Supplementary Material for "An Asymmetric Modeling for Action Assessment"

Jibin Gao[1,4], Wei-Shi Zheng[1,2,5*], Jia-Hui Pan[1], Chengying Gao[1*], Yaowei Wang[2], Wei Zeng[3], and Jianhuang Lai[1]

[1] School of Data and Computer Science, Sun Yat-sen University, China
[2] Peng Cheng Laboratory, Shenzhen 518005, China
[3] School of Electronics Engineering and Computer Science, Peking University, China
[4] Pazhou Lab
[5] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, China
{gaojb5,panjh7}@mail2.sysu.edu.cn;{zhwshi,mcsgcy,stsljh}@mail.sysu.edu.cn
wangyw@pcl.ac.cn;weizeng@pku.edu.cn

Abstract. We provide some video demos for the visualization of assessment process in our model. We also give a more detailed explanation for our attention fusion module, and details of dataset TASD-2 construction and data preprocessing in Experiments.

1 Video demos for the visualization of assessment process

We provide video demos for the visualization of assessment process in our model, which is presented as Fig. 6 in our paper as well.

2 The attention fusion module

Our attention formulation absorbs the merit of the attention function in [7]. It is defined as

$$\mathcal{A}(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{S1}$$

where Q, K and V represent the queries, keys and values, respectively, and d_k is a scaling factor of K. In computer vision, we generally learn the $\frac{K}{\sqrt{d_k}}$ (denoted as K_d) by a FC layer with a informative feature z, computed by

$$K_d = \mathcal{FC}_{key}(z). \tag{S2}$$

K_d corresponds to O_{key} in Eq. (5). Inspired by self-attention, we regard $(X_{wl}^{(t)} \oplus Y_{msi}^{(t)})$ as the queries and values, and thus Eq. (5) corresponds to $softmax(\frac{QK^T}{\sqrt{d_k}})$, as well as Eq. (4) corresponding to attention function. We used the whole-scene feature F_{wl} for the key of attention because it contains the whole-scene context. If replacing F_{wl} with X_{wl}, a little drop of performance will be found due to information loss, as shown in Table S1.

* corresponding authors

Table S1. Study on the choice of the features for the key of attention(%).

	Suturing	Needle Passing	Knot Tying	Avg. Corr.
Ours	63	**65**	**82**	**71**
Replace F_{wl} with X_{wl} for the key	**64**	63	81	70

3 Dataset TASD-2 construction details

We collected more than 600 samples from twenty valid and complete video-recordings of entire synchronized diving events on YouTube, including four in the Olympic Games, three in FINA, nine in the European diving competitions and four in the Southeast Asian Games, which could be categorized as synchronized 3-m springboard (SyncDiving-3m) and synchronized 10-m platform (SyncDiving-10m). To determine whether each diving video was taken from the front view, we watched almost the entire video and recorded the starting frame and ending frame to split out a sample video. Additionally, for a specific sample, some labels should be recorded for further study, especially the final scores for action quality assessment. The details of the dataset are given in Table S2. Note that "execution score_v2" is determined through calculating "difficulty score" multiplied by "execution score", since referees only give the "execution score" with value ranging 0 to 10, regardless of the difficulty of the action. Hence, in individual analysis for execution of synchronized diving, we prefer to use "execution score_v2" rather than "execution score" directly. The length of each video was uniformly modified to 102 frames with the format of 320×240 for each frame referring to *AQA-7* [5]. We have augmented the videos by left-right flipping and split them into a training set and a testing set with a ratio of 4:1, respectively, in a random fashion.

4 Data preprocessing in Experiments

On *JIGSAWS* [3], a dataset containing egocentric surgical videos, the primary and secondary information are fetched from the 3D kinetics feature in that

Table S2. Details of the *TASD-2 dataset*

Sport	**SyncDiving-3m**	**SyncDiving-10m**
#Frames of a sample	102	102
#Samples	119	184
#Augmented samples	238	368
#Training set	188	293
#Testing set	50	75
Sample's attributes	videos ID:	"v_id"
	difficulty score:	"diff_score"
	execution score:	"exec_score"
	synchronization score:	"sync_score"
	final score:	"final_score"
	execution score_v2:	"exec_score_diff"

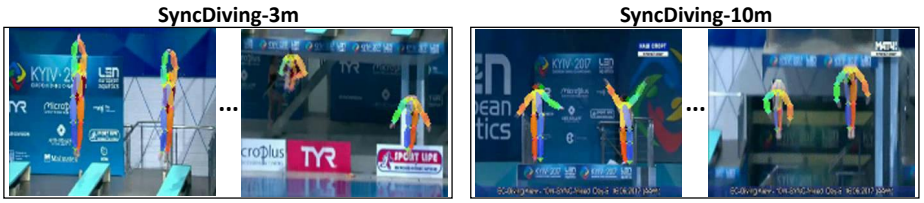

Fig. S1. Detection results by using Alphapose [2] on *TASD-2*. The colourful lines represent the skeleton of players.

dataset. To map different observed variables into a common space, DCT is operated on the 3D kinetics feature to obtain 50-dimensional expanding A_a, where it is intuitive to regard the master tool manipulators as the *primary* in the asymmetric interaction module and the patient-side manipulators as the *secondary*. On sport action assessment tasks, we extract human poses (i.e. the coordinate of each key-point of the poses) as A_a through AlphaPose [2], with denoising and linear interpolation for completion. Fig. S1 shows the example of detection results of applying AlphaPose [2] to our *TASD-2*. We extract the whole-scene feature via I3D pretrained on Kinetics [1], with RGB and optical flow [6] feature input. Referring to previous works [4], we uniformly divide every video into 10 segments, corresponding to 10 time steps. For each segment, 16 frames in sports videos are uniformly sampled as the input of I3D, while in egocentric surgical videos, 64 frames are sampled due to their longer duration than the sports videos used in our experiment. Except *TASD-2*, which was augmented during dataset construction, we augment the videos by left-right flipping used in [4]. The scores in each dataset are normalized to $[0, 100]$ as the labels to supervise our model.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2334–2343 (2017)
3. Gao, Y., Vedula, S.S., Reiley, C.E., Ahmidi, N., Varadarajan, B., Lin, H.C., Tao, L., Zappella, L., Béjar, B., Yuh, D.D., et al.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In: MICCAI Workshop: M2CAI. vol. 3, p. 3 (2014)
4. Pan, J.H., Gao, J., Zheng, W.S.: Action assessment by joint relation graphs. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
5. Parmar, P., Tran Morris, B.: Action quality assessment across multiple actions. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1468–1476 (Jan 2019). https://doi.org/10.1109/WACV.2019.00161
6. Pérez, J.S., Meinhardt-Llopis, E., Facciolo, G.: Tv-l1 optical flow estimation. Image Processing On Line pp. 137–150 (2013)
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf