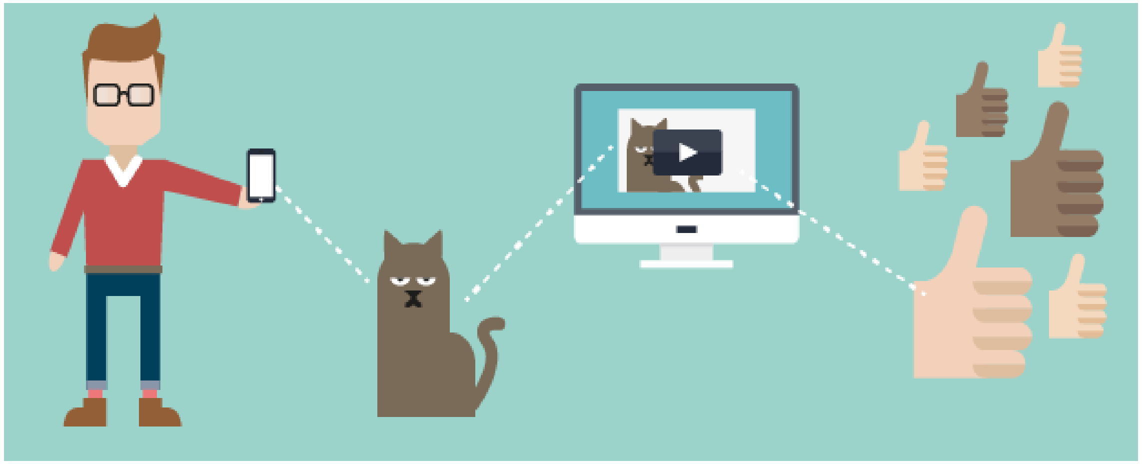
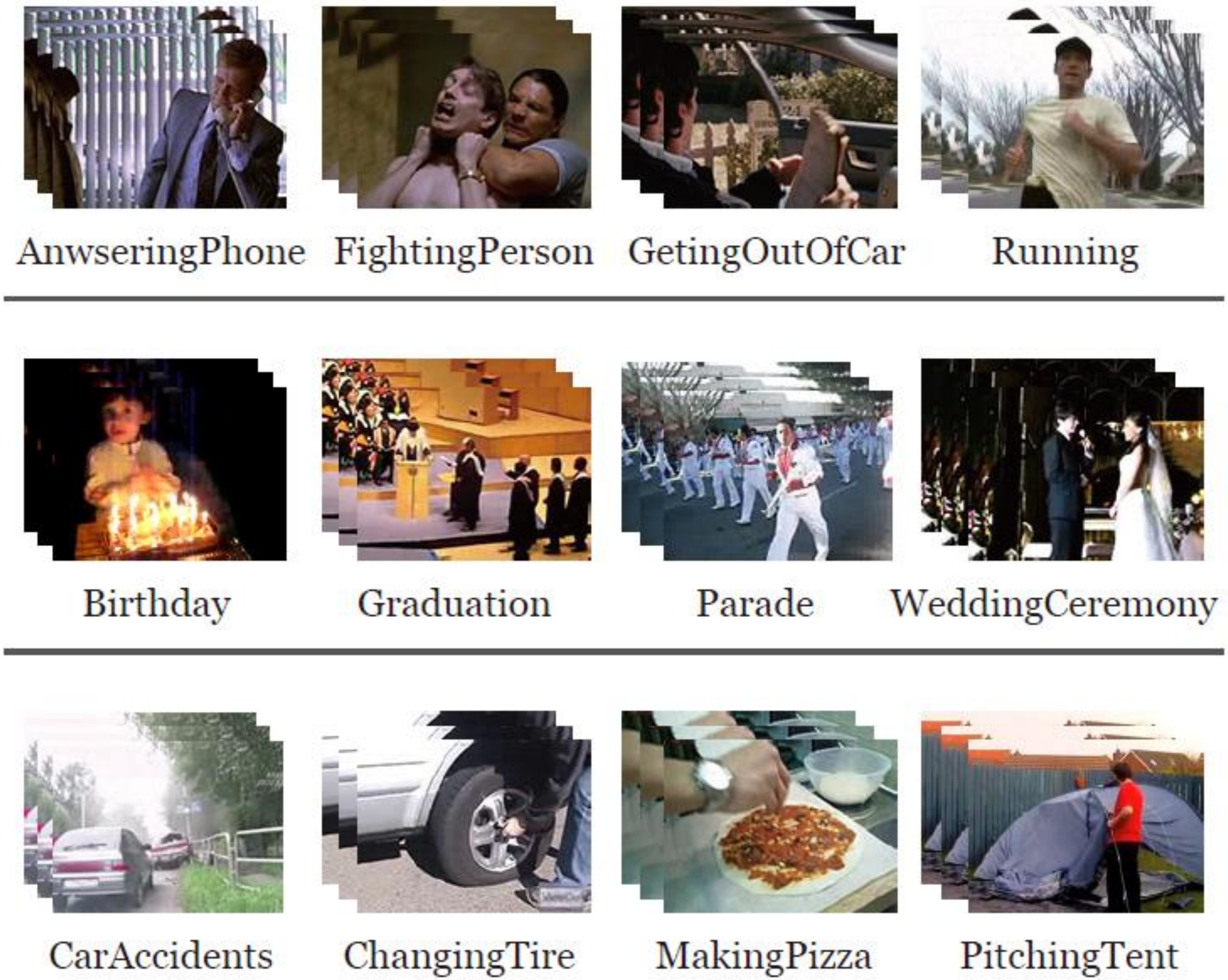
Action and Event Recognition

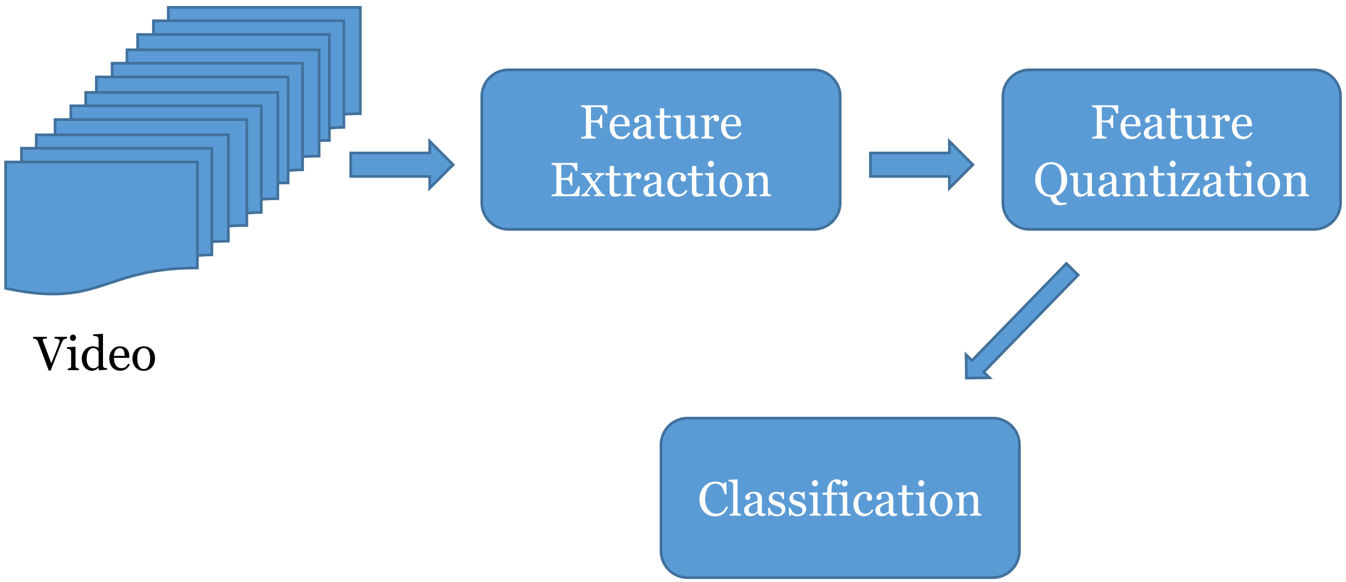
Zuxuan Wu

Massive Videos




AnwseringPhone
FightingPerson
GetingOutOfCar
Running
Birthday
Graduation
Parade
WeddingCeremony
CarAccidents
ChangingTire
MakingPizza
PitchingTent

