西安交通大學
XI'AN JIAOTONG UNIVERSITY

You Ought to Look Around: Precise, Large Span Action Detection

Ge Pan*, Han Zhang*, Fan Yu†, Yonghong Song*, Yuanlin Zhang#, Han Yuan†

*School of Software Engineering, Xi' an Jiaotong University,
#Institute of Artificial Intelligence and Robotics, Xi' an Jiaotong University,
†Distributed and parallel software lab, Huawei Technologies

Outline

1、Introduction

2、Background

3、Methodology

4、Results

Introduction

- Visual understanding of human actions is a challenging research area.
- For human action detection, one of the main challenges is the large diversity of action duration
- There are three major aspects for large span of durations: 1) Camera parameters ; 2) Subject factor ; 3) Action factor.

Background

- There are two spatiotemporal feature extraction methods, one is based on the two-stream network method in [1] and the other is the 3D convolution method in [2].
- There are end-to-end methods used in [3] and two-stage methods in [4] in action localization.
- The fusion of features is also something that needs attention in [5]

1、 Simonyan K , Zisserman A . Two-Stream Convolutional Networks for Action Recognition in Videos[J]. Advances in neural information processing systems, 2014, 1.

2、 D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," 2015 IEEE International Conference on Computer Vision (ICCV).

3、 H. Xu, A. Das and K. Saenko, "R-C3D: Region Convolutional 3D Network for Temporal Activity Detection," 2017 IEEE International Conference on Computer Vision (ICCV).

4、 Lin T , Zhao X , Su H , et al. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation[J]. 2018.

5、R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan, "Graph convolutional networks for temporal action localization," arXiv:Computer Vision and Pattern Recognition, 2019.

Methodology

Pipeline overview

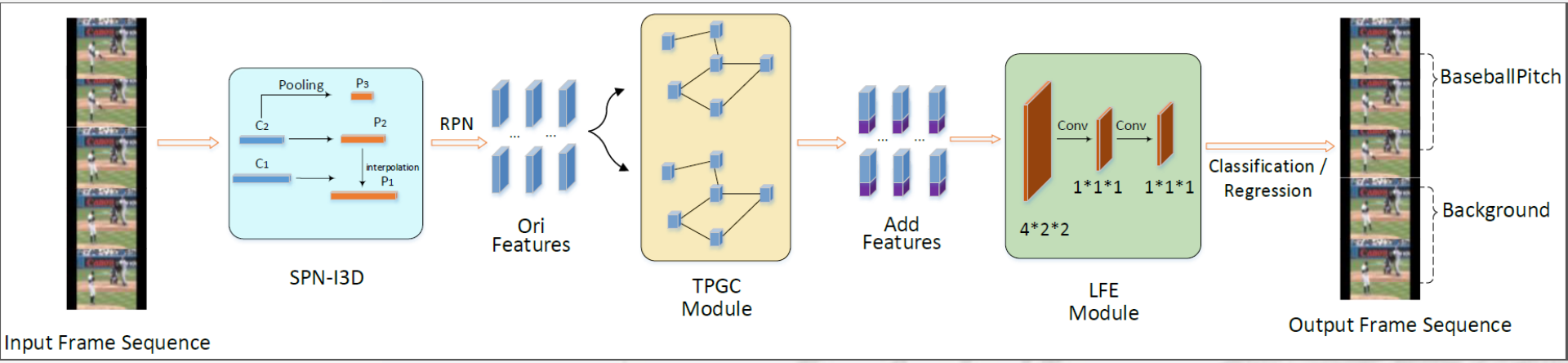


Fig.1. YOLA model architecture. We propose a more robust end-to-end framework for temporal detection task.

