Spatio-Temporal Fusion Networks for Action Recognition

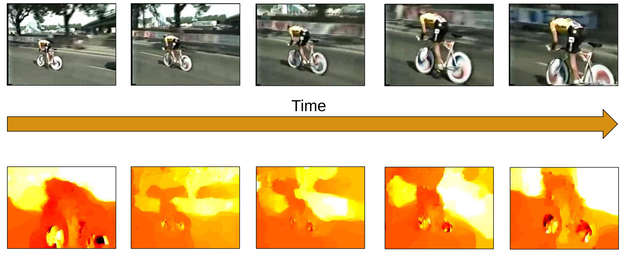
Computational Imaging Lab. (CIL)
Department of Computer Science

Sangwoo Cho, Hassan Foroosh (swcho@knights.ucf.edu, foroosh@cs.ucf.edu) Dept. of Computer Science, University of Central Florida

ACCV 2018
2 – 6 December 2018 Perth Western Australia

Motivation

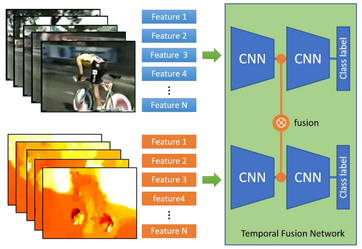
Temporal Dynamics



Actions can be represented with temporal evolution of features.

Q) How to extract and take advantage of them?
A) Temporal convolution (1D convolution)
A) Fusion strategy for better representation

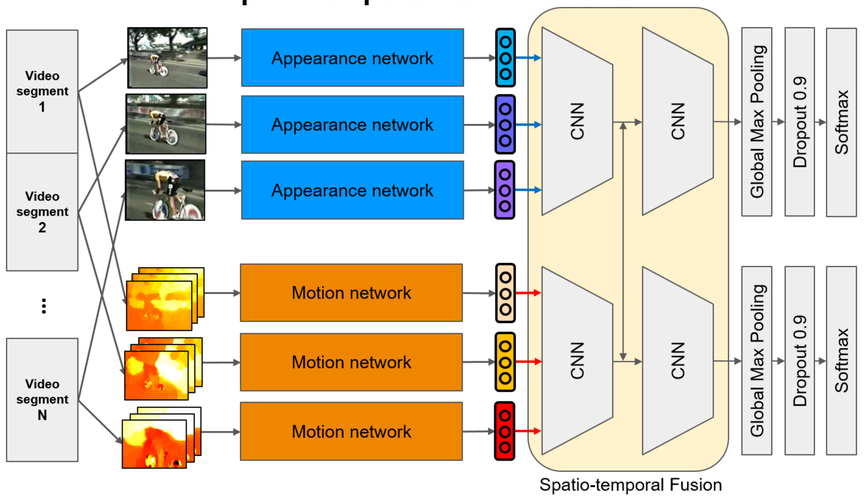
System overview



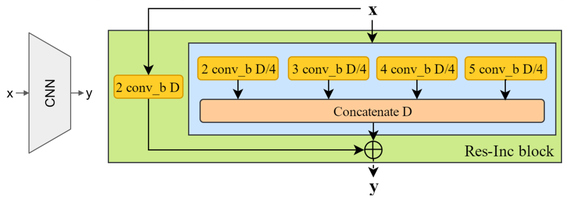
- **Feature learning and extraction using CNN (Resnet-101, Inception-V3)**
- **Temporal dynamic information extraction using Temporal res-inception modules**
- **Fusion of two temporal dynamics using consecutive res-inception blocks**

Method

Spatio-Temporal Fusion Networks

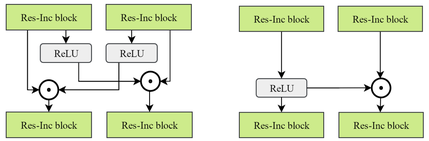


Temporal dynamic extraction module (Temporal Res-Inception)



- ✓ 1D convolution + BN + relu
  - ✓ Extract temporal info.
- ✓ Different kernel sizes (2,3,4,5)
  - ✓ Extract diff. temporal outputs
  - ✓ Resnet + Inception

Spatio-Temporal Fusion Block (STFB)



- ✓ **Bi-direction(best)** / Uni-direction
- ✓ Fusion operation
  - ✓ **Avg(best)**, Max, Multiply
- ✓ Num. of video segments
  - ✓ Num. of inputs for Res-Inc: **7(best)**

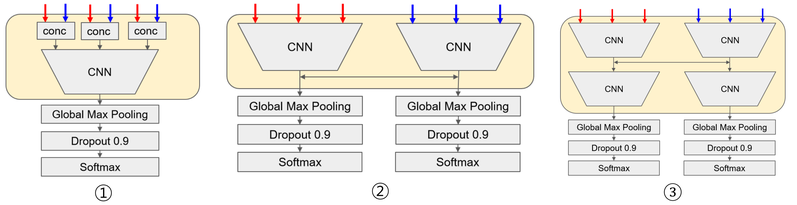
Experimental results

Baseline performance (no STFB)

✓ R: Resnet-101, I-v3: Inception-v3 / 5 segments

	HMDB51		UCF101	
	R	I-v3	R	I-v3
Spatial	48.2	51.2	83.5	84.8
Temporal	58.1	59.2	86.0	88.1
Late fusion	61.1	**62.7**	91.8	**92.3**

Fusion	HMDB51 R	UCF101 R
① (feat. fusion)	69.2	92.0
② (1 block)	69.6	93.2
③ (2 blocks)	**70.4**	**93.5**



State of the Art

	HMDB51	UCF101	Note
iDT+FV	57.2	85.9	Hand crafted feature
Two stream	59.4	88.0	Late fusion of RGB and OF(optical flow)
ActionVLAD	66.9 / 69.8	92.7 / 93.6	Avg. w/ iDT features
ST-Resnet	68.9 / 72.2	93.4 / 94.6	Resnet-50 fusion using addition
ST-Multiplier	68.9 / 72.2	94.2 / 94.9	Resnet-152 fusion using multiplication
I3D (Imagenet)	66.4	93.4	Inception-v1 w/ pre-trained on Imagenet
I3D (Kinetics)	80.9	97.8	pre-trained on Kinetics, 240k train videos
TSN	71.0	94.9	7 segments, 3 modalities
Four-Stream	72.5 / 74.9	95.5 / 96.0	RGB+OF+dynamic images(RGB, OF)
OFF	74.2	96.0	RGB+OF+OFF(RGB, OF)
STFN (Resnet-101)	71.2 / 73.3	94.3 / 95.1	RGB+OF (Bi-dir, Avg, 7 seg, 2 blocks)
STFN (Inception-v3)	**72.1 / 75.1**	**95.4 / 96.0**	Avg. w/ MIFS features

- ActionVLAD: Learning spatio-temporal aggregation for action classification, CVPR17
- St-Resnet: Spatiotemporal residual networks for video action recognition, NIPS16
- ST-Multiplier: Spatiotemporal multiplier networks for video action recognition, CVPR17
- I3D: Quo Vadis, Action Recognition, CVPR17
- TSN: Temporal Segment Networks, PAMI18
- Four-Stream: Action recognition with dynamic image networks, PAMI18
- OFF: Optical flow guided feature, CVPR18