Refining OpenPose with a new sports dataset for robust 2D pose estimation

Takumi Kitamura
Kyushu University, Fukuoka, Japan
kitamura.takumi.583@s.kyushu-u.ac.jp

Hitoshi Teshima
Kyushu University, Fukuoka, Japan
teshima.hitoshi.058@s.kyushu-u.ac.jp

Diego Thomas
Kyushu University, Fukuoka, Japan
thomas@ait.kyushu-u.ac.jp

Hiroshi Kawasaki
Kyushu University, Fukuoka, Japan
kawasaki@ait.kyushu-u.ac.jp

Abstract

3D marker-less motion capture can be achieved by triangulating estimated multi-views 2D poses. However, when the 2D pose estimation fails, the 3D motion capture also fails. This is particularly challenging for sports performance of athletes, which have extreme poses. In extreme poses (like having the head down) state-of-the-art 2D pose estimator such as OpenPose do not work at all. In this paper, we propose a new method to improve the training of 2D pose estimators for extreme poses by leveraging a new sports dataset and our proposed data augmentation strategy. Our results show significant improvements over previous methods for 2D pose estimation of athletes performing acrobatic moves, while keeping state-of-the-art performance on standard datasets.

1. Introduction

In recent years, there has been an increase in the use of cutting-edge technology in the field of sports. For example, high speed cameras and magnetic field sensors are used for goal scoring in soccer, drones are used to track players, and AI is used to analyze tactics in sports. Various sensors and devices are also used to measure and analyze the movements and strength of athletes to optimize their movements and manage their physical conditions. In particular, the estimated posture and 3D skeleton of the human body is a powerful data to analyze the motion and performance of athletes.

Existing methods for 3D motion capture of athletes can be classified into those that use motion sensors and those that use visual information only. The methods that use motion sensors can detect the accurate movement of the human skeleton by attaching markers to the human body. However, dedicated equipment such as motion sensors are expensive and the environment for measurement is limited.

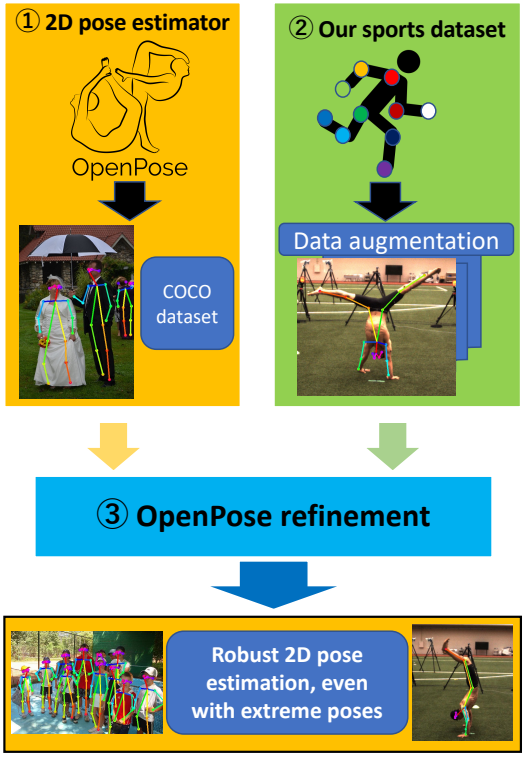
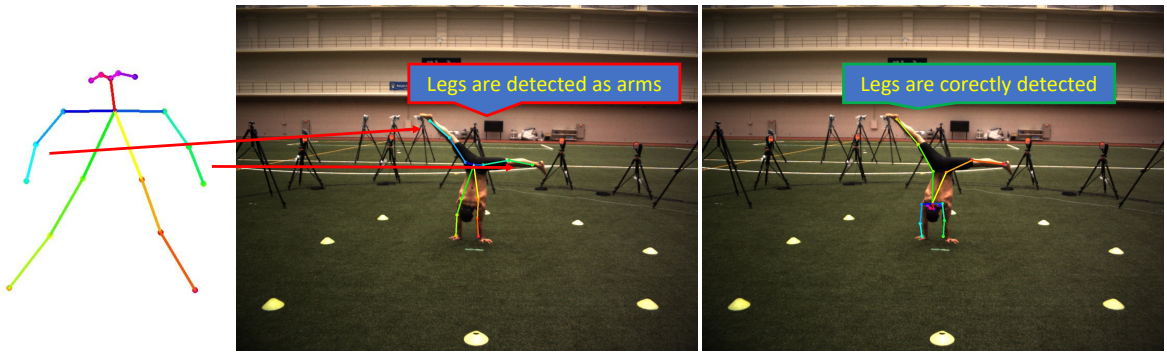


Figure 1. We refine OpenPose with our newly introduce sports dataset and a augmentation technique to improve the quality of 2D pose estimation in extreme poses.

