Complex & Intelligent Systems (2023) 9:1963–1974
https://doi.org/10.1007/s40747-022-00892-6

ORIGINAL ARTICLE

Gaussian guided frame sequence encoder network for action quality assessment

Ming-Zhe Li[1] · Hong-Bo Zhang[1] · Li-Jia Dong[1] · Qing Lei[2] · Ji-Xiang Du[3]

Received: 17 August 2021 / Accepted: 12 October 2022 / Published online: 27 October 2022

Abstract

Can a computer evaluate an athlete's performance automatically? Many action quality assessment (AQA) methods have been proposed in recent years. Limited by the randomness of video sampling and the simple strategy of model training, the performance of the existing AQA methods can still be further improved. To achieve this goal, a Gaussian guided frame sequence encoder network is proposed in this paper. In the proposed method, the image feature of each video frame is extracted by Resnet model. And then, a frame sequence encoder network is applied to model temporal information and generate action quality feature. Finally, a fully connected network is designed to predict action quality score. To train the proposed method effectively, inspired by the final score calculation rule in Olympic game, Gaussian loss function is employed to compute the error between the predicted score and the label score. The proposed method is implemented on the AQA-7 and MTL–AQA datasets. The experimental results confirm that compared with the state-of-the-art methods, our proposed method achieves the better performance. And detailed ablation experiments are conducted to verify the effectiveness of each component in the module.

Keywords Action quality assessment · Frame sequence encoder network · Gaussian loss function · Regression analysis

Introduction

Action Quality Assessment (AQA) aims to develop an automatic way to evaluate the quality of action performance in the video. Over the past few years, people have discovered its valuable practicability and proposed many methods to facilitate practical tasks of AQA [17,23–25,40], especially in sports video scoring [8,16,21,22,30,33,36,44], skill assessment [6,9,11,26,31], and medical rehabilitation [19,20,29,35,38]. Therefore, developing an AQA system that can generate accurate and objective scores for sport events has become an urgent demanding.

AQA in sport events specifically refers to building a system that can score the performance of athletes like a referee. The input of the AQA system is a sport video, such as diving, and the output is the athlete's performance score. However, how to evaluate the sport action performance more accurately and effectively is still a question and is worthwhile exploring. Unlike action recognition [1,10,18,32,37,39,41,43] in which rare keyframes are enough to classify a video action sequence, the difficulty of AQA is that people must obtain the subtle differences between the continuous motions which determine the level of final score. Therefore, the integrated temporal semantic plays a significant role in AQA task.

In most of previous methods for scoring sports video [25,30,40], 3D convolutional neural network (3D CNN) was applied as the feature extraction model to learn the temporal semantic in the video. In general, in these methods, short clip sampled from original video are used as the input of 3D CNN, and then, a regression model [17,23,28] is trained to predict the final scores. Due to the randomness of video frame sampling strategy, the features extracted from short clip may ignore critical action information and other factors that affect the score of an athlete. The performances of these AQA methods are unstable and unsatisfactory. As illustrated in Fig. 1, some keyframes of diving video are shown. The frames in the yellow box represent small splash, while the

✉ Hong-Bo Zhang
zhanghongbo@hqu.edu.cn

[1] School of Computer Science and Technology, Huaqiao University, Xiamen 361000, China

[2] Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen 361000, China

[3] Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen 361000, China

frames in the red box represent large splash. For diving sport assessment, how well you perform is just as important as the size of the splash, and similar characteristics are also reflected in other sports categories [30,33,36]. In the sampling process, if the f_4, f_6 frame with small splash size is selected as the input clip, while f_5 frame is neglected, it will mislead the evaluation of the final score. Therefore, to obtain the stable and intact temporal semantic feature, inspired by action detection and recognition method [15] which used the feature sequence of static video frame as input rather than video clip, a frame sequence-based temporal encoder convolutional network (TECN) is applied to capture the action quality feature in the proposed AQA method.

To train the AQA network effectively, it is necessary to define a reasonable loss function to optimize the weight learning of the network. For most traditional AQA methods, mean square error (MSE) loss function [21,40] was generally used in regression analysis. However, according to the objective scoring rules in sport events, the final score of each action is the average of the subjective scores provided by multiple referees. There is a reasonable deviation margin between the referee scores and the final score. Therefore, to model the statistical characteristics of score label, the action quality score is modeled as random variables and the label is regarded as the mean of the distribution of the action quality score in this work. The goal of network optimization is to make the score predicted by the model obtain the maximum probability in the score distribution with label mean. Specifically, we assume that the action quality score label is subject to Gaussian distribution, and a Gaussian loss function is utilized to guide the network training in AQA tasks. Finally, the framework of the proposed method is shown in Fig. 2.

In the proposed method, Resnet model [13] is applied as feature extracted backbone to obtain the appearance feature of each frame in the video. And frame sequence-based TECN which is built by spatial convolutional and temporal pooling [15] is used to obtain the temporal semantic features. Since we employ features that extracted from fully video frames as input information, the overall temporal feature is more stable and robust. And then, a fully connected regression network is designed in this work to predicted the final action quality score. In the training stage, the proposed Gaussian loss function is used to calculate the difference between the predicted scores and the ground-truth scores. In the Gaussian loss function, the label score is regarded as the mean of the Gaussian function, and the standard deviation is set as a hyper-parameter which is fine-tuned in the experiment.

In addition, the final score of an Olympic event is usually sophisticated. When some detailed score labels are available, such as execution score provided by each referee and the difficulty degree of the diving, we further design the different training strategies with these complex score labels, including overall score-based training strategy and execution score-based training strategy.

The contributions of our paper are summarized as follows:

- To obtain higher level and integrated feature information, we propose a frame sequence-based TECN to learn action quality feature for AQA task.
- We propose a novel Gaussian loss function for regression analysis, which optimizes certain defects of traditional MSE loss.
- We conduct experiments on the AQA-7 [24] and MTL–AQA [25] datasets. The experimental results show that the proposed method achieves the better performance compared with the existing methods. Ablation experiments are performed to verify the effectiveness of the various components in the proposed method.

The remainder of this paper is organized as follows. “Related work” section introduces the related works, and “Approach” section describes the algorithms. “Experiments” section presents and discusses the experimental results, and “Conclusion” section concludes the paper.

