AI6121: Computer Vision

Direct Reading and Literature Review

Ron Kow Kheng Hui

ID: G1903451J

1 October 2021

School of Computer Science and Engineering

Nanyang Technological University

Contents

1 Introduction

This report summarizes and describes the research presented in the paper *OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields* by Cao et al., 2019 [5]. This paper is an extension and improvement of the authors' highly cited earlier work, *Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields* 2017 [6]. These two papers present a novel approach to detect the different body parts (torso, arms, legs, facial features, etc.) of multiple persons in different poses in an image.

In the literature, this task is known as *multi-person pose estimation* [20]. For images known to contain only one person, an easier task is known as *single-person pose estimation.* A more general name is *human pose estimation* [7], which we will use in this report when we are not referring specifically to multiple persons. The name *monocular human pose estimation* is also used to refer to human pose estimation from a single image, as opposed to prediction from multiple images of the same scene.

This report is organized as follows:

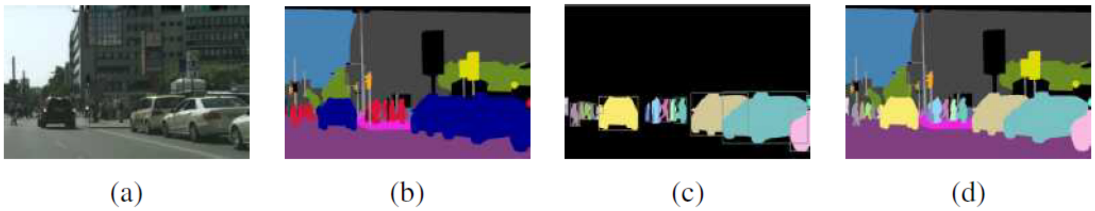
- In Section 2, we define the task, discuss the motivations and applications, and give a brief outline the contributions presented in the paper.
- In Section 3, we briefly describe two different approaches to human pose estimation and discuss the challenges of the task.
- In Section 4, we describe the authors' work in greater detail, provide a high level description of their five key contributions, and explain how they address gaps in existing research.
- In Section 5, we describe the benchmark datasets and evaluation metrics used.
- In Section 6, we discuss the constraints of the proposed methods and discuss possible future work.
- Lastly, in Section 7, we summarize and conclude our report.

2 Definitions, Applications and Research Contributions

2.1 Image Segmentation

Human pose estimation is related to *image segmentation*, which refers to the task of simultaneously performing object recognition and boundary segmentation. The approach to solving this problem is to formulate it as a classification problem: how do we classify and label every pixel in an image?

There are three types of related segmentation problems. *Semantic segmentation* is the classification and labeling of pixels into semantic categories (e.g., cars, buildings, road, sky, trees). *Instance segmentation* is the accurate delineation of individual objects in an image (e.g., marking out individual cars in a row of cars). *Panoptic segmentation* combines semantic and instance segmentation. Figure 1[1] shows an example of the results obtained from these three types of image segmentation.



(a) (b) (c) (d)

Figure 1: (a) Image, (b) Semantic segmentation, (c) Instance segmentation, (d) Panoptic segmentation

2.2 Human Pose Estimation

Human pose estimation restricts the semantic segmentation problem to external human body parts. The goal is to classify and label the external body parts of people in different poses. If the image is known to contain only one person, the task is known as *single-person pose estimation*. A more challenging task is *multi-person pose estimation*, in which the image contains multiple persons in different poses.

Each detected body part, and the line segment joining two points (or *keypoints*) on two different body parts, is labeled using a specific color in the output. The keypoint definitions depend on the benchmark dataset used. For multi-person pose estimation, two widely used benchmarks are the MPII Human Pose dataset [1] and the COCO dataset [13]. The MPII dataset has 16 labeled keypoints. The COCO dataset includes facial features and has 17 labeled keypoints. Figure 2 show the keypoints in the MPII and COCO datasets. Figure 3[2] shows an example of the output from the OpenPose system developed by the authors. We will describe the OpenPose system in Section 4.5.

2.3 Applications

Human pose estimation is a fundamental problem for many computer vision applications involving detection of people, including human-robot interaction, intelligent traffic lights, autonomous vehicle, virtual and augmented reality, and facial recognition.

[1] Images from: *Kirillov et al. Panoptic Segmentation, 2019*

[2] Image from: *Cao et al. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, 2019.*