In addition, markers must be attached on the human body, which requires expert knowledge and may not be comfortable for the athlete. As a consequence, non-invasive techniques such as 3D pose estimation from multi-view RGB images are preferred ([1, 17, 20, 28, 29],[13]). In general, existing methods first estimate the 2D pose of the person in each 2D image and then triangulate the 2D skeletons to create a 3D skeleton. By doing so, a marker-less, low-cost system that can be used anywhere can be built. In such sys-



a) OpenPose skeleton b) OpenPose failure c) OpenPose success (after our refinement)

Figure 2. From left to right: (a) format of the skeleton used in OpenPose; (b) example of a typical failure case of the original OpenPose network in sports video; (c) example of the success of our refined OpenPose network when trained with our new dataset and augmentation technique.

tems, accurate and robust 2D pose estimation is critical.

OpenPose [4] is the most popular method to obtain a 2D skeleton from a single color image. OpenPose (and follow up works [18, 27, 22, 7, 30]) consists of a deep convolutional neural network (CNN) that is trained on a large human database annotated with 2D skeletons (called COCO dataset [15]). Remarkably, OpenPose shows high accuracy even for images that contain multiple people. However, in the field of sports, there are many poses that are not seen in normal life, such as having the head down, and OpenPose fails in pose estimation using such images as input. This is because such extreme poses are few (or even not present at all) in the COCO dataset used for training OpenPose. In this paper, we introduce a new labeled sports dataset that contains many images with extreme poses that allows us to refine the OpenPose network to predict accurate pose estimates even in extreme cases.

In addition, we propose two solutions to avoid overfitting when retraining the network proposed in OpenPose. The first solution uses data augmentation while the second solution uses data pre-processing. This allows us to improve the accuracy of 2D pose estimation for complex poses common in sports, while keeping state-of-the-art accuracy with standard data (as seen in COCO dataset).

Figure 1 illustrates an overview of our proposed method. The contributions of our work are three-fold. (1) We refine OpenPose to handle extreme human poses in sports performance and performed comparative evaluation. (2) We introduce a new labelled sport dataset with annotations generated using our own annotation tool. Our new sports dataset contains many complex postures that are rare in normal datasets. (3) We propose and evaluate two different techniques to prevent over-fitting, which allows drastic improvement in sports scenarios while maintaining state-of-the-art performance on standard datasets.

2. Related Work

2.1. 3d Pose Estimation

In recent years, there has been a lot of research done on 3D pose estimation[17],[20],[28],[29],[13]. To supervise the training of deep networks, large datasets with 3D joint coordinates annotation are required[21],[8],[16],[32]. However, annotations are in general obtained with using motion capture systems, which must be manipulated by experts and strongly limits the amount of 3D training data available. As a consequence performance of direct 3D pose estimation methods are not on par with 2D pose estimation methods. In [23], a convolutional neural network (CNN) is used to improve 3D pose estimation by learning latent representations using an automatic encoder and considering the structure of the skeleton.

In [19], the authors propose to discretize the 3D space around the human body and build a 3D heat map that represent the occupancy probability of each joint in each voxel. Then a 3D CNN can be trained using the generated 3F heat maps to predict the body pose. In [31], the authors propose to simultaneously train 2D pose estimation and depth estimation networks. The predicted depth at each joint location allows to generate the 3D skeleton without relying on a 3D training dataset. In [5], a set of 3D pose data and corresponding 2D poses from multiple directions are used for training. At inference time, a 2D pose is lifted to its corresponding 3D pose by matching the 2D estimation results to a 3D pose library. This allows to generate 3D pose estimates even in occluded areas. In [9], a 3D skeleton is generated by triangulation from multiple 2D images, and two methods: Algebraic Triangulation and Volumetric Triangulation, are used to generate the 3D pose, which greatly exceeds the performance of existing methods. However, all these methods still require an accurate and robust 2D pose estimation. In cases when 2D pose estimation fails (like in

Figure 3. We have developed our own manual annotation tool to create data for 2D poses to be used as ground truth.

sports videos) these methods do not work at all.

2.2. 2d Pose Estimation

2D pose estimation from a single image or videos has been extensively researched in the past decade [18],[27],[22],[7],[30]. And there are many datasets of 2d human poses that are used for training[12],[6],[14],[10]. Toshev et. al. [25] was the first method that uses deep learning for 2D human pose estimation. This is a top-down method in which the person is detected by a detector and the pose is estimated independently for each person. This method demonstrated that deep learning is effective for pose estimation and that cascading is a successful strategy. In [26], which is also a top-down method, a confidence map is created for each part in the first stage, and in subsequent stages, the confidence map is refined by expanding the acceptance field of the relevance while learning the relevance between parts by referring to the confidence map of the previous stage. OpenPose [4] uses a bottom-up method for extracting key points in the image and then matching them, which improves the accuracy of pose estimation for multiple persons. In the bottom-up method, the amount of computation does not increase with the increase in the number of users, but there is a problem that matching takes a long time. Therefore, OpenPose solved the problem of matching by adopting inter-joint vectors called PAF (Part Affinity Field), and succeeded in performing high-performance posture estimation in real time. However, OpenPose fails in extreme cases such as sports videos.

