FollowMeUp Sports: New Benchmark for 2D Human Keypoint Recognition

Ying Huang[1(✉)], Bin Sun[2], Haipeng Kan[2], Jiankai Zhuang[3],
and Zengchang Qin[3]

[1] Alibaba Business School, Hangzhou Normal University, Hangzhou, China
yw155@buaa.edu.cn
[2] Keep Inc., Beijing, China
{sunbin,kanhaipeng}@keep.com
[3] Intelligent Computing and Machine Learning Lab, School of ASEE,
Beihang University, Beijing, China
{zhuangjk,zcqin}@buaa.edu.cn

Abstract. Human pose estimation has made significant advancement in recent years. However, the existing datasets are limited in their coverage of pose variety. In this paper, we introduce a novel benchmark "FollowMeUp Sports" that makes an important advance in terms of specific postures, self-occlusion and class balance, a contribution that we feel is required for future development in human body models. This comprehensive dataset was collected using an established taxonomy of over 200 standard workout activities with three different shot angles. The collected videos cover a wider variety of specific workout activities than previous datasets including push-up, squat and body moving near the ground with severe self-occlusion or occluded by some sport equipment and outfits. Given these rich images, we perform a detailed analysis of the leading human pose estimation approaches gaining insights for the success and failures of these methods.

Keywords: Pose estimation · Benchmark testing · Performance evaluation

1 Introduction

Human pose estimation is an important computer vision problem [1]. Its basic task is to find the posture of a person via recognising human joints and rigid parts from normal RGB images. The extracted pose information is essential to modelling and understanding the human behaviours, and can be used in many vision application problems, such as virtual/augmented reality, human-computer interaction, action recognition and smart perception.

In the psst few years, pose estimation methods based on deep neural network techniques have achieved great progress [2–4]. Although the performance of some human pose estimation models (e.g. [5–7]) is almost saturated on the above mentioned datasets, applying these high-precision algorithms to the other specific industrial tasks shows a degradation in accuracy. For instance, one application case is workouts or sports scoring. In this case, lots of activities have severe self-occlusion or unusual postures, such as push-up and crunch. We find out the models [8–10] trained on the MS-COCO dataset [11] cannot correctly detect body joints with atypical postures, as shown in Fig. 1. In the top-right image of Fig. 1, the right knee is falsely detected as left knee. In the top-left and lower-part images of Fig. 1, some body joints, such as shoulders, knees and ankles, are missed in prediction. Since the pose estimation results of the same person in the standing posture are correct, we argue the false predictions are caused by the abnormal postures. Current datasets lack the corresponding samples [12, 13].

We use the MS-COCO dataset [11] as an example to analyse the distribution of human postures. In our statistics, the number of human instances in standing posture achieves 102,495 (84.53%) while people in other postures only have 18,756 (15.47%) as shown in Fig. 2. The human instances in a horizontal position or an uncommon pose are extremely rare. This makes the model unable to learn the knowledge of irregular postures during training.

To improve the performance of human pose estimation in the certain sports situation, a large-scale human keypoints benchmark is presented in this paper. Our benchmark significantly advances state-of-the-art in terms of particular activities, and includes more than 16,000 images of people. We used the workout class videos as a data source and collected images and image sequences using queries based on the descriptions of more than 200 workout activity types. For each activity type, there are 3 different shot angles. This results in a diverse set of images covering not only different workout activities, but contrasting postures. This allows us to enhance the current human pose estimation methods.

2 Related Work

There are several human keypoints datasets presented in the past decades. Buffy dataset [14] and PASCAL stickmen dataset [15] only contain upper-bodies, but we need to process the full-body. In these two datasets pose variation is insignificant. The contrast of image frames is relatively low in the Buffy dataset.

The UIUC people dataset [16] contains 593 images (346 for training, 247 for testing). Most people in the images are playing badminton. Some people are playing jogging, Frisbee, standing, walking, etc. There are very aggressive pose and spatial variations. However, the activity type is limited in this dataset.

The sport categories of Sport image dataset [17] is more plentiful, which including soccer, cycling, acrobatics, American football, croquet, golf, horseback riding, hockey, figure skating, etc. The total number of images is 1299 (649 of them are split as training set and the rest as testing set).

Leeds Sports Poses (LSP) dataset [1] includes 2000 images, where one half for training and the other half for testing. The dataset shows people involved in various sports.

