**Paper Review Report:**

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, Zhe Cao Tomas Simon, Shih-En Wei and Yaser Sheikh, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1302-1310, Citation: **DOI:**[10.1109/CVPR.2017.143](https://doi.org/10.1109/CVPR.2017.143)

**1. Introduction**

This paper begins by introducing human 2D pose estimation as the problem of localizing anatomical key points, which for humans are body parts. It then describes three main obstacles faced by multi-person pose estimation. They include unknown number of people with different position and scale, complex spatial interference such as due to occlusion, and an increase in run time complexity caused by more people. Subsequently, the paper addresses common top-down approaches and notes further disadvantages like low computational efficiency. It then presents bottom-up approaches as a solution but points out additional problems like a costly global inference in the final parse. Therefore, this paper presents an efficient bottom-up approach that uses Part Affinity Fields (PAFs) to encode the location and orientation of limbs over the image domain for multi- person pose estimation.

**2. Proposed Method**

The overall pipeline of the proposed method is as follows. The method takes an image as an input into a two-branch convolutional neural network (CNN) to predict confidence maps for body part detection and PAFs for parts association. Next a parsing step is performed to conduct bipartite matchings to associate body parts to each person in order to assemble the full body poses for all people in the image.

**2.1. CNN for simultaneous detection and association**

In this method the CNN architecture is designed to simultaneously predict detection confidence maps and affinity fields that encode part-to-part association. The CNN has two branches, with the first branch used to predict the confidence maps and the other used to predict the PAFs. First, an image is analysed by a CNN (first 10 layers from [1]) to generate feature maps. These feature maps are then inputted into the two branches of the proposed CNN [2] to output the confidence maps and PAFs. This process is iterated *T* times. For the first iteration the input consists only of feature maps, but for iterations two and greater, the branches also take the outputs of the previous iterations as inputs. At the end of each iteration a L2 loss is calculated between the estimated predictions and the ground truth maps and fields.

**2.1.1 Confidence maps**

A confidence map is a 2D representation of the belief that a particular body part occurs at each pixel location. In order to calculate the L2 loss in the training stage (Section 2.1), ground truth confidence maps need to be generated. For multiple people, the *kth* ground truth confidence map will have a peak associated to each visible part *j* for each person *k*. For each location *p* in the confidence map of the *kth* person a value is calculated. It is calculated by taking the square of the Euclidian norm of the difference of the position of *p* and the ground truth position of body part *j* is divided by some value σ2 and then multiplied by -1. The parameter σ controls the spread of the peak which corresponds to a visible of a person. Then, Euler’s number is raised to the power of the value obtained previously to get the final value of location *p.* The final ground truth confidence map is obtained by applying the maximum operator to all confidence maps for each person.

**2.1.2 Part Affinity Fields**

Part affinity fields are 2D vector fields for each limb. The advantage of PAFs is that they preserve both the location and orientation information of the support region of a limb. In order to calculate the L2 loss in the training stage (Section 2.1), ground truth PAFs need to be generated. To do so, PAFs for every limb *c* of person *k* are created by computing a value equal to the difference of the ground truth positions of body parts *j1* and *j2*divided by theEuclidian norm of the difference of the ground truth positions of body parts *j1* and *j2.* This is computed at every point *p* in the PAF, if point *p* lies on the limb. If point *p* does not lie on the limb then the value of the PAF at point *p* equals zero. The final ground truth PAF is then determined by taking the average of all the PAFs for all *k* people.

**2.2 Multi-Person Parsing**

In order to associate the limbs to each person, the body part candidates detected by the confidence maps need to be paired to generate limbs that are associated with the correct person in an image. To do so bipartite graph matching is employed. The process is as follows, a variable ‘*z*’ is created to hold all possible pairs of body part candidates for each joint *j* and every candidate number *N.* To determine which candidates for joints *j1* and *j2* form the correct limb,the variable ‘*z*’ is multiplied by the line integral of the PAF for each candidate *m* of joint *j1* and candidate *n* of joint *j2*. Finally, the maximum function is applied to determine the pair of candidate parts with the highest weighting for the chosen joints. This is done for every pair of joints and the Hungarian algorithm [3] is used to obtain the optimum matching. Once all the limbs are established the full body pose is created for every person using a spanning tree skeleton.

**2.n. Novelty of the proposed method**

The novelty of the proposed method is in the use of a new association method using PAFs. As mentioned in the paper, PAFs are a new method that have not been used anywhere else and allow for the encoding of global context in a computationally efficient manner which other bottom-up methods like the ones in [4] and [5] do not do.

**3. Experimental Results**

The two benchmark datasets used to test the proposed method include the MPII human multi-person dataset (MPII) [6] and the COCO 2016 key points challenge dataset (COCO) [7]. When tested on the MPII dataset the proposed method greatly outperformed previous state-of-the-art methods like the one from Iqbal et al [8] by over 20%. More specifically, when tested on a subset of 288 images and the whole dataset, the average accuracy of the proposed method was 79.7% and 75.6%, respectively. When tested on the COCO dataset, the proposed method once again outperformed all the other state-of-the-art methods. It gained average accuracy of 60.5% and 61.8% for test-challenge and test-dev, respectively. However, the proposed method did not achieve the highest precision for people of smaller scales when tested on COCO. Lastly the proposed method’s runtime performance was tested using a laptop with one NVIDIA GeForce GTX-1080 GPU. It showed that the CNN processing time runtime complexity was O(1) and multi- person parsing time runtime complexity was O(n2). When tested on a video with 19 people, it achieved a speed of 8.8 fps.

**4. Conclusion / Future works**

In conclusion, the proposed method does seem better than other state-of-the-art methods not only because it outperformed many of them in terms of accuracy, but also because it is computationally efficient. Therefore, it seems reasonable to suggest that future research on this method ought to be conducted. More specifically, future work should focus on improving performance on people of smaller scales, and also on overall accuracy as well.

**References:**

[1] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[2] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Con- volutional pose machines. In *CVPR*, 2016. 1, 2, 3, 6

[3] H.W.Kuhn. The Hungarian method for the assignment problem. In *Naval research logistics quarterly*. Wiley Online Li- brary, 1955.

[4] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016

[5] E.Insafutdinov, L.Pishchulin, B.Andres, M.Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi- person pose estimation model. In *ECCV*, 2016

[6] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: new benchmark and state of the art analysis. In *CVPR*, 2014

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra- manan, P. Dolla ́r, and C. L. Zitnick. Microsoft COCO: com- mon objects in context. In *ECCV*, 2014

[8] U.Iqbal and J.Gall. Multi-person pose estimation with local joint-to-person associations. In *ECCV Workshops, Crowd Understanding*, 2016