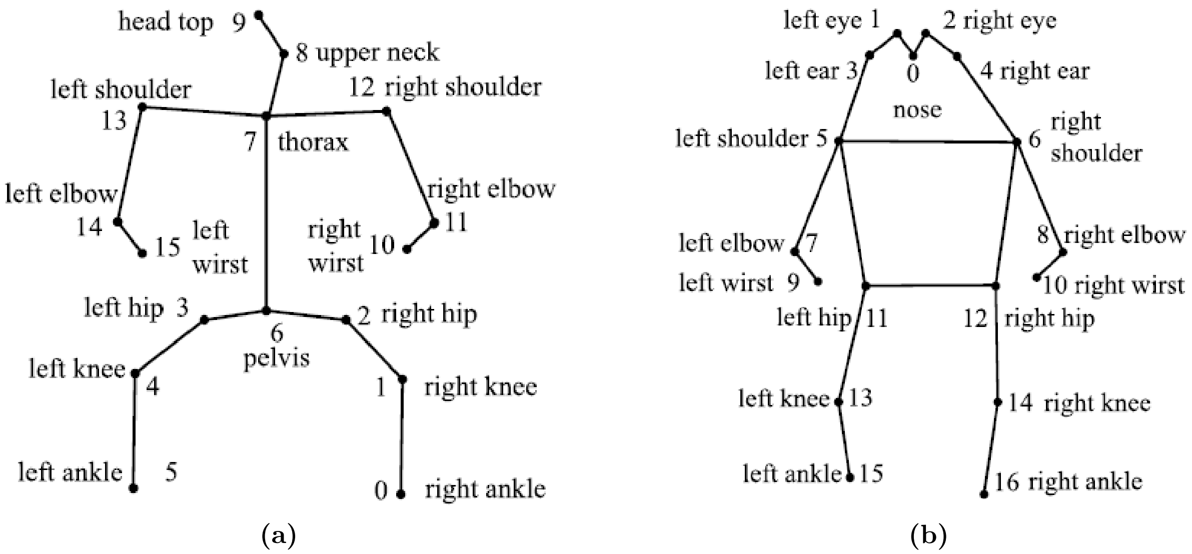


Figure 2: Keypoints: (a) MPII dataset, (b) COCO dataset

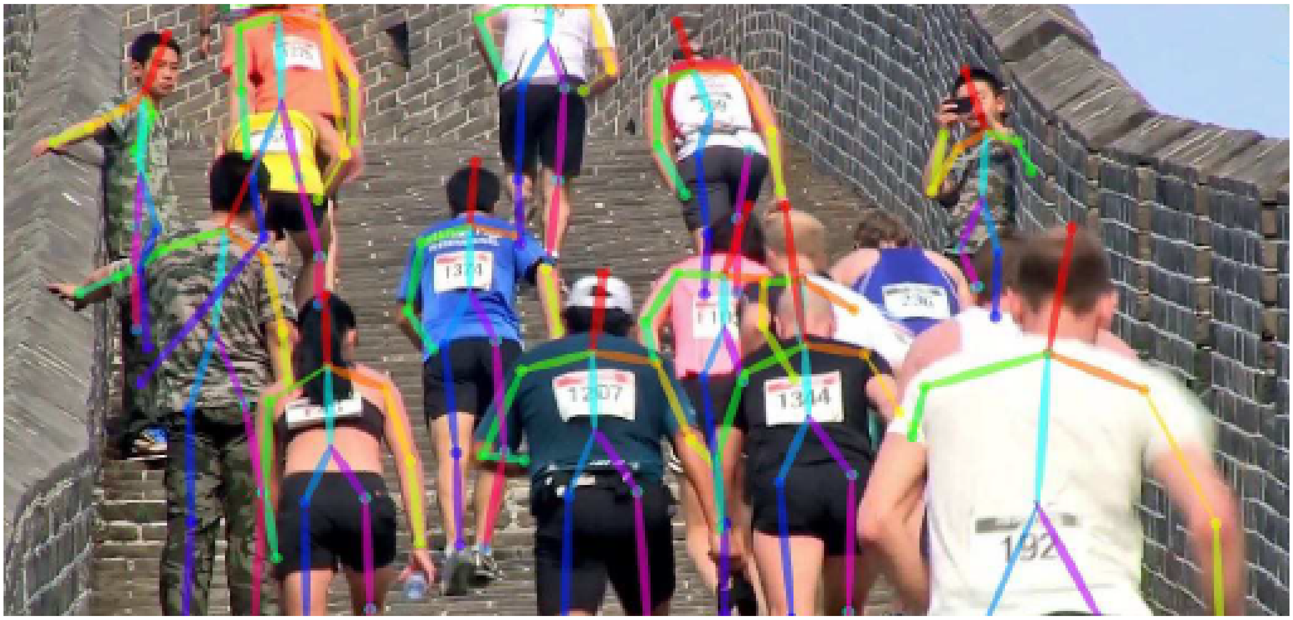


Figure 3: 2D pose estimation from OpenPose

2.4 Outline of Research Contributions

In the paper, the authors presented the following five key contributions:

- Proposal of a novel approach using *Part Affinity Fields (PAFs)* (first proposed by the authors in their earlier paper) that encode the orientations of body parts and the associations between two body parts.
- Proposal of a pipeline and a convolutional neural network model architecture that jointly learns body part detection and association.
- Empirical proof that a greedy multi-person parsing algorithm is sufficient to obtain accurate predictions of body poses, and also preserves inference speed regardless of the number of people.
- Release of a human foot dataset by using a subset of the COCO dataset.
- Open-source release of *OpenPose* library, a complete system for multi-person 2D pose estimation, including body, foot, hand and facial keypoints, and its inclusion in the OpenCV library.

In addition, the authors presented the following findings:

- Showing that Part Affinity Fields refinement (i.e., obtaining more and more accurate predictions over stages) is far more important for maximising overall prediction accuracy than confidence map prediction refinement to detect body parts.
- Showing that a combined model with body and foot keypoint estimation (instead of just body keypoints) can be trained without any increase in training speed and decrease in model accuracy.
- Showing that the proposed methods can be used to predict keypoints of objects in general, such as vehicle keypoints.

3 Approaches and Challenges

In this section, we explain very briefly two common approaches to human pose estimation, the top-down approach and the bottom-up approach. We also list a few highly cited recent work. A survey of all the methods in recent work is beyond the scope of this report. Readers may refer to recent surveys by Wang et al., 2021 [20] and Souza dos Reis et al.,2021 [7]. Figures 4 and 5[3] show the steps in the top-down and bottom-up approaches.

