Accessibility, Assistive Technology and Digital Environments, Vol. 121, 2024, 141–151
https://doi.org/10.54941/ahfe1004622

Leveraging Human Data for Avatar Action Recognition in Virtual Environments

Somaya Eltanbouly and Osama Halabi

Department of Computer Science and Engineering, Qatar University, Doha, Qatar

ABSTRACT

In the dynamic metaverse landscape, avatars serve as digital embodiments, crucial in representing user interactions. This research investigates the dynamics of avatars within the metaverse, with a specific focus on their movements and actions in virtual environments. Real-world pose estimation techniques, notably OpenPose, are applied to establish a correlation between human and avatar movements. A Multilayer Perceptron model, trained on the NTU RGB+D dataset, proves highly effective in accurately classifying avatar actions, achieving an impressive accuracy rate exceeding 95%. The study explores the impact of different avatar types on action recognition, revealing notable performance differences. Notably, a half-body avatar with only the upper part demonstrates comparable performance to a full-body avatar, while the absence of head joints adversely affects accuracy. Additionally, the research assesses the generalizability of human data for avatar recognition, highlighting its superiority over models trained exclusively on avatar data. The findings underscore the adaptability of these techniques to diverse avatar configurations and offer promising advancements in action recognition within the ever-evolving metaverse. This research contributes valuable insights into effectively integrating real-world action recognition techniques to understand avatar behaviors in virtual spaces comprehensively.

Keywords: Avatar, Pose estimation, Action recognition, Metaverse, Virtual environments

INTRODUCTION

In the evolving metaverse world, avatars play an essential role in representing users in the virtual environment. Serving as digital bodies controlled by the user, avatars traverse the metaverse, assuming various shapes such as humanoid avatars, half-body avatars, cartoonish avatars, and more. Regardless of their realism, all these different avatar shapes are fully controlled by the user. As digital counterparts of users, avatars can express different actions and behaviors during interactions with other users or the surrounding virtual environment. Recognizing the significance of treating digital bodies akin to physical ones is crucial for the development of an intelligent and safer metaverse (Qayyum *et al.*, 2022). Therefore, this study focuses on recognizing avatar actions and poses within the dynamic metaverse by exploring the application of techniques developed for humans in the real world to avatars in the virtual world.

The potential impact of recognizing avatar movements and actions extends across diverse domains, enriching interactions and refining simulations. Applications in healthcare are evident through the use of 3D pose estimation in rehabilitation evaluations (Wu *et al.*, 2020) and virtual rehabilitation training settings (Wu *et al.*, 2019). The utilization of pose estimation for behavior analysis spans real-world applications (Kwan-Loo *et al.*, 2022) and immersive VR environments (Mannan *et al.*, 2023), suggesting a novel approach—virtual behavior analysis based on users' digital representations. Consequently, integrating a robust avatar action recognition system that combines 3D pose estimation and virtual training scenes holds transformative potential for healthcare, training simulations, behavior analysis, and immersive VR environments.

Human pose estimation aims to identify body part positions and construct a representation, such as a skeletal structure, using input data like images and videos. Techniques like AlphaPose (Fang *et al.*, 2023), DeepPose (Toshev and Szegedy, 2014), HRnet (Sun *et al.*, 2019), and OpenPose (Cao *et al.*, 2021) have shown high performance. OpenPose is widely used in research due to its usability and significant contributions to pose estimation tasks.

OpenPose finds applications in various domains. For instance, OpenPose is used to model non-verbal behaviors in cooperative learning (Eiji and Ozeki, 2019). It can also be used to explore in-store consumer behavior (Li *et al.*, 2020). Moreover, students' unrelated actions during online lectures can also be detected using OpenPose (Kawamata and Akakura, 2022). OpenPose extends its utility to detecting complex behaviors like violence and abuse (Chu and Wang, 2022; Chang and Liao, 2023; Huang *et al.*, 2023). Another usage is in sports-related research, in which OpenPose was used for performance learning in Yoga (Lin *et al.*, 2021) and as a support system for home-based squat training (Hirasawa *et al.*, 2020). Additionally, OpenPose can be used practically to build a posture analysis model for basketball free throws (Masato and Tsunoda, 2019). Additionally, OpenPose and other specialized techniques in human motion capture, avatar character animation, recordings, pose classification, and sequential moving pose analyses were used to enhance the understanding of human complexity by detecting avatar movements (Jeong, Xu and Miller, 2020).

