MiniMesh: Real-Time 5,000-Node Anatomical Human Body Mesh Reconstruction for Portable Devices

Daniel Mathew

(i) Personal Section

I once met Iron Man.

Opening the towering fifteen-foot metallic doors, he entered a motion capture lab as my first patient during a biomechanics study. Without a word, he unscrewed his bone-integrated leg implant and handed it to me along with his prosthetic arm. Suddenly pausing, he looked up and joked, “Never thought you would meet Iron Man. Ha!” I sighed in relief.

As I assisted him in his grueling three-hour physical assessment, I soon discovered that better solutions for gait analysis were needed. My first taste was with 3D Body Pose Estimation, the task of predicting 17 joint locations of a person from a picture. My goal was to reduce the complexity of the task and run the algorithm on a phone. Without a mentor, I took this project to the International Science and Engineering Fair (ISEF) and presented my first mobile biomechanics device.

The very next year, I joined Lumo Imaging as a consultant for dermatology research. Their goal is to create state-of-the-art technology that assists doctors in their work area. With a diverse selection of students, every member of the research team works independently on a project that intellectually challenges them. After a brief discussion with the Managing Director, I discovered an unsolved research problem that aligned similarly with my background. This time, instead of predicting 17 three-dimensional points, my goal was to predict thousands.

I spent countless hours learning new mathematical techniques to break down this complex problem into easier tasks. As I gleaned from research papers and online lesson series, I

got glimpses into fascinating worlds of Linear Algebra, Differential Calculus, and Analytical Geometry. What initially seemed like such distant topics to a highschooler became more familiar with constant learning. Equipped with powerful tools, I was able to create a "pathway" to success in which I hoped to theoretically solve this problem. With weekly meetings to keep myself accountable, I inched step by step closer to a highly accurate solution. As I worked on this problem, I fell more in love with the scientific method. I enjoyed the rigorous cycle of hypothesis and testing in the pursuit of truth.

My biggest piece of advice for any highschooler looking or considering conducting research is to have endurance. I cannot claim to be the most persistent person in the world, but I can testify to the power of even a little bit of resilience. When I first looked for research, I was honestly very discouraged. I had cold emailed what felt like countless professors all which fell through immediately. It was after many redirects where I did land my first position at Lumo. After working on their assigned project, I then realized it was not as I had hoped for, so I made it my own. Diligently working, I was able to finish the original task as soon as possible and got the opportunity to design my own project (which is the one I presented). Even while researching, it's very common to face roadblocks along the way. What makes research exciting is figuring out ways to overcome them and pressing on towards the solution. In all this, resilience does certainly play a part in forming successful and enjoyable research experiences.

(ii) Research Section

Summary

When a person goes to check on a skin lesion or a runner wants to improve their form, a scanner is often used to track points on the body for measurements. Currently, there exists no solution that can instantaneously (in less than a second) compute the location of all these points

at once. MiniMesh is a novel, resource-efficient algorithm that can accomplish this task on a small computer (like a phone or laptop) in real time from a single image. The algorithm takes an unorthodox approach of splitting this complex task into two simpler problems: finding the location of only 119 landmarks and extracting an outline from the patient's image. After running these procedures, their output can be used to estimate the location of thousands of points which are displayed on the screen using a custom-made rendering engine. Overall, MiniMesh can process on average 20 images per second with high accuracy in all tasks. The speed of the algorithm can be improved to 50+ images per second when running each part of the algorithm parallelly. MiniMesh accomplishes what large motion capture systems can do using only a portable device, creating a fast, accurate, and inexpensive solution for all.

Introduction

This project explores a core component of biomechanics research and human-computer interaction: pose estimation—the usage of computer vision in localizing anatomical landmarks on the human body.