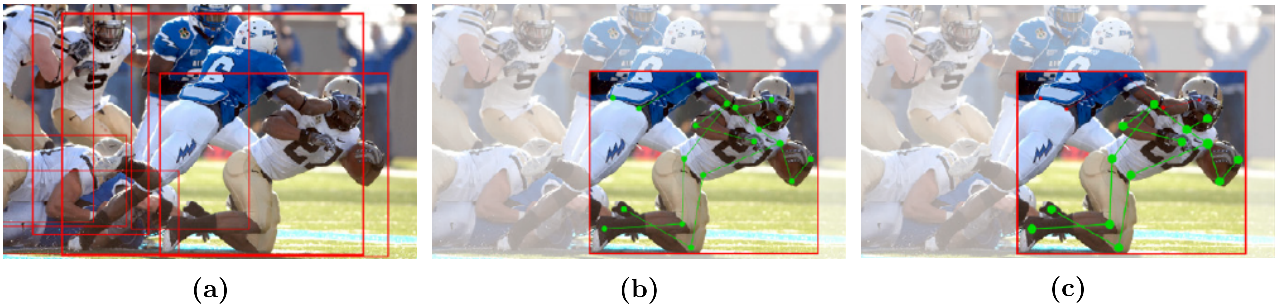


Figure 4: Top-down approach: (a) Segmentation by bounding boxes, (b) Body part detection within a box, (c) Refinement

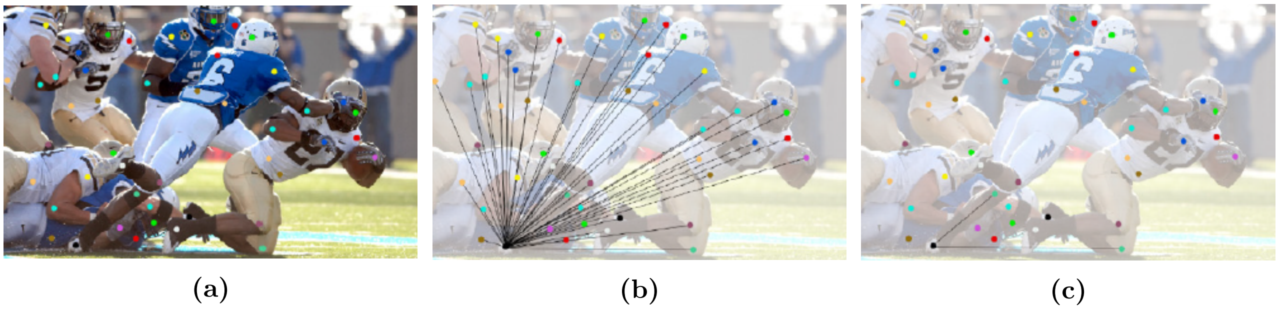


Figure 5: Bottom-up approach: (a) Detection of joint instances, (b) Graph of candidate associations between keypoints, (c) Refinement

3.1 Top-Down Approach

The top-down approach uses methods established in single-person pose estimation. It performs two steps: single-person detection followed by single-person pose estimation. In the first step, a person detector segments each person instance into a *bounding box* (a rectangle). A single person instance is cropped from the image and instance image is fed to the model. In the second step, the model performs body part detection and outputs the results. If the model detects body parts of other persons within the bounding box, it needs to be able to exclude these body parts.

3.2 Bottom-Up Approach

The bottom-up approach first predicts the locations of all body parts. Each keypoint becomes a node on a graph and the next step is to parse the graph to find the optimum association between two keypoints.

[3] Image from: *Souza dos Reis et al. Monocular Multi-Person Pose Estimation: A Survey ,2021.*

Bottom-up approaches generally give lower accuracy than top-down approaches. However, inference speed for top-down approaches is roughly proportional to the number of persons in the image, since it performs detection and pose estimation one person at a time.

3.3 Highly Cited Prior Work

Early research in pose estimation focused on single-person pose estimation and using non-neural models. Important research include work by Felzenszwalb et al., 2005 [9], Ramanan et al., 2005 [16], Andriluka et al., 2009 [2], Andriluka et al., 2010 [3], Johnson et al., 2010 [12] and Yang et al., 2012 [22].

More recent work on multi-person pose estimation use convolutional neural networks but top-down approaches are still more popular than bottom-up approaches. Highly cited research include work by Tompson et al., 2014 [18], Toshev et al., 2014 [19], Wei et al., 2016 [21] and Newell et al., 2016 [14].

3.4 Challenges of Multi-Person Pose Estimation

Compared to single person pose estimation, multi-person 2D pose estimation has additional challenges such as:

- The image may contain an unknown number of people in any position or scale (i.e., relative sizes).
- Two people may overlap in the image, resulting in occlusion of body parts.
- The presence of statues or animals alongside humans in the image may lead to false positives.
- Computational complexity may increase with the number of people in the image.

4 Summary of Research

We now present a high level description of the authors' research and their five key contributions. Most of the mathematical details are omitted. All images presented in this section are from the paper.

4.1 Pipeline and Network Architecture

The proposed method has the following steps:

1. The model takes a color image of size $w \times h$ as input.
2. A multi-stage convolutional neural network (CNN) predicts a set $S = (S_1, S_2, ..., S_J)$ of *confidence maps*. There is one confidence map $S_j \in \mathbb{R}^{w \times h}$ for each of the J body parts. A confidence map is the probability density function for the body part. Each point on the image has a predicted probability of the body part occurring at that point.
3. The CNN also predicts a set $L = (L_1, L_2, ..., L_C)$ of *Part Affinity Fields (PAFs)*. There is one PAF $L_c \in \mathbb{R}^{w \times h \times 2}$ for each of the C pairs of body parts (e.g., shoulder and elbow). Each PAF is a set of 2D vector fields representing the orientation of body parts and the degree of association between two body parts.
4. The confidence maps and PAFs are parsed by greedy inference algorithm.
5. When inference is complete, the system outputs the detected body parts for all persons in the image.