Methodology

Segment Feature Alignment

- TPGC Module

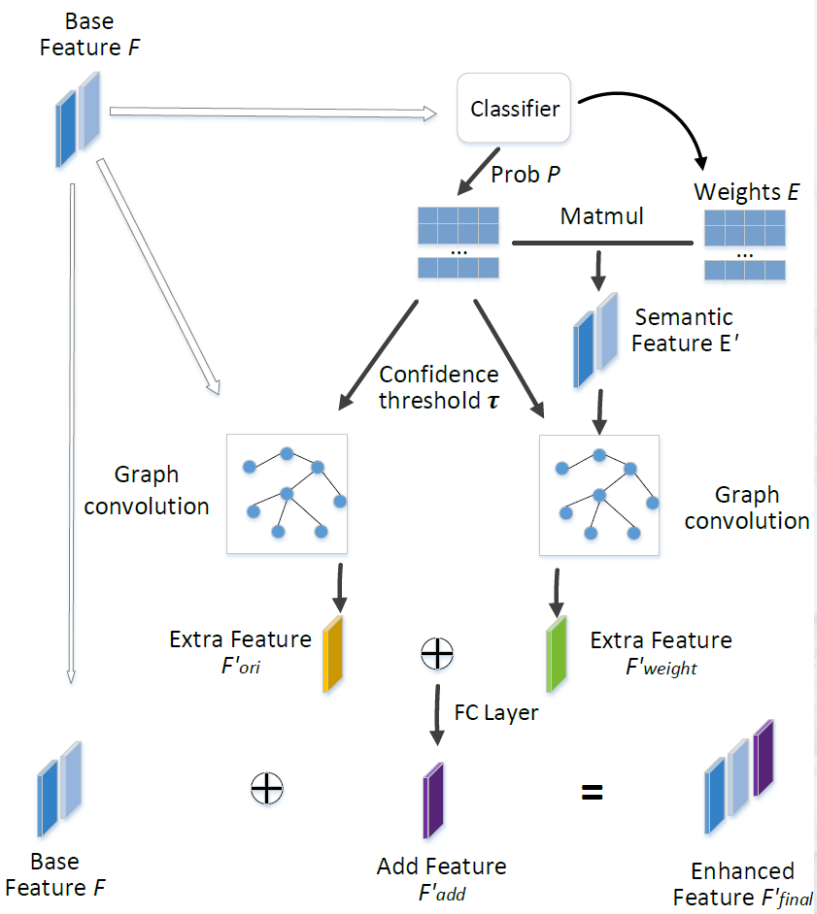


Fig.2. TPGC Module.

Methodology

Segment Feature Alignment

- Generate the adjacency matrix

$$A(X_p, X_q) = \begin{cases} 0, & \text{if } c_p < \tau \text{ or } c_q < \tau \\ \alpha A_1(X_p, X_q) + \beta A_2(X_p, X_q), & \text{if } c_p \geq \tau \text{ and } c_q \geq \tau \end{cases}$$

$$\begin{aligned} A_1(X_p, X_q) &= tIOU_{X_p, X_q} \\ &= \frac{(t_{start,p}, t_{end,p}) \bigcap (t_{start,q}, t_{end,q})}{(t_{start,p}, t_{end,p}) \bigcup (t_{start,q}, t_{end,q})} \end{aligned}$$

$$A_2(X_p, X_q) = Cos(X_p, X_q) = \frac{F_p^T F_q}{||F_p||_2 ||F_q||_2}$$

- Generate the feature

$$F'_{ori} = AF\,W_{ori}$$

$$F'_{weight} = AE'\,W_{weight}$$

Methodology

Segment Feature Alignment

- Feature fusion

$$F'_{final} = F \bigoplus FC(F'_{ori} \bigoplus F'_{weight})$$

- rectify the center coordinates

$$C_{new} = \begin{cases} gt_{start} + 0.5 * l_{org}, \text{if right part of instance lost} \\ gt_{end} - 0.5 * l_{org}, \text{if left part of instance lost} \\ \frac{gt_{start}+gt_{end}}{2}, \text{if left and right part of instance lost} \end{cases}$$

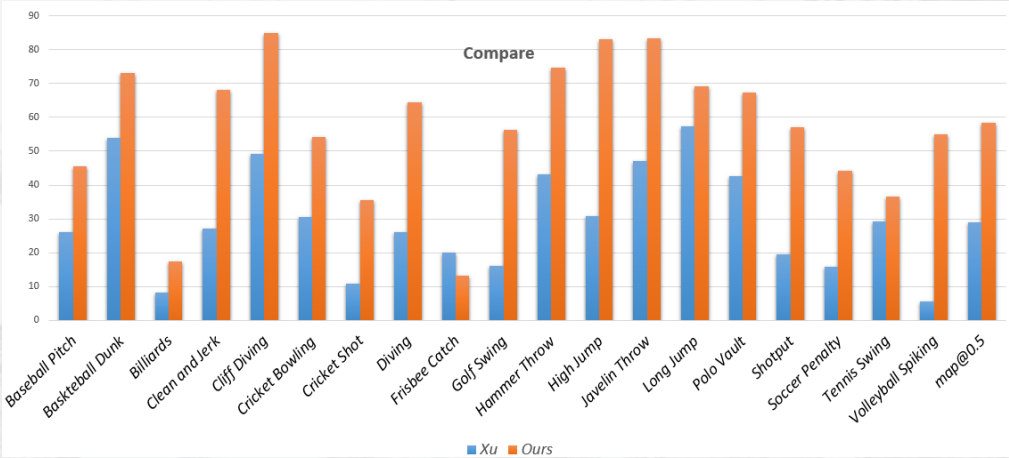
Results

State-of-the-art results on Thumos' 14 datasets

TABLE I THE EXPERIMENTAL RESULTS ON Thumos' 14 DATASET.