Fig. 1. Limitations of applying current pose estimation models on some workout postures, which have severe self-occlusion. Some body keypoints are falsely detected or missed in prediction even the background is plain.

The image parsing (IP) dataset [18] is a small dataset and contains 305 images of fully visible people, where 100 images for training and 205 images for testing. The dataset consists of various activities such as dancing, sports and acrobatics.

The MPII Human pose dataset [12] consists of 24,589 images, in which 17,408 images with 28,883 annotated people are split for training. During the testing stage, one image may contain multiple different evaluation regions that consist of a non-identical number of people. [20] defines a set of 1,758 evaluation regions on the test images with rough position and scale information. The evaluation metric deploys mean Average Precision (mAP) of the whole body joint prediction. The accuracy results are evaluated and returned by the staff members of the MPII dataset.
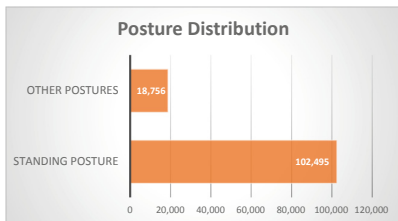
Fig. 2. The posture distribution of MS-COCO dataset. Around 85% human instances are standing with good, upright posture.

The MS-COCO keypoints dataset [11] includes training, validation and testing sets. On the COCO 2017 keypoints challenge, training and validation sets have 118,287 and 5000 images respectively, totally containing over 150,000 people with around 1.7 million labelled keypoints. In experiments, we perform ablation studies on the validation set. To analyse the effect of training, we also combine the COCO train set with the FollowMeUp train set to validate that new images will not affect the model's generality performance.

The DensePose-COCO dataset [19] has reannotated dense body surface annotations on the 50k COCO images. These dense body surface annotations can be understood as continuous part labels of each human body.

The PoseTrack dataset [13] includes both multi-person pose estimation and tracking annotations in videos. It can perform not only pose estimation in single frames, but also temporal tracking across frames. The dataset contains 514 videos including 66,374 frames in total. The annotation format defined 15 body keypoints. For the single-frame pose estimation, the evaluation metric uses mean average precision (mAP) as is done in [20].

3   The Dataset

3.1   Pose Estimation

The key motivation directing our data selection strategy is that we want to represent rare human postures that might be not easily accessed or captured. To this end, we follow the method of [21] to propose a two-level hierarchy of workout activities to guide the collection process. This hierarchy was designed according to the body part to be trained during the exercise. The first level is the body part interested to be trained, such as shoulder, whereas the second level is specific workout activities that can strengthen the muscles of shoulder.

Data Collection. We select candidate workout videos according to the hierarchy and filter out videos of low quality and those that people are truncated. This resulted in over 600 videos spanning over 200 different workout types with three shot angles. We also filter out the frames in which pose is not recognisable due to

poor image quality, small scale and dense crowds. This step resulted to a total of 110,000 extracted frames from all collected videos. Secondly, since different exercises have disparate periods, we manually pick key frames with people from each video. We aim to select frames that either depict the whole one exercise period in a substantially different pose or different people with dissimilar appearance. The repeated or no significant distinction postures are ignored. Following this step we annotate 16,519 images. We rough randomly split the annotated images for training and use the rest for testing. Images from the same video are either all in the training or all in the test set. We last obtain the train set of 15,435 images and test set of 1,084 images.

Data Annotation. We follow the keypoint annotation format of COCO dataset, where 17 body keypoints are defined. This design facilitates us to utilise the common samples of COCO dataset during training. Following [11] the left/right joints in the annotations refer to the left/right limbs of the person. Additionally, for all body joints the corresponding visibility is annotated. At test time both the accuracy of joints localisation of a person along with the correct match to the left/right limbs are evaluated. The annotations are performed by in-house workers and inspected by authors. For some unqualified and incorrect annotations are modified continuously until totally correct. To maintain the quality of annotations, we arranged a number of annotation training classes for all annotation workers to unify the standard of annotation. We also supervise and handle some uncertain cases for workers during annotation.