Visual Recognition Pipeline



Features

Videos are naturally *multimodal*

- Static Appearance Features
- Motion Features
- Acoustic Features
- High-level Features

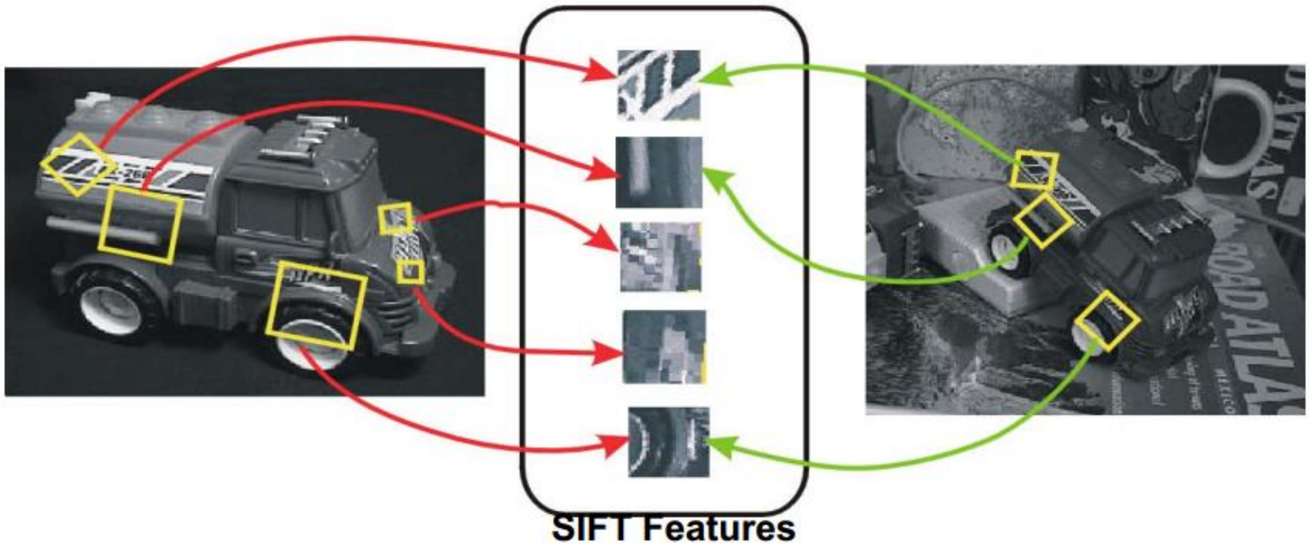
Static Appearance Features

- Captures Static Appearance Information In *Each Frame*
  - shape
  - edge
  - color
  - even high-level appearance information

- Frame-level Features are averaged to generate Video-level representation

Static Appearance Features: SIFT

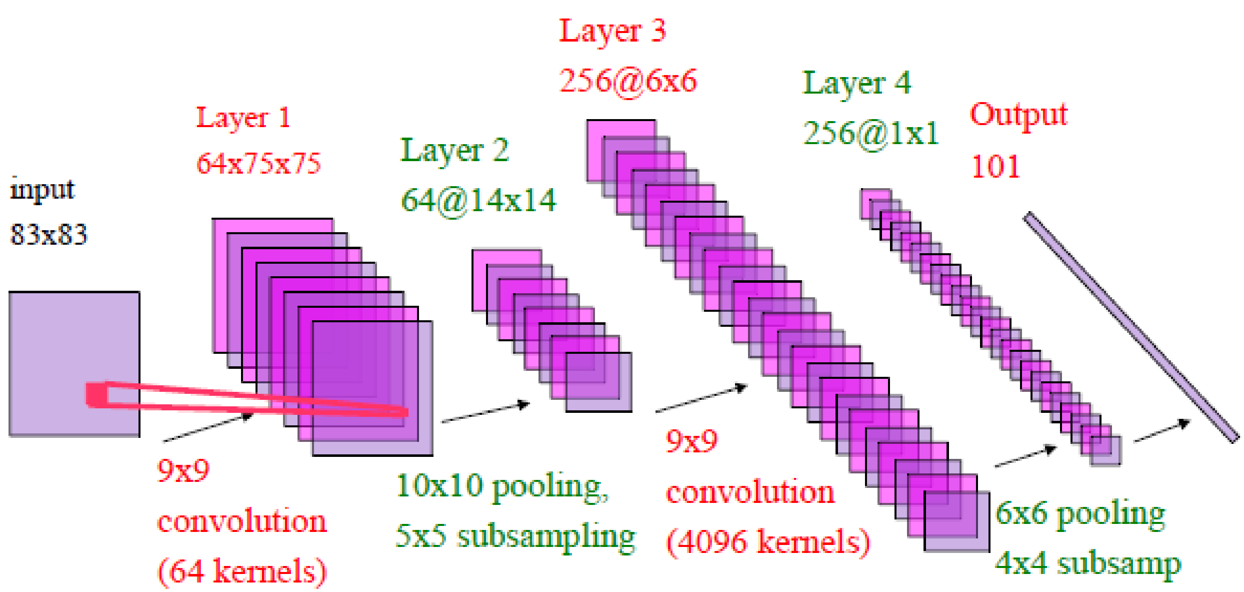
IDEA : Image content is transformed into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters



Static Appearance Features: CNN features

Convolutional Neural Network Recap

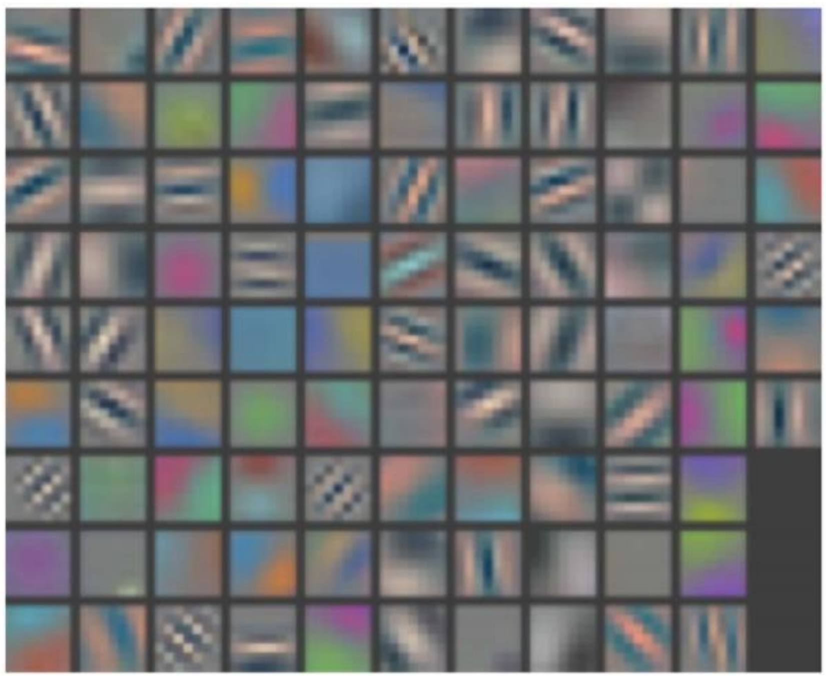
CNN is a model of Deep Learning; Unlike hand-crafted features, CNN learn features from raw images



