Multi-Person Pose Estimation for Pose Tracking with Enhanced Cascaded Pyramid Network

Dongdong Yu[1,†], Kai Su[1,2,†], Jia Sun[1], and Changhu Wang[1,*]

[1] ByteDance AI Lab, Beijing, China
{yudongdong,sukai,sunjia.ring,wangchanghu}@bytedance.com
[2] MOE Key Laboratory of Computer Network and Information Integration, Southeast University, China
sukai@seu.edu.cn

Abstract. Multi-person pose estimation is a fundamental yet challenging task in machine learning. In parallel, recent development of pose estimation has increased interests on pose tracking in recent years. In this work, we propose an efficient and powerful method to locate and track human pose. Our proposed method builds upon the state-of-the-art single person pose estimation system (Cascaded Pyramid Network), and adopts the IOU-tracker module to identify the people in the wild. We conduct experiments on the released multi-person video pose estimation benchmark(PoseTrack2018) to validate the effectiveness of our network. Our model achieves an accuracy of 80.9% on the validation and 77.1% on the test set using the *Mean Average Precision* (MAP) metric, an accuracy of 64.0% on the validation and 57.4% on the test set using the *Multi-Object Tracking Accuracy* (MOTA) metric.

Keywords: Pose Estimation · Pose Tracking.

1 Introduction

Multi-person pose estimation and tracking are important yet challenging problems for all persons in single RGB image, which are fundamental research topics for many visual applications like human action recognition [19], human-computer interaction [6] and so on.

Recently, the performance of multi-person pose estimation on standard benchmarks such as MPII Pose [11] and COCO [12] has been greatly improved with the rapidly development of convolution neural networks [18, 16, 20, 2, 14, 8, 15, 4, 13]. Existing methods can be classified into two kinds of approaches: the bottom-up approach and the top-down approach. The bottom-up approach detects human skeletons from all potential human candidates and then assemble these skeletons into each person. The top-down approach first adopt a detection module to get all the human boxes from the image, then apply a single-person human

[†] Equal contribution.
[*] Corresponding author.

pose estimator to detect human skeletons. Although impressive performance has been achieved, current state-of-the-art methods still have difficulty to deal with occluded keypoints, invisible keypoints, and crowed backgroud, which cannot be well localized. Most recent pose tracking methods track the human box over the entire video in terms of similarity between pairs of boxes measured with box iou or similarity between pairs of human keypoints measured with keypoint oks distance in adjacent frames [7, 9, 21].

In this work, we propose an efficient and powerful approach to multi-person keypoint detecting and tracking in videos. For the keypoints detecting stage, we propose an enhanced cascade pyramid network to accurately locate human keypoint in each frame of a video. For the keypoint tracking stage, we employs IOU tracker which is a lightweight frame-by-frame optimization method, allowing our model to be scalable to virtually any length videos.

2 Related Work

Our proposed approach is related to previous works involving with human pose estimation and tracking, as described as follows:

Multi-person pose estimation is an important task in computer vision. Existing approaches can be divided into two categories: bottom-up approaches and top-down approaches. Bottom-up approaches firstly predict all keypoints and then assemble them into multiple persons. For example, associate embedding simultaneously predict heatmaps and tagmaps to group the predicted keypoints to different persons [13]. Top-down approaches firstly detect all human boxes in an image, and then predict the keypoints within each box independently. For example, Cascaded Pyramid Network (CPN) predicts human bounding boxes first and then solve the single person pose estimation in the cropped person patches [4]. In general, top-down approaches perform more accurate than bottom-up approaches. However, with the number of humans increases in an image, top-down approaches perform more slower.

Based on the multi-person pose estimation architectures described above, it is natural to extend them from still image to video. Some online trackers simplify this tracking problem as a maximum weight bipartite matching problem and solve it with greedy or Hungarian Algorithm. Nodes of this bipartite graph are human bounding boxes in two adjacent frames. For example, PoseTrack [7] and ArtTrack [9] in CVPR17 primarily introduce multi-person pose tracking challenge and propose a new graph partitioning formulation, building upon 2D DeeperCut [10] by extending spatial joint graph to spatio-temporal graph.

