Temporal Segment Networks: Towards Good Practices for Deep Action Recognition

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, Luc Van Gool

ETH Zurich The Chinese University of Hong Kong Shenzhen Institutes of Advanced Technology, CAS

Motivation

- Modeling long-range temporal structure is crucial for human activity recognition.
- Frames in a video are highly redundant.

Modeling long-range temporal structure is not simply wrapping tons of frames.

Frames are dense, but contents are sparse!

Code Release

Temporal Segment Networks

An open source action recognition framework

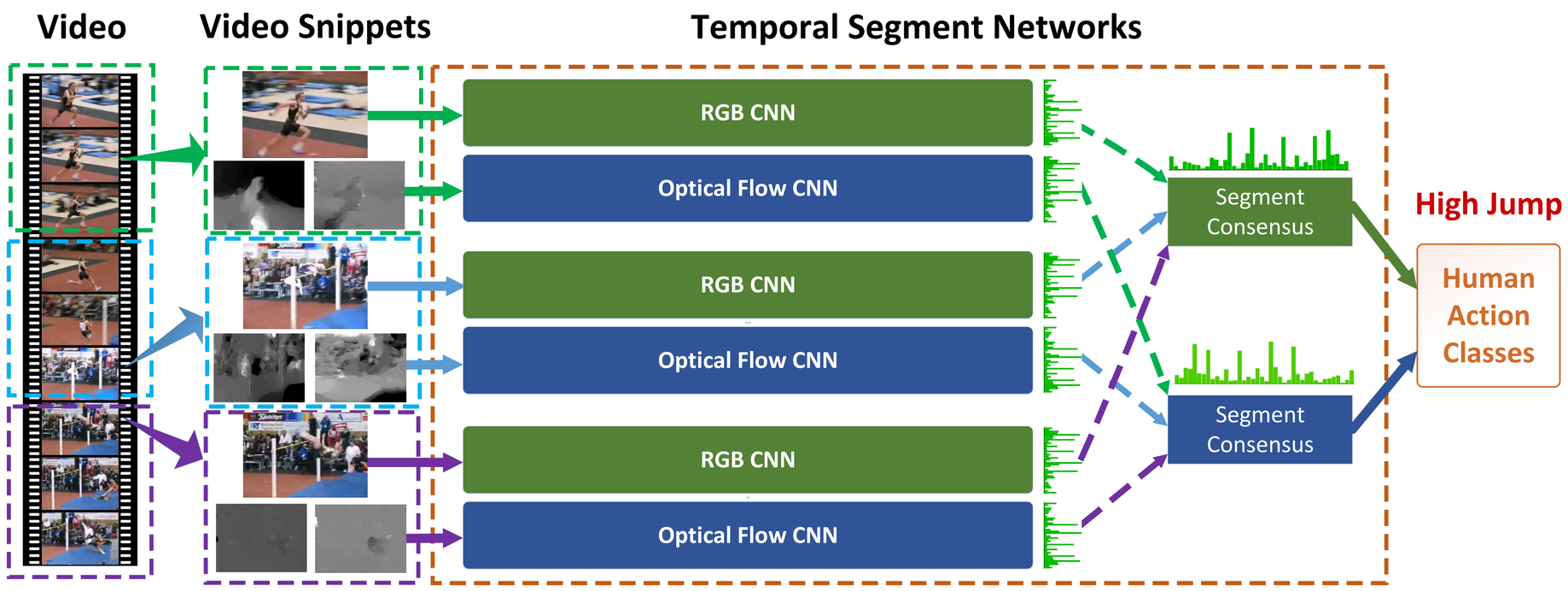
Winner of ActivityNet 2016 (93.2% mAP)

https://github.com/yjxiong/temporal-segment-networks

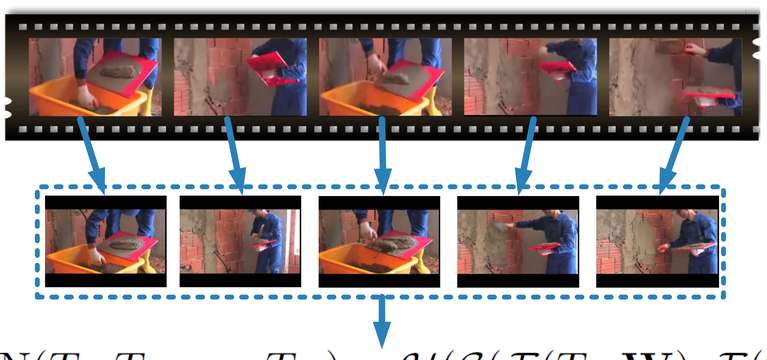
Temporal Segment Networks (TSN): The Model

