EAGLE-Eye: Extreme-pose Action Grader using detaiL bird's-Eye view

Mahdiar Nekoui
University of Alberta
nekoui@ualberta.ca

Fidel Omar Tito Cruz
Universidad Nacional de Ingeniería
ftitoc@uni.pe

Li Cheng
University of Alberta
lcheng5@ualberta.ca

Abstract

Measuring the quality of a sports action entails attending to the execution of the short-term components as well as overall impression of the whole program. In this assessment, both appearance clues and pose dynamics features should be involved. Current approaches often treat a sports routine as a simple fine-grained action, while taking little heed of its complex temporal structure. Besides, they rely solely on either appearance or pose features to score the performance. In this paper, we present JCA and ADA blocks that are responsible for reasoning about the coordination among the joints and appearance dynamics throughout the performance. We build our two-stream network upon the separate stack of these blocks. The early blocks capture the fine-grained temporal dependencies while the last ones reason about the long-term coarse-grained relations. We further introduce an annotated dataset of sports images with unusual pose configurations to boost the performance of pose estimation in such scenarios. Our experiments show that the proposed method not only outperforms the previous works in short-term action assessment but also is the first to generalize well to minute-long figure-skating scoring.

1. Introduction

From just a glance at a video or by looking into its key frames, you may infer what action is being carried out in the video. But what if you are asked to assess the quality of the action?

Judging a competitive sporting event and awarding scores to the performers needs a keen eye for details while still looking from the bird's eye view on the whole routine. This challenging problem has introduced a new branch to the human activity analysis field which is known as Action Quality Assessment (AQA). Due to the wide applications of AQA in many sports like gymnastics, diving, and etc., this new field has attracted considerable attention in recent years [15, 16, 17, 26]. However, the negligence of some factors has affected the performance of previous action assessors.

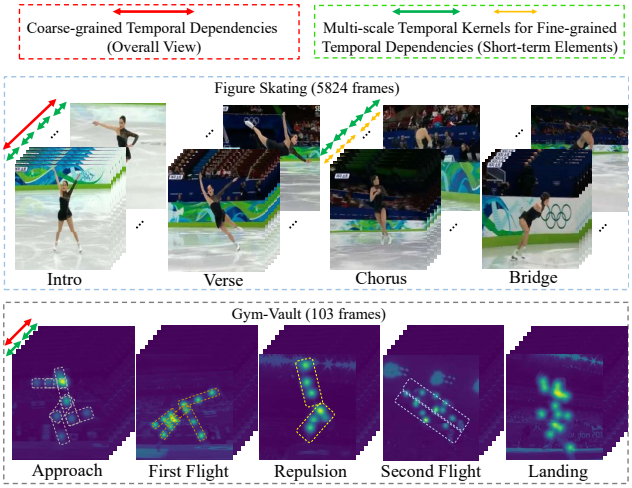


Figure 1: To assess the performance of an athlete both fine and coarse-grained temporal dependencies should be captured. The assessment is based on the coordination among the joints/body parts as well as appearance dynamics features.

The first factor regards how to model an activity. A sports routine can be considered as a long-term activity, comprising some medium-term phases. For example, in a figure-skating contest the athlete phrases his/her program into intro, verse, chorus, and bridge in unison with the music being played (see Fig.1). Each of these phases is composed of some short-term elements like jumps, spins, and footwork sequences. The judge should keep track of the execution of each short-term element to award the Grade of Execution (GOE) [24]. On the other hand, the athlete's overall skating skill and composition of the elements through each medium-term phase are assessed to provide the Program Component Score (PCS) [23]. The same thing can be applied to other sports fields. A gym-vault routine is constructed of the approach, first flight (a half twist on the ground with fully straight knees and body), repulsion (pushing off from the table with straight legs), second flight (airborne performance of saltos and twists in a tuck, pike, or free position), and landing phases (see Fig.1). The well-execution of each short-term element like a salto or a

twist as well as having an overall smooth performance in medium-term phases would result in a high score. Existing methods [16, 15, 26] mainly rely on short-term temporal relations of few consecutive frames and neglect the long-term temporal relations that create a holistic view over phases. This problem escalates in the case of longer activities like figure skating in which each medium-term phase may take a few minutes to unfold. This fact has led the most of the previous approaches to be only applicable to short-lasting activities like gym-vault [16, 15, 13].

The second factor regards where and when to attend more to award the score. In typical individual sports footage the cameraman tracks the athlete to locate him at the center of the video. Thus, there should be a higher attention on the middle pixels of each frame to assess the performance. In the temporal domain, the judge may deliberately place some components over some others. For example, in a figure skating contest as the time goes on the athlete gets more tired and performing each element becomes more difficult. In order to acknowledge the skill and stamina of the skater, the judge gives bonus marks to well-execution of short-term elements in later phases [22]. Thus, the later parts of the performances are most likely to make the differences between the athletes' scores. Another example is gym-vault in which having a perfect landing in the last frames has a higher contribution than other phases in the scoring schema. On the other hand, there are some parts in long-term activities on which the judge may unintentionally focus more on. The PCS score of each figure-skater is awarded after the whole program which takes about three minutes to complete. As a result, the composition of the components and the holistic view of the last medium-phases are remembered most. Overall, there has been a lack of exploration of these factors in awarding the final score.

