ScoringNet: Learning Key Fragment for Action Quality Assessment with Ranking Loss in Skilled Sports

Yongjun Li[1,3], Xiujuan Chai[1,2], and Xilin Chen[1,3](✉)

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
{yongjun.li,xiujuan.chai}@vipl.ict.ac.cn, xlchen@ict.ac.cn

[2] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[3] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. Nowadays, scoring athletes' performance in skilled sports automatically has drawn more and more attention from the academic community. However, extracting effective features and predicting reasonable scores for a long skilled sport video still beset researchers. In this paper, we introduce the ScoringNet, a novel network consisting of key fragment segmentation (KFS) and score prediction (SP), to address these two problems. To get the effective features, we design KFS to obtain key fragments and remove irrelevant fragments by semantic video segmentation. Then a 3D convolutional neural network extracts features from each key fragment. In score prediction, we fuse the ranking loss into the traditional loss function to make the predictions more reasonable in terms of both the score value and the ranking aspects. Through the deep learning, we narrow the gap between the predictions and ground-truth scores as well as making the predictions satisfy the ranking constraint. Widely experiments convincingly show that our method achieves the state-of-the-art results on three datasets.

Keywords: Action quality assessment · Key fragment segmentation · Ranking loss

1 Introduction

The sports in which athletes have to complete the specified actions are called skilled sport. In these sports, the process can be divided into several stages and only key stages determine scores according to sports rules. Referees will give three kinds of scores including "Difficulty" score (fixing agreed-upon value based

This work was partially supported by 973 Program under contract No. 2015CB351802, Natural Science Foundation of China under contracts Nos. 61390511, 61472398, 61532018.

C. V. Jawahar et al. (Eds.): ACCV 2018, LNCS 11366, pp. 149–164, 2019.
https://doi.org/10.1007/978-3-030-20876-9_10

on specified action type), "Execution" score (judging quality of an action) and the final score. The final score can be obtained with the other two scores, such as adding them. These sports includes all diving events, all gymnastics events, equestrianism, synchronised swimming and so on.

In real life, it is very time consuming to train a qualified referee of skilled sports because they must go through long-term training to get familiar with all specified actions. Hence replacing manual scoring with an automatic scoring system is a trend in the future. On the other hand, the manual judgement is subjective. The automatic scoring system could be used as a trusted impartial opinion to avoid scoring scandals where the partiality of judges is questioned [1]. Nowadays, there have been some organizations trying applying automatic scoring systems in real sports competitions. For example, the international gymnastics federation (FIG) plans to introduce artificial intelligence technology to assess the quality of gymnastic in the 2020 Tokyo Olympics. Although there is a pressing need for the automatic scoring system, its application is impeded by two major obstacles. (1) A skilled sport usually contains a series of complicated motion fragments. So how to model a skilled sport video and obtain effective features are difficult. (2) The predictions should not only have small difference from the ground-truth scores but also satisfy the ranking constraint. How to make reasonable predictions to accomplish the two goals at the same time is non-trivial.

To overcome the aforementioned obstacles, we introduce the ScoringNet, as shown in Fig. 1, which consists of KFS and SP to realize action quality assessment in skilled sports. Inspired by the deep learning breakthroughs in the video domain [13,23–25] where rapid progress has made in feature learning, we build the ScoringNet on top of the 3D convolutional neural network to extract discriminative features. But the sport videos are usually untrimmed and have some noise fragments, extracting features from an untrimmed video is unreasonable. What's more, not all fragments of trimmed video have contribution to the score, such as the fragment where athletes are in run-up of diving. Hence we design the learning-based KFS to perform semantic video segmentation in the ScoringNet. Then the 3D convolutional neural network will only extract features from the key fragments which determine scores. In [5], Pirsiavash *et al.* first proposes to divide a video into fragments in action quality assessment, but they segment a video along the temporal dimension evenly. Most importantly, we fuse ranking loss into traditional loss function to make the predictions satisfy the score value constraint and the ranking constraint at the same time effectively. Additionally, the ScoringNet assesses athletes' performance more carefully by generating "Difficulty" score, "Execution" score and the final score instead of only the final score.

To summarize, our main contributions are as follows:

- We design the learning-based KFS in the ScoringNet to filter irrelative fragments and obtain key fragments by semantic video segmentation to ensure the effectiveness of features.

- The ranking loss are integrated with traditional loss function to form a powerful combined loss function which takes account of both the score value constraint and the ranking constraint.

2 Related Works

In action quality assessment, previous works fall into two categories, i.e. sports [5–7,9,12,14] and medicine [8,10,11] according to the application scenarios. Concerning the method, there are regression-based method [5,6,9,14,31] and classification-based method [8,10,12,30].

The regression-based method directly predicts a continuous score to evaluate the quality of an action. In [5], Pirsiavash *et al.* introduce a general learning-based framework to automatically assess actions' quality from videos. The framework uses a pose estimator to extract both low level and high level features of each frame. Then a support vector regression model (SVR) is trained to predict the final score. The results show that their approach is significantly better at assessing an action's quality than a non-expert human. Literature [7] proposes a recursive neural network that leverages growing self-organization for efficient learning of body motion sequences. The quality of an action is then measured in terms of how much a performed action matches the correct continuation of a learned sequence. In [31], they first use the similar way to segment videos into five stages. Then they employ P3D [24] to extract the body-pose features of fragments. Finally, SVR is implemented to regress the score.

While for the classification-based method, the quality of an action will be classified into different grades. In [8], Parmar and Morris examine the assessment of quality of large amplitude movement (LAM) actions designed to treat cerebral palsy (CP) in an automatic fashion. They transform both joint positions and angles to frequency domain by taking the discrete cosine transform (DCT). Each data variation then be used as the input feature vector for classifier. The task is regarded as a classification problem that whether the quality of an action is good or not. Literature [10] presents an automatic framework for surgical skill assessment. They first use Spatio-Temporal Interest Points (STIPs) [28] to get the feature from video data. Then they also transform motion data into frequency domain and finally the skill classification.

There are also some works combining the two methods. In [16], Chai *et al.* develop a system on sign quality evaluation with both classification and regression model. The system first determines whether a sign is the appointed one by the classification model. For the sign which passes the verification, the regression model will give a score.

Both [5] and [9] are comprehensive and pioneering works on action quality assessment in sports. However, their methods are traditional. Although [9] using deep learning to obtain features, they divide the whole process into feature extraction and score prediction instead of an end-to-end process. In score prediction, they assess the quality of an action only by the final score and their loss functions ignore the ranking constraint. In [9], they also divide a video into fragments, but the segmentation is even along the temporal dimension and not semantic.

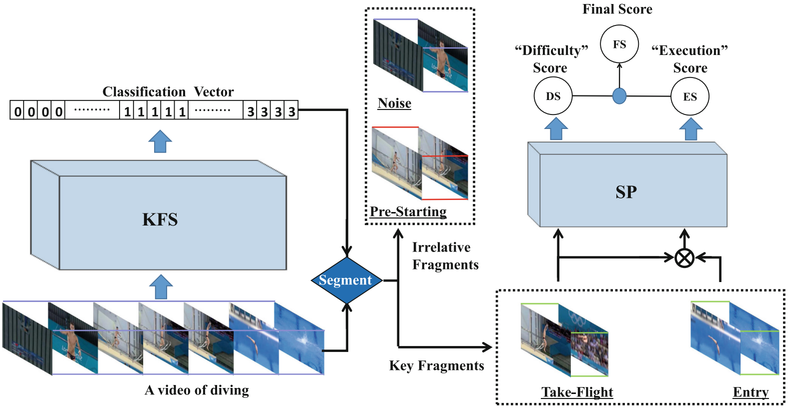


Fig. 1. The pipeline of the ScoringNet. DS, ES and FS represent "Difficulty" score, "Execution" score and the final score respectively. Pre-Starting, Take-Flight and Entry are the stages of a dive.

3 Our Method

In this section, we give detailed descriptions of the ScoringNet. The ScoringNet is composed of key fragment segmentation (KFS) and score prediction (SP). KFS is in charge of semantic video segmentation and obtaining key fragments. SP is employed to regress scores of the given sport videos. Firstly, the untrimmed video is sent into KFS to generate a classification vector whose value on each dimension represents the class. Then, all frames are classified into several fragments according to the vector. The irrelative fragments are dropped and the key fragments are sent to SP to predict "Difficulty" score and "Execution" score. Finally, the final score is obtained by fusion of the two scores.

3.1 Key Fragment Segmentation

Inspired by the idea that FCN [22] segments images by classifying each pixel, we design KFS for semantic video segmentation. KFS classifies all frames by sport stages and noise, the frames in the same class form a fragment. Only the key segments will be preserved. In general, a deep 2D convolutional neural network [19,26,27,29] is widely used for image classification. However, they may not generalize well to video-frame classification because of their limited access to temporal context. In a video, the contents of many frames at different times may be similar. The temporal context can help to distinguish these similar frames.

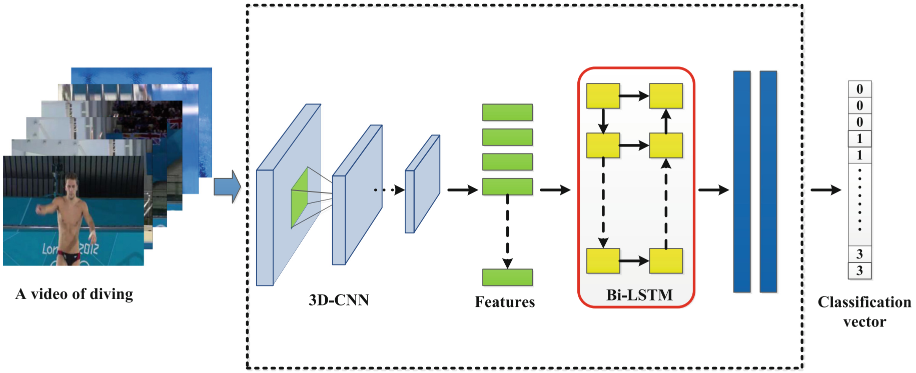


Fig. 2. The structure of KFS. The input and output of KFS are a video of diving and classification vector respectively.

Hence KFS consists of a 3D convolutional neural network, a Bidirectional LSTM (Bi-LSTM) network [21] and two fully connected layers, as Fig. 2. The 3D convolutional neural network processes short-temporal context and the Bi-LSTM is mainly for long-temporal context.

To clearly illustrate KFS. We can formulate KFS as Eq. 1,

$$C = T_{kfs}(X; W_{kfs}, B_{kfs}) \tag{1}$$

where $C = \{c_1, c_2, c_3 ... c_l\}, c_i \in \mathbb{R}^0, C \in \mathbb{R}^{l \times 1}$. c_i is the classification result of ith frame in the video. $X = \{x_1, x_2, x_3 ... x_l\}, x_i \in \mathbb{R}^{h \times w \times c}, X \in \mathbb{R}^{l \times h \times w \times c}$. x_i is the ith frame in the video. h, w and c are the height, width and channel of a frame respectively. l is the length of the video. T_{kfs} is a function representing operations on X in KFS. W_{kfs} and B_{kfs} are the hyper-parameters of T_{kfs}.

The Bi-LSTM we employ is the same as [21]. For the 3D convolutional neural network, we modify C3D [13] by the following two steps: **(1)** removing the fully connected layers of C3D and flatting the output of pool5 layer into a 4096-dim vector. **(2)** removing the temporal max-pooling operation to maintain the length of the video.

3.2 Score Prediction

In skilled sports, we just need to predict "Difficulty" score and "Execution" score. The final score can be obtained by them according to sports rules.

Thus SP, as Fig. 3, includes two branches and parallelly predicts the two kinds of scores. These two branches have the same structure but don't share weights. In each branches, a C3D extractor is adopted for the feature extraction, which consists of the first 12 layers of C3D. Once the feature is extracted, it will go through the score regressor which consists of two convolution layers (followed by Relu and $3 \times 3 \times 3$ Max-pooling) and three fully connected layers (followed by Relu and Dropout) to regress the score.

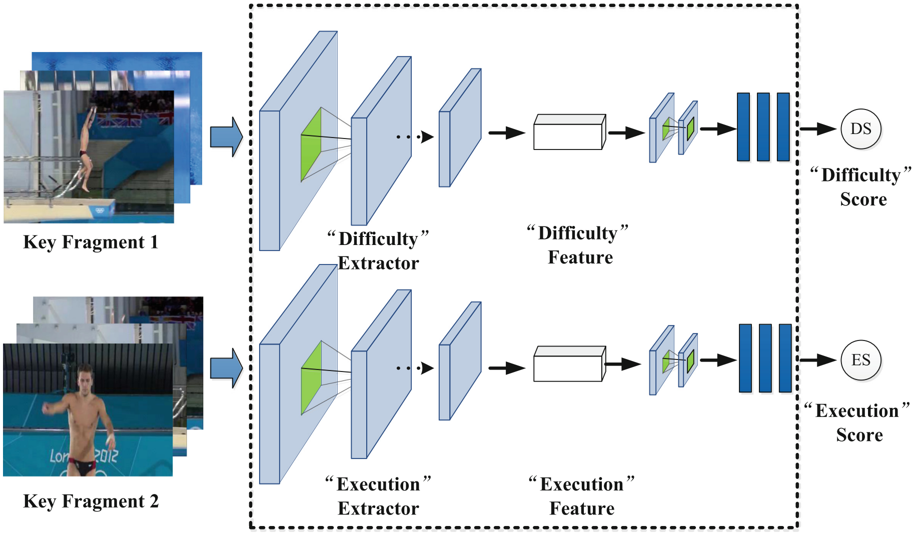


Fig. 3. The structure of SP. The input and output of SP are key fragments and scores respectively.

The SP can be formulated as Eqs. 2–4. For one branch predicting "Difficulty", we can formulate it as Eq. 2,

$$S_d = T_{spd}(D; W_{spd}, B_{spd}) \tag{2}$$

where $D = \{d_1, d_2, d_3...d_m\}, d_i \in \mathbb{R}^{h \times w \times c}, D \in \mathbb{R}^{m \times h \times w \times c}$. d_i is the ith frame in the key fragment which determines the "Difficulty" score. m is the length of the key fragment. T_{spd} is a function, with hyper-parameters W_{spd} and B_{spd}, which represents operations in SP. S_d is the "Difficulty" score prediction. Similarly, we can get another branch as Eq. 3,

$$S_e = T_{spe}(E; W_{spe}, B_{spe}) \tag{3}$$

where $E = \{e_1, e_2, e_3...e_t\}, e_i \in \mathbb{R}^{h \times w \times c}, E \in \mathbb{R}^{t \times h \times w \times c}$. e_i is the ith frame in the key fragment which determines the "Execution" score. t is the length of the key fragment. T_{spe} is a function, with hyper-parameters W_{spe} and B_{spe}, which represents operations in SP. S_e is the "Execution" score prediction.

Finally, according the rules, the finally score can be obtained by Eq. 4. In diving, F is the sum function of two scores while F represents quadrature function of them in vault.

$$S = F(S_d, S_e) \tag{4}$$

3.3 Loss Function

The loss function mainly consists of two parts in which one is for KFS and the other is for SP.

For KFS, we use standard categorical cross-entropy loss as the loss function, as Eq. (5),

$$L_{ce} = -\frac{1}{n \times l} \sum_{i=1}^{n \times l} (y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i)) \tag{5}$$

where n is the size of a batch. y_i and p_i are the ith ground-truth label and prediction label respectively. l is the length of the video.

For SP, MSE is widely adopted as the loss function to constrain score value, as Eq. (6),

$$L_{mse} = \frac{1}{2n} \sum_{i=1}^{n} (s_i - g_i)^2 \tag{6}$$

where g_i and s_i are the ith ground-truth score and prediction respectively. However, the performance of an automatic scoring system depends on whether the predictions satisfy both the score value constraint and the ranking constraint. So we add the ranking loss to make the predictions meet the ranking constraint. The ranking loss for a batch of data is defined as Eq. (7),

$$L_{rk} = \sum_{i=1}^{n} \sum_{j=1, j>i}^{n} RELU(-(s_j - s_i) sign(g_j - g_i) + \delta), \delta > 0 \tag{7}$$

where $RELU(\cdot)$ is a rectified linear unit activation and δ works as the margin for the ranking loss. When the predictions violate the ranking constraint, the ranking loss will generate a punishment term. On the contrary, the value of the ranking loss is zero. Further, the combined loss function for SP can be written by Eqs. (8) and (9),

$$L_d = L_{mse}^d + \alpha L_{rk}^d + \beta ||w^d||^2 \tag{8}$$

$$L_e = L_{mse}^e + \alpha L_{rk}^e + \beta ||w^e||^2, \alpha > 0, \beta > 0. \tag{9}$$

where L_d and L_e are the loss functions for two sub-architectures in SP respectively. $||w||^2$ is L2-regularization term, α and β are parameters used to balance these three terms.

Our final loss function includes two combined loss functions and the categorical cross-entropy loss as Eq. (10),

$$L = L_d + L_e + L_{ce} \tag{10}$$

4 Experimental Results

In this section, we evaluate our method on two sports (diving and vault) from three public datasets, i.e. Mit-Diving Dataset [5], UNLV-Diving Dataset [9] and

UNLV-Vault Dataset [9]. Firstly, we make a simple introduction of the three datasets. Secondly, the experimental configuration and the evaluation metric are given. Thirdly, we give the way of selecting key fragments and verify the correctness of this way. Then three experiments are conducted on the latter two datasets to evaluate our method. After this, the qualitative analysis is given to explain why our method works well. Finally, we compare our method with other state-of-the-art methods on the three datasets.

4.1 Datasets

Mit-Diving Dataset: This dataset contains 159 videos of London Olympic men's 10-m platform, each has roughly 150 frames. The size of all frames is 320×240. The annotation includes "Difficulty" score (varying between 2.7 and 4.1) and "Execution" score (varying between 6.0 and 29.0). The final score (ranging from 21.6 to 100.0) is determined by the product of "Difficulty" score multiplied by "Execution" score.

UNLV-Diving Dataset: This dataset, which is extended from the Mit-Diving Dataset, includes 370 videos from London Olympic men's 10-m platform. Apart from the number of videos, everything else is the same as the Mit-Diving Dataset.

UNLV-Vault Dataset: This dataset includes 176 videos with an average length of about 75 frames. The frame size is 320×240. The annotation includes "Difficulty" score (varying between 4.6 and 7.4) and "Execution" score (varying between 7.38 and 9.67). The final score (ranging from 12.30 to 16.87) is generated by adding "Execution" score and "Difficulty" score. These videos are shot from 5 international competitions and the view variation is quite large among these videos making it a more difficult dataset to score.

