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. Action assessment is a task of assessing the performance of an action. It is widely applicable to many real-world scenarios such as medical treatment and sporting events. However, existing methods for action assessment are mostly limited to individual actions, especially lacking modeling of the asymmetric relations among agents (e.g., between persons and objects); and this limitation undermines their ability to assess actions containing asymmetrically interactive motion patterns, since there always exists subordination between agents in many interactive actions. In this work, we model the asymmetric interactions among agents for action assessment. In particular, we propose an asymmetric interaction module (AIM), to explicitly model asymmetric interactions between intelligent agents within an action, where we group these agents into a primary one (e.g., human) and secondary ones (e.g., objects). We perform experiments on *JIGSAWS* dataset containing surgical actions, and additionally collect a new dataset, *TASD-2*, for interactive sporting actions. The experimental results on two interactive action datasets show the effectiveness of our model, and our method achieves state-of-the-art performance. The extended experiment on *AQA-7* dataset also demonstrates the generalization capability of our framework to conventional action assessment.

1 Introduction

Action assessment [4, 13, 9, 18, 1] has attracted much attention in recent years. It is widely applicable to many practical scenarios. For instance, action assessment models can be applied in sports events to assist the referee in scoring, as well as to assist athletes in training [18, 20, 1, 16]. Athletes can make reasonable corrections to their motions according to the feedback from the assessment model to achieve better training effects. In modern medical treatment, rehabilitation

* corresponding authors

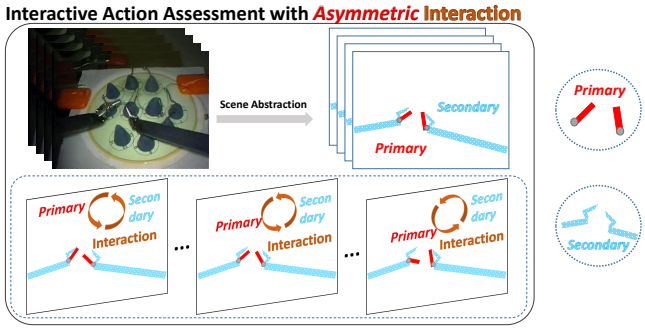

Fig. 1. Our asymmetric interaction module is designed to assess action performance. For egocentric surgical videos, we regard motions of the master tool-tips as the *primary* (in red), and those of the slave tool-tips and handles, which are relatively inactive, as the *secondary* (in blue). Best viewed in colour.

treatment has received increasing attention. Action assessment can be applied to the rehabilitation training of patients [13, 22, 28]. By monitoring and assessing the daily rehabilitation training of patients, doctors can give follow-up rehabilitation treatment suggestions to patients according to the assessment report, aiming to achieve efficient treatment [31, 32, 7, 14].

However, existing action assessment methods [20, 15, 17] are mostly designed for individual actions, such as diving and vaulting. In real-world applications, there are many non-individual actions, which are defined by interaction, especially there are subordination between agents in an interaction. For example, in the view of egocentric surgical videos, only motions of two tool-tips are captured. Accordingly, the actions involving interactions between human (featured as motion of tool-tips) and the "two tool-tips and handles" should be explored explicitly for the assessment. More importantly, such an interaction is asymmetric. The roles in asymmetrically interactive actions mentioned above can semantically be categorized into the primary agent and the secondary ones. While some work such as [15] can be applied to handle the assessment on interactive action, they treat all agents equally, and thus they cannot provide particular modeling on the subordination between agents (e.g. between human and objects).

In this work, we propose a new framework for addressing asymmetrically interactive action assessment with an asymmetric interaction module (AIM) that provides a general and novel proposal for action interaction among multiple people or parts. In this module, it is assumed that there is a primary agent that is dominant relative to the others it interacts with; correspondingly, the others are viewed as the secondary agents in interactive actions. This assumption makes sense, since the multiple parts involved in the interactive actions are always naturally semantically categorized as two parts, an important part (e.g., human) and secondary parts (e.g., objects). In AIM, we exploit the difference between the *primary* and the *secondary* in the same latent space, and utilize the

primary equipped with the difference to learn the interaction in the temporal domain, since the *primary* is dominant relative to *secondary* in representing an action. With this module, our framework can explicitly learn the latent criterion of interactive action assessment. Afterwards, we construct an attention fusion module inspired by the attention mechanism [24] to pay different amounts of attention to the whole-scene feature and AIM feature.

Moreover, apart from assessing interactive actions in strong asymmetric relation among each part, our method can also be applied to interactive actions whose agents are in weak asymmetric or equal relation, such as synchronous sports. We therefore generalize our model by a multi-task learning for operating general interactive action assessment.

To the best of our knowledge, *AQA-7* [18] and MTL-AQA [16] are the only two available datasets that contain the events involving two players; however, these events are from side view, and thus it is unsuitable to investigate the interaction between players as they overlap seriously most of the time. Therefore, we have additionally collected a new dataset, named Two-person Action Synchronized Diving dataset (*TASD-2*) for evaluating interactive action assessment.

