The SCUT-Samsung Team's Technical Report on the PoseTrack Challenge

Yuejia Shen[1, *], Lin Sun[2, *], Jinmiao Cai[1, 3], Yao Li[1, 3], and Kui Jia[1]

[1]South China University of Technology, 381 Wushan Road, Guangzhou, China
[2]Samsung Smart Machine, Strategy and Innovation Center, 3655 North First Street, San Jose, CA, USA
[3]Shenzhen Cloudream Technology Co., Ltd., 2388 Houhai Avenue, Shenzhen, China
[*] indicates equal contribution

Abstract

We present in this report the technical details of our entry of the PoseTrack challenge. Our entry is for the setting of multi-person pose estimation in individual video frames (Challenges 1 and 2).

To perform pose estimation for an unknown number of persons in an image/frame, we take a top-down strategy that first detects individual persons in the image and then estimates the pose keypoint positions inside each detected bounding box. We use the Single Shot Multibox Detection (SSD) method for person detection, and the hourglass model for pose estimation. Both SSD and hourglass models are trained using additional (publicly available) datasets that contain human images. We finally report experiments on the PoseTrack validation and test datasets.

1. Introduction

Human pose estimation is a challenging task with many applications. For example, it may be used to help analyze human activities in videos, which may further be used for human computer interaction or video surveillance. Going beyond the earlier and simpler setting of single-person pose estimation, multi-person pose estimation [1] addresses the more practical challenge that allows multiple persons to appear in a test image, and the task is to find the keypoint pose positions for each person in the image.

In the literature, two main approaches (i.e., top-down or bottom-up ones) are proposed for multi-person pose estimation. Bottom-up approach [2, 5] first predicts all the body keypoint positions in the image, and then groups them to form pose estimation for each specific person. While achieving promising results on benchmark datasets recently [1], we argue that such an approach is limited in leveraging global configuration of human pose for estimating any specific keypoint positions. In contrast, top-down approach [9, 4] first trains human detector to find individual regions each of which contains a single person, and then performs pose estimation inside each detected region/bounding box. As such, both the model for human detection and that for pose estimation can be trained using global (and plentifully available) human and pose annotations in public datasets.

In this work, we adopt the top-down approach for our entry of the PoseTrack challenge. Particularly, we use Single Shot Multibox Detection (SSD) method [7] to do human detection and thus provide person proposals, and use a modified hourglass network [8] for each person proposal that outputs the predicted keypoints.

2. Description of our used method

2.1. The human detection implementation

We use SSD as the human detection network due to its efficiency and accuracy. In [7], the authors propose two architectures, namely SSD300 and SSD512. We mainly use SSD300 for the controlled studies on the PoseTrack dataset.

The base network of SSD is VGG16 [10], which is pre-trained on ImageNet. We use training settings as in the original SSD paper [7] to train our human detection network: the initial learning rate is set as 10^{-3}, which is decreased by a factor of 10 at iterations 10000, 12000, and 14000 respectively; the whole training procedure stops at iteration 15000; the momentum is set as 0.9 and weight decay as 5^{-4}. Except the pre-trained VGG16 base network, all other network parameters are initialized using Gaussian distribution. We use the Pascal VOC [3] rules to evaluate the detection performance.

Training with the PoseTrack dataset PoseTrack is a multi-person pose estimation dataset which contains 500 video sequences, 20K frames, and 120K body pose annotations. The PoseTrack dataset does not provide ground-truth per-

son bounding boxes. However, such bounding boxes can be easily obtained from the ground-truth body keypoint annotations, e.g., by defining some proper offsets from body keypoints. Such obtained bounding box annotations are less accurate than those manually labeled. Nevertheless, we use them as good alternatives to train human detectors. Results of SSD300 trained from scratch on the PoseTrack data are presented as below.

Table 1. The human detection performance of SSD300 by training from scratch using the PoseTrack data.

confidence threshold	mAP
0.15	0.47
0.3	0.45
0.5	0.40

Bounding boxes with confidence scores higher than each confidence threshold in Table 1 are kept. If the confidence threshold is set to be higher than 0.5, detection accuracy (mAP) would drop dramatically. So experiments with confidence threshold higher than 0.5 are not discussed in this report. Note that mAP increases when confidence thresholds are relatively low. In the final experiments, we set confidence threshold as 0.15.

Training with the VOC 07/12 data We directly apply SSD300 pre-trained on Pascal VOC 07/12 on the PoseTrack validation dataset. Results are shown in Table 2. Since Pascal VOC contains less training images than PoseTrack does, results in Table 2 are worse than those in Table 1.