Related work

In this section, we first introduce action recognition methods in videos, especially the approaches of spatial–temporal structure modeling. Based on that, we provide a detail review about related works of AQA tasks in the practical problems, including pose-based methods and appearance-based methods.

Spatial–temporal feature learning methods

How to transform video information into higher spatio-temporal semantic representation is an important issue to improve the performance of action recognition. Many video feature learning methods were proposed in the field of action understanding research. 3D CNN models, like C3D [34] and I3D [3], have made remarkable achievements in the field of action recognition that directly learned the spatio-temporal feature from video clips by 3D convolution operation. [10] involved a multi-path network that the slow pathway was designed to grasp spatial semantics, and the fast pathway was used to capture changing motion at temporal scale. Based on that promoted, [43] proposed a feature-level temporal pyramid network, which utilized inflated 3D ResNet-50 [12] as backbone. To learn long-range temporal information, [37] used a sparse temporal sampling strategy and fused the categories of the sampled segments.

All above appearance-based action recognition methods explore the higher semantic information across temporal

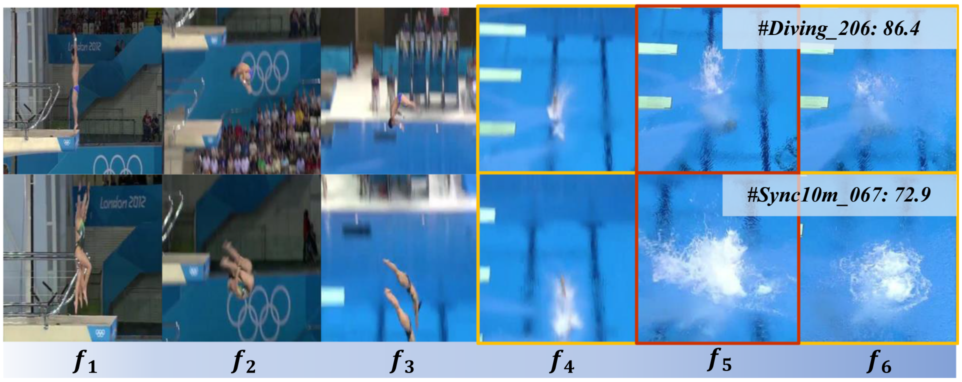


Fig. 1 Example of the critical frames in diving

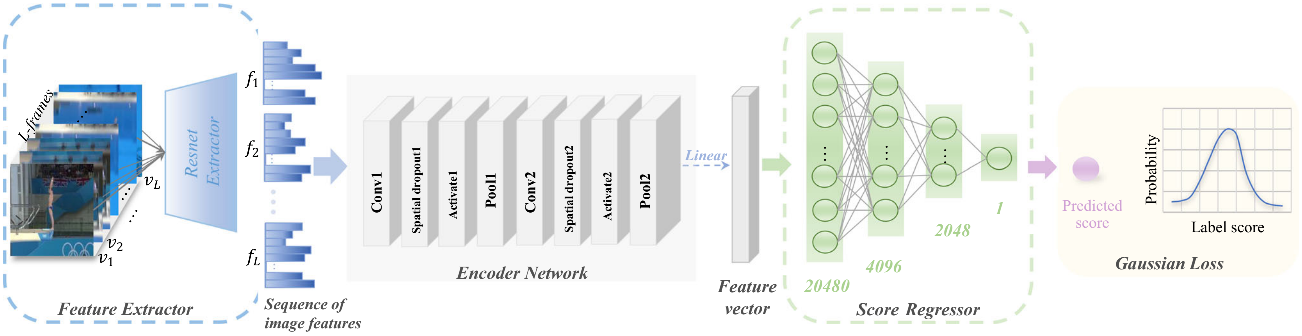


Fig. 2 Pipeline of the proposed Gaussian guided frame sequence encoded network. It contains four parts: feature extractor, temporal feature encoder network, score regression network, and loss function module

dimension. It is also the basic problem in AQA task which investigate the temporal structure modeling to assess the performance of the video.

Action quality assessment

In recent years, many attempts have been dedicated to make breakthrough on the AQA research [5,17,23–26,33,44]. Most existing algorithms can be roughly divided into two categories according to the form of input information: pose-based methods and appearance-based methods. Pose-based methods [21,22,28] tackle AQA works by extracting the action quality feature from human pose sequence. [28] proposed a framework for learning the spatio-temporal pose features from pose sequence. In this work, discrete cosine transform (DCT) was applied to extract the action features and a linear Support Vector Regressor (SVR) was used to predict the score. [22] tackled the detailed joint motion based on the conjoint relations and integrated the joint commonality module with the joint difference module for joint motion learning. The learned features were fed into the score regressor to predict the score. However, it is difficult to represent other critical factors completely that affect scores in AQA research only using the human pose information. For example, the final splash concerns a significant contribution about the score in diving, which serves as the important visual clues from background.

The appearance-based approach aims to use information learned from visual representations to score action quality. [23] performed score regression based on C3D spatio-temporal features and used SVR or LSTM to regress the score. To explore how fragment-level features affect scoring in a long video, [17] segmented each video into several fragments and extracted fragment-level features by stacked C3D networks, while the correlation between each fragment is ignored. Furthermore, [25] proposed a largest AQA dataset to date, utilized C3D to learn features, and optimized the proposed model using three loss functions which account for incorporate three related tasks (i.e., fine-grained action recognition, commentary generation, and AQA score prediction) to provide better feedback to the scoring task. [40] applied P3D model on each segment of a video and then fused the stage-wise features to obtain the score. However, due to the random characteristics of video sampling in the input stage of each prediction, the estimation result is unstable. To learn more discriminative representations for videos, [44] presented a hybrid dynamic–static context-aware attention network which learn the video dynamic information and the static postures, respectively, using the two streams mode. The final scores were generated by leveraging the combination features of these two streams. [33] devoted to explore the inherently ambiguity in the score labels thus proposed an uncertainty-aware score distribution learning approach, which described different evaluated scores as probability distribution and generated a corresponding predicted one, and then optimized the model by the Kullback–Leibler divergence between these two distributions.

Compared with previous work, we propose a modular network for AQA research that is applicable to most events in sport competitions. We argue that in sports, by employing full video with all critical moments, it performs better than simple sampling of video frames in assessing action quality. A comparison between the proposed method and the existing AQA methods is summarized in Table 1.