Some sports datasets have been made publicly available, such as the Leeds Sport Pose [11] or the 2D SkiPose [3], but they only contain single views of the person and cannot be used for 3D pose learning. Our proposed dataset contains multi-view images, which makes it promising for improving performance of 3D pose estimators.

3. 2D pose estimation in sports video

2D pose estimation in-the-wild from images or videos of single or multiple people has been extensively studied and robust solutions such as OpenPose exist. However, efficient 2D pose estimation in sports video remains difficult because of extreme poses of the athletes that are not represented in the available annotated 2D human pose datasets. We demonstrate the limitations of the original OpenPose in sports scenarios and present a new annotated dataset together with an efficient data augmentation technique to refine the network.

3.1. Limitations of OpenPose

OpenPose [4] can detect the poses of multiple people in real time with high accuracy from a single image (figure 1). However, when a person is in an extreme pose such as inverted position (head down), OpenPose fails to estimate the correct pose (figure 2 b)). For example, the hands and feet are estimated reversely. This is because the COCO dataset [15] used for training the network does not contain enough unusual poses such as people being upside down.

OpenPose is a method that learns a confidence map called PAF (Part Affinity Field) that represents the label of a body part and a vector between the different body parts. In this work, we propose to refine OpenPose so that the network can accurately predict difficult poses such as upside down poses. Figure 2 b) shows an example of OpenPose failure and figure 2 c) shows success after applying our refinement. As shown in figure 2 a), the red skeleton represents the left foot, light green represents the right foot, dark green represents the left hand, and blue represents the right hand, but in figure 2 b), the colors of the hands and feet are

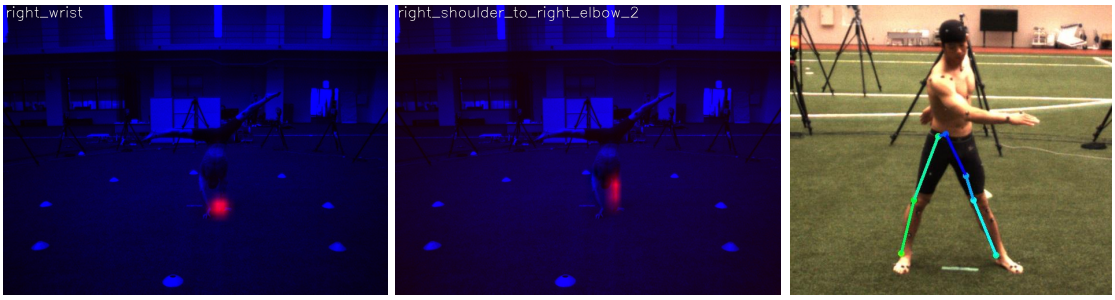


Figure 4. From left to right: (a) example of a heatmap for the right wrist; (b) example of a PAF (Part Affinity Field) for the bone (right shoulder, left elbow); (c) example of a failure due to over-fitting.

switched, indicating that the pose estimation failed.

3.2. Our annotated sports dataset

The original OpenPose network is trained is the publicly available COCO dataset [15]. The COCO dataset consists of a training dataset that contains 64115 images (including human bodies with annotations) and a validation dataset that contains 2693 images with annotations. In this dataset (and other publicly available human datasets such as [2]), complex poses such as people with the head upside down and their annotations are rare. In this paper we introduce a new human pose dataset specifically tailored to sports scenario that contains many images of extreme poses common in sports videos with precise ground truth annotations obtained manually. We call this dataset the Kanoya dataset.

Our Kanoya dataset consists of multiple series of videos of gymnasts performing acrobatic movements (see the supplemental material for some examples). Each video consists of 25 seconds of the athletes performing multiple movements such as backflips. These videos were shot simultaneously from multiple directions (all cameras were calibrated and synchronized). As a consequence the estimated 2D poses can also be easily triangulated with the known camera parameters to generate ground truth 3D poses (although this is out of the scope of this paper). The videos from the front and the back were divided into 750 frames each (1500 frames in total), and 18 key points were annotated for each frame using our own annotation tool (figure 3). Since OpenPose requires only 17 annotations (excluding the neck), we converted the annotations after correcting them in the GUI. Of the 1500 annotations modified, 1200 were used for training and 300 for validation.

Pose annotation tool OpenPose often fails to estimate extreme poses such as backflips and joints with occlusions. Therefore, we have developed our own annotation tool to correct such OpenPose estimation errors (figure 3). All that is required to use our tool is to select the original images and the JSON files that is the initial output of OpenPose. We developed the tool in Python and so that we can easily modify the position of any wrongly estimated joint by dragging it with the mouse. As other specific functions, we can copy and paste a pose from other frame and undo. As a consequence, our annotation tool allows us to quickly and intuitively fix OpenPose errors and generate high quality ground truth annotations. We will make our annotation tool publicly available upon acceptance

3.3. OpenPose Refinement on sports dataset

As shown in figure 1, we use the pre-trained weights of OpenPose at initialisation and refine the weights in the network using the Sports sports dataset introduced in the previous section. We optimize the network by adjusting the hyper-parameters based on the output loss information, heatmap (figure 4 a)), PAF (Part Affinity Field) (figure 4 b)) and inference results.