Pose Estimation Evaluation Metrics. Some previous keypoints evaluation metrics rely on the calculation of body limbs' length, such as PCP, PCK and PCK_h used in [12]. However, the workout activities usually have specific postures where the limb's length may be near 0 if the limb is perpendicular to the image plane and the evaluation is not numeric stable in these cases. Therefore comparing the distance between points of groundtruth and prediction directly is more sensible. Here we follow the COCO keypoints dataset, using 5 metrics to describe the performance of a model. They are AP (i.e. average precision), $AP^{0.5}$, $AP^{0.75}$, AP^M and AP^L, as illustrated in Table 1. In the matching between predictions to groundtruth, a matching criterion called object keypoint similarity (OKS) is defined to compute the overlapping ratio between groundtruth and predictions in terms of point distribution [11]. If OKS is larger than one threshold value (e.g. 0.5), the corresponding groundtruth and prediction are considered as a matching pair and the correctness of predicted keypoint types is further analysed. Here OKS is similar to the intersection over union (IoU) in the case of object detection. Thresholding the OKS adjusts the matching criterion. Notice that in general applications, $AP^{0.5}$ gives a good accuracy already. When computing AP (averaged across all 10 OKS thresholds), 6 thresholds exceed 0.70 are over strict due to unavoidable jittering in annotations.

Table 1. Evaluation metrics on the COCO dataset.

Metric	Description
AP	AP at OKS[*] $= 0.50 : 0.05 : 0.95$ (primary metric)
AP$^{0.5}$	AP at OKS $= 0.50$
AP$^{0.75}$	AP at OKS $= 0.75$
AP^{M}	AP for medium objects: $32^2 < area < 96^2$
AP^{L}	AP for large objects: $area > 96^2$

[*]OKS–Object Keypoint Similarity, same role as IoU

4   Analysis of the State of the Art

In this section we first compare the leading human pose estimation methods on the COCO keypoints dataset, and then analyse the performance of these approaches on our benchmark.

The basis of the comparison is that we note that there is no uniform evaluation protocol to measure the performance of existing methods from a view of practical application. Although human pose estimation is one of the longest-lasting topics, and significant performance improvement has been achieved in the past few years, some reported accuracies in these approaches are obtained through several post-processing steps or some strategies used in the dataset challenge. For example, performing multi-scale evaluation, refining results by a different method, or precision is evaluated at one image scale while speed is recorded at another scale. These post-processing steps interfere the judgement in identifying the strength and weakness of an algorithm. Therefore, evaluating a method without any post-processing steps and strategies is more objective and more valuable for the research and practical application.

The aim of the analysis is to evaluate the generality of the current models on the different datasets and their performance to the unseen samples, identify the existing limitations and stimulate further research advances.

Currently, there are two main categories of solutions: top-down methods [7,22–26] and bottom-up methods [9,10,27–30]. Top-down methods can be seen as a two-stage pipeline from global (i.e. the bounding box) to local (i.e. joints). The first stage is to perform human detection and to obtain their respective bounding boxes in the image. The second stage is to perform single person pose estimation for each of the obtained human regions. [7] deploys multiple high-to-low resolution subnetworks with repeated information exchange across multi-resolution subnetworks. This design obtains rich high-resolution representations and leading more accurate result. [22] utilises a Symmetric Spatial Transformer Network to handle inaccurate bounding boxes. [24] uses simple deconvolution layers to obtain high-resolution heatmaps for human pose estimation. On the side of bottom-up methods, [9] proposes a limb descriptor and an efficient bottom-up grouping approach to associate neighbouring joints. [10] modifies the network architecture of [9] and optimises the post-processing steps to achieve real-time speed on the CPU devices. [30] designs two new descriptors based on [9] for body
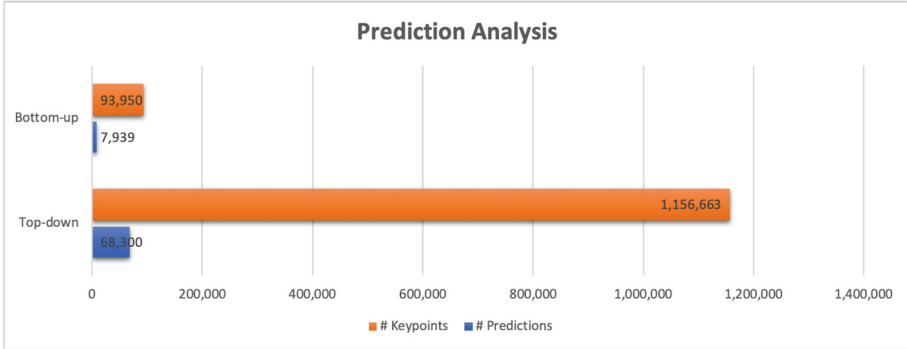
Fig. 3. The comparison of the numbers of effective instance predictions and body keypoints between top-down and bottom-up methods. The prediction number of top-down method is around 10 times higher than bottom-up method.

