Pairwise Contrastive Learning Network for Action Quality Assessment

Mingzhe Li[1], Hong-bo Zhang[1], Qing Lei[2], Zongwen Fan[2], Jinghua Liu[3], Ji-Xiang Du[3]

([1]College of Computer Science and Technology, Huaqiao University, Xiamen, China)

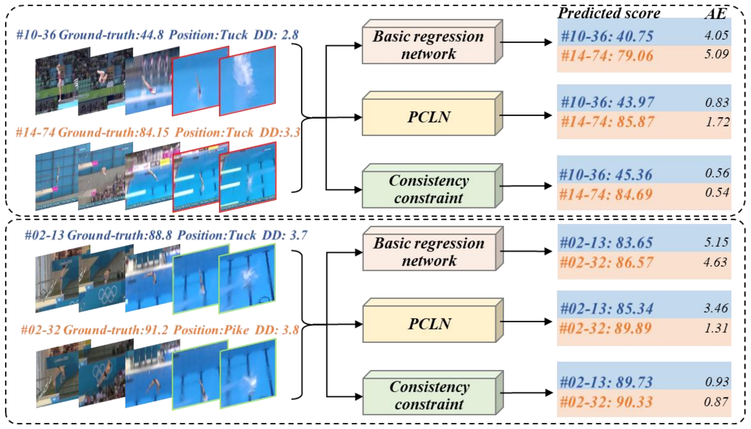
([2]Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Xiamen, China)

([3]Fujian Key Laboratory of Big Data Intelligence and Security, Xiamen, China)

Introduction

Motivation:

- Previous strategies ignored the subtle difference between various videos, which is the key reminder to predict action quality score.
- Learning to rank task compares each pair of data to obtain the ranking.
- Previous methods have low efficiency and computation power when processing large quantities of data.

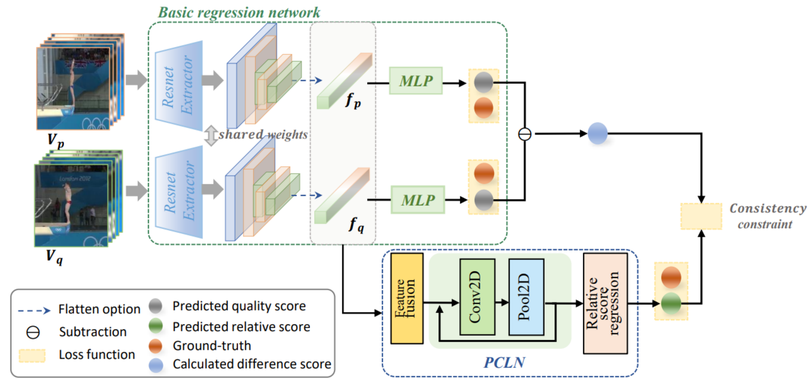


Qualitative comparison results.

Contributions:

- We extend the quality score regression to relative score prediction and propose a new end-to-end AQA model to enhance the performance of the basic regression network. The basic regression network and PCLN are combined during the training, but in the testing, only basic regression network is employed, which makes the proposed method simple but high accuracy.
- A novel pairwise LTR-based model PCLN is proposed to concern the subtle difference between videos. A new consistency constraint between PCLN and basic regression network is defined.
- The experimental results based on the public datasets show that the proposed method achieves the better performance compared with existing methods. Ablation experiments are also conducted to verify the effectiveness of the each component of proposed method.

Method

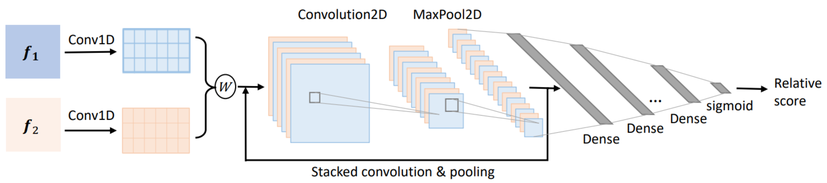


Pipeline of our proposed model. The video pair is fed into the feature extractor as the input, then the score regressor is used to predict the score of these two videos, and PCLN is designed to learn the mapping from the video pair to the relative score. We optimize the overall model by three constraints. During testing, only basic regression network is employed to predict the quality score of the input video.

$$\mathcal{L}_{bs} = \frac{1}{2}\sum_{i=p,q}^{N}(\tilde{S}_i - S_i)^2 \qquad \mathcal{L}_{rs} = (\Delta S - |S_p - S_q|)^2$$

$$\mathcal{L}_{ds} = (\Delta S - |\tilde{S}_p - \tilde{S}_q|)^2 \qquad \mathcal{L} = \mathcal{L}_{bs} + \mathcal{L}_{ds} + \mathcal{L}_{rs}$$

To train the proposed AQA model, we formulate three constraints to learn effective relative scores between different videos and accurate athlete quality scores simultaneously.



In order to learn the differences between videos to assist the final scoring task, we build a separate branch for the pairwise video based LTR network named PCLN to learn the relative scores.

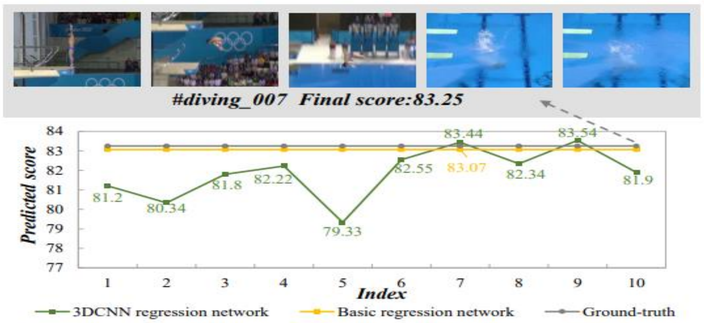
Results

Network	Diving	Gym Vault	Skiing	Snowboard	Sync. 3m	Sync. 10m	Avg. SRC
ST-GCN	0.3286	0.5770	0.1681	0.1234	0.6600	0.6483	0.4433
C3D-SVR	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937
JRG	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849
USDL	0.8099	0.7570	0.6538	0.7109	0.9166	0.8878	0.8102
EAGLE-Eye	0.8331	0.7411	0.6635	0.6447	0.9143	0.9158	0.8140
Adaptive	0.8306	0.7593	0.7208	0.6940	0.9588	0.9298	0.8500
Ours	**0.8697**	**0.8759**	**0.7754**	**0.5778**	**0.9629**	**0.9541**	**0.8795**

Comparison on AQA-7.

Methods	Pose+ DCT	C3D-LSTM	MSCAD C-MTL	C3D-AVG-MTL	USDL	SA& HMreg	Ours(FSP)	Ours(ESP)
Sp. Corr.	0.2682	0.8489	0.8612	0.9044	0.9066	0.8970	**0.8798**	**0.9230**

Comparison on MTL-AQA.



Comparison results of temporal feature encoder and 3D convolution network.

Batch Size	Video Pairs	Score Label	
		FSP	ESP
8	28	0.8729	0.9094
16	120	0.8750	0.9118
32	496	0.8777	0.9188
64	2016	**0.8798**	**0.9230**

Comparison results of different video pair number on the MTL-AQA dataset.

Basic regression network	PCLN	Consistency constraint	Score Label	
			FSP	ESP
✓	×	×	0.8745	0.9095
✓	✓	×	0.8788	0.9196
✓	✓	✓	**0.8798**	**0.9230**

Ablation study on different assessment structures on the MTL-AQA dataset, all of the models use 2016 pairs of video.