International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 10 | October - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Comparative study on pose estimators such as MoveNet Lighting,MoveNet Thunder, and OpenPose (MobileNet) model for HumanPose Estimation over Real-Time Feed

Akshay A [1], Sai Varun T[2], Yesvanthraja D[3], Nithyapriya S[4]

[1]Student , Department of Artificial Intelligence and Data Science,Bannari Amman Institute Of Technology, Sathyamangalam,
[2]Student , Department of Artificial Intelligence and Data Science, Bannari Amman Institute Of Technology, Sathyamangalam,
[3]Student , Department of Artificial Intelligence and Data Science, Bannari Amman Institute Of Technology, Sathyamangalam,
[4]Assistant Professor, Department of Artificial Intelligence and Data Science,Bannari Amman Institute Of Technology, Sathyamangalam.

Abstract - Human pose estimation is a crucial task in various domains, including fitness and motion analysis, and sports performance evaluation. Existing technologies have limitations in terms of accuracy and real-time performance, which highlights the need for advanced solutions. The aim of this paper is to display the comparison between models for human pose estimation over real-time feed, with high accuracy and real-time performance. The problem statement involves developing a model and testing it to check if it can handle multiple subjects, different poses, and various lighting conditions. The proposed methodology involves using state-of-the-art models, such as MoveNet Lightning and MoveNet Thunder, along with the comparison to pre-existing models such as OpenPose model, which are slow and inaccurate. The key findings of this work include real-time performance comparison, high accuracy, and applicability in various domains. Nevertheless, it is imperative to acknowledge and confront specific constraints, such the issue of pose ambiguity and the challenge of effectively controlling occlusion. The suggested study presents a potentially effective approach for real-time human pose assessment, which has the capacity to yield advantages for humans.

Keywords: Human pose estimation, real-time feed, MoveNet Lightning, MoveNet Thunder, OpenPose.

1 INTRODUCTION

This paper aims to conduct a comparative study on various pose estimators, namely MoveNet Lightning, MoveNet Thunder, and OpenPose (MobileNet) model, for human pose estimation over real-time feed. The study involves analyzing the performance of these models on static images as well as real-time video feeds from webcams.

The paper provides an overview of the overall process involved in pose estimation, including data collection, key point definition, model selection, and image resizing. It also discusses the architecture of MoveNet and MobileNet models, highlighting their unique features that make them suitable for real-time applications.

The study compares the performance of these models based on factors such as accuracy, speed, and resource utilization.

The results of the comparative study can help developers and researchers choose the most suitable pose estimator for their specific use case.

Overall, this paper aims to provide valuable insights into the strengths and weaknesses of different pose estimators and their potential applications in computer vision research and development.

1.1 Evolution of Pose Estimation

Pose estimation is a fascinating computer vision technique that involves determining the positions and orientations of body parts in humans from visual input, such as images or videos.

Over time, this field has witnessed significant advancements, with early approaches relying on manual feature engineering and simple models.

However, recent progress has been primarily driven by deep learning methods, particularly convolutional neural networks (CNNs), which have shown exceptional accuracy and adaptability in capturing intricate poses from real-time video feeds.

1.2 Dealing with Occlusions

A crucial challenge in real-time pose estimation lies in handling occlusions. In humans, body parts maybe partially or fully obscured by various objects or individuals, making it difficult for the algorithm to accurately infer the complete pose.

1.3 Ensuring Real-Time Performance

Real-time pose estimation requires the algorithm to process video frames efficiently and swiftly. Striking a balance between high accuracy and low-latency performance is essential, especially for interactive systems and real-time monitoring applications.

1.4 Datasets for Training

The foundation of robust pose estimation models relies on high-quality, diverse, and well-annotated data. Data plays a critical role in helping algorithms learn the complex relationships between body joints and their appearances in various real-world scenarios.

1.5 Annotation Techniques

Manually annotating pose data can be time-consuming and labor-intensive. To streamline the process and ensure accuracy, researchers have developed various annotation techniques, including key point annotation tools and semi-automated methods. Key point annotations not only helps us deduce the structural dependencies into measurable factors but also helps us keep track of how the model fares in different conditions through its mapping techniques.

