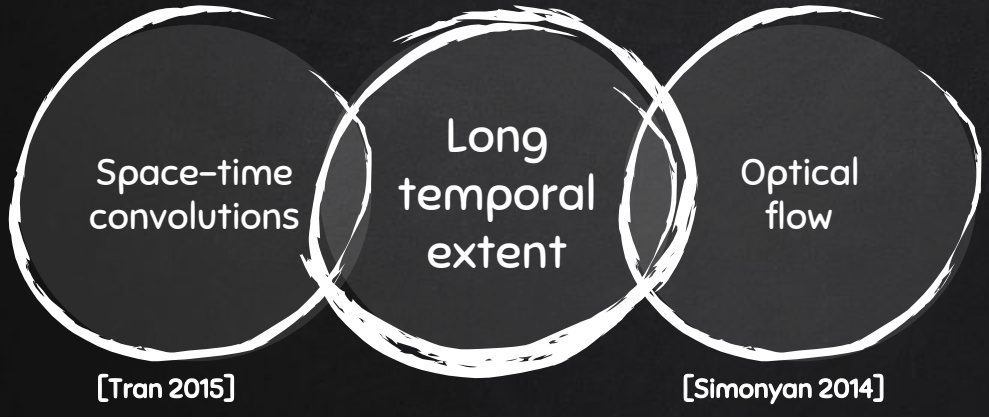
# Long-term Temporal Convolutions for Action Recognition

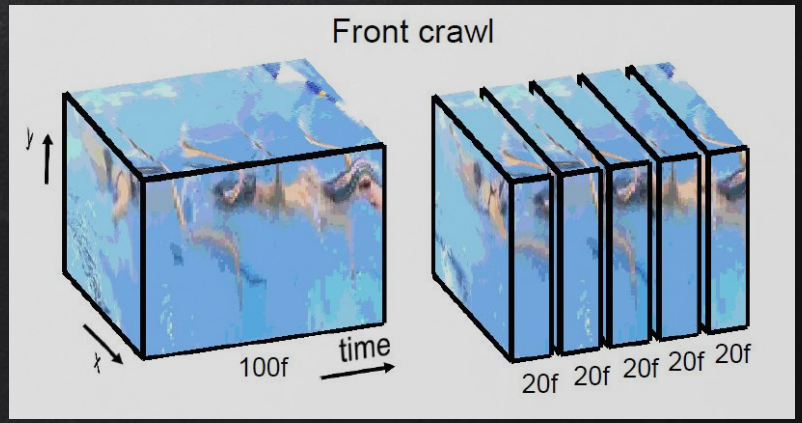
Gul Varol  Ivan Laptev  Cordelia Schmid

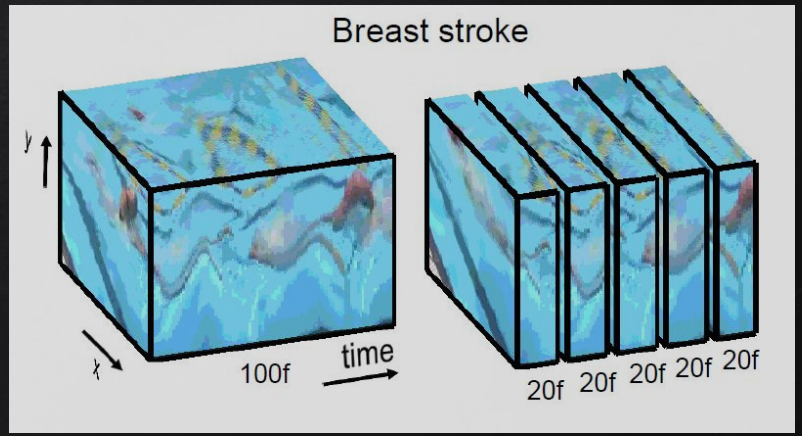
INRIA

# Motivation

- Current CNN methods for action recognition learn representations for short intervals (1–16 frames).
- Typical actions last several seconds.
- Actions contain characteristic patterns with specific long-term temporal structure.



