CS231N Section

Video Understanding

5/29/2020

Outline

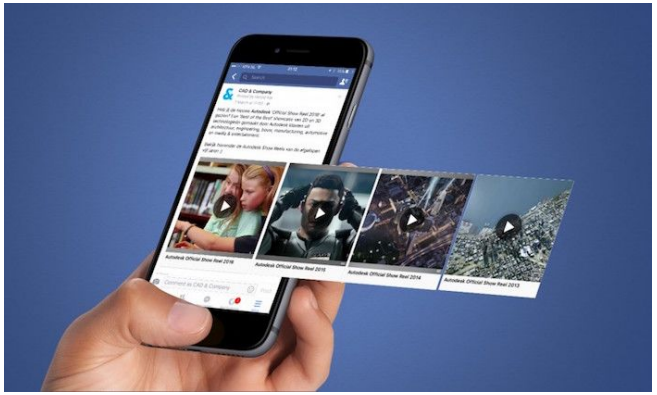
- Background / Motivation / History
- Video Datasets
- Models
  - Pre-deep learning
  - CNN + RNN
  - 3D convolution
  - Two-stream

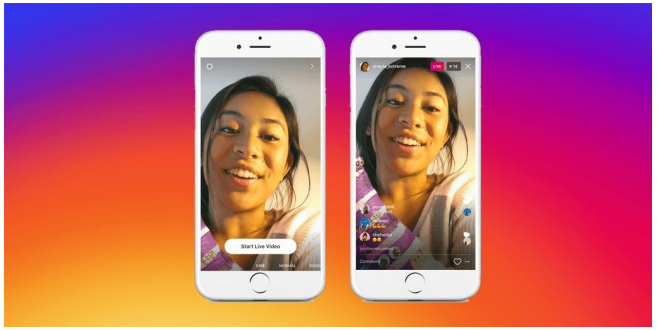
What we've seen in class so far...

- Image Classification
- CNNs, GANs, RNNs, LSTMs, GRU
- ...

What's missing → videos!

