**Critical Review on “OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields”**

# Introduction

This paper proposes a real-time approach to detect the 2D pose of multiple people in an image. OpenPose is an open-source multi-person detection system supporting the body, hand, foot, and facial key points. Human pose estimation is focused on finding individuals and their pose from an image. This problem is quite challenging and is quite complex to determine the number of people that can appear in an image. Moreover, the interferences caused due to contact or proximity between people makes the association of parts quite challenging.

The current state of the art techniques uses a top-down strategy that estimates the pose of each person on each of the detected region. This approach suffers from performance issues because the run time is directly proportional to the number of people in the image.

The method used in the paper uses a nonparametric representation called Part Affinity Fields (PAFs). This approach uses the location and orientation of body parts to encode unstructured pairwise relationships between the body parts of the individuals in the image. In the earlier version of the same paper, they refined both the confidence maps and the part affinity fields (PAFs) at each stage using a branched multistage Convolutional neural network (CNN) and this resulted in much more computation and time at each stage. [1]

This paper focused on refining the existing techniques and the architecture of the CNN [1]. They found that PAF refinement is crucial for maximising accuracy, while body part prediction refinement is not that important. In the paper, the author used the first 10 layers of the VGG-19 model as a baseline network to extract feature maps. These feature maps are input to the model to derive the PAFs that define the degree of association between the parts and the confidence maps, which are then refined using the greedy approach.

The paper **OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields** is published by Zhe Cao, Student Member, IEEE, Gines Hidalgo, Student Member, IEEE, Tomas Simon, Shih-En Wei, and Yaser Sheikh under IEEE Transactions on Pattern Analysis and Machine Intelligence journal.

# Summary of the work, results, and findings

As the first stage, the feature maps of the input image are processed by passing it through a baseline network. In this paper, the author uses the first 10 layers of the VGG-19 model for generating the feature maps. Then the feature maps are processed with multiple stages of CNN to generate a set of Part Confidence Maps and a set of Part Affinity Fields. The system uses a multi-stage CNN. It consists of two types of stages. The first set of stages predicts part affinity fields i.e., degree of associations between the body parts the last set of stages predicts the confidence maps. By replacing each 7x7 convolutional kernel with 3 consecutive 3x3 kernels, the current model was able to preserve the receptive field while computation expense is reduced. Using an approach like DenseNet [11] the output of the 3 convolution kernels is concatenated. This approach tripled the non-linearity layers and helped the model to keep both higher and lower-level features.

A loss function is applied at the end of every step in both branches. Finally, the Confidence Maps and Part Affinity Fields are processed by a greedy algorithm where optimal partite graph matching is used to obtain the poses for each person in the image

The results and findings are presented, and the accuracy of the multi-person key point detection has been compared to the existing methods [2][3]. The real-time performance metric for the system has been specified by mentioning the number of frames it can process each second. All findings are visualized, and graphs are provided for better understanding. The author finds that the newly refined approach increased both the speed and accuracy by approximately 200% and 7% respectively.

The paper contributed to the field of foot detection by presenting an annotated foot dataset with 15K human foot instances that has been publicly released for further study and development. The author open-sourced their work and the OpenPose system provide easy-to-use pipelines with command-line interfaces, python API, Unity plugin. It has also been included in the OpenCV library for ease of access. The system runs on CPUs and GPUs such as CUDA and OpenCL. They also provide portable executable binaries which can easily be downloaded and run with minimal setup.

