End-To-End Learning for Action Quality Assessment

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, action quality assessment has attracted more and more attention of the researchers in computer vision. In this paper, an end-to-end framework is proposed based on fragment-based 3D convolutional neural network to realize the action quality assessment in videos. Furthermore, the ranking loss integrated with the MSE forms the loss function to make the optimization 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. The proposed network can indeed learn the evaluation criteria of actions and works well with limited training data. Widely experiments conducted on three public datasets convincingly show that our method achieves the state-of-the-art results.

Keywords: Action quality assessment
3D convolutional neural network · Deep learning · Ranking loss

1 Introduction

Nowadays, action quality assessment has received more attention in the community of computer vision. It aims at measuring the difference between the standard actions and the actions we perform. In practice, the action quality assessment has many potential applications, such as scoring athletes' performance, medical rehabilitation tests, dancing teaching and so on. This paper takes scoring athletes'performance as the instance to explore the action quality assessment algorithm. In sports competitions, the inefficiency and subjectivity of manual judgement diminish the interest seriously. However an automatic scoring system which is efficient and objective can enormously improve this situation [2].

This work was partially supported by 973 Program under contract No2015CB351802, Natural Science Foundation of China under contracts Nos. 61390511, 61472398, 61532018.

R. Hong et al. (Eds.): PCM 2018, LNCS 11165, pp. 125–134, 2018.
https://doi.org/10.1007/978-3-030-00767-6_12

Although the automatic scoring system can bring a lot of benefits, some challenges prevent it from applications, which mainly lie in the following aspects:

1. Each kind of action has its own characteristics and designing a general framework for various kinds of actions is difficult.
2. Some criteria for action quality assessment are not quantifiable and hard to be represented by hand-crafted features.
3. The automatic scoring system should make predictions close to the ground-truth scores and satisfy the ranking constraint. In the training, accomplishing the two goals synchronously is difficult.

To overcome these problem, we propose an end-to-end framework to realize an automatic scoring system. Inspired by the breakthroughs of feature learning with deep learning in video domain [10,16,17], we leverage the 3D convolutional neural network to extract features and these features are the better representation of actions than man-crafted features. However extracting discriminative features from a long sequence is difficult, we segment each video into fragments and extract fragment-features as [7]. Meanwhile, extracting features by fragments ensures that our framework has the capacity to learn the evaluation criteria of actions. Synchronously, we consider the score value constraint and the ranking constraint in the optimization objective. To this end, we integrate the ranking loss with MSE to form the final loss function. Additionally, our framework is common for various kinds of actions.

To summarize, our main contributions are as follows:

1. We propose an end-to-end framework which can indeed learn the evaluation criteria of actions.
2. The ranking loss is integrated with the traditional MSE to form the loss function making the predictions meet the score value constraint and the ranking constraint synchronously.

2 Related Works

There are only a few prior researches in the action quality assessment. The methods can be classified into regression-based method [4,5,7,11] and classification-based method [6,8,9].

The regression-based method directly predicts continuous score to evaluate the quality of an action. In [4], Pirsiavash *et al.* propose a learning-based framework that takes steps towards assessing how well people perform actions in videos. Their approach works by training a regression model from spatiotemporal pose features to scores obtained from expert judges. In [7], they employ the 3D Convolutional Network (C3D) [10] to extract the body-pose features and the C3D+SVR achieves the state-of-art results.

While for the classification-based method, the quality of an action will be classified into different grades. In [6], Parmar *et al.* explore the problem of exercise quality measurement. They collect negative exercise data by asking subjects to deliberately make subtle errors during the exercise. Following that, they use adaboost to classify exercise into "good" or "bad". Literature [8] presents an automatic framework for surgical skill assessment. They introduce video analysis techniques to obtain features.

There are also some researches combining two methods. In [13], Chai *et al.* develop a system on sign evaluation with both classification and regression models. The system first determines whether a sign is the appointed one by the classification model. For the sign which passes the verification, the system will give a score by the regression model.

In summary, most of previous researches tackle this problem by traditional machine learning techniques. They treat the problem as a two-phase task instead of an end-to-end process and only emphasize the score value constraint with ignoring the ranking constraint.

