Enhanced Two-Stage Multi-person Pose Estimation

Hiroto Honda(✉), Tomohiro Kato, and Yusuke Uchida

DeNA Co., Ltd., Tokyo, Japan
{hiroto.honda,tomohiro.kato,yusuke.a.uchida}@dena.com

Abstract. In this paper we introduce an enhanced multi-person pose estimation method for the competition of the PoseTrack [6] workshop in ECCV 2018. We employ a two-stage human pose detector, where human region detection and keypoint detection are separately performed. A strong encoder-decoder network for keypoint detection has achieved 70.4% mAP for PoseTrack 2018 validation dataset.

Keywords: Multi-person pose estimation · Keypoint detection

1 Introduction

The progress of human pose estimation is significant owing to the success of convolutional neural networks. However, the multi-person pose estimation problem is still challenging in the situations where there are various amounts of scale, rotation and overlapping (occlusion). We employ a top-down two-stage detector, where human region detection and keypoint detection are separately performed. For the first stage detector, we choose bounding box (or region of interest, ROI) regression output of a two-stage multi-person keypoint detector. To make the keypoint detection more accurate, we train the second stage detector that performs single-person pose estimation for each ROI.

The contributions of this report are twofold:

- We empirically show the effectiveness of a two-stage detector.
- We investigate the optimal design of the keypoint detector.

2 Related Work

The recently proposed approaches are categorized into two types: bottom-up and top-down. Bottom-up methods such as [1] first detect keypoints of multiple persons simultaneously and group them into individuals afterwards. On the other hand, top-down methods such as [3,10] detect each person's location first and detect keypoints afterwards. Our method is based on [10] which first detects the person regions, crops the regions from the input image, and localizes the keypoints using the keypoint detection network.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11130, pp. 217–220, 2019.
https://doi.org/10.1007/978-3-030-11012-3_18

3 Method

Our method detects human keypoints in a top-down and two-stage manner. At the first stage, the detector takes the whole image as an input and returns region-of-interests (ROIs) of persons. At the second stage, the keypoint detector takes the detected ROIs and locates each person's keypoints. In this section we describe the details of the two detectors using Fig. 1.

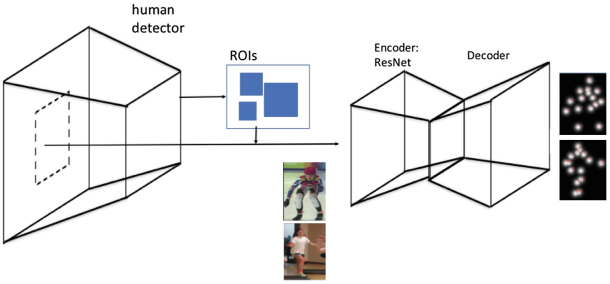

Fig. 1. Our two-stage network.

3.1 Person Detection

For the first stage, we adopt a multi-task detector that localizes bounding boxes and human keypoints at the same time. The detector is pretrained using images with bounding boxes and keypoints, thus already works as a multi-person keypoint detector. We pick the bounding box regression output of the detector and do not use the keypoint output. Compared with single-task (bounding box) detectors like Faster R-CNN [9], the bounding box regression results of the multi-task detector are more accurate due to the benefit from keypoint supervision.

3.2 Keypoint Detection

As the second-stage single-person keypoint detector, we employ an encoder-decoder network which is often referred to as an 'hourglass' structure. Human regions are cropped from the input whole image with margins and resized to a fixed image size. The hourglass network takes the cropped image and gives the heatmaps of each keypoint. The target is a set of K heatmaps $H_1 ... H_k$, each of which is generated with a 2D gaussian with $\sigma = 3.0$, centered at each keypoint.

We employ ResNet152 [4] for encoder and the simple decoder that has three sequential deconvolution - batchnorm [5] - ReLU blocks and one convolution layer. The intermediate channel width is 256 and deconvolution kernel size is 4×4.