joints and limbs with the additional variable of object's spread. [28] presents a network to simultaneously output keypoint detections and the corresponding keypoint group assignments. [31] designs a feedback architecture that combining the keypoint results of other pose estimation methods with the original image as the new input to the human pose estimation network. In our analysis we consider 8 state-of-the-art multi-person pose estimation methods, which are listed in Table 2.

We compare the performance of each approach in terms of accuracy and speed on the COCO dataset and our novel FollowMeUp dataset. All the experiments are performed on a desktop with one NVIDIA GeForce GTX-2080Ti GPU. Since all testing approaches are trained and optimised on the COCO dataset, their open source codes have the corresponding configurations, we directly use their default parameters in our testing.

4.1 Comparisons of Approaches on the COCO Dataset

Table 2 presents the comparison results of testing approaches on the COCO dataset. The upper part of Table 2 are top-down approaches. [7] has the highest AP precision of 0.753. Note that the runtime costs around 50 ms as this only includes the part of pose estimation since this open source library uses the groundtruth of human bounding box as the human detection results on the COCO validation set. [24] and [22] have a relatively lower accuracy than [7] using smaller input sizes, which illustrates that the high-resolution and detailed representation is important for the task of human pose estimation. Note that some post-processing strategies, such as multi-scale and flip, are ignored to obtain the actual performance in the real application environments.

For the bottom-up methods, [9] achieves the fastest speed. [30] attains the highest precision in this group. The joint grouping part of [30] costs much longer time than [9]. [10] has around 7% degradation compared with [9] due to using a light-weight network architecture. We also see that the precision of bottom-up

Table 2. Comparisons of pose estimation results on the COCO 2017 validation set.

Type	Method	AP	$AP^{0.5}$	$AP^{0.75}$	AP^{M}	AP^{L}	Input size	Runtime
Top-down	HRNet [7]	**0.753**	**0.925**	**0.825**	**0.723**	**0.803**	384×288	0.049^*
	Xiao [24]	0.723	0.915	0.803	0.695	0.768	256×192	0.110
	RMPE [22]	0.735	0.887	0.802	0.693	0.799	320×256	0.298
Bottom-up	PAF [9]	0.469	0.737	0.493	0.403	0.561	432×368	**0.081**
	Osokin [10]	0.400	0.659	0.407	0.338	0.494	368×368	0.481
	PifPaf [30]	0.630	0.855	0.691	0.603	0.677	401×401	0.202
	AE [28]	0.566	0.818	0.618	0.498	0.670	$\mathbf{512 \times 512}$	0.260
	PoseFix [31]	0.411	0.647	0.412	0.303	0.559	384×288	0.250

*: without human detection

algorithms are lower than top-down methods. After detailed analysis, we find that the numbers of predicted effective keypoints of bottom-up methods are around 10 times less than top-down methods as illustrated in Fig. 3. We note that top-down methods correspond to performing single-person pose estimation on each detected human region. Single-person pose estimation can output all types of keypoints even the keypoint is occluded or truncated. However, for multi-person bottom-up methods, two or more overlapping keypoints with the same type can only be detected one due to depth information is not available on the RGB image. For the COCO dataset, there are a lot of crowded and occluded human instances. Therefore, the performance of bottom-up methods is weakened. In the FollowMeUp dataset, the crowding case is rare while most human instances have self-occlusion. We perform the same comparison on the FollowMeUp dataset and validate that bottom-up methods have comparable performance to top-down approaches in this circumstance.

4.2   Comparisons of Approaches on the FollowMeUp Dataset

Table 3 provides the comparison results of testing approaches on the COCO dataset. Since the open source libraries of [7] and [24] do not provide default human detection algorithm, using different human detector may bias the precision distribution, thus we do not test [7] and [24] on the FollowMeUp dataset. We are surprised that [22] obtains a very high precision value. However, the training set only including the COCO dataset of [9] just achieve the precision of 0.778. We argue that the training set of [22] may include other samples except the COCO dataset with particular postures. In this dataset, the precision of [10] decreases by 13% in $AP^{0.5}$ compared with [9], which indicates that the generality of [10] is also narrowed. We use the results of [9] as the initial poses of [31]. Through pose refinement, [31] improved the pose estimation results by 0.4%.