Moreover, unlike action recognition in which appearance-based features are informative enough for most cases to classify an action, integrating pose features in the process of assessing a performance is of great significance. Nevertheless, due to the difficulty of estimating pose features in the extreme contortions of the body throughout the performance, most of the previous methods [16, 11, 26] resorted to the whole-scene appearance features to regress the score. As a result, the coordination and dynamics of the body's joints, an important criterion of assessing an action, is ignored. Recently Pan *et al*. [15] proposed a graph-based method to capture the local motion of some certain body parts as well as coordination among joints in each frame. The resulted features together with the raw whole-scene appearance-based features are encoded to regress the final score. However, splitting the skeleton into some predefined patches may lead to ignoring the dependencies between some others. Besides, the blurriness of image frames, as well as extreme contortions of the body, led to underperformance of the pose estimator and the whole network respectively. To facilitate the estimation of pose in such extreme cases, Nekoui *et al*. introduced the *ExPose* dataset that covers such scenarios in a diving routine. However due to the domain discrepancy between different contortive sports, the dataset has limited utility in other fields.

In this paper, we propose EAGLE-Eye a modular two-stream network that sits on top of extracted appearance-based and pose-based features of a sports activity and evaluates the quality of the performance. The first stream is responsible for assessing the coordination among the joints and a variety of body parts with the help of a stack of JCA blocks (see the upper part of Fig.2). The first blocks capture the short-term temporal dependencies of the individual joints at different tempos with the help of multi-scale temporal kernels and temporal-wise channel convolutions. By stacking more of these JCAs both the temporal and semantic receptive field gets more broad, contributing to capture the holistic view of the performance as well as dependencies of different body parts/super-joints. Likewise, the second stream captures the both fine and coarse-grained appearance dynamics with the help of its stacked ADA blocks (see the lower part of Fig.2). The network is also supplied with some spatial and temporal attention blocks to increase the contribution of the frames' middle pixels and last medium-term phases in assessing the sports action.

We summarize the contributions of the paper as follows:

- In order to handle estimation of the pose in the extreme contortions of the body in different sports, we have extended the *ExPose* dataset [13] to cover other sports than diving like synchronized diving, snow-boarding, and skiing. It is demonstrated that training the pose estimator on this dataset improves its performance in the extreme pose configurations of such sports.

- We propose a modular network that quantifies how well an action has been performed based on both fine and coarse-grained temporal dependencies. Same as the case of human judges grading schema, both visual and pose clues have been involved in the assessment.

- The proposed network not only outperforms the previous works in short-term actions assessment but also is the first to demonstrate a good generalization to the case of long-term sports activities like figure-skating. We further provide a thorough ablation study to evaluate the effectiveness of each block of the network.

2. Related Work

Action Quality Assessment: Previous works in the literature of AQA can be mainly divided into two categories.

Appearance-based methods only use the whole-scene appearance features to regress the score of each performance. Parmar *et al*. [16] feed C3D [21] features of the routine into a regression network (SVR or LSTM) to get the score. Li *et al*. [11] explore the contribution of dividing each sample into some fragments to get the most distinctive C3D features of each routine. In [17], taking average over C3D features of multiple clips in the temporal domain is shown to be effective. It was further demonstrated that leveraging commentary and detailed description of each diving routine would help have a better action evaluator.

A few works have focused on the pose features to regress the score. Pirsiavash *et al*. [19] apply DCT on the pose sequence of a routine and feed the result to a linear SVR. Recently, Pan *et al*. [15] proposed joint difference and commonality modules to represent the coordination among the joints and motion of certain body parts for consecutive timesteps. The resulted features together with the raw appearance I3D features are ultimately fed to a score regressor. Very recently, Nekoui *et al*. [13] introduced a dataset of extreme poses to alleviate the underperformance of the pose estimator in a diving routine. They further regressed the score of a routine by late fusion of assessing the appearance features and local motion of some hand-crafted body parts.

The significance of long-term temporal modeling in a minute-long activity has made all of the aforementioned methods only applicable to short-term routines like diving. A simple way to score a minute-long activity like figure-skating is to treat it as a short-term one and simply feed the C3D features of the routine to a SVR [18]. Recently, Xu *et al*. [26] proposed multi-scale skip LSTMs to cover a more broad receptive field. The resulted features are further fused with compacted local feature representation provided by a self-attentive LSTM to regress the score of a figure-skating routine. However, not only the pose features are completely ignored but also the rigid structure of the proposed method is not applicable to short-term activities. Besides, skipping some frames/clips to have a broad receptive field would result in ignoring some useful visual clues.

