Towards a Markerless 3D Pose Estimation Tool

Amaan Rahman
amaan.rahman@cooper.edu
M.Eng. and B.E. in Electrical Engineering '23, The Cooper Union
USA

ABSTRACT

Evaluation of exoskeleton performance benefits from standards to verify proper functionality and safety. Currently, there are limited evaluation methods for exoskeletons. Measurement methods to evaluate human-exoskeleton kinematics include optical tracking systems (OTS) and inertial measurement units (IMUs). However, OTS and IMUs can be intrusive, requiring the attachment of markers or sensors. This research focuses on investigating markerless 3D pose estimation algorithms with low-cost red, green, blue (RGB) cameras to determine their viability as methods for tracking human joint positions and deriving skeletal frame orientations. We present a tool that utilizes state-of-the-art 3D pose estimation algorithms to generate 3D pose estimation data. Future experiments will be performed to evaluate the viability of 3D pose estimation algorithms as markerless methods for joint position and orientation estimation.

CCS CONCEPTS

• **Human-centered computing** → **User models**; • **Computing methodologies** → **Object detection**; • **Software and its engineering** → *Application specific development environments.*

KEYWORDS

exoskeleton/human kinematics, monocular computer vision, 3D human pose estimation, deep neural networks

ACM Reference Format:
Amaan Rahman. 2023. Towards a Markerless 3D Pose Estimation Tool. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3544549.3583950

1 INTRODUCTION

Wearable robotics are manufactured for the purpose of physical load reduction on the subject's body [11]. In an industrial setting, exoskeletons are utilized to reduce work fatigue and provide heavy load assistance. However, limited exoskeleton standards and certifications cause difficulty in adopting wearable devices in industrial settings [20].

The current technologies for quantifying human-exoskeleton performance include optical tracking systems (OTS) and inertial measurement units (IMUs). The primary benefit of marker-based OTS is high precision results[27]. However, marker-based methods suffer from uncertainties such as erroneous placement and movement, and especially for exoskeleton evaluations, physical constraints in marker placement [10]. The OTS and IMUs require adhering markers or sensors on the subject. Exoskeletons introduce physical limitations to marker-based technologies such as an increase in constraint for marker placements due to physical obstruction by the exoskeleton. Given that the subject is equipped with an exoskeleton, variations in marker and sensor placement, to avoid occlusions and to adjust to fit with the exoskeleton frame, can both introduce additional systematic and random errors[22]. Similar to OTS, IMUs are also susceptible to soft tissue artifacts[22]. In addition, IMUs are limited in precision due to drift, interference, and variations in the rotational reference frames[9].

Although marker-based technologies are a standard for evaluating human biomechanics, markerless methods have the potential of measuring 3D data of the subject [8] through a non-intrusive means of extracting joint positions and orientation data from the subject and can result in less limiting and more cost-effective experimental setups. We propose the investigation of markerless methods, specifically 3D monocular pose estimation algorithms. This study investigates whether low-cost RGB cameras can be utilized as the primary medium of measurement through the application of 3D pose estimation algorithms. This research introduces a measurement tool comprised of state-of-the-art 3D monocular pose estimation algorithms for evaluating markerless methods for human-kinematic evaluations, with the OTS measurement method as a baseline.

2 BACKGROUND

OTS and IMUs have been used for clinical evaluation of human biomechanics [1, 4, 15, 22, 27]. OTS yield Mean Per Joint Position Error (MPJPE) less than 0.3 mm [3, 24] and between $1°$ and $6°$ approximately [7, 27] in joint angle error for human kinematic analysis, specifically for the knee joint. OTS have been applied towards synchronous human-exoskeleton pose tracking as a potential measurement method for exoskeleton performance [7]. However, the measurement quality is limited to accurate and precise placement of anatomical landmark markers, artifacts, and sensors for high-precision joint position and orientation estimation.

Markerless methods introduce a non-intrusive and relatively low-cost solution to pose tracking, but have limited validation of precision and accuracy [10]. One prominent markerless method is pose estimation with deep learning techniques. This pose estimation algorithm's building blocks comprise of convolution neural networks (CNNs), which extract learned feature sets given multi-dimensional data, such as images, for classification or regression [17]. CNNs are basic building blocks utilized for estimating the pose or positional information of a particular object, or in the case of this research, human joints [26]. Different algorithms have innovative frameworks that achieve increased positional precision and computational efficiency.

