Multiple-Branches Faster RCNN for Human Parts Detection and Pose Estimation

Kaiqiang Wei and Xu Zhao(✉)

Institute of Image Processing and Pattern Recognition,
Shanghai Jiao Tong University, Shanghai 200240, China
zhaoxu@sjtu.edu.cn

Abstract. In this work, we primarily address multiple people pose estimation challenge by exploring the performance of Faster RCNN on human parts detection. We develop a multiple-branches Faster RCNN model for our specific task of detecting persons and their parts. Our model can improve the performance of detecting human parts and the whole persons, meanwhile speeding up detection process with shared weights. A part-based method is proposed to estimate multiple people poses, bringing recent advances on object detection to this task. Experiments demonstrate that our model achieves better performance than the original Faster RCNN model on our task. Compared with other pose estimation approaches, our approach achieves fair or better results.

1 Introduction

Human pose estimation has long been a concentrated concerned research topic in computer vision. Fast and robust human pose estimation has extensive application prospects such as human-computer interaction, virtual reality and intelligent monitoring. Pose estimation is still a tough task despite the long history of efforts. Complex variation in limb orientation, large occlusions, truncation and distractions from clothes or other overlapping objects make human pose estimation a challenging problem.

Most of the work on human pose estimation focuses on single person, which makes the task less intractable because one does not need to search over people's positions. We try to deal with multiple people pose estimation, taking an image with multiple people as input and outputting people's poses in an end-to-end way. This is a much tougher task than single person pose estimation, because neighboring people frequently overlap with each other with strong interference.

In this work, we adopt Faster RCNN [17] and explore its performance on human parts detection. We develop a multiple-branches Faster RCNN model for simultaneously human parts and whole person detection. The multiple-branches model has two branches, one for person detection and the other for human parts detection. We take a three-stage training strategy to make the two branches sharing the basic convolutional layers. The multiple-branches model outperforms the original Faster RCNN on our task by a large margin, while maintaining

C.-S. Chen et al. (Eds.): ACCV 2016 Workshops, Part III, LNCS 10118, pp. 453–462, 2017.
DOI: 10.1007/978-3-319-54526-4_33

the time efficiency of Faster RCNN as much as possible. In addition, a part-based method is proposed for multiple people pose estimation. We set a series of simple but practical rules to infer joints location from detected part boxes and connect the joints orderly to get human poses. We naturally take the predicted people location boxes as the basis of multiple people pose estimation, because person detection result is accurate and reliable. Experiments on public datasets demonstrate that our approach achieves fair or better results when compared with other pose estimation approaches.

2 Related Work

To deal with pose estimation problem, the Pictorial Strictures Model (PSM) [6] and many other PSM-based pose models [1] were developed in the early days. PSMs represent human parts with rectangle boxes and model parts connection relationship with tree structure. Yang [22] introduces part types based on PSM and uses several mixture part types to model human parts, which improves pose estimation performance. Part-based models such as poselets [3] and their variants [8] also play an import role in human pose estimation and action recognition. Random trees [18], graphical model [4] and convolutional neural networks [20] are also applied on pose estimation. A large amount of studies based on these learning-based methods had been presented and obtain competitive performance and good tractability. Most research focus on single person pose estimation only. There is not much literature addressing multiple people pose estimation, which is a more realistic problem.

Recently, tremendous progress has been achieved on object classification and detection due to convolutional neural network (CNN) [14]. The latest proposed Faster RCNN [17], one of the state-of-the-art object detection algorithms, shows amazing performance on PASCAL VOC object detection challenge. It dates back to RCNN [10], a region-based CNN detection algorithm, followed by SPP [12]. Fast RCNN [11] simplifies multi-level spatial pyramid pooling layers in SPP to be single-level RoI layer and adopts joint training with a multi-task loss. Faster RCNN introduces Region Proposal Network (RPN) to replace the selective search [21] to generate high-quality proposals and merges RPN with Fast RCNN to speed up detection process.

Many researchers apply convolutional neural networks on pose estimation and action recognition. DeepPose [20] trained a three-stage cascade of pose regressors to predict joints location in a holistic manner with high precision, each stage using the same 7-layer network and taking the output of previous stage as the input of next stage. DeepPose proves the feasibility and shows the power of convolutional neural networks on joints location task. Gkioxari et al. [9] train an RCNN-based detector with a multi-task loss to jointly optimize the task of key points location and action recognition. Recently, they [7] also proposed a part-based method by training part detectors for action and attributes classification, which shows that adding parts has essential contribution. Pfister et al. [15] introduce a novel network architecture to regress confidence heatmap of joint position to estimate human pose in videos, capitalizing the temporal information in videos.