Approach

In this work, we develop a novel network for action quality assessment. In this section, we outline the proposed AQA framework in Olympic events, which generates the action quality score of the athletes from the input sport videos. As shown in Fig. 2, our overall framework can be divided into four parts: full-video frame feature extraction, frame sequence-based temporal feature encoder, fully connected network for score regression, and Gaussian loss function-based learning mechanism. In this work, the AQA problem in sports event can be defined in Eq. (1)

$$S = TECN(V), \tag{1}$$

where V represents the input video, $TECN(\cdot)$ means the proposed algorithm, and S represents the predicted action quality score.

Full-video frame feature extraction

Given an input video $V = \{v_1, v_2, ..., v_L\}$, where v_i represents i-th frame in a video and L is the length of input video. The image sequence is fed into a Resnet model [13] pretrained on large-scale image classification dataset [4] for offline feature extraction. Different from these previous works [8,33,40] which used sampled or segmented video clip as input, the proposed method adopts the unsegmented frames as the input of AQA network to obtain more comprehensive feature representation. In traditional Resnet network, the last two pooling and fully connected layers correspond to the original image classification task. Therefore, they are removed and the output of the remaining network is used as the image feature in this work. In the experiment, Resnet-50 [13] structure is adopted. After image feature extraction, the input video is transferred into the image feature sequence, as defined in Eq. (2)

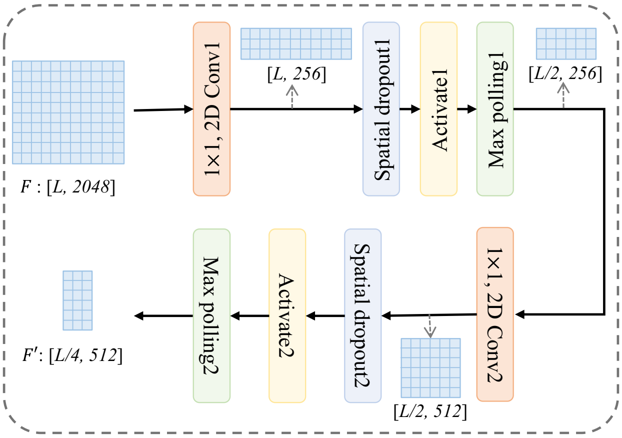


Fig. 3 The structure of TECN

$$\begin{aligned} F &= \{f_1, f_2, \ldots, f_L\} \\ f_i &= R(v_i), \end{aligned} \tag{2}$$

where F denotes the obtained feature sequence, f_i represents the feature of ith frame in the video, and $R(\cdot)$ represents the feature extraction network. The dimension of the feature extracted from Resnet is 2048. Due to the limitation of computing capability and excessive parameters will lead to memory consumption, Resnet network is only used as the offline feature extractor without training in the proposed AQA model.

Frame sequence-based temporal encoder

Temporal semantic information is significant feature for motion analysis, such as diving, skating, and gym vault. The task of frame sequence-based temporal encoder convolutional network in this study is to capture the sequence relationship of a continuous action to obtain higher level temporal and spatial semantic information [10,43]. Consider the frame-level feature sequence $F^{L\times2048}$ that extracted from Resnet as input, a temporal encoder convolutional network is applied as the encoder module to learn the temporal information of action. The structure of TECN is shown in Fig. 3.

In TECN, each encoding block is composed of an 1 × 1 temporal convolution, specific activation function and Max pooling across temporal series. We utilize $E_c \in \mathbb{R}^{F_c\times T_c}$ to depict the structure of encoder, where F_c is the number of convolutional filters in the cth layer and T_c is the number of corresponding time steps. For the encoder network, the inter-frame temporal convolution is expressed as Eq. (3)

$$E_c = h(W_c \otimes p(E_{c-1}) + b), \tag{3}$$

Table 1 Comparisons between the proposed method and the existing AQA methods

Method	Year	Input	Feature extractor	Score predictor	Learning strategy
[28]	2014	S	DCT	SVR	Linear
[42]	2018	S	ST-GCN	FCN	MSE loss
[22]	2019	S+V	Joint relation graph + I3D	FCN	MSE loss
[21]	2021	S+V	Joint stream + appearance stream	FCN	MSE loss
[17]	2018	V	Temporal segmentation + stacked C3D	FCN	MSE + ranking loss
[40]	2018	V	Temporal segmentation + P3D	FCN	MSE loss
[24]	2019	V	C3D+LSTM	FCN	MSE loss
[25]	2019	V	Multi-stage C3D + average aggregation	FCN	Multi-task learning
[33]	2020	V	Multi-stage I3D + temporal pooling	Softmax	KL loss
[11]	2020	V	Asymmetric interaction module	FCN	MSE loss
[36]	2020	V	Spatial + temporal convolution + attention	FCN	MSE + correlation loss
[44]	2020	V	Static and dynamic stream + context-aware attention	FCN	MSE loss
[5]	2021	V	Temporal segmentation+P3D	Multi-substage FCN	MSE loss
Ours	2022	V	Image feature + Temporal encoder	FCN	Gaussian loss

FCN means fully connection network, *S* represents skeleton data, and *V* represents video data

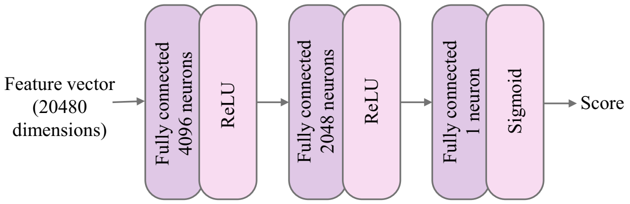


Fig. 4 The architecture of score regression network

where the collected convolution filters are denoted as $W = \{W_i\}_{i=1}^{F_c}$ for $W_i \in \mathbb{R}^{d_c \times F_{c-1}}$, $b \in \mathbb{R}^{F_c}$ for being a corresponding bias vector, $h(\cdot)$ being the activation function, d_c being the filter length at the cth layer, $\otimes$ being the convolution operator, and $p(\cdot)$ being the temporal pooling operation.

Score regression

Most of the existing methods [8,40] consider AQA task as a regression problem, this process is roughly as aggregating the features extracted by convolutional neural network [3,12,34] to form a video-level feature representation, and perform regression analysis on the feature vector. As shown in Fig. 4, a score regression network is defined by three-layer fully connected network in this work, and each layer consists a fully connected operation and a non-linear activation layer.

