Towards Multi-Person Pose Tracking: Bottom-up and Top-down Methods

Sheng Jin[1] Xujie Ma[1] Zhipeng Han[2] Yue Wu[3] Wei Yang[4]
Wentao Liu[5] Chen Qian [5] Wanli Ouyang[4,6]
[1] Tsinghua University [2] Peking University [3] Wuhan University
[4] The Chinese University of Hong Kong
[5] SenseTime Group Limited [6] The University of Sydney
[1]{js17, mxj14}@mails.tsinghua.edu.cn [2] hanzhipeng@pku.edu.cn [3] yuewu@whu.edu.cn
[4]wyang@link.cuhk.edu.hk [5]{liuwentao, qianchen}@sensetime.com [6] wanli.ouyang@sydney.edu.au

Abstract

In this paper, we focus on the challenging problem of multi-person pose tracking in the wild. Recent multi-person articulated tracking methods can be categorized into the top-down and bottom-up approaches. We investigate the advantages and disadvantages of both bottom-up and top-down methods on various datasets. We propose a novel bottom-up joint detector, termed as MSPAF to extract multi-scale features and implement a human detector based on recent development of object detection. Incorporating the global context, we use a human detector to rule out bottom-up false alarms which significantly improves the tracking results. Following the commonly used graph partitioning formulation, we construct a spatio-temporal graph and solve a minimum cost multicut problem for human pose tracking. Our proposed method achieves the state-of-the-art performance on both "Multi-Person PoseTrack" dataset and "ICCV 2017 PoseTrack Challenge" dataset.

1. Introduction

The PoseTrack Challenge focuses on multi-person articulated tracking in the wild, which remains a challenging problem due to large variability in appearance and scales, complex poses, occlusion and crowding.

The pose tracking approach mainly follows two stages, pose estimation and grouping. Recent work on multi-person pose estimation can be grouped into bottom-up and top-down approaches. Figure 1 compares the advantages and disadvantages of both bottom-up and top-down methods. Bottom-up methods are more robust to occlusion and complex poses. However, most bottom-up methods do not directly benefit from human body structural information, leading to many false positives. Top-down methods utilize global context and strong structural information. However, they are not able to handle complex poses. Moreover, the performance of the top-down model is closely related to person detection results. Images with crowding people may lead to detection errors and thus resulting in poor pose estimation results. We therefore propose to use human detections with body structural information to rule out false alarms by bottom-up joint detectors. As for the grouping stage, we follow the commonly used minimum cost multicut formulation [1, 9] and construct a spatio-temporal graph to facilitate human pose tracking.

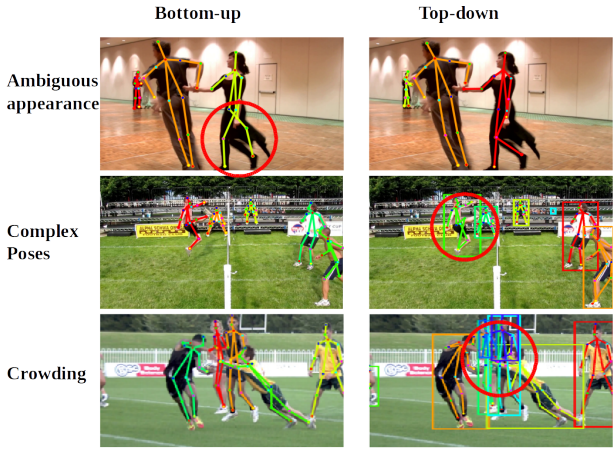

Figure 1. Comparison of bottom-up and top-down methods. Problems of bottom-up and top-down methods are highlighted with red circles.

In this paper, our main contributions are threefold. (1) We investigate and compare the performance of both top-down and bottom-up pose estimation models on various datasets. Human bounding boxes contain the global context which is able to rule out false alarms generated by bottom-up joint detectors. (2) We propose a multi-person joint detector with multi-scale features (MSPAF) to generate joint proposals in a bottom-up manner. The proposed

MSPAF significantly improves over [4]. (3) We evaluate our approach on the PoseTrack dataset [11] and ICCV 2017 PoseTrack Challenge dataset [1]. We show that our results significantly outperform the baseline approaches.

