CNN for Video Recognition

Liangliang Cao

http://llcao.net/cu-deeplearning15/

Papers at a glance

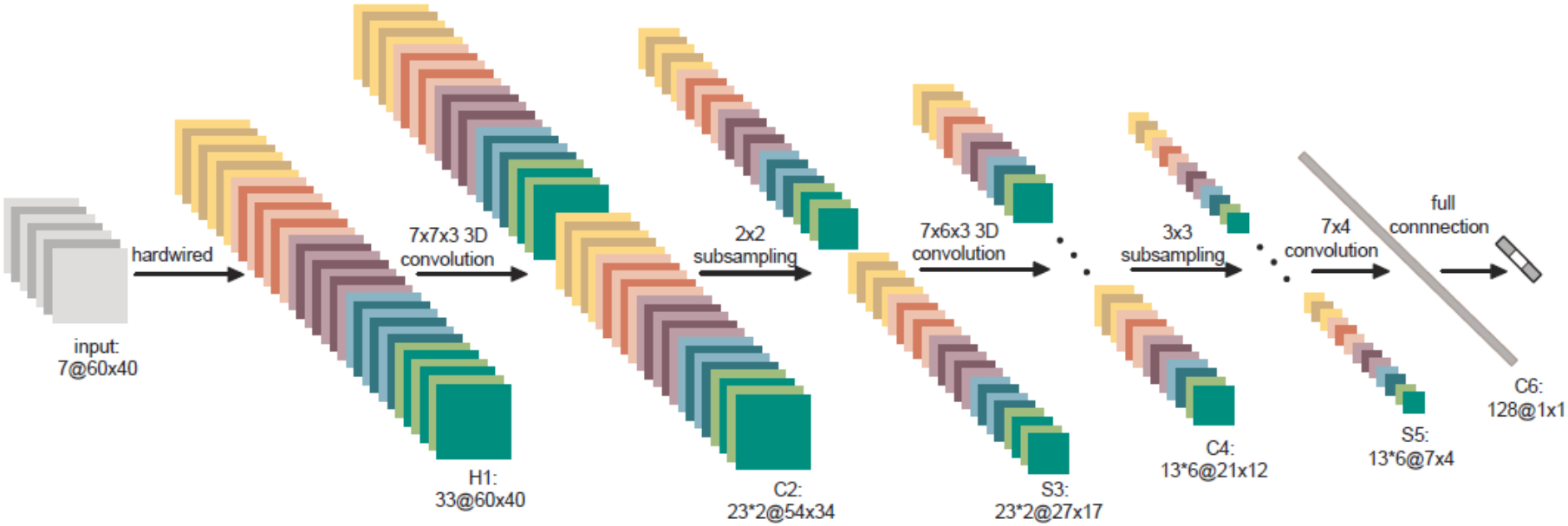
- Shuiwang Ji, Wei Xu, Ming Yang, Kai Yu, 3D Convolutional Neural Networks for Human Action Recognition, ICML 2010

- Andrej Karpathy et al, Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

- Karen Simonyan and Andrew Zisserman, Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014

[Ji et al, ICML 2010]

Network over 7-sequential frames:



[Ji et al, ICML 2010] TRECVID

Method	FPR	Measure	CellToEar	ObjectPut	Pointing	Average
3D CNN	0.1%	Precision	**0.6433**	**0.6748**	0.8230	**0.7137**
		Recall	**0.0282**	**0.0256**	0.0152	**0.0230**
		AUC($\times 10^3$)	**0.0173**	**0.0139**	0.0075	**0.0129**
	1%	Precision	**0.4091**	**0.5154**	0.7470	**0.5572**
		Recall	**0.1109**	**0.1356**	0.0931	**0.1132**
		AUC($\times 10^3$)	**0.6759**	**0.7916**	0.5581	**0.6752**

Observation:

