Paper ID #1510

TSA-Net: Tube Self-Attention Network for Action Quality Assessment

Shunli Wang[1,3], Dingkang Yang[1,2], Peng Zhai[1,4], Chixiao Chen[1], Lihua Zhang[2,1,3,4*]

Academy for Engineering and Technology, Fudan University[1] Ji Hua Laboratory, Foshan, China[2]
Engineering Research Center of AI and Robotics, Shanghai, China[3]
Jilin Provincial Key Laboratory of Intelligence Science and Engineering, Changchun, China[4]

https://github.com/Shunli-Wang/TSA-Net

Introduction

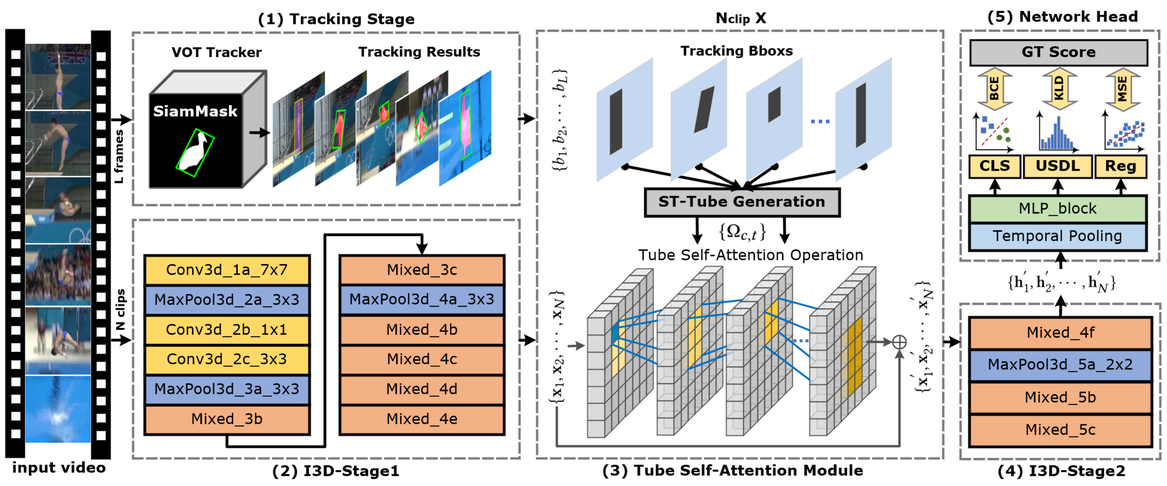
Models in HAR require distinguishing subtle differences between different actions.

Models in AQA require evaluating a specific action's advantages and disadvantages.

Human Action Recognition (HAR) → **Huge GAP** ← **Action Quality Assessment (AQA)**

- Most existing methods in AQA usually directly migrating the model from HAR tasks, which ignores the intrinsic differences within the feature map such as foreground and background information.
- To address this issue, we propose a Tube Self-Attention Network (TSA-Net) for action quality assessment tasks with the following merits:
  - ✔ High Computational Efficiency
  - ✔ High Flexibility
  - ✔ State-of-the-art Performance
- A dataset named FR-FS is proposed to explore the basic action assessment in the figure skating scene.

Framework



Overview of the proposed TSA-Net for AQA

- **1.Tracking**. VOT tracker is adopted to generate tracking results $\mathbf{B}$.
- **2.Feature Extraction-s1**. The input video is divided into N clips and the feature extraction is performed by I3D-Stage1 to generate $\mathbf{X}$.
- **3.Feature Aggregation**. ST-Tube is generated given $\mathbf{B}$ and $\mathbf{X}$, and then the TSA mechanism is used to complete the feature aggregation, results in $\mathbf{X}'$.
- **4.Feature Extraction-s2**. Aggregated feature $\mathbf{X}'$ is passed to I3D-Stage2 to generate $\mathbf{H}'$.
- **5.Network Head**. The final scores are generated by *MLP_Block*.

Experiments

- **Quantitative analysis on AQA-7 and MTL-AQA**

Comparison with SOTAs on AQA-7

Method	Diving	Gym Vault	Skiing	Snowboard	Sync. 3m	Sync. 10m	Avg. Corr.
Pose+DCT [27]	0.5300	-	-	-	-	-	-
ST-GCN [41]	0.3286	0.577	0.1681	0.1234	0.6600	0.6483	0.4433
C3D-LSTM [23]	0.6047	0.5636	0.4593	0.5029	0.7912	0.6927	0.6165
C3D-SVR [23]	0.7902	0.6824	0.5209	0.4006	0.5937	0.9120	0.6937
JRG [22]	0.7630	0.7358	0.6006	0.5405	0.9013	0.9254	0.7849
USDL [33]	0.8099	0.757	0.6538	**0.7109**	0.9166	0.8878	0.8102
NL-Net	0.8296	0.7938	**0.6698**	0.6856	0.9459	0.9294	0.8418
TSA-Net (Ours)	**0.8379**	**0.8004**	0.6657	0.6962	**0.9493**	**0.9334**	**0.8476**

Computational cost analysis on AQA-7

Method	NL-Net	TSA-Net	Comp. Dec.	Corr. Imp.
Diving	2.2G	0.864G	-60.72%	↑0.0083
Gym Vault	2.2G	0.849G	-61.43%	↑0.0066
Skiing	2.2G	0.283G	-87.13%	↓0.0041
Snowboard	2.2G	0.265G	-87.97%	↑0.0106
Sync. 3m	2.2G	0.952G	-56.74%	↑0.0034
Sync. 10m	2.2G	0.919G	-58.24%	↑0.0040
Average	2.2G	0.689G	-68.70%	↑0.0058