In summary, our contributions are three-fold: (1) A novel module, called the asymmetric interaction module (AIM), is constructed to reasonably extract the interactive relation for asymmetric interaction; (2) a general framework for interactive action assessment is proposed that can be easily generalized to different kinds of action assessment tasks; (3) a new dataset, *TASD-2*, is collected in our work containing two-person interactive actions captured from the front view. We have reported experiments to validate the effectiveness of the proposed method. Project homepage: `https://www.isee-ai.cn/~gaojibin/ProjectAIM.html`.

2 Related Work

Action Quality Assessment. Action assessment is the evaluation of how well an action is performed. For the tasks of action assessment, the existing works mainly modelled the problem in three manners: 1) casting it as a classification task to classify the action performance as expert or novice [33, 32]; 2) casting it as a regression problem that fits the scores of multiple action performances [20, 26, 17, 18, 30]; 3) casting it as a pair-wise ranking task [4, 1, 5, 29]. Our work follows the second branch. However, few works have assessed the action quality by explicitly exploring the interaction in actions, and especially lack of modeling on asymmetric interaction for the assessment. To learn the relation among joints of performers' skeleton, Pan et al. [15] computed action quality based on joint relation modeling by GNN [21]. Compared to [15], our asymmetric interaction module is a non-symmetric modeling (i.e. we treat primary and secondary non-equally), while existing method [15] treats them equally. Nevertheless, they can be applied to model the actions in joint-based interaction, but they ignore the subordination modeling in the asymmetric interaction.

Interactive Models. Recently, an increasing number of interactive models [10, 25, 15, 17] have been proposed. Wang et al [25] constructed a channel-wise inter-

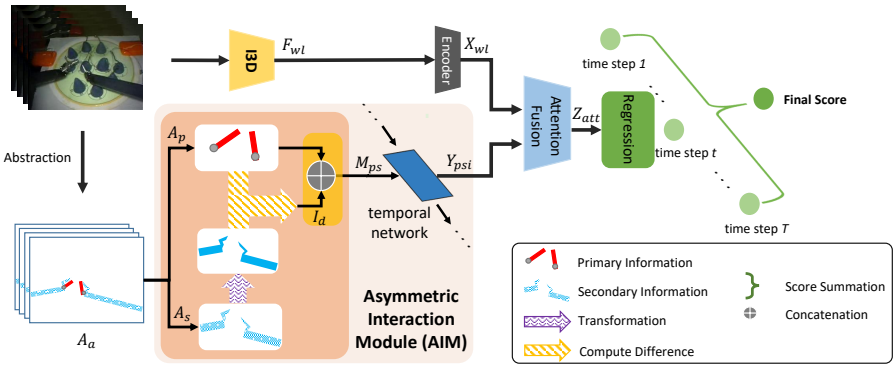

Fig. 2. An overview of our proposal. We uniformly divide an input video into T time steps, and present the process of the asymmetric interaction module (AIM) in a clear manner at time step t. The kinetic information of mobile objects is extracted, including the *primary* and the *secondary* ones. We perform asymmetric interaction between the *primary* and the *secondary* and obtain the AIM feature. Afterwards, we perform attentive contextual interaction between the whole-scene feature, which is extracted via I3D [2], and the AIM feature with an attention fusion. Finally, a regression module is utilized to learn regressing the action quality.

action learning method to evaluate the interaction of each part of an image to preserve information with a binary feature map through prior knowledge graphs. Li and Cai [10] paid different attention to the interaction of each individual part extracted from the images. In action assessment, many related works [17, 20] fed the key-points feature into an LSTM [8] framework to explore the interaction in the temporal domain, while other works [15, 27] have exploited the interactive relations among the human skeleton through GNNs. However, these methods either yield poor interpretability of the interaction process or are limited by the connection of nodes, resulting in poor generalization to various assessment tasks.

3 Approach

In this section, we will introduce our model for asymmetrically interactive action assessment in detail. The overall structure of our model is shown in Fig. 2. In this framework, the asymmetric interaction module is particularly designed to explore the asymmetric interaction between primary agent) and secondary agents. An attentive textual interaction with an attention fusion module is developed to further fuse the AIM feature and whole-scene feature. Finally, a regression module is used to learn regressing the action quality.