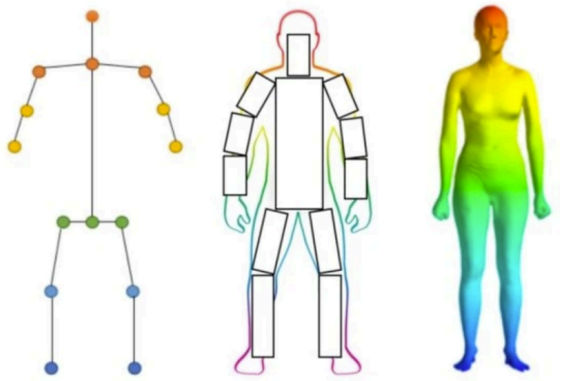

Fig 1: Diagram of Pose Estimation Types (Odemakinde, 2023)

The overarching research of pose estimation, however, is split into three broad categories in increasing computational difficulty: kinematic, planar, and volumetric. The first of which,

kinematic, aims to predict the 2D or 3D location of specific joints of a human subject in the global frame (Josyula & Ostadabbas, 2021). Kinematic pose estimation is often characterized as determining locations in a pointwise method, where each joint corresponds to a single dot. The second, planar pose estimation, which increases the difficulty, attempts to estimate the counters of the human body. Oftentimes, this research predicts silhouettes of the human subject from a camera feed. Finally, volumetric, the most computationally expensive of them all, aims to render the entire three-dimensional mesh of the human subject.

In the most basic form of pose estimation, kinematic, there are a plethora of problems that make this challenge especially difficult: occlusions from various objects or the subjects themselves, varying numbers of human subjects at different scales, motion-blur/focus issues arising from the hardware itself, etc (Osokin, 2018). This task becomes increasingly more complicated as the number of points increases from an average of 18 joints, in kinematic pose estimation, (Sarafianos & Boteanu, 2016) to over 5,000 in a full volumetric pose estimation algorithm.

While much research has been done on creating efficient pose estimation on the kinematic and planar scale, there are few solutions for an anatomically-consistent full-body volumetric pose estimation that has the potential to be ported to a mobile device. Much-less, solutions that run on a GPU in real-time don't have vertex correspondence (also known as anatomical consistency). Vertex correspondence allows that regardless of the person or pose, vertices that have the same ID correspond to the same anatomical landmark.

MiniMesh aims to accomplish modified full-body volumetric pose estimation for portable devices in real-time. Additionally, the mesh should maintain vertex correspondence, so that for any vertex, it is located at the same anatomical location regardless of the human shape or pose.

Applications

Pose Estimation has many applications in the modern world in which analysis of the human body is required.

Biomechanics

In biomechanical research and industry, a standard model for the human body is often needed for computational purposes. For instance, in gait analysis, the study of walking motion, large motion capture setups have been often employed to track anatomical landmarks/joints on the body as a patient goes through a series of tests. By having a quantifiable location for each joint, irregularities can be found in an individual's gait and treatments can be administered.

Oftentimes, these expensive motion capture systems can have costs of upwards of thousands to tens of thousands of dollars and are found sparsely in laboratories around the nation. Thus a solution that uses computer vision techniques that run solely on an individual's mobile device could have a tremendous impact on the availability of such tools for a consumer.

While motion capture only covers kinematic pose estimation (location of individual joints), MiniMesh has the computational capabilities to estimate the entire mesh of the human subject, containing thousands of points. With an abundance of information from volumetric pose estimation, a patient’s joints, silhouettes, texture information, and shape parameters can be extracted with ease, opening the door to many more biomechanical applications from fall detection to security/surveillance systems.

Dermatology and Anatomy

Another very promising application of volumetric pose estimation is in the field of dermatology and anatomy. By maintaining vertex correspondence on landmarks, MiniMesh

allows for the creation of a "coordinate system" of the human body where individual anatomical locations can be indexed in constant time.

One good example of this in dermatology is feature correspondence, where dermatologists need to track the location/shape of a surface feature (skin cancer, lesions, etc.) as it progresses in time. Current solutions have accomplished this by calculating distances from easily visible landmarks such as the nipples, naves, hip joints, and many others; however, these only serve as a rough guess and are labor-intensive. With MiniMesh, any surface feature can be represented by the set of vertices it is encapsulated by.

Similar to its application in Dermatology, MiniMesh's capability to act as a coordinate system for any point on the body has a significant impact on Anatomy. As the mesh is created to fit anatomical landmarks, it may be used to locate subcutaneous features such as organs, bones, veins, and more. This feature has implications for robotic surgery and rapid anthropometric measurements

Animation

As VR/AR and the prospective meta-verse are on the horizon, MiniMesh offers substantive improvements in techniques for animation and texture mapping (Kumarapu & Mukherjee, 2021). By keeping track of thousands of anatomical landmarks with high accuracy, various meshes other than the standard human mesh may be applied. By increasing the fidelity of the mesh model, highly detailed animated characters may be displayed.