3 Multiple-Branches Faster RCNN Model

3.1 Multiple-Branches Faster RCNN

In this work, we develop a multiple-branches Faster RCNN model for the task of detecting people's whole bodies and parts in the full image. The architecture of our model is shown in Fig. 1. Our model employs the 16-layer VGG network [19] as the basic model and has two branches above the shared convolutional layers, with each branch for one specific detection task. In the part branch, we detect human parts, including head, torso, forearm, upper arm, lower leg, upper leg as six different classes and train a Faster RCNN model to detect these parts. In the body branch, we train another Faster RCNN model to detect persons. The reason why we do not train one model to detect human body together with human parts is that they are not independent objects. Human body location box should cover all the human parts location boxes. This situation would affect the training of region proposal network, the proposal generating process, leading to inferior proposal quality and worse detection performance. The following experiments also prove our consideration, which will be described in detail next.

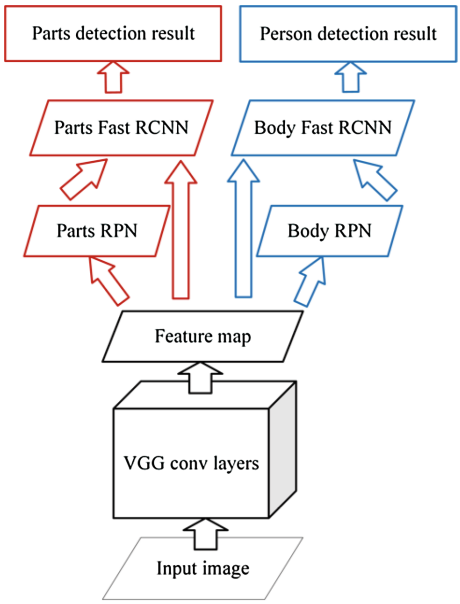

Fig. 1. Multiple-branches faster RCNN architecture.

Furthermore, we take similar ways as Faster RCNN to make our human parts detection model and human body detection model to share convolutional layers. In this way, the detection process can be accelerated. Our training strategy is elaborated as follows.

In the first stage, we train the initial human parts detection model. The human parts RPN model is trained to generate part proposals. Then these proposals are fed into Fast RCNN model to train part detection model. In this stage, both RPN and Fast RCNN are initialized with an ImageNet-pre-trained model.

In the second stage, human parts RPN and Fast RCNN are retrained in the same way as in the first stage, but initialized with the first stage Fast RCNN human parts detection network. We fix all the shared convolutional layers, only fine-tune the RPN and Fast RCNN's respective unique layers. Now the human parts detection model is ready, as the parts branch plotted in our network architecture (see Fig. 1).

In the third stage, we train human body detection model. The RPN is initialized with the first stage Fast RCNN human parts detection network. We only fine-tune the layers which are unique to RPN. Then we use the proposals generated by this stage's RPN to train our human body Fast RCNN detection model. The Fast RCNN in this stage is also initialized with the first stage's Fast RCNN human parts detection network, only updating the fully connected layers. Now we have the body detection branch.

Finally, the parts detection model and body detection model share the basic VGG convolutional layers. During testing process, we only need to do forward inference once to get human parts detection and person detection results, which would improve time efficiency. We call our model multiple-branches Faster RCNN model. Such modifications have advantages on typical tasks such as object and parts detection.

3.2 Multiple People Pose Estimation

For pose estimation, the task is to locate and connect body joints to form a kinematic tree structure. In order to transfer the task of pose estimation to object detection, we firstly set a series of rules to get parts location boxes to train our model. Then we try to reverse the process and set another series of inverse rules to get joints location from predicted parts location boxes. But the two processes are not completely reversible. Our rules are described in detail next.

From Joints to Boxes. To get parts location box annotations for training, we use two joints of a part to construct a tight box at first. If the ratio of the location rectangle box is less than 0.3, the shorter side is extended to be 0.3 times of the longer side. Then the box we get is still too tight to contain the whole part, so we resize the box to be 1.2 times larger, which leads to better detection performance by about 1% improvement of AP.