Figure 6 shows colored visualizations of a confidence map and Part Affinity Fields for an image with two persons. Figure 7 shows the CNN architecture.

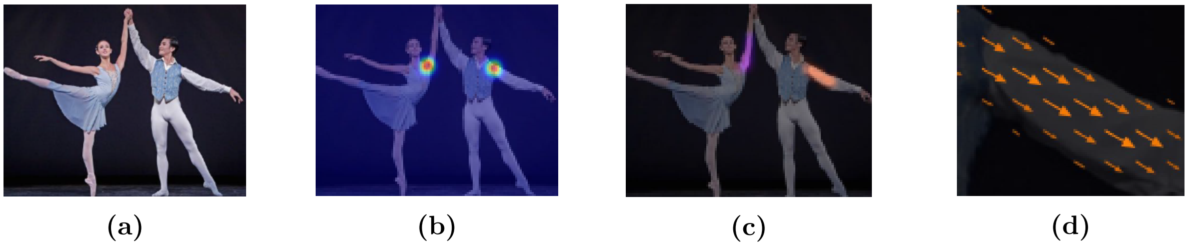


Figure 6: (a) Image, (b) Confidence map, (c) Part Affinity Fields, (d) Close-up of Part Affinity Fields

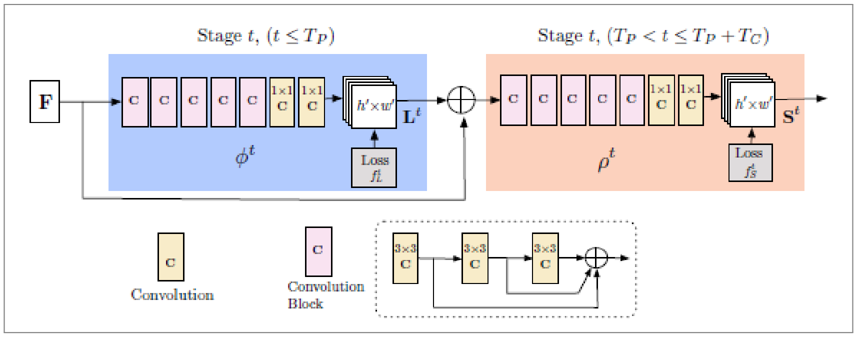


Figure 7: Convolutional neural network architecture

The network in the blue rectangle predicts the PAFs and the network in the beige rectangle predicts the confidence maps. Predictions are refined over a number of stages (T_P stages for PAF prediction followed by T_C stages for confidence map prediction). Prior to the first stage, the image is analyzed by a CNN which produces a set of feature maps F. F is input into the first stage, and a set of PAFs $L^1 = \phi^1(F)$ is produced. In subsequent stages, the predictions from the previous stage and the original feature map are concatenated and fed to the current stage. At stage t:

$$L^t = \phi^t(F, L^{t-1}) \ \forall \ 2 \leq t \leq T_P$$

After T_L stages, the process is repeated for confidence map prediction:

$$S^{T_P} = \rho^{T_P}(F, L^{T_P})$$

$$S^t = \rho^t(F, L^{T_P}, S^{t-1}) \ \forall \ T_P \leq t \leq T_P + T_C$$

At the end of each stage, the authors applied an L_2 loss function between the predicted confidence maps and PAFs and their ground truths. Ground truth confidence maps are generated from the labeled keypoints. If the image contains a single person and the body part is visible, there should be a single peak in the confidence map. If there are multiple persons, there should be a peak corresponding to each visible body part for each person.

4.2 Part Affinity Fields

The second, and arguably the most important contribution by the authors is the use of what they called Part Affinity Fields, first proposed in their earlier paper. The confidence map for a body part predicts the location of the body part in the image. The PAF for a pair of body parts predicts whether the two parts are associated. To illustrate how PAFs work, the authors compared it with another method. This is illustrated in Figure 8.

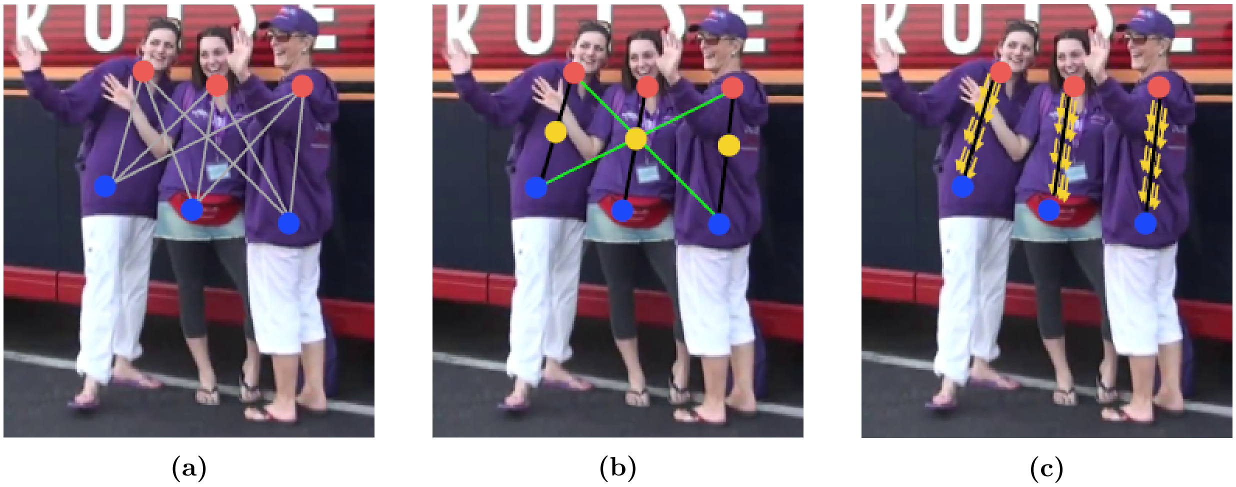


Figure 8: Body part association methods: (a) Detection (red and blue dots) of two body part types and association candidates (grey lines) (b) Using midpoints (yellow dots) between two body parts, (c) Using Part Affinity Fields (yellow arrows)