Specifically, the output feature of the last layer in the encoder model is flattened into a vector as the input of score regression network, and then passes through 4096, 2048, and 1 neurons subsequently. For purpose of eliminating the influence of dimensions on the final predictions, making different features comparable and improve the convergence speed of the model, we add a sigmoid activation function to normalize the output of the last FC-layer. This calculation process is expressed as Eq. (4)

$$S = Sigmoid(FC(E')), \tag{4}$$

where S is the predicted score of the pipeline, $FC(\cdot)$ is descending FC-layer operation, and E' is the output of the temporal encoder network. The final prediction is deemed as an athlete's score.

Gaussian loss function

According to the theory of deep learning [2], in the training stage, we have to tell the model what we expect the network to minimize to guide the parameters' learning. Therefore, an effective loss function should be considered to train the proposed network. MSE loss is one of the most extensively used and most straightforward loss function for regression network. It is also widely used in many AQA tasks. However, MSE loss is oversimplified that ignores the characteristic of the sports scoring and the performance of the model trained by this loss function should be further improved.

According to the objective scoring rules in most sports event, the final score is generated by multiplying the execution score and the difficulty degree. And the execution score is the average score of various referees' score. For example, the overall score of diving is based on the referee's score (how was that done) and the difficulty degree (what was done) of an athlete. There are uncertainty between the referee's evaluation score and the label score. To solve the shortcomings of previous loss functions used in AQA task, inspired by [33] using uncertainty-aware score distribution learning for AQA, we take the score of athletes as random variables which obey the normal distribution. The score label provided by the dataset is regarded as the mean of the distribution. The probability density function of the quality score is written as follows:

$$f(S) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(S-\mu)^2}{2\sigma^2}}, \tag{5}$$

where μ and σ denotes the mean and standard deviation, respectively. We assume that $f^{'}(S) = \mathrm{e}^{-\frac{(S-\mu)^2}{2\sigma^2}}$, while consider it in terms of our practical task, μ is regarded as ground-truth score and S is the predicted score by the proposed method. We expect that the score predicted by the network corresponds to the maximum probability of this distribution. Therefore, the overall loss function can be calculated as Eq. (6)

$$\begin{aligned} \mathcal{L}_{gau} &= \frac{1}{N}\sum_{n=1}^{N}(1 - f^{'}(S)) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left(1 - \mathrm{e}^{-\frac{(S_n-\mu_n)^2}{2\sigma^2}}\right), \end{aligned} \tag{6}$$

where σ serves as the hyper-parameter that need to be fine-tuned in the experiment. When the predicted score is exactly equal to the ground-truth score, identically $S_n = \mu_n$, the loss takes the minimum value.

In addition, some datasets will provide more detailed labels for sports event assessment. For example, MTL–AQA dataset [25] provides the overall score label, seven execution scores from judges, and difficulty degree. The overall execution score is calculated by the sum of the three referee scores after removing the two highest referee scores and two lowest scores. Therefore, two different training scenarios based on the aforementioned scoring rules are built to verify the effectiveness of the proposed method in this work.

Overall score training strategy (OS)

In the training scenario where only overall score is counted as the label, the influencing factors of execution score and difficulty degree have not been considered, so that the output of the network is directly used as the overall predicted score to calculate loss with the score label. Therefore, the loss function can be expressed as Eq. (7)

$$\mathcal{L}_{gau}^{os} = \frac{1}{N}\sum_{n=1}^{N}\left(1 - \mathrm{e}^{-\frac{(S_n^{os}-\mu_n^{os})^2}{2\sigma^2}}\right), \tag{7}$$

where S_n^{os} denotes the predicted overall score and μ_n^{os} denotes the ground-truth overall score of nth video sample.

Execution score training strategy (ES)

While sophisticated score labels are available, we consider a training scenario closer to the objective rules to explore the potential influence factors in the model. In this scenario, execution score is used as the training label to optimize each component of the model. Therefore, in the inference stage, it is necessary to multiply the output score by the difficulty degree to obtain the overall score. We can describe the calculation of loss function as Eq. (8)

$$\mathcal{L}_{gau}^{es} = \frac{1}{N}\sum_{n=1}^{N}\left(1 - \mathrm{e}^{-\frac{(S_n^{es}-\mu_n^{es})^2}{2\sigma^2}}\right), \tag{8}$$

where S_n^{es} is the predicted execution score and μ_n^{es} is the ground-truth execution score of nth video sample.

Experiments

We evaluate the proposed method on the AQA-7 [24] and MTL–AQA [25] datasets, which are the most commonly used in AQA task. The comparison results are applied to verify the effectiveness of the proposed method. Moreover, some ablation experiments are included to further illustrate the effectiveness of each component proposed in this work.

Datasets and experiment settings

AQA-7 dataset [24]. This dataset contains 1189 videos from seven sport categories captured during Summer and Winter Olympics: 370 videos from single diving-10 m platform competition, 176 videos from gymnastic vault competition, 175 videos from big air skiing competition, 206 videos from big air snowboarding competition, 88 videos from synchronous diving-3 m springboard competition, 91 videos from synchronous diving-10 m platform competition, and 83 videos from trampoline competition. The length of the trampoline video is 650 frames, while all of the other sports videos in this dataset are 103 frames. Refer to the previous work [24,33] and to make a fair comparison, the trampoline videos are excluded in the experiment, that is, we evaluate the performance of the proposed method on the remaining six competitions. In the experiment, we build the AQA model for each sport category independently. The division of training set and testing set follows the official setting. AQA-7 dataset only provides the overall score label for each video.

MTL–AQA dataset [25]. As the largest dataset in AQA field, MTL–AQA dataset contains 1412 videos that only focus on Diving category. The dataset is collected from 16 different events instead of the single main event, such as 2012 Olympics Men's 10m Platform Diving competition in AQA-7 dataset. This dataset covers various view and different camera angles, includes the 10m Platform as well as 3 m Springboard, both male and female athletes, and individual or pairs of synchronized divers. For each video, MTL–AQA dataset not only provides the overall score label, but also offers the difficulty degree of diving and the execution score of each referee. According to the previous work [25], in this dataset, 1059 videos are applied as the training set, and 353 videos are set as the testing set.