3.4. Data augmentation method

By re-training OpenPose using the strategy described above, pose estimation accuracy can be drastically improved for the case of sports videos (head upside down etc...). However, the drawback of this approach is that the network over-fits the sports dataset and performance significantly degrades for standard poses (as shown in figure 4 c) for example). To prevent the network to over-fit the Sports dataset and maintain state-of-the-art performance on standard datasets, we propose to re-train OpenPose by combining both COCO and Sports datasets.

The straightforward strategy would be to simply mix the COCO dataset and the Sports dataset and re-train OpenPose on this combined dataset (we call this approach ”Mix”). However, while the COCO dataset contains about 65000 images, the Sports dataset contains only 1500 images. The number of raw data in our sports dataset is insufficient to balance the number of image in the original dataset and therefore the ”Mix” approach does not allow to learn extreme poses.

To balance the two datasets, the solution is to perform

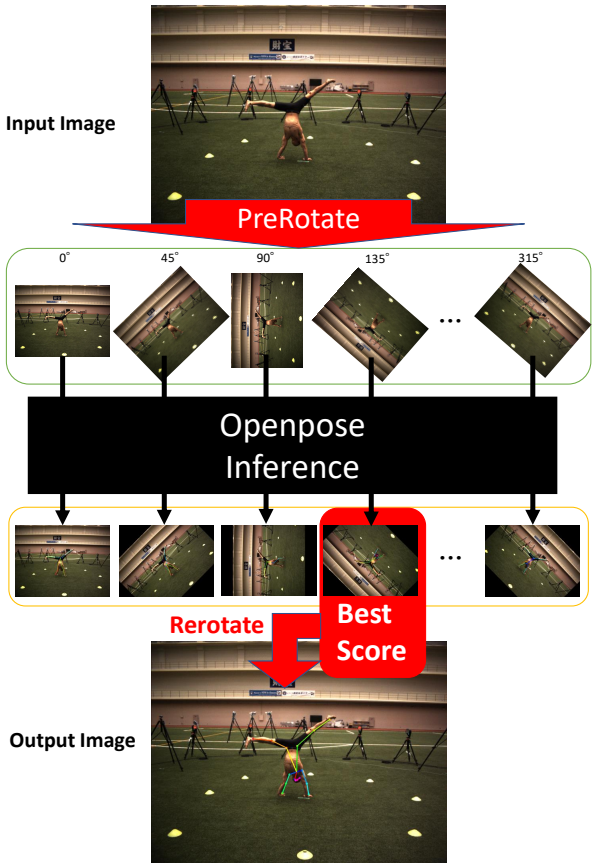


Figure 5. Illustration of our proposed pre-rotation method

data augmentation. This means to generate synthetic annotated images of sports video from the captured data. This can be done by rotating the input images with various angles. With data augmentation, we increase the number of training data for sports scenarios, which allows us to re-train OpenPose with a balanced dataset to achieve accurate and robust pose estimation in various situations.

3.5. Pre-rotation method

We also propose an alternative approach that does not require re-training of OpenPose or data augmentation. The key idea here is to pre-process all input images before feeding them to OpenPose so that the input images fit the distribution of the COCO dataset (i.e., the head is positioned at the top of the image and the feet are positioned at the bottom of the image). By doing so the original pre-trained OpenPose network can be directly use as it is. This also allows us to clearly evaluate the advantages of building new datasets tailored for sports activity and data augmentation technique.

Each input image is rotated several times with different angles and fed to the (original) OpenPose network. The difficulty here is that we do not know where are the feet and the head, so we do not know the best rotation to correct the image. Instead, we propose to try several rotations and retain the best one according to the output confidence of OpenPose. Namely, we rotate the images in eight directions as input for OpenPose's inference, and obtain eight output results. We use the sum of the confidence scores of the key points (which is one of the outputs of OpenPose) and select the most appropriate one among the 8 as the final output. Figure 5 illustrates the pipeline for our pre-rotation method.

4. Experiments

We perform both quantitative and qualitative comparative experiments to evaluate the advantages of our proposed method and the benefits of using our newly introduced sports dataset.

4.1. Validation method

We used two test datasets for validation: the validation set of COCO dataset and a subset of our newly introduced sports dataset. First, we use the COCO dataset validation dataset, which was used to validate the conventional training of OpenPose, and by using this dataset for validation we confirm whether the accuracy of the original OpenPose is maintained without over-fitting to extreme poses (such as head upside down) after re-training. Second, we use our proposed sports test data set, which consists of 81 images and annotations extracted from a different video than the one used for training. By using this dataset, we evaluate to what extent OpenPose can estimate complex postures such as those taken by athletes in acrobatic figures.