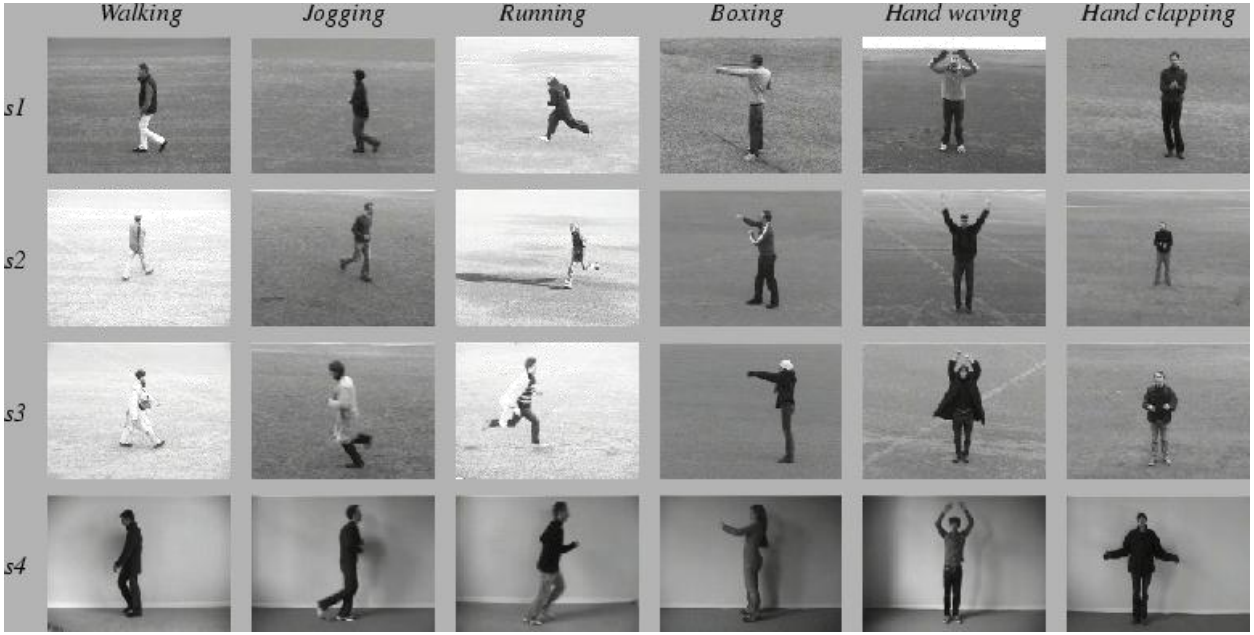
- Detection- framework
- TRECVID SED is difficult:
  - A lot of negatives
  - Overwhelming false alarms

An example of cellphone2ear

[Ji et al, ICML 2010] KTH

Acc on KTH

METHOD	AVERAGE
3D CNN	90.2
SCHÜLDT	71.7
DOLLÁR	81.2
NIEBLES	83.3
JHUANG	91.7
SCHINDLER	92.7



Why not good enough?

- xyt-CNN is not powerful enough?
- Small training samples (2.3K training samples)

[Karpathy CVPR14]

- Results on Youtube sports 1M

Model	Clip Hit@1	Video Hit@1	Video Hit@5
Feature Histograms + Neural Net	-	55.3	-
Single-Frame	41.1	59.3	77.7
Single-Frame + Multires	**42.4**	**60.0**	**78.5**
Single-Frame Fovea Only	30.0	49.9	72.8
Single-Frame Context Only	38.1	56.0	77.2
Early Fusion	38.9	57.7	76.8
Late Fusion	40.7	59.3	78.7
Slow Fusion	**41.9**	**60.9**	**80.2**
CNN Average (Single+Early+Late+Slow)	41.4	63.9	82.4

10 frames video clips

[Karpathy CVPR14] UCF 101

- Transfer learning results are not great:

Model	3-fold Accuracy
Soomro et al [22]	43.9%
Feature Histograms + Neural Net	59.0%
Train from scratch	41.3%
Fine-tune top layer	64.1%
Fine-tune top 3 layers	**65.4**%
Fine-tune all layers	62.2%

Table 3: Results on UCF-101 for various Transfer Learning approaches using the Slow Fusion network.