In regards to texture mapping, as the mesh is "universal" (uses the same number of vertices regardless of the user) the associated UV Map of the mesh doesn't change in topology. Thus, transforming and applying texture to MiniMesh is extremely straightforward and only

requires one texture generation sequence to apply texture for any number of individuals regardless of shape or pose.

Related Work

The following is a brief summary of the related work in pose estimation that has been considered to create MiniMesh.

Kinematic Pose Estimation

2D Pose Estimation. The two most common methods of 2D pose estimation include regression-based and detection-based techniques. Solutions that follow the regression-based paradigm directly predict the 2D joints of the individual (Mao & Ge, 2022). These solutions encode the entire input image and directly regress the *2n* coordinates of the points. Direct regression is extremely fast as the output vector is significantly smaller than the input, so large convolutions may be used. However, spatial relationships between joints are not used effectively, and thus often fail further testing when large parts of the body are occluded (partial body pose estimation). In other words, this method is "all or nothing" in that the model only produces viable output if the entire subject is seen.

Detection-based models, also known as heatmap regression, have been found to perform significantly better than their pure regression counterparts as they maintain high accuracy. This paradigm initializes a randomly generated heatmap and regresses a predicted heatmap through several filters to generate locations that have a high probability for joints (Bulat & Tzimiropoulos, 2016) This solution does involve much more memory and requires longer processing times as the output and input vector are the same size; however, accuracy is improved and is suitable for partial body pose estimation as spatial information is retained.

Planar Pose Estimation

The types of planar pose estimation are split into semantic segmentation and instance segmentation.

Semantic segmentation computes the pose estimation task on a pixel-by-pixel level by determining whether or not any given pixel is part of the object in question. In this scenario, a heatmap is produced with each value representing the probability of that pixel being part of the class label. For multi-part detection, several channels are created to accommodate various labels.

Instance segmentation, on the other hand, is more geared towards pixel-perfect silhouette segmentation. This methodology is generally split into classification, detection, and segmentation. There is a high level of granularity in the prediction but results in extremely large and slow models that are unsuitable for real-time applications (Zhang & Li, 2019). The most common solutions for high-accuracy instance segmentation are YOLACT and MaskRCNN.

Volumetric Pose Estimation

Model-Based Volumetric Pose Estimation. Most current volumetric pose estimation models rely on model-based approaches that use prior information about human anatomy and morphology for mesh generation. These solutions regress predefined parameters that control the shape, pose, and position of these volumetric models—the most popular ones being SMPL for body pose estimation and MANO for hand pose estimation (Boukhayma & de Bem, 2019).

Early models used CNNs to directly regress parameters model-based solutions. Kanazawa et al. (2018) proposed a trainable human mesh recovery (HMR) system that has a defined loss function to ensure a plausible output mesh. Kolotouros et al. (2019) tried a different solution to iteratively predict model parameters in an error minimization problem with high accuracy and lower times (as adjacent frames require small parameter adjustments). Moon et al.

(2020) suggest a voxel-based approach (similar to heatmap regression discussed in kinematic pose estimation) and use five loss functions to regress model parameters.

Model-Free Volumetric Pose Estimation. Model-free solutions aim to predict all vertices on human mesh without the use of parameter-based models. Kolotouros et al (2019) suggested a graph convolution network-based approach to pose estimation with a template human mesh in a standard pose. MEsh TRansfOrmer (Lin & Wang, 2021) directly regresses all joints of the mesh with a transformer that performs at a high accuracy but at slower speeds.

Methodology

The following is a brief summary of the methodology towards the creation of the MiniMesh algorithm.

While model-based pose estimation reduces the processing time of a machine learning model, most spatial relationship information is lost, resulting in less robust predictions. On the other hand, model-free pose estimation, which predicts individual vertex locations, suffers from larger calculations and slow speeds. To combat these issues, MiniMesh novelly breaks down the expensive volumetric pose estimation problem into a combination of several lesser kinematic and planar pose estimation tasks (*Fig 2*). This new solution maintains the high-speed benefits of model-based pose estimation while allowing for spatial information to be incorporated.