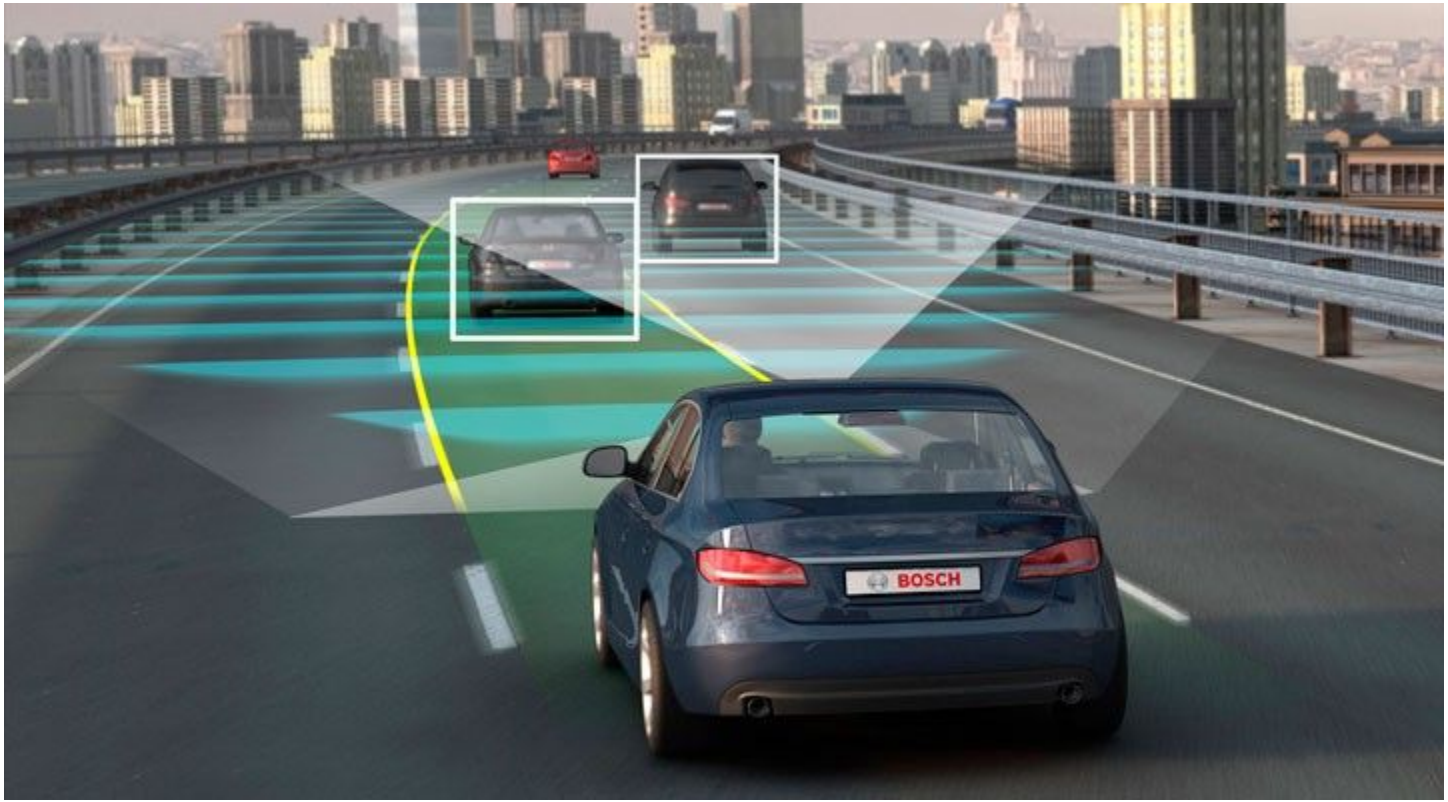
Videos are Everywhere





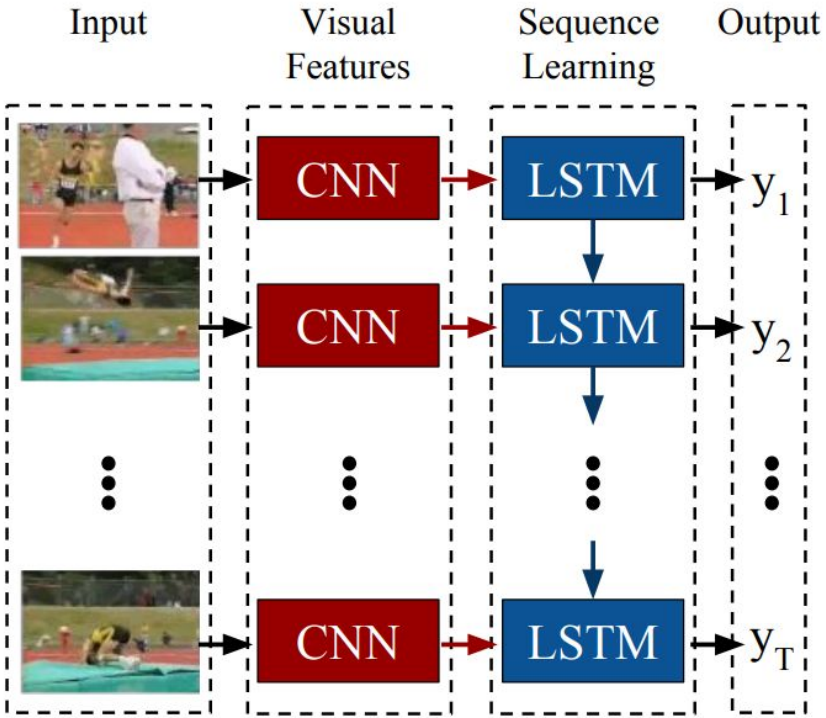
Robotics / Manipulation

Self-Driving Cars



Collective Activity Understanding

Video Captioning



Donahue et al.,Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR, 2015

...and more!

- Video editing
- VR (e.g. vision as inverse graphics)
- Video QA
- ...