Supposed that three thorax keypoints and three pelvis keypoints have been detected in an image, as shown in Figure 8(a). How could the model determine which thorax keypoint should be associated with which pelvis keypoint? One way is to detect an additional keypoint on each person midway between the thorax and pelvis keypoints, and then check if the additional keypoint lies on candidate association lines, as shown in Figure 8(b). However, if the three persons are close to one another, this method could produce false positives, as shown by the green lines in Figure 8(b). This method fails because the additional keypoints encodes only position and not orientation.

PAFs not only encodes location but also orientation. Each PAF is a 2D vector field for a pair of body parts. The 2D vector encodes the direction from one body part to the other body part, as shown in Figure 8(c), and helps in predicting association correctly. The authors measure association by computing the line integral over the PAF along the line segment connecting the two body parts. We refer the reader to the paper for the mathematical computations.

4.3 Multi-Person Parsing Algorithm

The third important contribution by the authors is their proposed multi-person parsing algorithm which is effectively a greedy algorithm. As illustrated in Figure 9, for a set of three pairs of body parts, all possible associations between them form a complete graph. Finding the optimum association is equivalent to finding a graph edge with the maximum weight. When there are multiple persons, each with many keypoints, this is a NP-Hard problem which requires relaxations to solve. The authors used two relaxations, as shown in Figure 9(c). First, they selected a minimal number of edges to obtain a graph that connects only adjacent body parts. For example, shoulders connect to elbows but do not connect to wrists. Second, they decomposed the graph into a set of independent bipartite graphs. The authors showed empirically that this greedy parsing algorithm is sufficient to obtain accurate predictions of body poses, and preserves training speed regardless of the number of people.

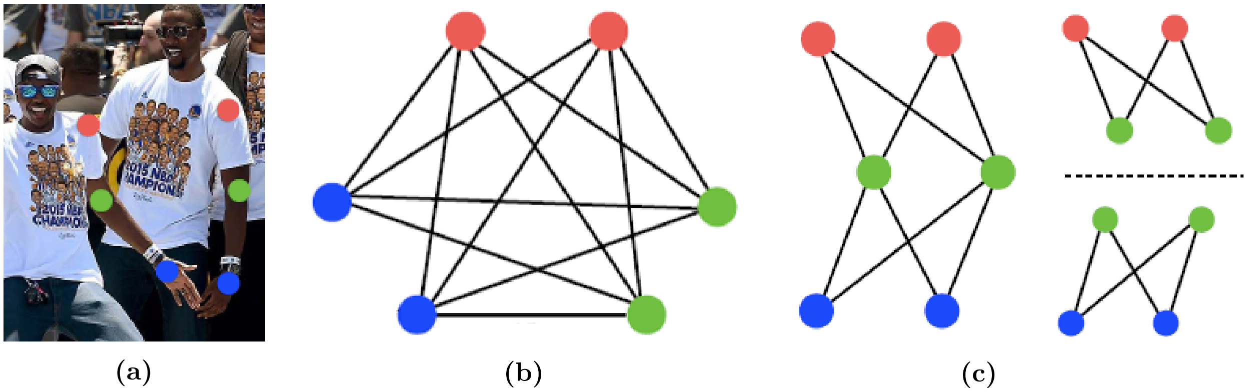

(a) **(b)** **(c)**

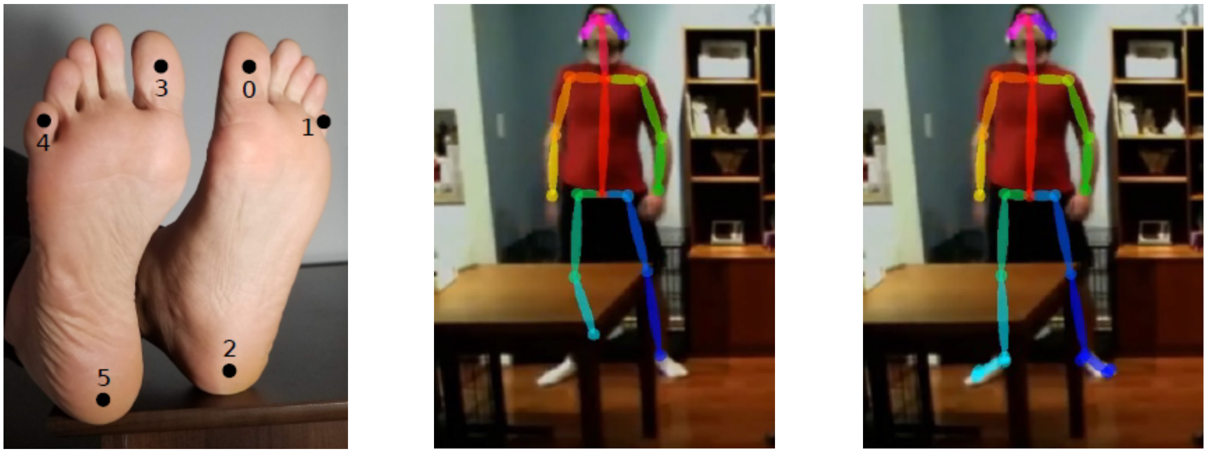
Figure 9: Body part association graph: (a) Image with detection points, (b) Complete graph of all candidate associations, (c) Graph with a minimal number of candidate associations and decomposed into bipartite graphs

4.4 Foot Keypoint Detection

The fourth contribution from the authors is the creation of the foot dataset using a subset of 14K person instances from the COCO dataset. The MPII and COCO datasets contain ankle keypoints but no keypoints on the foot. A dataset with foot keypoints is useful for training models used in graphics applications such as 3D human shape reconstruction. Without foot data, existing approaches may give poor results such as deformed human shapes.

