Comparision Of ML Models For Posture

Utsav Kaim
Dept. of Electronics and Telecommunication Engineering
Sardar Patel Institute of Technology
Mumbai, India

Vyoum Khare
Dept. of Electronics and Telecommunication Engineering
Sardar Patel Institute of Technology
Mumbai, India

Aryan Jaiswal
Dept. of Electronics and Telecommunication Engineering
Sardar Patel Institute of Technology
Mumbai, India

Prof. Manish M Parmar
Dept. of Electronics and Telecommunication Engineering
Sardar Patel Institute of Technology
Mumbai, India

***Abstract*—**In computer vision, Human pose estimation (HPE) has come up as an important study area, with applications ranging from sports movement analysis, health support to video surveillance. Over the past couple of decades, the need for HPE has created a number of HPE libraries. In current time, skeleton-based HPE algorithms have gained momentum leading to development of libraries that facilitate their use by researchers. Therefore, these libraries' performance is key in their adaptation into practical use. Despite this fact, there lacks an extensive benchmarking assessment on these libraries presently. This article seeks to fill this void by examining the pros and cons of two skeletal-based HPE libraries: YOLOv7 and MediaPipe Pose. For both image and video datasets, we conduct a comparative study between these libraries.

***Index Terms*—**Human Pose Estimation; Yolov7; MediaPipe Pose

I. Introduction

Human Pose Estimation (HPE) means identification and localization of all body parts from input images or videos. A major area in computer vision is HPE, which has possibilities in video surveillance [1-6] and medical assistance [7-15]. These fields employ human key-points for pose classification and movement accuracy assessment. For instance, intelligent video surveillance systems can use human key points to identify poses during suspicious or criminal activities. Similarly, in medical assistance, detected key-points can evaluate posture accuracy for fall determination, at-home rehabilitation, and physical therapy activities. Also, sports analysis could involve comparing detected key-points with reference poses in order to analyse an athlete's performance.

According to quantity of people in the original picture, HPE can be divided into 2D and 3D HPE as well as single-person and multi-person HPE. Skeleton key-points detection method separates single-person from multi-person HPE which further classifies them into top-down and bottom-up[29]. This research is dedicated to 2D single-person HPE: a comparison of different algorithms.

The mushrooming requirement of HPE resulted in the emergence of several skeletal-based HPE techniques that were created into libraries for usage by scientists. These libraries' performance is important so as not to compromise their integration into real-world applications. Take for instance the case of an HPE library utilized in a home rehabilitation system that should accurately detect patient poses in different environments within homes for it to work effectively. This becomes more complicated when dealing with typical difficulties like wrong camera placement and self-occlusion [11].

II. Literature Review

A. Human Pose Estimation Based on Deep Learning

The field of pose estimation based on deep learning has made significant progress since Google introduced DeepPose in 2014. Usually, such algorithms typically operate in two phases.

- Person determination
- key-point Location

These algorithms can be divided into top-down and bottom-up approaches based on which stage occurs first.

1) Top-Down Approach: Under the top-down approach, each person is identified first, and then each person's landmarks are located. The more people there are, the more difficult the computation becomes. These methods provide good accuracy performance on widely-used benchmarks and are scale-invariant. However, obtaining real-time inference is highly computational because of these models' complexity.

2) Bottom-Up Approach: The bottom-up method identifies identity-free landmarks (key-points) for each person in an image at the same time, then groups these key-points into distinct persons. This method estimates the probability that each pixel contains a certain landmark (key-point) using a probabilistic map called a heat-map. Next, non-maximum suppression is used to eliminate the best landmark. While these techniques are often less accurate than top-down ones, they are also less complex.

B. Human Pose Estimation in Real Time

The performance of various pose estimation frameworks can vary significantly depending on the hardware used, such as CPU, GPU, or TPU. Many two-stage pose estimation models, including Alpha Pose, OpenPose, and Deep Pose, excel in benchmark tests. However, real-time performance is highly

computational due to the advanced nature of these models, particularly on CPUs. They tend to run efficiently on GPUs but struggle with speed on CPUs.

MediaPipe stands out as a well-balanced framework in terms of efficiency and accuracy, achieving continuous detection on CPUs. Hence, we evaluated the YOLOv7 Pose to compare its performance against MediaPipe.