In this paper, we propose a modular network that is applicable to both short-term and long-term activities. By stacking more JCA and ADA layers, our short-term action assessor turns into a long-term one (see Fig.2).

Temporal Structure Modeling: Extracting the temporal structure of a video has been extensively studied for action recognition. Niebles *et al*. [14] model each complex action as a composition of some motion segments and use Latent SVM to get the parameters of the model. To capture the long-term temporal relations, TSN [25] evenly divides the video into multiple segments and fuses the class score of sampled snippets from each segment. TRN [28] learns the pair-wise long-term temporal relation between sampled frames at different scales and fuses the resulted features.

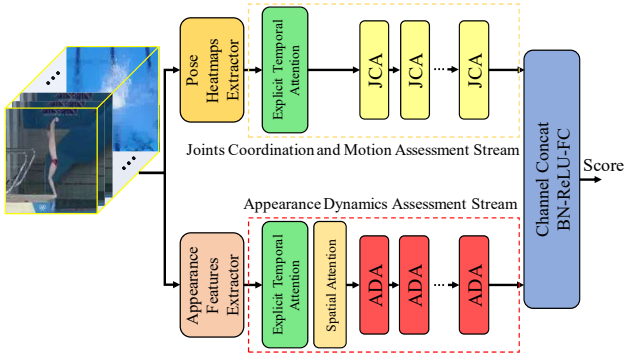


Figure 2: Overview of our pipeline. The network regresses the score of each performance based on both pose and appearance clues with attending to short-term elements as well as holistic view of the performance.

Recently, Hussein *et al*. [7] proposed a modular layer that sits on top of appearance features of a complex action video to learn its long-term temporal dependencies.

All above appearance-based methods have explored the significance of temporal modeling in action recognition. Here we investigate the temporal structure of an action to assess it based on both appearance and pose features.

3. Method

This section outlines our proposed action quality assessor. The overview of our pipeline is depicted in Fig.2. Here we delve into each block of the network and describe the intuition behind them.

3.1. Explicit Spatio-temporal Attention

Given the small size of existing datasets in AQA, proposing a simple attention mechanism that does not make the network deeper is of great importance. Here, we introduce our simple yet effective AQA-specific attention block to capture the most important parts of a routine. The first objective of this block is to model the deliberate higher attention of a judge to the last parts' short-term elements of a routine in the temporal domain. To this end, we propose an explicit temporal attention block that gradually attenuates the contribution of the first phases of a performance. Let's consider the input of this block as X with the dimension of $T\times H\times W\times C$ in which T is the number of timestamps, H and W are the spatial size of each frame, and C is the number of channels. Then the block produces $\tilde{X}$ by element-wise multiplication of the input feature with explicit temporal importance mask (M^e):

$$\tilde{X}_{h,w,c} = X_{h,w,c} \odot M^e \tag{1}$$

$$M^e_t = a + (1-a)\frac{t}{T} \tag{2}$$

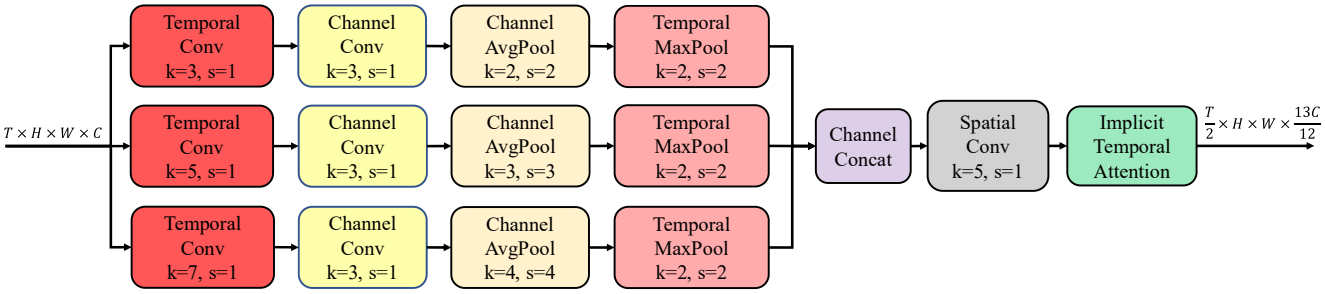


Figure 3: The overview of a JCA block.

In which $1 \leq c \leq C, 1 \leq t \leq T, 1 \leq h \leq H, 1 \leq w \leq W$ and a is a constant coefficient $0 \leq a \leq 1$. As a result, $a \leq M_t^e \leq 1$.