In this study, we employ OpenPose for avatar pose estimation due to its proven effectiveness, as demonstrated in previous works. Our investigation focuses on recognizing avatar movements within the metaverse. Similar to the objectives outlined by previous work (Jeong, Xu and Miller, 2020), our study is investigating the avatar's actions. However, we address observed limitations in their implementations, including constraints related to training and testing data and the absence of evaluative results.

This paper presents a methodology for leveraging real-world action recognition techniques in avatar action recognition. Our research is divided into three main objectives. The first objective is to test the feasibility of detecting avatar actions in a virtual environment. The second objective is to study the impact of avatar types on the performance of action recognition and pose estimation. The last objective is to test the generalizability of human action data. In summary, the research will address the following questions:

- How effective are pose estimation and action recognition techniques when applied to avatars?
- What impact do avatar types with different joint configurations have on action recognition performance?
- Can models trained on human data effectively generalize to avatars, and how do they compare to avatar-trained models?

The subsequent sections of this paper are organized as follows: The Methodology section explains the research methodology, with the Pose Estimation section dedicated to the pose estimation part and the Action Recognition Section for the action recognition part. Subsequently, the evaluation is expounded upon in the Evaluation Section, including subsections discussing Dataset, Experimental Setup, and Experiments. The Experiments subsection contains details of the results for e, Impact of avatar type on action recognition (RQ2), Generalization of human data on avatar action (RQ3). Finally, our findings are summarized, and the paper is concluded in the Conclusion and Future Work section, discussing potential future directions.

METHODOLOGY

The system aims to recognize the actions and movements of avatars in virtual environments. This is achieved by employing real-world techniques commonly used for analogous tasks. The process begins by using the avatar video as input for the pose estimation model, followed by using the resulting pose as input for the action recognition model. An overview of the system is illustrated in Figure 1.

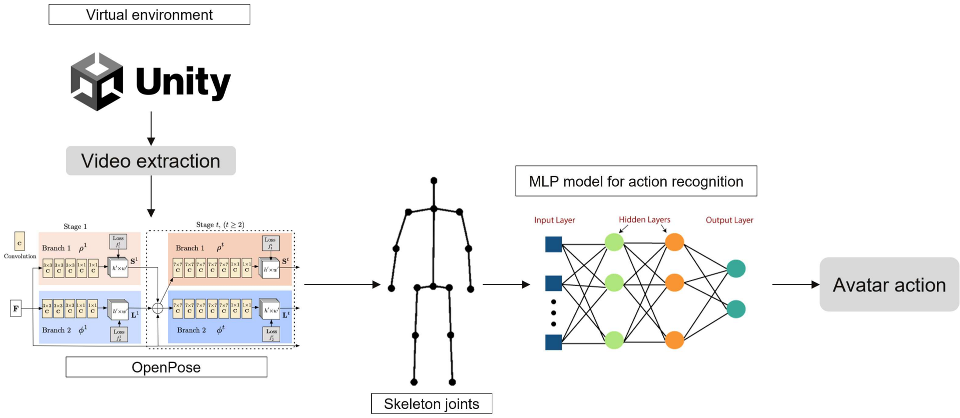

Figure 1: System overview.

