Critical Review : OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields

Sarathkumar P S [2359859]
Course Code [06-32257]
University of Birmingham
sxp208@student.bham.ac.uk

March 14, 2022

I. Introduction

i. Provide name of the article and publication details

This document conducts a critical review of the journal article, OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields[2], published in 2018 at arXiv. The journal article follows the original paper [3] presented in the 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR) and extends the original work by improving the runtime performance and accuracy.

ii. What the paper is about

Realtime multi-person 2D pose estimation is a key component in enabling machines to have an understanding of people in images and videos, however, inferring the pose of multiple people in images presents a unique set of challenges such as position and scale of people in the image, spatial interference and runtime performance. In this paper authors present a realtime approach to detect the 2D pose of multiple people in an image; moreover, the authors claim that the same approach can be used to any keypoint annotation task and demonstrates this by running the same network architecture for the task of vehicle keypoint detection.

iii. how does it fit into current work in that field

The research studies in the area of pose estimation evolved from identifying human pose on a single person image to multi-person image; thus, initial approaches to solve the multi-person pose estimation relied heavily on identifying persons in the image first and then applying available single person pose estimation solutions for each person in the image, this approach is called top-down approach. There are several challenges to this approach, firstly the runtime is proportional to the number of people in the image, secondly this approach doesn't capture dependencies across different people in the image, finally accuracy of this approach depends heavily on the person detection model used.

iv. how does it approach the problem.

In this paper, authors approach the multi-person pose estimation problem in a bottom-up manner, ie, identify all the keypoints/parts present in the image first, then identify the limbs and finally construct the human pose. Though there were studies using bottom-up approach to solve the multi-person pose estimation problem, they all suffered with high inference time, in the order of minutes and hours. To improve the inference time and accuracy, authors introduces a representation, Part Affinity Fields(PAFs), consisting of a set of flow fields that encodes unstructured pairwise relationships between body parts of a variable number of people.

II. Summary of the work, results & findings

Below are the main steps involved in the proposed method

1. Firstly, the image is analyzed by a Convolutional Neural Network (CNN) to extract feature maps.

2. Secondly, the feature maps are processed with multi-stage CNN to generate:

    - A set of Part Confidence Maps (Assign each pixel in the image a confidence score indicating the chance that part j is present at that location, this is done for all the part types)

  - A set of Part Affinity Fields (PAFs) which encodes the degree of association between the parts.

3. Finally, the Confidence Maps and Part Affinity Fields are processed by a greedy algorithm to obtain the poses for each person in the image.

The authors evaluates the above model with 3 different datasets, MPII Human Pose Dataset[1], COCO Keypoint Challenge Dataset[13] and Human Foot Keypoint Dataset.

i. What does the paper present in terms of results, their conclusions and contribution to the science

The authors claim that the proposed method outperforms [11] in MPII Human Pose Dataset[1] by 8.5% on mAP(Mean Average Precision) and the inference time is 6 orders of magnitude less, approximately 0.005 seconds per image. On COCO Keypoint Challenge Dataset[13] authors approach is less accurate but the inference time is lowest; thus, authors approach is best suitable for scenarios where accuracy is less important than the speed.

Authors put forth an argument that current benchmarks evaluates the performance of a model mainly by looking at the accuracy, however, inference time is also an important aspect. Authors state that the current gap in accuracy between top-down and bottom-up approaches is due to the input image resolution limitations; nevertheless, authors are optimistic that in future when hardware gets faster and increases its memory, bottom-up methods might be able to work with higher resolution images and reduce the accuracy gap with respect to top-down approaches.

Additionally, authors have created a foot keypoint dataset consisting of 15K foot keypoint instances and have open-sourced this work as OpenPose library, the first realtime system for body, foot, hand, and facial keypoint detection.

ii. How does it fit into other published works in this field

Accuracy of the authors model is not best among the other published works in this field such as [9], [6],[4],[17],[14],[7], but, the runtime analysis shows that it outperforms most of these models still providing comparable/usable level of accuracy. Moreover, the runtime for the authors approach doesn't depend on the number of people in the image, most of the other model run times depends linearly on the number of people in the image. Hence I believe this paper will have wider acceptance among certain industries as it provides human pose estimation at a usable level of accuracy with small response time.

iii. What does it contribute to better understanding of the problem

Authors claims that combining body and foot estimation into a single model boosts the accuracy of each component individually and reduces the inference time of running them sequentially. This observation means that adding more keypoints to the human pose estimation problem actually helps the network to learn better and reduce spatial interference related issues. Furthermore, authors demonstrate that a greedy parsing algorithm is sufficient to get results with comparable accuracy while reducing the runtime significantly.