Training TSN

1. Divide one video into a fixed number of segments
2. Sample snippets from the segments
3. Optimize the classification loss based on segment consensus



Segment-Based Sparse Sampling



$\mathcal{H}$ Softmax function

$\mathcal{G}$ Segment consensus function

$\mathcal{F}$ ConvNet (RGB/optical flow)

$$\mathrm{TSN}(T_1, T_2, \cdots, T_K) = \mathcal{H}(\mathcal{G}(\mathcal{F}(T_1; \mathbf{W}), \mathcal{F}(T_2; \mathbf{W}), \cdots, \mathcal{F}(T_K; \mathbf{W})))$$

Segment Consensus

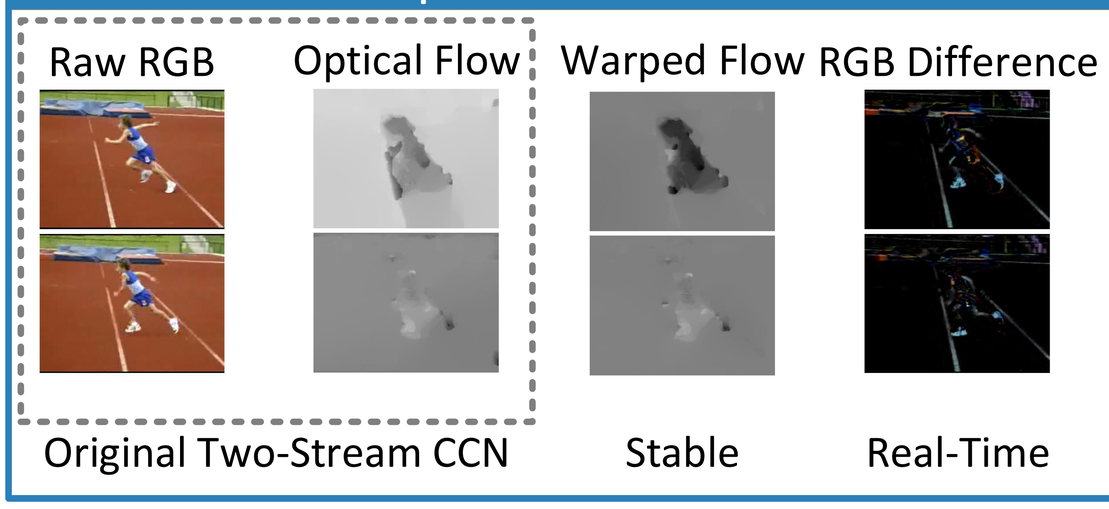
- Predict after observing all segments
- video level supervision instead of frame-wise

$$\mathcal{L}(y, \mathbf{G}) = -\sum_{i=1}^{C} y_i \left(G_i - \log \sum_{j=1}^{C} \exp G_j \right) \qquad \frac{\partial \mathcal{L}(y, \mathbf{G})}{\partial \mathbf{W}} = \frac{\partial \mathcal{L}}{\partial \mathbf{G}} \sum_{k=1}^{K} \frac{\partial \mathcal{G}}{\partial \mathcal{F}(T_k)} \frac{\partial \mathcal{F}(T_k)}{\partial \mathbf{W}}$$