1.6 Sports Analysis

Pose estimation plays a vital role in sports analysis, where it provides valuable insights into athletes' movements. Coaches and analysts can use this information to assess players' performance, identify areas for improvement, and devise effective training strategies.

2 OBJECTIVES AND METHODOLOGY

2.1 Overall Process

The overall process of our pose estimation model involves collecting the necessary data, such as images or videos with subjects in them. We then install the necessary dependencies and define our workflow environment.

Next, we define a set of key points based on the subject in the frame. We list out the pairing conditions for interconnection and set a threshold by which the model finalizes key point pairing.

For model selection, we use different blocks to segregate loading the feed, followed by estimation and drawing key point references on the frame. We also resize the image to fit the model build and input requirements.

The result is a 3-color frame image with points mapped on an x-y scale, along with confidence-checked points that are filtered out to map the pose on the given frame.

This proposed work highlights the functionality of our model, which achieves good performance while running at a high frame rate of over 50 FPS on most modern laptops. These findings contribute to the advancement of pose estimation techniques and their potential applications in various fields.

The process as mentioned above performed over a static image as shown in Fig.2.1, the key points are mapped, and the pose estimation is performed over the image.

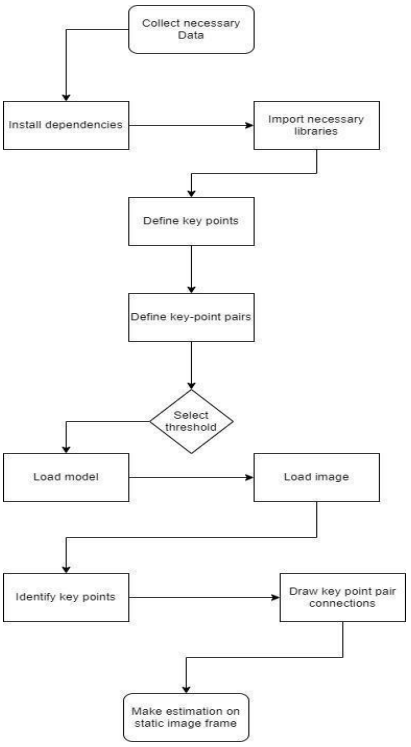


Fig.2.1 Pose estimation over static image

2.2 Over real-time feed

In Real time feed the usage of a web cam comes into play where the pose estimation work on the object rendering over real time and the key points are mapped and pose estimation is performed as shown in Fig.2.2.

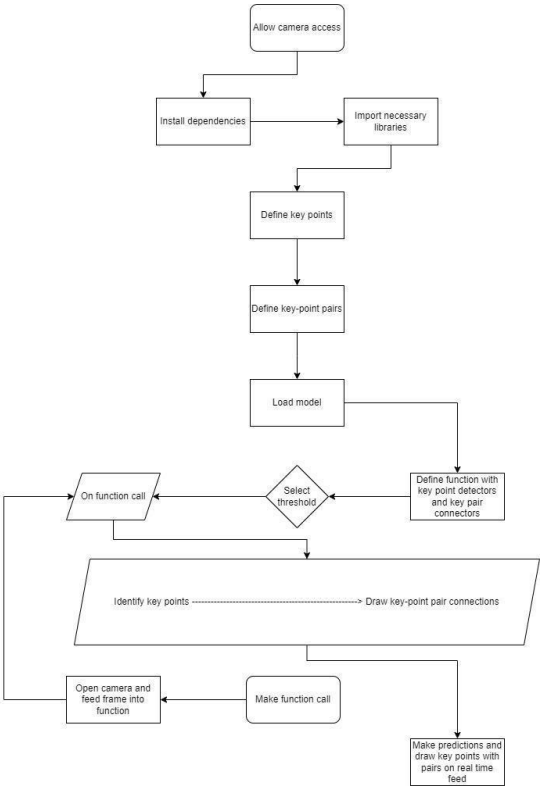


Fig.2.2 Pose estimation over real time-feed

2.3 Movenet Lightning and Thunder architecture

MoveNet is an advanced model utilized for the purpose of pose identification, which has been specifically developed to accurately and swiftly identify 17 keypoints representing various parts of the human body. The development of this model has been conducted through rigorous study and comprehensive testing, resulting in its shown efficacy across a diverse array of applications.