3.1 Asymmetric Interaction Module

It is common that the interactions among multiple people or multiple agents, in particular the asymmetric interaction among humans and objects, play impor-

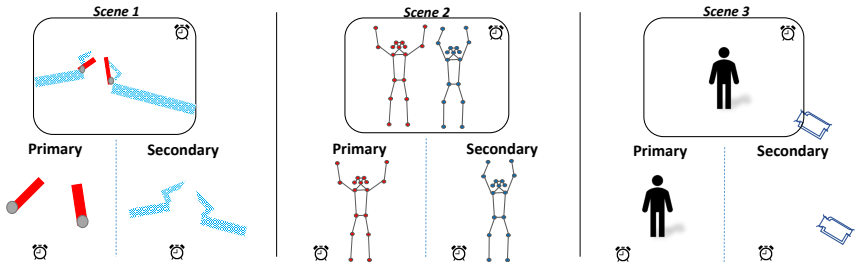

Fig. 3. Examples of the primary information and secondary information partitioning. The clock icon indicates the motion of that part. Specially, there exist actions in which two performers are in weak asymmetric relation or equal. In these cases, any one of the performers can be assigned as the *primary* and the other as the *secondary*.

tant roles. In order to reduce noise interference, we extract subtle but informative feature at an abstraction level, which we denote as A_a; that is only indispensable kinetic information for describing an action is considered, such as human pose. For example, on surgery tasks in egocentric views, we assign A_a with the kinetic information of the tool-tips which contains the object information (e.g., tool orientations) and speed information; while for most types of action performance, the entire human body should be considered, so A_a is the pose information provided by some pose estimator.

Before performing interaction, we divide A_a into two parts according to their semantics, the primary information, denoted as A_p, and the secondary information, denoted as A_s. An example diagram is shown in Fig. 3. For egocentric surgical videos, where only motions of two tool-tips are captured, it is more intuitive since each tool-tip consists of a master part and a slave one. For semantic consistence, we regard motions of the master tool-tips as the *primary*, and those of the slave tool-tips and handles, which are relatively inactive, as the *secondary*.

To explicitly explore the asymmetric relations between primary information and secondary information, we design the asymmetric interaction module (AIM), as shown in Fig. 2. With different semantics, it is natural that the primary information and secondary information come from different domains. Thus, to explore the potential relation and asymmetric interaction between the *primary* and the *secondary*, we first pass secondary information through a transformation module to map it into a latent space, the same as that of the primary. When the *primary* and *secondary* are from the same domain, the transformation will tend to learn an identity function [3]. Afterwards, we determine the difference between the *primary* and the *secondary* after the transformation, where the difference operation is an effective operation to explore the relation between visual instances [15], and the process can be formed as

$$I_d^{(t)} = \mathcal{D}(A_p^{(t)}, \mathcal{T}(A_s^{(t)})), \tag{1}$$

where $A_*^{(t)}$ denotes a certain feature A_* in time step t in Fig. 2, $I_d^{(t)} \in \mathbb{R}^N$ and $\mathcal{T}(\cdot)$ is a function to conduct the transformation operation, and $\mathcal{D}(\cdot)$ is a function to determine the difference between the *primary* and the *secondary*. Here, N denotes the dimension of the $I_d^{(t)}$ and $A_p^{(t)}$.

According to the discussion above, the primary information is dominant relative to the secondary in the capability of representation for interactive actions. To utilize the superiority of the primary information, we then concatenate the difference feature and primary information, called the primary-secondary information and denoted as M_{ps}. We present it as

$$M_{ps}^{(t)} = A_p^{(t)} \oplus I_d^{(t)}, \tag{2}$$

where $M_{ps}^{(t)} \in \mathbb{R}^{2N}$ and $\oplus$ represents the concatenating operator.

The process mentioned above can be regarded as the interaction in the spatial domain. Moreover, since the interactions occur over time, temporal relations for asymmetrically interactive action assessment are essential. Then, we use a temporal network to learn the temporal interaction and obtain the complete AIM feature, which can be expressed as

$$Y_{psi}^{(t)} = \mathcal{P}(M_{ps}^{(t)}), \tag{3}$$

where $Y_{psi}^{(t)} \in \mathbb{R}^d$ and $\mathcal{P}(\cdot)$ is a temporal network, for which we use LSTM in this work, and d is the dimension of the hidden layer in the LSTM model.

3.2 Attentive Contextual Interaction

To assist the learned AIM feature, we further employ I3D [2] to extract the whole-scene feature of videos, denoted as F_{wl}. To some extent, the whole-scene feature contains extra information complement to our AIM feature, even though noise exists. Now, we obtain two-stream outputs, the whole-scene one F_{wl} and the AIM one Y_{psi}. Before fusion of these outputs, we pass F_{wl} through an encoder to map F_{wl} into the latent space the same as the AIM feature, Y_{psi}. Then, X_{wl} is obtained, where $X_{wl}^{(t)} \in \mathbb{R}^d$ and d is the dimension of the encoder feature.

In our fusion modeling, we perform attentive contextual interaction between the whole-scene feature and the AIM feature; that is the whole-scene feature F_{wl} is utilized to learn a key map as attention for fusion of the whole-scene feature and our AIM feature because it contains the whole-scene context. Inspired by self-attention [24], we regard $(X_{wl}^{(t)} \oplus Y_{psi}^{(t)})$ as the queries and values of attention mechanism. To be detailed, we form the fusion process as follows:

$$Z_{att}^{(t)} = W^{(t)} \circ (X_{wl}^{(t)} \oplus Y_{psi}^{(t)})', \tag{4}$$

$$W^{(t)} = softmax((X_{wl}^{(t)} \oplus Y_{psi}^{(t)})' \circ O_{key}^{(t)}),\ O_{key}^{(t)} = \mathcal{FC}_{key}(F_{wl}^{(t)}), \tag{5}$$

where $\circ$ represents the matrix multiplication, $\oplus$ represents the concatenating operator, and $softmax(\cdot)$ is the softmax function, and $\mathcal{FC}_{key}(\cdot)$ is a fully connected layer to learn the key mapping. Here, $X_{wl}^{(t)}, Y_{psi}^{(t)}, Z_{att}^{(t)}, O_{key}^{(t)} \in \mathbb{R}^d$, and A' denotes the transpose of matrix A.