Pose Estimation

To accurately estimate the pose of avatars, we strategically chose to leverage the OpenPose library. OpenPose is a real-time multi-person human pose detection library known for its high performance in identifying key points for the human body, feet, hands, and face in individual images. The decision to employ OpenPose is due to its proven effectiveness in diverse real-world scenarios, making it a reliable choice for our pose estimation task in virtual environments. OpenPose utilizes a robust and efficient algorithm that combines computer vision techniques to provide accurate joint estimation. Its ability to handle diverse poses aligns well with the complex movements exhibited by avatars in our virtual environment. Moreover, the library's real-time capabilities ensure that pose estimation can be performed seamlessly, which suits the dynamic nature of avatar actions.

We captured avatar movements using the Unity Engine Recorder to facilitate our experiments. This method allows us to collect realistic and diverse avatar data representing various actions. The recorded videos serve as input for the OpenPose model, which has been pre-trained on the COCO pose estimation dataset (Lin Tsung-Yi and Maire, 2014). The OpenPose model extracts a total of 36 values, representing x and y coordinates for each of the 18 joints in the avatar's pose. Utilizing OpenPose on videos from the virtual environment ensures accurate and dynamic pose estimation, enabling evaluation of the model's performance in a virtual environment closely mirroring real-world scenarios.

Action Recognition

In classifying actions performed by avatars, we utilized a multilayer perceptron (MLP) model. The decision to use an MLP stems from its ability to effectively model the complex relationships between joint positions in the avatar's pose data, which may exhibit nonlinear dependencies crucial for recognizing different actions. In our case, the spatial relationships between the different joints in the avatar's pose are intricate and nonlinear, making the power of an MLP beneficial for learning and capturing these nuances. Additionally, the simplicity and interpretability of MLPs make them well-suited for our classification of simple actions. By leveraging an MLP, we strike a balance between model complexity and performance, aligning with the nature of our action recognition problem.

The architecture details of the MLP are outlined as follows: three dense layers with ReLU activation functions, progressively decreasing the number of units to refine essential features. Batch normalization is applied after each dense layer to enhance training stability. The output layer with SoftMax activation facilitates probability-based classification of the avatar's actions into distinct classes. This model takes as input the skeleton data extracted from the Pose Estimation step and predicts the action class of the avatar.

EVALUATION

In this Section, we will provide details about the datasets employed for training and testing our model. This will be accompanied by an explanation of the system configurations, experimental procedures, and the outcomes of our system.

Dataset

For the training and validation of the Action Recognition model, the NTU RGB+D dataset (Shahroudy et al., 2016) was employed. This dataset encompasses a total of 60 classes, offering a diverse range of actions. However, owing to limitations in the avatar's movements, the system was exclusively trained and tested using data from two classes: jumping and standing. Multiple videos with known labels were recorded from the virtual environment for testing purposes. Avatar movements within the environment were created using the Ready Player Me platform[1], a platform for virtual reality (VR) avatar creation and customization. The primary testing involved the use of full-body avatars with multiple representations, as illustrated in Figure 2.

The data collected from the videos were processed through OpenPose to extract joints, each labeled accordingly, to be utilized in testing. The dataset comprises over 500,000 samples. However, integrating data from OpenPose, trained on the COCO dataset with 18 joints, into an action recognition model trained on the NTU RGB+D dataset with 25 joints posed a challenge. Consequently, the training data was modified to align with the representation of OpenPose, involving the removal of key points corresponding to most of the spine, hands, and legs.

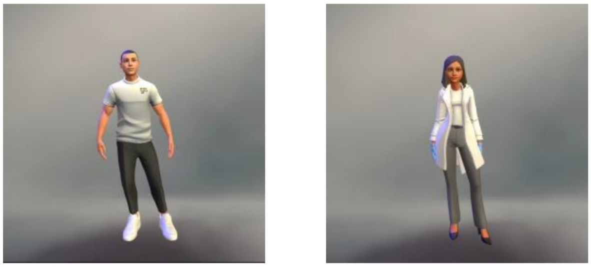

Figure 2: Examples of avatars created to perform actions in the virtual environment.