Static Appearance Features: CNN features

Convolutional Neural Network Recap

CNN is a model of Deep Learning; Unlike hand-crafted features, CNN learn features from raw images



Static Appearance Features: CNN features

Convolutional Neural Network Recap

ImageNet Challenge 2012

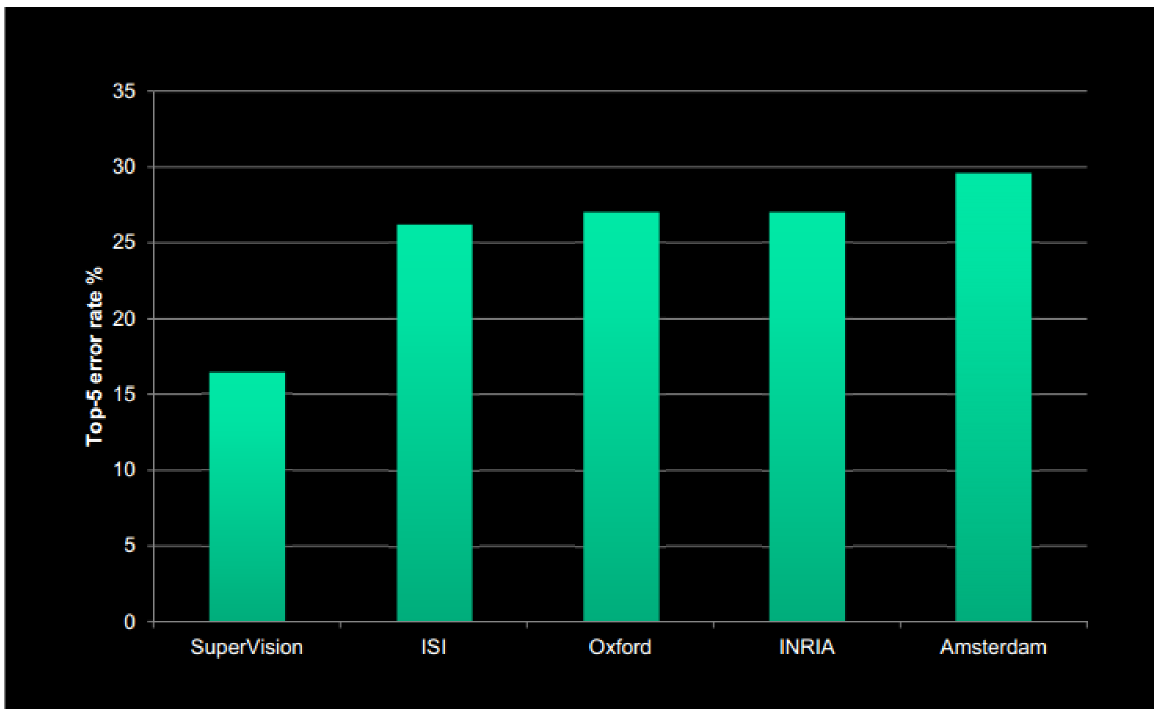
[Deng et al. CVPR 2009]

- ~14 million labeled images, 20k classes
- Images gathered from Internet
- Human labels via Amazon Turk
- Challenge: 1.2 million training images, 1000 classes

Static Appearance Features: CNN features

Convolutional Neural Network outperforms significantly

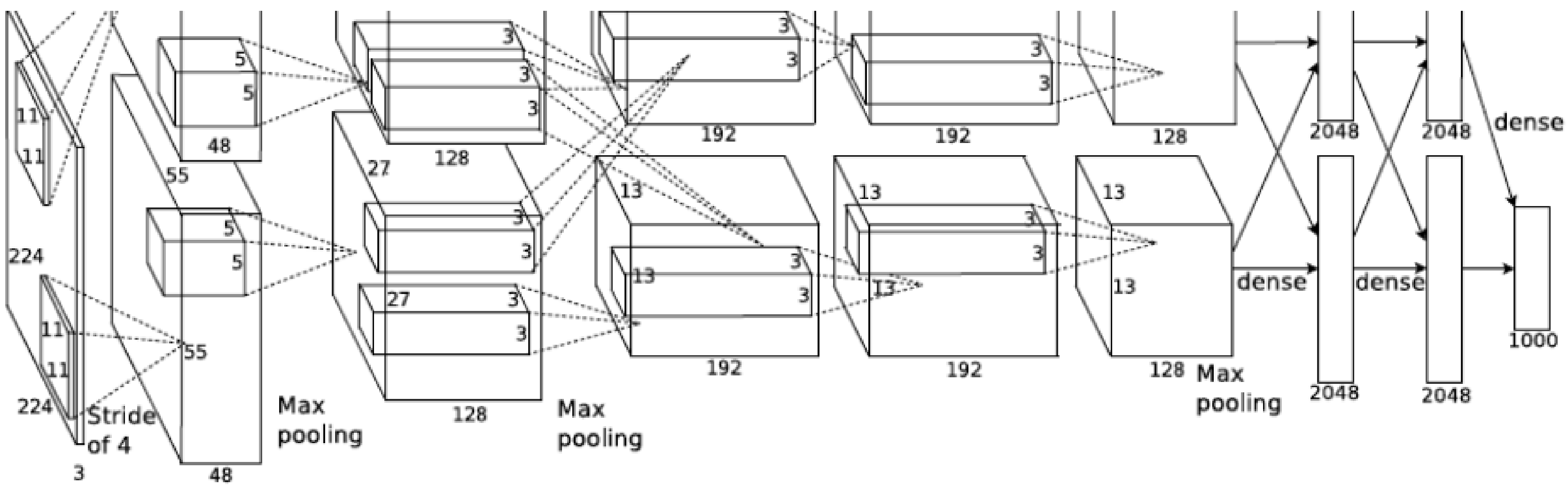
- Krizhevsky et al. -- 16.4% error (top-5)
- Next best (non-convnet) – 26.2% error



Static Appearance Features: CNN features

Convolutional Neural Network outperforms significantly

- Krizhevsky et al. -- 16.4% error (top-5)
- Next best (non-convnet) – 26.2% error



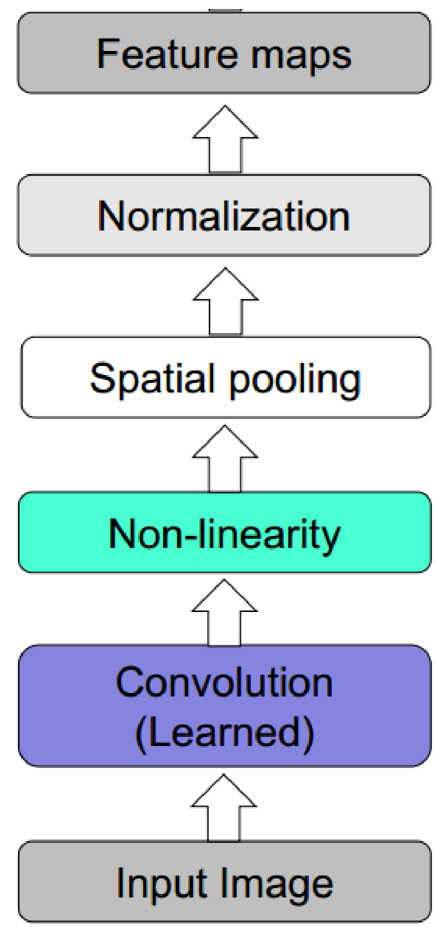
Pre-trained On ImageNet
The Last 3 layers can be viewed as *features*