MoveNet possesses a notable attribute in the form of its exceptional processing speed, rendering it highly suitable for real-time applications. The aforementioned model exhibits the ability to efficiently handle substantial volumes of data within a limited timeframe, hence enabling it to achieve exceptional precision in the detection and tracking of bodily motions.

There are two variations of MoveNet, namely Lightning and Thunder, which may be accessed on TensorFlow Hub. The aforementioned versions have been strategically tuned to cater to various use cases, hence providing developers with the flexibility to select the most suitable option based on their own requirements. Lightning is particularly well suited for applications that necessitate precise and prompt data processing, with minimal delay. Conversely, Thunder is more advantageous for applications that prioritize the handling of large volumes of data at a rapid pace.

In general, MoveNet can be regarded as a remarkable model that signifies a notable progression in the domain of pose detection. The combination of its high speed, precise accuracy, and adaptable nature renders it a highly commendable option for a diverse array of applications, spanning from the study of sports to the realm of virtual reality and beyond. The model's inclusion on TensorFlow Hub facilitates convenient access for developers and researchers worldwide, rendering it a crucial resource for individuals engaged in computer vision research and development.

The arhitecture for Movenet Lightning and Thunder model is mentioned in Fig.2.3.

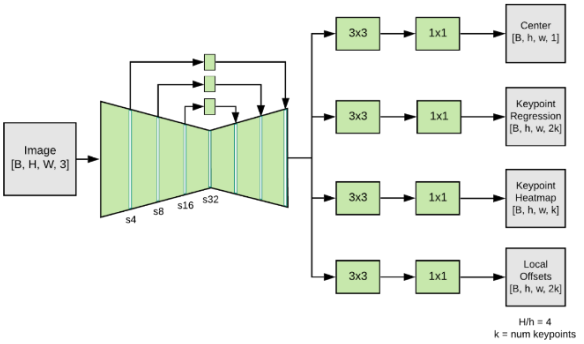


Fig.2.3 Movenet Lightning and Thunder architecture

2.4 Mobilenet architecture

The MobileNet architecture is a convolutional neural network that has been specifically developed to optimize accuracy while minimizing processing resources. The architectural design incorporates multiple layers, including depthwise separable convolution layers, which serve the purpose of minimizing the number of

parameters and computational expenses, while simultaneously preserving the correctness of the model.

The architecture of MobileNet incorporates depth wise separable convolution layers as a crucial element as shown in Fig 2.4. The layers are composed of two sequential operations: depth wise convolution and pointwise convolution. During the depth wise convolution process, the input channels undergo independent convolutions using a small kernel size, resulting in a reduction in the number of parameters needed for this operation. During the pointwise convolution operation, the output channels undergo convolution with a 1x1 kernel, hence enhancing the model's capacity for expression.
MobileNet utilizes depth wise separable convolution layers to provide superior accuracy with reduced parameters and processing resources compared to conventional convolutional neural network designs. This characteristic renders it a very suitable option for applications that necessitate real-time processing or are constrained by computational resources.

In addition to its efficient architectural design, MobileNet incorporates supplementary characteristics that significantly enhance its overall accuracy. Some of the strategies commonly used in deep learning models are batch normalization, rectified linear unit (ReLU) activation functions, and skip connections. These strategies contribute to enhancing the resilience and steadiness of the model, guaranteeing its ability to operate effectively even in demanding circumstances.

In general, the MobileNet architecture displays the ability to attain significant precision while utilizing little processing resources. The effectiveness of the model can be attributed to the utilization of depthwise separable convolution layers, which have gained significant popularity across several domains such as image classification and object recognition.

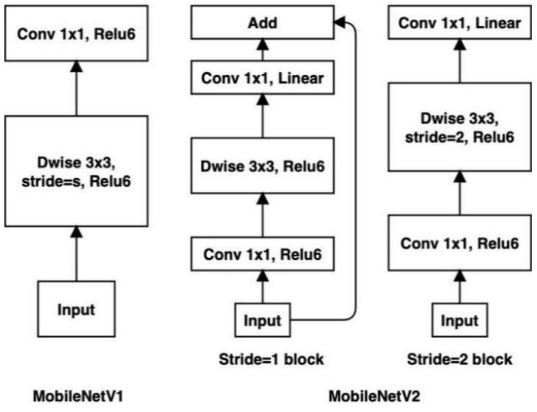


Fig.2.4 MobileNet architecture

3 PROPOSED WORK MODULES

3.1 Data Collection and Preparation

3.1.1 Data Sources

Our journey commences with the crucial task of acquiring pertinent datasets. Effective model training and evaluation necessitate the availability of high-quality data. Any given model is going to fare well in a situation where it encounters a decent dataset dedicated to a use case, but we set the bar a little bit different and tried to encompass how the model performs when fed with real time data and images that was not particularly meant for pose estimation.