The other objective of this module is to increase the spatial attention on the center of each frame. We propose a spatial attention block that applies a gaussian importance mask to its feature map input. As a result, the middle pixels of the appearance features which are more likely to represent the athlete's body would have a higher contribution in awarding the score.

$$\tilde{X}_{t,c} = X_{t,c} \odot M^s \tag{3}$$

$$M^s_{x,y} = e^{-\frac{(x-\mu)^2+(y-\mu)^2}{2\sigma^2}} \tag{4}$$

3.2. Joints Coordination Assessor (JCA)

Our proposed JCA block is responsible for extracting the temporal pattern of body joints and parts which is depicted in Fig.3. Firstly it takes the $T \times H \times W \times C$ pose heatmaps of the routine in which C represents the number of joints. This input is fed to a set of multi-scale channel-wise separable temporal convolutions to capture temporal dependencies of joints at different scales. Utilizing fixed-size temporal kernels seem to be too rigid to model the complex temporal structure of short-term elements in a routine. A diver may perform a somersault and a twist together. These two elements may have different tempos and it is important to capture temporal dependencies throughout each element.

The next step is to capture the coordination among the joints. To this end, a temporal-wise separable channel convolution extracts the dependencies in the semantic subspace in which each channel represents an individual joint.

In assessing an action, the motion and coordination of different body parts is also monitored consistently. The symmetry of different body parts during the performance makes it aesthetically pleasant. In order to systematically capture such features we employ a set of average pooling with different kernel and stride sizes along channels. As a result, a variety of body parts/super-joints at multiple scales are formed. Consequently, the convolution filters of the next JCA block would capture the motion and coordination of these super-joints.

As discussed before, having a holistic view over the performance is of great importance for assessment purposes. Therefore, a temporal max pooling block at the end of each branch increases the temporal receptive field. Thanks to this, the next JCA block would be able to capture longer temporal dependencies. Obviously, minute-long activities like figure-skating require stacking more of these JCA blocks to capture the dependencies between distant frames (see Fig.2).

In the case of group activities, the pose heatmap contains more than one instance for each joint. In such cases the dependence between the joints of each performer with another should be extracted. For example, in a synchronized diving contest capturing the symmetry between the performers' joints is an important criterion of assessment. The spatial convolution block of JCA is responsible for catching such dependencies.

Finally, an implicit temporal attention block models the automatic fade of the first parts of a performance in the judge's short-term memory. A judge awards the PCS score of each figure-skating routine after the completion of the whole performance. Each routine takes two minutes and forty seconds on average. With the passage of time, the first parts of the performance become attenuated in the judges memory. To model this effect we propose a sigmoidal implicit temporal importance mask (M^i):

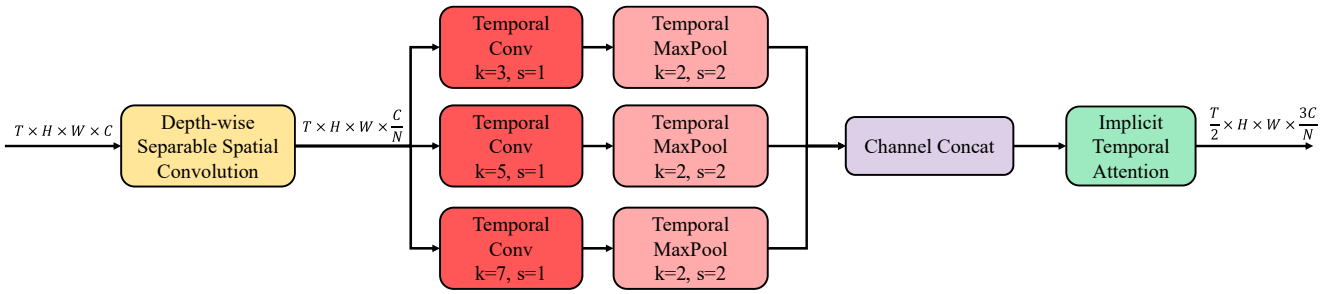
$$\tilde{X}_{h,w,c} = X_{h,w,c} \odot M^i \tag{5}$$

$$M^i_t = \frac{1 + be^{\frac{-T}{d}}}{1 + be^{\frac{-t}{d}}} \tag{6}$$

In which b and d are constant coefficients ($0 \leq b \leq 1$, $1 \ll d$). As a result $\frac{1+be^{\frac{-T}{d}}}{1+b} \leq M^i_t \leq 1$.

Consequently, the next JCA blocks which are responsible for capturing a holistic view of the performance would perceive a higher attention on the last parts. It should be noted that the explicit temporal attention block impacts the assessment of fine-grained dependencies in both short and

Figure 4: The overview of an ADA block.



long-term actions. However, the implicit temporal attention block contributes to attending more on last phases coarse-grained dependencies in a long-term action assessment.

3.3. Appearance Dynamics Assessor (ADA)

We further propose ADA blocks to capture the dynamics of appearance features. The architecture of an ADA block is depicted in Fig.4.