Experimental Setup

Our experiments were conducted on a system equipped with an Nvidia GeForce RTX 3070 GPU. For the implementation of the multilayer perceptron (MLP) in the action recognition task, we utilized TensorFlow with the Keras API. A TensorFlow-based version of the OpenPose library was integrated into our framework for pose estimation. In tuning the hyperparameters for the action recognition models, we adjusted the learning rate by varying it between 0.000001 and 0.1. We reduced the learning rate when there were no improvements for three epochs, using a decreasing factor of 0.5. The number of epochs was fixed at 100 for all models, with an early stopping mechanism if there was no improvement for five epochs. Additionally, the choice of optimization function was another hyperparameter we tested. Both Adam and SGD (Stochastic Gradient Descent) optimizers were examined to evaluate their effectiveness in enhancing model performance. This experimental setup allowed us to thoroughly assess the impact of various configurations

[1] https://readyplayer.me/hub

on the performance of both pose estimation and action recognition models, leading to a more informed analysis of our results.

Experiments

In this Section, we will present the experimental steps and the outcomes derived from our system corresponding to each research question.

Effectiveness of Recognizing Avatar's Actions (RQ1)

The experiment for this point was conducted by collecting the avatar movement data and testing it using the model trained on human data. We demonstrated its effectiveness with avatar data compared to human data. We also tested the model on real-time human actions. The results confirmed that the model trained on human data can effectively be used on avatars. The results were also similar to those obtained by testing on real-time human action data. The results in detail are shown in Table 1. The results demonstrate effectiveness in testing avatar data, with an F1 score of 0.95, which is even higher than testing on real-time human actions, primarily due to limitations on the recall score. Thus, we can conclude that avatar movements can be accurately detected and classified using techniques similar to those employed for human movements.

Table 1. Results of action recognition on avatar data and human data.

	Testing on avatar data	Testing on human data
Accuracy	0.955	0.996
Precision	0.950	0.998
Recall	0.962	0.780
F1-score	0.954	0.857

Impact of Avatar Type on Action Recognition (RQ2)

We divided the avatars into four main categories to test the impact of different avatar joint configurations. The first category is the full-body avatar, same as avatar represented in Figure 2; the second is the half-body avatar, typically lacking the lower body, as shown in Figure 3a; the third is an avatar without any face joints; and lastly, an avatar with only one head joint, these conditions can be seen in Figure 3b, as the avatar can either be used without any face joints, or by taking one joint that represents the head position. The full-body avatar has a total of 18 joints, the half-body avatar has 12 joints, the avatar without face joints has 13 joints, and the avatar with one head joint has 14 joints. The human data obtained from the NTU RGB+D dataset was altered accordingly to match the avatar joint categories, and different models were trained for each joint configuration.

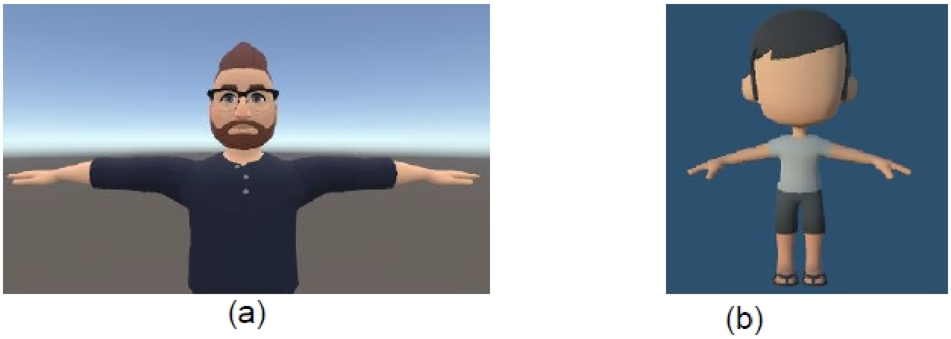

Figure 3: Avatars with different joint configurations, (a)[2] represents an example of the half-body avatars, and (b)[3] represents an example of avatars without face details.