4.3   The Effect of Training on the FollowMeUp Dataset

To validate the effectiveness of samples with particular postures, we retrain the model on the COCO + FollowMeUp train set using the method of [9]. Testing

**Table 3.** Comparisons of pose estimation results on the FollowMeUp dataset.

Type	Method	$AP^{0.5}$	$AP^{0.6}$	$AP^{0.7}$	$AP^{0.8}$	$AP^{0.9}$
Top-down	RMPE [22]	0.975	0.948	0.885	0.787	0.421
Bottom-up	PAF [9]	0.778	0.728	0.625	0.474	0.326
	Osokin [10]	0.645	0.585	0.520	0.370	0.215
	PoseFix [31]	0.782	0.716	0.621	0.466	0.334

Table 4. Comparisons of pose estimation results on the FollowMeUp dataset.

Method	Train set	Test set	$AP^{0.5}$	$AP^{0.6}$	$AP^{0.7}$	$AP^{0.8}$	$AP^{0.9}$
PAF [9]	COCO	FollowMeUp	0.778	0.728	0.625	0.474	0.326
PAF [9]	COCO + FollowMeUp	FollowMeUp	0.964	0.959	0.926	0.876	0.691

Table 5. Comparisons of pose estimation results on the COCO dataset.

Method	Train set	Test set	AP	$AP^{0.5}$	$AP^{0.75}$	AP^{M}	AP^{L}
PAF [9]	COCO	COCO	0.465	0.740	0.447	0.379	0.597
PAF [9]	COCO + FollowMeUp	COCO	0.465	0.748	0.454	0.373	0.605
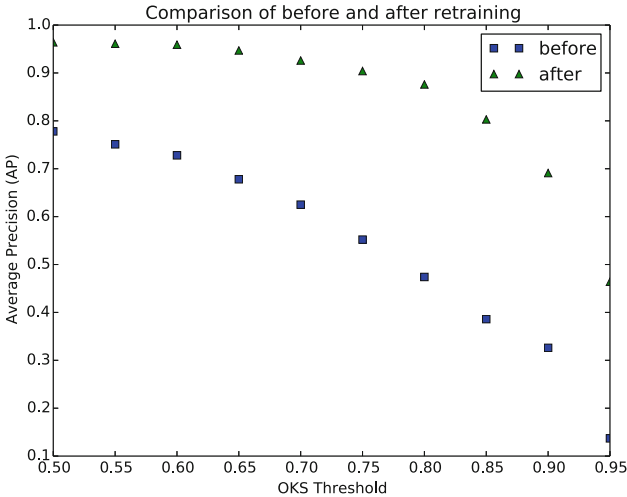
Fig. 4. Comparison of estimation accuracy before and after retraining on the FollowMeUp dataset. The accuracy of retrained model (marked as green triangles) has an obvious improvement. (Color figure online)

is performed both on the FollowMeUp test set and COCO validation set. The results of testing are provided in Table 4. We notice that the performance of the retrained model is greatly improved by around 20% in $AP^{0.5}$. While the

threshold of AP becomes more strict, the AP value is decreased. Even in the most strict threshold of 0.9, the AP value attains 0.691, which is higher than the model before retraining by 37%. The accuracy comparison of before and after retraining on the FollowMeUp dataset is shown in Fig. 4. We also perform testing on the COCO validation set using before and after retraining models to check whether the model can maintain the performance on the COCO dataset. In Table 5 we see that before and after retraining the precision has no change. The generality of the retrained model is preserved. These results show that increasing some unusual samples which had not been learnt by the model before is an effective way to improve the accuracy in some specific scenes.