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3583950

Practiced markerless methods applying pose estimation algorithms, in literature, require depth-sensing camera systems or active camera systems [10]; therefore, a hardware constraint is generated. Although, depth-sensing or active cameras can be relatively low-cost compared to marker based systems, a more cost effective markerless method applies pose estimation algorithms with low-cost RGB cameras in a monocular configuration.

Recently, monocular 3D pose estimation algorithms have yielded promising results that are competitive with stereophotogrammetric algorithms (utilizing multiple cameras, each with 2D pose estimation algorithms to estimate 3D pose). This research applies three monocular 3D pose estimation algorithms, GAST-Net[18], VIBE [16], and Blazepose [5], for an alternative markerless method of pose evaluations.[1]

3 METHODOLOGY

3D monocular pose estimation models were selected based on contributions in dynamic joint detection and tracking while limiting the effects of occlusions and optimizing computational efficiency. Furthermore, algorithms were selected based on the Mean Per Joint Position Error (MPJPE) metric.

The selected models have been configured into an open-source[2] tool to execute repeatable tests on input human locomotion video sequences. The tool was designed to easily integrate state-of-the-art models and to generate analysis experiments to gain insights about human kinematic studies, which motivates exoskeleton kinematic studies. Integration is motivated from Continuous Integration (CI) standards in industry regarding software management to create a modularized pipeline framework. The merit of such a framework is the ease in motivation of future kinematic studies.

3.1 Models

3.1.1 ***GAST-Net.*** 3D pose estimation algorithms can be significantly limited by joint occlusions. GAST-Net is a model that proposes a novel method of recognizing spatial patterns in the spatio-temporal domain by modeling local and global spatial information with convolution and graph attention mechanisms, while recognizing temporal features.

The pipeline of the algorithm is initiated with human detection to capture the bounding box of detected humans utilizing a high-end object detection algorithm, YOLOv3 [23], which is then fed into a 2D pose estimation algorithm, HRNet [25]. Given 2D pose estimation data, GAST-Net utilizes temporal convolution networks (TCNs) to capture features over long temporal sequences. Spatial information is estimated through a local spatial attention network and a global spatial attention network. The local spatial attention network models the kinematic structure of the pose. The global spatial attention network encodes non-local joint relationships to control depth ambiguities and limb occlusions.

The average MPJPE over the Human3.6M dataset (large dataset of human locomotions with various viewpoints) [13] is documented to be about 44 mm.

3.1.2 ***VIBE.*** The VIBE 3D pose estimation algorithm addresses the occlusion problem by generating natural temporal pose sequences, mitigating irregular or unnatural human locomotion. Instead of generating a skeletal pose structure, a 3D pose mesh is generated called Skinned Multi-Person Linear model (SMPL) [19].

A pretrained CNN is utilized to generate 2D poses given a frame, which is then fed into a gated recurrent neural network (GRN). The GRN enables utilization of past pose estimation data to constrain the pose in future frames. The output is latent feature data utilized to regress SMPL body parameters through an iterative 3D regression model analogous to Human Mesh Recovery (HMR), an end-to-end algorithm to estimate 3D joint data [14]. However, constraining poses in future frames do not address discontinuous temporal flow of estimated poses. To increase continuity to generate natural estimated pose locomotion, a discriminator network is utilized to penalize improbable poses or motions compared to a ground-truth 3D motion capture dataset called Archive of Motion Capture As Surface Shapes (AMASS)[21].

Compared to GAST-Net, the average MPJPE is observed to be greater (about 66 mm) [16]. However, VIBE can be leveraged to yield reproducible results due to stability improvements of pose data.

3.1.3 ***Blazepose.*** Another key issue with 3D pose estimation algorithms is computational inefficiency. Blazepose introduces an architecture yielding improvements in computational efficiency [5].

Blazepose architecture starts off with a person detector motivated by BlazeFace [6]. The architecture is analogous to a convolutional block structure in ResNets [12], which reduces depth in overall neural network architecture yielding faster training time and improved accuracy. BlazeFace has BlazeBlocks [6] comprised of depth-wise convolutions with a skip connection taking the input, processing it through maximum pooling and channel dimension pooling, to the output. Blazepose utilizes the aforementioned structure for human detection and estimates alignment parameters in association to the structure of the Vitruvian man [5].