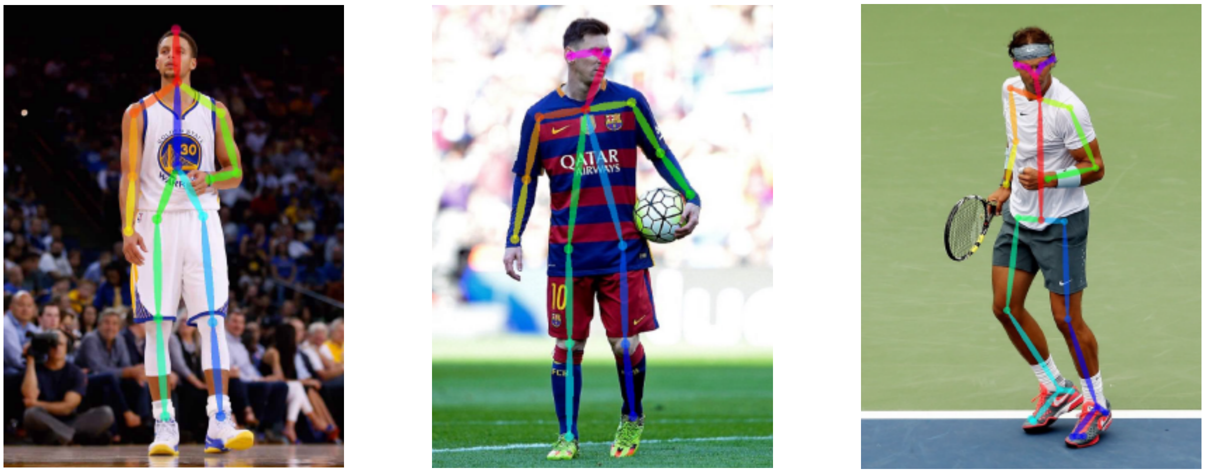
Figure 10(a) shows the six keypoints in the foot dataset. Figures 10(b) and (c) show an example of how foot keypoints help the model to detect leg and ankle keypoints correctly. Adding foot keypoints to existing body keypoints in model training results in a more accurate model.

The authors trained a foot keypoint detector using the model architecture proposed. Figure 11 shows the keypoint distribution of the MPII, COCO and COCO+Foot datasets.



(a) (b) (c)

Figure 10: (a) Foot keypoints, (b) Detection without using foot keypoints, (c) Detection using foot keypoints



(a) (b) (c)

Figure 11: Keypoint distributions: (a) MPII, (b) COCO, (c) COCO+Foot

4.5 OpenPose

Lastly, the authors developed and released an open-source system using their model, named OpenPose[4]. The system is able to detect a total of 135 human body, foot and facial keypoints.

Unlike other similar libraries such as AlphaPose [8][5] and Mask R-CNN [10][6], OpenPose detects all body parts, including foot and facial keypoints. OpenPose also provides a complete pipeline, including modules to read the image, visualize the results, write the results to a file, support for different platforms with CPU or GPU, etc.

The current release of OpenPose provides 2D multi-person keypoint detection and 3D single-person keypoint detection. The 2D multi-person keypoint detector has three modules: a body and foot detector, a hand detector, and a face detector.

Performance-wise, the strength of OpenPose is its fast inference time while giving high prediction accuracy. It is included in the OpenCV[7] [4] library.

[4] Source code at: `https://github.com/CMU-Perceptual-Computing-Lab/openpose`

[5] Source code at: `https://github.com/MVIG-SJTU/AlphaPose`

[6] Source code at: `https://github.com/facebookresearch/Detectron`

[7] Source code at: `https://github.com/opencv/opencv`

5 Benchmarks and Evaluation

5.1 Benchmark Datasets

The authors used the following two benchmark datasets and the foot dataset they created:

- MPII Human Pose Dataset[8] [1] containing 14 body and 2 head and neck labeled keypoints in 25K images and 40K person instances.
- COCO (Complete Objects in Context) dataset[9] [13] released by Microsoft, containing 12 body and 5 facial labeled keypoints in 100K person instances and more than 1 million keypoints in total.

MPII and COCO are the two standard benchmarks in pose estimation. Both datasets contain images of real-world human interactions in a wide range of contexts, scales and number of persons. A yearly COCO challenge awards prizes for the best results. The keypoints in the MPII and COCO datasets are shown in Figure 2.

5.2 Evaluation Metrics

The authors of the MPII benchmark proposed the evaluation metric PCKh, a variant of the PCK (Probability of Correct Keypoint) metric from [22]. The overall performance metric is *mean average precision* (mAP), the mean of average precision of all body parts.

COCO defined the OKS (Object Keypoint Similarity) metric and uses a series of average precisions (AP) over 10 thresholds as the COCO challenge metric.