3.3 Scoring for Action Assessment

In the final step, our method should give a final score for the action performance through the regression module shown in Fig. 2. The overall assessment result will be presented in a score given by

$$S = \sum^{T} \mathcal{R}(Z_{att}^{(t)}), \tag{6}$$

where S denotes the predicted score for the action performance, $Z_{att}^{(t)}$ is the output of attentive contextual interaction, T is the number of time steps in the video, and $\mathcal{R}(\cdot)$ represents the regression module implemented with two FCs.

In the training stage, we use the Mean-Squared Error(MSE) as loss function for model optimization, which is defined as $\delta = \frac{1}{C}\sum_{i}^{C}(y_i - \hat{y}_i)^2$, where y and $\hat{y}$ represent the ground truth and the predicted value respectively, and C denotes the number of samples.

4 Extension to General Interactive Action Assessment: A Multi-task Training

The asymmetric interaction module can be generalized to general interactive action assessment, even though there is no explicit primary and secondary roles between performers. The *second row* in Fig. 3 does not show a strong asymmetric relation between two performers[6], and then we choose any one of them as the *primary* and the other as the *secondary*.

To be detailed, we generalize our model by a multi-task training. The multiple tasks can naturally align to the two-stream features in reasonable semantics; the whole-scene feature can be utilized for learning action assessment on overall performance, and the AIM feature can be designed for learning action assessment on interactive actions. For instance, for synchronized diving, the execution score and synchronization score are given by referees during scoring for the entire action performance. We could assess the execution of action using the whole-scene feature, which several existing methods [17, 18] have conducted, while feature extracted by AIM is capable to be utilized for learning the synchronization of action reasonably since AIM mainly explores the interaction between two players. Thus, we can use the whole-scene feature X_{wl} to learn scoring for the execution and Y_{psi} for synchronization of the action. Their assessment results are given by

$$S_{ex} = \sum^{T} \mathcal{R}_A(X_{wl}^{(t)}),\ S_{sn} = \sum^{T} \mathcal{R}_B(Y_{psi}^{(t)}), \tag{7}$$

where S_{ex} and S_{sn} denote the predicted execution score and synchronization score, respectively. $\mathcal{R}_*(\cdot)$ represents the regression module implemented with two fully connected layers.

[6] For interactive actions involving more than two performers, the important people detection [23, 11, 12] can be utilized to divide performers into the *primary* (the most important one) and the *secondary* (the rest).

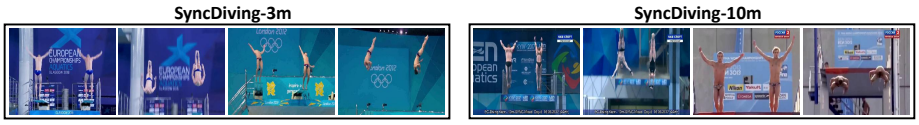

Fig. 4. Samples of the *TASD-2 dataset*.

The loss function in this setting could be formulated as

$$\mathcal{L} = \mathcal{L}_{fn} + \theta * \mathcal{L}_{ex} + (1 - \theta) * \mathcal{L}_{sn}, \tag{8}$$

where $\mathcal{L}_{fn}$, $\mathcal{L}_{ex}$ and $\mathcal{L}_{sn}$ denote the loss functions of regression for final scores, execution scores and synchronization scores, respectively. θ denotes a trade-off weight. Similarly, the loss function is the Mean-Squared Error(MSE).

The overall loss function shown in Eq. (8) is meaningful, since in synchronized diving, a great performance must be excellent in both synchronization and execution. Therefore, apart from the final score, the execution score and synchronization score are also utilized to perform a multi-task training.

4.1 TASD-2 dataset

For assessing general interactive action, we also collect a new dataset. While *AQA-7* [18] contains two events involving two performers, namely the synchronized 3-m springboard and 10-m platform, these events were captured from the side view, and thus it is hard to investigate the interaction between two performers as they overlap seriously for most of the time. Therefore, we collected a new diving dataset whose videos provide a better view to capture the interaction between two performers on synchronized diving videos. The construction details of our TASD-2 dataset can be found in the supplementary materials.

5 Experiments

We mainly conducted experiments on assessment of interactive action on *JIGSAWS* and *TASD-2*. In addition, conventional action assessment for single person can be regarded as a special extension of our method, and an evaluation of it on *AQA-7* [18] was also conducted.