3.1.2 Data Pre-processing

To ensure effective model training, we commit ourselves to a meticulous process of data pre-processing. This critical step ensures that our data is primed for optimal utilization.

We employ several key tasks during data pre-processing, including resizing images to meet specific input dimensions required for our models. This step ensures uniformity and compatibility within our training pipeline.

We also prioritize data consistency and quality, leaving no stone unturned in ensuring that our data is not only consistent but also of the highest quality. This entails meticulous quality control measures to eliminate anomalies that could hinder the training process.

Recognizing the power of data diversity, we apply data augmentation techniques during pre-processing. These techniques breathe life into our training dataset, exposing our models to a broader range of scenarios and thereby enhancing their robustness.

Overall, these pre-processing steps pave the way for effective model training and contribute to the advancement of computer vision research and development.

3.2 Model Selection and Architecture

3.2.1 Model Variants

Our research embarks on a journey that involves the selection of two primary model variants for pose estimation:

MoveNet.SinglePose.Lightning (4.2.1.1): This model

variant is engineered for real-time performance, offering a streamlined architecture ideal for modern laptops. Its unique selling point lies in its ability to maintain excellent performance while ensuring real-time capabilities.

MoveNet.SinglePose.Thunder (4.2.1.2): For those who demand uncompromised prediction quality, the Thunder variant is the choice.
It offers a higher capacity model that still maintains real-time performance, albeit at a slightly reduced speed compared to its Lightning counterpart.

3.2.2 Model Architecture

The MobileNetV2 architecture is thoroughly ingrained in both of our model versions. This basic structure acts as the blank canvas on which we draw our vision.

These models use a use Pyramid Network decoder with a stride of 4, which is an essential component that improves their accuracy.

The CenterNet prediction heads further enhance the posture estimation capabilities of our model.

Our models are capable of accurate position estimation thanks to these heads and proprietary post-processing logic (MobileNets, Howard et al., 2017).

3.3 Model Training and Evaluation

3.3.1 Training Process

Our model-training journey is a multi-faceted process characterized by several key steps, each contributing to the refinement of our models:

Data loading and Augmentation: Our preprocessed data is loaded with care, and data augmentation techniques are expertly applied. This step is pivotal in enhancing the diversity of our training dataset, thereby bolstering our model's robustness.

Hyperparameter Tuning: We embark on the fine-tuning of model hyperparameters, a critical endeavor that optimizes training performance and convergence. It is a delicate dance of parameter adjustments, where precision and expertise are paramount.

Depth Multiplier Setting: Depending on the selected model variant (Lightning or Thunder), we carefully calibrate the depth multiplier.

A value of 1.0 for Lightning and 1.75 for Thunder ensures efficient training (MobileNets, Howard et al., 2017).

3.3.2 Evaluation Metrics

To evaluate our model's prowess, we employ two key evaluation metrics:

Keypoint Mean Average Precision (mAP) with Object

Keypoint Similarity (OKS): This industry-standard metric serves as the yardstick to measure the quality of our keypoint predictions. It's a metric that has stood the test of time and is commonly used in the highly regarded COCO competition (Cao et al., 2018).

Inference Time: In the era of real-time applications, quantifying our model's efficiency is of paramount importance. We meticulously measure the time required for model inference on a single image, expressed in milliseconds. This metric speaks volumes about our model's efficiency (Cao et al., 2018).

3.4 Experimentation and Results

3.4.1 Experimental Setup

We execute experiments using our trained models on diverse datasets, including:

Active Dataset Evaluation Set: Comprising images sampled from YouTube fitness, yoga, and dance videos, this dataset captures various poses, motions, and self-occlusions pertinent to fitness and human motion analysis (Cao et al., 2018).