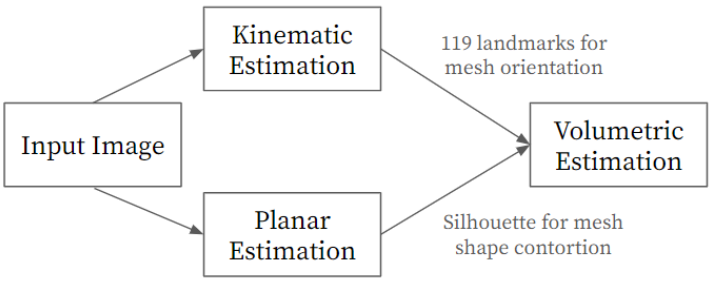

Fig 2: Workflow Diagram of MiniMesh

The first stage of MiniMesh is a 2D kinematic pose estimation of the entire body using pre-trained models, including models from MoveNet. The calculation of these landmarks allows later algorithms to use them to fit and align the mesh onto the body by using these marks as references. Rather than computing all body landmarks at once, landmarks are calculated individually for the three groups of the body: the head, the hands, and the torso/legs. These landmarks are then combined to form the 119 2D kinematic landmarks for the entire body.

Concurrent with the 2D kinematic landmark detection is planar pose estimation—the second stage of MiniMesh. A combination of pre-trained machine learning models (BodyPix) and image contour analysis is used to extract high-accuracy silhouettes of the human subject in a real-time capacity. Many training and run-time techniques are used to reduce computational complexity and allow for faster predictions on portable devices. These silhouettes are used to calculate the width of various parts of the body.

Finally, the third and final stage of MiniMesh is a fitting algorithm to fit face, hand, and body meshes to the human subject based on information from the prior stages—volumetric pose estimation. The 2D landmarks are used to assist in the alignment of the mesh onto the picture and an iterative solver is deployed to use the silhouette information to contort the mesh to fit the body shape of the individual. MiniMesh uses a custom-created 3D engine to render and manipulate all points on the mesh as there does not yet exist a formal interface for this task.

Speed Testing

As the MiniMesh algorithm is designed to run on smaller devices in a real-time capacity, having sufficiently quick processing speeds is mandatory. The speed is tested by measuring both the detection rate (the number of images the model can detect in a single second) and the frame rate (rendering time) of a camera feed. For all tests, 3,400 random images were fed in as a stream

and processed on one core of a CPU. Below are some of the graphs measuring the speed of the various algorithms.

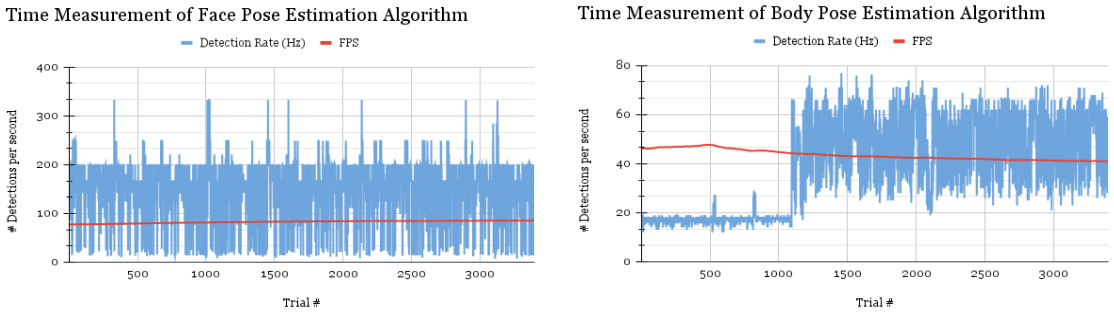

Fig 3: Execution Time and Frame Rate of Face and Body Kinematic Pose Estimation

Both images exhibit the same patterns in regards to detection time. Over all trials and estimation tasks, the frame rate for prediction is consistently less than the detection rate for each image. This phenomenon can be explained by rendering delays due to the output of the algorithm. As kinematic pose estimation algorithms produce disjoint points that are independently plotted, these must be processed iteratively. This methodology is heavy on the renderer, causing delays in that portion of the algorithm. While detection is fairly quick, the frame rate is ultimately limited by the inability to keep rendering at the same speed.

Two peculiar patterns are noticeable in the detection rate for all algorithms: large jumps in overall processing time and noise between individual trials. The large overall jumps (found in *Body 2D Kinematic Pose Estimation*) can be accounted for by the individual processor and not the algorithm itself. As the programs are running on a single core, resource management on the processor ultimately allows for redundant operations to be eliminated. This optimization allows for the remaining tests to be done at a significantly faster speed. The second pattern, the large noise, is simply due to variability in model processing (found in all tests). At random points during the algorithm, resources may be allocated to the wrong areas causing delays for any

individual frame. This again, is not a result of the algorithm but resource allocation from the CPU.