5   Conclusion

The problem of human pose estimation has obtained a great progress in recent years. This progress cannot be done without the development of large-scale human pose datasets. However, the existing human pose datasets are not sufficient for some particular application environments. In this paper, we propose a new large-scale workout activity human pose dataset, which provides a wide variety of sport exercise postures. We select 8 state-of-the-art multi-person pose estimation approaches and compare their performance on both the popular COCO keypoints dataset and our FollowMeUp dataset. The comparison results show that most methods trained on the COCO dataset do not have ideal performance on the FollowMeUp dataset. We also test the generality of the model using the data of both COCO and FollowMeUp datasets. The test results show that training on the data of both COCO and FollowMeUp datasets will not affect the performance of the model on the COCO dataset but the performance of the model on the FollowMeUp dataset is greatly improved. In the future, we will continue investigate pose tracking [32], multi-view action recognition [33], and light-weight network design [34] approaches on the FollowMeUp dataset.

References

1. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: British Machine Vision Conference (BMVC), p. 5 (2010)
2. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4732 (2016)
3. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
4. Luvizon, D.C., Picard, D., Tabia, H.: 2D/3D pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5137–5146 (2018)

5. Chu, X., Ouyang, W.L., Li, H.S., Wang, X.G.: Structured feature learning for pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4715–4723 (2016)
6. Chu, X., Yang, W., Ouyang, W.L., Ma, C., Yuille, A.L., Wang, X.G.: Multi-context attention for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1831–1840 (2017)
7. Sun, K., Xiao, B., Liu, D., Wang, J.D.: Deep high-resolution representation learning for human pose estimation. arXiv preprint arXiv:1902.09212 (2019)
8. He, K.M., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017)
9. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7291–7299 (2017)
10. Osokin, D.: Real-time 2D multi-person pose estimation on CPU: lightweight Open-Pose. arXiv preprint arXiv:1811.12004 (2018)
11. 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
12. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3686–3693 (2014)
13. Andriluka, M., et al.: PoseTrack: a benchmark for human pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5167–5176 (2018)
14. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
15. Eichner, M., Ferrari, V., Zurich, S.: Better appearance models for pictorial structures. In: British Machine Vision Conference, p. 5 (2009)
16. Tran, D., Forsyth, D.: Improved human parsing with a full relational model. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 227–240. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_17
17. Wang, Y., Tran, D., Liao, Z.C.: Learning hierarchical poselets for human parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1705–1712 (2011)
18. Ramanan, D.: Learning to parse images of articulated objects. In: Neural Information Processing Systems (NIPS) (2006)
19. Alp Güler, Rı., Neverova, N., Kokkinos, I.: DensePose: dense human pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7297–7306 (2018)
20. Pishchulin, L., et al.: DeepCut: joint subset partition and labeling for multi person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4929–4937 (2016)
21. Ainsworth, B.E., et al.: 2011 compendium of physical activities: a second update of codes and MET values. Med. Sci. Sports Exerc. **43**(8), 1575–1581 (2011)
22. Fang, H.S., Xie, S.Q., Tai, Y.W., Lu, C.w.: RMPE: regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2334–2343 (2017)

23. Papandreou, G., et al.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4903–4911 (2017)
24. 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
25. Chen, Y.L., Wang, Z.C., Peng, Y.X., Zhang, Z.Q., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7103–7112 (2018)
26. Su, K., Yu, D.D., Xu, Z.Q., Geng, X., Wang, C.H.: Multi-person pose estimation with enhanced channel-wise and spatial information. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5674–5682 (2019)
27. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: DeeperCut: a deeper, stronger, and faster multi-person pose estimation model. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 34–50. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_3
28. Newell, A., Huang, Z.A., Deng, J.: Associative embedding: end-to-end learning for joint detection and grouping. In: Proceedings of the Neural Information Processing Systems (NIPS), pp. 2277–2287 (2017)
29. Papandreou, G., Zhu, T., Chen, L.-C., Gidaris, S., Tompson, J., Murphy, K.: PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 282–299. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_17
30. Kreiss, S., Bertoni, L., Alahi, A.: PifPaf: composite fields for human pose estimation. arXiv preprint arXiv:1903.06593 (2019)
31. Moon, G., Chang, J.Y., Lee, K.M.: PoseFix: model-agnostic general human pose refinement networkz. arXiv preprint arXiv:1812.03595 (2018)
32. Raaj, Y., Idrees, H., Hidalgo, G., Sheikh, Y.: Efficient online multi-person 2D pose tracking with recurrent spatio-temporal affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4620–4628 (2019)
33. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3D human pose regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3425–3435 (2019)
34. Zhang, F., Zhu X.T., Ye, M.: Fast human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3517–3526 (2019)