3.4.2 Results and Findings

We present the results of our experiments, including keypoint mAP scores for different attributes and categories such as gender, age, and skin tone.

These findings highlight the robustness and efficiency of our models across various scenarios, with a particular focus on the Active Single Person Image Set (Cao et al., 2018).

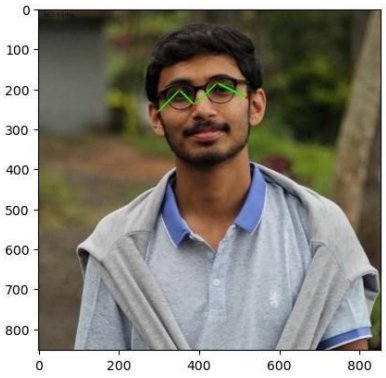
4 RESULTS AND DISCUSSION

4.1 Result

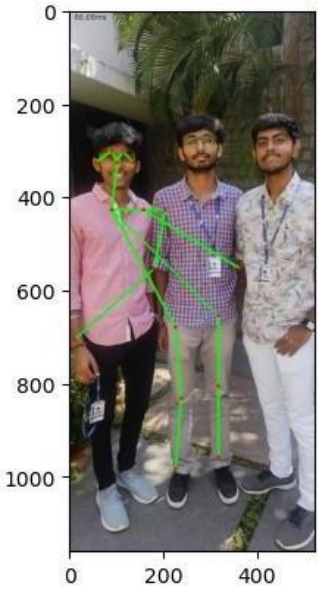
A photograph of results generated by the respective models is given below. These photographs represent the key point mapping that took place under static and dynamic feeds with single or multiple subjects at hand.

Human Pose estimation:

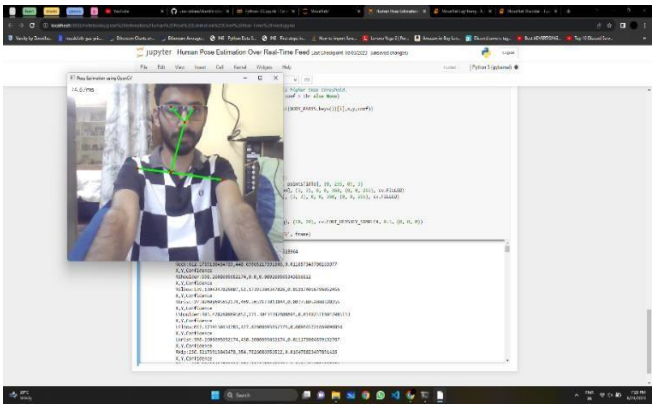
OpenPose model:



Static image frame with single subject;
Visible key points: 8
Mapped: 5
No issue finding pairs to map using available pair matrix
Background influence: NIL
Precise mapping: Moderate



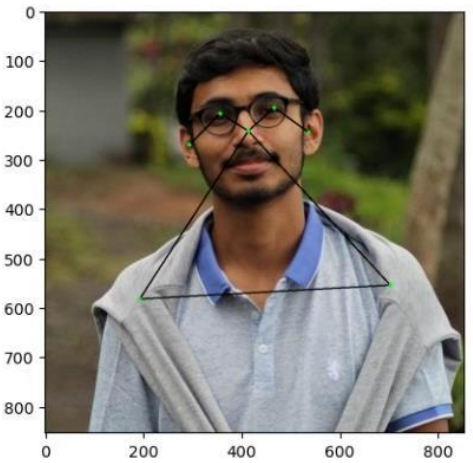
Static image frame with multiple subjects;
Visible key points: ~18*3
Mapped: 16
Has issue finding pairs to map using available pair matrix for individual subject.
Background influence: High due to availability of multiple subjects.