The input to each ADA block is either the output of its previous ADA block or the output of appearance features extractor backbone (like I3D[4]). At the first step, a depth-wise separable spatial convolution captures the cross-channel correlations of the input feature map. It also shrinks the semantic subspace by a factor of N. As a result, stacking ADAs wouldn't lead to the explosion of the number of channels. The proposed ADA head relies heavily on its appearance features extractor backbone to capture the spatial dependencies. Thus, given the small changes that the spatial attention block has made, this lightweight spatial convolution layer suffices in the new setting. Secondly, same as the case of JCA blocks, multi-scale temporal kernels are employed to capture the different visual tempos of appearance clues. Finally, as depicted in Fig.2, the resulted feature map of the ADA stream is concatenated with the JCA stream over semantic subspace and fed to a $BN - ReLU - FC$ layer to get the final score.

4. Dataset

In order to support experiments in estimating human body pose with extreme contortions and involving pose features in AQA we introduce *G-ExPose* (Generalized ExPose). Most of existing datasets for in-the-wild pose estimation only cover normal daily activities like walking, sitting, etc. captured by static cameras [12, 1]. As evident in Fig.5 these datasets are not suitable for estimating the contortive poses of a competitive sports activity which is taken by a moving camera. To address this issue, recently Nekoui *et al*. introduced *ExPose*, a collection of 3000 diving and 1000 gym-vault 2D annotated images sourced from moving camera videos. Despite showing acceptable pose estimation

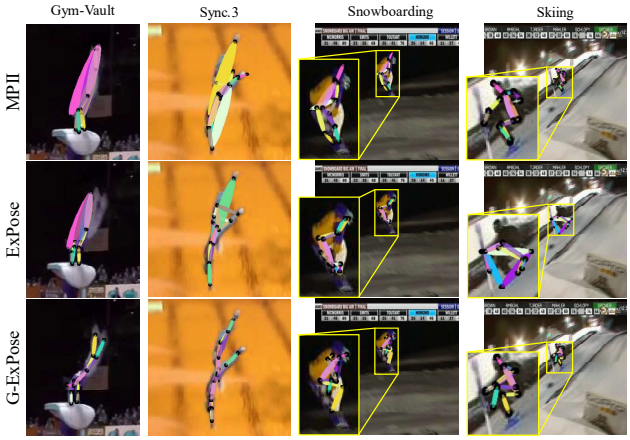


Figure 5: Qualitative test-time pose predictions of HRNet[20] model trained on MPII (first row), ExPose (second row), and G-ExPose (third row), respectively.

results in diving, utilizing this dataset leads to failure of the estimator in the case of other contortive sports (see Fig.5). To alleviate this problem, we extend *ExPose*[13] by 7500 annotated images from four different sports.

G-ExPose contains 2500 snowboarding, 2000 skiing, 1500 synchronized diving, and 1500 gym-vault 2D annotated images. The snowboarding images are obtained from 2018, 2019, and 2020 X-games competitions at Aspen. The skiing images are taken from X-games 2020 ski big air contests at Aspen and Norway. We extended *ExPose* to also cover highly occluded synchronized diving images by introducing a set of 1500 annotated images from women's 3 meter springboard and men's 10 meter platform synchronized diving finals at 2016 European diving championships in London. We further enlarged gym-vault samples of *ExPose* by annotating 1500 images from Rio 2016 Olympics and Stuttgart 2019 world championships women's vault finals. In order to collect the dataset, we first queried YouTube to get the original video of each event. Secondly, we filtered out the irrelevant parts of the video (like the opening ceremony and medal presentation) and extracted the frames of

Method	Diving	Vault	Skiing	Snowboard	Sync. 3m	Sync. 10m	**Avg. Corr.**	Skating
Pose-DCT-SVR [19]	53.00	—	—	—	—	—	—	35.00
ConvISA [10]	—	—	—	—	—	—	—	45.00
ST-GCN [27]	32.86	57.70	16.81	12.34	66.00	64.83	44.33	—
C3D-LSTM [16]	60.47	56.36	45.93	50.29	79.12	69.27	61.65	51.07*
C3D-SVR [16]	79.02	68.24	52.09	40.06	59.37	91.20	69.37	53.00
JR-GCN [15]	76.30	73.58	60.06	54.05	90.13	92.54	78.49	—
AIM [6]	74.19	72.96	58.90	49.60	92.98	90.43	77.89	—
C3D-(S+M)LSTM [26]	—	—	—	—	—	—	—	57.69*
Ours	83.31	74.11	66.35	64.47	91.43	91.58	81.40	60.10

Table 1: Detailed results on both AQA-7[16] dataset that contains short-term activities and long-term figure skating videos of extended MIT-Skate dataset (171 samples)[18]. First and second best are shown in color. Following [16, 15], we use Fisher's z-value to compute the average correlation between the short-term sports of AQA-7 dataset. The results marked with * are obtained by reimplementing the correspondent method. [26] reported 59.00 Sp. Corr. for the old MIT-Skate dataset (150 samples) in the original paper.

