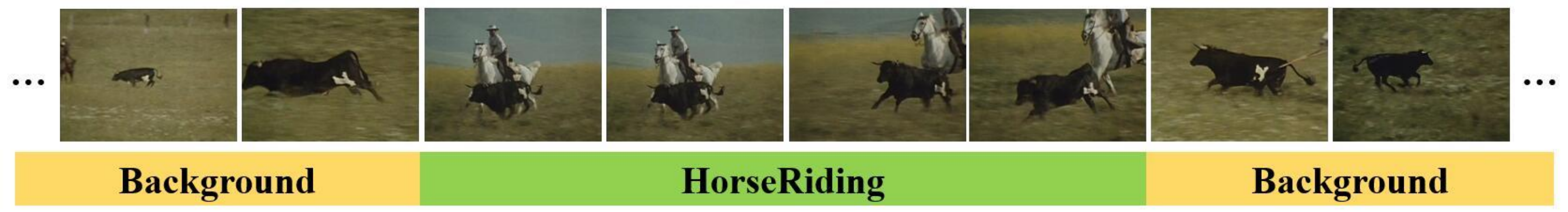
SHANGHAI JIAO TONG UNIVERSITY

TEMPORAL ACTION LOCALIZATION WITH TWO-STREAM SEGMENT-BASED RNN

Tianwei Lin, Xu Zhao*, Zhaoxuan Fan
Department of Automation, Shanghai Jiao Tong University, China

ICIP 2017
IEEE International Conference on Image Processing

Introduction



Temporal action localization

- ➢ Task: localize action instances in long untrimmed videos and classify their categories
- ➢ Challenges
- Videos to be analyzed are usually long and untrimmed
- A video may contains multiple action instances with different categories

● Our main contributions

- ➢ Propose a two-stream segment-based recurrent neural network (TSS-RNN) framework for temporal action localization task
- ➢ Propose a temporal segment proposal method combining multi-scale sliding window and temporal selective search
- ➢ Reach state-of-the-art performance in THUMOS'14 dataset

Proposed Approach

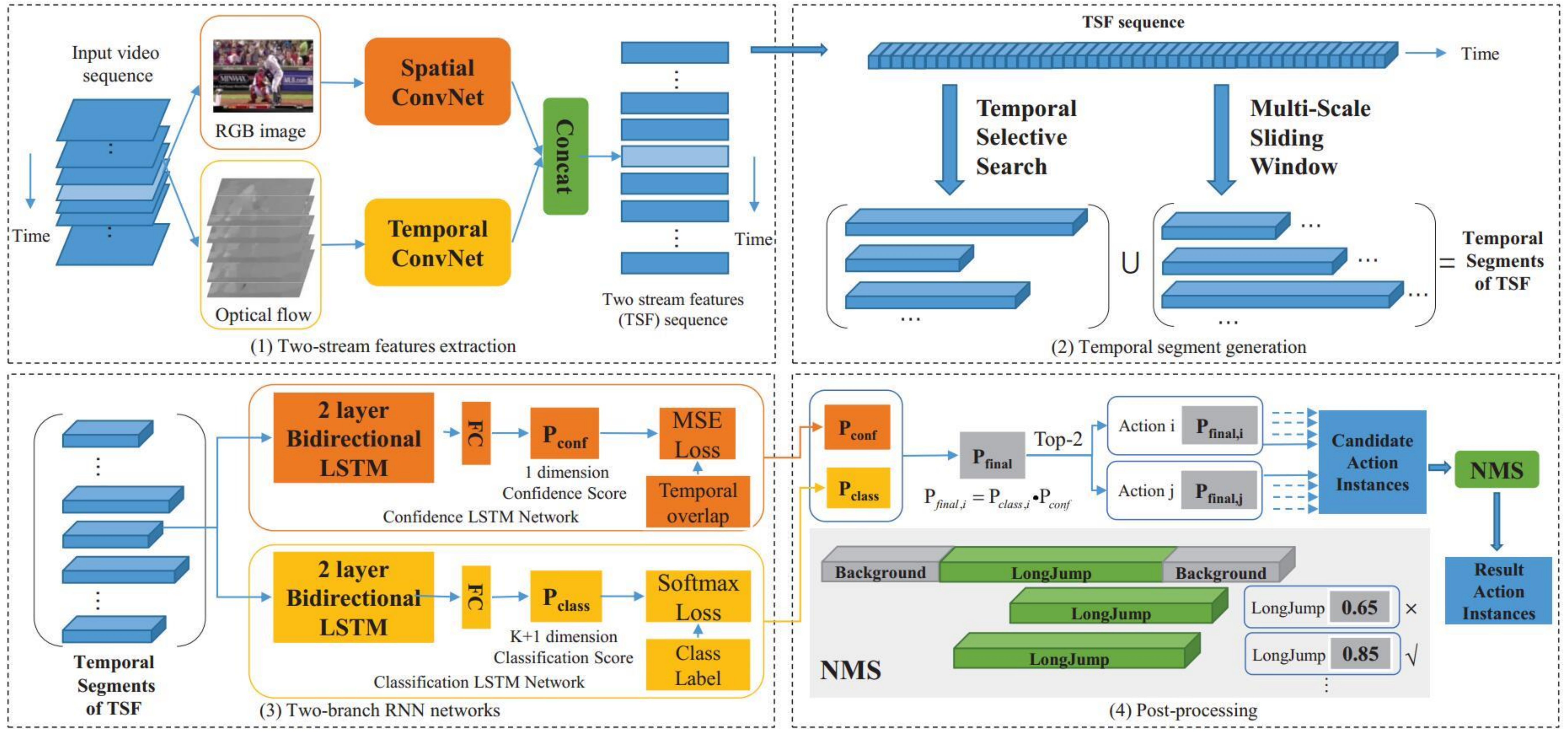
● Framework of TSS-RNN

(1) ***Two-stream features extraction***

- FC8 layer outputs of two stream networks are concatenated as Two-Stream Feature (TSF)
- Spatial network extracts appearance feature
- Temporal network extracts motion feature