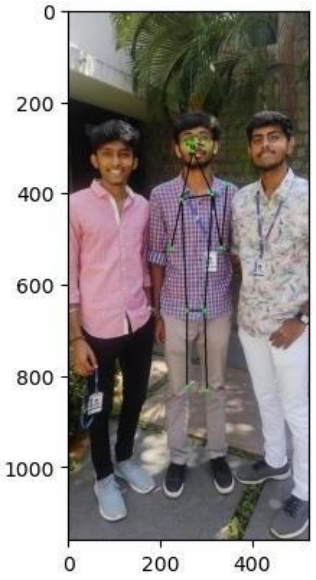
Dynamic image frame over video feed with single subject;
Visible key points: 10
Mapped: 6(at captured instance may vary)
Is slow in finding pairs to map using available pair matrix
Background influence: NIL

Accuracy on average over single subject (static/dynamic): 61.25%

MoveNet Lightning:



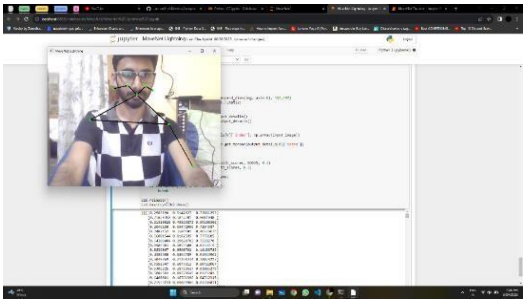
Static image frame with single subject;
Visible key points: 8
Mapped: 8
No issue finding pairs to map using available pair matrix
Background influence: NIL
Precise mapping: Better than OpenPose Model

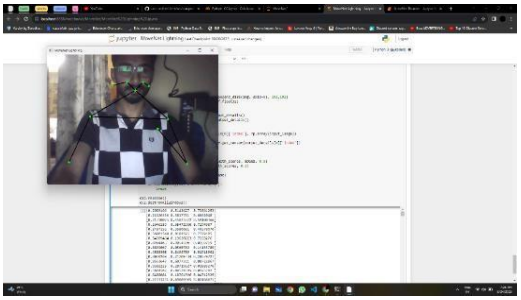


Static image frame with multiple subjects;
Visible key points: ~19*3
Mapped: 14

Has issue finding pairs to map using available pair matrix for individual subject but is better at segregating subjects and finding points on single subject rather than cross key point pairing.

Background influence: Low in comparison to OpenPose Model.

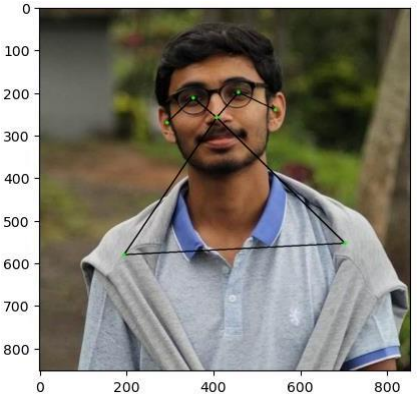




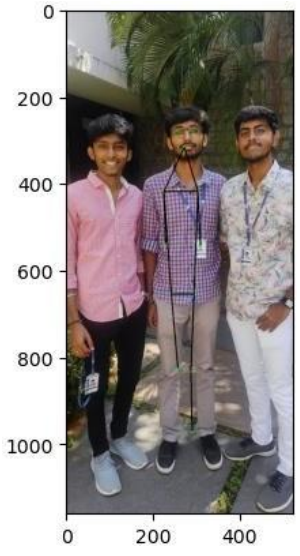
Dynamic image frame over video feed with single subject;
Visible key points: 9
Mapped: 8(at captured instance may vary)
Very fast in finding pairs to map using available pair matrix and does as better a job in low light conditions
Background influence: NIL

Accuracy on average over single subject (static/dynamic): 94.44%.

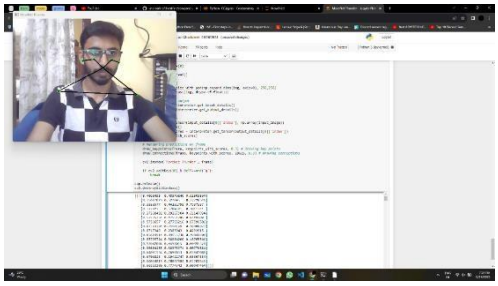
MoveNet Thunder:

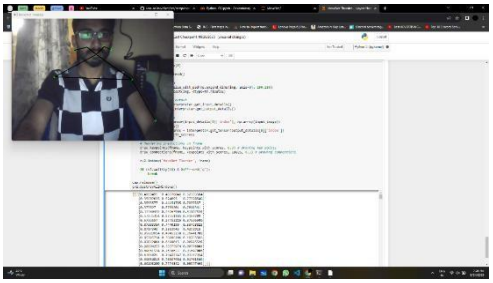


Static image frame with single subject;
Visible key points: 8
Mapped: 8
No issue finding pairs to map using available pair matrix
Background influence: NIL
Precise mapping: Best amongst the three