5.3 Results

OpenPose achieved 75.6% mAP on the MPII benchmark testing dataset, which outperformed the state-of-the-art (59.5%) for bottom-up methods. The state-of-the-art for top-down methods was 78.0% [15]. Using inference time per image as a metric to measure inference speed, OpenPose achieved 0.005 seconds compared to 485 seconds for the state-of-the-art for top-down methods.

OpenPose achieved 64.2% mAP on the COCO benchmark, compared to the state-of-the-art of 70.5% for bottom-up methods and 78.1% on COCO for top-down methods.

5.4 Comparison of Inference Times With Other Libraries

The authors compared their model, OpenPose, with two other widely used libraries which use top-down methods: AlphaPose and Mask R-CNN. As expected for top-down methods, the inference times for AlphaPose and Mask R-CNN are proportional to the number of persons that their person detectors extract.

The authors presented a detailed analysis of the inference time for OpenPose. The inference time depends on the CNN processing time which is constant (complexity of $O(1)$) and the multi-person

[8] Website: http://human-pose.mpi-inf.mpg.de/

[9] Website: https://cocodataset.org/

parsing time (complexity of $O(n^2)$, for n persons). Since the parsing time is two orders of magnitude less than the CNN processing time, the inference time is almost constant regardless of the number of persons in the image.

6 Constraints and Future Work

6.1 Constraints and Failure Cases

The authors observed that top-down methods give higher accuracy but lower inference speed. For bottom-up methods, accuracy is generally lower. Model accuracy is affected by the per person image resolution fed to the model. Top-down methods individually crop each detected person before feeding the image to the model. Bottom-up methods use the entire image, resulting in lower resolution per person. The authors noted that AlphaPose has the greatest accuracy while OpenPose has the fastest runtime speed. They also observed that ambiguous poses cause the model to fail or give false positives. Figure 12[10] shows examples of common failure cases.

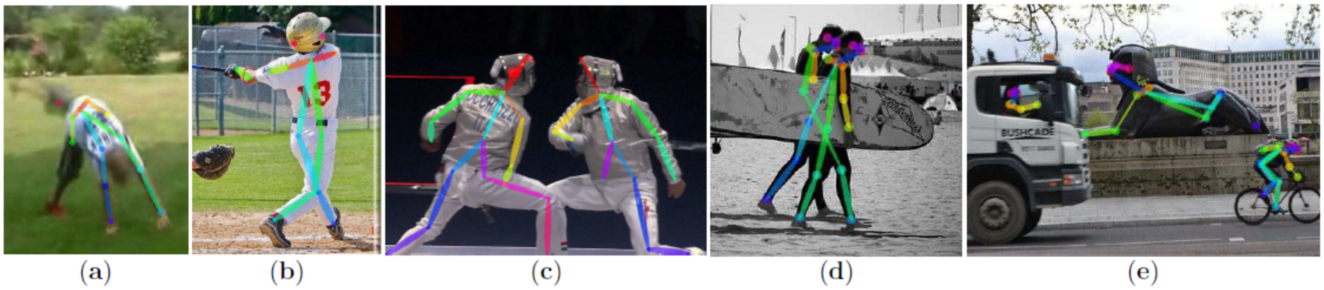


Figure 12: Common failure cases: (a) Rare poses, (b) Occlusion (missing parts), (c) Overlapping parts, (d) Overlapping bodies, (e) False positives (non-human objects)

6.2 Future Work

Deep convolutional neural networks (CNN) have resulted in increasingly more accurate models on the MPII and COCO benchmarks of static color images [7]. Since OpenPose was first released, Sun et al. [17] achieved 92.3% mAP on the MPII benchmark and 75.5% mAP on the COCO benchmark by using a CNN model architecture that maintains high-resolution representations throughout the training process. Further improvement on the benchmarks are likely in the next few years.

3D human pose estimation (inferring the (x, y, z) coordinates in 3D space) from images or videos is a more challenging problem due to insufficient labeled training data and 3D-considerations such as depth ambiguities [11]. Current methods map body part pixels in a 2D image to the 3D surface of the same human body. Figure 13[11] illustrates the extension of 2D pose estimation to 3D pose estimation.

The long-term goal of computer vision is human-like *scene understanding*. Other than detecting and recognizing objects and people in images, the more challenging task is understanding the *context* of a complex scene the way humans can do so instinctively. How can machines recognize different hand and arm gestures? For instance, how can machines distinguish between a person waving goodbye with his arm and another person stretching his arm during physical exercise? Thus, 2D or 3D human pose estimation is only the beginning of a long research journey in computer vision.

[10] Images from: *Cao et al. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, 2019.*

[11] Image from *Ji et al. A Survey on Monocular 3D Human Pose Estimation, 2020.*