tIoU	0.1	0.2	0.3	0.4	0.5
Richard et al.	39.7	35.7	30.0	23.2	15.2
Shou et al.	47.7	43.5	36.3	28.7	19.0
Yeung et al.	48.9	44.0	36.0	26.4	18.8
Yuan et al.	51.4	42.6	33.6	26.1	18.8
Buch et al.	-	-	45.7	-	29.2
Gao et al.	60.1	56.7	50.1	41.3	31.0
Dai et al.	-	-	-	33.3	25.6
Gao et al.	54.0	50.9	44.1	34.9	25.6
Xu et al.	66.0	59.4	51.9	41.0	29.8
Lin et al.	-	-	53.5	45.0	36.9
Chao et al.	59.8	57.1	53.2	48.5	42.8
Zeng et al.	69.5	67.8	63.6	57.8	49.1
YOLA	**76.6**	**75.3**	**72.1**	**67.4**	**58.3**

Fig.1. Per-class AP at tIoU threshold = 0.5 on THUMOS' 14.



Results

Competitive results on ActivityNet v1.3 dataset.

TABLE III THE EXPERIMENTAL RESULTS ON ActivityNet v1.3 DATASET.

tIoU	0.5	0.75	0.95	Average
Singh et al.	26.01	15.22	2.61	14.62
Singh et al.	22.71	10.82	0.33	11.31
Dai et al.	36.44	21.15	3.90	-
Chao et al.	38.23	18.30	1.30	20.22
Xu et al.	26.80	-	-	12.70
YOLA	**38.67**	19.74	1.73	**21.14**

Results

State-of-the-art results on Thumos' 14 datasets

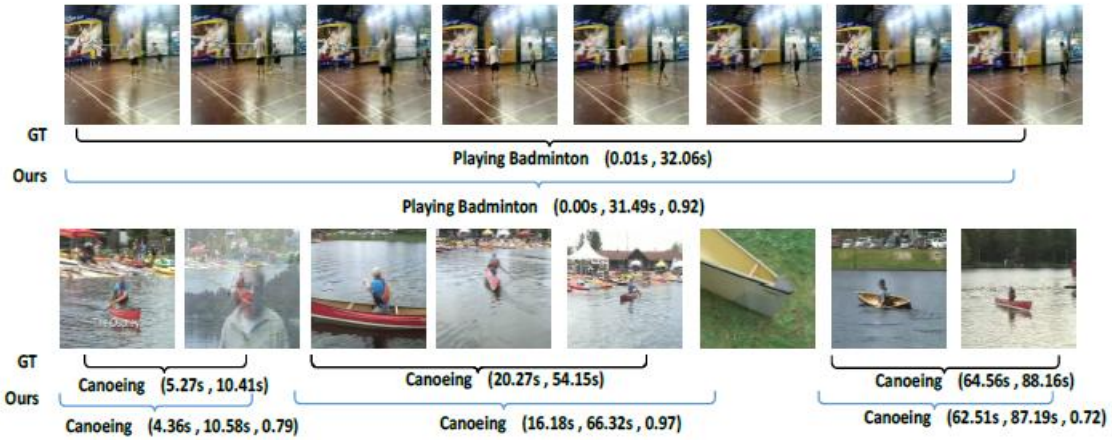
TABLE II RESULTS FOR INCORPORATING DIFFERENT DESIGNS ON Thumos' 14 DATASET.

I3D+LFE	√	√	√	√	√	√
SPN		√	√	√	√	√
Rectify			√	√	√	√
F' ori				√		√
F' weight					√	√
RGB	50.07	51.85	54.06	55.27	54.96	56.68
Flow	-	-	49.11	49.56	50.33	50.41
RGB+Flow	-	-	56.89	-	-	58.32

Results

Visualization of Thumos’ 14 and ActivityNet v1.3 datasets

(a) THUMOS’14



(b) ActivityNet v1.3

西安交通大學
XI'AN JIAOTONG UNIVERSITY

Thanks for listening!