Static Appearance Features: CNN features

Convolutional Neural Network feature extraction

• Feed-forward feature extraction:

1. Convolve input with learned filters
2. Non-linearity
3. Spatial pooling
4. Normalization



Static Appearance Features: CNN features

Convolutional Neural Network achieved a great success

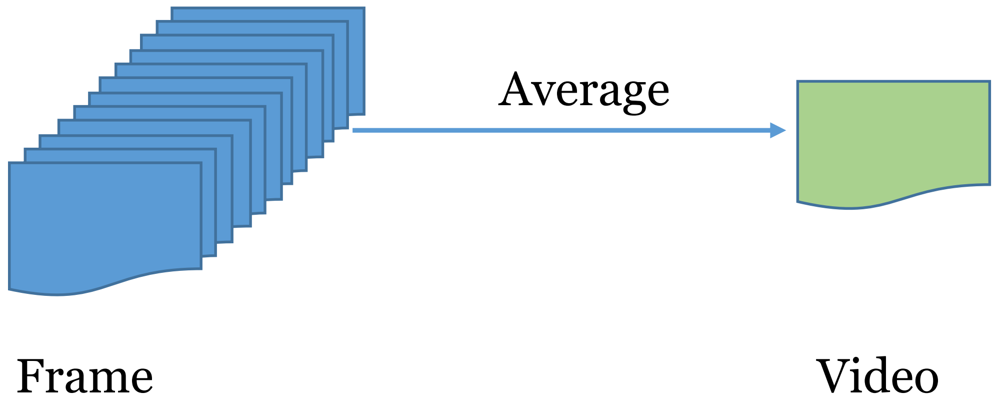
Astounding Baseline with CNN features

Comparing Best State of the Art Methods with Deep Represenations

	VOC07c	VOC12c	VOC12a	MIT67	SUN397	VOC07d	VOC10d	VOC11s	200Birds	102Flowers	H3Datt	UIUCatt	LFW	YTF	Paris6k	Oxford5k	Sculp6k	Holidays	UKB
best non-CNN results	70.5	82.2	69.6	64.0	47.2	34.3	40.4	47.6	56.8	80.7	69.9	~90.0	96.3	89.4	78.2	81.7	45.4	82.2	89.3
off-the-shelf ImageNet Model	80.1[13] 80.1[10] 77.2[1]	82.7[10] 79.0[6]	-	69.0[1]	40.9[4]	46.2[2] 46.1[11] 44.9[13]	44.1[11]	-	61.8[1] 58.8[4]	86.8[1]	73.0[1]	91.5[1]	-	-	79.5[1]	68.0[1]	42.3[1]	84.3[1]	91.1[1]
off-the-shelf ImageNet Model + rep learning	-	-	-	68.9[3]	52.0[3]	-	-	-	65.0[4]	-	-	-	-	-	-	-	-	80.2[3]	-
fine-tuned ImageNet Model	82.42[10] 77.7[5]	83.2[10] 82.8[5]	70.2[5]	-	-	60.9[13] 58.5[2]	53.7[2]	47.9[2]	75.7[12]	-	-	-	-	-	-	-	-	-	-
Other Deep Learning Models	-	-	-	-	-	-	-	-	-	-	79.0[7]	-	97.35[8]	91.4[8]	-	-	-	-	-

Static Appearance Features

Frame-level Features are averaged to generate Video-level representation



Motion Features

Plays *the most important* role for video classification

We introduce

- Dense Trajectories today: state-of-the-art features nowadays (CVPR 11, citation 455 now), still beats DL till now.

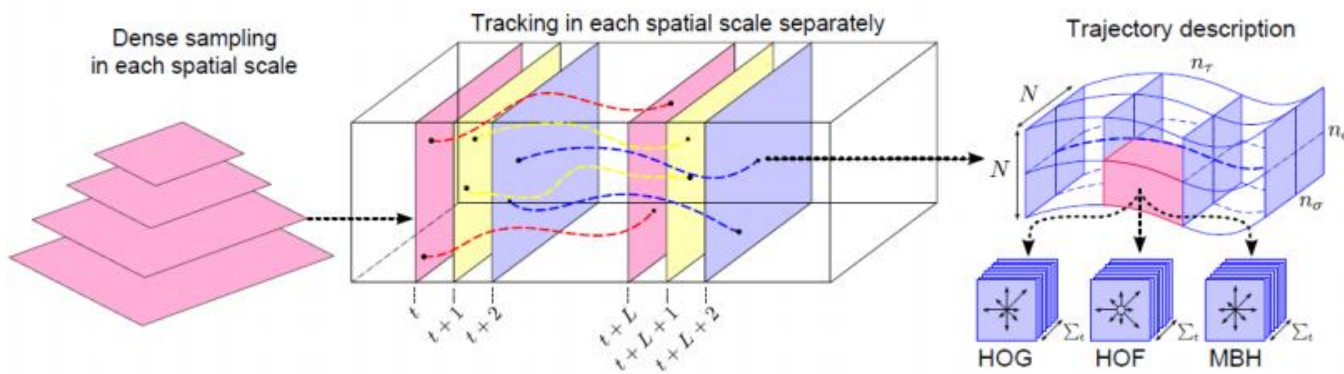
Motion Features

Plays *the most important* role for video classification

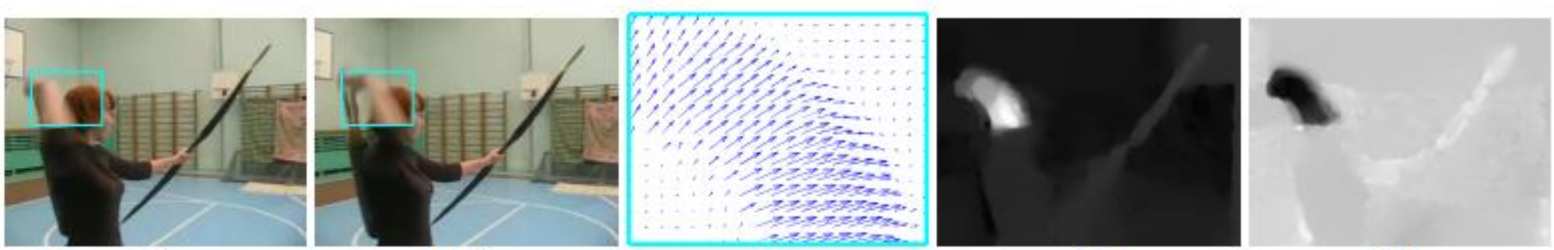
We introduce

- Dense Trajectories today: state-of-the-art features nowadays (CVPR 11, citation 455 now), still beats DL till now.