C. Current Comparative Analysis

In two investigations using comparative analysis, using image datasets, scientists evaluated the effectiveness of YOLOv7 and MediaPipe for human pose estimation. In study [13], the AR dataset was utilized to evaluate YOLOv7 and MediaPipe, with PCK@0.2 as the evaluation metric. YOLOv7 achieved slightly better results, scoring 87.8 compared to MediaPipe's 84.1.

A more extensive comparison was done in reference [14] which compared YOLOv7 and MediaPipe against other HPE libraries such as PoseNet, OpenPose, MoveNet Thunder, and MoveNet Lightning. The study employed the COCO and MPII image datasets then introduced a novel evaluation metric for measuring performance. It turned out that YOLOv7 outperformed all of them with Mediapipe coming second. These analyses went beyond image sets into evaluating both video as well as image data set.

These comparisons show how effective YOLOv7 and MediaPipe are in human pose estimation across different datasets including images and videos.

D. HPE Libraries

The next part of the article will provide an overview of two advanced Human Pose Estimation (HPE) solutions: MediaPipe Pose and YOLOv7. This section features a comparison of the specifications of these libraries with other state-of-the-art HPE solutions.

In total, 17 common key-points are detected in this library called YOLOv7. These are further subdivided into the following types: head-keypoints (ears, eyes, and nose – 5 keypoints), upper body keypoints (shoulders, elbows, wrists – 6 keypoints), and lower body keypoints (hips, knees and ankles – 6 keypoints).

On the contrary, MediaPipe Pose has more comprehensive annotations that capture up to 33 key-points.

There are two methods to approach key-point recognition in HPE libraries: top-down and bottom-up. MediaPipe Pose utilizes a top-down method, whereas YOLOv7 employs a unique single-stage multi-person key-point detection approach. Each library uses different underlying networks for pose estimation: MediaPipe Pose utilizes a Convolutional Neural Network (CNN), while YOLOv7 implements its own specialized architecture, demonstrating the diversity in their design and capabilities.

YOLOv7 Pose: YOLOv7 Pose showcases itself as a bottom-up approach-inspired single-level multiple person key-point identifier. Unlike traditional methods, this model excludes heat-maps from its key-point detection process. An

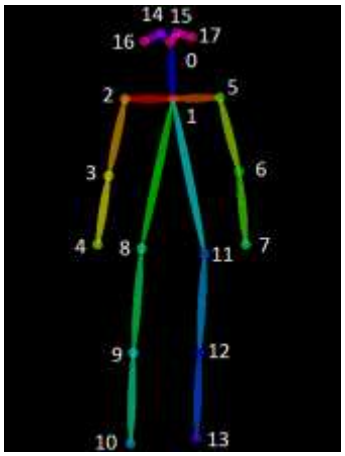

Fig. 1. key-points detected by YOLOv7

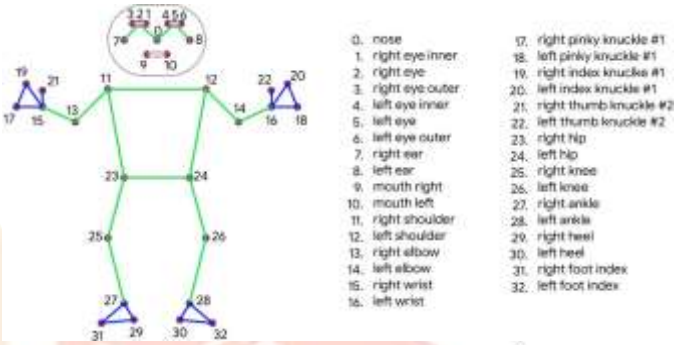

Fig. 2. key-points mapped by MediaPipe

evolution of YOLO-Pose, it integrates aspects from both top-down and bottom-up methodologies. YOLOv7 Pose, which is implemented in PyTorch and offers flexibility for code customisation according to individual demands, has been taught on the data set COCO with 17 reference topologies. Users can leverage a pre-trained key-point detection model, yolov7-w6-pose.pth, to streamline their workflow.