The AP score provided in the COCO dataset measures the truth of Oks, which is a value indicating the degree of association between the estimated value of the key point and the true value. To compute the AP we use multiple thresholds, and the integral of Precision and Recall with respect to the ground truth. The AP score is the integral value of Precision and Recall. For each threshold value, comparisons are generally made using three indices: AP50 is the AP score for a threshold value of 0.5, AP75 is the AP score for a threshold value of 0.75, and AP is the average of AP50, AP55, AP60, ..., AP95. The AP scores were computed for each test dataset and compared to each other to evaluate the results. In the computation of Oks, the area of the human part is required, but the COCO validation data set also holds the annotation information of the area, while the sports data set does not. Therefore, we computed the Oks by using the value of half of the Bounding box as the person instead. For this reason, we can not directly compare the AP scores between the two test data sets.

4.2. Results

We compared six methods: the original OpenPose, OpenPose re-trained using the sports dataset, OpenPose re-

Table 1. Comparative quantitative results of 2D pose estimation when using different refinement strategies.

	COCO validation dataset			Sports test dataset		
	AP	AP50	AP75	AP	AP50	AP75
Original [24]	0.457	0.712	0.475	0.194	0.427	0.081
1. Retrain	0.245	0.478	0.225	0.461	0.828	0.446
2. 1 + Augmentation	0.143	0.304	0.116	**0.521**	**0.903**	**0.545**
3. 1 + Mix	0.378	0.645	0.37	0.29	0.725	0.165
4. 2 + 3	**0.413**	**0.675**	**0.419**	0.483	0.892	0.412
Pre-rotate	0.374	0.639	0.37	0.221	0.627	0.061

trained with using the augmented Sports dataset (with using our proposed method for data augmentation), OpenPose re-trained with mixing the Sports dataset with the COCO dataset (Mix), OpenPose re-trained with mixing COCO dataset and the augmented Sports dataset, and the pre-rotation method. Table 1 shows the results of AP, AP50, and AP75 for the two test data sets (sports dataset and COCO dataset).

Because the two datasets differ in the size of the data and the method of estimating the area of the person area necessary for computing the AP score, it is difficult to compare the values between the two datasets. Nevertheless, from the results, it can be seen that by re-training using the sports data set, the value for the sports test data set improved significantly from 0.194 to 0.461, indicating that the estimation accuracy for extreme poses has improved. On the other hand, the result for the COCO test data set dropped from 0.475 to 0.225, which means that there is over-fitting.

We proposed two methods to counter the problem of over-fitting: data augmentation and pre-rotation. From our results we observed that the accuracy for extreme poses was the best when using the data augmentation strategy. Moreover, the Mix method had the effect of maintaining the accuracy of the original OpenPose on conventional dataset (COCO). The combination of the two datasets with the data augmentation method produced results with accuracy of 0.483 for the sports test dataset and 0.413 for the COCO test dataset. The accuracy of the pre-rotation method was 0.221 for the sports test dataset and 0.374 for the COCO test dataset. The accuracy of the pre-rotation method was inferior to that of the method using re-training. The reason for this is that simply rotating the images so that the upper body is positioned at the top of the image still results in extreme poses, such as legs open with 180 and feet not touching the ground. Such poses are not frequent in the COCO dataset and 2D pose estimation remains difficult. In addition, the time required for 2D pose estimation increases by a factor of 8. As a consequence, the method using re-training is more effective than the pre-rotation method when considering real-time applications.

Figure 6 shows the changes in 2D pose estimation accuracy with respect to the number of images from the COCO

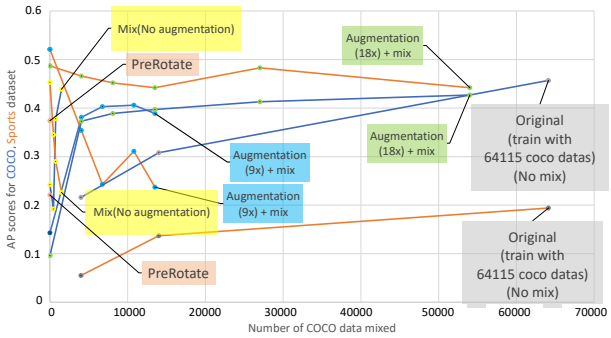


Figure 6. 2D pose estimation accuracy with respect to the number of data from the COCO dataset using in mixing with our sports dataset.