2. Method

2.1. Body Part Detector

We explore the CNN based appearance feature for part detection. We implement both bottom-up and top-down body part detectors and make comparisons on different datasets.

We implemented our bottom-up joint detector based on Part Affinity Fields (PAFs) [4] and improve it by employing a novel MSPAF structure with inception-residual blocks. The inception-residual blocks have a multi-path structure and are able to extract multi-scale features. Multi-scale features preserve both high-level knowledge and low-level cues to localize joint candidates and the multi-path networks may lead to better convergence. Our MSPAF network also contains multiple stages, allowing for intermediate supervision. The PAF model outputs both the part confidence maps and the part affinity fields. The part confidence maps are used to generate part proposals. The part affinity fields are a set of 2D vector fields encoding the location and the orientation of the limbs which can be used for data association. In [4], per-frame greedy association algorithm is employed which may lead to inaccuracy in pose estimation, especially in the cases of occlusion and overlapping. We instead build a spatio-temporal graph, which will take both temporal and spatial information into consideration, effectively handling occlusion.

Our top-down pose detector is based on Mask R-CNN [7]. In order to handle complex poses in the PoseTrack dataset, we replace the simple head structure in Mask R-CNN [7] with a separately-trained single person pose estimator. However, the performance of the pose estimator is sensitive to the person bounding boxes used in training. We only get poor results when training with detection boxes. To get better performance, we use RPN proposal boxes to train the single person pose estimator. In detail, we use an RPN for person bounding boxes trained on MSCOCO dataset [14]. We generate around 2000 region proposals on each training image as Faster R-CNN [15] does, and only keep proposals which have an IOU over 0.5 with people detection boxes. Then the single-person joint detector is trained based on these proposal boxes.

Training In our implementation, we make full use of MSCOCO [14], MPII [2] and PoseTrack Challenge [1] datasets to train our model. We adopt the idea of curriculum learning [6] to train the model with gradually increasing datasets. We observe that compared with MPII and PoseTrack, the MSCOCO dataset contains a smaller number of people and simpler poses. Thus, we first use MSCOCO dataset to pre-train the MSPAF pose estimator. The simpler dataset will help the model to learn general concept of joints and limbs. To handle the annotation difference, we automatically label the nose on MPII with the pre-trained MSCOCO model. Then the pre-trained model is fine-tuned on a mixture of PoseTrack and relabeled MPII datasets.

2.2. Person Detector

We built our person detector based on the Faster R-CNN method [15]. We use the pre-trained Faster R-CNN models introduced in [5]. In detail, we use ResNet-152 model, FPN ResNet-101 model [13], and FPN ResNet-152 model. To get better performance, we ensembled the three Faster R-CNN models with different backbone structures following [8]. Model ensembling greatly enhanced our detection accuracy.

Following [3], We replace NMS with Soft-NMS in the second stage of Faster R-CNN. Instead of using hard overlap threshold in NMS, Soft-NMS algorithm decays the score of detection bounding boxes that have a high overlap with other detection boxes. Soft-NMS encourages more nearby objects to be taken into consideration, which further improves the recall of our person detector.

2.3. Tracker

Following the minimum cost multicut formulation [12, 11, 9], the multi-person articulated tracking problem is formulated by a spatial-temporal graph partitioning. We first use our MSPAF bottom-up joint detector to generate a set of joint proposals and use top-down person detector to generate bounding box candidates in each frame. Then we construct a spatio-temporal graph based on the human body part candidates and introduce spatial edges within a frame and temporal edges across the video frames. We compute the potentials for nodes and edges following [11] and introduce several constraints to get feasible solutions. Since our human detector is robust and contains global context, we make use of the human bounding boxes to rule out the outlier body part proposals.

The tracking problem is formulated as an integer linear programming (ILP) problem. We follow [10, 11] to employ branch-and-cut algorithm of the ILP solver Gurobi to optimize it.

