LONG-TERM TEMPORAL CONVOLUTIONS FOR ACTION RECOGNITION

Gül Varol, Ivan Laptev, Cordelia Schmid - INRIA, France

{gul.varol,ivan.laptev,cordelia.schmid}@inria.fr

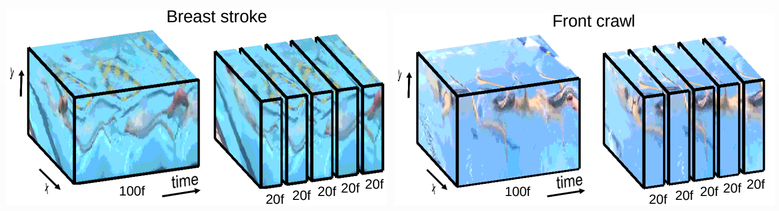
1. Introduction

Goal.

- Human action recognition in video.

Motivation.

- Human actions contain long-term temporal structure.
- Current CNNs learn spatio-temporal structure at the level of a few video frames → failing to model actions at their full temporal extent.

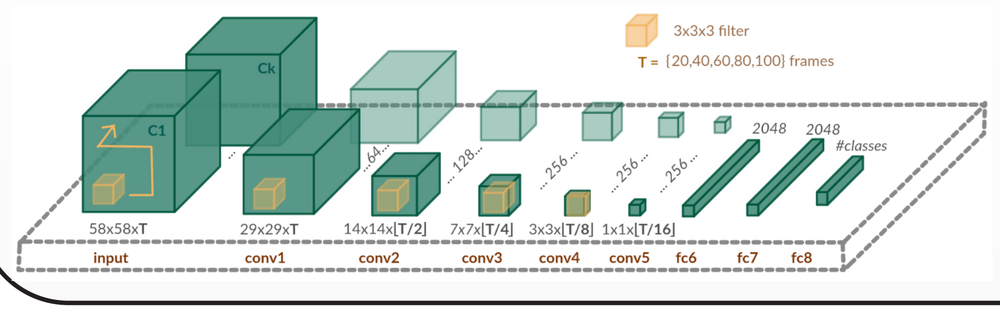


Contributions.

- Learning video representations using 3D CNNs with *long-term temporal convolutions* (LTC).
- Studying the impact of alternative inputs: Optical Flow of different quality and RGB.
- State-of-the-art results: UCF101 (92.7%), HMDB51 (67.2%)

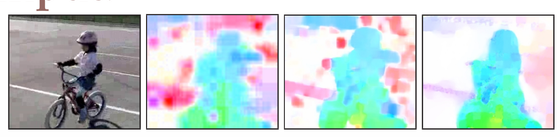
2. Network

- 3D convolutions over large number of video frames.
- Increased temporal extent by the cost of decreased spatial resolution.
- 2-channel optical flow or 3-channel RGB as input.



3. Experiments

Input.



RGB	MPEG	Farneback	Brox
57.0%	58.5%	66.3%	74.8%
59.9%	63.8%	71.3%	79.6%

UCF101 (split 1).

Data augmentation.

Random clipping	Multiscale cropping	Dropout	Clip	Video
-	-	0.5	71.6	76.5
✓	-	0.5	74.8	79.6
-	✓	0.5	72.5	78.1
-	-	0.9	74.4	78.5
✓	✓	0.9	76.3	80.5

UCF101 (split 1).

16f vs 60f networks.

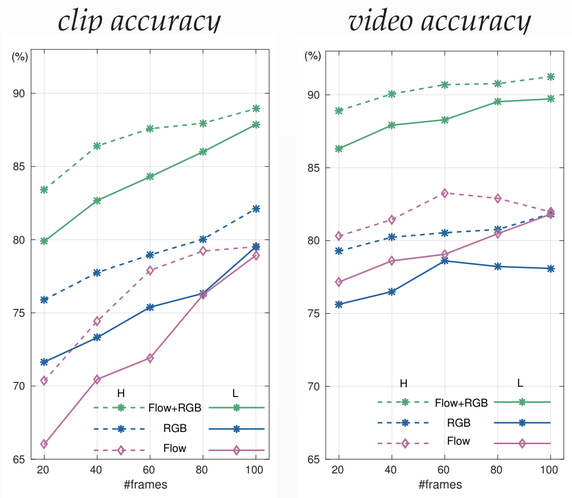
Input		16f	60f	gain
RGB	clip	48.4	57.0	+ 8.6
	video	51.9	59.9	+ 8.0
Flow	clip	67.1	76.3	+ 9.1
	video	78.7	80.5	+ 1.8

UCF101 (split 1).

Pre-training		16f	60f	gain	[3]
Flow	clip	37.0	52.6	+ 15.6	
from scratch	video	43.9	52.9	+ 9.0	46.6
Flow	clip	40.6	56.1	+ 15.5	
from UCF101	video	48.3	57.1	+ 8.8	49.0

HMDB51 (split 1).

Temporal resolution.

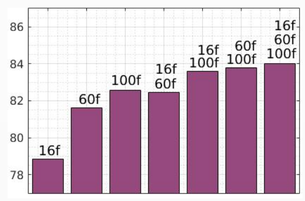


UCF101 (split 1).

Combining networks.