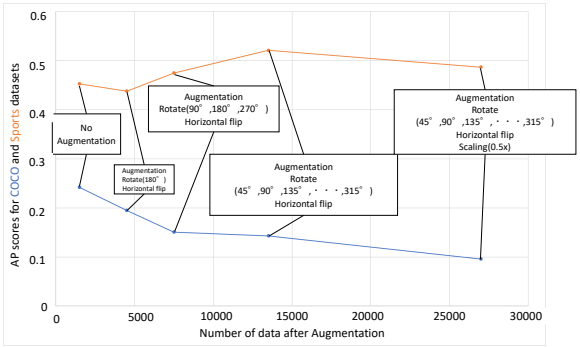


Figure 7. 2D pose estimation accuracy with respect to different number of rotation angles used for data augmentation.

dataset used when mixing the COCO dataset with the sports dataset. Experiments were conducted under various conditions: with pre-rotation method and with different data augmentation. The horizontal axis represents the number of images from the COCO dataset that are mixed and the vertical axis representing the AP score. The orange line shows the results obtained on the sports test dataset, and the blue line shows the results obtained on the COCO test dataset. From the graph, we can see that the results obtained on the COCO test data set improved as the number of images from the COCO dataset increased. We also found that combining mixing with data augmentation improves results obtained on both the COCO and sports datasets.

As shown in figure 7, data augmentation increases the

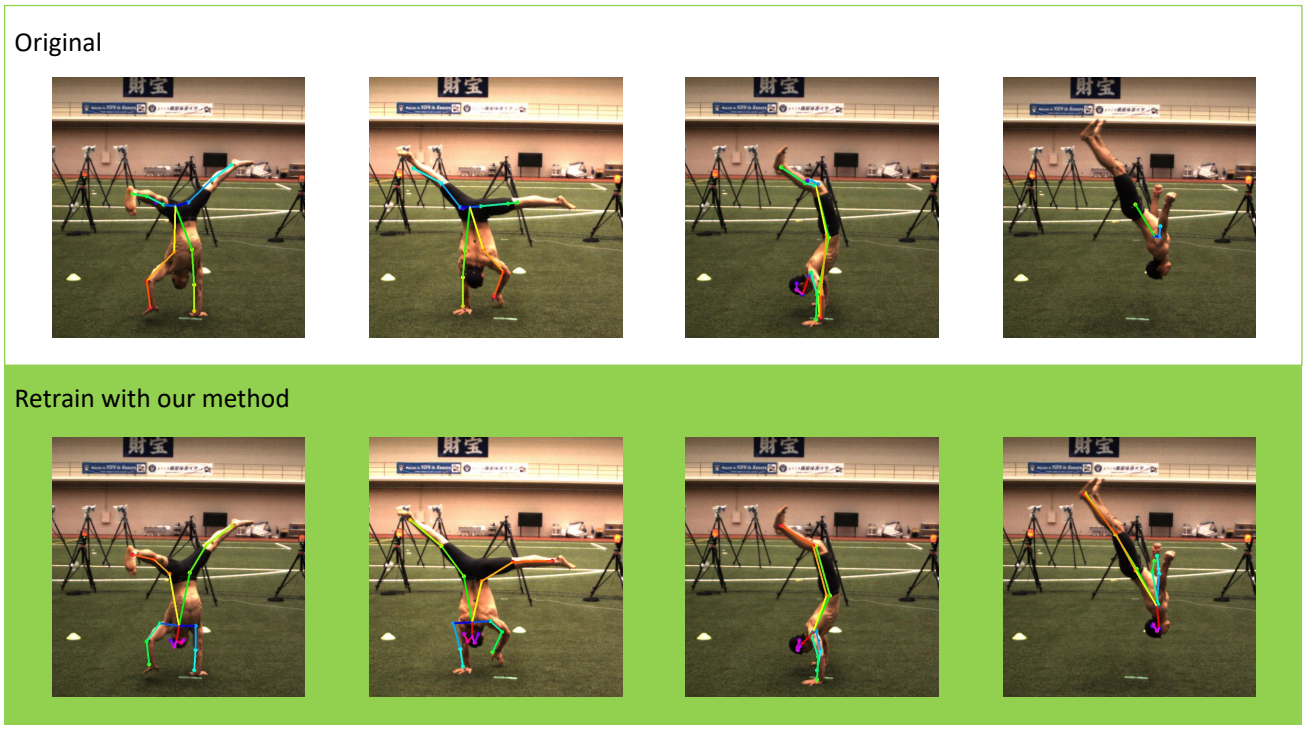


Figure 8. Qualitative comparison between results obtained by original OpenPose and our method on the sports test dataset

variation of data used for re-training and significantly improves the estimation accuracy for extreme poses such as those seen in sports videos. On the other hand, the accuracy for 2D pose estimation on the COCO test dataset tends to decrease. This is because in our sports dataset introduced in this paper, there is only one person in the images, while in the COCO test dataset, there are multiple people in the image. One of our future work will be to expand our sports dataset with multiple athletes in the videos.

Figures 8 and 9 show qualitative comparison between results obtained with the original OpenPose method and our proposed method (data augmentation + Mix) with images sampled from both our sports dataset and the COCO dataset. As shown in figure 8, the original OpenPose method failed in extreme poses such as head upside down, but our proposed method was able to output accurate 2D poses estimates on these images. Note that the color of the skeleton output by OpenPose is wrong (see figure 2 a)): the legs are detected as arms. In addition, as shown in figure 9, we can see that our Mix + augmentation method effectiely prevent the network to over-fit to the sports dataset and maintain the accuracy of the original OpenPose on standard poses.