From Boxes to Joints. After the multiple-branches detection model is trained, we use it to detect human parts and get parts labels and location boxes. To restore joints location of parts, the box is resized to be 1/1.2 times smaller at first. Then, if the height-width ratio or width-height ratio is less than 0.5,

we choose the middle point of the shorter side as the predicted joint. Otherwise, we take the image patch inside the box to judge the part orientation. We compute the image patch's gradient along its two diagonal to determine whether the part lies in the left-top to right-bottom diagonal direction or in the right-top to left-bottom diagonal direction. Then the two joints of the part are set to be the corresponding two vertexes of the image patch.

From Joints to Pose. The last step is to link the joints to get pose. We need to connect lower leg to upper leg to form leg, and connect forearm to upper arm to form arm. Since our model does not distinguish left leg from right leg, we match lower leg with its nearby upper leg according to their distance. Once a whole leg is obtained, we modify the knee joint to be the middle point of upper leg's low joint and lower leg's up joint, and connect the whole leg with hip, knee, and ankle joints. Same ways are taken to match forearm with upper arm. Then we get modified elbow joint and connect a whole arm with shoulder, elbow, and wrist joints.

Multiple People Pose Estimation. Since our model detects not only all persons' parts location boxes, but also all persons' whole body location boxes, we naturally take the predicted human body location box as one person's ground truth location, and regard the parts boxes that overlap with the body box as the person's parts. Then we can use the rules described above to get multiple people's poses. By the way, our model performs very well on detecting persons so it is reliable to use the predicted person boxes as the basis of multiple person pose estimation.

4 Experiments

To evaluate our model, we conduct experiments on three public datasets and report part detection and pose estimation performance respectively with standard metrics. We describe the details of our experiments in the next.

4.1 The MPII Human Pose Dataset

The first dataset we use to train our model is the MPII Human Pose Dataset [2], a state-of-the-art benchmark for evaluation of articulated human pose estimation. The dataset includes around 25K images containing over 40K people with annotated human part joints totally. We select images with enough annotated part joints to construct ground truth part boxes from the training set only, without data augmentation. Finally we pick out about 17K images in total, excluding some incomplete annotations. We split the data into two sets, four fifths for training and one fifth for testing. We quantify the performance of our model by computing the average precision (AP), a standard metric for object detection, on our test dataset.

Table 1. Parts and person detection performance on the MPII human pose dataset under different settings for train-time and test-time proposals.

AP(%)	Head	Torso	Upper arm	Forearm	Upper leg	Lower leg	Person
Train 2k, test 300	71.8	69.0	30.1	17.7	29.9	32.3	89.1
Train 6k, test 1k	75.1	71.3	32.5	18.7	31.7	34.4	89.0

We train our Faster RCNN model based on the 16-layer VGG network [19]. The parameters we fine-tuned are the number of train-time region proposals and test-time region proposals. The original setting outputs 2k RPN proposals for Fast RCNN training and selects 300 proposals for testing. We give more proposals during training and testing process because our model needs to output much more boxes of human parts in one single image comparing with the task of object detection on PASCAL dataset. During testing, we set the score thresh of person detection to be 0.75, while the threshes for parts remain as 0.6. We also set the nms threshes for arms and legs to be 0.4 because these limbs overlap with each other more easily, while keeping the nms threshes for head, torso and person as 0.3 (Table 2).

Table 2. Comparison of our multiple-branches Faster RCNN and original Faster RCNN on parts and person detection on MPII human pose dataset. Train-time proposals of both are set to be 2k and test-time proposals of both are both set to be 1k.

AP (%)	Head	Torso	Upper arm	Forearm	Upper leg	Lower leg	Person
Faster RCNN [17]	54.7	56.2	22.5	13.3	20.0	23.8	85.9
Ours	74.3	70.2	30.5	18.1	30.3	32.4	89.2

In order to prove that our Multiple-branches Faster RCNN outperforms the original Faster RCNN on our task of detecting parts and persons, we also train a original Faster RCNN model to treat head, torso, forearm, upper arm, lower leg, upper leg and person as seven different classes. We set the train-time proposals to be 2k and test-time proposals to be 1k, and report the AP metric in Table 1. As can be seen, we get significant improvement on parts and person detection. Figure 2 shows the head and torso detection comparison. The AP of our model is 74.3% for head and 70.2% for torso, while the AP of the original model is 54.7% for head and 56.2% for torso. The improvement of AP is mainly caused by higher recall of our multiple-branches Faster RCNN model, due to its better-trained Region Proposal Network.