# Contributions

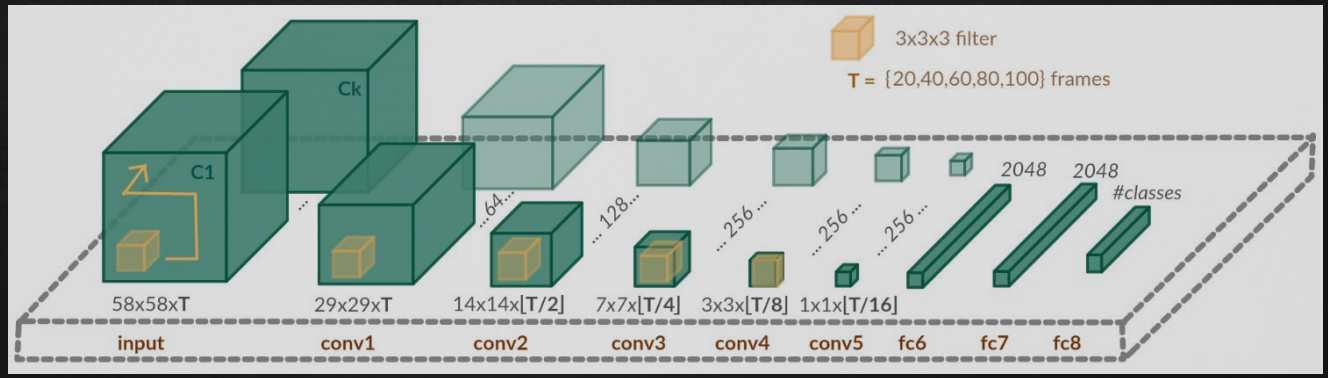
- The advantages of **long-term temporal convolutions**
- The importance of **high-quality optical flow** estimation

for learning accurate video representations.

# Approach

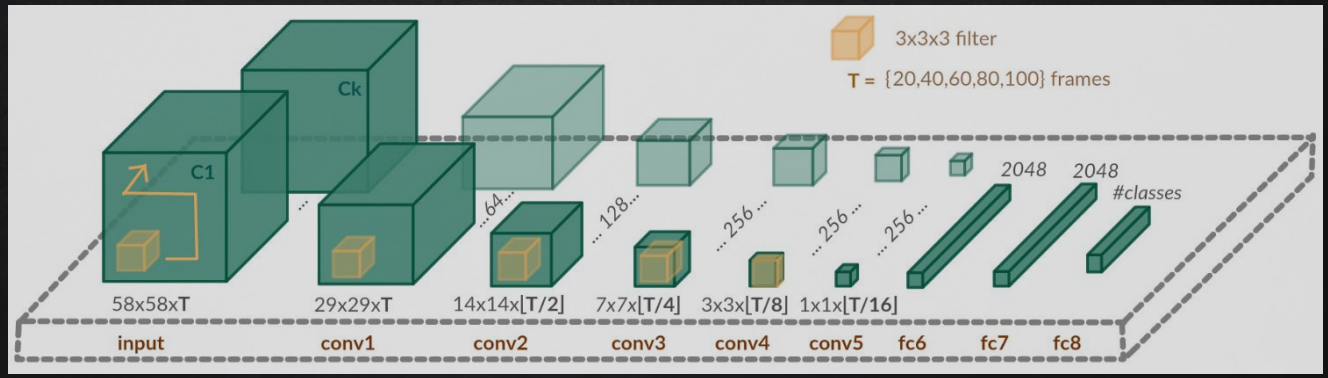
# Network Architecture

- 3D convolutions with 3x3x3 filters
- ReLU
- 3D max-pooling of 2x2x2
- Experiments with T = {16, 20, 40, 60, 80, 100}



# Network Architecture

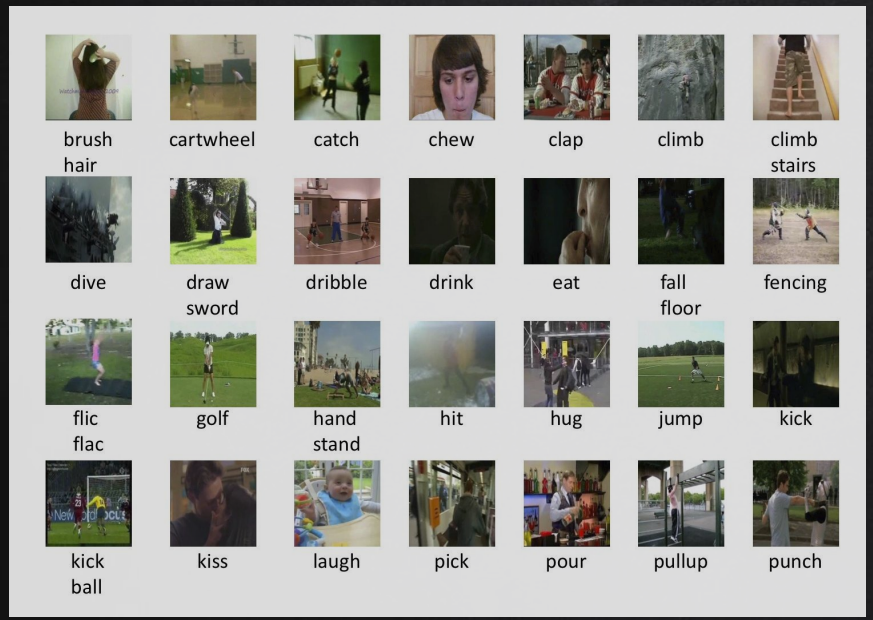
- Optical flow : 2-channel input (original [x, y] values)
- RGB : 3-channel input
- Increased temporal extent by the cost of decreased spatial resolution.



# Experiments

# Datasets

## HMDB51 (Kuehne et al. 2011)




## UCF101 (Soomro et al. 2012)

# Optical Flow

- 60-frame training from scratch
- With different input modalities

Conclusions

1. Even low-quality MPEG flow outperforms RGB.
2. Quality of flow impacts the results significantly.

RGB | MPEG flow | Farneback flow | Brox flow

| Input | Clip | Video |
|---|---|---|
| RGB | 57.0 | 59.9 |
| MPEG flow | 58.5 | 63.8 |
| Farneback | 66.3 | 71.3 |
| Brox | 74.8 | 79.6 |

*60-frame networks from scratch on UCF101 (split 1)*

# 16- vs 60-frame networks

spatial res. 112x112 58x58

16f network has the same architecture as Tran 2015.

| 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 | [Simonyan 2014] |
|---|---|---|---|---|---|
| Flow from scratch | Clip | 37.0 | 52.6 | +15.6 | - |
| | Video | 43.9 | 52.9 | +9.0 | 46.6 |
| Flow from UCF101 | Clip | 40.6 | 56.1 | +15.5 | - |
| | Video | 48.3 | 57.1 | +8.8 | 49.0 |

*HMDB51 (split 1)*

# RGB Network Finetuning

- RGB from scratch is difficult to learn
- We need pre-training

| | Clip | Video |
|---|---|---|
| 16f | 48.4 | 51.9 |
| 60f | 57.0 | 59.9 |
| 100f | 62.9 | 65.4 |

*UCF101 (split 1) RGB from scratch*

- C3D → 16f 3D convnet trained on Sports-1M (Tran 2015)
- We extend C3D to longer temporal convolutions as follows:
  - Conv5 layer output has T/16 temporal resolution.
  - Max-pool conv5 output over time to re-cycle pre-trained fc layers.
  - Finetune whole network.

# Varying Temporal and Spatial Resolutions

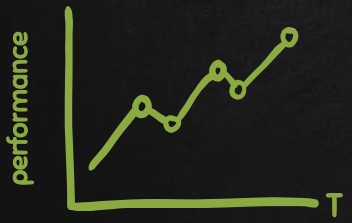
(dotted ..) High resolution (71x71)

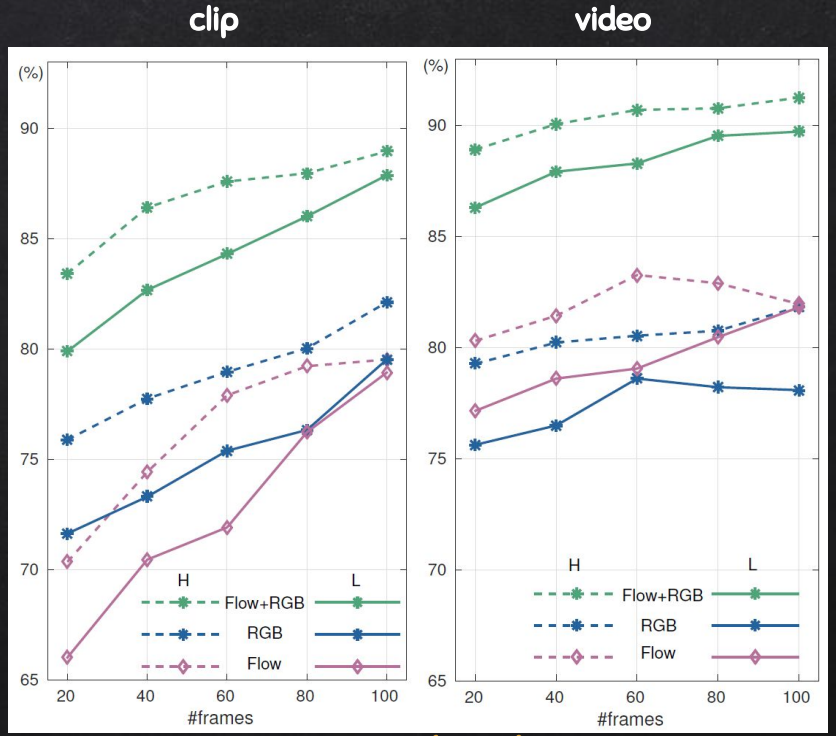
(plain –) Low resolution (58x58)