3 End-To-End Score Prediction

Our end-to-end framework is composed of feature extraction and score prediction, as shown in Fig. 1. In order to get discriminative features which can represent complex movement information, we employ the 3D convolutional neural network extractor (C3D extractor), which consists of the first 13 layers of C3D [10], in our framework. Since an action video sequence is too long, it is difficult for C3D extractor to generate a good representation over the whole action. Hence we divide the video into fragments evenly along the temporal dimension. Specifically, one extractor is applied for one fragment and the weights are not

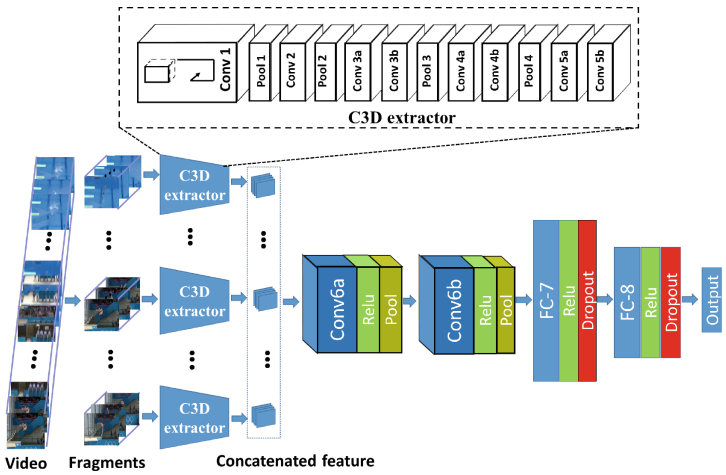


Fig. 1. Overview of our proposed end-to-end learning framework.

shared among C3D extractors, making each C3D extractor be sensitive to just one special movement which is a part of an action, such as the take-off in diving.

Once the fragment-based features are extracted, they will form a concatenated feature immediately. Then the concatenated feature goes through two convolution layers (followed by Relu and $3 \times 3 \times 3$ Max-pooling) to get higher level feature maps. The receptive fields of both convolution layers are $3 \times 3 \times 3$. Finally, two fully-connected layers (followed by Relu and Dropout = 0.8) and an output layer are used to regress scores.

4 Loss Function

In our loss function, there are two complementary terms. One is MSE (Eq. 1) which is general applied to constrain score value and the other is ranking loss (Eq. 2) severing to ensure the right order of the predictions. 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.

$$L_1 = \frac{1}{2n}\sum_{i=1}^{n}(s_i - g_i)^2 \tag{1}$$

$$L_2 = \sum_{i=1}^{n}\sum_{j=1,j>i}^{n} RELU(-(s_j - s_i)sign(g_j - g_i) + \delta) \tag{2}$$

where g_i and s_i are the ground-truth score and prediction of the ith sample in a batch of data respectively. n is batch size. $RELU(\cdot)$ is a rectified linear unit activation and $\delta = 2$ works as the margin.

The final loss function can be written by Eq. 3,

$$L = L_1 + \alpha L_2 + \beta\|w\|^2, \alpha > 0, \beta > 0 \tag{3}$$

where $\|w\|^2$ is L2-regularization item, α and β are used to balance these three items.

For the better optimization, we should balance L_1 and L_2. If one adjusts weights too fast, the other will be weakened. So we make them update weights with similar speeds. The gradients of L_1 and L_2 with respect to s_i can be denoted as Eqs. 4 and 5:

$$\nabla_{s_i} L_1 = \frac{1}{n}(s_i - g_i) \tag{4}$$

$$\begin{aligned}\nabla_{s_i} L_2 = \alpha &\sum_{j=1,j>i}^{n} H(-(s_j - s_i)sign(g_j - g_i) + \delta) \\ &-\alpha \sum_{j=1,j<i}^{n} H(-(s_i - s_j)sign(g_i - g_j) + \delta)\end{aligned} \tag{5}$$

$$H(x) = \begin{cases} 1, x > 0 \\ 0, x \leq 0 \end{cases} \tag{6}$$

Let $|\nabla_{s_i} L_1| = |\nabla_{s_i} L_2|$. Thus we can get α by Eq. 7 roughly.

