Supplementary Material for “Action Assessment by Joint Relation Graphs”

Jia-Hui Pan[1], Jibin Gao[1], Wei-Shi Zheng[1,2,3,*]
[1]School of Data and Computer Science, Sun Yat-sen University, China
[2]Peng Cheng Laboratory, Shenzhen 518005, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
panjh7@mail2.sysu.edu.cn, gaojb5@mail2.sysu.edu.cn, wszheng@ieee.org

Abstract

We provide more implementation details including the local patch cropping and video feature extraction. We also provide further evaluations of our method on different numbers of video segments.

1. Implementation Details

Local Patch Cropping. We extract human poses and bounding boxes using Mask-RCNN [2]. In each frame, we take the person with the largest bounding box and remove the rest, preventing the interference of audience, coaches and referees. Linear interpolation is applied on the temporal dimension to complement the pose estimation results on those frames where the athletes are not detected. In each frame, we crop a local patch around each joint with a square bounding box centered on the joint position. The side length of the square bounding box is 1/10 the average of the width and height of the human bounding box. All patches on the temporal dimension of a certain joint form a *local patch video* which records the detailed joint movement.

Video Feature Extraction. We extract video features with I3D, pre-trained on Kinetics[1]. The features of RGB and optical flow are added together as the video features (i.e. the joint features and the whole-scene features) in our experiments. For each video, the optical flow data is obtained by the TV-L1 algorithm [3]. All images including the whole frames and the local patches are resized to 224 × 224 before feeding into the I3D network. The whole-scene feature q^t is obtained with the whole-scene video, while the joint feature matrix F^t is obtained with the *local patch videos*. We divide the all videos into 10 segments, and 16 frames are uniformly sampled out in each segment as the input to the I3D network. A whole-scene video is divided according to its starting and ending time-stamps, while a *local patch video* is divided according to the starting and ending of the human detection. To prevent over-fitting, we perform data augmentation on the videos by left-right flipping.

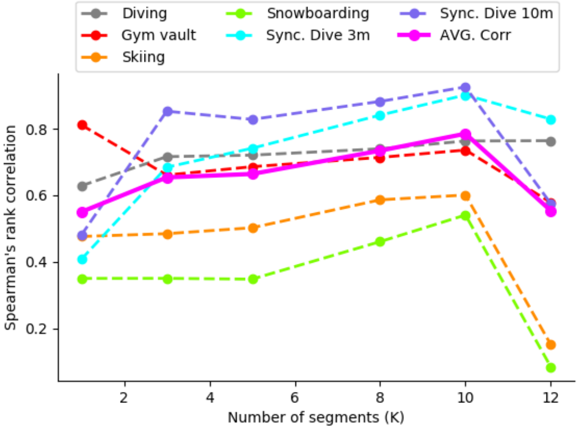

Figure S1. The performance of our method on different numbers of video segments

2. Further Evaluations

Number of Video Segments. Fine-grained temporal division of the action videos helps to achieve better results of our method. We test our method on different numbers of video segments, and the results are shown in Figure S1. On average, the performance of our method improves as the the number of segments increases from 1 to 10, although the performance growth tends to get slower. The performance gain is not surprising because dividing the videos into more segments enables finer-grained action analysis. The fine-grained action assessment models exploit more effective short-term motion patterns. However, the performance of our method drops when the number of segment exceeds 10. Therefore, we choose to divide the videos into 10 segments for experiments in our manuscript.

[*] Corresponding author

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[3] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. Tv-l1 optical flow estimation. *Image Processing On Line*, pages 137–150, 2013.