Since the Mit-Diving Dataset is a subset of the UNLV-Diving Dataset, we perform the detailed evaluation on the UNLV-Diving Dataset while only give the comparison results on the Mit-Diving Dataset. In experiments on the UNLV-Diving Dataset and UNLV-Vault Dataset, we follow the testing scheme in [9]. In [9], they perform a random data split and release the split result. The training/testing split is 300/70 on the UNLV-Diving Dataset and 120/56 on the UNLV-Vault Dataset.

4.2 Experimental Setting and Evaluation Metric

To mitigate the risk of over-fitting from the limited training data source, we pre-train the C3D in KFS and two extractors in SP with UCF-101 [15]. Further, all videos are augmented by shifting the start frame with a random number within [0, 5]. In the training process, the learning rate is initialed as $10e{-4}$ and decreases to its 0.45 every 600 iterations. The optimization algorithm is Adam [18]. In the loss function, α is set to 5 and 10 empirically for two diving datasets and the vault dataset respectively. Meanwhile β is set to 0.0005.

To measure the performance of our method, the spearman rank correlation (SRC, as Eq. 11) and mean euclidean distance (MED, as Eq. 12) are adopted as the our criterion.

$$SRC = 1 - \frac{6 \times \sum_{i=1}^{n} {h_i}^2}{n \times (n^2 - 1)} \tag{11}$$

$$MED = \frac{1}{n} \sum_{i=1}^{n} |s_i - g_i|, \tag{12}$$

where h_i is the difference between the two ranks of each observation. $|\cdot|$ represents absolute function. SRC is a nonparametric measure of rank correlation. We utilize it to measure the statistical dependence between the ranking of the predictions and the ground-truth scores. The larger the SRC, the higher the rank correlation. For MED, the smaller MED represents that the predictions are more close to the ground-truth scores.

In all experiments, we calculate "Difficulty" score SRC (D-SRC), "Execution" score SRC (E-SRC) and the final score SRC (F-SRC). For MED, there are also three kinds of MED (D-MED, E-MED, F-MED) accordingly.

4.3 Which Are Key Fragments

KFS is a learning-based method and need to be trained with videos which have labelled fragments. So we have to determine which fragments are key fragments.

In diving rules [3], the diving can be divided into 4 consecutive stages, namely the Pre-Starting, the Take-Off (including the starting position and the approach), the Flight and the Entry. We regard the Take-Off and the Flight as one stage called the Take-Flight because they determine "Difficulty" score together. Then, we divide each diving video into four fragments including all three stages and the noise. According to diving rules [3,4], the Take-Flight determines "Difficulty" score while the Take-Flight and the Entry contribute to "Execution" score together, so the Take-Flight and Entry are two key fragments while the Pre-Starting and the noise are irrelative fragments. Consequently, we employ the Take-Flight to predict "Difficulty" score while the Take-Flight and Entry are for "Execution" score prediction. Similarly, a vault video can be divided into four fragments which are the Pre-Flight, the Posting-Flight (including the Support), the Landing and the noise [2]. The Posting-Flight and Landing are two key fragments while the Pre-Flight and noise are irrelative fragments. According to vault rules [2], we use the Posting-Flight to predict "Difficulty" score. Meanwhile the Posting-Flight and Landing are for "Execution" score prediction.

To validate the way of selecting key fragments, we employ different ground-truth fragments to predict the two kinds of scores and compare performances of them. In this experiment, the loss function for SP is MSE for simplicity. As Table 1 shows, The Take-Flight and Take-Flight+Entry are the most suitable for "Difficulty" score prediction and "Execution" score prediction respectively on the UNLV-Diving Dataset. The similar results are reported in Table 2 on the UNLV-Vault Dataset. So, in the following subsections, we select key fragments according to sports rules.

Table 1. Validation on the way of selecting key fragments on the UNLV-Diving Dataset.

SRC	Pre-Starting	Take-Flight	Entry	Take-Flight+Entry
D-SRC	0.12	**0.64**	0.3208	0.56
E-SRC	0.08	0.52	0.82	**0.85**

Table 2. Validation on the way of selecting key fragments on the UNLV-Vault Dataset.