After human or pose detection, the pose tracker is triggered. The pose tracker utilizes heatmap and offset loss for supervised training of the lightweight embedded network. The heatmap and offset loss aspect of the network are removed during inference. The regression module utilizes the lightweight embedding to estimate the 3D pose of each joint. Throughout the network, skip connections are utilized for high-level and low-level feature balance.

Instead of the MPJPE, Blazepose utilizes PCK@0.2, Percent Correct Keypoints evaluated at a threshold of 0.2 times the torso diameter. Blazepose achieved about 84% PCK@0.2, which is about 1% increase for about 10 times increase in output Frames per Second (FPS) [5].

3.2 Measurement Tool

The measurement tool utilizes GAST-Net, VIBE, and Blazepose to generate 3D pose estimated data to evaluate each algorithm against the ground-truth data to estimate the measurement uncertainty for each method. The tool is comprised of two parts: pose data generation and evaluation. The experiment conducted with

[1]Certain commercial products or software are identified here to describe our study adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the products or names identified are necessarily the best available for the purpose.
[2]https://github.com/amaan4152/3DPoseEvaluator

the measurement tool consists of one subject for demonstration purposes.

3.2.1 ***Data Generation***. An input video sequence is processed by a specified model, and raw pose data is generated. Joints comprising two adjacent limbs are selected to extract estimated 3D pose data, joint angle between limbs is computed, and the orientation of each joint is computed as a quaternion. Given a scenario of computing the joint angle between the femur, $\mathbf{v_{fem}}$, and the tibia, $\mathbf{v_{tib}}$, the joint angle, θ, is computed as follows:

$$\theta = arccos\,(\langle \mathbf{v_{fem}},\, \mathbf{v_{tib}} \rangle) \tag{1}$$

Where $\langle \cdot \rangle$ represents the inner product operator. The orientation of each joint is computed through axis-angle representation of quaternions. The cross product of $\mathbf{v_{fem}}$ and $\mathbf{v_{tib}}$, is the axis of rotation in 3D space, $\mathbf{u}$, and the angle, θ, represents the angle of rotation. For valid representation of 3D orientation, the quaternion must be normalized; normalized axis of rotation, $\hat{\mathbf{u}}$, yields a normalized quaternion. Given the aforementioned scenario, the orientation of each joint is computed as follows:

$$\hat{\mathbf{q}} = e^{\frac{\theta}{2}\hat{\mathbf{u}}} \tag{2}$$

3.2.2 ***Evaluation***. Given pose estimation data and ground truth data, the pose estimation data must be temporally aligned and spatially registered with the ground truth data for proper evaluation. Since the pose estimation data is sampled at 60 FPS and the ground truth data is sampled at 120 FPS, downsampling the ground truth data by half yields optimal temporal alignment, as can be seen in **Figure 2**.

Spatial registration is achieved with a Least-Squares (LS) method of fitting, called the Arun method [2].

Given $\mathbf{X} \in \mathbb{R}^{3\times N} = \{\mathbf{x}_i\}$ (pose estimation data points) and $\mathbf{Y} \in \mathbb{R}^{3\times N} = \{\mathbf{y}_i\}$ (ground truth data points), where N is the number of points, the LS solution is $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{T}}$, which can be used to register $\mathbf{X}$ to $\mathbf{Y}$. Thus, the spatial registration algorithm is as follows:

$$\mathbf{X}' = \mathbf{x}_i - \mathbb{E}[\mathbf{x}_i],\ \ i = 1, 2, \ldots, N \tag{3}$$

$$\mathbf{Y}' = \mathbf{y}_i - \mathbb{E}[\mathbf{y}_i],\ \ i = 1, 2, \ldots, N \tag{4}$$

$$H = \mathbf{X}'\mathbf{Y}'^{T} = U\Sigma V^{T} \tag{5}$$

$$\tilde{\mathbf{R}} = UV^{T} \tag{6}$$

$$\tilde{\mathbf{T}} = \mathbb{E}[\mathbf{y}_i] - \tilde{\mathbf{R}}\mathbb{E}[\mathbf{x}_i],\ \ i = 1, 2, \ldots, N \tag{7}$$

$$\tilde{\mathbf{X}} = \tilde{\mathbf{R}}\mathbf{X} + \tilde{\mathbf{T}} \tag{8}$$