Motion Features: Dense Trajectories



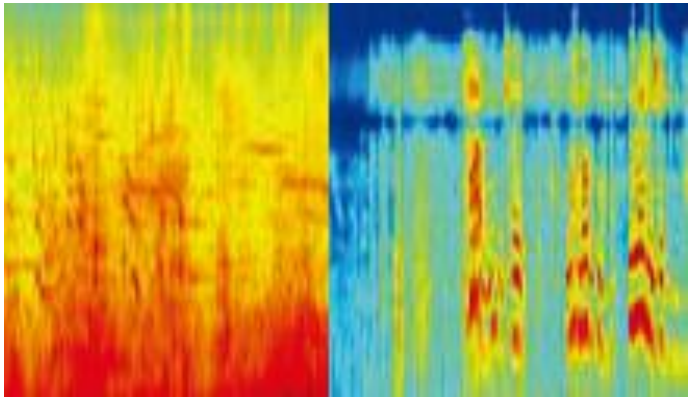
Wang et al, CVPR 2011, IJCV 2012


(a) (b) (c) (d) (e)

Acoustic Features: MFCC

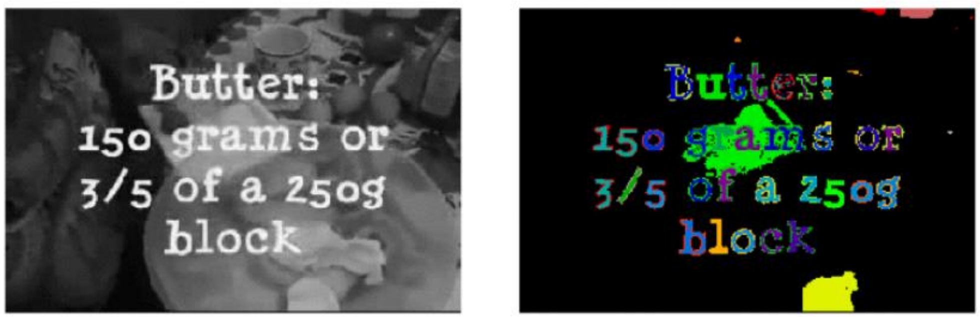
- MFCC (Mel-frequency Cepstral Coefficients)