$$\alpha = \frac{|\nabla_{s_i} L_1|}{|\nabla_{s_i} L_2|} \tag{7}$$

5 Experimental Results

In this section, we evaluate our method on three kinds of actions (diving, vault and figure skating) from three public datasets, i.e. UNLV-Diving Dataset [7], UNLV-Vault Dataset [7] and Mit-Skating Dataset [4]. First is the simple introduction of these three datasets. Secondly, the experiment configuration and the evaluation metric are given. Then, we explore the learning capacity of our framework. After that we verify the effectiveness of our loss function. Finally, we compare our method with others.

5.1 Dataset

UNLV-Diving Dataset: This dataset contains 370 Olympic men's 10-meter platform videos, each has roughly 150 frames. The scores vary between 0 (the worst) and 100 (the best). A diving score is determined by the product of execution score (judging quality of a diving) multiplied by the diving difficulty score (fixing agreed-upon value based on diving type).
UNLV-Vault Dataset: This dataset includes 176 videos. These videos are short relatively with an average length of about 75 frames. The scores range from 0 (the worst) to 20 (the Best). A vault score is determined by the sum of the execution score and the difficulty score.
Mit-Skating Dataset: 150 Olympic Figure Skating videos are included in this dataset. Each one has an average of 4200 frames. The scores range from 0 (the worst) to 100 (the Best). In [7], Parmar *et al.* augment the dataset by adding another 21 videos. The figure skating score is obtained as the same way of vault.

In experiments on diving and vault, the training/test data splits are 300/70 and 120/56 respectively. The two data splits are the same as that in [7]. In diving, all videos are padded with zero frames to length 151 and each video is divided into 9 fragments (16 frames per fragment, the rest of frames are used for data augmentation). In vault, all videos are padded to length 100 and the number of fragments is 6 (16 frames per fragment). Each fragment in the two datasets includes 16 frames. The redundant frames are used to augment data by shifting the start frame.

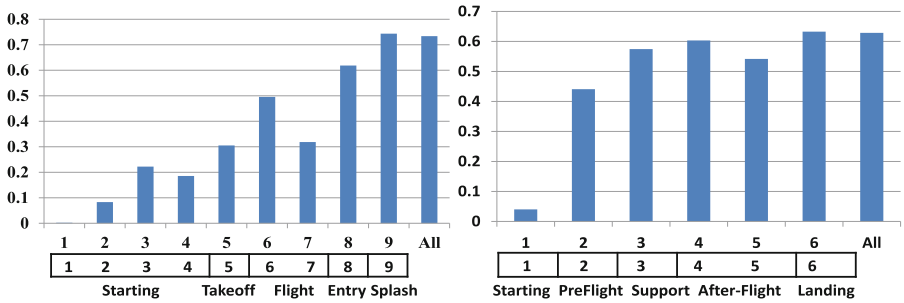
For the Mit-Skating Dataset, 100 videos are used for training and the remaining 71 videos for testing and all videos are padded to length 4500. We drop four out of every five frames because of limited computing resources and divide each video into 9 fragments. Each fragment includes 100 frames.

Because of the limited computing resources and high rate of down-sample, we perform the detailed evaluation on the UNLV-Diving Dataset and the UNLV-Vault Dataset while only give the comparison results on the Mit-Skating Dataset.

5.2 Experiments Setting and Evaluation Metric

In our task, only several hundreds of data are available. In order to address the problem of limited training data, we pre-train the C3D extractor with UCF-101 [12] and augment data by shifting the start frame with a random number within [0,5]. During the training, we adopt different learning rate for the feature extraction (lr = 10e–4) and the score prediction (lr = 10e–3). Learning rate decay is set to 0.45 per 600 iterations. The optimization algorithm is Adam [15].

To evaluate the performance, commonly used spearman rank correlation (SRC) [7] is adopted as our measurement. SRC is a nonparametric measure of rank correlation (statistical dependence between the ranking of two groups of variables). The larger the SRC, the higher the rank correlation between the ground-truth scores and the predictions. Also the high SRC means that the predictions meet the ranking constraint well.



