinates of body keypoints is a complex task due to the large number of possible human poses, the large number of degrees of freedom, changes in appearance such as clothing and lighting, changes in the environment, and occlusions [6],[7]. Despite these challenges, many robust models have been constructed that perform reasonably well for applications such as athletic training, rehabilitation, and sign language [8]–[10].

While pose estimation models have succeeded in many applications, the ability to correctly identify keypoints is necessary for clinical decision-making. In a previous study conducted by Zhang et al.[11], OpenPose BODY25 more often correctly labeled body keypoints than HyperPose COCOv2. This study extends this initiative by comparing another pose estimation model, BlazePose, to OpenPose.

The BlazePose model, developed by Google (Mountain View, CA), utilizes a two-step detector-tracker inference pipeline [12]. The detector runs on the first frame, or until a person is detected, then the tracker is employed to track the person in subsequent frames [12]. In this model, Bazarevsky et al. "use an encoder-decoder network architecture to predict heatmaps for all joints followed by another encoder that regresses directly to the coordinates of all joints" [12, p. 1]. This architecture enables real-time inference, combined with its lightweight nature, making BlazePose favorable for smartphone applications.

978-1-6654-0810-3/21/$31.00 ©2021 IEEE

Real-time capabilities offer considerable advantages in the healthcare industry where immediacy is often desired. This advantage could further improve virtual motion assessment outcomes by providing the patient and clinician with real-time feedback. This feedback could be used to correct patient movements and positioning in the frame to collect more "AI-friendly" data and provide the clinician with meaningful data to make immediate recommendations.

This research aimed to determine the viability of the newer, lightweight CNN architecture for human pose estimation, BlazePose, and compare correct keypoint identification to OpenPose.

II. Methods

A. Video Procurement

A series of ten videos were procured for this study with discrete emphasis on the upper and lower extremities (Table I). Videos were divided into five demonstrating upper extremity movements and five demonstrating lower extremity gait. Two experienced physicians selected movements with clinical and functional relevance to gross motor function. The upper extremity videos encompassed a range of motion across the shoulder and elbow joints obtained from a single point of view. The lower extremity videos consisted of an individual walking with a consistent gait while different video capturing vantages and techniques were used. The physicians recorded each video containing a single person in frame. The smartphone videos were captured using an iPhone XS (2018) and iPhone XR (2018) using default settings. All videos passed to the AI models were in MP4 format.

B. Data Acquisition and Processing

The videos were passed to BlazePose and OpenPose to obtain pose estimation coordinates for all body keypoints. BlazePose returned 33 3-Dimensional (3D) pose landmarks (coordinates in the frame's x-y plane and a z-coordinate representing landmark depth from the camera) and a visibility score (i.e., keypoint "in-frame likelihood") [13]. In contrast, OpenPose output 25 2-Dimensional (2D) keypoints and a confidence score (i.e., confidence in keypoint location) [11]. Only 2D pixel coordinates of the 17 keypoints from the standard COCO topology [14] were analyzed because these points were shared across both models (Fig.1).

The video processing and rendering techniques used in the study conducted by Zhang et al. [11] were replicated with a few differences. The OpenPose keypoint data was processed by initially gap-filling cases where the confidence score was below a chosen threshold of 10% and interpolated using a cubic spline [11]. Filtering was then applied to achieve the best performance, corresponding to the use of a zero-phase second-order low-pass Butterworth filter with a 12Hz cutoff frequency [11]. There remained few instances where OpenPose did not return keypoint data. Missing keypoints occurred when the body landmark was occluded for multiple frames, spanning beyond the interpolation window set to 5 frames. In these cases, OpenPose returned -1 instead of keypoint coordinates.

The BlazePose keypoints were filtered using the same technique. Prior to this, the visibility score was used in place of the confidence score for the thresholding and interpolation process. Therefore, this process only occurred when the visibility score was less than 10%, which was uncommon across datasets. Unlike OpenPose, during periods of heavy occlusion, BlazePose returned coordinates for all keypoints if a person was detected in the frame by inferring pose positions from previous frames. Only frames with both OpenPose and BlazePose complete keypoint sets were included in the analysis.