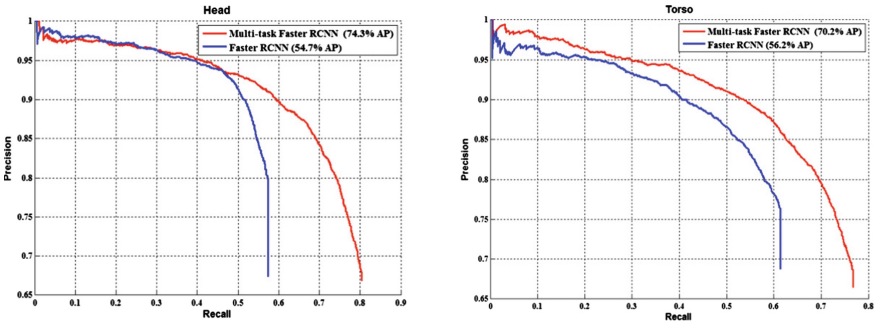

Fig. 2. AP comparison for head and torso detection between our multiple-branches faster-RCNN model and the original faster-RCNN model.

4.2 The PASCAL VOC Dataset

The second dataset we applied on to evaluate our trained model is PASCAL VOC09 person detection val dataset. This dataset has 1446 images with human joints annotated. So we are able to construct ground truth boxes with the annotated joints. The difference is that there are no head ground truth boxes in the PASCAL VOC09 dataset, but some facial points such as eyes, nose, ears and so on are provided. So we have to construct the head ground truth boxes with facial points. Torso box is also different. The MPII dataset provides pelvis annotation so we use pelvis joint and head's low joint to construct torso box, and the width box of ground truth box is about 0.3 times of its height. In PASCAL VOC09 dataset, we have to construct torso box using two joints of shoulders and two joints of hips, then we shrink the box's width to be 0.5 times of its original width.

Table 3 shows the AP comparison of our model and the original faster-RCNN model on this dataset. Our model gains better detection performance then the original faster-RCNN model on parts and person detection. Gkioxari et al. [7] also trained CNN models to detect human parts, including head, torso, legs on this dataset. We make comparison with their result in Table 4. Higher threshold of intersection-over-union (σ in Table 4) means stricter detection measure. Usually the threshold is set to be 0.5. As can be seen, our model outperforms Gkioxari's at high thresholds, which proves our model's good generalization ability.

Table 3. Comparison of our multiple-branches Faster-RCNN and original Faster-RCNN on parts and person detection on PASCAL VOC09 person detection dataset. Train-time proposals of both are set to be 2k and test-time proposals of both are set to be 1k.

AP (%)	Head	Torso	Upper arm	Forearm	Upper leg	Lower leg	Person
Faster RCNN [17]	32.6	13.0	14.6	7.7	8.6	11.9	54.2
Ours	39.7	16.5	17.8	9.6	14.5	17.9	56.0

Table 4. Comparison of parts detection performance on the PASCAL VOC09 person detection dataset.

AP(%)		$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$
Ours	Head	51.8	49.3	45.9	39.8
	Torso	47.0	41.9	31.2	17.5
	Upper leg	29.8	26.4	21.4	15.5
	Lower leg	32.9	29.6	24.8	17.5
Gkioxari's [7]	Head	55.2	51.8	45.2	31.6
	Torso	42.1	36.3	23.6	9.4
	Legs	34.9	27.9	20.0	10.4

4.3 The LSP Dataset

The third dataset we used is the Leeds Sport Dataset (LSP) [13] and its extension. The LSP dataset contains about 2k pose images annotated with 14 joints, one half for training and the other half for testing. The LSP extension dataset contains 10k images, which can be used for training. We do data augmentation with image horizontal flip on the training set. The anchor size of the region proposal network is changed to $[16, 32, 64]$ to fit the smaller image scale. We train our model and then get human pose according to the rules as described in the previous section. Percentage of Correct Parts (PCP) is reported on the dataset. PCP is a widely accepted metric to measure pose estimation performance. We report PCP for single person. We match left or right predicted part joints with left or right ground truth part joints according to their distance. Then the distance between two predicted joints of a part and two corresponding ground-truth joints is computed. A part is considered detected correctly if the distance is less than half of the part length.

