WearRAcon Europe 21 Virtual
European Exoskeleton Conference
October 5-7, 2021

MARKERLESS HUMAN KINEMATIC EVALUATION WITH 3D MONOCULAR POSE ESTIMATION ALGORITHMS FOR EXOSKELETON EVALUATIONS STANDARD

Amaan Rahman[1]

[1] **Undergraduate Researcher**: The Cooper Union, 41 Cooper Square, NY

ABSTRACT

Evaluation of exoskeleton performance analysis introduces standards for exoskeletons to ensure proper functionality and safety for the subject. Currently, there are limited standard evaluation methods for exoskeletons. Methods that are under investigation by the National Institute of Standards and Technology (NIST) Exoskeletons and Exosuits Research and Standard Test Methods team are applications of Optical Tracking System (OTS), markerless pose estimation methods, and Inertial Measurement Units (IMUs) to analyze exoskeleton user kinematics and to infer kinetic information to understand the impact of exoskeletons on a subject's joint movement and forces. This research focuses on investigating markerless 3D pose estimation algorithms to determine their viability as methods for tracking human joint position and deriving skeletal frame orientations.

Keywords: Markerless, Exoskeleton, 3D pose estimation, Convolutional Neural Networks

1. INTRODUCTION

3D pose estimation applies a Convolutional Neural Network (CNN) structure to estimate the 3D pose and joints of individual(s). Convolutional neural networks are a special type of neural network that extracts features from images utilizing trained filters for either classification or regression. A survey was conducted of various state-of-art 3D pose estimation algorithms such as, but not limited to: BlazePose [1], 3DMPPE [2], VideoPose [3], and HMR [4]. Given a survey of state-of-art 3D pose estimation algorithms, a selection of algorithms to be implemented was performed that focuses on human kinematic awareness and yields low error metrics based on each algorithm's reported errors.

3D monocular pose estimation algorithms are investigated to measure error between the estimated 3D pose and ground truth data (OTS data) by examining human limb orientations and joint positions. Current methods under investigation: Graph Attention Spatio-Temporal Convolution Network (GAST-NET) [5] and Video Inference for Human Body Pose and Shape Estimation (VIBE) [6]. The project consisted of evaluating the algorithms on videos of gait and hurdle tests; the hurdle test had additional challenges of limb occlusion.

The calculation of error metrics, Percent Detected Joints (PDJ) and Mean Per Joint Position Error (MPJPE), between various 3D monocular pose estimation algorithms and ground truth data provide a proper standard to determine precision and accuracy of pose estimation algorithms. Stereo based algorithms yielded MPJPE values between 10 mm to 50 mm and monocular based algorithms yielded MPJPE values between 30 mm to 90 mm. However, stereophotogrammetry requires synchronized cameras and calibration data, whereas most monovision algorithms do not. Spatio-Temporal models such as GAST-NET yield MPJPE values as low as 30 mm; through visual inspection, the algorithm detected joint positions even with occlusions by the hurdle test apparatus. The initial algorithm evaluation demonstrated the method was a strong candidate for markerless 3D joint detection and position estimation.

2. MATERIALS AND METHODS

Due to the COVID-19 pandemic, the research was conducted online. A Dell G7 laptop running Ubuntu 20.04 Linux distribution with 12-core Intel i7 CPU, RTX 2070 GPU, and 16GB of RAM was utilized to implement and execute various 3D pose estimation algorithms. The algorithms were tested on gait and hurdle test videos provided by the NIST. This research focused on the right leg limbs and joints, front view standing motions.

A program was developed to generate a seamless evaluation procedure against the provided NIST videos with the algorithms to be tested (GAST-NET and VIBE). Viewport extraction is conducted at the initial stages due to the provided videos being a composite of various synchronized viewports (front, back, side).

The quaternion orientation of the right leg limbs and 3D joint positions in the associated OTS data file is extracted to generate the ground truth data.

The orientation of the limbs and 3D joint positions of the right leg is extracted from each respective 3D pose estimation algorithm to then perform a comparison study based on the following error metrics: MPJPE and PJD.

MPJPE is the mean of the Euclidean distance between the estimated joint position and the ground truth joint position. PJD is the percent of joints that have a Euclidean distance between the estimated and ground truth joint position less than 0.2 of the torso diameter.

3. GAST-NET ARCHITECTURE

The GAST-NET architecture is broken down into 5 parts:

1. **Human(s) detection**: Uses YOLOv3, a robust object-detection algorithm, to track an arbitrary number of humans specified in the video sequence.
2. **2D Pose Estimation**: HRNet a high resolution 2D pose estimation algorithm used to generate the 2D image coordinates of the detected joints of poses.
3. Detect long-term patterns utilizing for an arbitrary set of frames with their associated 2D pose data.
4. **Local Attention Graph:** Use a Graph Convolution Network to construct 2D skeletal graph considering the symmetrical structure of the human anatomy and the bone-to-bone kinematic relations.
5. **Global Attention Graph:** Determine disconnected joint relations to determine global semantic data and pose constraints; this resolves issues with depth ambiguities and occlusions that most 3D pose estimation algorithms largely suffer from.

4. VIBE ARCHITECTURE

The VIBE architecture in 3 sections:

1. Extract features per frame of the input video using a pretrained CNN.
2. Extract information from past and future frames and then use features to regress the mesh parameters (SMPL parameters). The parameters contain relative bone orientations data, 3D joint position data relative to midpoint of the hips.
3. Discriminate fake and real motions based on a ground truth set of motions called AMASS.