(pink) Flow from scratch

(blue) RGB from C3D

Conclusion



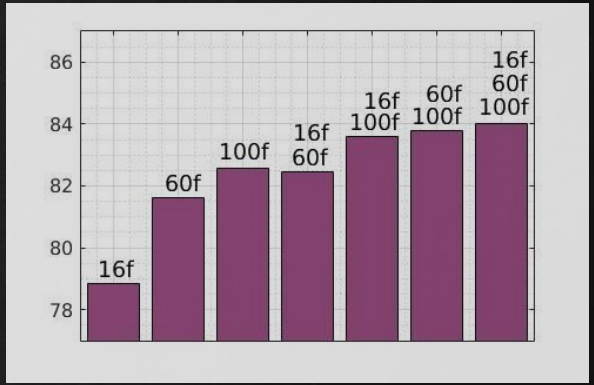

UCF101 (split 1)

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

# Multiple Networks Combined

| Input | 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 | – |
| $LTC_{Flow+RGB}$ | 91.0 | 65.6 |
| $LTC_{Flow+RGB}$ + IDT | 91.8 | 67.7 |

*split 1*




*UCF101 (split 1) flow*

# Comparison to the State-of-the-art

*3 splits average*

| | Method | UCF101 | HMDB51 |
|---|---|---|---|
| Hand-crafted | [Wang 2013] IDT+FV | 85.9 | 57.2 |
| | [Peng 2014] IDT+HSV | 87.9 | 61.1 |
| | [Lan 2015] IDT+MIFS | 89.1 | 65.1 |
| | [Peng 2014] IDT+SFV | – | 66.8 |
| CNN (RGB) | [Karpathy 2014] Slow fusion (scratch) | 41.3 | – |
| | [Tran 2015] C3D (scratch) | 44 | – |
| | [Karpathy 2014] Slow fusion | 65.4 | – |
| | [Simonyan 2014] Spatial stream | 73.0 | 40.5 |
| | [Tran 2015] C3D (1 net) | 82.3[1] | – |
| | **$LTC_{RGB}$** | **81.5** | **49.7[2]** |
| | [Tran 2015] C3D (3 nets) | 85.2 | – |
| CNN (Flow) | [Simonyan 2014] Temporal stream | 83.7[3] | 54.6[3] |
| | **$LTC_{Flow}$** | **85.2** | **59.0** |
| Fusion | [Simonyan 2014] Two-stream(avg) | 86.9 | 58.0 |
| | [Simonyan 2014] Two-stream(SVM) | 88.0 | 59.4 |
| | [Ng 2014] LSTM (flow+RGB) | 88.6 | – |
| | [Wang 2015] TDD | 90.3 | 63.2 |
| | [Tran 2015] C3D+IDT | 90.4 | – |
| | [Wang 2015] TDD+IDT | 91.5 | 65.9 |
| | **$LTC_{Flow+RGB}$** | **91.7** | **64.8** |
| | **$LTC_{Flow+RGB}$ + IDT** | **92.7** | **67.2** |

[1] Our implementation is 80.2% [2] No finetuning [3] Uses multi-task learning

# Qualitative Analysis

# Classes with Largest Improvement

| | 16f | 60f |
|---|---|---|
| JavelinThrow | 54.8 | 96.8 |

*JavelinThrow is mostly confused with FloorGymnastics in 16f.