- Spectrogram SIFT

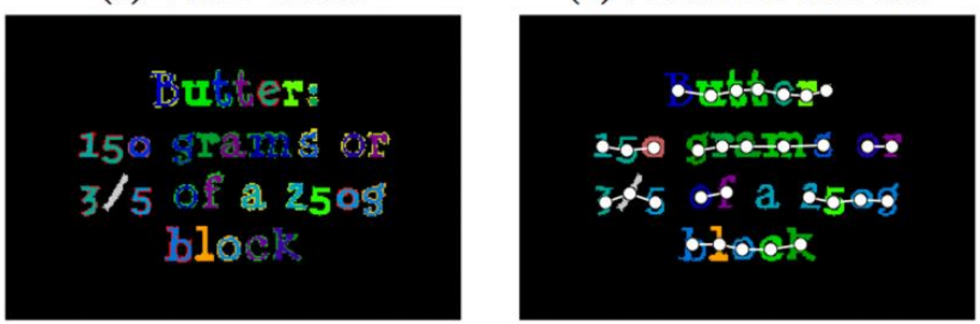


High-level Features

- ASR (Automatic Speech Recognition)
- OCR

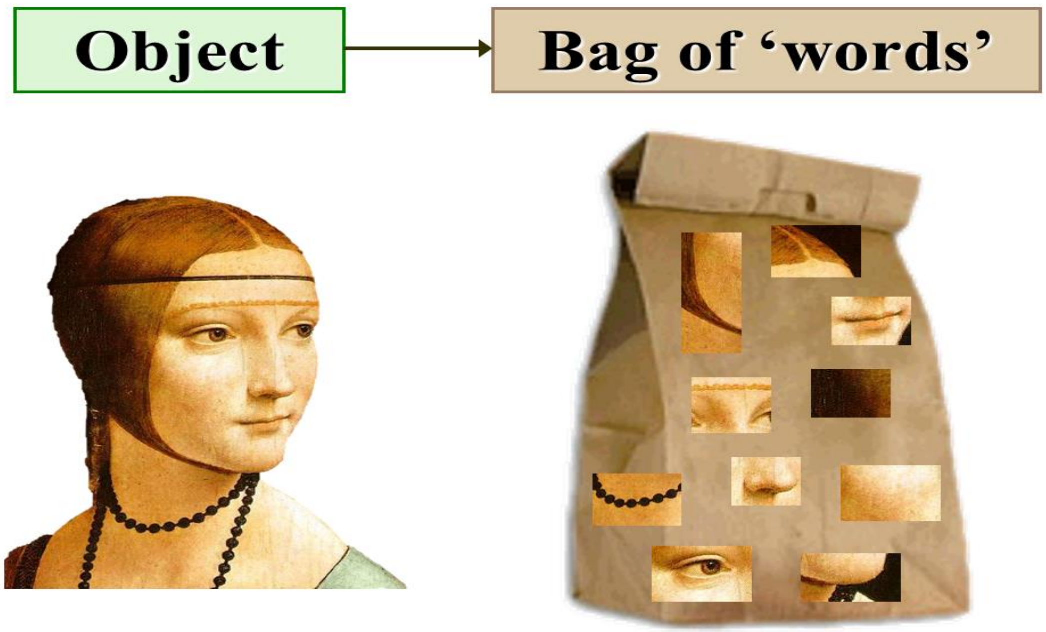


(a) Video frame (b) Extracted MSERs



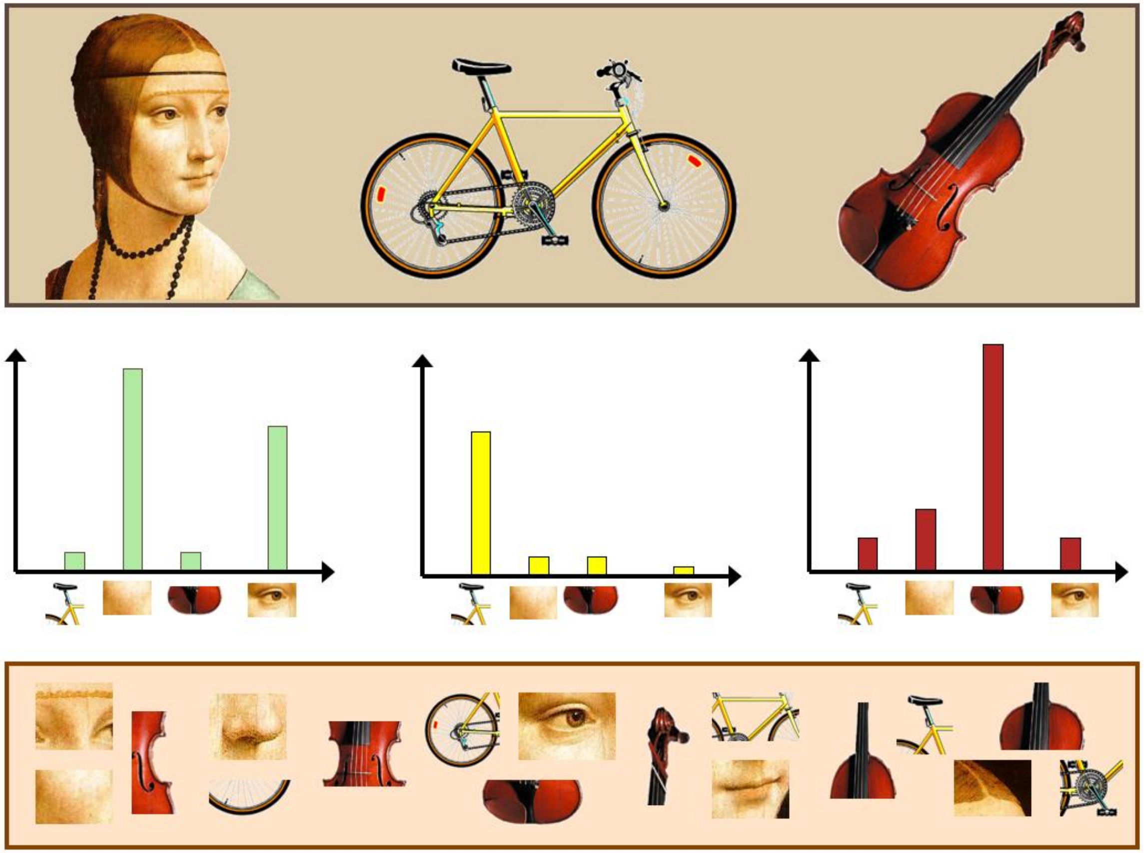
Feature Encoding Strategies

- Bag-of-words Representation

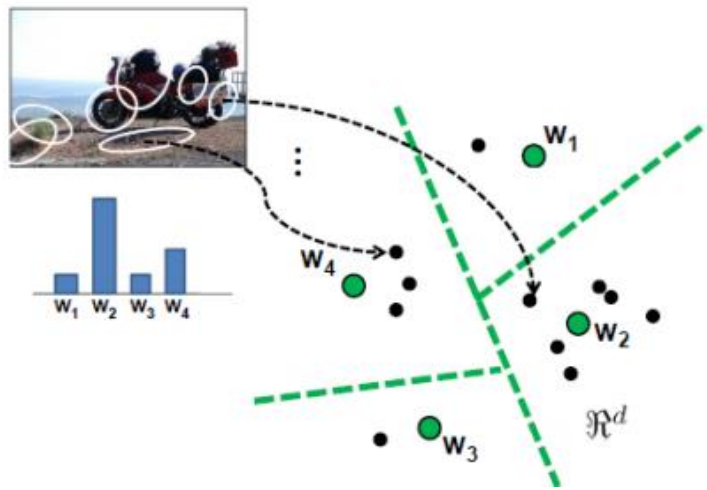


Feature Encoding Strategies

- Bag-of-words Representation



Feature Encoding Strategies

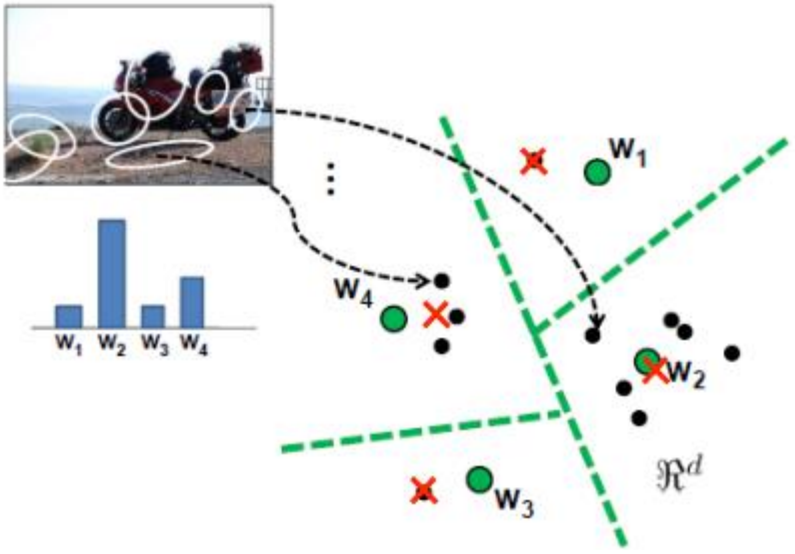


- BoW is only about counting the number of local descriptors assigned to each region
- Why not including other statistics?

Feature Encoding Strategies

- Fisher Vector Representation
- Including other statistics

➢ mean of local descriptors

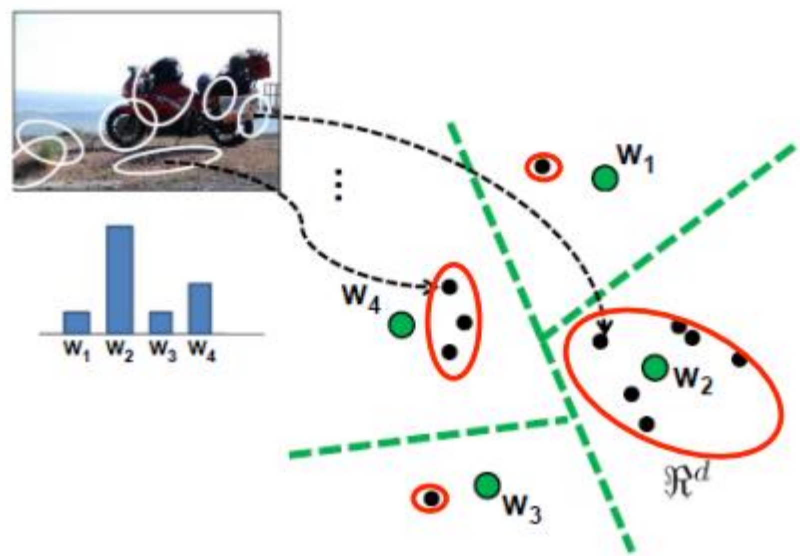


Feature Encoding Strategies

- Fisher Vector Representation
- Including other statistics

➢ mean of local descriptors

➢ variance of local descriptors



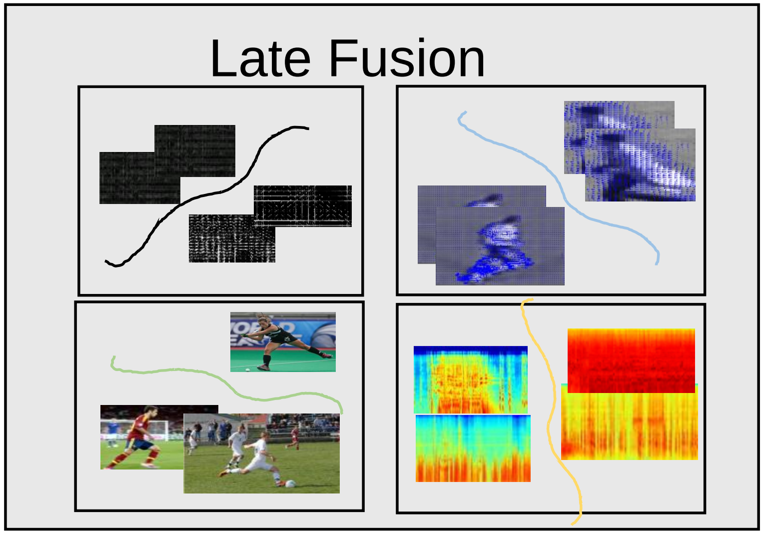
Classification

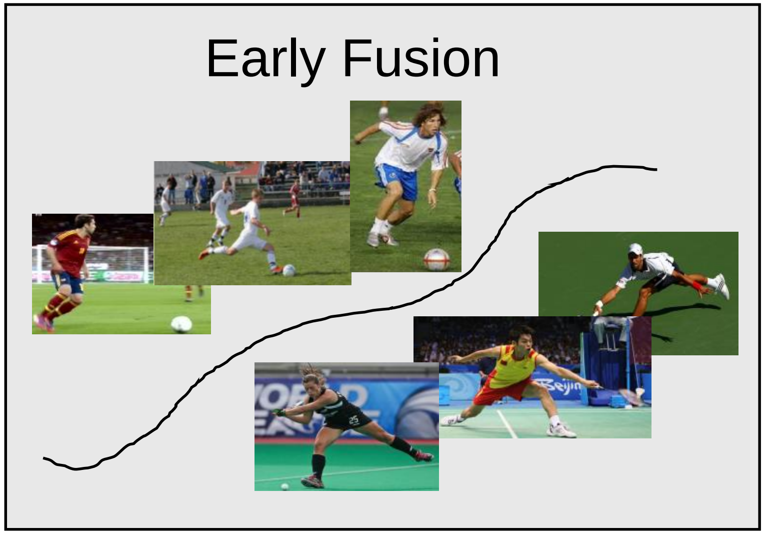
- For histogram-based features, non-linear \chi^2-kernel SVMs

- For Fisher Vector based features, linear SVMs is good enough.