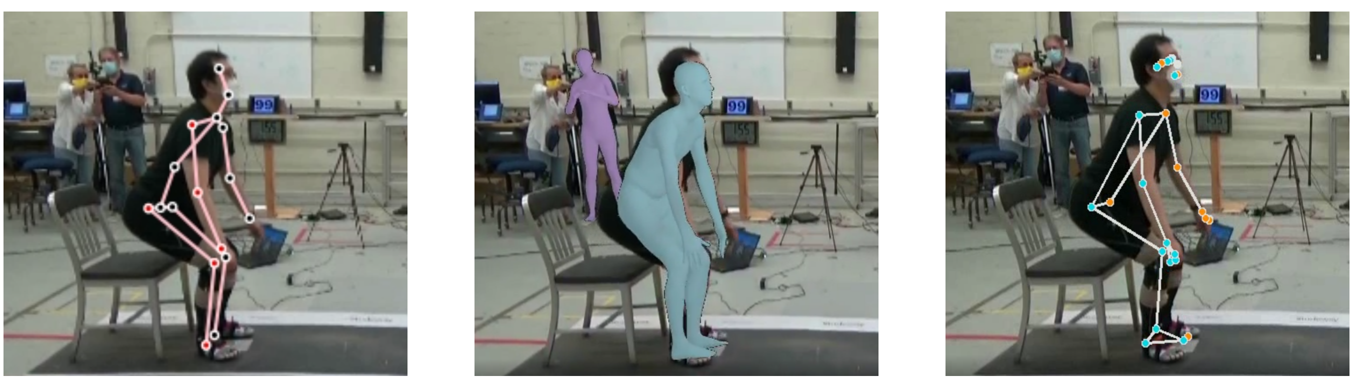
Here, $\mathbb{E}[\,\cdot\,]$ is the expectation operator and T represents the real transpose of a matrix. The estimated and ground-truth points are normalized with respect to the mean, and the singular value decomposition (SVD) of their outer product is utilized to compute the rotation matrix. The rotation matrix can then be applied using Eq. 8 to compute the translation matrix. Thus, the rotation matrix and translation matrix can be utilized to map the estimated points to the ground-truth data coordinates, given a video of human locomotion for analysis and a specified range of frame numbers.

Provided that the pose estimation data has been registered to the ground truth data, two evaluation metrics to determine the effectiveness of 3D pose estimation data are computed: Mean Per Joint Position Error (MPJPE) and Percent Detected Joints (PDJ). MPJPE is computed as the mean Euclidean distance between ground truth data and pose estimation data across all joints, and PDJ is computed as the percentage of joints that have Euclidean distance between ground truth data and pose estimation data less than 0.2 of the true torso diameter of the subject.

4 RESULTS

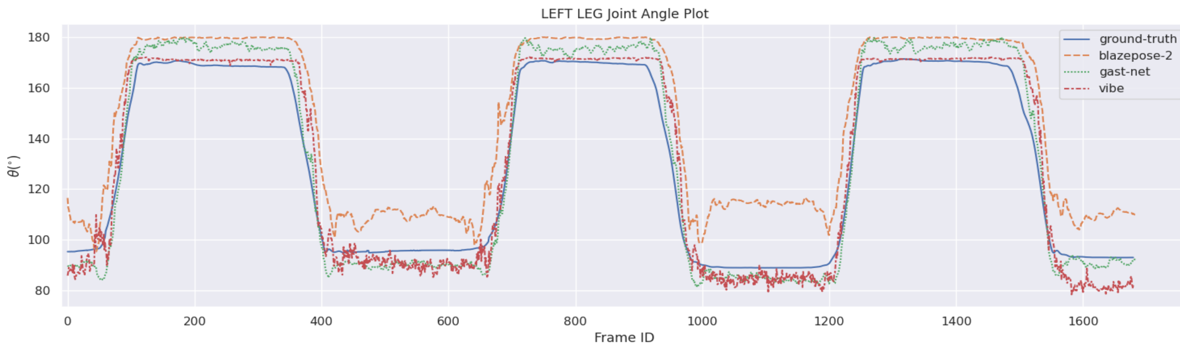
The measurement tool was tested[3] on sample side-view video sequence of three sit-stand-sit motions of a single individual as can be seen in **Figure 1**. All tests where executed in CPU mode with

[3]The datasets used to evaluate the software tool were collected in accordance with the Institutional Review Board (IRB) at the National Institute of Standards and Technology (NIST).