MediaPipe Pose: MediaPipe Pose is designed with single person pose estimate in mind. It utilizes BlazePose with a 33-landmark topology, which encompasses COCO key-points, BlazeFace, and BlazePalm. There are two phases to this model's operation: location and tracing, optimizing inference speed by employing a tracker post-initial detection. Notably, MediaPipe Pose is engineered for CPU-only support, ensuring efficient real-time inference on devices lacking dedicated GPUs. Furthermore, it seamlessly integrates segmentation, enabling smooth transition between pose estimation and segmentation based on a user-defined flag. Serving as a superset of COCO key-points, BlazePose within MediaPipe incorporates additional landmarks to enhance pose estimation accuracy.

III. METHODOLOGY

The datasets used in the experiments are described in this section. Following dataset selection, data pre-processing steps were executed. Subsequently, the Human Pose Estimation

Fig. 3. Sample Images from the COCO and PENN Dataset

(HPE) process was performed. An evaluation was carried out to test the effectiveness of the HPE library.

A. Datasets

The Microsoft COCO (Common Objects in Context) picture dataset and the Penn Action video dataset are the two datasets used in this investigation. Six critical sites for the upper body and six for the lower body are shared by both datasets. However, the key distinction lies in the annotation of head key-points. While Penn Action offers a single key-point at the head position, COCO has five key-points that cover the ears, nose, and eyes. Figure 3 shows sample photos from the Penn Action and COCO datasets.

COCO is an extensive dataset that is frequently used for tasks related to object identification, segmentation, and captioning.It includes 330,000 photos, 1.5 million item instances, 80 item categories, 91 item categories and annotations for 250,000 humans with key-points. Each person instance in COCO is annotated with 17 key-points, making it suitable for body key-point detection experiments. Various versions of COCO exist, with COCO 2017 being a common choice for Human Pose Estimation (HPE) experiments.

Conversely, the video dataset Penn Action has 2326 video sequences that depict 15 distinct human actions. In HPE experiments, it is often used. In addition to annotations describing human behaviors, 2D bounding boxes for human localization, and skeleton key-points denoting body sections, every video sequence has RGB image frames. Penn Action annotates each human experience with thirteen critical points.

B. Data Pre-processing

Data pre-processing was carried out to remove unnecessary data from both datasets before assessing the performance of the HPE libraries. In the COCO dataset, three types of images were identified: those featuring a single person, those with multiple people, and those without any individuals. Given the focus on single-person HPE in this experiment, images without people were removed, leaving only relevant images for analysis.

The dataset's 17 annotations served as the ground truth, and these were compared with the 17 often detected critical spots from the human body in order to obtain a good differentiation of effectiveness of the HPE libraries.

Since the video frames in the Penn Action movies only showed the topmost section of the human body, the action of strumming a guitar was not included. Furthermore, the head annotation was left out of the trials to guarantee uniformity across all libraries because the Penn Action dataset only included one annotation for a key-point for the head, which was different from the annotations utilized by the HPE libraries. Consequently, the remaining 12 key-points were utilized as the ground truth for evaluation purposes.

IV. Results

We selected two posture estimation machine learning models for our study: YOLOv7 and MediaPipe. Both models were evaluated on a laptop with following configurations: CPU: Intel Core i5-1135G7 (11th Gen Octa-core processor), RAM: 8GB, GPU: None. The evaluation of the models encompassed various scenarios, and the obtained results are outlined below.

A. Normal Conditions

Under normal sitting conditions, both models successfully detected the key-points. However, YOLOv7 exhibited slower performance, achieving a lower frames per second (FPS) rate of 4-5. In contrast, MediaPipe demonstrated faster performance, achieving an FPS range of 16-18, as depicted in **fig. 4**.

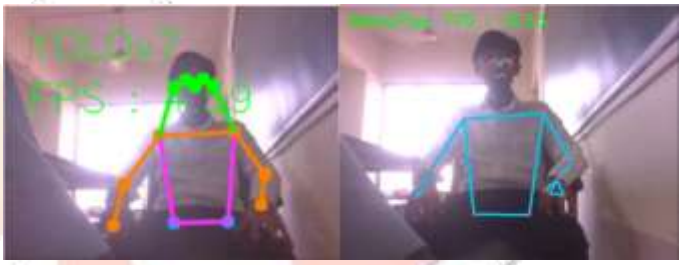

Fig. 4. Yolov7 vs MediaPipe (normal)

B. Dark Environment

In the experiment conducted in a low-light environment with the subject standing, both models successfully detected the key-points on the subject. However, YOLOv7 encountered difficulties in capturing the key-points compared to MediaPipe, as illustrated in **fig. 5**.