Figure 4 illustrates the differences in performance among the various joint configurations. The half-body avatar was less affected, performing similarly to the full-body avatar with only a 4% difference. Meanwhile, the other two configurations experienced more than a 20% decrease in performance, with slight improvement using only one head joint. The detailed results are shown in Table 2. These findings indicate that leg joints, such as the hip and knee, do not significantly impact action classification for avatars. On the other hand, head joints are shown to be more critical in avatar action classification. However, overall, the performance for all configurations demonstrates good classification of avatar actions, suggesting that avatar action classification can be applied in any virtual environment with diverse types of avatars.

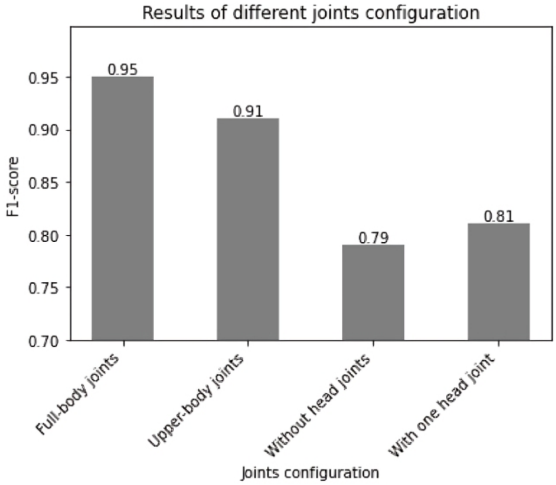

Figure 4: The performance of action recognition with different avatar joint configurations.

[2] https://assetstore.unity.com/packages/tools/integration/meta-avatars-sdk-271958

[3] https://assetstore.unity.com/packages/3d/characters/humanoids/mini-modular-character-free-demo-256389

Table 2. Results of different avatar joint configurations.

	Full-body	Upper-body	Without face joints	With one face joint
Accuracy	0.955	0.912	0.792	0.805
Precision	0.950	0.907	0.831	0.822
Recall	0.962	0.919	0.824	0.828
F1score	0.954	0.910	0.791	0.805

Generalization of Human Data on Avatar Action (RQ3)

In our paper, we utilized the human actions dataset to generalize the action recognition model for different avatar representations. This approach differs from the previous work (Jeong, Xu and Miller, 2020), where researchers trained their model on avatar data but did not report any results for their methodology. To assess the generalization ability of both human-trained and avatar-trained models, we trained an additional model with avatar data and compared the performance of both types of training. We tested both models on a total of three avatar representations; one was the same as the avatar-trained model, and the other two were different.

The avatar-trained model demonstrated very high performance when tested on the same avatar data, with almost zero errors, outperforming the human-trained model. However, it exhibited very low generalizability when tested on the other two avatars, while the human-trained model showed high performance for both. This establishes the superiority of the human-trained model in terms of generalization to different shapes with varying skeleton scales over the avatar-trained model. Detailed results for the three types of avatars for both models are presented in Table 3. The human-trained model is denoted as M1 in the table, while the avatar-trained model is denoted as M2.

The first avatar type is the ReadyPlayerMe avatar, on which M2 is trained, explaining the model's high performance on this type. The second[4] and third[5] types of avatars are imported from the Unity Asset store. The models used for this experiment lack head joints to ensure usability for the three different types of avatars. The human-trained model exhibited an average F1-score of 0.853, surpassing the average of the avatar-trained model by more than 20%. From these results, we can conclude that training the model with human data provides an opportunity to detect actions in a broader range of avatars, benefiting from the diversity present in the human-collected dataset.

[4] https://assetstore.unity.com/packages/3d/characters/humanoids/characterpack-free-sample-79870

[5] https://assetstore.unity.com/packages/3d/animations/basic-motions-free154271

Table 3. Results comparison between human-trained model and avatar-trained model.