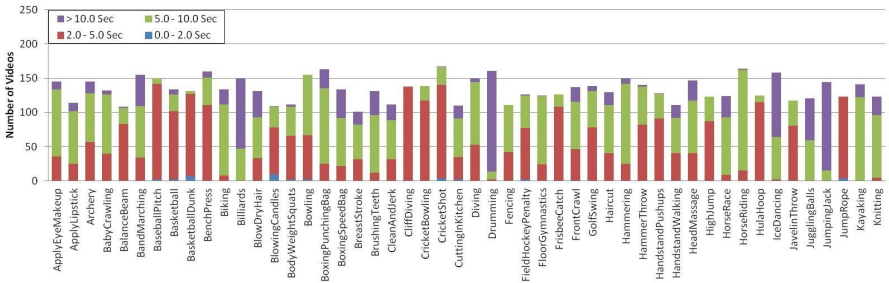
Datasets

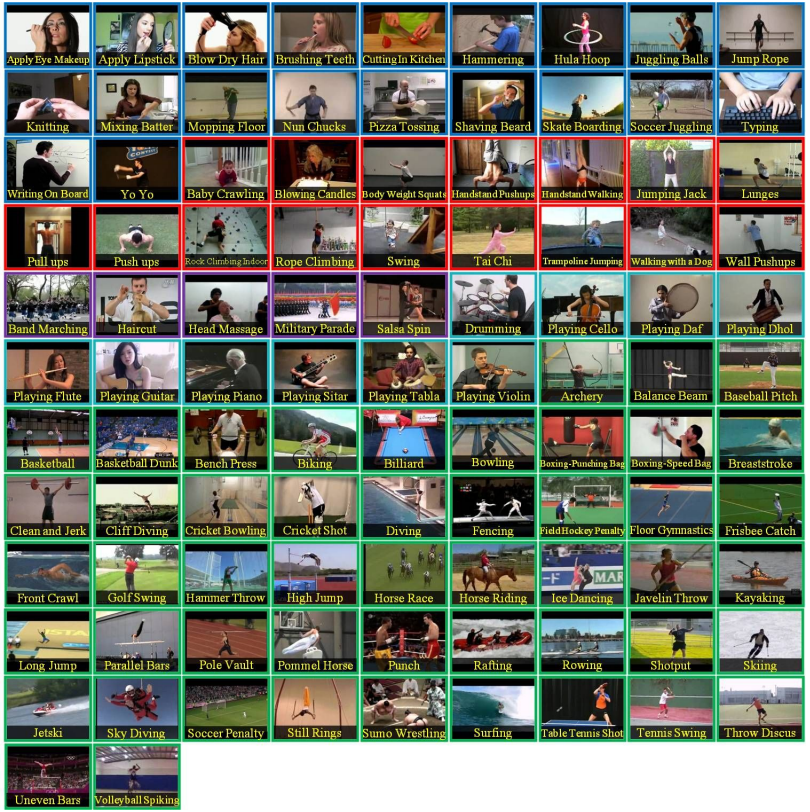
- Video Classification
- Atomic Actions
- Video Retrieval

Video Classification

UCF101

- YouTube videos
- 13320 videos, 101 action categories
- Large variations in camera motion, object appearance and pose, viewpoint, background, illumination, etc.