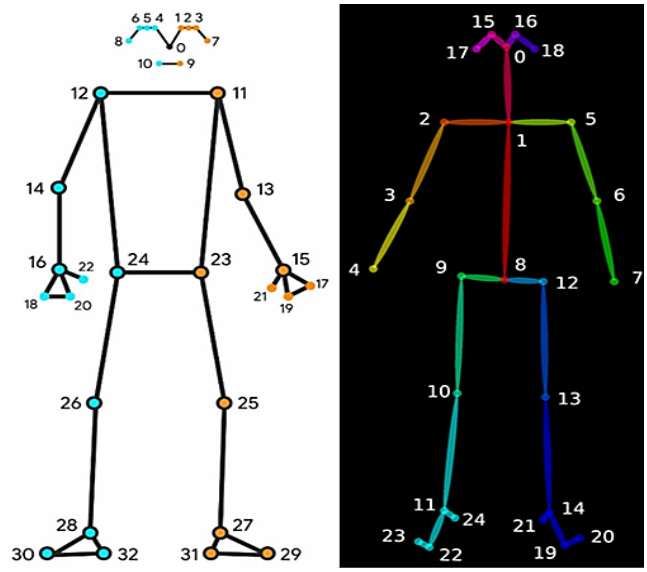

Fig. 1. 2-Dimensional skeleton keypoint topology for BlazePose (left) and OpenPose (right). BlazePose includes 33 keypoints in comparsion to OpenPose that includes 25 keypoints. BlazePose contains more keypoints for the hands and face.

For both models, the processed data was used to overlay a skeleton on each video frame. Keypoints were denoted with a colored circle and connecting lines along body segments (Fig.1). These overlays provided a visual reference for how well each model performed. Frames were qualitatively assessed based on how well the keypoints corresponded to the appropriate body landmarks.

C. Data Analysis

Using the same 4-point scale as Zhang et al. [11], each frame of the rendered OpenPose videos was scored by two reviewers. This process involved determining what keypoints were present in the frame, and how closely the overlayed coordinates aligned with the appropriate joint centers. Utilizing the evaluated performance of OpenPose, comparison metrics were computed to infer how closely the BlazePose model predicted keypoints.

Using Microsoft Excel, Pearson correlations were computed for each of the mutual 2D pixel coordinate trajectories. Any keypoint coordinate with a correlation coefficient of at least +/-0.8 was deemed highly correlated, and no further analyses were conducted. For the remaining keypoints with correlations below the threshold, the trajectories were plotted, and the rendered videos were examined to qualitatively interpret the deviations between models. To quantify these deviations, the Root Mean Square Error (RMSE) was calculated and converted to a percentage of the pixels along the respective axis of the frame.

III. Results

Ten videos, including five upper extremity and five lower extremity demonstrating clinically relevant motions were used to compare the performance of the OpenPose and BlazePose human pose estimation models. The composition of the videos for each movement are outlined in Table I.

TABLE I. INCLUDED VIDEO COMPOSITION

ID	*Movement*	*Plane*	*Orientation*	*# Frames*
1	Shoulder Abduction	Frontal (AP)	Portrait	250
2	Shoulder External Rotation	Frontal (AP)	Portrait	195
3	Hands-to-neck (External Rotation)	Frontal (AP)	Portrait	345
4	Hands-to-back (Internal Rotation)	Frontal (AP)	Portrait	198
5	Hands-to-mouth (Elbow Flexion)	Frontal (AP)	Portrait	131
6	Walking	Oblique (AP)	Landscape	140
7	Walking	Frontal (AP)	Portrait	145
8	Walking	Oblique (PA)	Landscape	128
9	Walking	Frontal (PA)	Portrait	160
10	Walking	Sagittal	Landscape	116

Only keypoints corresponding to joints that contributed to the respective movement were considered for comparison. For all upper extremity videos, the keypoints distal to the hips were omitted since the person remained standing in place throughout the video. For the lower extremity walking videos, all keypoints, except the keypoints proximal to the neck were considered since both the upper and lower extremities contributed to the gait cycle.