Sports Field	MPII[1]	ExPose[13]	G-ExPose
Sync. 3m	27.2	45.2	**56.7**
Sync. 10m	29.3	53.6	**63.9**
Skiing	18.8	5.5	**30.5**
Snowboarding	20.1	9.1	**31.0**
Gym Vault	24.0	38.8	**53.2**

Table 2: The quantitative results of HRNet pose estimator[20] on the 100 annotated images of each extreme sports field when trained on MPII, ExPose and our G-ExPose dataset. The evaluation metric is the standard PCKh@0.5[1, 20]. The position of a joint is correctly estimated if its distance with the ground truth is within 50% of the head segment length.

each video with the frame rate of 10 fps. Finally, we mostly held out the normal poses of each routine (like approaching to the springboard or after-landing in a gym-vault routine) to focus more on the main part of the execution in which the performer contorts his body in some unusual configurations. We follow the same pose annotation format as MPII dataset that considers 16 joints for the body.

Besides the qualitative evaluation of our dataset, we further quantitatively assessed the effectiveness of *G-ExPose* in extreme pose configurations in comparison with *ExPose* and an in-the-wild normal activities pose dataset like *MPII*. To this end, we first picked 10 videos from each field of AQA-7[16] dataset and annotated 10 images from each video with a focus on the main parts of the execution. This dataset contains 1106 sports routine videos as well as their correspondent score from diving, synchronized diving (3 and 10 meters), gym-vault, snowboarding, and skiing. As a result, a set of 600 annotated images from 6 different fields got collected. We then evaluated the performance of the SOTA HRNet [20] pose estimator on the 100 images of each field when it is trained on *G-ExPose*, *ExPose*, and *MPII*. As it can be seen from Tab.2, the HRNet which is trained on G-ExPose, outperforms others in the extreme pose estimation task. It should be noted that there is no conflict between G-ExPose and AQA-7 source events.

5. Experiments

5.1. Datasets and Implementation Details

For short-term AQA we follow recent works [16, 15] and evaluate our approach using the AQA-7 dataset. Each video of the dataset is originally normalized to 103 frames. We follow the same train-test data split as [16, 15]. In order to get the appearance features of each video, we use the output of `mixed-5c` layer of an I3D network pretrained on the Kinetics dataset[4]. For our pose features extractor backbone we entangled the DiMP [2] visual object tracker with the HRNet pose estimator (trained on *G-ExPose*). The channel shrinkage factor of the ADA blocks (N) is set to 2. The short-term attenuation temporal coefficient (a in Eq.2) is set to 0.9. The mean and the standard deviation of the spatial importance mask (see Eq.4) are set to 4 and 5 respectively. As discussed before, the implicit temporal attention block should only impact the long-term activities assessing. Thus, we set the long-term temporal attenuation coefficient (b in Eq.6) to 0 in short-term action assessment.

We further assess the effectiveness of the proposed model on long-term activities by evaluating it on the extended version of the MIT-Skate [19] by [18]. This dataset contains the awarded scores of 171 single figure-skating videos that take 2.5 minutes on average. In order to normalize all videos to a fixed number of frames we first extract the frames of all videos at 25 fps . We then zero-pad the first frames of each video to fit to 5824 frames which is the longest video's number of frames. We follow [18] and randomly split the dataset into 100 samples for training and 71 for testing. Since a figure skating routine does not involve extreme pose configurations, we train the pose

Ablated Model	Diving	Vault	Skiing	Snowboard	Sync. 3m	Sync. 10m	Skating
W/o Explicit Temporal Att. Block	82.66	72.91	62.11	63.23	90.59	89.72	59.66
W/o Spatial Att. Block	82.83	72.11	72.11	61.16	**91.50**	90.76	—
Fixed-size Temporal Kernels	76.71	67.17	55.09	51.99	87.17	86.81	54.53
W/o Channel Avg	81.87	71.24	63.73	62.15	89.28	90.43	57.22
W/o JCA Stream	80.95	70.96	60.46	60.63	88.70	90.18	55.49
W/o ADA Stream	74.30	70.58	57.90	57.90	85.28	85.64	51.11
# JCA and ADA Blocks = $K-1$	79.86	72.24	62.55	59.83	88.16	89.20	57.72
# JCA and ADA Blocks = $K+1$	81.53	71.18	64.94	62.11	90.24	88.32	58.41
Only Whole-Scene Appearance	63.39	68.72	51.79	50.53	87.83	88.32	44.23
W/o Implicit Temporal Att. Block	—	—	—	—	—	—	58.73
Ours	**83.31**	**74.11**	**66.35**	**64.47**	91.43	**91.58**	**60.10**

Table 3: Ablation study results on AQA-7 dataset[16] for short-term actions and MIT-Skate dataset[18] for long-term figure-skating sport: We systematically removed the components of our network to evaluate their contribution to the full model. K is equal to 2 for short-term and 4 for long-term assessment.

estimator backbone on the COCO+Foot dataset [3]. In the long-term action assessing we set b and d coefficients to 0.5 and 1000 respectively.