Evaluation metric. To be comparable with previous AQA methods, Spearman's rank correlation (SRC) is utilized to evaluate the correlation between two statistical variables (i.e., ground-truth scores and predicted scores). When the two variables are completely monotone correlated, the Spearman's correlation coefficient is +1 or −1. The calculation can be expressed as Eq. (9)

$$\rho = \frac{cov(r_X, r_Y)}{\sigma_{r_X}\sigma_{r_Y}}, \tag{9}$$

where ρ denotes the Spearman's correlation coefficient, r_X and r_Y are the rank variables of the predicted score and label score, respectively, $cov(r_X, r_Y)$ is the covariance of the rank variables, and σ_{r_X} and σ_{r_Y} are the standard deviations of the rank variables.

We conducted experiments on the six sports categories in the AQA-7 dataset. To make a comprehensive comparison with other methods, the average correlation coefficient of the six categories is necessary. Because of the non-additive character of correlation coefficients, they must be converted into a new variable, named the Fisher's z value [7], it is available for measuring the overall average correlation coefficient ρ between variables X and Y.

Implementation details. PyTorch [27] is used to implement the proposed model and the proposed method is trained on single Nvidia RTX 3090 GPU. The Resnet network is pretrained on ImageNet [4] dataset as the feature extractor. All the video frames are resized to 224 × 224. We adopt Adam [14] optimizer to train our network with initial learning rate of 0.0001 and the decay rate is set as 0.5.

In our experiments, to prevent over-fitting and unstable factors, all of the score labels are normalized into the interval of 0 to 1. For overall scores, we utilize min–max normalization as below

$$S' = \frac{S - \min(S)}{\max(S) - \min(S)}. \tag{10}$$

Here, S' is the normalized score, and $\min(S)$ and $\max(S)$ denote the minimum and maximum score in the training set. Since the execution score is ranging from 0 to 30, we normalize them by dividing 30 directly. To make a fair comparison with other methods, in the experiment, all of the experimental results are calculated based on the overall score.

Results on AQA-7 dataset

Comparison with state-of-the-art methods. We first compare the performance of the proposed method with other state-of-the-art methods on AQA-7 dataset. Since the lack of difficulty degree labels, only the normalized overall score label is used to train our network on AQA-7 dataset. The results can be seen in Table 2, and the proposed method outperforms the existing AQA methods in all sports categories except Snowboard. We argue that the reason is that in the Snowboard sport, the camera view and displacement of athletes are diversified, and it is difficult to capture accurate human region and significant motion features, resulting in poor score prediction. This result also means when the resolution of object is too small or the distance between camera and object, the proposed method is hard to learn the key features of an athlete under such condition, making the prediction poor. This is also the limitation of the proposed model. In the future, it needs to further explore an AQA model that is more suitable for small-scale objects in practical application.

But even so, the result of the proposed method on Snowboard has exceeded the JRG method [22] which uses the excessive optical flow information. It is worth noticing that in Skiing category, we gain the most significant improvement about 6.79%, and there have been varying degrees of growth in other categories similarly. The average correlation of the TECN approach is improved by 3.64% compared with the state-of-the-art methods, which verifying the strong effectiveness of the method proposed in this paper.

Ablation study on different loss functions. The performances of four different loss functions are concluded in Table 3. We can see that even conducting with the MSE loss function or L1 loss function, the proposed TECN model could achieve 83.07% and 82.68% Spearman's correlation coefficient, respectively, better than other previous methods as mentioned in Table 2 with the same loss function [23,40]. We also use a sum of MSE and L1 loss between the predicted score and ground-truth score as [25] suggested. Compared with these loss functions, the proposed method trained by Gaussian loss performs the best SRC in all sport categories and achieves the best average correlation coefficient. The SRC is higher than MSE loss, L1 loss, and a combination of them by 1.97%, 2.36%, and 1.19%, respectively. The ablation results indicate that our Gaussian loss could boost the training of regression model in AQA field.

Table 2 Comparisons of action quality assessment accuracy on the AQA-7 dataset

Network	Diving	Gym vault	Skiing	Snowboard	Sync. 3 m	Sync. 10 m	Avg. corr.
Pose + DCT [28]	0.5300	–	–	–	–	–	–
ST-GCN [42]	0.3286	0.5770	0.1681	0.1234	0.6600	0.6483	0.4433
C3D-LSTM [24]	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165
C3D-SVR [24]	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937
AIM [11]	0.7419	0.7296	0.5890	0.4960	0.9298	0.9043	0.7789
JRG [22]	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849
USDL [33]	0.8099	0.7570	0.6538	**0.7109**	0.9166	0.8878	0.8102
MSRM [5]	0.8129	–	–	–	–	–	–
EAGLE-Eye [21]	0.8331	0.7411	0.6635	0.6447	0.9143	0.9158	0.8140
SCN+ATCN [36]	0.8500	0.7600	–	–	–	–	–
Ours(TECN)	**0.8604**	**0.8156**	**0.7314**	0.5755	**0.9432**	**0.9417**	**0.8504**

Bold values represent the best performance

Table 3 Study on different loss function. We use consistent backbone with only the type of loss function changed

Loss function	Diving	Gym vault	Skiing	Snowboard	Sync. 3 m	Sync. 10 m	Avg. corr.
MSELoss	0.8360	0.7852	0.7223	0.5559	0.9351	0.9274	0.8307
L1Loss	0.8343	0.8034	0.7042	0.5505	0.9218	0.9304	0.8268
MSELoss + L1Loss	0.8498	0.8155	0.7156	0.5389	0.9381	0.9321	0.8385
Gaussian loss	**0.8604**	**0.8156**	**0.7314**	**0.5755**	**0.9432**	**0.9417**	**0.8504**

Bold values represent the best performance

Results on MTL–AQA dataset

Comparison with state-of-the-art methods. We further verify the proposed TECN model on MTL–AQA dataset. The comparison results are shown in Table 4. MTL–AQA dataset provides the score of each referee, execution score, and difficulty degree labels; therefore, in this experiment, two different scenarios (OS and ES) are included. As we can see that the proposed TECN with ES model obtains a competitive result compared with the existing methods. To some extent, it proves that exploring integrated frames' sequence and sophisticated temporal representations have successfully bring promotion and new inspiration on the action quality assessment task.