	UCF101	HMDB51
LTC_{Flow} (100f)	82.6	56.7
LTC_{Flow} (60f+100f)	83.8	60.5
LTC_{RGB} (100f)	81.8	-
LTC_{RGB} (60f+100f)	81.5	-
$\text{LTC}_{Flow+RGB}$	91.0	65.6
$\text{LTC}_{Flow+RGB}$+IDT	91.8	67.7

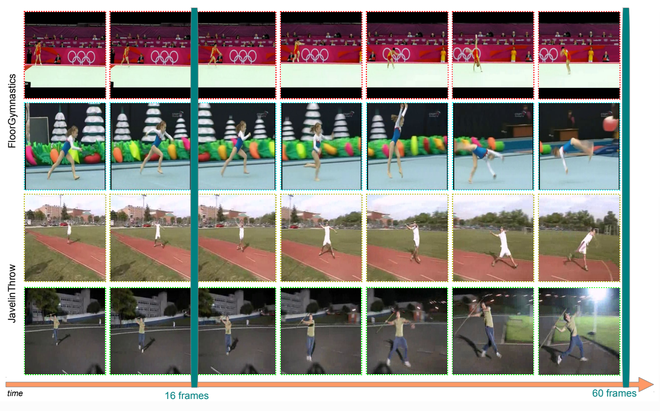
- Long temporal extent 👍
- High spatial resolution 👍
- RGB+Flow complementary
- RGB > Flow (clips)
- Flow > RGB (videos)
- Curves less steep for video



UCF101 (split 1).

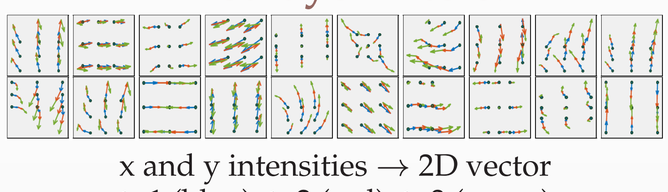
4. Qualitative analysis

Classes w/ largest improvement.



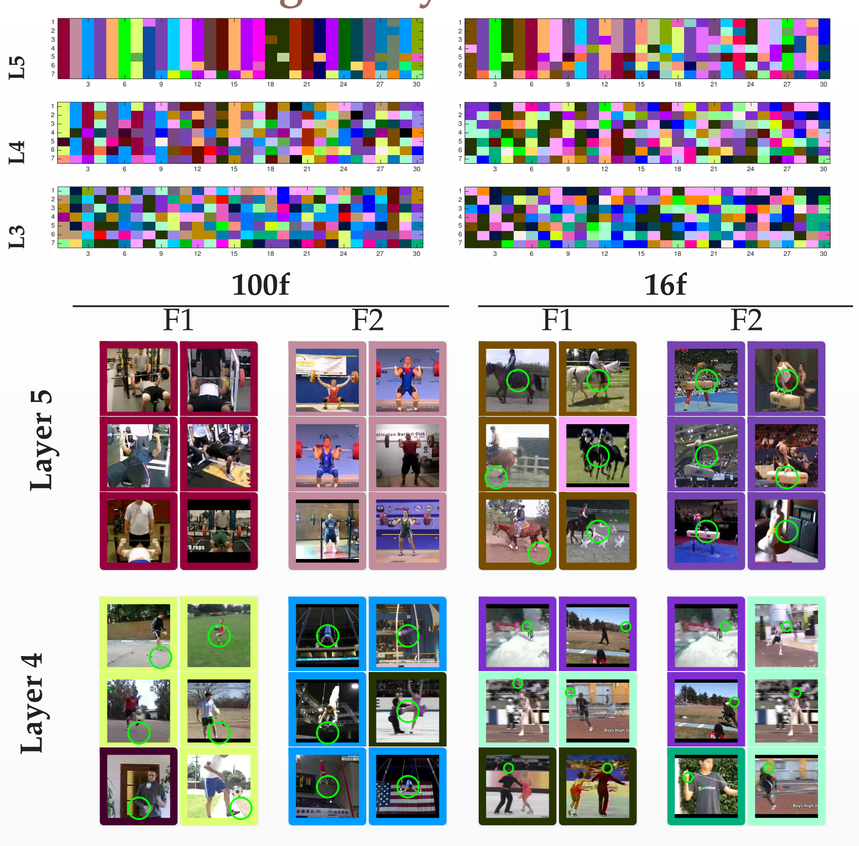
Accuracy of the 'JavelinThrow' class increased from 54.8% (16f) to 96.8% (60f), while mostly being confused with 'FloorGymnastics' in 16f.

First layer filters.



x and y intensities → 2D vector
t=1 (blue), t=2 (red), t=3 (green)

Higher layer filters.



5. Results

State-of-the-art.

LTC outperforms previously published results.

		Method	UCF101	HMDB51
Hand crafted	[5]	IDT+FV	85.9	57.2
	[1]	IDT+MIFS	89.1	65.1
CNN (RGB)	[3]	Spatial stream	73.0	40.5
	[4]	C3D (1 net)	82.3	-
		LTC_{RGB}	**82.4**	-
	[4]	C3D (3 nets)	85.2	-
CNN (Flow)	[3]	Temporal str.	83.7	54.6
		LTC_{Flow}	**85.2**	**59.0**
Fusion	[3]	Two-stream (avg. fusion)	86.9	58.0
	[3]	Two-stream (SVM fusion)	88.0	59.4
	[2]	LSTM	88.6	-
	[4]	C3D+IDT	90.4	-
		$\text{LTC}_{Flow+RGB}$	**91.7**	**64.8**
		$\text{LTC}_{Flow+RGB}$+IDT	**92.7**	**67.2**

3 splits average.

Conclusions.

We show

1) the advantages of learning long-term temporal convolutions,

2) the importance of high-quality optical flow estimation

for learning accurate video representations.

References

[1] Lan et al. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In *CVPR*, 2015.

[2] Ng et al. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015.

[3] Simonyan et al. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014.

[4] Tran et al. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015.

[5] Wang et al. Action recognition with improved trajectories. In *ICCV*, 2013.

Project page

www.di.ens.fr/willow/research/ltc

> code in Torch
> pre-trained CNN models
> paper, slides, poster
> supplementary video