3. Experiments

3.1. Datasets and Evaluation

We evaluate our proposed method on "Multi-Person PoseTrack" dataset [11] and "ICCV 2017 PoseTrack Challenge" dataset [1]. We show that our approach outperforms the state-of-the-art PoseTrack method [11].

MOTA Head	MOTA Shou	MOTA Elb	MOTA Wri	MOTA Hip	MOTA Knee	MOTA Ankl	MOTA Total	MOTP Total	Prec Total	Rec Total
71.5	70.3	56.3	45.1	55.5	50.8	37.5	56.4	61.9	86.8	68.0

Table 1. Multiple Object Tracking (MOT) evaluation on PoseTrack Challenge validation set.

MOTA Head	MOTA Shou	MOTA Elb	MOTA Wri	MOTA Hip	MOTA Knee	MOTA Ankl	MOTA Total	MOTP Total	Prec Total	Rec Total
63.2	61.8	48.9	41.0	46.7	41.6	31.1	48.8	40.5	80.9	66.0

Table 2. Multiple Object Tracking (MOT) evaluation on PoseTrack Challenge (partial) test set.

MSCOCO Dataset contains over 66k images with 150k people and 1.7 million labeled keypoints.

The Multi-Person PoseTrack Dataset contains 30 videos for testing, where the length of the videos ranges between 41 and 151 frames. The number of persons ranges between 2 to 16 with more than 5 persons on average. The annotation includes 14 keypoints for each person.

ICCV 2017 PoseTrack Challenge Dataset contains 514 short video clips annotated for multi-person pose estimation and multi-person pose tracking. The dataset is very challenging compared to previous ones. The images often contain large crowd of people with various poses which lead to severe overlapping and occlusion. Isolated limbs and joints are common because of truncation.

We use Total AP to evaluate the multi-person pose estimation results and standard MOTA metric to evaluate the tracking performance.

3.2. Pose Estimation

In this section, we explore the performance of our pose detectors. We have implemented the state-of-the-art bottom-up PAF methods to achieve 58.5% mAP and top-down Mask R-CNN methods to achieve 63.1% mAP on MSCOCO dataset listed in Table 3. We show that our MSPAF outperforms PAF by a large margin, indicating the effectiveness of multi-scale features. However, we find that the top-down method outperforms bottom-up methods on MSCOCO, since top-down methods are able to directly utilize global information leading to less false-positives.

Method	Model	AP
BU	PAF	58.5
	MSPAF	60.8
TD	Mask-RCNN (Resnet-50)	62.7
	Mask-RCNN (Resnet-101)	**63.1**

Table 3. Pose Estimation Performance on MSCOCO dataset. We use BU to denote bottom-up methods and TD to denote top-down approaches.

We further test our models on ICCV 2017 PoseTrack Challenge dataset. We implement a two-stage baseline, which is to first use a person detector then apply single-person stacked hourglass pose estimation model. To our surprise, different from MSCOCO, the bottom-up method outperforms top-down methods on PoseTrack dataset by a large margin. From the table 4, we can see that our MSPAF achieves over 7% higher mAP than top-down methods on PoseTrack validation set.

We looked into the data to investigate the reason behind the bad performance of top-down methods on PoseTrack. We find that the PoseTrack dataset contains much larger number of people with occlusion and overlapping, which leads to inaccurate person detection. Although Mask R-CNN is able to provide accurate human keypoints on MSCOCO but fails on PoseTrack. Instead of using a simple head network in Mask R-CNN, we trained a single person pose detector following Yang et al. [16]. With the complex pose detector, we achieved a better performance of 60.3% mAP, as shown in Table 4.

Method	Model	AP
BU	MSPAF	**67.8**
TD	Mask-RCNN (Resnet-101)	57.4
	Person Detector + PRMs [16]	60.3

Table 4. Pose Estimation Performance on PoseTrack Challenge Validation Set. We use BU to denote bottom-up methods and TD to denote top-down approaches.

3.3. Person Detection Evaluation

We trained and tested our detection model with MSCOCO dataset.