Testing set	Model	Precision	Recall	F1-score	Accuracy
Avatar type 1	M1	0.831	0.824	0.791	0.792
	M2	0.996	0.997	0.996	0.996
Avatar type 2	M1	0.818	0.799	0.807	0.838
	M2	0.370	0.496	0.244	0.317
Avatar type 3	M1	0.936	0.930	0.928	0.928
	M2	0.257	0.5	0.339	0.514
		Average	M1	0.842	0.853
			M2	0.526	0.609

CONCLUSION AND FUTURE WORK

In conclusion, this research explores the dynamic metaverse landscape, emphasizing the pivotal role of avatars in representing user interactions. The study uses real-world pose estimation techniques to establish a correlation between human and avatar movements. The investigation unfolds in three key dimensions: assessing the feasibility of employing human-centric techniques for avatar actions, evaluating the impact of diverse avatar configurations on action recognition, and examining the generalizability of human-trained models to avatars. Results reveal the effectiveness and generalizability of human-trained models in recognizing avatar actions. Additionally, different avatar joint configurations showcase varying impacts on performance. This research contributes valuable insights, demonstrating the adaptability of real-world techniques to metaverse scenarios. The findings promise advancements in action recognition to aid in understanding and effectively leveraging avatar behaviors, which will be crucial for safer and context-aware virtual interactions as the metaverse evolves.

Future work holds promising avenues for refining action recognition techniques in the metaverse. Implementing specialized models like ST-GCN tailored for skeleton action recognition and conducting detailed analyses on the influence of individual joints on specific actions will deepen our understanding of avatar movements. Exploring alternative pose estimation models, expanding the range of training classes, and experimenting with diverse avatar representations are crucial steps in advancing the adaptability and generalizability of our approach. This ongoing evolution is poised to significantly enhance action recognition methodologies, fostering safer and more context-aware virtual interactions in the dynamic metaverse realm.

ACKNOWLEDGMENT

This work was made possible by GSRA grant GSRA9L-2-0605-22108 from the Qatar National Research Fund (a member of Qatar Foundation) and by NPRP14C-0916-210015-SP-470 from the Qatar National Research Fund (a member of Qatar Foundation). The findings herein reflect the work and are solely the authors' responsibility.

REFERENCES

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. and Sheikh, Y. (2021) 'OpenPose: Real-time Multi-Person 2D Pose Estimation Using Part Affinity Fields', *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1), pp. 172–186. Available at: https://doi.org/10.1109/TPAMI.2019.2929257.

Chang, K.-C. and Liao, Y.-C. (2023) 'Design of Violence Event Detection System Based on CCTVs by Human Body Pose Recognition', in *2023 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, pp. 695–696. Available at: https://doi.org/10.1109/ICCE-Taiwan58799.2023.10226669.

Chu, S.-W. and Wang, C.-M. (2022) 'Combining OpenPose with BiLSTM for Violence Detection in Long-Term Care', in *2022 IEEE/ACIS 23rd International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, pp. 212–215. Available at: https://doi.org/10.1109/SNPD54884.2022.10051807.

Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L. and Lu, C. (2023) 'AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-time', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6), pp. 7157–7173. Available at: https://doi.org/10.1109/TPAMI.2022.3222784.

Hirasawa, Y., Gotoda, N., Kanda, R., Hirata, K. and Akagi, R. (2020) 'Promotion System for Home-Based Squat Training Using OpenPose', in *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pp. 984–986. Available at: https://doi.org/10.1109/TALE48869.2020.9368366.

Huang, Y., Zhang, Z., Zhao, Z., Chen, Z., Zeng, K. and Li, Y. (2023) 'Identification of Child Physical Abuse Based on Openpose and ST-GCN', in *2023 8th International Conference on Image, Vision and Computing (ICIVC)*, pp. 682–687. Available at: https://doi.org/10.1109/ICIVC58118.2023.10270667.

Jeong, D., Xu, J. and Miller, L. (2020) 'Inverse Kinematics and Temporal Convolutional Networks for Sequential Pose Analysis in VR', in. Available at: https://doi.org/10.1109/AIVR50618.2020.00056.