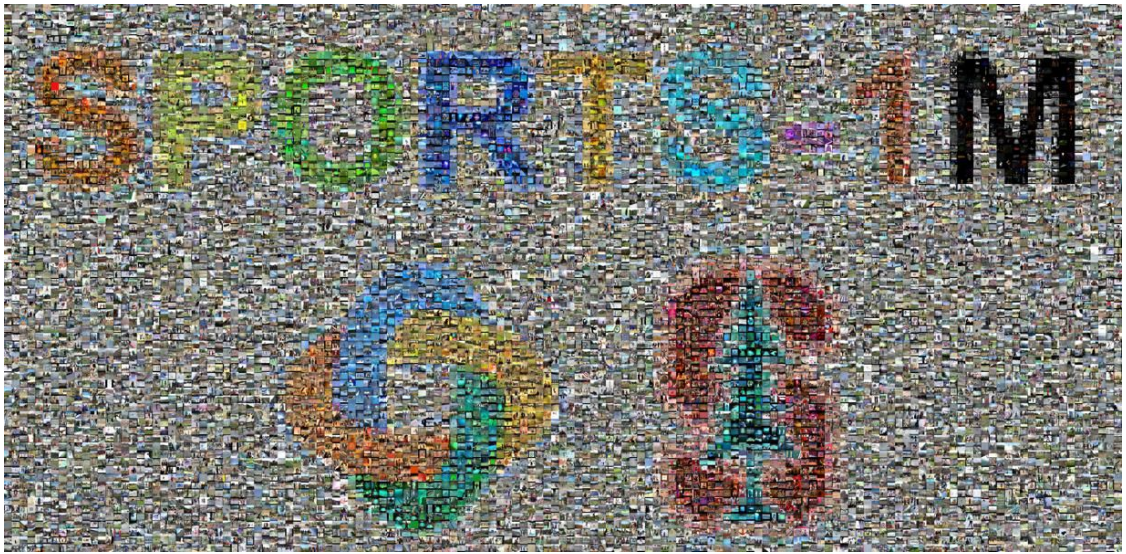




Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild., CRCV-TR-12-01, 2012

Sports-1M

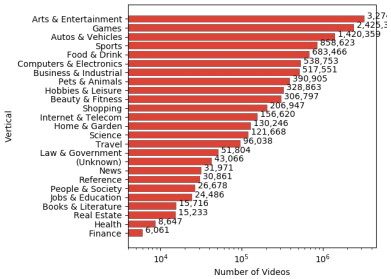
- YouTube videos
- 1,133,157 videos, 487 sports labels

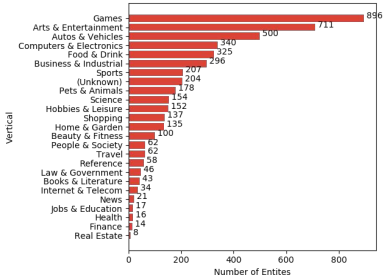


Karpathy et al., Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014

YouTube 8M

- Data
  - Machine-generated annotations from 3,862 classes
  - Audio-visual features





Atomic Actions

Charades

- Hollywood in Homes: crowdsourced “boring” videos of daily activities
- 9848 videos
- RGB + optical flow features
- Action classification, sentence prediction
- Pros and cons
  - Pros: Objects; video-level and frame-level classification
  - Cons: No human localization

Gunnar et al., Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding, ECCV 2016

Atomic Visual Actions (AVA)

- Data
  - 57.6k 3s segments
  - Pose and object interactions
- Pros and cons
  - Pros: Fine-grained
  - Cons: no annotations about objects

Gunnar et al., Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding, ECCV 2016

Moments in Time (MIT)

- Dataset: 1,000,000 3s videos
  - 339 verbs
  - Not limited to humans
  - Sound-dependent: e.g. clapping in the background
- Advantages:
  - Balanced
- Disadvantages:
  - Single label (classification, not detection)

Monfort et al., Moments in Time Dataset: one million videos for event understanding, PAMI 2019

Video Retrieval: Movie Querying

M-VAD and MPII-MD

- Video clips with descriptions. e.g.:
  - SOMEONE holds a crossbow.
  - He and SOMEONE exit a mansion. Various vehicles sit in the driveway, including an RV and a boat. SOMEONE spots a truck emblazoned with a bald eagle surrounded by stars and stripes.
  - At Vito's the Datsun parks by a dumpster.

AD: Another room, the wife and mother sits at a window with a towel over her hair.

She smokes a cigarette with a latex-gloved hand.

Putting the cigarette out, she uncovers her hair, removes the glove and pops gum in her

She pats her face and hands with a wipe, then sprays herself with perfume.

AD: They rush out onto the street.

A man is trapped under a cart.

Valjean is crouched down beside him.

Javert watches as Valjean places his shoulder under the shaft.

LSMDC (Large Scale Movie Description Challenge)

- Combination of M-VAD and MPII-MD
- https://sites.google.com/site/describingmovies/

Tasks

- Movie description
  - Predict descriptions for 4-5s movie clips
- Movie retrieval
  - Find the correct caption for a video, or retrieve videos corresponding to the given activity
- Movie Fill-in-the-Blank (QA)
  - Given a video clip and a sentence with a blank in it, fill in the blank with the correct word

AD: Abby gets in the basket.