We train the model for 500 epochs with the learning rate of 0.005 and the batch size of 20 using Adam optimizer[9]. We use the MSE loss function to train the model and award the scores, following what other regression-based AQA methods[16, 15, 18] have done. To be consistent with the previous works, we use the Spearman's Rank correlation to evaluate the performance of the model and compare the predicted scores with the ground-truth. For further details refer to the supplementary document.

5.2. Results

We first evaluate the performance of our network on short-term activities of AQA-7 dataset. It should be noted that we have used two ADAs and two JCAs for assessing a short-term action. As it can be seen in Tab.1, the proposed method outperforms the existing SOTA AQA methods. Its worth mentioning that the JR-GCN [15] and AIM[6] methods have used the excessive optical flow information while our method resorts to the RGB frames input. The largest gaps belong to skiing and snowboarding sports. In these sports' video footage, the size of the athlete is much smaller than the size of the whole frame. Therefore, it is not surprising that methods like [16] which only rely on whole-scene appearance features to regress the score of the performers, are underperforming significantly. Besides, in such fields the position of each individual joint is as important as the symmetry of different body parts during the execution. Failure to grab the board or any insecurity that requires hand movements to remain stable affects the score negatively. Thus, extracting the features of some predefined local patches around the joints (which has been done in[15]) results in neglecting the individual joints position and motion of some body parts to award the score. On the other hand, our pose-based assessment stream not only works intuitively as an action localizer, it also judges the position of the joints as well as the motion of a variety of body parts.

We further evaluate the performance of our model on the long-term figure-skating sports activity. In order to have a fair comparison with the existing works, we changed the appearance features extractor backbone to C3D to have the same backbone as theirs. Following [26], we feed the output of `fc6` layer of C3D network which is pre-trained on Sports-1M dataset[8] to our ADA stream. As discussed before, long-term temporal reasoning is crucial to model the judge's impression of the overall performance of the figure-skater. To this end, more JCAs and ADAs (here we use 4) are stacked in the long-term activities assessment. As demonstrated in Tab1, our model generalizes well to long-term action assessment task.

We conduct a comprehensive ablation study to evaluate the effectiveness of our models components (see Tab.3). We first removed the explicit temporal attention block to uniformly attend to all frames in a short-term activity. As a result, we are neglecting the fact that having a clean landing in a snowboarding routine or performing a vertical entry to the water with the least amount of splash in a diving performance are the most distinctive features of the execution[5]. In the second set of experiments we removed the spatial attention block of the ADA stream. Therefore, we equally attend to each pixel of the extracted appearance features, no matter whether it belongs to the background or the athlete's body. The third row of Tab.3 refers to the results of using fixed-size temporal kernels instead of multi-scale ones. Consequently, the same temporal kernel size for capturing complex temporal dependencies of the two short-term elements that are performed together would be used. Fourthly, we removed the temporal-wise average pooling of the JCA

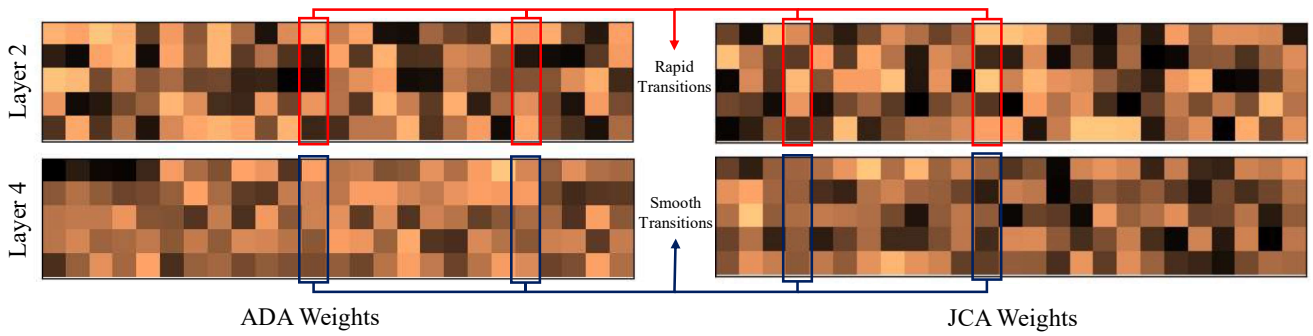


Figure 6: Learned weights of the k=5 temporal convolution in ADA (left) and JCA(right) blocks for figure-skating. The upper plot refers to the second ADA(JCA) and the lower one refers to the fourth ADA(JCA). Due to limited space the first 25 channels have been visualized.