3 Method

In this work, we take the top-down method to estimate multi-person pose in each frame. Firstly, we apply as human detector on the RGB image to generate human bounding-boxes. Secondly, we predict the detailed localization of the keypoints for each candidate human bounding-boxes by a single-person pose estimator. Finally, we we simplify the tracking problem to bipartite matching the candidate bounding-boxes between a pair of frames.

3.1 Person Detector

In order to detect more people from image, we adopt the Deformable Convolutional Networks (with detection MAP of 44.4 on the COCO minival dataset) [5] and SNIPER (with detection MAP of 46.5 on the COCO minival dataset) [17] methods to generate our human bounding-boxes.

3.2 Pose Estimator

In order to get accurate person keypoints, we adopt the state-of-the-art single person pose estimator [4](Cascade Pyramid Network) to detect the human skeletons. In addition, we have enhance the cascade pyramid network to make it more robust and accurate to handle large pose variations, changes in clothing and lighting conditions, severe body deformations, heavy body occlusions and so on. For the Global-Net, we design a shuffle unit to cross the information from all feature scales. For the Refine-Net, we design an attention unit to extract more representative feature to predict the keypoint localization.

3.3 Pose Tracker

Following the ICCV 2017 winner [7], these detections are presented as a graph, where every detected person bounding box in every frame is a node. And the edges are defined to connect each human bounding-box in a frame to each human bounding-box in the next frame. The cost of each edge is defined as the iou metric of the two human bounding-boxes linked on that edge to belong to the same person. To compute tracks, we simplify the problem to bipartite matching between a pair of frames, and propagate the labels forward, one frame at a time, starting from the first frame to the last.

4 Experiments

4.1 Dataset and Evaluation Metric

Our single person pose estimation model is trained with three datasets: MSCOCO dataset [12], AI challenge dataset [3], and PoseTrack challenge 2018 dataset [1]. MSCOCO dataset contains over 66k images with 150k people, AI challenge dataset has more than 270k images with 449k people, and PoseTrack challenge 2018 dataset contains 667 short video clips annotated for multi-person pose estimation and multi-person pose tracking.

We evaluate our proposed method on PoseTrack Challenge 2018 dataset. We use Total AP to evaluate the multi-person pose estimation results and standard MOTA metric to evaluate the tracking performance.

Table 1. The performance of the MAP metric on PoseTrack challenge 2018 dataset.

Dataset	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
validation	82.4	88.8	86.2	79.4	72.0	80.6	76.2	80.9
partial test	79.0	84.6	81.7	75.5	68.8	77.4	72.0	77.1

Table 2. The performance of the MOTA metric on PoseTrack challenge 2018 dataset.

Dataset	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
validation	68.8	73.5	65.6	61.2	54.9	64.6	56.7	64.0
partial test	61.4	65.1	58.4	55.0	49.0	59.0	51.5	57.4

4.2 Training Details

Our single person pose estimation model is trained using adam algorithm with an initial learning rate of 5e-4. Note that we also decrease the learning rate by a factor of 2 every 3600000 iterations. We use a weight decay of 1e-5 and the training batch size is 32.In the training for pose estimation, 4 V100 GPUs on a GPU server are used.

4.3 Testing Details

Following same testing strategies used in CPN, we apply a gaussian filter on the predicted heatmaps. We also predict the pose of the corresponding flipped image and average the heatmaps to get the final prediction. A quarter offset in the direction from the highest score response to the second highest response is used to obtain the final location of the keypoints. In order to get the best performance on the MAP metric, we first use the SoftNMS on the candidate human bounding-boxes generated by the Deformable Convolutional Networks and SNIPER. Second, we use the Pose-OKS method with the threshold of 0.4 to filter out the redundant human keypoints. Finally, we filter out the human bounding boxes which area is smaller than 3600. In order to achieve the best performance on the MOTA metric, two more rules added. The score of human-bounding box must be higher than 0.35 and the score of the predicted keypoint must be higher than 0.85.