III. Pros/Cons of the Paper

i. What do you like about the paper?

- Authors has provided a detailed analysis on runtime of the model and how it compares with the other state of the art models.
- Authors demonstrate that their model can be used to any keypoint annotation task which opens a lot of application possibilities.
- Authors released foot keypoint dataset and open sourced the OpenPose library as part of the work on this paper which has helped a lot of studies and applications since 2018.
- As of today, OpenPose library has got 23.5K stars on github with 92 contributors and is well maintained, which makes this paper still relevant for the industry as well as research studies.

ii. What stands out as being significant?

- Inference time of the model is extremely good such that it can be used in industry level applications.
- PAF representation (encodes both location & direction of a point) introduced in this paper has potential applications in other research areas as well.
- Greedy parsing approach introduced in this paper can be applied to other similar problems to reduce running time.

iii. What don't you like about the paper?

- The authors could have added accuracy & inference time comparison with more models, especially considering there exists other models with higher accuracy such as [15],[12] etc released at the same of time of publishing this paper.
- It would have been helpful if the authors wrote a section describing the model from an explainability point of view.

iv. Was there any significant gaps in what has been presented?

- Authors hasn't given a clear rationale about why they chose first 10 layers of VGG-19 as the baseline CNN network to extract features before passing the image to PAF stage.
- Though the authors mentioned about accuracy of the Associative Embedding architecture[14] in the paper, authors does not mention anything about the runtime of this architecture. Considering that Associative Embedding approach slightly higher accuracy than authors model, it would have been nice to get a comparison on time as well.
- Figure 13 in the paper shows 2 data point for the OpenPose algorithm and has mentioned that it is for images with different resolutions, but, there is no information available on the figure about the resolution of image.

IV. Potential Advancement & Future Work

i. Based on your literature search above and wider reading, what could have been done to make the work more substantial?

- More comparisons and experiments on the model based on images of different resolution would have given the reader a better understanding of the model and its application scenarios.
- Authors could have tried initializing the baseline network with a different approach, such as ResNet50, instead of VGG-19 and checked whether it improves the result.

ii. How would you recommend the work to continue on this topic?

- Since the release of the paper *Attention is All You Need*[16], there were some studies, [5], experimenting to use attention based mechanism in computer vision tasks. The authors approach to get the PAF & confidence maps currently relies on CNN network only, we can study this architecture further conduct experiment to see whether attention based mechanism improves the accuracy or speed of the authors approach.
- Like papers [10], [8], there is a scope for evaluating the performance of the authors approach from an edge device perspective and can be checked for improvement specifically for edge devices.

iii. What alternative approaches, analysis and testing would you recommend.

- The performance can be evaluated on additional datasets such as PoseTrack
- Analyze the performance on images with different resolutions.
- As the approach is multistage, Part Confidence Maps stage and Part Affinity Fields (PAFs) stage, an analysis of the loss function at each stage would be helpful to understand the gaps & scope for improvement.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1812.08008, 2018.

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017.

[7] Mihai Fieraru, Anna Khoreva, Leonid Pishchulin, and Bernt Schiele. Learning to refine human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 205–214, 2018.

[8] Daniel Groos, Heri Ramampiaro, and Espen Alexander Fürst Ihlen. Efficientpose: Scalable single-person pose estimation. *Applied Intelligence*, 51:2518–2533, 2021.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[10] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *ArXiv*, abs/1704.04861, 2017.

[11] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European conference on computer vision*, pages 34–50. Springer, 2016.

[12] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In *Proceedings of the european conference on computer vision (ECCV)*, pages 713–728, 2018.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[14] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017.

[15] Guanghan Ning, Zhi Zhang, and Zhiquan He. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia*, 20(5):1246–1259, 2017.

[16] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.

[17] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018.