Results on MTL-AQA

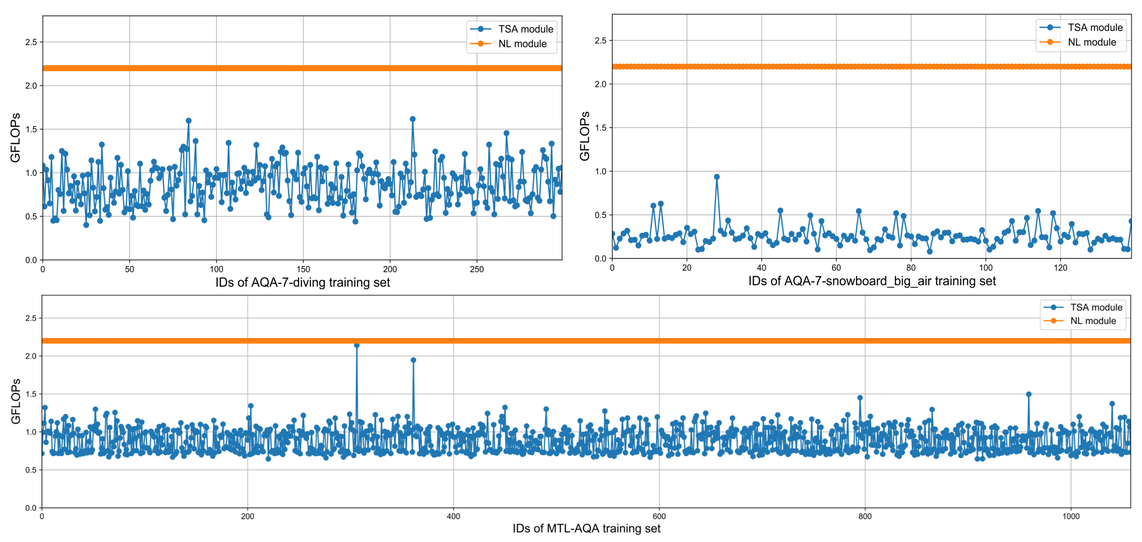
Method	Avg. Corr.
Pose+DCT [27]	0.2682
C3D-SVR [23]	0.7716
C3D-LSTM [23]	0.8489
C3D-AVG-STL [25]	0.8960
C3D-AVG-MTL [25]	0.9044
MUSDL [33]	0.9273
NL-Net	**0.9422**
TSA-Net	0.9393

TSA-Net achieves SOTA performance on AQA-7 and MTL-AQA datasets.

- **Visualization on AQA-7 and MTL-AQA**

Single object tracking strategy can handle these difficult situations perfectly:
★*Camera rotation* ★*Small target* ★*Synchronized diving* ★*Target loss*

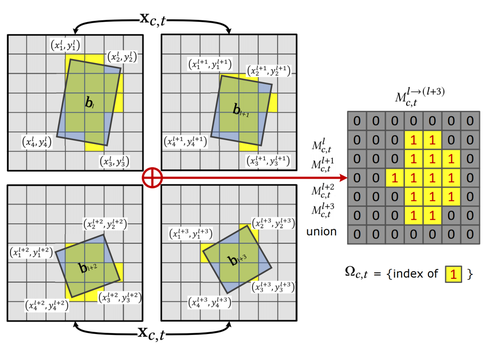
- **Computational cost comparison: TSA module *vs* Non-local**



TSA-Net can obtain better performance while reducing the computational cost.

TSA Module

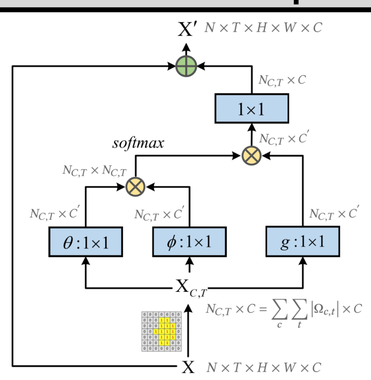
Step 1: Spatio-temporal Tube Gen.



Mask Generation

$$M_{c,t}^{l}(i,j)=\begin{cases}1, S(b_l,(i,j)) \geqslant \tau \\ 0, S(b_l,(i,j)) < \tau\end{cases}$$

Step 2: Tube Self-attention Operation



TSA mechanism

$$\mathbf{y}_p=\frac{1}{C(\mathbf{x})}\sum_{\forall c}\sum_{\forall t}\sum_{\forall (i,j)\in\Omega_{c,t}} f\left(\mathbf{x}_p,\mathbf{x}_{c,t}(i,j)\right)g(\mathbf{x}_{c,t}(i,j))$$

Computational Complexity Reduction $O\left((N\times T\times H\times W)\times(N\times T\times H\times W)\right) \Rightarrow O\left(\left(\sum_c\sum_t|\Omega_{c,t}|\right)\times\left(\sum_c\sum_t|\Omega_{c,t}|\right)\right)$

Conclusion

- In this paper, we present TSA-Net for AQA tasks, which is able to capture rich spatiotemporal contextual information in human motion.
- Experiments on AQA-7, MTL-AQA, and proposed FR-FS demonstrate that TSA-Net can capture long-range information and achieve high performance with less computational cost.
- An adaptive mechanism of ST-Tube will be explored to avoid the sensitivity of the TSA-Net to the size issue in future.

Acknowledgements

This work was supported by Shanghai Municipal Science and Technology Major Project 2021SHZDZX0103 and National Natural Science Foundation of China under Grant 82090052.

Contact

Shunli Wang
slwang19@fudan.edu.cn
Lihua Zhang*
lihuazhang@fudan.edu.cn