4 Experiments

Training on the COCO Dataset. Firstly, the hourglass network is trained on the COCO train2017 dataset [8] with the Adam optimizer [7] for 90k iterations with batch size 64 and learning rate 1E-3. The learning rate is scheduled to be dropped by ×0.1 at 60k and 80k iterations. The duration of training is approximately 32 h on NVIDIA Tesla V100 GPU. We use horizontal flip, rotation within 40°, and scale variation within 30% as data augmentation.

Training on the PoseTrack2018 Dataset. The model trained on COCO is fine-tuned on Posetrack2018 dataset. We use the same setting as training on COCO, except that the initial learning rate is set to 1E-4.

4.1 Performance on PoseTrack 2018 Dataset

The pretrained Keypoint R-CNN network named X-101-32x8d-FPN, which is available at [2], is used for ROI detection. Each detected ROI is expanded by 60 pixels in every direction and resized to $(h, w) = (384, 288)$. Horizontal flip ensembling is used for the second-stage detection on each ROI. As shown in Table 1, our final result has achieved 70.4% and 65.9% of mAP with ResNet152 and ResNet50 respectively, on the PoseTrack2018 validation dataset. The visualization results are shown in Fig. 2.

Table 1. Performance on the PoseTrack 2018 validation dataset. PT and PT* stands for fine-tuning on the PoseTrack2018 dataset for 76k and 1k iterations respectively.

Encoder network	Trained on	Head	Shou	Elb	Wri	Hip	Knee	Ankl	Total
ResNet152	COCO, PT*	77.0	76.1	73.2	66.2	67.6	66.5	63.1	70.4
ResNet152	COCO, PT	76.1	74.3	71.6	65.4	62.3	64.7	61.5	68.5
ResNet152	COCO	27.8	77.1	73.4	65.7	68.5	67.8	62.8	60.9
ResNet50	COCO, PT*	75.4	73.5	67.4	59.7	62.0	61.3	57.0	65.9
ResNet50	COCO	26.8	69.7	62.5	54.1	57.8	54.5	50.8	51.9

4.2 Discussion

We observe that the AP result is improved by fine-tuning on the PoseTrack 2018 dataset but start to decay after 1000 iterations. More appropriate data pre-processing and data augmentation are considered to be necessary for the dataset. The difference between ResNet152 and ResNet50 is significant. There is a possibility that the second-stage network could be further improved by optimizing the network size or architecture.

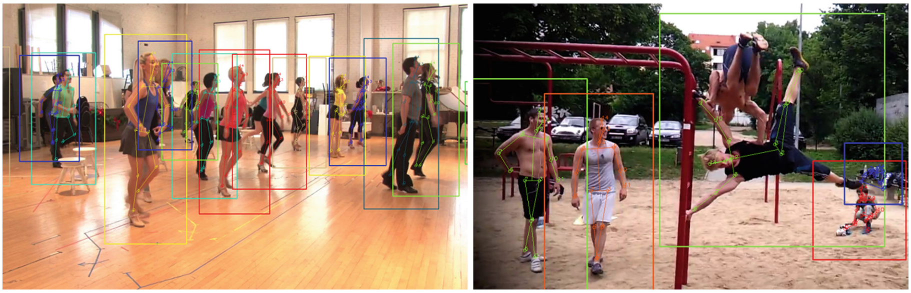

Fig. 2. Inference result on PoseTrack 2018 validation dataset. The right image includes person detection and keypoint detection failures.

5 Conclusions

We have proposed the enhanced multi-person pose estimation exploiting a two-stage human pose detector. The individual strong networks are employed for person region detection (first stage) and keypoint localization (second stage) respectively and the latter is trained on the COCO and PoseTrack2018 datasets. Finally, our whole pipeline achieves 70.4% mAP for PoseTrack 2018 validation.

References

1. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of CVPR (2017)
2. Girshick, R., Radosavovic, I., Gkioxari, G., Dollár, P., He, K.: Detectron (2018). https://github.com/facebookresearch/detectron
3. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: Proceedings of ICCV (2017)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of CVPR (2016)
5. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of ICML (2015)
6. Iqbal, U., Milan, A., Gall, J.: PoseTrack: joint multi-person pose estimation and tracking. In: Proceedings of CVPR (2017)
7. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of ICLR (2015)
8. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
9. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 1137–1149 (2015)
10. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 472–487. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_29