4.4 PoseTrack Challenge Results

We evaluate our method on the whole validation set and partial of test set of the PoseTrack challenge 2018 dataset. The performance of the MAP metric is shown in the Table 1. And, the performance of the MOTA metric is shown in Table 2. We also show some sample keypoints detection results of our model on the PoseTrack challenge 2018 dataset in Fig. 1.

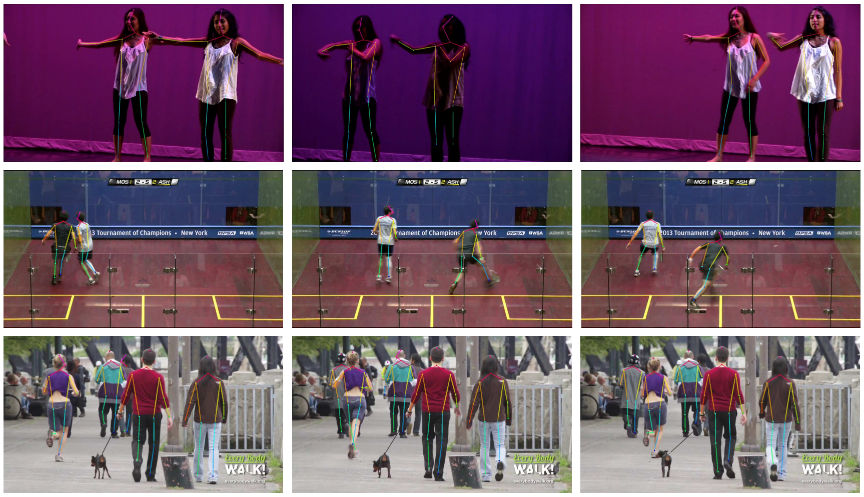

Fig. 1. Some results of our model on the PoseTrack challenge 2018 dataset.

5 Conclusions

In this paper, we propose an efficient and powerful method for the multi-person pose estimation and tracking. For the multi-person pose estimation, based on the Cascaded Pyramid Network, we design a shuffle unit to fuse the pyramid feature maps and an attention unit to extract more representative feature maps. For the multi-person pose tracking, we simplify the problem as a bipartite matching problem between a pair of the frames. Experimental results show that our method achieves an accuracy of 80.9% on the validation and 77.1% on the test set using the *Mean Average Precision* (MAP) metric, an accuracy of 64.0% on the validation and 57.4% on the test set using the *Multi-Object Tracking Accuracy* (MOTA) metric.

References

1. challenge 2018, P.: Posetrack challenge 2018 dataset. https://posetrack.net/
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. vol. 1, p. 7 (2017)
3. challenger, A.: Ai challenger dataset. https://challenger.ai/
4. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. arXiv preprint arXiv:1711.07319 (2017)
5. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. CoRR, abs/1703.06211 **1**(2), 3 (2017)
6. Dix, A.: Human-computer interaction. In: Encyclopedia of database systems, pp. 1327–1331. Springer (2009)
7. Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M., Tran, D.: Detect-and-track: Efficient pose estimation in videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 350–359 (2018)

8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on. pp. 2980–2988. IEEE (2017)
9. Insafutdinov, E., Andriluka, M., Pishchulin, L., Tang, S., Levinkov, E., Andres, B., Schiele, B.: Arttrack: Articulated multi-person tracking in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). vol. 4327. IEEE (2017)
10. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision. pp. 34–50. Springer (2016)
11. MPII: Mpii human pose dataset. http://human-pose.mpi-inf.mpg.de/
12. MS-COCO: Coco keypoint leaderboard. http://cocodataset.org/
13. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems. pp. 2274–2284 (2017)
14. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. pp. 483–499. Springer (2016)
15. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: CVPR. vol. 3, p. 6 (2017)
16. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4929–4937 (2016)
17. Singh, B., Najibi, M., Davis, L.S.: Sniper: Efficient multi-scale training. arXiv preprint arXiv:1805.09300 (2018)
18. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1653–1660 (2014)
19. Wang, C., Wang, Y., Yuille, A.L.: An approach to pose-based action recognition. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. pp. 915–922. IEEE (2013)
20. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4724–4732 (2016)
21. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose flow: Efficient online pose tracking. arXiv preprint arXiv:1802.00977 (2018)