In total, 200 keypoint coordinates were considered in the comparison. 26 keypoint coordinate trajectories had correlations below the +/-0.8 threshold. These cases included 15 coordinates from the lower extremity videos and 7 from the upper extremity videos. The Pearson correlations and RMSE for these instances are presented in Table II.

TABLE II. CORRELATIONS BELOW THRESHOLD AND ROOT MEAN SQUARE ERROR

Video ID	*Keypoint*	*Correlation*	*RMSE (% pixels)*
1	Right Shoulder, y	0.71	2.77
	Right Elbow,y	0.62	3.49
	Right Wrist, y	0.68	4.16
2	Right Elbow,y	0.79	0.70
	Left Shoulder, x	0.59	0.64
3	Highly correlated for all upper extremity keypoints		
4	Right Shoulder, y	0.72	0.24
5	Right Shoulder, x	0.68	1.03
6	Left Shoulder, y	0.23	0.39
	Right Shoulder, y	0.77	0.29
7	Right Hip, x	0.60	5.03
	Right Knee, x	0.73	1.50
	Right Ankle, x	0.55	1.79
	Left Hip, x	0.41	9.56
8	Right Shoulder, y	0.71	1.03
	Left Shoulder, y	0.76	1.00
9	Right Hip, x	0.71	4.06
	Left Hip, x	-0.66	7.71
	Left Knee,x	0.73	6.67
	Left Ankle,x	0.78	4.97
10	Left Shoulder, y	0.79	0.52
	Left Elbow, y	0.72	0.65
	Left Wrist, y	0.63	1.51
	Left Knee, y	0.21	11.95
	Left Hip, y	0.55	25.87
	Right Ankle,y	0.77	11.11
	Right Knee, y	0.68	13.14

Of all considered keypoint coordinates, BlazePose had an average correlation of 0.88 +/- 0.20 in comparison to the OpenPose baseline keypoints. The average RMSE of the keypoints with correlations less than the threshold was 4.68 +/- 5.82% pixels in each row.

From qualitative assessments of the joint trajectories on the resulting skeleton overlay videos, BlazePose keypoints did not always correspond to the respective joint center. These situations were more likely to occur when the person was moving between frames, the environment had high contrasting backgrounds or additional objects were in the frame, creating a more challenging task for the AI models (Fig 2).