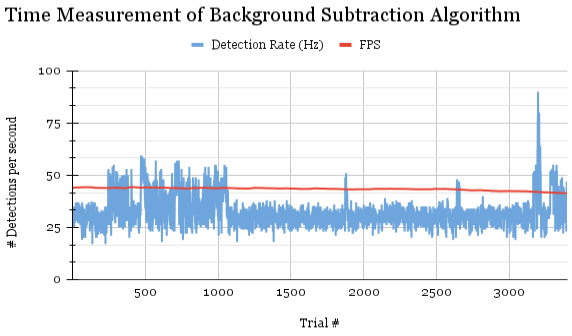

Fig 4: Execution Time and Frame Rate of Planar Pose Estimation

In terms of detection rate and frame rate, the planar pose estimation displays a completely different pattern than the kinematic pose estimation counterpart. Rather than having the frame rate on average lower than the detection rate, the pattern is reversed. Yet the explanations for both phenomena are similar. In this case, the output of the model is a large mask, containing thousands of boolean values. To be rendered, the program only has to apply the mask to the image using an efficient function for this task. Overall, this puts a heavier emphasis on the image detection aspect of the algorithm rather than the rendering. Thus, the frame rate is found to have, on average, a higher value than the detection rate.

Demonstrations

MiniMesh, and its components, are evaluated in both the minimization of prediction time and the accuracy of the prediction in comparison to a ground truth. While the various pre-trained models use different datasets for testing and training purposes, every set contains a variety of images from diverse scenarios. For instance, every image contains challenges in real-world datasets including occlusion, contact, and crowding. This complexity in data allows for a set of more robust models for usage outside of the typical laboratory settings.

a) Original Image; b) 68 2D Face Landmarks; c) 3D Face Box; d) Orientation of Head Calculation; e) Silhouette Extraction; f) Face Detection Bounding Box g); 17 2D Body Landmarks; h) 34 2D Hand Landmarks; i) Kinematic and Planar Together