Comparison of different loss functions with various training strategy. To verify the affection of the detailed execution score label mentioned in Sect. 3.4 and different loss functions for accurate scoring, we compared the performances of overall score training strategy and the execution score training strategy in our model. As shown in Table 5, when taking the Gaussian loss into consideration, the performance of our network improved by 0.81% under the execution score training strategy compared to MSE loss, and improved by 0.58% under the overall score training strategy. With the contrast of the performance of four different loss functions in our model, a similar improvement is observed as in AQA-7 dataset. It is worth noting that when taking the execution score into consideration in our model could boost the correlation coefficient by 3.5% with Gaussian loss. These experiment results further illustrate the fact that detailed score label can improve the accuracy of action quality scoring, and the Gaussian loss is responsible for training AQA model effectively.

Table 4 Comparison of our approach with existing methods on the MTL–AQA dataset

Methods	Sp. corr.
Pose + DCT [28]	0.2682
C3D-SVR [23]	0.7716
C3D-LSTM [23]	0.8489
MSCADC-STL [25]	0.8472
MSCADC-MTL [25]	0.8612
USDL-regression [33]	0.8905
C3D-AVG-STL [25]	0.8960
C3D-AVG-MTL [25]	0.9044
Ours (TECN + OS)[a]	0.8745
Ours (TECN + ES)[b]	**0.9095**

Bold value represents the best performance

[a] Using the overall score as the training label in our network

[b] Using the execution score as the training label in our network

Ablation study on various σ in Gaussian loss function. As is mentioned in Sect. 3.4, there exists a hyper-parameter in

Table 5 Comparison of using different score labels and loss function on the MTL–AQA dataset

Loss function	Score label	
	OS	ES
MSELoss	0.8687	0.9014
L1Loss	0.8666	0.8996
MSELoss + L1Loss	0.8683	0.9002
Gaussian loss	**0.8745**	**0.9095**

Bold value represents the best performance

Table 6 Comparison of various hyper-parameters σ in Gaussian loss function

σ	Score label	
	OS	ES
1	0.8564	0.8983
2	**0.8745**	**0.9095**
3	0.8695	0.9021
4	0.8724	0.8981
5	0.8648	0.9035

Bold value represents the best performance

Table 7 Comparison of different layers of score regressor

Layers	Score label	
	OS	ES
fc-layers{20480, 1}	0.8545	0.8885
fc-layers{20480, 4096, 1}	0.8688	0.8937
fc-layers{20480, 4096, 2048, 1}	**0.8745**	**0.9095**
fc-layers{20480, 4096, 2048, 512, 1}	0.8645	0.9018

Bold value represents the best performance

Eq. (6), which serves as the standard deviation in the Gaussian function. Here, σ is a hyper-parameter which takes for the level of a significant deviation from the average value. We choose five possible values at {1, 2, 3, 4, 5} intervals, as shown in Table 6. The results show that the hyper-parameter in Gaussian loss function has a strong affection on scoring performance, in particular because it concerns the deviation degree of the scoring distribution issues. As we can see that when the σ is set to 2, the TECN model has obtained the best results in ES and OS scenarios.

Comparison of different structure of score regression network. Finally, we explore the performance of several different structures of *fc-layers*. Four different fully connected structures are designed in Table 7. First, remove all the remaining layers to exploit the features strengthened by encoder network directly, but such predicted results are relatively poor. Therefore, we utilize the descending FC network to predict the action quality score. The number in each block means the sum of the neurons in corresponding layer. We observe that the model with *fc-layers*{20480, 4096, 2048, 1} structure obtain the best performance in the overall score training strategy and execution score training strategy.

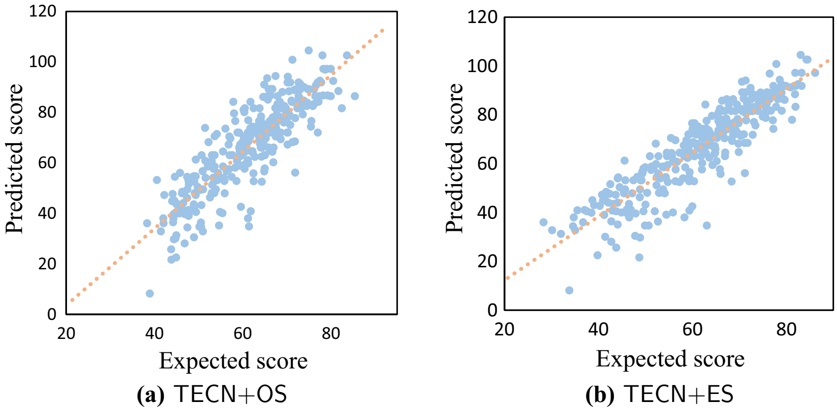
Visualization and qualitative analysis

To further verify which training strategy can fit the data better, we conduct a regression analysis to construct a mathematical model that examines the relationship between the predicted score and ground-truth. The comparison results of TECN + OS and TECN + ES methods are illustrated by scatter plots in Fig. 5. The predicted scores are represented by scatter points in the plot and the ground-truth scores are plotted in trend dotted line. The horizontal axis denotes the ground-truth score and vertical axis is the predicted score. Unsurprisingly, the superiority of TECN+ES on MTL–AQA is quite impressive. We observe that the proposed TECN+ES model converges the ground-truth score trendline preferable. The fewer outliers and the denser distribution of intermediate points confirm the validity of execution score training strategy with Gaussian loss function.

Two qualitative examples of the TECN model on MTL–AQA dataset are shown in Fig. 6. The dotted arrow in the figure represents the label that comes with the sport video. The ground-truth and predicted results of “Sample 26-19” are labeled by red color, while the data of “Sample 03-08” is labeled by blue color. Then, we compare the performance of the difference training strategies. Obviously, the result using detailed score labels such as execution scores performs better. And more notably, it is noteworthy that our model can accurately predict both the high and low scores of athletes, which indicates that the proposed method can capture variety of spatio-temporal information representation of diverse performance levels.

Conclusion

In this paper, we propose a novel Gaussian guided frame sequence encoder network for action quality assessment, which regards the AQA task as regression problem. Specifically, the introduced TECN network includes the feature extraction module, temporal encoder, and score regression module. A video frame-level framework is applied to model full-video frame features rather than clip-level features. Furthermore, we employ Gaussian loss function to optimize the network that devotes to model the score label using a Gaussian distribution instead of mean square error. In addition, the training scenarios followed by the objective rules of diving sport are designed to improve the accuracy of the score prediction. The experimental results on AQA-7 and MTL–AQA datasets superior to the state-of-the-art methods and demonstrate the strong effectiveness of the proposed method.