We make comparison with other mainstream pose estimation approaches in Table 5. As can be seen, the average PCP of our approach is 63.0%, which is comparable to the other approaches. We achieve higher PCP on head (87.4%) and lower leg (74.2%) detection, which are the best among these approaches. Head and lower legs are large parts and overlap less easily with other parts. So our part-based approach is suitable. While upper arm and forearm are small

Table 5. Comparison of PCP performance at 0.5 of other approaches and ours on the LSP dataset.

Method	Head	Torso	Upper arm	Lower arm	Upper leg	Lower leg	Average
Andriluka [1]	74.9	80.9	46.5	26.4	67.1	60.7	55.7
Dontone [5]	79.2	81.6	45.1	24.7	66.5	61.0	55.5
Yang [22]	79.3	82.9	56.0	39.8	70.3	67.0	62.8
Pishchulin [16]	78.1	87.5	54.2	33.9	75.7	68.0	62.9
Ours	87.4	84.9	47.5	34.6	72.6	74.2	63.0

Fig. 3. Pose estimation results for single person and multiple persons.

and easily covered, other inference approaches [16,22] based on the relationship between parts achieves better results. In addition, the test time for one image is about 350 ms on K40 GPU. Figure 3 shows some pose estimation examples for single person and multiple persons. Images in the first row come from MPII dataset and images in the second row come from the LSP test set.

5 Conclusion

Our work explores the performance of Faster RCNN on human parts and whole body detection. The multiple-branches Fast RCNN model we developed shows comparative advantages on human parts and whole body detection. We speed up the detection process by making the parts detection model and person detection model to share weights. Our multiple-branches model can also be applied to other object and object part detection tasks with similar logic relationship of parts and wholes. We provide a practical part-based method to estimate multiple people poses. Compared with other pose estimation approaches, our method achieves comparative or better results.

Acknowledgement. We gratefully acknowledge that this work is supported by the fundings from National Natural Science Foundation of China (61273285, 61375019).

References

1. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: people detection and articulated pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1014–1021. IEEE (2009)
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3686–3693. IEEE (2014)
3. Bourdev, L., Malik, J.: Poselets: body part detectors trained using 3D human pose annotations. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1365–1372. IEEE (2009)

4. Chen, X., Yuille, A.L.: Articulated pose estimation by a graphical model with image dependent pairwise relations. In: Advances in Neural Information Processing Systems, pp. 1736–1744 (2014)
5. Dantone, M., Gall, J., Leistner, C., Van Gool, L.: Human pose estimation using body parts dependent joint regressors. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3041–3048. IEEE (2013)
6. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Comput. Vis. **61**, 55–79 (2005)
7. Gkioxari, G., Girshick, R., Malik, J.: Actions and attributes from wholes and parts. arXiv preprint arXiv:1412.2604 (2014)
8. Gkioxari, G., Hariharan, B., Girshick, R., Malik, J.: Using k-poselets for detecting people and localizing their keypoints. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3582–3589. IEEE (2014)
9. Gkioxari, G., Hariharan, B., Girshick, R., Malik, J.: R-CNNs for pose estimation and action detection. arXiv preprint arXiv:1406.5212 (2014)
10. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 580–587. IEEE (2014)
11. Girshick, R.: Fast R-CNN. arXiv preprint arXiv:1504.08083 (2015)
12. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8691, pp. 346–361. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10578-9_23
13. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC, vol. 2, p. 5 (2010)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp. 1097–1105 (2012)
15. Pfister, T., Charles, J., Zisserman, A.: Flowing convNets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1913–1921 (2015)
16. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 588–595 (2013)
17. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)
18. Rogez, G., Rihan, J., Ramalingam, S., Orrite, C., Torr, P.H.: Randomized trees for human pose detection. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014)
20. Toshev, A., Szegedy, C.: DeepPose: human pose estimation via deep neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1653–1660. IEEE (2014)
21. Uijlings, J.R., Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. Int. J. Comput. Vis. **104**, 154–171 (2013)
22. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE Trans. Pattern Anal. Mach. Intell. **35**, 2878–2890 (2013)