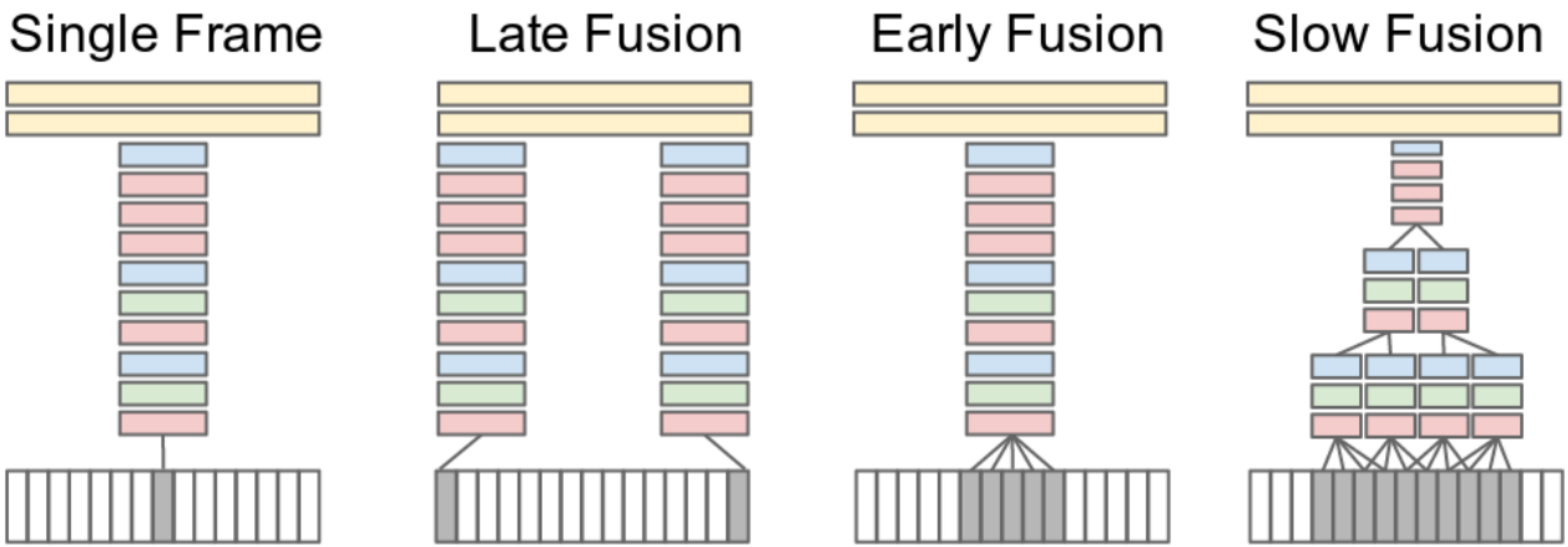
Fusion

- Late Fusion (Classifier)
- Early Fusion (Feature)





Deep Learning Approach



Deep Learning Approach

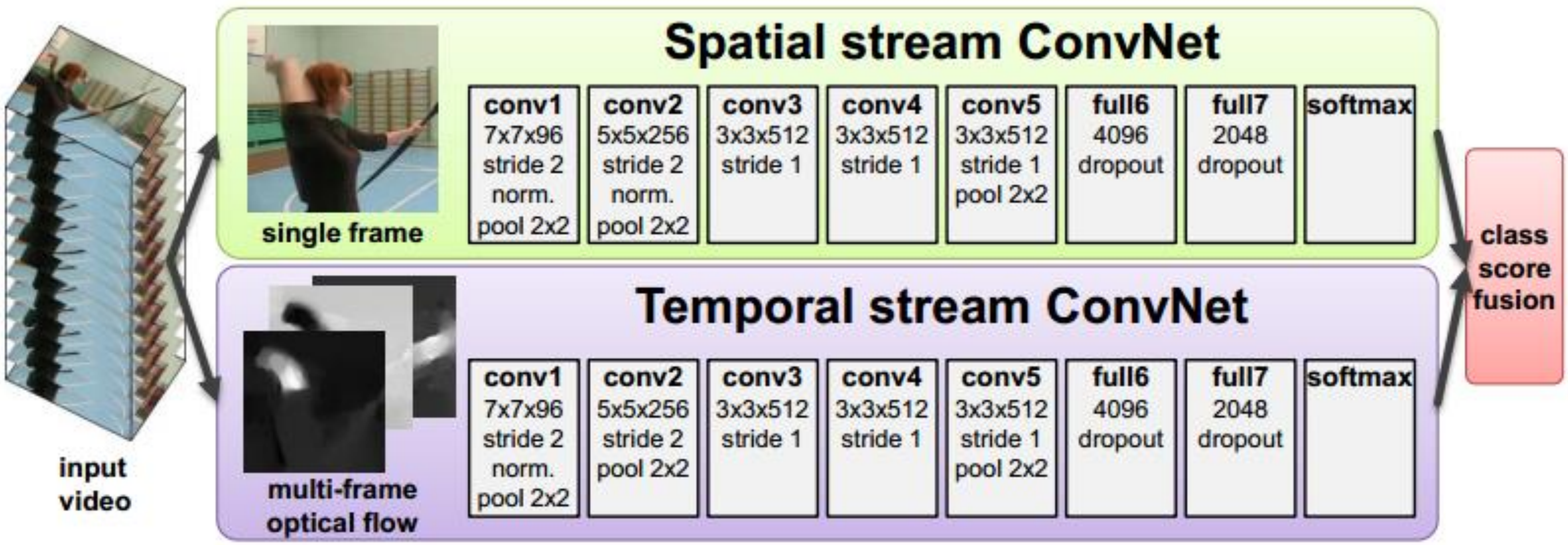


Figure 1: **Two-stream architecture for video classification.**

Challenge

DATA extremely LARGE!

- MED 14 : TEST set, 200,000 video clips, about 4T

Challenge

DATA extremely LARGE!

- MED 14 : TEST set, 200,000 video clips, about 4T

Challenge

DATA extremely LARGE!

- MED 14 : TEST set, 200,000 video clips, about 4T

Data Management
Data transfer
Classifier Training
Slow Feature Extraction

Scp?? 30m/s?
1T needs 9h

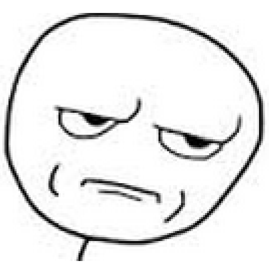
Challenge

DATA extremely LARGE!