FloorGymnastics =
running + gymnastics

JavelinThrow =
running + throwing javelin

# First Layer Filters

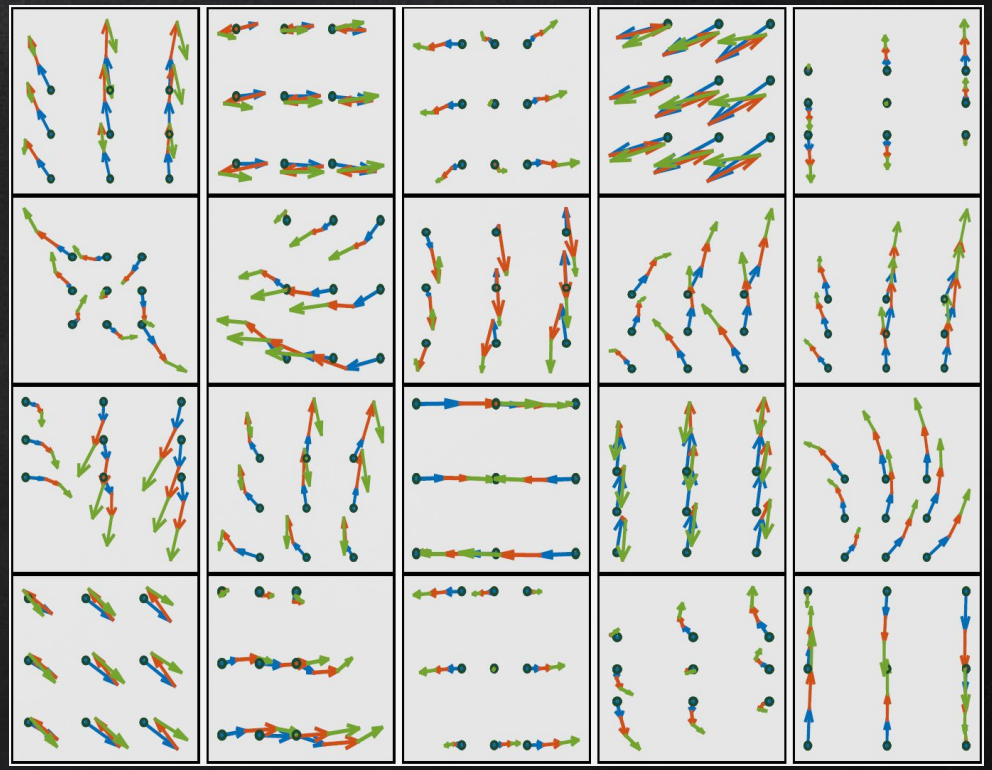
Complex motion patterns in local neighborhoods

x and y intensities → 2D vectors

t=1 blue

t=2 red

t=3 green



*60f Flow on UCF101 (split 1)*

# Higher Layer Filters

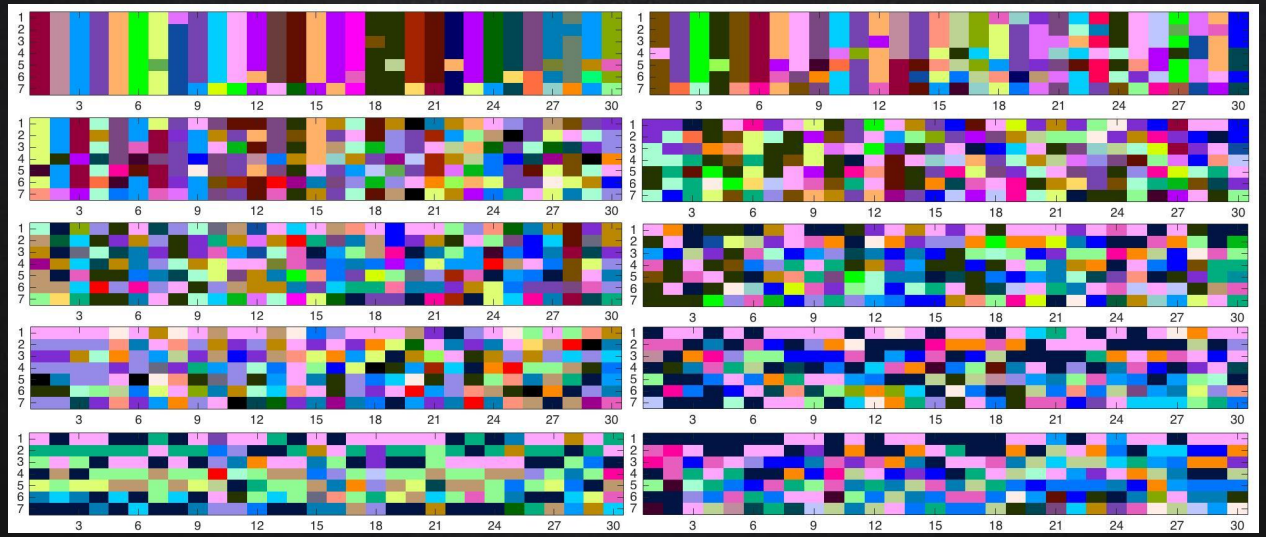
Video

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

Top activations of filters at conv layers.

Colors: classes, Rows: maximum responding test videos, Columns: filters.

100f 16f



# THANKS!

## Questions?

project page : www.di.ens.fr/willow/research/ltc/
contact : gul.varol@inria.fr

# Credits

Special thanks to all the people who made and released these awesome resources for free:

- Presentation template by SlidesCarnival
- Photographs by Unsplash