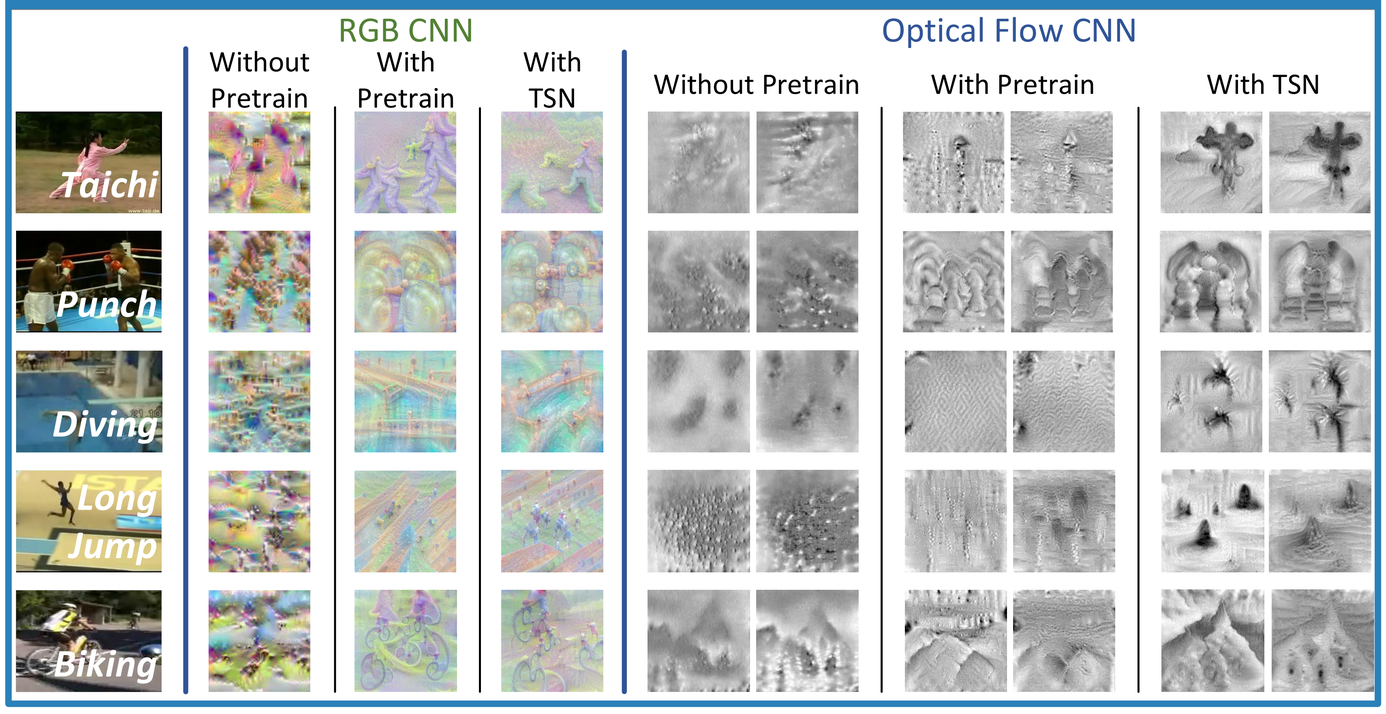
Choices for $\mathcal{G}$

Max Average Weighted More?

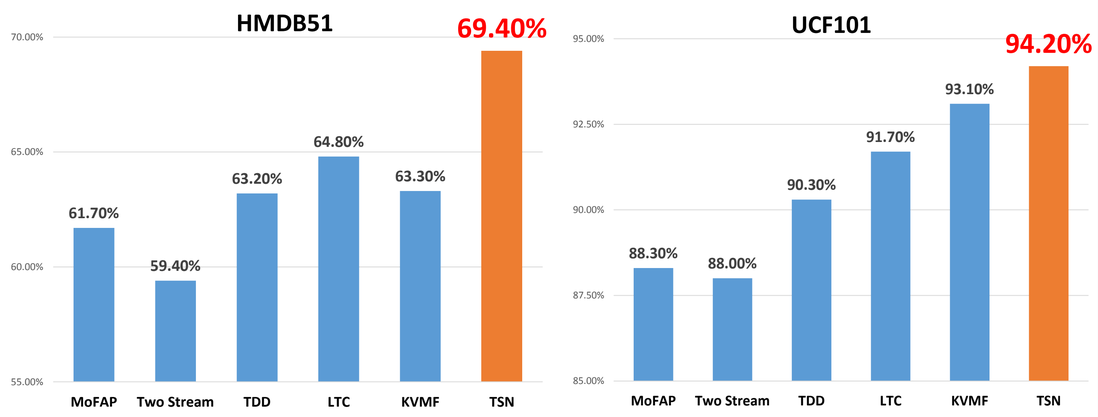
Input Modalities



Visualized Trained CNNs



Experimental Results



HMDB51		UCF101	
TSN (2 modalities)	68.5%	TSN (2 modalities)	94.0%
TSN (3 modalities)	**69.4%**	TSN (3 modalities)	**94.2%**

Component Analysis: Identify the Good Practices

On UCF101-Split 1, adding components one by one

Component	Basic Two-Stream [1]	Cross-Modality Pre-training	Partial BN with dropout	Temporal Segment Networks
Accuracy	90.0%	91.5	92.0%	93.5%

[1] Simonyan Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos.", NIPS 2014.
[2] Joe Yue-Hei Ng, et al. "Beyond short snippets: Deep networks for video classification." CVPR 2015.
[3] Jeffrey Donahue, et al. "Long-term recurrent convolutional networks for visual recognition and description." CVPR. 2015.