SRC	Pre-Flight	Posting-Flight	Landing	Posting-Flight+Landing
D-SRC	0.20	**0.53**	0.25	0.43
E-SRC	0.13	0.3475	0.38	**0.41**

4.4 Effect of Key Fragment Segmentation

In this subsection, we compare KFS against other networks and verify the effectiveness of KFS on the UNLV-Diving Dataset and the UNLV-Vault Dataset. Since there are no frame-level labels in these two datasets, we assign the label for each frame manually by stages and noise. Firstly, we compare KFS with ResNet50 and the Modified C3D respectively. The Modified C3D is obtained by two steps in Sect. 3.1. Figure 4 reports that the Modified C3D outperforms ResNet50 and KFS achieves the best performance. The results suggest both the short and the long contextual information can indeed help to classify frames. Additionally, we observe that the three results are very close on diving. The reason is that the contents of diving frames in different stages are distinct and easy to be classified. Further on, we show the accuracy on each class of two actions in Table 3.

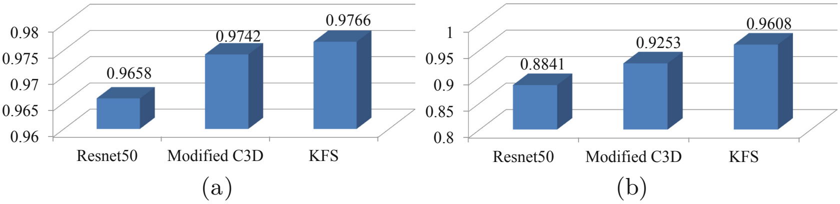


Fig. 4. Exploration of different networks for semantic video segmentation on the UNLV-Diving Dataset and the UNLV-Vault Dataset. (a) The accuracy of classification on UNLV-Diving Dataset. (b) The accuracy of classification on the UNLV-Vault Dataset.

Table 3. The accuracy on each class of two actions.

UNLV-Diving	Noise	Pre-Starting	Take-Flight	Entry
Acc	0.97	0.95	0.99	0.97
UNLV-Vault	Noise	Pre-Flight	Posting-Flight	Landing
Acc	0.96	0.81	0.97	0.97

Secondly, we verify the effectiveness of KFS. In this experiment, we drop KFS and take the untrimmed video as the input to SP directly. We also employ MSE as the loss function for SP in this experiment. Table 4 shows that the performance drops sharply without KFS. The results show that KFS has a strong positive effect for score prediction.

Table 4. Validation on the effectiveness of KSF on UNLV-Diving Dataset (represented by Diving) and UNLV-Vault Dataset (represented by Vault). −KFS refers to the experiment without KFS and +KFS refers to the experiment with KFS.

Method	D-MED	E-MED	F-MED	D-SRC	E-SRC	F-SRC
−KFS (Diving)	0.16	1.75	6.78	0.35	0.78	0.71
+KFS (Diving)	0.13	1.50	5.95	**0.57**	**0.82**	**0.79**
−KFS (Vault)	0.53	0.86	1.35	0.48	0.33	0.61
+KFS (Vault)	0.49	0.64	1.04	**0.51**	**0.38**	**0.68**

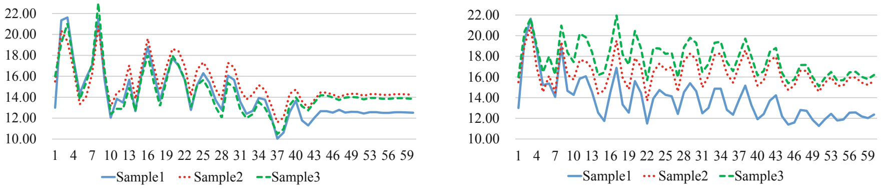
4.5 Evaluation on Different Loss Functions

We compare the performances of different loss functions which are MSE and the combined loss on the UNLV-Diving Dataset and the UNLV-Vault Dataset. This experiment is carried out with KFS. Table 5 reports the results. Our results show SRCs obtain the significant improvement while MEDs retain relatively small value, suggesting that the combined loss is capable of making the predictions satisfy the score value constraint and the ranking constraint as the same time.

In order to illustrate the effectiveness of the ranking loss explicitly, we choose three samples from UNLV-Diving Dataset (their ground-truth “Execution” scores: sample1 13.00, sample2 15.50, sample3 16.00) and record their “Execution” score predictions every epoch in the experiment with MSE (Fig. 5(a)) and the experiment with the combined loss (Fig. 5(b)) respectively. As shown in Fig. 5(a), only using MSE, three curves are interlaced. What’s more, the predictions of sample2 and sample3 are in wrong order as they have close ground-truth scores. When MSE are combined with the ranking loss, as shown in Fig. 5(b), these three curves are separated and the predictions are in right order from very early and keep it afterwards. That shows the powerful effectiveness of the ranking loss in the ranking constraint.

Table 5. Comparison of different loss functions on the UNLV-Diving Dataset (represented by Diving) and the UNLV-Vault Dataset (represented by Vault).