Static image frame with multiple subjects;
Visible key points: ~19*3
Mapped: 10
Has issue finding pairs to map using available pair matrix for individual subject but is better at segregating subjects and finding points on single subject rather than cross key point pairing and is precise.
Background influence: Low in comparison to OpenPose Model.
Moderate in comparison to MoveNet Lightning





Dynamic image frame over video feed with single subject;
Visible key points: 9
Mapped: 8(at captured instance may vary)
Very fast in finding pairs to map using available pair matrix and does as better a job in low light conditions
Background influence: NIL

Accuracy on average over single subject (static/dynamic): 96.29%.

4.2 Significance, Strengths and Limitations

The proposed work represents a significant advancement in the field of pose estimation and vision applications. Its contributions lie in the development of real-time pose estimation models that can be applied across a range of domains, from fitness and motion analysis to wildlife conservation and sports performance evaluation. The key significance of this work includes:

Real-time Performance: The models, especially the MoveNet variants, excels in real-time performance, enabling instantaneous pose analysis that can be valuable in various applications.
Accuracy: The use of advanced architectures and training techniques ensures high accuracy in key point detection, making the system reliable for precise pose estimation.

Versatility: The model's adaptability across different scenarios, such as fitness, demonstrates their versatility and potential to revolutionize multiple fields.

Several strengths characterize the proposed work include:
- State-of-the-Art Models
- Real-world Applicability

State-of-the-Art Models: The utilization of state-of-the-art models, including MoveNet and MobileNet, forms a robust foundation for accurate pose estimation. As exemplified by the MoveNet variants, ensures that the models can operate seamlessly on resource-constrained devices.

Real-world Applicability: The extensive experimentation and case studies demonstrate the practical applicability of the models in real-world scenarios, reinforcing their utility.

4.2.1 Limitations:

Limitations and Negative Cases:
• Multi-Person Pose Estimation - when multiple individuals are present in the frame, especially in close proximity or with overlapping body parts, the model may struggle to accurately associate body parts with the correct individuals.

• Occlusions - where one person partially or fully covers another, can lead to inaccuracies in key point detection
• Interactions - instances where subjects interact or overlap, such as during physical activities or team sports, can pose challenges for the model.

• Pose Ambiguity - where the model misinterprets the poses due to overlapping body parts or similar key point configurations.
The main limitation of proposed work at this moment in time is that all the modules have various ways in which they process multiple subject in a given frame.
The OpenPose model has very low threshold even that point is the best case for the model, on average, it is very low in comparison to any threshold standard and the model is a poor performer in estimation with precision.

While the MoveNet model has a better threshold value it adheres by in comparison to OpenPose Model, it is still low by any standard but is a significant improvement and is better that its counterparts.
We do have a significant development in model optimization everyday so, we can expect this to be higher in a few years

down the line.
The model fails in multi subject framed inputs, MoveNet seems to do a bit optimization to figure out a singular subject to focus on and estimate key points over it but the OpenPose model does cross key point pairing for best available points with greater threshold.

4.2.2 Summary:

The results of this project underscore the effectiveness of the proposed real-time pose estimation models, namely MoveNet and MobileNet, in a variety of practical applications. Through extensive experimentation on diverse data the models consistently demonstrated high accuracy in keypoint detection across various scenarios. These scenarios include real time feed motion key point mapping and static frame/dynamic frame over static instance mapping. Furthermore, the discussion of our findings delved into the significance of the models' real-time capabilities, their adaptability across domains, and their potential to drive transformative advancements in computer vision. However, it's important to acknowledge the limitations, particularly in multi-person scenarios, where the model's accuracy can be challenged by occlusions, interactions, and pose ambiguity. These findings emphasize the need for ongoing research and improvement in addressing such challenges and expanding the models' capabilities.

In conclusion, this project's results highlight the practical utility of real-time pose estimation models in diverse applications while recognizing the importance of ongoing refinement to enhance their performance in complex multi-person scenarios.