Table 1. 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

Fig. 5. Frames of samples in *JIGSAWS*.

- Dataset introduction. We conduct experiments on *JIGSAWS* [6] and *TASD-2*. *JIGSAWS* contains egocentric videos of three surgical tasks, including suturing, needle passing and knot tying. There are 206 videos in this dataset, of which 78 are for suturing, 56 for needle passing, and 72 for knot tying. Samples are shown in Fig. 5. The videos are captured in stereo recordings with left views and right views by using two cameras. All videos will be used in our experiments. For each video in *JIGSAWS*, 3D kinetics information of the master tool manipulators and patient-side manipulators is provided. The details of *TASD-2* can be found in Section 4.1.

5.1 Implementation Details

- Model training setting.[7] Our model is implemented in PyTorch. Without specific explanation, our model uses Adam Optimizer with a weight decay rate of 0.5. In the training process, the batch size is 64. We use cyclic learning rates of {1e-4, 1e-5, 1e-6} changing according to the {20, 50, 100}th epoch in every 100 epochs. For each task, we train our model for 3000 epochs. T is set to 10. The encoder was implemented with a fully connected layer of shape 400×512 with ReLU activation, and the LSTM is a single layer with a 512-dimensional output. In AIM, we used fully connected layers with input and output in the same dimension as the learnable transformation operation. The difference operation is vector subtraction. In the regression module, two FC layers were utilized. The first has a shape of 512×128 with ReLU activation, and the second has a shape of 128×1 without an activation function to avoid dead ReLU during score regression. The dropout parameter is set to 0.2 and θ is 0.4.

- Evaluation Metric. For comparison with previous works [15, 18, 17, 20], we use Spearman's rank correlation as the evaluation metric of our model. It is defined as $\rho = \frac{\sum_i (p_i - \bar{p})(q_i - \bar{q})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (q_i - \bar{q})^2}}$, where p and q represent the ranking of two sequences and $-1 \leq \rho \leq 1$. The higher the Spearman's rank correlation is, the more positive the ranking relation between two sequences is. It will be used to evaluate the ranking relation between the predicted and ground-truth assessment results of our model. In order to better reflect the performance of our methods, we run the model 10 times and report the average as the final model performance. Moreover, for multiple actions in a dataset, we compute

[7] Details of data preprocessing can be found in the supplementary materials.

Table 2. The results(%) of our proposal compared with the state-of-the-art methods and our baseline on*JIGSAWS*.

	Sutur-ing	Needle Passing	Knot Tying	Avg. Corr.
ST-GCN [27]	31	39	58	43
TSN [4]	34	23	72	46
JR-GCN [15]	36	51	75	57
Baseline	5	9	11	8
Baseline+Kinetic	17	37	73	46
Ours	**63**	**65**	**82**	**71**

the average Spearman's rank correlation across actions from individual action correlations by using Fisher's z-value, as in [18].

5.2 Comparison

Experiments on interactive actions. We first evaluate our model on *JIGSAWS*, and the results are shown in Table 2, comparing with the state-of-the-art methods and our baseline. To the best of our knowledge, the methods proposed in [15, 4, 27] had achieved state-of-the-art performance for skill action assessment on *JIGSAWS*, and we used 4-fold cross validation on *JIGSAWS* by following [15]. In comparison, the results show that our model outperforms the previous state-of-the-art methods and achieves the best results, with an improvement of 14% on average. According the structure of our model, it is common to find that the great performance of our method partially benefits from the well-performed I3D [2] and partially profits from the asymmetric interaction. Thus, we remove the AIM part in Fig. 2, and evaluate our baseline by only using the I3D feature. We also concatenate I3D feature and kinetics feature as a stronger baseline. The results in Table 2 (the last three rows) indicate that the asymmetric interaction is much important in our model. The effectiveness of AIM is confirmed. Moreover, ablation study in Section 5.3 demonstrates that the roles of *primary* and *secondary* could not be exchanged for their asymmetric relation on *JIGSAWS*.

We also compared our method with the best non-deep learning approach reported in [30] using leave-one-user-out(LOUO) in Table 3.As shown, both JR-GCN [15] and ours have their own strength. However, since the LOUO setting is demanding for the model's gerenration ability, our model is better and less specialized than [15], in which each joint is modelled in a specialized manner.

To confirm the generalization of our framework to actions in weak relations between the *primary* and the *secondary*, experiments on *TASD-2* are performed, and the results are shown in Table 4. Since *TASD-2* is brand new, we utilize a naive model (RANDOM) that predicts scores for actions performance randomly

Table 3. Evaluation(%) on *JIGSAWS* with LOUO.

	Sutur-ing	Needle Passing	Knot Tying	Avg. Corr.
DTC+DFT +ApEn [30]	37	25	60	41
JR-GCN [15]	35	**67**	19	40
Ours	**45**	34	**61**	**47**

Table 4. Results (%) of our model on *TASD-2*.