(a) TECN+OS (b) TECN+ES

Fig. 5 A comparison of different training strategy in scatter plot

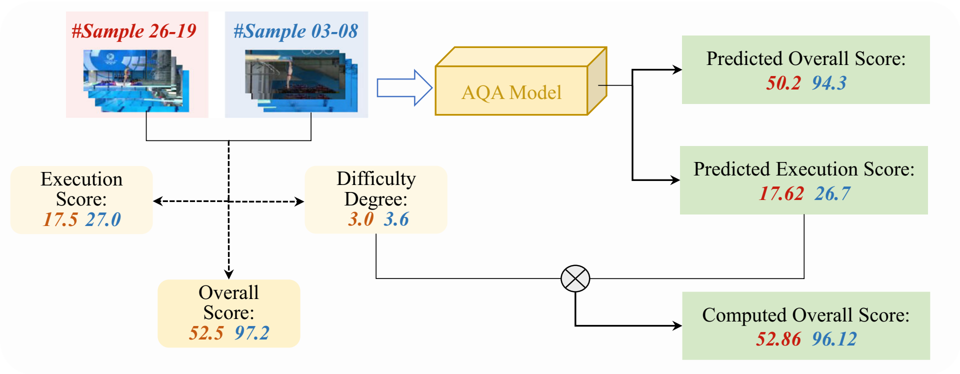


Fig. 6 Case study with qualitative results, which present the comparisons of TECN+OS, TECN+ES and ground-truth

Although the proposed work has achieved valuable results, the research of AQA still needs to further discuss. Since the proposed method only explore the contribution of a single branch in time series, there are limitations to some extent; therefore, we hypothesize that the AQA task can be improved from the following two aspects in the future:

- Multi-branch model for the score labels of several referees should be considered to construct a more objective and reasonable scoring mechanism.
- It needs to further explore an AQA model that is more suitable for small-scale datasets in practical application.

Acknowledgements This work was supported by the National Natural Science Foundation of China (61871196, and 62001176), National Key Research and Development Program of China (2019YFC1604700), Natural Science Foundation of Fujian Province of China (2020J01085 and 2022J01317), and the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (ZQN-YX601).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Basak H, Kundu R, Singh PK, Ijaz MF, Woźniak M, Sarkar R (2022) A union of deep learning and swarm-based optimization for 3d human action recognition. Sci Rep 12(1):1–17
2. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35(8):1798–1828. https://doi.org/10.1109/TPAMI.2013.50
3. Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: 2017 IEEE conference on computer vision and pattern recognition, pp 4724–4733. https://doi.org/10.1109/CVPR.2017.502
4. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
5. Dong L-J, Zhang H-B, Shi Q, Lei Q, Du J-X, Gao S (2021) Learning and fusing multiple hidden substages for action quality assessment. Knowl Based Syst 107388. https://doi.org/10.1016/j.knosys.

2021.107388. https://www.sciencedirect.com/science/article/pii/S095070512100650X (**ISSN 0950-7051**)