Kawamata, T. and Akakura, T. (2022) 'Automatic Evaluation of Learning Behaviors for Online Lectures by OpenPose', in *2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech)*, pp. 384–385. Available at: https://doi.org/10.1109/LifeTech53646.2022.9754894.

Kwan-Loo, K. B., Ortíz-Bayliss, J. C., Conant-Pablos, S. E., Terashima-Marín, H. and Rad, P. (2022) 'Detection of Violent Behavior Using Neural Networks and Pose Estimation', *IEEE Access*, 10, pp. 86339–86352. Available at: https://doi.org/10.1109/ACCESS.2022.3198985.

Li, J., Zeng, J., Hou, K., Zhou, J. and Wang, R. (2020) 'Application of Openpose algorithm to detect consumer behavior in store', in, pp. 317–322. Available at: https://doi.org/10.24264/icams-2020.III.11.

Lin, C.-H., Shen, S.-W., Anggraini, I. T., Funabiki, N. and Fan, C.-P. (2021) 'An OpenPose-Based Exercise and Performance Learning Assistant Design for Self-Practice Yoga', in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, pp. 456–457. Available at: https://doi.org/10.1109/GCCE53005.2021.9621907.

Lin Tsung-Yi and Maire, M. and BS and HJ and PP and RD and DP and ZCL (2014) 'Microsoft COCO: Common Objects in Context', in T. and SB and T. T. Fleet David and Pajdla (ed.) *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, pp. 740–755.

Mannan, F. A., Porffy, L. A., Joyce, D. W., Shergill, SS and Celiktutan, O. (2023) ‘Automatic Detection of Cognitive Impairment with Virtual Reality’, *Sensors*, 23(2). Available at: https://doi.org/10.3390/s23021026.

Nakai Masato and Tsunoda, Y. and H. H. and M. H. (2019) ‘Prediction of Basketball Free Throw Shooting by OpenPose’, in M. and M. K. and S. K. Kojima Kazuhiro and Sakamoto (ed.) *New Frontiers in Artificial Intelligence*. Cham: Springer International Publishing, pp. 435–446.

Qayyum, A., Butt, M. A., Ali, H., Usman, M., Halabi, O., Al-Fuqaha, A., Abbasi, Q. H., Imran, M. A. and Qadir, J. (2022) ‘Secure and Trustworthy Artificial Intelligence-Extended Reality (AI-XR) for Metaverses’, *ACM Computing Surveys* [Preprint]. Available at: https://doi.org/10.1145/3614426.

Shahroudy, A., Liu, J., Ng, T.-T. and Wang, G. (2016) ‘NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis’, *CoRR*, abs/1604.02808. Available at: https://arxiv.org/abs/1604.02808.

Sun, K., Xiao, B., Liu, D. and Wang, J. (2019) ‘Deep High-Resolution Representation Learning for Human Pose Estimation’, *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5686–5696. Available at: https://api.semanticscholar.org/CorpusID:67856425.

Toshev, A. and Szegedy, C. (2014) ‘DeepPose: Human Pose Estimation via Deep Neural Networks’, in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1653–1660. Available at: https://doi.org/10.1109/CVPR.2014.214.

Watanabe Eiji and Ozeki, T. and K. T. (2019) ‘Modeling of Non-verbal Behaviors of Students in Cooperative Learning by Using OpenPose’, in H. and C. I.-A. and T. H. and I. S. and H. U. Nakanishi Hideyuki and Egi (ed.) *Collaboration Technologies and Social Computing*. Cham: Springer International Publishing, pp. 191–201.

Wu, Q., Xu, G., Zhang, S., Li, Y. and Wei, F. (2020) ‘Human 3D pose estimation in a lying position by RGB-D images for medical diagnosis and rehabilitation’, in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 5802–5805. Available at: https://doi.org/10.1109/EMBC44109.2020.9176407.

Wu, X., Liu, H., Zhang, J. and Chen, W. (2019) ‘Virtual reality training system for upper limb rehabilitation’, in *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1969–1974. Available at: https://doi.org/10.1109/ICIEA.2019.8834288.