4.3 Cost Benefit Analysis

OpenPose Model:

The OpenPose model does poor at a threshold of 0.2 whereas when lowered down to 0.1 it does a decent job of mapping out points in any given frame even if not precise enough; if you're looking for more key point mapping during an instance you need to lower the threshold even further which speaks to how inferior it is when compared to its pears.

Single	Multiple	Real- Time
X, Y, Confidence 389.41304 34782609 241.065217 39130434 0.62592738 86680603 X, Y, Confidence	X, Y, Confidence 102.13043 47826087 327.826086 95652175 0.80166751 14631653 X, Y, Confidence	X,Y,Confidence Nose:570.4347826086956,459.1304347826087,0.016880771 14522457 X,Y,Confidence Neck:598.2608695652174,448.69565217391306,0.05249322 950839996 X,Y,Confidence RShoulder:598.260869565217

OpenPose model

MoveNet:

Both the MoveNet Lightning and Thunder model do fairly well when it comes to key point mapping under a better threshold; depending on the version and the type of model make you use the threshold may vary between 0.3 and 0.4, which is a recommended standard. In comparison, it does very well even at a base threshold of 0.2, which seems over valued for the OpenPose model.

MoveNet Lightning is the faster model and MoveNet Thunder is the more precise model when the counterparts are weighed against each other.

Each offer a tradeoff, which works seamlessly during integration.

Below listed table shows, an example from above results, which includes X, Y and Confidence parameters, used to map out the frame and each distinct key point.

These points describe the confidence level of each body point in the frame and is available for all 18 points during a single estimation cycle.

X and Y describe where these points are in the frame; be it Static/ Dynamic with single or multiple subjects in the frame.

Single	Multiple	Real- Time
[[[[0.2851104 0.48624766 0.6109643] [0.23056468 0.5485563 0.49754462] [0.24392894 0.41881627 0.5488925] [0.28912222 0.6306698 0.5785037]	[[[[0.26692003 0.5205884 0.35118547] [0.2523526 0.52841175 0.5113513] [0.25077128 0.5047841 0.41297564] [0.25783587 0.5761813 0.45540148]	[[[[0.27691334 0.5372794 0.70683086] [0.2170013 0.5926046 0.6885197] [0.20651405 0.48605072 0.6622773]

MoveNet Lightning

Single	Multiple	Real- Time
[[[[0.30414122 0.47837102 0.73837686] [0.23499544 0.53803414 0.67633843] [0.25028116 0.4154021 0.6559211] [0.28303066 0.6371125 0.77646536]	[[[[0.2733925 0.5219095 0.31457877] [0.25937876 0.5411877 0.2148313] [0.26320675 0.5052686 0.28163186] [0.27070993 0.5581452 0.28184232]	[[[[0.26390132 0.58781195 0.8084677] [0.1876953 0.6516216 0.7058042] [0.18516369 0.5117545 0.63673234] [0.23534952 0.6836863 0.6703287]

MoveNet Thunder

5 CONCLUSION

In conclusion, the proposed work offers a promising solution for human pose estimation over real-time feed. The models seem to have significant strengths and would be a great asset to many key developments in the field of Computer Vision and its applications. It has several strengths that make it an ideal choice for various domains. It is proven to be effective and robust over real time feed when we are dealing with single subjects; with that in mind we also need to take into consideration that there are certain limitations that need to be addressed, the proposed work provides us with an effective method for addressing key issues in human posture correction, posture estimation, and prosthetic body part development. With further advancements in tech architecture and hardware, we can expect this model to be even more accurate and versatile in the future.

6 References

[1] MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications (Howard et al., 2017)
[2] OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields (Cao et al., 2018)
[3] Newell, A., Yang, K., & Deng, J. (2016). Stacked Hourglass Networks for Human Pose Estimation. In European Conference on Computer Vision (pp. 483-499). Springer.
[4] Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional Pose Machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4724-4732).
[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).
[6] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In European Conference on Computer Vision (pp. 740-755). Springer.
[7] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 834-848.
[8] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.
[9] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single Shot MultiBox Detector. In European conference on computer vision (pp. 21-37). Springer.
[10] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779-788).
[11] Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of residual networks using large learning rates. arXiv preprint arXiv:1708.07120.
[12] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 88(2), 303-338.