Script: After a moment a frazzled Abby pops up in his place.

Mike leans over and sees how high they are.

Mike looks down to see – they are now fifteen feet above the ground.

Abby clasps her hands around his face and kisses him passionately. For the first time in her life, she stops thinking and grabs Mike and kisses the hell out of him.

Challenges in Videos

- Computationally expensive
  - Size of video >> image datasets
- Lower quality
  - Resolution, motion blur, occlusion
- Requires lots of training data!

What a video framework should have

- Sequence modeling
- Temporal reasoning (receptive field)
- Focus on action recognition
  - Representative task for video understanding

Models

Pre-Deep Learning

Pre-Deep Learning

Features:

- Local features: HOG + HOF (Histogram of Optical Flow)
- Trajectory-based:
  - Motion Boundary Histograms (MBH)
  - (improved) dense trajectories: good performance, but computationally intensive

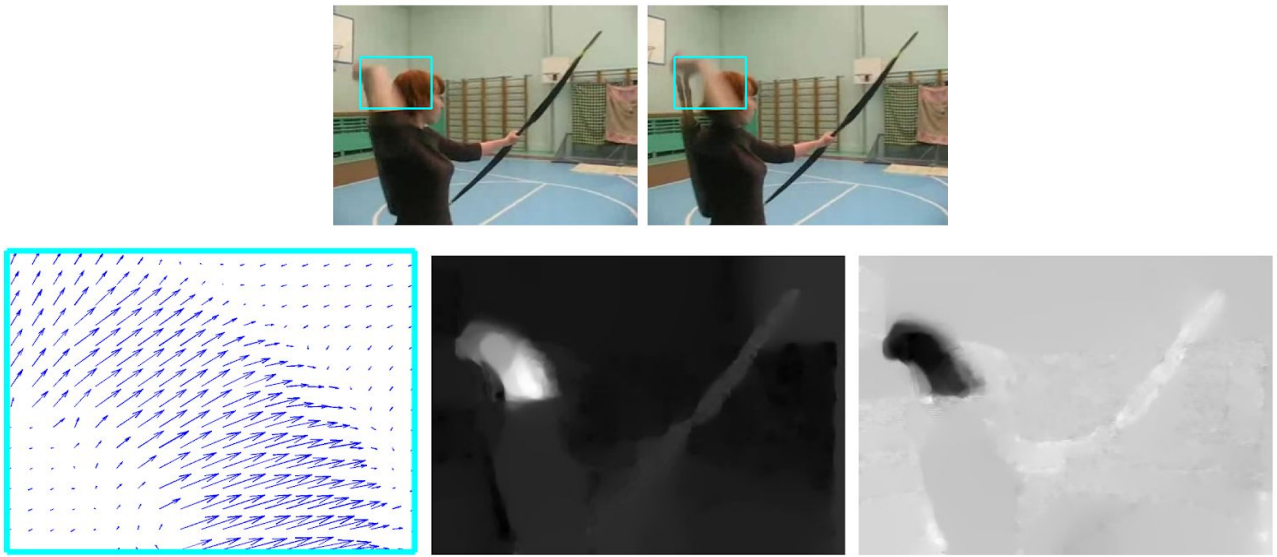
Ways to aggregate features:

- Bag of Visual Words (Ref)
- Fisher vectors (Ref)