(2) ***Temporal segment generation***

- Temporal selective search generate segments by merging mini-segments continually based on Manhattan distance
- Temporal selective search and multi-scale sliding window are combined to generate temporal segments



(3) ***Two-branch RNN networks***

- Both branches adopt bi-directional 2-layers LSTM
- Confidence network is trained for the IoU score regression

$$L_{conf} = \frac{1}{N}\sum_{1}^{N}(y_{pred} - y_u)^2 + \lambda \cdot L_2(\Theta_{conf})$$

- Classification network is trained for classifying action category

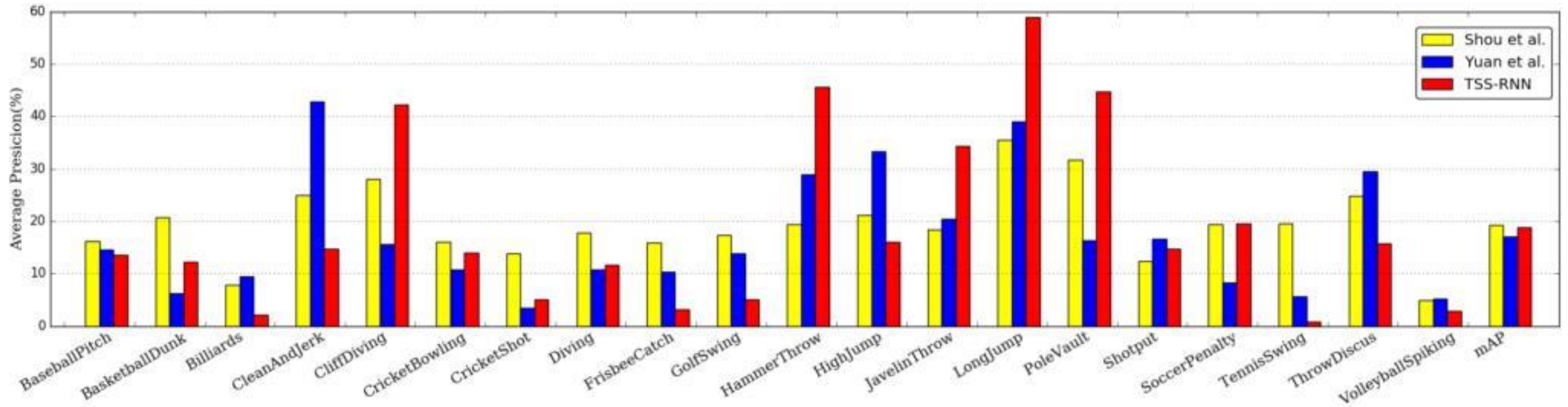
$$L_{class} = L_{softmax} + \lambda \cdot L_2(\Theta_{class})$$

(4) Post-Processing

- Combine classification and confidence score

$$P_{final} = P_{class} \cdot P_{conf}$$

- Non-maximum suppression (NMS) is used for removing redundant detections



Detection AP over different action classes of THUMOS'14 with overlap=0.5

Experiments

- ➢ Comparison of our approach with state-of the art on THUMOS '14 with variable IoU threshold (mAP %)

θ	0.5	0.4	0.3	0.2	0.1
Wang et al. [7]	8.5	12.1	14.6	17.8	19.2
Oneata et al. [5]	15.0	21.8	28.8	36.2	39.8
Yeung et al. [13]	17.1	26.4	36.0	44.0	48.9
Yuan et al. [6]	18.8	26.1	33.6	42.6	**51.4**
Shou et al. [9]	**19.0**	28.7	36.3	**43.5**	47.7
TSS-RNN(Ours)	18.8	**28.9**	**36.9**	42.9	46.1

- ➢ Comparison between two-branch TSS-RNN and one-branch classification TSS-RNN on THUMOS'14.

Networks	mAP ($\theta = 0.5$)
TSS-RNN(w/o confidence network)	9.8
TSS-RNN	**18.8**

- ➢ Comparisons between different components of our temporal segment generation method on THUMOS'14.

Method	mAP ($\theta = 0.5$)
Temporal Selective Search	14.3
Multi-Scale Sliding Window	17.8
Combined two methods	**18.8**

Conclusions

- The two-branch TSS-RNN architecture we proposed shows great performance in THUMOS'14 dataset
- The quality of temporal segment proposals have a great impact on the accuracy of temporal action localization

Recent Works

- ➢ We won the first place in both Temporal Action Proposal task and Temporal Action Localization task of ActivityNet Large Scale Activity Recognition Challenge 2017!
- ➢ Tianwei Lin, Xu Zhao, Zheng Shou. **Single Shot Temporal Action Detection.** ACM International Conference on Multimedia (**ACMMM**). Mountain View, U.S., 2017.