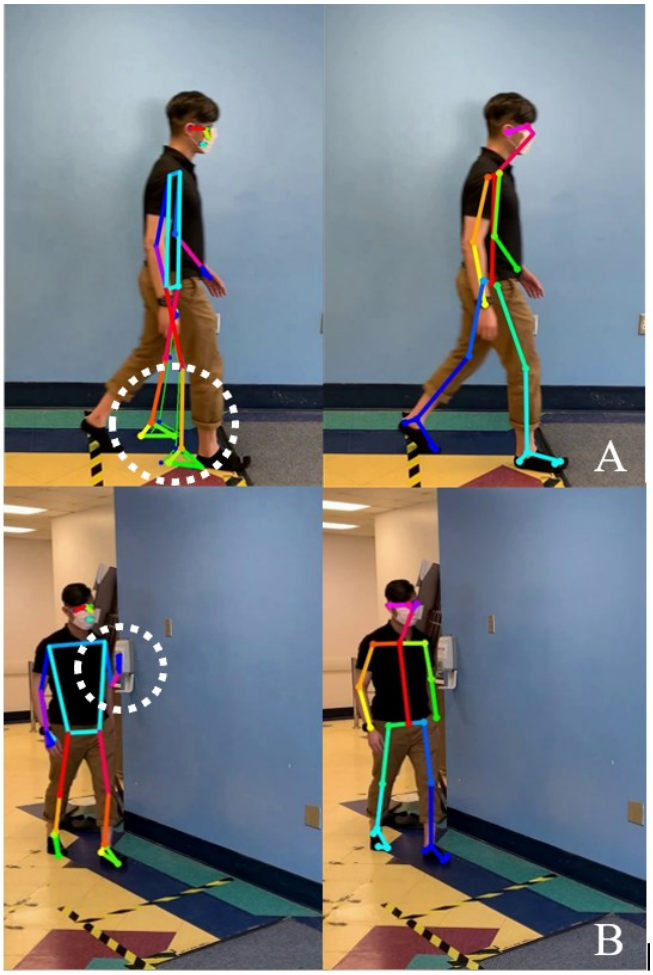

Fig. 2. Examples when the BlazePose (left) model incorectly predicted keypoint coordiantes, compared to OpenPose (right). In video 10 (A), the foot keypoints are located away from the respective joint centers. In video 6 (B), the model misinterperts an object in the environemnt for body landmarks.

IV. DISCUSSION

OpenPose was better than BlazePose at providing appropriate keypoint inference from videos of movements relevant to clinicians. The newer, lightweight model BlazePose had real-time capabilities and 3D outputs that are attractive features to implement in a remote assessment tool for motion analysis. However, correct identification of the predicted joint centers is imperative for a clinically viable solution. For this application, incorrect keypoint coordinates could result in inaccurate patient assessment by the clinicians interpreting the outcome measures.

Deviations were identified between keypoint coordinates of BlazePose when compared to OpenPose. These deviations were quantified by the RMSE and identified qualitatively through the rendered videos, indicating decreased quality in the predicted keypoint coordinates. BlazePose occasionally misidentified objects in the environment as landmarks, labelling keypoint coordinates away from the respective joint centers. These instances occurred more frequently in the lower extremity gait videos when all keypoints moved between frames creating a more challenging computer vision task. Improvement was seen for videos where the person was standing still indicated by a lower number of keypoints with correlations below the threshold. This was expected and corresponded to similar results by Zhang et al. [11].

BlazePose can process videos in real-time, having this advantage over OpenPose. OpenPose can process videos in 22 fps [15] compared to BlazePose's speed of 140fps[16]. For the videos included in the study, BlazePose could perform approximately six times faster than OpenPose.

The average correlation between models was above the correlation threshold of 0.8. Considering that BlazePose can run much faster than OpenPose, this is a notable result. Because the quality of predicted keypoints was not as high, however, the BlazePose AI model was not clinically viable for completing pose estimation. The model does offer several advantages that could be incorporated into other aspects of a virtual motion assessment tool:

- BlazePose can process input videos in real-time. The immediate feedback could guide the person taking the video to record higher quality videos for later processing by a more accurate pose estimation model.
- The visibility score relating to "in-frame likelihood" could provide feedback of the patient's position in the frame. This information could trigger prompts to the patient to change position to ensure all keypoints are within the frame, thus, improving the "AI friendliness" of the video.
- The 3D-coordinate output could provide increased capabilities compared to 2D models. Outcome measures that rely on a reference of the patient's depth in the frame could be included. Future research should determine the accuracy and viability of the z-coordinate for incorporation into virtual motion assessments.

V. Conclusion

This study demonstrated that the BlazePose model did not perform as well compared to OpenPose when predicting coordinates for human pose estimation. However, there exists a trade-off between the speed and correct keypoint identification of the two models. The BlazePose model had faster runtime performance and lightweight nature that can be executed on-device. Considering the requirement for correct identification of the predicted keypoints to be clinically viable, it is recommended to use OpenPose for pose estimation. BlazePose could be incorporated to aid in real-time feedback for positioning the subject in the frame to increase the "AI-friendliness" of the patient-recorded video.

Virtual motion assessments have the potential to overcome many barriers that patients and clinicians face to complete traditional motion assessments. The drawback of requiring the patient to be physically present for evaluation was heightened by COVID-19 and expedited the need for these technologies. The overall efficiency and accessibility of these movement measurement procedures can be improved by implementing accurate AI-powered pose estimation models.