Method	D-MED	E-MED	F-MED	D-SRC	E-SRC	F-SRC
MSE (Diving)	0.13	1.50	5.95	0.57	0.82	0.79
The combined loss (Diving)	0.11	1.55	5.36	**0.79**	**0.86**	**0.84**
MSE (Vault)	0.49	0.64	1.04	0.51	0.38	0.68
The combined loss (Vault)	0.55	0.45	1.11	**0.57**	**0.57**	**0.70**



(a) (b)

Fig. 5. Illustration about effectiveness of the ranking loss (a) The predictions of three samples in the experiment with MSE. (b) The predictions of three samples in the experiment with the combined loss. In the two diagrams, Y-axis represents the value of the predictions and X-axis represents the number of epochs.

4.6 Qualitative Analysis

In order to further clarify the effect of KFS and Ranking loss. We make the following qualitative analysis. There are three models which are baseline model (M-b), the model with KFS (M-k) and the model with both KFS and ranking loss (M-kr). We predict the "Difficulty" score of a test video by these three models respectively. The ground truth score is 3.7 and three predictions are M-b:3.40, M-k:3.47 and M-kr:3.64 respectively. Then we show the feature maps (Fig. 6, the lighter the gray scale, the larger the activation value.) from conv1-layer of the three models. We infer that KFS removes redundant temporal information and makes the model focus on current frame (the legs in current feature map of M-b include the information of two frames before and after). Meanwhile, ranking loss is a powerful constrain and is able to eliminate the redundant spatial information, such as background.

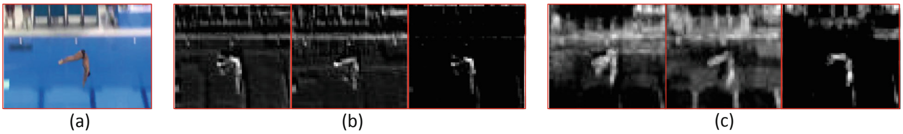


Fig. 6. (a) A frame from the test video. (b) The 8th feature maps of conv1-layer from three models (left to right: M-b, M-k, M-kr). (c) The 44th feature maps of conv1-layer from three models (left to right: M-b, M-k, M-kr).

4.7 Stability of Our Method

We perform 6 random data splits to exclude the influence of data split and verify the stability of our method. This experiment is carried out with KFS and the combined loss is the loss function for SP. Table 6 shows the average value and standard deviation over 6 random data splits on the UNLV-Diving Dataset and the UNLV-Vault Dataset respectively. The high average value and small standard deviation of SRCs suggest the strong stability of our method.

Table 6. Verification on the stability of our method on the UNLV-Diving Dataset (represented by Diving) and the UNLV-Vault Dataset (represented by Vault). The AVG and the STD mean average value and standard deviation respectively over the results of 6 random data splits.

Method	D-MED	E-MED	F-MED	D-SRC	E-SRC	F-SRC
AVG (Diving)	0.10	1.32	5.60	0.77	0.85	0.79
STD (Diving)	0.01	0.11	0.46	0.05	0.03	0.04
AVG (Vault)	0.48	0.81	1.08	0.61	0.50	0.68
STD (Vault)	0.07	0.16	0.12	0.04	0.07	0.05

4.8 Comparison with Other Methods

We compare our method with other state-of-the-art methods on the three Datasets. In previous methods, they only predict the final score, so we adopt F-SRC as evaluation metric. All results are shown in Table 7.

For the UNLV-Diving Dataset and the UNLV-Vault Dataset, we adopt the same split as [9]. They also test Pose+DCT with the same split. The results of Pose+DCT on the UNLV-Diving Dataset and the UNLV-Vault Dataset are extracted from [9].

For MIT-Diving Dataset, [5] uses 200 random data splits on this dataset and averages the results. The training/testing split is 100/59. The results of ConsISA

Table 7. Comparison of our method with other state-of-the-art methods on the three datasets

	UNLV-Diving	UNLV-Vault	MIT-Diving
Pose+DCT [5]	0.53	0.10	0.35
ConsISA [17]	-	-	0.19
ApEnFT [14]	-	-	0.45
C3D+LSTM [9]	0.27	0.05	0.36
C3D+SVR [9]	0.78	0.66	0.74
Ours (MSE+Ranking loss)	**0.84**	**0.70**	**0.78**

and ApEnFT are also extracted from [5]. Since the limited computing resources, our results are obtained by averaging the results from 6 random data splits.

As shown in Table 7, our method outperforms others tremendously and achieve the state-of-the-art results on these three datasets.

5 Conclusion

Automatically assessing an action's quality is a pressing need in skilled sports. But the two obstacles of extracting effective feature and making reasonable predictions prevent it being realized. In this paper, we present the ScoringNet, an end-to-end framework that aims to realize the action quality assessment in skilled sports. As demonstrated on three public datasets, our work has brought the state-of-the-art to a new level. This is largely ascribed to KFS and the combined loss, which overcome the two obstacles. The former provides an effective way to remove irrelative fragments and obtain key fragments to ensure the effectiveness of features, while the latter makes more reasonable predictions with satisfying the score value constraint and the ranking constraint as the same time.