Representing Motion

Optical flow: pattern of apparent motion

- Calculation: e.g. TVL1, DeepFlow,

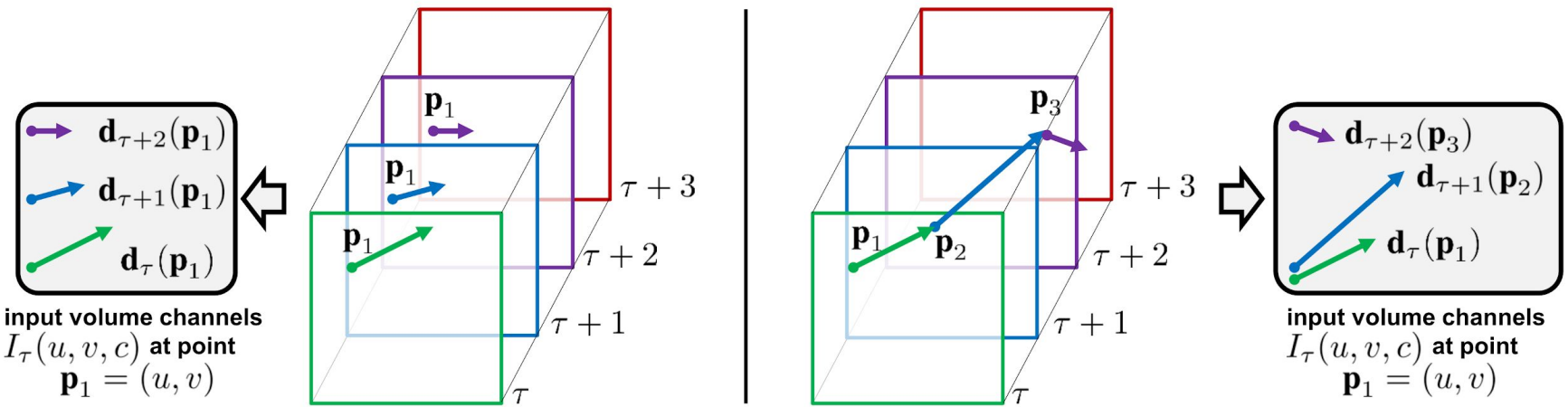


Simonyan and Zisserman, Two-Stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

Representing Motion

1) Optical flow

2) Trajectory stacking



Simonyan and Zisserman, Two-Stream Convolutional Networks for Action Recognition in Videos, NeurIPS 2014

Deep Learning 😊

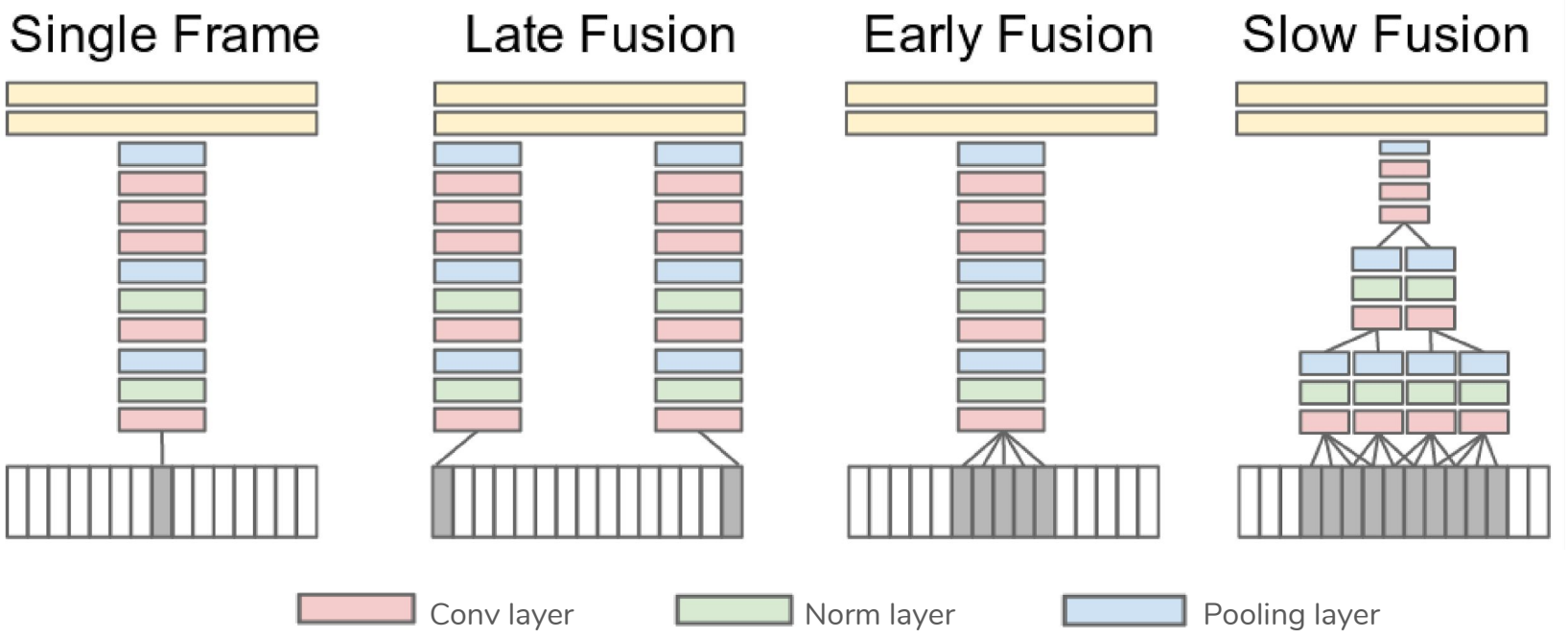
Large-scale Video Classification with Convolutional Neural Networks (pdf)