(a) Diving (b) Vault

Fig. 2. Exploration on the learning capacity of our framework on the UNLV-Diving Dataset and the UNLV-Vault Dataset. The vertical axis indicates SRC. The horizontal axis indicates index of the fragment. In the bottom of each graph, there are 5 stages and the corresponding fragments of each action respectively.

5.3 Learning Capacity of Our Framework

In this subsection, we explore the learning capacity of our framework by predicting the score with different fragments and all fragments. The loss function is MSE.

Every time, we only employ one fragment to predict scores and the other fragments are set to zero. At last, we employ all fragments for score prediction. The results are shown in Fig. 2. According to diving rules [3], a diving can be divided into 5 stages which are Starting, Take-Off, Flight, Entry and Splash. Among these stages, Starting has no contribution to the score and Take-Off determines the score slightly. Flight and Entry are important to the score. Although Splash is irrelevant to the judgement, the small size of splash is a sign of a successful diving. Because, in the end, the accumulation of more minor deviations in preceding movements will result in a larger splash. For vault, it also has 5 stages

which are Starting, Pre-Flight, Support, After-Flight and Landing [1] respectively. Similarly, Starting doesn't determine the score and Pre-Flight has small contribution to the score. The other three stages mainly determine the score and the stable Landing can also be regarded as the sign of a perfect vault. Meanwhile we match fragments to different stages roughly at the bottom of Figs. 2(a) and (b). We observe that the fragments corresponding to the stages determining the score have higher SRC than that of other fragments. Specifically, the fragments corresponding to Splash and Landing have the best performance, indicating their important role in judgement. The Splash and Landing can be used to evaluate the quality instead of all fragments. So we only use the last fragment in the following experiments.

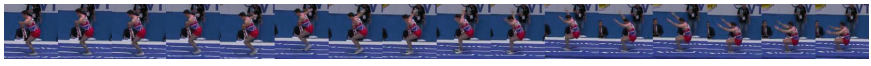
The results suggest that our framework can actually learn the evaluation criteria of actions. Figures 3 and 4 show the example frames in the last fragments of two actions and also the corresponding predictions and ground-truth scores.

(a) ground-truth score: 99.75 prediction: 99.73

(b) ground-truth score: 41.25 prediction: 40.40

Fig. 3. Example frames in the last fragments of two diving videos and the corresponding predictions and ground-truth scores. The less the splash, the higher the score.

(a) ground-truth score: 16.67 prediction: 15.77



(b) ground-truth score: 14.96 prediction: 13.06

Fig. 4. Example frames in the last fragments of two vault videos and the corresponding predictions and ground-truth scores. The steadier the landing, the higher the score.

5.4 Effectiveness of Loss Function

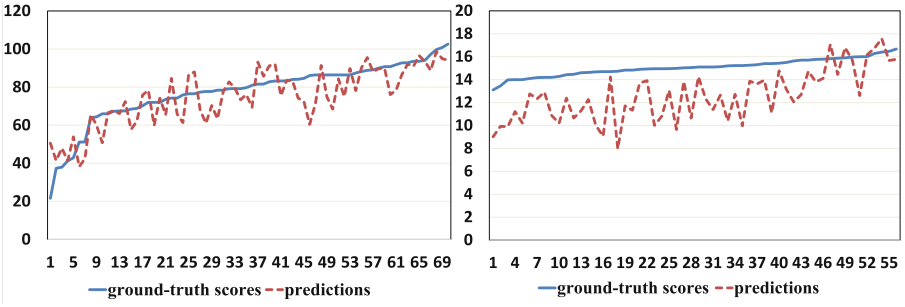
In this subsection, we verify the effectiveness of our loss function. The results in Table 1 show that the SRC of the ranking loss is higher than that of MSE significantly, indicating the ranking loss is more effective for the ranking constraint than MSE. Further, the combined loss function obtains an extra improvement compared with each single loss term, suggesting that the ranking loss can help MSE to make the prediction meet the ranking constraint.