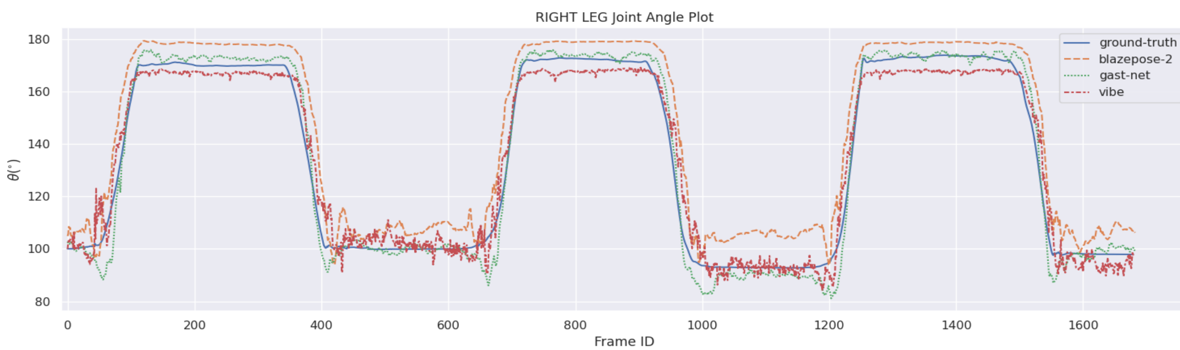


(a) (b) (c)

Figure 1: Pose visualization of three algorithms at a particular frame of a video sequence of three sit-stand motions: (a) GAST-Net pose visualization: The pose does not suffer from occlusion and all keypoints are visible. The evolution of the pose as the number of frames increased displayed minimal jitter. (b) VIBE pose visualization: The SMPL mesh structure applies a constraint on the pose, thus detects key points even with occlusions. The pose exhibited the most jitter of all three algorithms as the number of frames increased. (c) Blazepose pose visualization: Missed detection of left leg keypoints. Blazepose suffers from occlusion, but yields the most stable pose as the number of frames increase.



(a)



(b)

Figure 2: Plot of left (a) and right (b) knee joint angle between right tibia and right femur generated by GAST-Net, VIBE, and Blazepose compared against ground-truth data. Joint angle data represents three full sit-stand-sit motions of 1683 frames. As the subject rises from sitting to standing position, the knee joint angle rises during the extension phase, whereas the falling edge indicates knee flexion from standing to sitting position. The ground-truth data is sampled at 120 FPS, whereas the model data is sampled at roughly 60 FPS; the ground-truth data is therefore downsampled by half for proper data alignment.

Model	Kinematic Chain	MPJPE (mm)	PDJ (%)	Execution Time (s)
Blazepose Full	Right Leg	155.94	100	76.90
	Left Leg	162.83	100	76.08
Blazepose Heavy	Right Leg	147.02	100	254.64
	Left Leg	157.79	100	269.92
VIBE	Right Leg	50.69	100	3786.66
	Left Leg	51.41	100	3791.76
GAST-Net	Right Leg	35.89	100	3363.50
	Left Leg	37.95	100	3357.21

Table 1: Left and right leg evaluation metrics summary for each model for video sequence of three full sit-stand-sit motions of 1683 frames.

the following hardware specifications: 64-bit Ubuntu 20.04.4 LTS OS, Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz CPU, 16 GB RAM, 1.3 TB, NVIDIA GeForce RTX 2070 GPU. Each model required sufficient amount of shared memory in order to be operational. Blazpose required at most 1 GB, and GAST-Net and VIBE required at most 10 GB.

Furthermore, models such as Blazepose and VIBE have parameters that determine performance. Blazepose parameters (min_detection_confidence, min_tracking_confidence, model_complexity) where altered from `(0.5, 0.5, 1)` to `(0.75, 0.95, 1/2)`, respectively. The implication of raising min_detection_confidence and min_tracking_confidence is the increase in robustness levels for the cost of latency. The parameter model_complexity was tested as either `1/2`, where model_complexity=1 is Blazepose Full and model_complexity=2 is

Blazepose Heavy; higher the model_complexity, the greater the accuracy and latency of the model.

VIBE has parameters (tracker_batch_size, vibe_batch_size) with default values (12, 450), respectively. The former parameter is set for batch processing size for the bounding box tracker, and the latter parameter is set for the batch processing size for the VIBE model. The parameters where reduced to (1, 64) to drastically reduce shared memory requirements.