blocks. As a result, the coordination among the virtual super-joints/body parts would not be captured. In the next set of experiments we removed the whole JCA stream and ADA streams to validate their contribution in the action assessment. The drastic drop of the performance is because of solely relying on either appearance dynamics or joints coordination and motion features. We further changed the number of JCA and ADA blocks to confirm the optimality using two blocks. If we only use one JCA block the dependencies among the formed body parts/virtual super-joints would not be captured. Furthermore, given the small number of frames in short-term activities using three JCA and ADA does not seem beneficial since it leads to the increase of the number of parameters and overfitting problem. Next we evaluated the performance of the network when it only uses the appearance features of the backbone, by removing the JCA and ADA blocks and completely neglecting the pose features. We further evaluated the performance of the network in different numbers of JCAs and ADAs blocks and it turned out that using 4 blocks leads to the best performance. We finally removed the implicit temporal attention blocks, assuming that the all medium-term phases of the long-term action have the same contribution in the overall impression. It should be noted that this block is already deactivated in short-term action assessment. As listed in Tab.3, the network achieves its full potential when all of its components are utilized.

Finally, we visualize the learned weights of our model in Tab.6 following [7]. For brevity, we resorted to the temporal kernel with size of 5 and compared the transitions among the learned weights in each channel between two ADA(JCA) blocks. The upper plot in Fig.6 represents the learned weights of the second ADA(JCA) block. The rapid transitions among the weights in each channel demonstrate that this block is capturing the fine-grained temporal dependencies. On the other hand, as depicted in lower plot of Fig.6, the transitions among the 4th block learned weights are smoother, confirming the fact that this block is responsible for capturing coarse-grained temporal dependencies.

6. Conclusion

In this paper, we argue that evaluating the quality of an action requires incorporating appearance and pose features of the performance in both fine and coarse-grained temporal scales. To this end, we present a modular two-stream network that sits on top of extracted appearance and pose features of an action to assess it. The first stream is composed of a stack of JCA blocks that are responsible for evaluating the configuration of the joints and body parts throughout the performance. The other stream assesses the appearance dynamics of the action owing to its constituting ADA blocks. Empowering the network with more JCA and ADA blocks leads to capturing long-term coarse-grained temporal dependencies that represent overall impression of the program. Furthermore, we present a dataset of 7500 contortive sports images annotated with 2D human body pose. It is shown that training the pose estimator backbone on this dataset helps to have a more accurate pose estimation of sports actions that usually involve lots of unusual pose configurations. Our experimental evaluation demonstrates that our method achieves the state-of-the-art results on short-term action assessment in comparison to prior works. Moreover, the proposed modular network adapts simply to assess long-term actions by stacking more JCA and ADA blocks and outperforms the previous works on this task as well.

Acknowledgment

This work is supported by the University of Alberta Start-up grant, the NSERC Discovery Grant No. RGPIN-2019-04575, and the University of Alberta-Huawei Joint Innovation Collaboration grants.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] International Ski Federation (FIS). Judges handbook (snowboard and freeski). https://assets.fis-ski.com/image/upload/v1559714418/fis-prod/assets/Draft_SBFK_Judges_Handbook.pdf.

[6] Jibin Gao, Wei-Shi Zheng, Jia-Hui Pan, Chengying Gao, Yaowei Wang, Wei Zeng, and Jianhuang Lai. An asymmetric modeling for action assessment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[7] Noureldien Hussein, Efstratios Gavves, and Arnold W.M. Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[8] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[10] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR 2011*, pages 3361–3368. IEEE, 2011.

[11] Yongjun Li, Xiujuan Chai, and Xilin Chen. End-to-end learning for action quality assessment. In Richang Hong, Wen-Huang Cheng, Toshihiko Yamasaki, Meng Wang, and Chong-Wah Ngo, editors, *Advances in Multimedia Information Processing – PCM 2018*, pages 125–134, Cham, 2018. Springer International Publishing.

[12] 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.

[13] Mahdiar Nekoui, Fidel Omar Tito Cruz, and Li Cheng. Falcons: Fast learner-grader for contorted poses in sports. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[14] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European conference on computer vision*, pages 392–405. Springer, 2010.

[15] Jia-Hui Pan, Jibin Gao, and Wei-Shi Zheng. Action assessment by joint relation graphs. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[16] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1468–1476. IEEE, 2019.

[17] Paritosh Parmar and Brendan Tran Morris. 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.

[18] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[19] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 556–571, Cham, 2014. Springer International Publishing.

[20] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[22] International Skating Union. Bonus and deduction rules. https://www.isu.org/figure-skating/rules/sandp-handbooks-faq/17823-s-p-who-is-responsible-for-deductions-2019-20/file.

[23] International Skating Union. Program component chart for single skating. https://www.isu.org/isu-statutes-constitution-regulations-technical-rules-2/isu-judging-system/single-and-pair-skating/17596-program-component-chart-id-sp-2019-20/file.

[24] International Skating Union. Technical panel handbook for single skating. https://www.isu.org/isu-statutes-constitution-regulations-technical-rules-2/isu-judging-system/single-and-pair-skating/24781-tphb-single-skating-2020-21-final/file.

[25] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[26] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[27] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition, 2018.

[28] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.