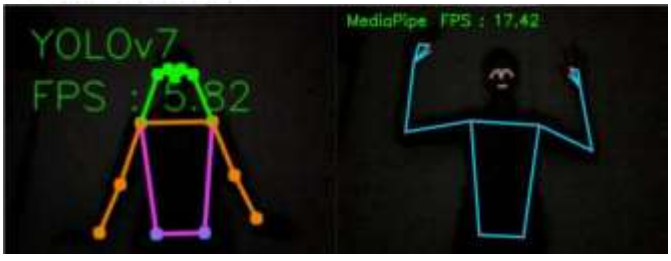

Fig. 5. Yolov7 vs MediaPipe (low-light)

C. Occlusion

In this experiment, where the subject concealed their right arm, YOLOv7 successfully detected the hand, whereas MediaPipe failed to do so, as evidenced in **fig. 6**.

Fig. 6. Yolov7 vs MediaPipe (occlusion)

D. Sitting Posture Estimation

Additionally, we implemented MediaPipe Pose to evaluate a person's sitting posture, distinguishing between good and bad positions. To achieve this, we measured the neck inclination and torso inclination relative to a vertical line. When the person tilts their head or torso in a manner that becomes uncomfortable, it is indicative of poor posture. These observations are depicted in **Fig. 7, 8, 9,** and **10**.

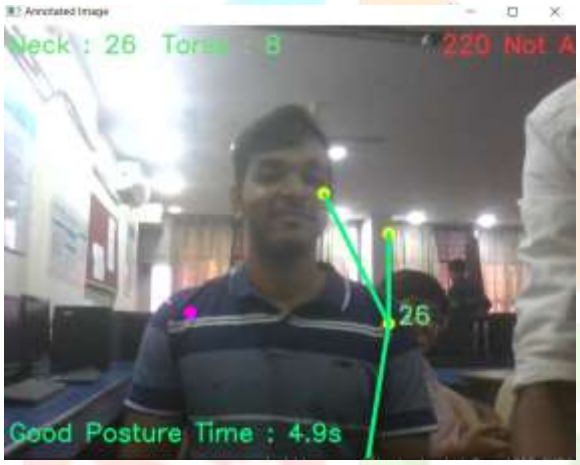

Fig. 7. Good sitting posture(front-view)

Fig. 8. Bad sitting posture(front-view)

Fig. 9. Good sitting posture(side-view)

Fig. 10. Bad sitting posture(side-view)

V. Conclusion

The current popularity of Human Pose Estimation (HPE) in computer vision stems from its wide range of real-world applications. Consequently, the effectiveness of HPE libraries holds significant importance. This paper has conducted a differential analysis of two HPE libraries, YOLOv7 and MediaPipe, with a focus on their strengths and weaknesses in live video HPE processing.

The findings suggest that MediaPipe offers certain advantages over YOLOv7, particularly in scenarios involving low-spec devices and the specific task of detecting sitting posture. MediaPipe demonstrates superior performance in processing low-resolution inputs, making it well-suited for situations with constrained computational resources. Its efficiency in CPU inference further enhances its applicability for deployment on devices with limited processing capabilities. While YOLOv7 exhibits faster performance due to its utilization of GPU acceleration, MediaPipe compensates by providing faster processing on CPU. The choice between GPU acceleration and CPU efficiency depends on the device's capabilities and the specific requirements of the posture detection application.

Moreover, MediaPipe's effectiveness in detecting far-away objects aligns well with the typical use case of monitoring sitting posture. However, it is important to acknowledge that YOLOv7 outperforms MediaPipe in scenarios involving occlusion, indicating its suitability for diverse environments. MediaPipe's limitation in handling multiple people concurrently contrasts with YOLOv7's ability to detect and track multiple individuals simultaneously, which becomes crucial in applications requiring monitoring of group activities.

In the context of the specific focus on sitting posture detection, especially on low-spec devices, MediaPipe emerges as a favorable choice due to its efficiency, CPU speed, and suitability for single-person scenarios. Nevertheless, the decision between MediaPipe and YOLOv7 ultimately hinges on the specific requirements, device capabilities, and acceptable trade-offs for the intended application.