The kinematic chains of interest were the left and right legs of the dataset in **Figure 1** because it has bone markers at the hip, knee, and ankle regions for the left and right leg. According the **Figure 2**, the joint angles of the left and right legs are computed by each model and the ground truth. Based on **Figure 1**, the left leg is observed to be occluded, however all models detect joints successfully, which can be measured as PDJ as seen in **Table 1**. Although Blazepose is observed to not detect the left leg keypoints according to **Figure 1**, the pose data contains no null points.

Based on **Figure 2** and **Table 1**, GAST-Net performs the best based on joint angle tracking and joint position error relative to ground truth data. Furthermore, the model exhibits minimal jitter relative to VIBE, which yields competitive results as well. However, VIBE performs better tracking the left-leg joint angle than GAST-Net when the subject is standing. VIBE can be seen to exhibit greater jitter when the subject is sitting down compared to GAST-Net and Blazepose. Although Blazepose is observed to have the largest MPJPE errors and poorer joint angle tracking, the execution time or latency is significantly less compared to VIBE and GAST-Net as low as 2 orders of magnitude.

5 CONCLUSION

3D monocular pose estimation algorithms are promising non-intrusive low-cost markerless methods for evaluating human kinematics. We have developed a tool to apply 3D pose estimation algorithms to estimate the measurement uncertainty of 3D pose estimation algorithms compared to ground truth data. The measurement tool is intended to aid in efficient non-intrusive evaluations of human kinematics using a low-cost monocular system. Future experiments are needed for validating whether 3D pose estimation algorithms are a viable markerless method in comparison with marker-based methods. Additional evaluation of the measurement tool in human kinematic experiments such as gait or hurdle tests, can be conducted.

ACKNOWLEDGMENTS

Many thinks to my mentors Dr. Mili Shah, Department Chair of Mathematics at The Cooper Union, and Ya-Shian Li-Baboud, who is also affiliated with NIST, for guidance. Special thanks to the NIST Associates Roger Bostelman and Soocheol Yoon for developing and executing the exoskeleton test methods as well as Kam Saidi, Soocheol Yoon, Joe Chalfoun, and Peter Bajcsy for their review and feedback.

REFERENCES

[1] Mohammad Al-Amri, Kevin Nicholas, Kate Button, Valerie Sparkes, Liba Sheeran, and Jennifer L Davies. 2018. Inertial Measurement Units for Clinical Movement Analysis: Reliability and Concurrent Validity. *Sensors (Basel, Switzerland)* 18, 3 (Feb. 2018), 719. https://doi.org/10.3390/s18030719

[2] K. S. Arun, T. S. Huang, and S. D. Blostein. 1987. Least-Squares Fitting of Two 3-D Point Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-9, 5 (Sept. 1987), 698–700. https://doi.org/10.1109/TPAMI.1987.4767965

[3] Alexander Aurand, Jonathan Dufour, and William Marras. 2017. Accuracy map of an optical motion capture system with 42 or 21 cameras in a large measurement volume. *Journal of Biomechanics* 58 (05 2017). https://doi.org/10.1016/j.jbiomech.2017.05.006

[4] Christopher A. Bailey, Thomas K. Uchida, Julie Nantel, and Ryan B. Graham. 2021. Validity and Sensitivity of an Inertial Measurement Unit-Driven Biomechanical Model of Motor Variability for Gait. *Sensors* 21, 22 (Jan. 2021), 7690. https://doi.org/10.3390/s21227690

[5] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. *arXiv:2006.10204 [cs]* (June 2020). http://arxiv.org/abs/2006.10204 arXiv: 2006.10204.

[6] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. 2019. BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs. *arXiv:1907.05047 [cs]* (July 2019). arXiv:1907.05047 [cs]

[7] Roger Bostelman, Ya-Shian Li-Baboud, Karl Van Wyk, and Mili Shah. 2020. *Development of a Kinematic Measurement Method for Knee Exoskeleton Fit to a User.* Technical Report. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.TN.2107

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *arXiv:1812.08008 [cs]* (May 2019). arXiv:1812.08008 [cs]

[9] Yawen Chen, Chenglong Fu, Winnie Suk Wai Leung, and Ling Shi. 2020. Drift-free and self-aligned IMU-based human gait tracking system with augmented precision and robustness. *IEEE Robotics and Automation Letters* 5, 3 (2020), 4671–4678.