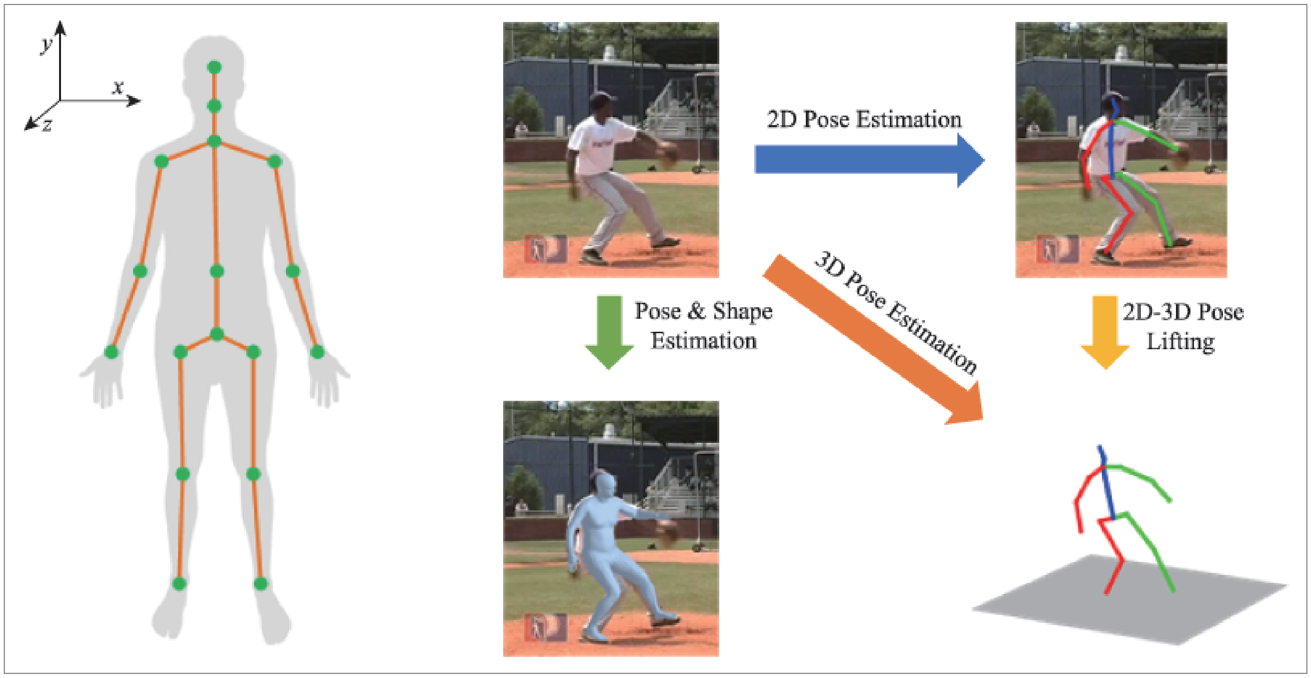


Figure 13: 2D and 3D pose estimation

7 Conclusion

We presented a summary of the research in the paper *OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields* by Cao et al., 2019. This paper is an extension and improvement of the authors' earlier work, *Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields*, 2017.

We also provided a very brief description of two approaches for 2D multi-person pose estimation and discussed the challenges of the problem. For the authors' work, we summarized their five key contributions and the gaps in existing work their research addressed. We also described the benchmarks and evaluation metrics used, and the constraints of their model.

In the paper, the authors proposed a novel bottom-up approach using Part Affinity Fields (PAFs) that encode the locations and orientations of body parts and the associations between two body parts. Second, their convolutional neural network model jointly learns body part detection and association. Third, the authors showed that the accuracy of Part Affinity Fields is crucial to the overall prediction accuracy. They showed that a greedy parsing algorithm is sufficient to obtain accurate predictions of body poses, and preserves training speed regardless of the number of people. Fourth, the authors released a human foot dataset by using a subset of the COCO dataset. Fifth, they released the *OpenPose* library for multi-person 2D pose estimation. Their OpenPose model outperformed the state-of-the-art for bottom-up methods and its inference time was six orders of magnitude less than the state-of-the-art.

Lastly, we discussed the research progress in 2D human pose estimation, its extension to 3D human pose estimation and its importance to the long-term goal of complete scene understanding in computer vision research.

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, 12

[2] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial Structures Revisited: People Detection and Articulated Pose Estimation. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1014–1021. IEEE, 2009. 6

[3] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3D Pose Estimation and Tracking by Detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 623–630. IEEE, 2010. 6

[4] Gary Bradski. The OpenCV library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 11

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019. 1

[6] 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*, pages 7291–7299, 2017. 1

[7] Eduardo de Souza Reis, Lucas Adams Seewald, Rodolfo Stoffel Antunes, Vinicius Facco Rodrigues, Rodrigo da Rosa Righi, Cristiano André da Costa, Luiz Gonzaga da Silveira Jr, Bjoern Eskofier, Andreas Maier, Tim Horz, et al. Monocular Multi-Person Pose Estimation: A Survey. *Pattern Recognition*, page 108046, 2021. 1, 5, 14

[8] 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. 11

[9] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial Structures for Object Recognition. *International journal of computer vision*, 61(1):55–79, 2005. 6

[10] 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. 11

[11] Xiaopeng Ji, Qi Fang, Junting Dong, Qing Shuai, Wen Jiang, and Xiaowei Zhou. A Survey on Monocular 3D Human Pose Estimation. *Virtual Reality & Intelligent Hardware*, 2(6):471–500, 2020. 14

[12] Sam Johnson and Mark Everingham. Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation. In *bmvc*, volume 2, page 5. Citeseer, 2010. 6

[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. 2, 12

[14] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 6

[15] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint Subset Partition and Labeling for Multi-Person Pose Estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016. 12

[16] Deva Ramanan, David A Forsyth, and Andrew Zisserman. Strike a Pose: Tracking People by Finding Stylized Poses. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 271–278. IEEE, 2005. 6

[17] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 14

[18] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. *Advances in neural information processing systems*, 27:1799–1807, 2014. 6

[19] Alexander Toshev and Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 6

[20] Chen Wang, Feng Zhang, and Shuzhi Sam Ge. A Comprehensive Survey on 2D Multi-Person Pose Estimation Methods. *Engineering Applications of Artificial Intelligence*, 102:104260, 2021. 1, 5

[21] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 6

[22] Yi Yang and Deva Ramanan. Articulated Human Detection With Flexible Mixtures of Parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012. 6, 12