	SyncDiv-ing-3m	**SyncDiv-ing-10m**	**Avg. Corr.**
RANDOM	-3	3	0
C3D-LSTM [17]	-14	1	-7
I3D [2]-SVR-L	77	73	75
I3D [2]-SVR-P	84	83	83
I3D [2]-SVR-RBF	71	77	74
JR-GCN [15]	89	81	86
Baseline	84	79	82
Baseline+Pose	88	80	84
Ours (Single-task)	89	**85**	87
Ours (Multi-task)	**92**	**85**	**89**

in the range of $[0, 100]$. The results illustrate that the distribution of samples in *TASD-2* is relatively reasonable. We also evaluate C3D-LSTM [17] on *TASD-2*, but it did not work based on the experimental setting in [18, 17]. Then, we use I3D [2] and SVR with different kernels, including linear polynomial and RBF kernels, on *TASD-2*. The results show that I3D-SVR models gain great performance, which reflects the strong ability of I3D to some extent. With the multi-task training in our model, our proposal achieved state-of-the-art performance on *TASD-2*, with a more than 3% improvement on average.

5.3 Ablation study

Table 5 shows the results of an ablation study on our model. To explore the contributions of each main module in our model, we conduct experiments by removing one of the components from our full model, including the attention fusion module and AIM. When replacing the attention fusion module with a fusion in each half, the model performance decreases by 4% on average. This result implies that paying different amounts of attention to whole-scene feature and AIM feature exactly makes a positive difference. Removing the transformation or difference module respectively, a 3% reduction was observed, indicating that these modelings are necessary. Moreover, when simply removing AIM part, the the performance decreases by 31%. These results indicate the significance of the asymmetric interaction and the effectiveness of AIM structure.

We also exchanged the *primary* and the *secondary* when performing model training and evaluating. The resulting performance reduction of 4% implies that

Table 5. Ablation study (%) for exploring the effectiveness of each main module of our model on *JIGSAWS*.

	Sutur-ing	**Needle Passing**	**Knot Tying**	**Avg. Corr.**
Full model	**63**	**65**	**82**	**71**
w/o AIM	7	41	64	40(-31)
w/o attention fusion module	61	55	80	67(-4)
Exchange *primary* and *secondary*	55	62	80	67(-4)
w/o transformation module	61	62	79	68(-3)
w/o difference module	60	61	80	68(-3)
Whole-scene(Baseline)	5	9	11	8(-63)

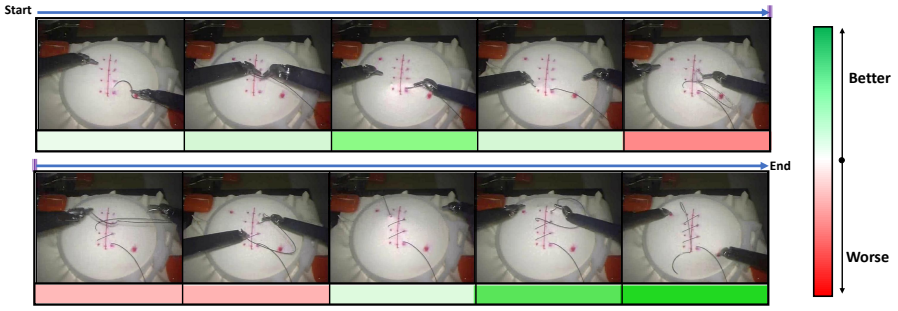

Fig. 6. The action assessment results of our model on a suturing case. The assessment results of our model indicate good (in green) and bad (in red) action performance for each time step. Best viewed in colour.

the *primary* and the *secondary* really play their semantic roles with asymmetric interaction in the model evaluation. Moreover, from the last two rows of Table 4, we find that our proposal with multi-task training increases by more than 2% in model performance on average, compared to that with single-task setting. Thus, the results indicate that the multi-task training is effective.

Moreover, we exchanged the *primary* and the *secondary* in our modeling when evaluating on *TASD-2*. The results are shown in Table 6. There was little difference as compared to the performance without exchange. Thus, it indicates that our proposal is adapted to interactive actions in weak asymmetric relation in semantics, such as synchronized diving.

Table 6. Results (%) of exchanging the *primary* and the *secondary* on *TASD-2*.

	SyncDiving-3m	**SyncDiving-10m**
Before exchanging *primary* and *secondary*	**91.50**	**85.13**
After exchanging *primary* and *secondary*	**91.75**	**85.10**

5.4 Visualization of the assessment process

In order to view the process of assessment, we output the predicted sub-score defined in Eq. (6). Fig. 6 shows an example about scoring in each time step[8]. We find that our model could give a reasonable score for each time step. Before accomplishing the first passing of the line used for suturing, it is difficult for most of us to control the surgical line expertly with tool-tips. Thus, it is not suitable to judge clearly a good or bad performance at this stage. Accordingly, we could observe that the proposed model gave relatively neutral judgement in

[8] Videos can be found in the supplementary materials.