[10] Steffi L. Colyer, Murray Evans, Darren P. Cosker, and Aki I. T. Salo. 2018. A Review of the Evolution of Vision-Based Motion Analysis and the Integration of Advanced Computer Vision Methods Towards Developing a Markerless System. *Sports Medicine - Open* 4 (June 2018), 24. https://doi.org/10.1186/s40798-018-0139-y

[11] Michiel P. de Looze, Tim Bosch, Frank Krause, Konrad S. Stadler, and Leonard W. O'Sullivan. 2016. Exoskeletons for Industrial Application and Their Potential Effects on Physical Work Load. *Ergonomics* 59, 5 (May 2016), 671–681. https://doi.org/10.1080/00140139.2015.1081988

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]* (Dec. 2015). arXiv:1512.03385 [cs]

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (July 2014), 1325–1339. https://doi.org/10.1109/TPAMI.2013.248

[14] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-End Recovery of Human Shape and Pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, Salt Lake City, UT, 7122–7131. https://doi.org/10.1109/CVPR.2018.00744

[15] Dylan Kobsar, Jesse M. Charlton, Calvin T.F. Tse, Jean-Francois Esculier, Angelo Graffos, Natasha M. Krowchuk, Daniel Thatcher, and Michael A. Hunt. 2020. Validity and Reliability of Wearable Inertial Sensors in Healthy Adult Walking: A Systematic Review and Meta-Analysis. *Journal of NeuroEngineering and Rehabilitation* 17, 1 (Dec. 2020), 62. https://doi.org/10.1186/s12984-020-00685-3

[16] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

[17] Yann LeCun, Koray Kavukcuoglu, and Clement Farabet. 2010. Convolutional Networks and Applications in Vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE, Paris, France, 253–256. https://doi.org/10.1109/ISCAS.2010.5537907

[18] Junfa Liu, Juan Rojas, Zhijun Liang, Yihui Li, and Yisheng Guan. 2020. A Graph Attention Spatio-temporal Convolutional Networks for 3D Human Pose Estimation in Video. *arXiv preprint arXiv:2003.14179* (2020).

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics* 34, 6 (Nov. 2015), 1–16. https://doi.org/10.1145/2816795.2818013

[20] Brian D. Lowe, William G. Billotte, and Donald R. Peterson. 2019. ASTM F48 Formation and Standards for Industrial Exoskeletons and Exosuits. *IISE transactions on occupational ergonomics and human factors* 7 (2019), 10.1080/24725838.2019.1579769. https://doi.org/10.1080/24725838.2019.1579769

[21] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5442–5451.

[22] Simone Pasinetti, Cristina Nuzzi, Nicola Covre, Alessandro Luchetti, Luca Maule, Mauro Serpelloni, and Matteo Lancini. 2020. Validation of Marker-Less System for the Assessment of Upper Joints Reaction Forces in Exoskeleton Users. *Sensors (Basel, Switzerland)* 20, 14 (July 2020), 3899. https://doi.org/10.3390/s20143899

[23] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]* (April 2018). arXiv:1804.02767 [cs]

[24] Angela Sorriento, Maria Bianca Porfido, Stefano Mazzoleni, Giuseppe Calvosa, Miria Tenucci, Gastone Ciuti, and Paolo Dario. 2020. Optical and Electromagnetic Tracking Systems for Biomedical Applications: A Critical Review on Potentialities and Limitations. *IEEE Reviews in Biomedical Engineering* 13 (2020), 212–232. https://doi.org/10.1109/RBME.2019.2939091

[25] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep High-Resolution Representation Learning for Human Pose Estimation. *arXiv:1902.09212 [cs]* (Feb. 2019). arXiv:1902.09212 [cs]

[26] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2022. Deep Learning-Based Human Pose Estimation: A Survey. *arXiv:2012.13392 [cs]* (Jan. 2022). arXiv:2012.13392 [cs]

[27] Roland Zügner, Roy Tranberg, John Timperley, Diana Hodgins, Maziar Mohaddes, and Johan Kärrholm. 2019. Validation of Inertial Measurement Units with Optical Tracking System in Patients Operated with Total Hip Arthroplasty. *BMC Musculoskeletal Disorders* 20 (Feb. 2019), 52. https://doi.org/10.1186/s12891-019-2416-4