Pose-ShuffleNet is a lightweight multi-person pose estimation method developed by integrating the concepts of this paper [4]. This method is highly suited for resource-constrained scenarios. It is based on the building blocks of the ShuffleNet unit [5] for image analysis in place of VGG-19. An architecture of multi-stages with two branches, jointly learning parts detection and parts association has been used. Pose-ShuffleNet has a significantly lower resource (network size and FLOPs) consumption than some of the available leading methods. Using the concepts of OpenPose the research has been further extended to the 3D human pose estimation [6]. Further research on Intention Recognition of Pedestrians and Cyclists by 2D Pose Estimation has been carried out. Overall, the proposed pipeline provides promising results on the intention recognition of vulnerable road users. [7]

# Pros and Cons of the paper:

Explanation with pictorial references clarifies the topic in a better way and was effective in clarifying relationships, simplifying complex topics and was one of the main key factors that had me inclined towards the paper. All the edge cases have been acknowledged and future actions that can mitigate the failure cases has also been suggested. The accuracy and performance of the combined model with body and foot key points is also an important contribution by the paper.

The generality of the method by applying it to the task of vehicle key point estimation and the use of the approach has contributed to the wide applications and competitive performance of the model. The paper also showcased three benchmarks performed on reputed and easily available datasets such as MPII [9], COCO [10] and a subset of COCO used as a foot dataset. And provide runtime analysis comparison against Mask R-CNN and Alpha-Pose to quantify the efficiency of the system and analyses the main failure cases.

The model performs up to 3 times faster on GPUs when compared to CPUs and in case of Body and foot model the performance difference is up to 30 times. Making the system more optimised on CPUs can help to make the model more accessible in smaller and less powerful systems and increase the use cases. [8]   
  
Potential Advancement and Future Work  
There have been many advancements in the field of pose estimation, and it is being used in various fields and applications, but at present, the practical applications are limited due to the challenges of crowded people and occluded body parts. A hybrid system of both top-down and bottom-up approaches along with Using anthropometric ratios for determining context can make the model more efficient [7]. Future developments can be made to generate a similar performance on CPUs using lightweight CNNs such as MobileNetV1, EfficientNet or a combination and can help the system be more efficient so that it is less reliant on GPUs and work on embedded systems. It can have various applications in security cameras, child monitors, automating warehouses, gesture control etc.

Data plays a key role in improving the robustness of the network. Adding complex scenarios of human poses along with samples of human statues and paintings would boost the accuracy by avoiding the problem of occluded body parts.

Improvement in the greedy approach can help to make the pose estimation more accurate. The greedy approach used in the current system suffers from limited structural context when subjects are occluded. Also, the Hungarian algorithm [8] employed is having a cubic complexity. By using a larger structural context of a subset of pre-assigned parts rather than using only the immediate predecessors as in [7], this time complexity can become linear in the number of candidates of any single part-class.

Refining the algorithm to reject less likely or complex poses and adding an animal dataset to train the system can have wide applications in pest control, animal farming, analysis of the bees as shown in Honeybee Detection and Pose Estimation using Convolutional Neural Networks [10].

# References

[1] Cao, Zhe, et al. "Realtime multi-person 2d pose estimation using part affinity fields." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[2] Fang, Hao-Shu, et al. "Rmpe: Regional multi-person pose estimation." Proceedings of the IEEE international conference on computer vision. 2017.

[3] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

 [4] Guan, Chen-zhi. "Realtime multi-person 2d pose estimation using shufflenet." 2019 14th International Conference on Computer Science & Education (ICCSE). IEEE, 2019.

 [5] Zhang, Xiangyu, et al. "Shufflenet: An extremely efficient convolutional neural network for mobile devices." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[6] Schwarcz, Steven, and Thomas Pollard. "3d human pose estimation from deep multi-view 2d pose." 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018.

[7] Fang, Zhijie, and Antonio M. López. "Intention recognition of pedestrians and cyclists by 2d pose estimation." IEEE Transactions on Intelligent Transportation Systems 21.11 (2019): 4773-4783.

[8] Osokin, Daniil. "Real-time 2d multi-person pose estimation on cpu: Lightweight openpose." arXiv preprint arXiv:1811.12004 (2018).

[9] Andriluka, Mykhaylo, et al. "2d human pose estimation: New benchmark and state of the art analysis." Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014.

[10] Rodríguez, I. F., et al. "Honeybee detection and pose estimation using convolutional neural networks." Congres Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP). 2018.

[11] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.