As shown in Table 5, without ensemble, our best model can achieve 38.3% mAP on COCO minival set. Ensemble of three models brings an improvement of 2.3 mAP, and using Soft-NMS brings another 0.3 mAP gain.

We finally use ensemble and Soft-NMS model with 40.9% mAP and 54.4% person mAP for person detection in PoseTrack Challenge.

Model	mAP	mAP(person)
resnet152	35.9	47.7
resnet101-fpn	37.3	50.4
resnet152-fpn	38.3	51.6
ensemble	40.6	54.1
ensemble + SNMS	**40.9**	**54.4**

Table 5. Person Detector Performance

3.4. Multi-Person Pose Tracking

To investigate the performance of multi-person pose tracking, for fair comparison, we first use the same joint detector as [11]. The results are demonstrated in Table 6. Our

human detection results contain global context and structural information which help to rule out the false alarms. With human detection, we achieve a significant improvement (32.7 vs. 28.2 MOTA). The refinement process further improves the results (32.9 vs. 32.7 MOTA) by removing the outliers. We further show that our MSPAF model outperforms the state-of-the-art approach by a large margin(36.3 vs. 28.2 MOTA).

Method	Rcll	Prcn	MOTA	MOTP
PoseTrack	63.0	64.8	28.2	55.7
+ Human Detection	65.5	66.9	32.7	55.5
+ Refine	65.6	67.0	32.9	55.5
+ MSPAF	**68.4**	**68.5**	**36.3**	**55.7**

Table 6. Evaluation on PoseTrack dataset.

3.5. ICCV PoseTrack Challenge Results

Here we show our ICCV 2017 PoseTrack Challenge results. Table 7 and Table 8 demonstrates the results of per-frame multi-person pose estimation performance (AP). Our approach achieves the result of 67.8% and 63.6% for the validation set and (partial) test set respectively. Table 1 and Table 2 show the articulated tracking results. We achieve 56.4% MOTA for the validation set and 48.8% MOTA for the (partial) test set.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
79.1	77.3	69.9	58.3	66.2	63.5	54.9	67.8

Table 7. Per-frame multi-person pose estimation performance (AP) on PoseTrack Challenge validation set.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
74.7	71.9	65.6	56.4	62.2	57.5	51.0	63.6

Table 8. Per-frame multi-person pose estimation performance (AP) on PoseTrack Challenge (partial) test set.

4. Conclusion

In this work, we investigate the performance of the top-down and bottom-up methods on different datasets. We find that bottom-up methods are able to handle complex poses and crowded scenes but suffer from false alarms. We therefore propose to use a robust human detector to rule out those false alarms to get better performance. We propose a MSPAF body part detector to extract multi-scale features, which significantly improves the performance. We evaluate our approach on "Multi-Person PoseTrack" dataset and "ICCV 2017 PoseTrack Challenge" dataset. Our method achieves the state-of-the-art performance on both datasets.

References

[1] M. Andriluka, U. Iqbal, A. Milan, E. Insafutdinov, L. Pishchulin, J. Gall, and B. Schiele. Posetrack: A benchmark for human pose estimation and tracking. *arXiv preprint arXiv:1710.10000*, 2017.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. 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.

[3] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5561–5569, 2017.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.

[5] X. Chen and A. Gupta. An implementation of faster rcnn with study for region sampling. *arXiv preprint arXiv:1702.02138*, 2017.

[6] J. L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. *arXiv preprint arXiv:1703.06870*, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, B. Schiele, and S. I. Campus. Arttrack: Articulated multi-person tracking in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 4327, 2017.

[10] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[11] U. Iqbal, A. Milan, and J. Gall. Pose-track: Joint multi-person pose estimation and tracking. *arXiv preprint arXiv:1611.07727*, 2016.

[12] M. Keuper, E. Levinkov, N. Bonneel, G. Lavoué, T. Brox, and B. Andres. Efficient decomposition of image and mesh graphs by lifted multicuts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1751–1759, 2015.

[13] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. *arXiv preprint arXiv:1612.03144*, 2016.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[16] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. *arXiv preprint arXiv:1708.01101*, 2017.