Table 2. The human detection performance on the PoseTrack validation by training SSD300 using the PASCAL VOC 07/12 data.

confidence threshold	mAP
0.15	0.43
0.3	0.43
0.5	0.39

Training with the MS-COCO data We also train SSD300 with the MS-COCO data [6]. Results are shown in Table. 3.

Table 3. The human detection performance on the PoseTrack validation by training SSD300 using the MS-COCO data.

confidene threshold	mAP
0.15	0.46
0.3	0.52
0.5	0.40

Joint training on the MPII and PoseTrack datasets The quality of training data in MPII [1] is better than that of PoseTrack. We thus combine both datasets and train SSD300. Results in Table 4 demonstrate the benefits.

Table 4. The human detection performance on the PoseTrack validation by training SSD300 using both the MPII and PoseTrack data.

confidence threshold	mAP
0.15	0.52
0.3	0.50
0.5	0.43

MS COOC dataset provides ground truth bounding boxes of person, however the "ground truth" bouding boxes of MPII and PoseTrack dataset are generated us, so the detection mAP is just a reference but not an reality evalution result. When test on the PoseTrack dataset, we use the SSD model trained on COCO dataset.

Effect of overlap threshold on performance We analyze the effect of different overlap thresholds when training SSD300 with the PoseTrack data. Results in Table 5 show that overlap thresholds higher than 0.5 do not help. We thus set the overlap threshold as 0.5 in all experiments.

Table 5. Human detection performance of SSD300 on the PoseTrack validation with overlapping thresholds higher than 0.5.

overlap threshold	confidene threshold	mAP
0.6	0.1	0.45
0.6	0.2	0.45
0.6	0.4	0.33
0.7	0.1	0.01
0.7	0.2	0.01
0.7	0.4	0.08
0.8	0.1	0.19
0.8	0.2	0.19
0.8	0.4	N/A

2.2. The pose estimation implementation

We follow the hourglass model [8] for (single-person) pose estimation in each detected bounding box. Our model architecture is presented in Fig. 1. We use four blocks with two stacks in the hourglass.

3. Experiments on the PoseTrack dataset

We present in this section results on the PoseTrack validation set using different combinations of SSD300 and hourglass training settings. In all experiments, we set the confidence threshold of SSD as 0.15. Baseline results by OpenPose [2] are also shown in Tables 6 and 7 for comparison.

Table 6. The pose estimation performance on the PoseTrack validation ***using OpenPose trained on the MPII dataset***.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
61.2	59.0	50.1	37.7	51.3	42.6	37.4	49.3

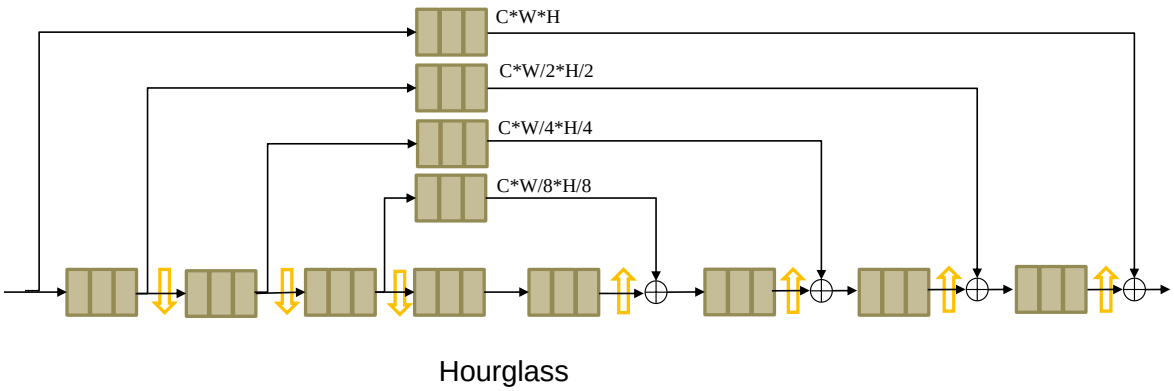

Figure 1. Illustration of the hourglass architecture used for (single-person) pose estimation in each detected bounding box. $\uparrow$ represents the up-sampling and $\downarrow$ is down-sampling. $\oplus$ means the features will be added and fed into the next block. W, H is the resolution of the feature maps of the pre-processing networks. Through the combination of feature maps with different resolutions, Hourglass can make a good prediction for the joint location.