Acknowledgment

This research was supported by the Natural Sciences and Engineering Research Council of Canada Discovery and CREATE programs. The authors acknowledge Caitlin Roe, Leah McDonell, Victoria Wyman and Christopher Rizkallah for their contributions to the review of rendered videos.

References

[1] G. Zoppo *et al.*, "AI technology for remote clinical assessment and monitoring," *J. Wound Care*, vol. 29, no. 12, Dec. 2020, Accessed: Jul. 28, 2021. [Online]. Available: https://www.magonlinelibrary.com/doi/full/10.12968/jowc.2020.29.12.692

[2] S. Rushabh and A. Chircu, "IoT and AI in healthcare: A systematic literature review," *Issues Inf. Syst.*, vol. 19, no. 3, pp. 33–41, 2018.

[3] S. Bhaskar *et al.*, "Designing Futuristic Telemedicine Using Artificial Intelligence and Robotics in the COVID-19 Era," *Front. Public Health*, vol. 0, 2020, doi: 10.3389/fpubh.2020.556789.

[4] J. T. Long and G. F. Harris, "Pediatric gait and motion analysis: Current limitations and emerging opportunities for quantitative assessment," *Technol. Disabil.*, vol. 22, no. 4, pp. 199–205, Jan. 2010, doi: 10.3233/TAD-2010-0304.

[5] R. B. Davis *et al.*, "A minimum standardized gait analysis protocol: development and implementation by the Shriners Motion Analysis Laboratory network (SMALnet)," in *Pediatric Gait: A New Millennium in Clinical Care and Motion Analysis Technology*, Jul. 2000, pp. 1–7. doi: 10.1109/PG.2000.858868.

[6] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in *Computer Vision – ECCV 2016*, Cham, 2016, pp. 483–499. doi: 10.1007/978-3-319-46484-8_29.

[7] A. Bulat and G. Tzimiropoulos, "Human Pose Estimation via Convolutional Part Heatmap Regression," in *Computer Vision – ECCV 2016*, Cham, 2016, pp. 717–732. doi: 10.1007/978-3-319-46478-7_44.

[8] J. Wang, K. Qiu, H. Peng, J. Fu, and J. Zhu, "AI Coach: Deep Human Pose Estimation and Analysis for Personalized Athletic Training Assistance," in *Proceedings of the 27th ACM International Conference on Multimedia*, New York, NY, USA, Oct. 2019, pp. 374–382. doi: 10.1145/3343031.3350910.

[9] Y. Li, C. Wang, Y. Cao, B. Liu, J. Tan, and Y. Luo, "Human pose estimation based in-home lower body rehabilitation system," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–8. doi: 10.1109/IJCNN48605.2020.9207296.

[10] A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan, "Real-Time Sign Language Detection Using Human Pose Estimation," in *Computer Vision – ECCV 2020 Workshops*, Cham, 2020, pp. 237–248. doi: 10.1007/978-3-030-66096-3_17.

[11] F. Zhang *et al.*, "Comparison of OpenPose and HyperPose artificial intelligence models for analysis of hand-held smartphone videos," in *2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, Jun. 2021, pp. 1–6. doi: 10.1109/MeMeA52024.2021.9478740.

[12] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," *ArXiv200610204 Cs*, Jun. 2020, Accessed: Aug. 06, 2021. [Online]. Available: http://arxiv.org/abs/2006.10204

[13] "Pose Detection | ML Kit," *Google Developers*. https://developers.google.com/ml-kit/vision/pose-detection (accessed Aug. 30, 2021).

[14] T.-Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," *ArXiv14050312 Cs*, Feb. 2015, Accessed: Aug. 09, 2021. [Online]. Available: http://arxiv.org/abs/1405.0312

[15] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[16] "High Fidelity Pose Tracking with MediaPipe BlazePose and TensorFlow.js." https://blog.tensorflow.org/2021/05/high-fidelity-pose-tracking-with-mediapipe-blazepose-and-tfjs.html (accessed Aug. 11, 2021).