- MED 14 : TEST set, 200,000 video clips, about 4T

Data Management
Data transfer
Classifier Training
Slow Feature Extraction

Scp?? 30m/s?
1T needs 9h

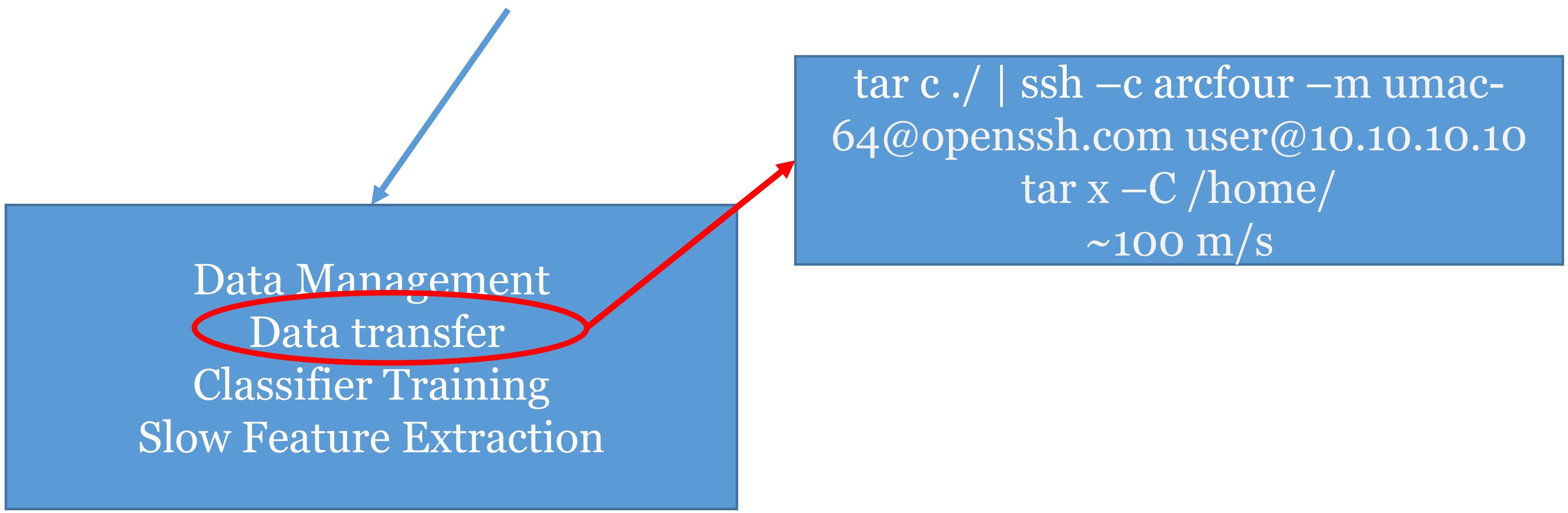


Too slow!!!!

Challenge

DATA extremely LARGE!

- MED 14 : TEST set, 200,000 video clips, about 4T



Challenge

DATA extremely LARGE!

- MED 14 : TEST set, 200,000 video clips, about 4T

Data Management
Data transfer
Classifier Training
Slow Feature Extraction

tar c ./ | ssh –c arcfour –m umac-
64@openssh.com user@10.10.10.10
tar x –C /home/
~100 m/s

Challenge

DATA extremely LARGE!

- MED 14 : TEST set, 200,000 video clips, about 4T

Data Management
Data transfer
Classifier Training
Slow Feature Extraction

LibLinear , LibSVM, WILL all collapse

Challenge

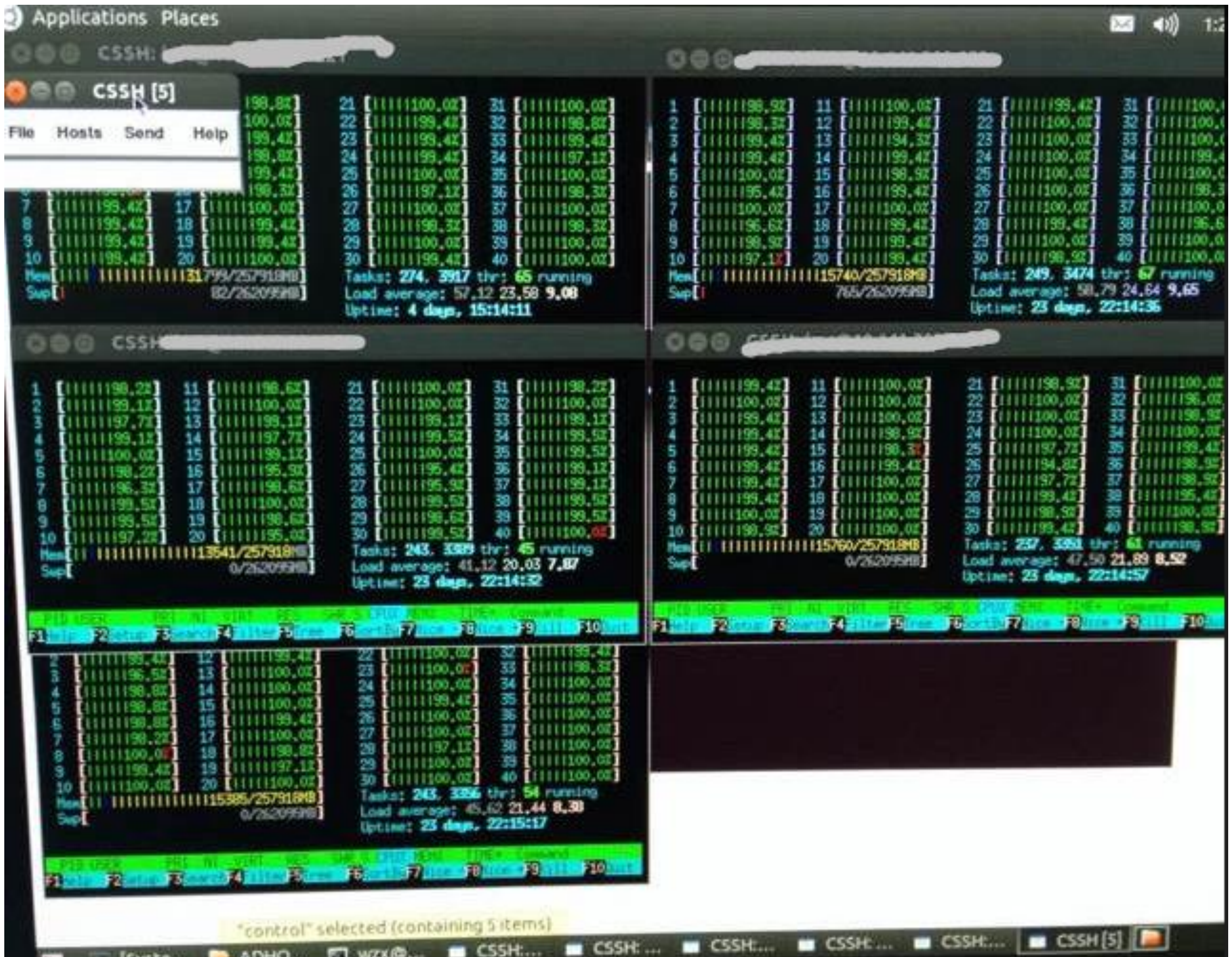
DATA extremely LARGE!

- MED 14 : TEST set, 200,000 video clips, about 4T

Data Management
Data transfer
Classifier Training
Slow Feature Extraction

LibLinear , LibSVM, WILL all collapse

Challenge


Applications Places
CSSH [5]
File Hosts Send Help
"control" selected (containing 5 items)