Table 7. The pose estimation performance on the PoseTrack validation ***using OpenPose trained from scratch on the PoseTrack***.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
55.7	43.9	32.5	21.3	30.0	24.3	20.1	34.1

Tables 6 and 7 suggest that although the PoseTrack dataset has more training images than the MPII multi-person dataset does, the OpenPose trained on MPII and tested on the PoseTrack validation performs better than that trained directly using the PoseTrack training data. The possible reason is that the quality of training images (and annotations) in MPII is better than that in PoseTrack.

Table 8. The pose estimation performance on the PoseTrack validation ***using SSD300 (trained on VOC 07/12) + Hourglass (trained on PoseTrack)***.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
51.9	47.1	41.9	34.4	36.2	33.1	24.4	39.3

Table 9. The pose estimation performance on the PoseTrack validation ***using SSD300 (trained on PoseTrack) + Hourglass (trained on PoseTrack)***.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
55.2	49.3	45.4	35.7	38.5	36.3	25.3	41.8

Tables 8 and 9 tell that SSD trained from the scratch on this specific dataset performs better.

Table 10. The pose estimation performance on the PoseTrack validation ***using SSD300 (trained on PoseTrack) + Hourglass (trained on MPII)***.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
53.2	52.6	48.9	43.4	43.6	44.0	39.2	46.9

Table 11. The pose estimation performance on the PoseTrack validation ***using SSD300 (trained on MPII and PoseTrack) + Hourglass (trained on PoseTrack)***.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
58.4	50.8	46.9	38.0	39.0	37.5	26.6	43.5

Tables 10 and 11 tell that hourglass trained on MPII is better, particularly for the keypoints of shoulder, elbow, wrist, hip, knee, and ankle.

Table 12. The pose estimation performance on the PoseTrack validation ***using SSD300 (trained on MPII and PoseTrack) + Hourglass (trained on MPII and PoseTrack)***.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
62.1	61.7	60.0	53.4	54.1	51.7	44.3	55.8

Table 13. The pose estimation performance on the PoseTrack validation ***using SSD300 (trained on COCO) + Hourglass (trained on MPII and PoseTrack)***.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
67.0	66.8	63.6	56.5	55.6	54.6	47.7	59.4

Tables 10 and 12 tell that the performance of SSD can be further boosted when more training data are available.

3.1. Final results on the PoseTrack validation and (partial) test sets

We also train an SSD512 on the COCO dataset. For the final submission, we use such trained SSD512 as the person detector.

3.2. Visualization

We illustrate some of our multi-person pose estimation results in Fig. 2 and Fig. 3. In Fig. 2, even some parts of the

Table 14. The final pose estimation performance on the PoseTrack validation ***using SSD512 (trained on COCO) + Hourglass (trained on MPII and PoseTrack)***.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
69.3	68.5	65.4	57.7	57.9	55.8	50.3	61.3

Table 15. The final pose estimation performance on the PoseTrack partial test set ***using SSD512 (trained on COCO) + Hourglass (trained on MPII and PoseTrack)***.

Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
66.6	65.8	62.3	56.4	54.1	53.7	48.2	58.7

specific person can not be well visualized or the detection bounding boxes can not locate the person well, out networks can output the reasonable predictions.

Figure 2. The sample images from our architecture. We can detection the person well and model their pose.

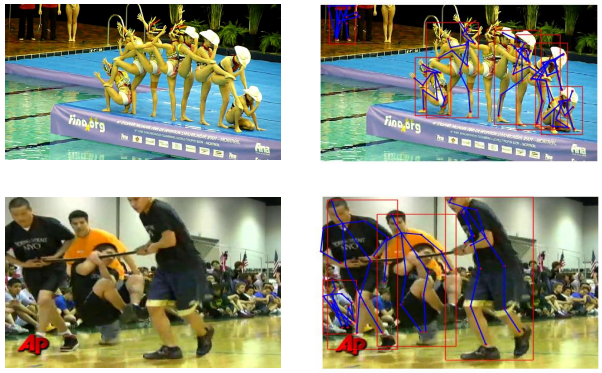

Figure 3. The sample images from our architecture. We can detection the person well and model their pose.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] 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.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[4] H. Fang, S. Xie, and C. Lu. Rmpe: Regional multi-person pose estimation. *arXiv preprint arXiv:1612.00137*, 2016.

[5] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, B. Andres, and B. Schiele. Articulated multi-person tracking in the wild. *arXiv preprint arXiv:1612.01465*, 2016.

[6] 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.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[8] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[9] T. Pfister, J. Charles, and A. Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.