5. RESULTS AND DISCUSSION

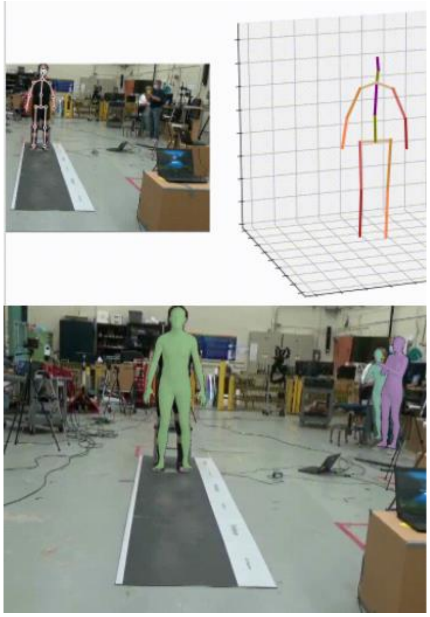

Figure 1: GAST-NET (top) and VIBE (bottom) 3D pose estimation visualization of a sample frame from standing motion frame set

GAST-NET and VIBE 3D pose estimation algorithms were executed against 300 frames of front-view standing motion. The processing time for GAST-

NET and VIBE was 70.290s ± 5.761s and 246.287s ± 2.377s respectively. **Figure 1** depicts the 3D pose estimation visualization of each respective algorithm. The VIBE visualization shows multiple persons evaluated, which is an expected result; the individual in the middle of the frame is the person of interest algorithm evaluations.

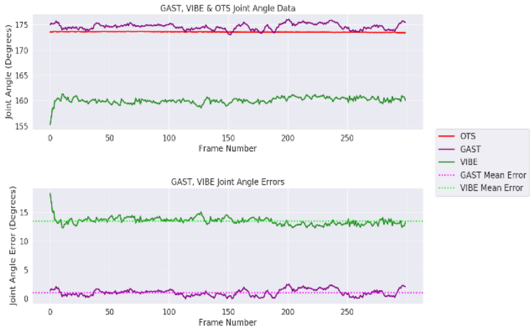

Figure 2: GAST-NET (Purple) and VIBE (Green) joint angle results (top) and errors (bottom) respective to the OTS data (red) over 300 frames of standing motion.

The joint angle data for GAST-NET is closer to the OTS joint angle data with low errors with respect to VIBE according to **Figure 2**. However, VIBE's joint angle data and errors are more constrained proving reliable repeatability. GAST-NET having larger extremes compared to VIBE is an expected result because GAST-NET's 3D pose estimation suffers if the video sequence has low resolution, which is true for the video input in this case; viewport extraction is essentially zooming into the region of interest of the provided NIST video to extract said viewport region (in this case the front view), thus reducing the resolution of the extracted viewport video sequence.

The MPJPE for GAST-NET and VIBE are 8.128mm ± 3.067mm and 4.277mm ± 2.742mm respectively. GAST-NET with a larger MPJPE compared to VIBE is due to the low-resolution video input. The PJD for GAST-NET and VIBE are both 100%, which is expected because there are 3 joints of interest: hip, knee, and ankle.

6. CONCLUSION

Examining standing motions provides a stable analysis of GAST-NET and VIBE against the ground truth and provides promise to utilizing markerless methods for kinematic and kinetic evaluations. Concluding that the markerless alternative for exoskeleton kinematic and kinetic evaluations is premature based on the research conducted; next steps are to evaluate a diverse set of motions and viewports with a larger set of 3D pose estimation algorithms against the ground truth data. Advanced kinematic evaluations are to be conducted in the future such as angular velocity, displacement, and acceleration, which leads into kinetic evaluations utilizing the Inertial Measurement Units as a ground truth basis.

The program that was developed to generate the experiment conducted and discussed is to be packaged for repeatability for large data generation.

ACKNOWLEDGEMENTS

Special thanks to my mentors Mili Shah (Cooper Union Research Mentor and NIST SURF Mentor) and Ya-Shian Li-Baboud (NIST SURF mentor), and the rest of the NIST Exoskeleton Evaluations team: Roger Bostelman, Soocheol Yoon, and Ann Virts.

REFERENCES

[1] V. Bazarevsky, I. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, and M. Grundmann, "BlazePose: On-device Real-time Body Pose tracking," p. 4.

[2] G. Moon, J. Y. Chang, and K. M. Lee, "Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation From a Single RGB Image," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), Oct. 2019, pp. 10132–10141. doi: 10.1109/ICCV.2019.01023.

[3] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli, "3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7745–7754. doi: 10.1109/CVPR.2019.00794.

[4] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-End Recovery of Human Shape and Pose," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, Jun. 2018, pp. 7122–7131. doi: 10.1109/CVPR.2018.00744.

[5] J. Liu, *A Graph Attention Spatio-temporal Convolutional Networks for 3D Human Pose Estimation in Video (GAST-Net)*. 2021. Accessed: Aug. 22, 2021. [Online]. Available: https://github.com/fabro66/GAST-Net-3DPoseEstimation

[6] M. Kocabas, *VIBE: Video Inference for Human Body Pose and Shape Estimation [CVPR-2020]*. 2021. Accessed: Aug. 22, 2021. [Online]. Available: https://github.com/mkocabas/VIBE