2 Questions:

- Modeling perspective: what **architecture** to best capture temporal patterns?
- Computational perspective: how to reduce **computation cost** without sacrificing accuracy?

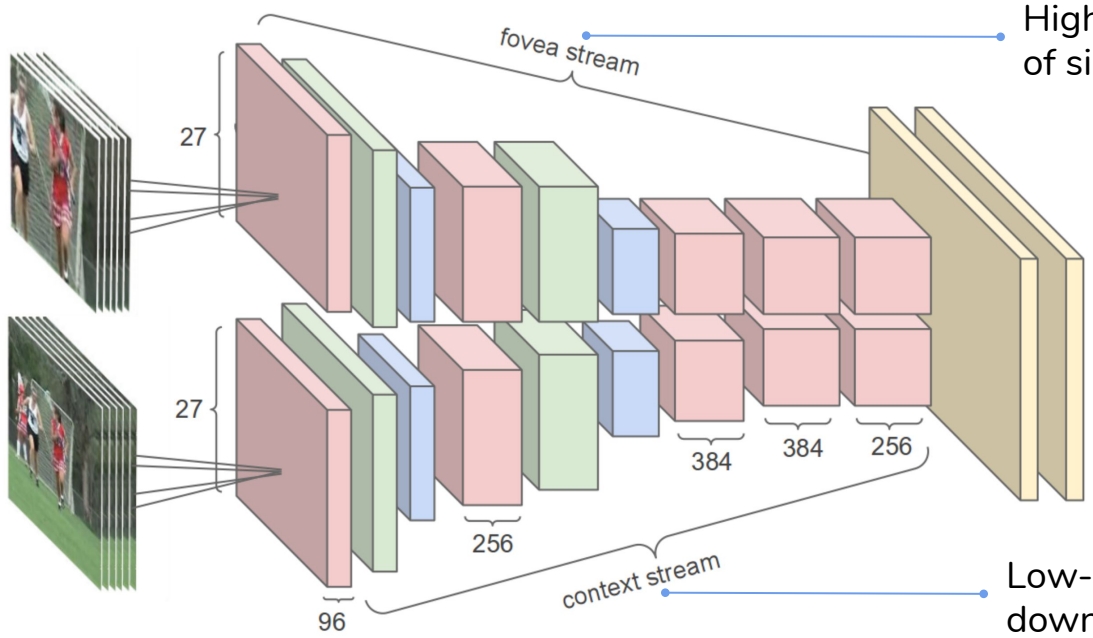
Large-scale Video Classification with Convolutional Neural Networks (pdf)

Architecture: different ways to fuse features from multiple frames



Large-scale Video Classification with Convolutional Neural Networks (pdf)

Computational cost: reduce spatial dimension to reduce model complexity
→ multi-resolution: low-res ***context*** + high-res ***foveate***



Large-scale Video Classification with Convolutional Neural Networks (pdf)

Results on video retrieval (Hit@k: the correct video is ranked among the top k):

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

Next...