References

1. List of Olympic Games Scandals and Controversies. https://en.wikipedia.org/wiki/List_of_Olympic_Games_boycotts. Accessed 26 Mar 2018
2. Vault. https://en.wikipedia.org/wiki/Vault_(gymnastics). 2.1.2. Accessed 4 June 2018
3. FINA Diving Rules. http://www.fina.org/sites/default/files/2017-2021_diving_16032018.pdf. D8.1.3. Accessed 12 Sept 2017
4. FINA Diving Rules. http://www.fina.org/sites/default/files/2017-2021_diving_16032018.pdf. APPENDIX 4. Accessed 12 Sept 2017
5. Pirsiavash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 556–571. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_36
6. Tao, L., et al.: A comparative study of pose representation and dynamics modelling for online motion quality assessment. Comput. Vis. Image Underst. **148**, 136–152 (2016)
7. Parmar, P., Morris, B.: Human motion assessment in real time using recurrent self-organization. In: 25th IEEE International Symposium on Robot and Human Interactive Communication, New York, USA, pp. 71–76 (2016)
8. Parisi, G., Magg, S., Wermter, S.: Measuring the quality of exercises. In: 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Florida, USA, pp. 2241–2244 (2016)
9. Parmar, P., Morris, B.: Learning to score olympic events. In: 30th IEEE Conference on Computer Vision and Pattern Recognition Work Shop, pp. 76–84. IEEE, Hawaii (2017)
10. Zia, A., Sharma, Y., Bettadapura, V., Sarin, E.L., Clements, M.A., Essa, I.: Automated assessment of surgical skills using frequency analysis. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9349, pp. 430–438. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24553-9_53

11. Baptista, R., Antunes, M., Aouada, D., Ottersten, B.: Video-based feedback for assisting physical activity. In: 12th International Conference on Computer Vision Theory and Applications, Porto, Portugal, pp. 430–438 (2017)
12. Carvajal, J., Wiliem, A., Sanderson, C., Lovell, B.: Towards Miss Universe automatic prediction: the evening gown competition. In: 23rd International Conference on Pattern Recognition, pp. 1089–1094. IEEE, Cancun (2016)
13. Du, T., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: International Conference on Computer Vision, pp. 4489–4497. IEEE, Santiago (2015)
14. Venkataraman, V., Vlachos, I., Turaga, P.: Dynamical regularity for action analysis. In: 26th British Machine Vision Conference, pp. 67.1–67.12. British Machine Vision Association, Swansea (2015)
15. Soomro, K., Zamir, A., Shah, M.: UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
16. Chai, X., Liu, Z., Li, Y., Yin, F., Chen, X.: SignInstructor: an effective tool for sign language vocabulary learning. In: 4th Asian Conference on Pattern Recognition, Nanjing, China (2017)
17. Le, Q., Zou, W., Yeung, S., Ng, A.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: 24th IEEE Conference on Computer Vision and Pattern Recognition, pp. 3361–3368. IEEE, Colorado Springs (2011)
18. Kingma, D.,Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv preprint arXiv:1412.6980 (2014)
19. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
20. Szegedya, C., et al.: Going deeper with convolutions. In: 28th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9. IEEE, Boston (2015)
21. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv preprint arXiv:1508.01991 (2015)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: 28th IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440. IEEE, Boston (2015)
23. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 30th IEEE Conference on Computer Vision and Pattern Recognition, pp. 4724–4733. IEEE, Hawaii (2017)
24. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: International Conference on Computer Vision, pp. 5534–5542. IEEE, Venice (2017)
25. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: 21st Conference on Neural Information Processing Systems, pp. 568–576. MIT Press, Montreal (2014)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: 19th Conference on Neural Information Processing Systems, pp. 1097–1105. MIT Press, Lake Tahoe (2012)
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
28. Laptev, I., Lindeberg, T.: On space-time interest points. Int. J. Comput. Vis. **64**, 107–123 (2005)

29. Huang, G., Liu, Z., Laurens, V.D.M., Weinberger, K.Q.: Densely connected convolutional networks. In: 30th IEEE Conference on Computer Vision and Pattern Recognition, pp. 2261–2269. IEEE, Hawaii (2017)
30. Doughty, H., Damen, D., Mayol-Cuevas, W.: Who's better? Who's best? Pairwise deep ranking for skill determination. In: 31st IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Salt Lake City (2018)
31. Xiang, X., Tian, Y., Reiter, A., Hager, G.D., Tran, T.D.: S3D: stacking segmental P3D for action quality assessment. In: IEEE International Conference on Image Processing. IEEE, Athens (2018)