6. Doughty H, Damen D, Mayol-Cuevas W (2018) Who's better? who's best? pairwise deep ranking for skill determination. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 6057–6066. https://doi.org/10.1109/CVPR.2018.00634
7. Faller AJ (1981) An average correlation coefficient. J Appl Meteorol Climatol 20(2):203–205. https://doi.org/10.1175/1520-0450(1981)020<0203:AACC>2.0.CO;2
8. Farabi S, Himel HH, Gazzali F, Hasan B, Kabir M, Farazi M et al (2021) Improving action quality assessment using resnets and weighted aggregation. arXiv preprint arXiv:2102.10555
9. Fard Mahtab J, Sattar A, Darin Ellis R, Chinnam Ratna B, Pandya Abhilash K, Klein Michael D (2018) Automated robot-assisted surgical skill evaluation: predictive analytics approach. Int J Med Robot Comp Assist Surg 14(1):e1850. https://doi.org/10.1002/rcs.1850
10. Feichtenhofer C, Fan H, Malik J, He K (2019) Slowfast networks for video recognition. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 6201–6210. https://doi.org/10.1109/ICCV.2019.00630
11. Gao J, Zheng W-S, Pan J-H, Gao C, Wang Y, Zeng W, Lai J (2020) An asymmetric modeling for action assessment. In: Vedaldi A, Bischof H, Brox T, Frahm J-M (eds) Computer vision–ECCV 2020. Springer International Publishing, Cham, pp 222–238 (**ISBN 978-3-030-58577-8**)
12. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 6546–6555. https://doi.org/10.1109/CVPR.2018.00685
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778. https://doi.org/10.1109/CVPR.2016.90
14. Kingma DP, Ba JL (2015) Adam: a method for stochastic optimization. In: 3rd international conference on learning representations, San Diego, CA, USA
15. Lea C, Flynn MD, Vidal R, Reiter A, Hager GD (2017) Temporal convolutional networks for action segmentation and detection. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 1003–1012. https://doi.org/10.1109/CVPR.2017.113
16. Lei Q, Du J-X, Zhang H-B, Ye S, Chen D-S (2019) A survey of vision-based human action evaluation methods. Sensors. https://doi.org/10.3390/s19194129 (**ISSN 1424-8220**)
17. Li Y, Chai X, Chen X (2018) End-to-end learning for action quality assessment. In: Hong R, Cheng W-H, Yamasaki T, Wang M, Ngo C-W (eds) Advances in multimedia information processing–PCM 2018. Springer International Publishing, Cham, pp 125–134 (**ISBN 978-3-030-00767-6**)
18. Li Y, Ji B, Shi X, Zhang J, Kang B, Wang L (2020) Tea: temporal excitation and aggregation for action recognition. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 906–915. https://doi.org/10.1109/CVPR42600.2020.00099
19. Liu D, Li Q, Jiang T, Wang Y, Miao R, Shan F, Li Z (2021) Towards unified surgical skill assessment. In: 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 9517–9526. https://doi.org/10.1109/CVPR46437.2021.00940
20. Malpani A, Vedula SS, Chen CCG, Hager GD (2014) Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In: Stoyanov D, Collins DL, Sakuma I, Abolmaesumi P, Jannin P (eds) Information processing in computer-assisted interventions. Springer International Publishing, Cham, pp 138–147 (**ISBN 978-3-319-07521-1**)
21. Nekoui M, Tito CFO, Cheng L (2021) Eagle-eye: extreme-pose action grader using detail bird's-eye view. In: 2021 IEEE winter conference on applications of computer vision (WACV), pp 394–402. https://doi.org/10.1109/WACV48630.2021.00044
22. Pan J-H, Gao J, Zheng W-S (2019) Action assessment by joint relation graphs. In: 2019 IEEE/CVF international conference on computer vision (ICCV), pp 6330–6339. https://doi.org/10.1109/ICCV.2019.00643
23. Parmar P, Morris BT (2017) Learning to score olympic events. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 76–84. https://doi.org/10.1109/CVPRW.2017.16
24. Parmar P, Morris B (2019) Action quality assessment across multiple actions. In: 2019 IEEE winter conference on applications of computer vision (WACV), pp 1468–1476. https://doi.org/10.1109/WACV.2019.00161
25. Parmar P, Morris BT (2019) What and how well you performed? a multitask learning approach to action quality assessment. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 304–313. https://doi.org/10.1109/CVPR.2019.00039
26. Parmar P, Reddy J, Morris B (2021) Piano skills assessment. In: 2021 IEEE 23rd international workshop on multimedia signal processing (MMSP), pp 1–5. https://doi.org/10.1109/MMSP53017.2021.9733638
27. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017) Automatic differentiation in pytorch. In: NIPS 2017 workshop on Autodiff,.https://openreview.net/forum?id=BJJsrmfCZ
28. Pirsiavash H, Vondrick C, Torralba A (2014) Assessing the quality of actions. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer vision–ECCV 2014. Springer International Publishing, Cham, pp 556–571 (**ISBN 978-3-319-10599-4**)
29. Reiley CE, Hager GD (2009) Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In: Yang G-Z, Hawkes D, Rueckert D, Noble A, Taylor C (eds) Medical image computing and computer-assisted intervention–MICCAI 2009. Springer, Berlin Heidelberg, pp 435–442 (**ISBN 978-3-642-04268-3**)
30. Roditakis K, Makris A, Argyros A (2021) Towards improved and interpretable action quality assessment with self-supervised alignment. In: The 14th PErvasive technologies related to assistive environments conference, PETRA 2021, pp 507-513. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3453892.3461624. https://doi.org/10.1145/3453892.3461624 (**ISBN 9781450387927**)
31. Sardari F, Paiement A, Hannuna S, Mirmehdi M (2020) Vi-net-view-invariant quality of human movement assessment. Sensors. https://doi.org/10.3390/s20185258 (**ISSN 1424-8220**)
32. Shi Q, Zhang H-B, Li Z, Du J-X, Lei Q, Liu J-H(2022) Shuffle-invariant network for action recognition in videos. ACM Trans. Multimedia Comput Commun Appl, 18(3). https://doi.org/10.1145/3485665. https://doi.org/10.1145/3485665. ISSN 1551-6857
33. Tang Y, Ni Z, Zhou J, Zhang D, Lu J, Wu Y, Zhou J (2020) Uncertainty-aware score distribution learning for action quality assessment. In 2020 IEEE/CVF conference on computer vision and pattern recognition (cVPR), pp 9836–9845. https://doi.org/10.1109/CVPR42600.2020.00986
34. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In 2015 IEEE international conference on computer vision (ICCV), pp 4489–4497. https://doi.org/10.1109/ICCV.2015.510
35. Varadarajan B, Reiley C, Lin H, Khudanpur S, Hager G (2009) Data-derived models for segmentation with application to surgical assessment and training. In: G-Z Yang, D Hawkes, D Rueckert, A Noble, and C Taylor, editors, Medical image computing and computer-assisted intervention—MICCAI, pp 426–434. Springer, Berlin, Heidelberg (**ISBN 978-3-642-04268-3**)

36. Wang J, Du Z, Li A, Wang Y (2020) Assessing action quality via attentive spatio-temporal convolutional networks. In: Peng Y, Liu Q, Lu H, Sun Z, Liu C, Chen X, Zha H, Yang J (eds) Pattern recognition and computer vision. Springer International Publishing, Cham, pp 3–16 (**ISBN 978-3-030-60639-8**)
37. Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016) Temporal segment networks: towards good practices for deep action recognition. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision–ECCV 2016. Springer International Publishing, Cham, pp 20–36 (**ISBN 978-3-319-46484-8**)
38. Wang T, Wang Y, Li M (2020) Towards accurate and interpretable surgical skill assessment: a video-based method incorporating recognized surgical gestures and skill levels. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L (eds) Medical image computing and computer assisted intervention-MICCAI 2020. Springer International Publishing, Cham, pp 668–678 (**ISBN 978-3-030-59716-0**)
39. Wieczorek Michał, Siłka Jakub, Woźniak Marcin, Garg Sahil, Hassan Mohammad Mehedi (2021) Lightweight convolutional neural network model for human face detection in risk situations. IEEE Trans Ind Inf 18(7):4820–4829
40. Xiang X, Tian Y, Reiter A, Hager GD, Tran TD (2018) S3d: stacking segmental p3d for action quality assessment. In: 2018 25th IEEE international conference on image processing (ICIP), pp 928–932. https://doi.org/10.1109/ICIP.2018.8451364
41. Yan G, Woźniak M (2022) Accurate key frame extraction algorithm of video action for aerobics online teaching. In: Mobile networks and applications, pp 1–10
42. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the thirty-second AAAI conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and eighth AAAI symposium on educational advances in artificial intelligence, AAAI'18/IAAI'18/EAAI'18. AAAI Press (**ISBN 978-1-57735-800-8**)
43. Yang C, Xu Y, Shi J, Dai B, Zhou B (2020) Temporal pyramid network for action recognition. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 588–597. https://doi.org/10.1109/CVPR42600.2020.00067
44. Zeng L-A, Hong F-T, Zheng W-S, Yu Q-Z, Zeng W, Wang Y-W, Lai J-H (2020) Hybrid dynamic-static context-aware attention network for action assessment in long videos. In: Proceedings of the 28th ACM international conference on multimedia. Association for Computing Machinery, New York, NY, USA, pp 2526–2534. https://doi.org/10.1145/3394171.3413560 (**ISBN 9781450379885**)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.