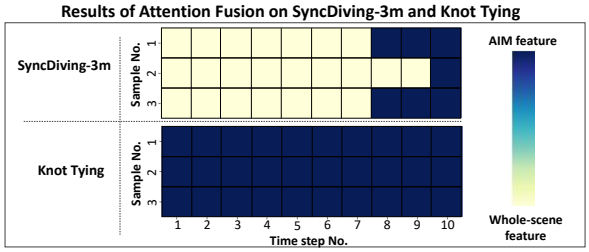

Fig. 7. Visualization of the attention fusion. We output the results of attention fusion on different actions, including synchronized 3-m springboard and knot tying. "Sample No." represents number of three randomly selected samples, and "Time step No." represents number of ten time steps for each video sample. The results indicate that our attention fusion could pay different amounts of attention on different time steps.

the first few time steps in Fig. 6. However, in the middle stage of the suturing case, we found that two tool-tips performed relatively abnormally, causing the surgical line to be staggered in the air; thus, bad judgements were obtained during this time. Correspondingly, when approaching to finishing the suturing task, our model scored with positive judgements for great performance in this process. Therefore, the visualization also confirms that our framework is effective and interpretable.

In addition, we also visualized the attention fusion through observing the computed results of Eq. (4) in Fig. 7. For the synchronized 3-m springboard, the attention fusion module could pay different amounts of attention on different time steps in a sample. The AIM feature is more important after time step 8 for SyncDiving-3m, because the interaction between two actors when they were approaching entry is more importance for synchronized diving assessment. It was obvious that our attention fusion also did make a difference by comparing different actions. It indicates that our attentive contextual interaction with an attention fusion is effective.

5.5 Extended experiment on single-person actions

The secondary information is relatively difficult to determine for single-person actions due to semantically only one motion in videos. For generalization, we define a condition that if the secondary information is ambiguous, we can use the motion of the camera capturing the action performance for replacement, as shown in the third row in Fig. 3. We additionally evaluated our framework on *AQA-7* [18] under such an assumption; this dataset is collected from summer and winter Olympics and contains 1106 videos in total composed by six actions. As discussed in Section 4.1, *AQA-7* [18] contains two-person actions, but only captured from the side view. The performers are not visually separable. Thus visually there is only one agent in the videos, and we regard it as the *primary* without other choices. Then, we extract the motion feature of the camera as the

Table 7. Results (%) of our model applied to *AQA-7*. To illustrate the competitive results, the average of the rank among existing methods is used.

	diving	Gymvault	skiing	snowboard	sync. 3m	sync.10m	Avg. Corr.	Avg. Rank.
Pose+DCT [20]	53.00**(5)**	–	–	–	–	–	–	**5**
ST-GCN [27]	32.86**(6)**	57.70**(4)**	16.81**(5)**	12.34**(5)**	66.00**(4)**	64.83**(5)**	44.33**(5)**	**4.9**
C3D-LSTM [17]	60.47**(4)**	56.36**(5)**	45.93**(4)**	50.29**(2)**	79.12**(3)**	69.27**(4)**	61.65**(4)**	**3.7**
C3D-SVR [17]	79.02**(1)**	68.24**(3)**	52.09**(3)**	40.06**(4)**	59.37**(5)**	91.20**(2)**	69.37**(3)**	**3**
JR-GCN [15]	76.30**(2)**	73.58**(1)**	60.06**(1)**	54.05**(1)**	90.13**(2)**	92.54**(1)**	78.49**(1)**	**1.3**
Ours	74.19**(3)**	72.96**(2)**	58.90**(2)**	49.60**(3)**	92.98**(1)**	90.43**(3)**	77.89**(2)**	**2.3**

secondary, by computing the optical flow (using the TV-L1 algorithm [19]) at the region near the edge of images. In this task, we fix the weight decay rate of Adam Optimizer in our model to 0.8. For consistency, the performance results that we report in Table 7 are obtained and the experimental setting follows [15]; the results demonstrate that our method is competitive compared with current state-of-the-art methods, with the best performance on sync. 3m action assessment. Our proposal outperforms most of the state-of-the-art methods except JR-GCN [15], and its performance score is only 0.6% less than that of JR-GCN [15] on average. Therefore, the extended experiment demonstrates that our framework is capable to generalize effectively to common action assessment tasks.

6 Conclusion

In this work, we proposed a novel asymmetric interaction model for asymmetrically interactive action assessment. In our model, we categorize the roles in an asymmetrically interactive action as a primary agent and secondary ones. With the asymmetric interaction, we can model the interactive actions in strong asymmetric relation. We evaluated our model on *JIGSAWS* [6] and our method achieved the state-of-the-art performance. Moreover, experimental results on *TASD-2*, a new dataset (to be released) collected in our work, also demonstrated our method could be generalized to general interactive actions in weak asymmetric relation. The extra experiments on *AQA-7* [18] have also indicated that our model can be adapted to perform conventional action assessment. For future development, our method can also be extended to actions involving more than two people, with the help of the important people detectors [23, 11, 12]. It will be explored in the future work along with constructing relevant datasets.