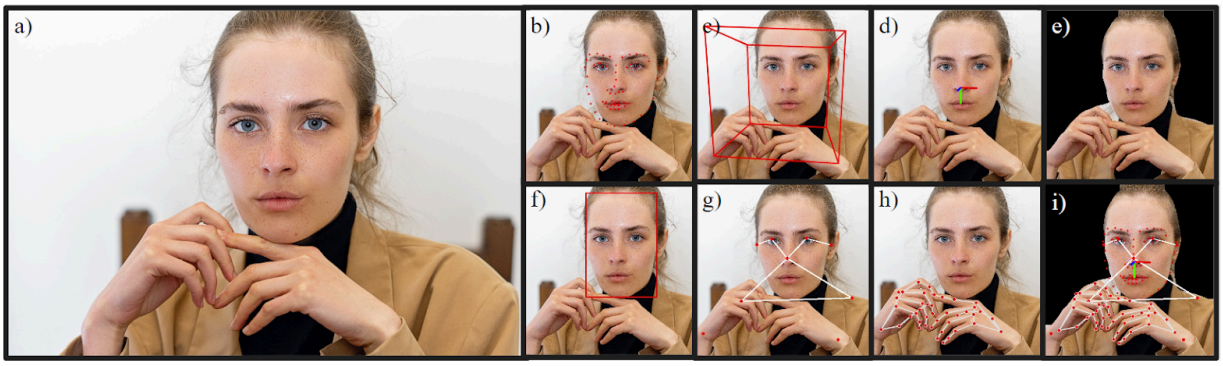

Fig 5: Headshot Image of Woman Processed Through Various Algorithms

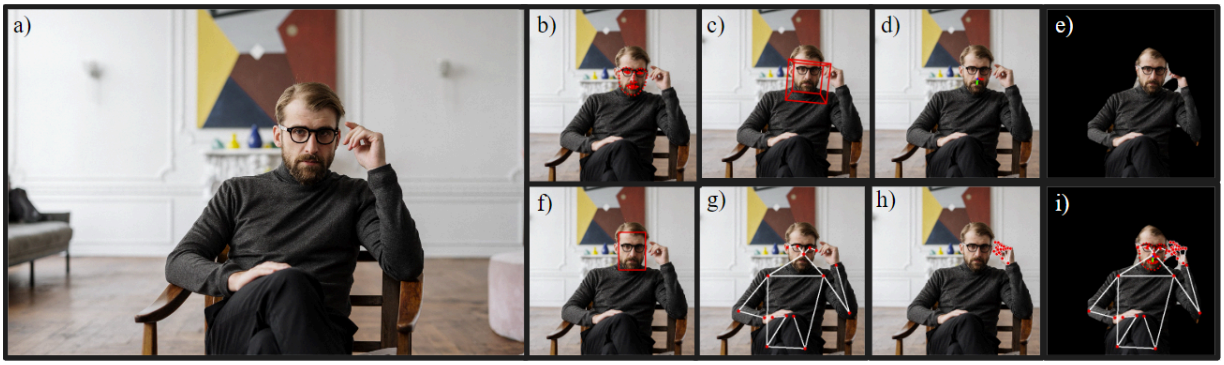

Fig 6: Image of Man Sitting on a Chair Processed Through Various Algorithms

Due to the variety of data, these algorithms can produce fairly accurate results in each of these components in real-time. Inspection of *Fig 5* and *Fig 6* reveal eight tests on two images run through several pose estimation algorihtms. *Fig 5* shows a close-up image of a woman with much of her body occluded from the camera. The presence of her hand occludes her body and her face; however, MiniMesh can accurately predict 3D orientation, kinematic pose estimation (face, hand, and body), and planar pose estimation from a single image to be used by the mesh fitting algorithm. *Fig 6* performs similarly with a patient in a complex pose, zoomed-out, and artifacts (chair in this instance) to confuse the algorithm.

References

Note: All images not cited were created by the student or royalty-free from Pexels

Boukhayma, A., de Bem, R., Torr, P.H.: 3D hand shape and pose from images in the wild. In: CVPR (2019)

Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)

Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: CVPR (2019)

Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: ICCV (2019)

Lin, K., Wang, L., & Liu, Z. (2021). End-to-end human pose and mesh reconstruction with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1954-1963).

Moon, G., & Lee, K. M. (2020). I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16 (pp. 752-768). Springer International Publishing.

Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z., & den Hengel, A. V. (2022, October). Poseur: Direct human pose regression with transformers. In European Conference on Computer Vision (pp. 72-88). Cham: Springer Nature Switzerland.

Sarafianos, N., Boteanu, B., Ionescu, B., & Kakadiaris, I. A. (2016). 3d human pose estimation: A review of the literature and analysis of covariates. Computer Vision and Image Understanding, 152, 1-20.

Osokin, D. (2018). Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. arXiv preprint arXiv:1811.12004.

Bulat, A., & Tzimiropoulos, G. (2016). Human pose estimation via convolutional part heatmap regression. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The

Netherlands, October 11–14, 2016, Proceedings, Part VII 14 (pp. 717-732). Springer International Publishing.

Kumarapu, L., & Mukherjee, P. (2021). Animepose: Multi-person 3d pose estimation and animation. Pattern Recognition Letters, 147, 16-24.

Josyula, R., & Ostadabbas, S. (2021). A review on human pose estimation. arXiv preprint arXiv:2110.06877.

Chen, C. H., & Ramanan, D. (2017). 3d human pose estimation= 2d pose estimation+ matching. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7035-7043).

Zhang, S. H., Li, R., Dong, X., Rosin, P., Cai, Z., Han, X., ... & Hu, S. M. (2019). Pose2seg: Detection free human instance segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 889-898).

Odemakinde, E. (2023a, March 10). Human pose estimation with deep learning - ultimate overview in 2023. viso.ai. https://viso.ai/deep-learning/pose-estimation-ultimate-overview/

Man in black sweater sitting on Brown Wooden Chair - pexels. (n.d.). https://www.pexels.com/photo/man-in-black-sweater-sitting-on-brown-wooden-chair-4100672/

Close up shot of woman in brown blazer. - pexels. (n.d.). https://www.pexels.com/photo/close-up-shot-of-woman-in-brown-blazer-8730379/

Suzuki, S., & Be, K. (1985). Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing, 30(1), 32-46. https://doi.org/10.1016/0734-189X(85)90016-7

Votel, R. & Li N. (2021) Next-Generation Pose Detection with MoveNet and TensorFlow.js. https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html