In Table 1, we also give the mean euclidean distance (MED) between the predictions and the ground-truth scores. It is observed that MSE is able to constrain the score value while the ranking loss doesn't work at all. Only the ranking loss are combined with MSE, the MED get to be small relatively.

So the two loss terms are complementary and their combination contributes to our optimization objective (high SRC and relatively small MED). Finally, we also show the predictions and the ground-truth scores of all test samples, which are sorted from low to high by the ground-truth score, from the two datasets in Fig. 5. As Fig. 5 shows, the predictions and the ground-truth scores have the similar trend.

Table 1. Experimental results of different loss functions on the UNLV-Diving Dataset and the UNLV-Vault Dataset.

Loss function	SRC (Diving)	MED (Diving)	SRC (Vault)	MED (Vault)
MSE	0.7337	9.01	0.6325	1.33
Ranking loss	0.7870	75.62	0.6648	14.63
MSE+Ranking loss	0.8009	7.78	0.7028	2.60



(a) Diving (b) Vault

Fig. 5. The predictions and ground-truth scores of all test samples in the UNLV-Diving Dataset and the UNLV-Vault Dataset. The Y-axis represents the score value. The X-axis represents index of sample (sorting from low to high by the ground-truth score).

5.5 Comparison with Other Methods

We also compare our method with other state-of-the-art methods [4,7,14] on three datasets, as shown in Table 2.

For diving and vault, literature [7] uses a fixed data split and also implement Pose+DCT by this data split. The results of Pose+DCT here are from [7]. We adopt the same data split as [7].

For figure skating, literature [4] repeats the experiment 200 times with different random data splits and average the results. We just do one random split for the heavy computing cost.

In Table 2, we achieve the state-of-the-art results on the three datasets by using MSE+Ranking loss. Although our method outperforms [4,7,14], the results on the UNLV-Vault Dataset and the Mit-Skating Dataset are not breathtaking. The reason may come down to the view variation among the vault videos and down-sample on the skating videos.

Table 2. Comparison of our method with other state-of-the-art methods on the three datasets.

	SRC (Diving)	SRC (Vault)	SRC (Skating)
Pose+DCT [4]	0.53	0.10	0.35
ConvISA [14]	-	-	0.45
C3D+LSTM [7]	0.27	0.05	-
C3D+SVR [7]	0.78	0.66	0.53
Ours (MSE)	0.7435	0.6325	0.4167
Ours (Ranking loss)	0.7870	0.6648	0.2129
Ours (MSE+Ranking loss)	**0.8009**	**0.7028**	**0.5753**

6 Conclusion

This paper proposes an end-to-end learning-based framework to realize action quality assessment. In this framework, we divide videos into fragments and employ C3D extractor to obtain fragment-based features. The fragment-based features ensures our framework has the capacity to learn the evaluation criteria of actions. Specifically, in our loss function, both the score value and the ranking are considered. The combination of MSE and the ranking loss contributes to the better performance. Finally, wide experiments show that our method outperforms other state-of-the-art methods on three public datasets.

References

1. Vault. https://en.wikipedia.org/wiki/Vault-(gymnastics). 2.1.2. Accessed 2018
2. List of Olympic Games Scandals and Controversies. https://en.wikipedia.org/wiki/List-of-Olympic-Games-boycotts
3. FINA-DIVING RULES. http://www.fina.org/content/diving-rules. D8.1.3
4. 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
5. 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)

6. Parmar, P., Morris, B.: Measuring the quality of exercises. In: 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), pp. 2241–2244 (2016)
7. Parmar, P., Morris, B.: Learning to score olympic events. In: 30th IEEE Conference on Computer Vision and Pattern Recognition, pp. 76–84 (2017)
8. 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
9. 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 (2016)
10. 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 (2015)
11. Venkataraman, V., Vlachos, I., Turaga, P.: Dynamical regularity for action analysis. In: 26th British Machine Vision Conference, pp. 67.1–67.12 (2015)
12. 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)
13. 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 (2017)
14. 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 (2011)
15. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization arXiv preprint arXiv:1412.6980 (2014)
16. 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 (2017)
17. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: International Conference on Computer Vision, pp. 5534–5542 (2017)