Figure 10 shows comparative results on images on-the-wild sampled from videos taken from YouTube. These results show that our proposed method is robust and can be used efficiently on real scenarios, which is promising for various applications in sports.

5. Conclusion

In this paper, we proposed a method for estimating 2D poses from a single image, even in extreme cases such as athletes performing acrobatic moves. We proposed to refine the original OpenPose network with our newly introduced sports dataset and an efficient data augmentation and mixing strategy. Our proposed method allowed to drastically improve the accuracy and robustness of 2D pose estimation in extreme cases, while maintaining state-of-the-art performance on standard poses. We demonstrated that our proposed method can be used to estimate complex 2D poses from sports videos in-the-wild. This could also be applied for winter sports like freestyle skiing where the athletes often have extreme posture. Our method does not require any markers or special equipment, which is promising to make motion data collection and analysis of athletic activities more accessible.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP20H00611, JP18H04119 in Japan. The dataset used in the paper is part captured by Sports Performance Research Center at National Institute of Fitness and Sports in KANOYA with support of Professor Tomohito Wada and Yoshie Motoshima in 2019.

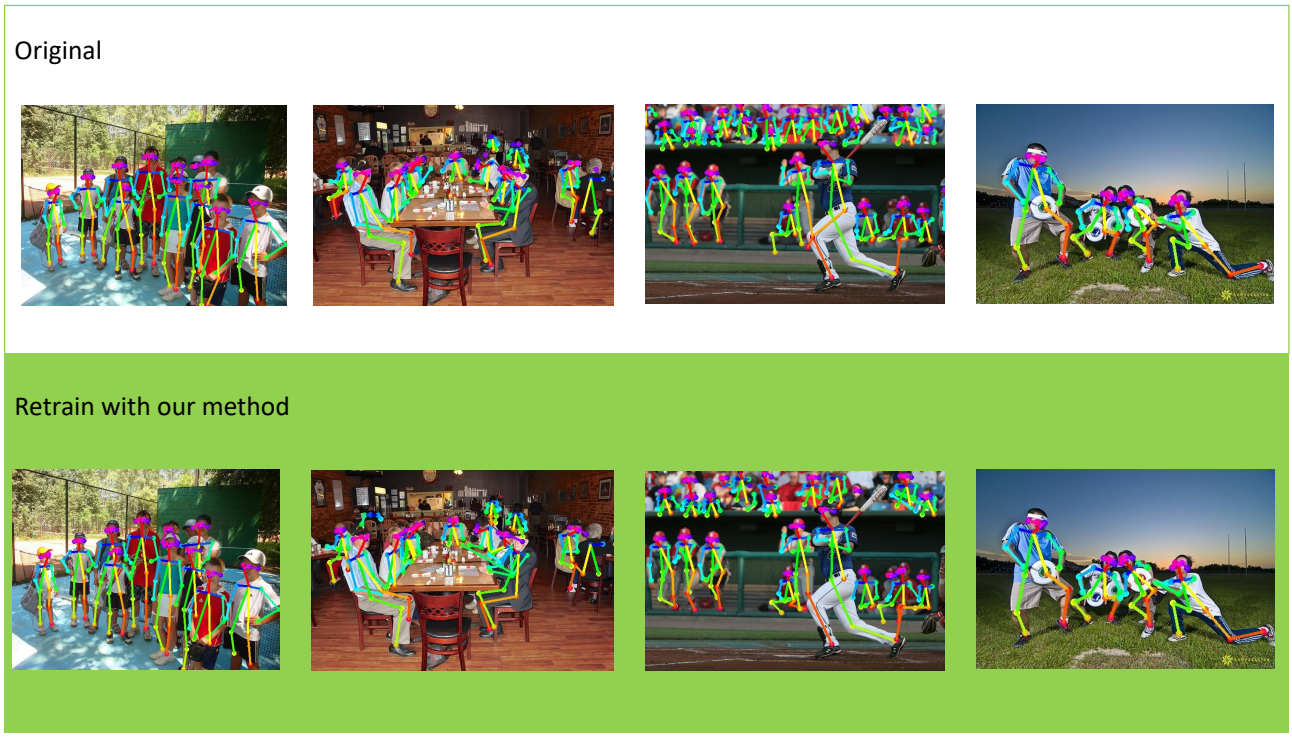


Figure 9. Qualitative comparison between results obtained by original OpenPose and our method on the COCO validation dataset