[Simonyan and Zisserman, NIPS 2014]

- Best performed CNN for video recognition
- Easiest to implement
- Not really 3D-CNN but 2D

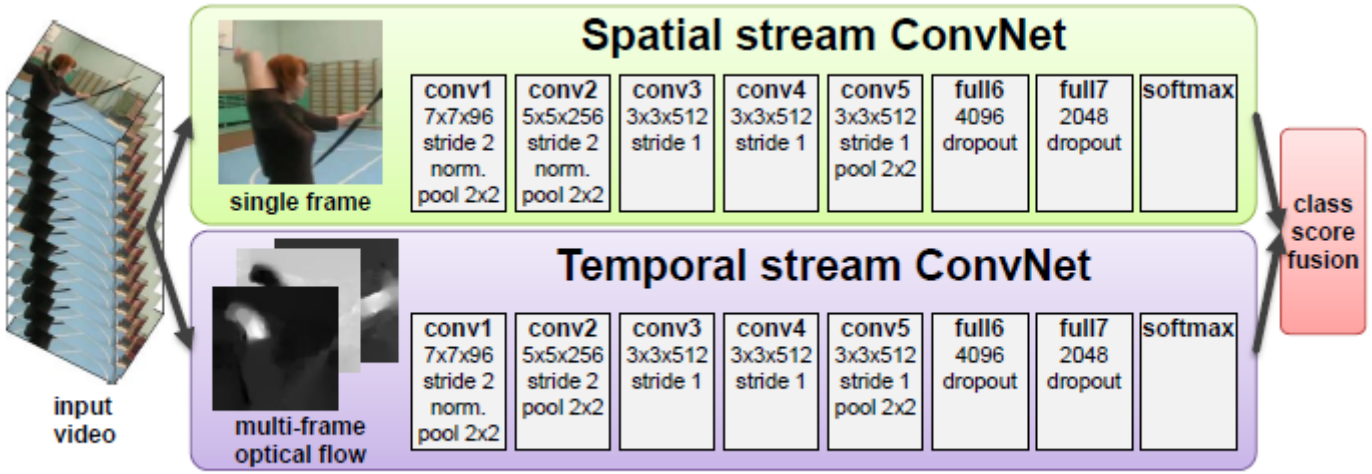


Figure 1: **Two-stream architecture for video classification.**

[Simonyan and Zisserman, NIPS 2014]

- Excellent performance on UCF 101

Table 1: **Individual ConvNets accuracy on UCF-101 (split 1).**

(a) **Spatial ConvNet.**

Training setting	Dropout ratio	
	0.5	0.9
From scratch	42.5%	52.3%
Pre-trained + fine-tuning	70.8%	**72.8%**
Pre-trained + last layer	**72.7%**	59.9%

(b) **Temporal ConvNet.**

Input configuration	Mean subtraction	
	off	on
Single-frame optical flow ($L = 1$)	-	73.9%
Optical flow stacking (1) ($L = 5$)	-	80.4%
Optical flow stacking (1) ($L = 10$)	79.9%	**81.0%**
Trajectory stacking (2)($L = 10$)	79.6%	80.2%
Optical flow stacking (1)($L = 10$), bi-dir.	-	**81.2%**

Table 3: **Two-stream ConvNet accuracy on UCF-101 (split 1).**

Spatial ConvNet	Temporal ConvNet	Fusion Method	Accuracy
Pre-trained + last layer	bi-directional	averaging	85.6%
Pre-trained + last layer	uni-directional	averaging	85.9%
Pre-trained + last layer	uni-directional, multi-task	averaging	86.2%
Pre-trained + last layer	uni-directional, multi-task	SVM	**87.0%**

Questions

- How to use temporal information effectively?

- Factors to consider
  - Small amount of training samples?
  - Video-level label or frame-level label?
  - What is the limitation of single frame CNN?