ACKNOWLEDGMENT

It is with deep gratitude that we acknowledge the assistance and support that have contributed to the completion of this thesis. We would like to sincerely thank our professors for their crucial guidance. Without their encouragement, this work would not have come to fruition. We are truly grateful for the freedom they provided, allowing us to thoroughly enjoy the process under their mentorship.

We would like to express our gratitude to the Electronics and Telecommunication Engineering Department Head and Faculty for providing us with all the resources required to carry out this research.

Additionally, we would like to sincerely thank all of our family members and well-wishers for their continuous support throughout the years. This job would not have been possible without their assistance.

REFERENCES

[1] Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Pro- ceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 29 October 2017; pp. 3960–3969.

[2] Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; Ouyang, W. Attention-aware compositional network for person re-identification. In Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 23 June 2018; pp. 2119– 2128

[3] Thyagarajmurthy, A.; Ninad, M.G.; Rakesh, B.G.; Niranjan, S.; Manvi, B. Anomaly detection in surveillance video using pose estimation. In Proceedings of the Emerging Research in Electronics, Computer Science and Technology, 2019; Springer: Singa- pore, 2019; pp. 753–766. Available online: https://link.springer.com/chapter/10.1007/978-981-13- 5802-9 66/ (accessed on 27 Oc- tober 2022).

[4] Lamas, A.; Tabik, S.; Montes, A.C.; Pe´rez-Herna´ndez, F.; Garc´ıa, J.; Olmos, R.; Herrera, F. Human pose estimation for mitigating false negatives in weapon detection in video-surveillance. Neurocomputing 2022, 489, 488–503.

[5] Yoo, H.R.; Lee, B.H. An openpose-based child abuse decision system using surveillance video. J. Korea Inst. Inf. Commun. Eng. 2019, 23, 282–290.

[6] Park, J.H.; Song, K.; Kim, Y.-S. A Kidnapping Detection Using Human Pose Estimation in Intelligent Video Surveillance Sys- tems. J. Korea Soc. Comput. Inf. 2018, 23, 9–16. https://doi.org/10.9708/JKSCI.2018.23.08.009

[7] Chang, Y.J.; Chen, S.F.; Huang, J.D. A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. Res. Dev. Disabil. 2011, 32, 2566–2570.

[8] Hassan, H.A.; Abdallah, B.H.; Abdallah, A.A.; Abdel-Aal, R.O.; Nu- man, R.R.; Darwish, A.K.; El-Behaidy, W.H. Automatic Feed- back For Physiotherapy Exercises Based On PoseNet. FCAI-Inform. Bull. 2020, 2, 10–14.

[9] Shapoval, S.; Garc´ıa Zapirain, B.; Mendez Zorrilla, A.; Mugueta- Aguinaga, I. Biofeedback applied to interactive serious games to monitor frailty in an elderly population. Appl. Sci. 2021, 11, 3502.

[10] Chua, J.; Ong, L.Y.; Leow, M.C. Telehealth using PoseNet-based system for in-home rehabilitation. Future Internet 2021, 13, 173

[11] Kim, W.; Sung, J.; Saakes, D.; Huang, C.; Xiong, S. Ergonomic postural assessment using a new open-source human pose esti- mation technology (OpenPose). Int. J. Ind. Ergon. 2021, 84, 103164

[12] Jawale, C.D.; Joshi, K.A.; Gogate, S.K.; Badgujar, C. Elcare: Elderly Care With Fall Detection. J. Phys. Conf. Ser. 2022, 2273, 012019.

[13] Jo, B.; Kim, S. Comparative Analysis of OpenPose, PoseNet, Yolov7, Mediapipe Pose and MoveNet Models for Pose Estimation in Mobile Devices. Trait. du Signal 2022, 39, 119–124

[14] Gadhiya, R.; Kalani, Tinier-YOLO: A Real-Time Object Detection Method for Constrained Environments Available online: https://ieeexplore.ieee.org/abstract/document/8941141

[15] Chen, W.; Jiang, Z.; Guo, H.; Ni, X. Fall detection based on key points of human-skeleton using openpose. Symmetry 2020, 12, 744