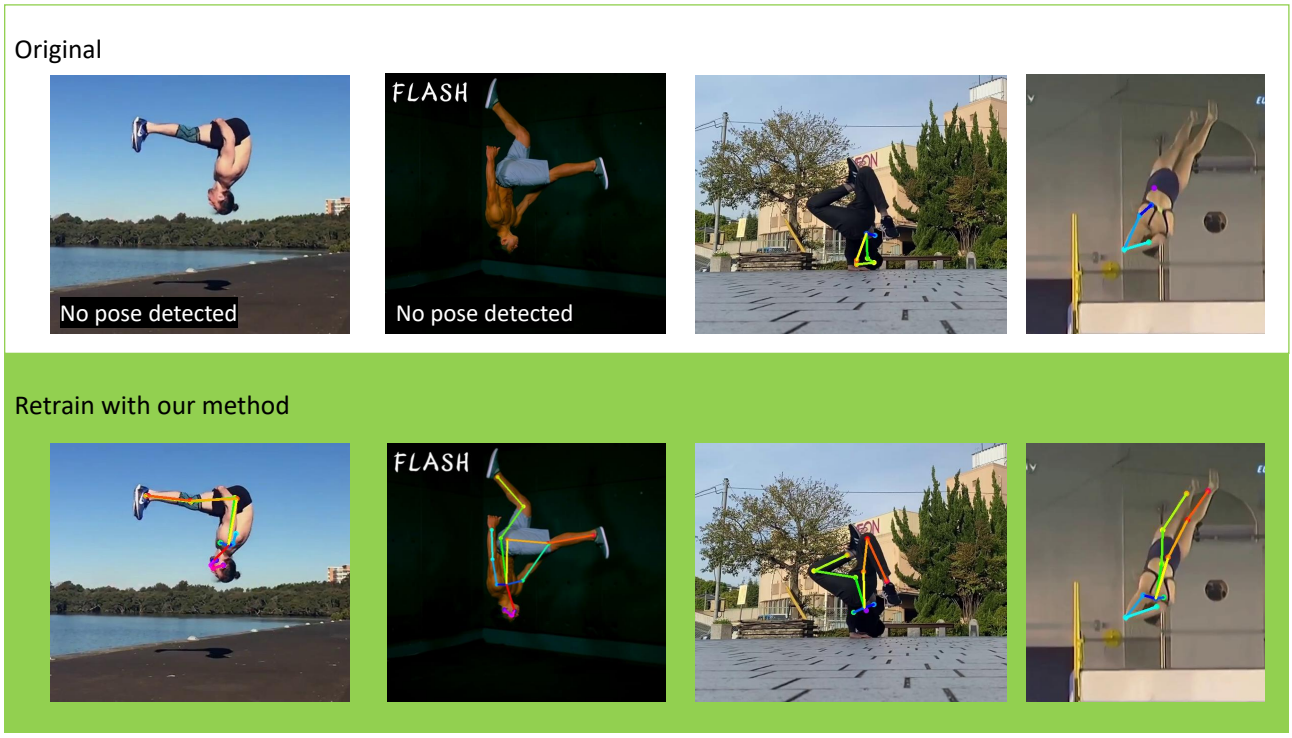


Figure 10. Qualitative comparison between results obtained by original OpenPose and our method on on-the-wild images taken from YouTube

References

[1] Sayo Akihiko. Synthesis of Pose-dependent Entire Human Shape using a Single RGB-D Camera and 3D Human Pose Refined by DeepNet .

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[3] Roman Bachmann, Jörg Spörri, Pascal Fua, and Helge Rhodin. Motion capture from pan-tilt cameras with unknown orientation. In *2019 International Conference on 3D Vision (3DV)*, pages 308–317. IEEE, 2019.

[4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019.

[5] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7035–7043, 2017.

[6] Haodong Duan, KwanYee Lin, Sheng Jin, Wentao Liu, Chen Qian, and Wanli Ouyang. Trb: A novel triplet representation for understanding 2d human body, 2019.

[7] Gines Hidalgo, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Single-network whole-body pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[8] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[9] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose.

[10] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, pages 196–214. Springer, 2020.

[11] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.

[12] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10863–10872, 2019.

[13] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6173–6183, 2020.

[14] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[16] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017.

[17] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[18] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *arXiv preprint arXiv:1611.05424*, 2016.

[19] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *CoRR*, abs/1611.07828, 2016.

[20] Saurabh Sharma, Pavan Teja Varigonda, Prashast Bindal, Abhishek Sharma, and Arjun Jain. Monocular 3d human pose estimation by generation and ordinal ranking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[21] Leonid Sigal and Michael J Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity TR*, 120(2), 2006.

[22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[23] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. *arXiv preprint arXiv:1605.05180*, 2016.

[24] tensorboy. pytorch-Realtime-Multi-Person-Pose-Estimation. `https://github.com/tensorboy/pytorch_Realtime_Multi-Person_Pose_Estimation`.

[25] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. *CoRR*, abs/1312.4659, 2013.

[26] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. *CoRR*, abs/1602.00134, 2016.

[27] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[28] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[29] Zhenbo Xu, Wei Zhang, Xiaoqing Ye, Xiao Tan, Wei Yang, Shilei Wen, Errui Ding, Ajin Meng, and Liusheng Huang. Zoomnet: Part-aware adaptive zooming neural network for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12557–12564, 2020.

[30] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[31] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[32] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019.