Acknowledgement

This work was supported partially by the National Key Research and Development Program of China (2018YFB1004903), NSFC(U1911401,U1811461), Guangdong Province Science and Technology Innovation Leading Talents (2016TX03X157), Guangdong NSF Project (No. 2018B030312002), Guangzhou Research Project (201902010037), and Research Projects of Zhejiang Lab (No. 2019KD0AB03).

References

1. Bertasius, G., Soo Park, H., Yu, S.X., Shi, J.: Am i a baller? basketball performance assessment from first-person videos. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2177–2185 (2017)
2. 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)
3. Chen, J., Wang, Y., Qin, J., Liu, L., Shao, L.: Fast person re-identification via cross-camera semantic binary transformation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
4. Doughty, H., Damen, D., Mayol-Cuevas, W.: Whoś better, whoś best: Skill determination in video using deep ranking. proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
5. Doughty, H., Mayol-Cuevas, W., Damen, D.: The pros and cons: Rank-aware temporal attention for skill determination in long videos (June 2019)
6. 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)
7. Gattupalli, S., Ebert, D., Papakostas, M., Makedon, F., Athitsos, V.: Cognilearn: A deep learning-based interface for cognitive behavior assessment. In: Proceedings of the 22nd International Conference on Intelligent User Interfaces. pp. 577–587. ACM (2017)
8. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with lstm. IET Conference Proceedings pp. 850–855(5) (January 1999)
9. Ilg, W., Mezger, J., Giese, M.: Estimation of skill levels in sports based on hierarchical spatio-temporal correspondences. In: Joint Pattern Recognition Symposium. pp. 523–531. Springer (2003)
10. Li, H., Cai, Y., Zheng, W.S.: Deep dual relation modeling for egocentric interaction recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
11. Li, W.H., Hong, F.T., Zheng, W.S.: Learning to learn relation for important people detection in still images. In: Computer Vision and Pattern Recognition (2019)
12. Li, W.H., Li, B., Zheng, W.S.: Personrank: detecting important people in images. In: International Conference on Automatic Face & Gesture Recognition (FG 2018) (2018)
13. Malpani, A., Vedula, S.S., Chen, C.C.G., Hager, G.D.: Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In: International Conference on Information Processing in Computer-Assisted Interventions. pp. 138–147. Springer (2014)
14. Paiement, A., Tao, L., Hannuna, S., Camplani, M., Damen, D., Mirmehdi, M.: Online quality assessment of human movement from skeleton data. In: British Machine Vision Conference. pp. 153–166. BMVA press (2014)
15. 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)
16. Parmar, P., Morris, B.T.: What and how well you performed? a multitask learning approach to action quality assessment. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

17. Parmar, P., Tran Morris, B.: Learning to score olympic events. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 20–28 (2017)
18. 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
19. Pérez, J.S., Meinhardt-Llopis, E., Facciolo, G.: Tv-l1 optical flow estimation. Image Processing On Line pp. 137–150 (2013)
20. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: European Conference on Computer Vision. pp. 556–571. Springer (2014)
21. Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE Transactions on Neural Networks **20**(1), 61–80 (2009)
22. Sharma, Y., Bettadapura, V., Plötz, T., Hammerla, N., Mellor, S., McNaney, R., Olivier, P., Deshmukh, S., McCaskie, A., Essa, I.: Video based assessment of osats using sequential motion textures. Georgia Institute of Technology (2014)
23. Solomon Mathialagan, C., Gallagher, A.C., Batra, D.: Vip: Finding important people in images. In: Computer Vision and Pattern Recognition (2015)
24. 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`
25. Wang, Z., Lu, J., Tao, C., Zhou, J., Tian, Q.: Learning channel-wise interactions for binary convolutional neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
26. Xu, C., Fu, Y., Zhang, B., Chen, Z., Jiang, Y.G., Xue, X.: Learning to score the figure skating sports videos. arXiv preprint arXiv:1802.02774 (2018)
27. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
28. Zhang, Q., Li, B.: Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. In: Proceedings of the 2011 international ACM workshop on Medical multimedia analysis and retrieval. pp. 19–24. ACM (2011)
29. Zhang, Q., Li, B.: Relative hidden markov models for video-based evaluation of motion skills in surgical training. IEEE transactions on pattern analysis and machine intelligence **37**(6), 1206–1218 (2015)
30. Zia, A., Essa, I.: Automated surgical skill assessment in rmis training. Int J CARS **13**, 731–739 (2018)
31. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Clements, M.A., Essa, I.: Automated assessment of surgical skills using frequency analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 430–438. Springer (2015)
32. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Essa, I.: Video and accelerometer-based motion analysis for automated surgical skills assessment. International journal of computer assisted radiology and surgery **13**(3), 443–455 (2018)
33. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Ploetz, T., Clements, M.A., Essa, I.: Automated video-based assessment of surgical skills for training and evaluation in medical schools. International journal of computer assisted radiology and surgery **11**(9), 1623–1636 (2016)