- CNN + RNN
- 3D Convolution
- Two-stream networks

CNN + RNN

Videos as Sequences

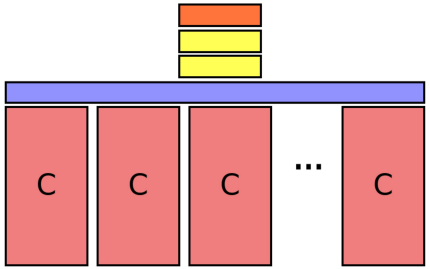
Previous work: multi-frame features are temporally local (e.g. 10 frames)

Hypothesis: a **global description** would be beneficial

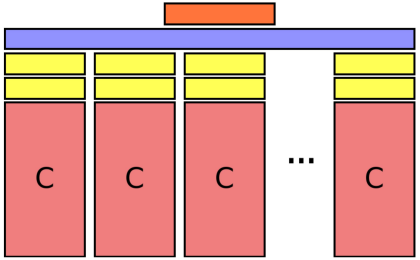
Design choices:

- *Modality*: 1) RGB 2) optical flow 3) RGB + optical flow
- *Features*: 1) hand-crafted 2) extracted using CNN
- *Temporal aggregation*: 1) temporal pooling 2) RNN (e.g. LSTM, GRU)

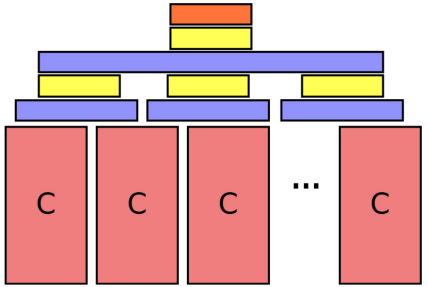
Beyond Short Snippets: Deep Networks for Video Classification (arXiv)



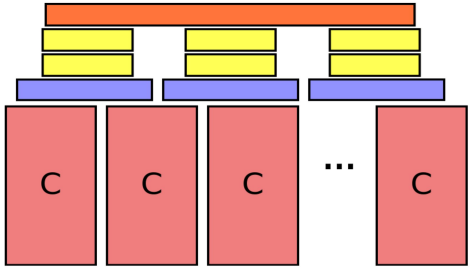
1) Conv Pooling



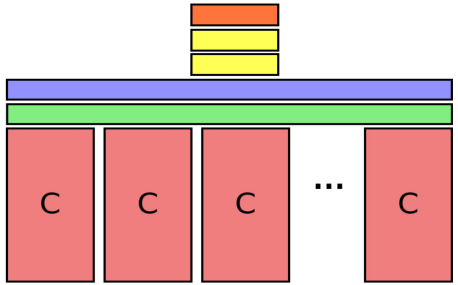
2) Late Pooling



3) Slow Pooling



4) Local Pooling



5) Time-domain convolution

Beyond Short Snippets: Deep Networks for Video Classification (arXiv)

Learning **global description**:

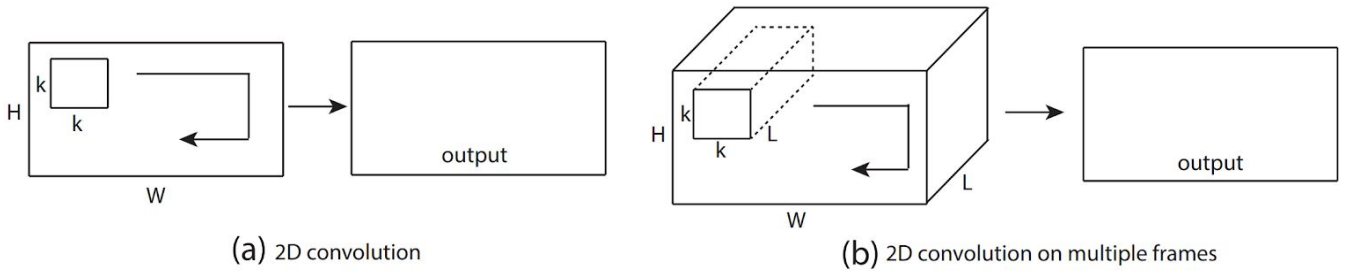
Design choices:

- *Modality*: 1) RGB 2) optical flow 3) **RGB + optical flow**
- *Features*: 1) hand-crafted 2) **extracted using CNN**
- *Temporal aggregation*: 1) temporal pooling 2) **RNN (e.g. LSTM, GRU)**

3D Convolution

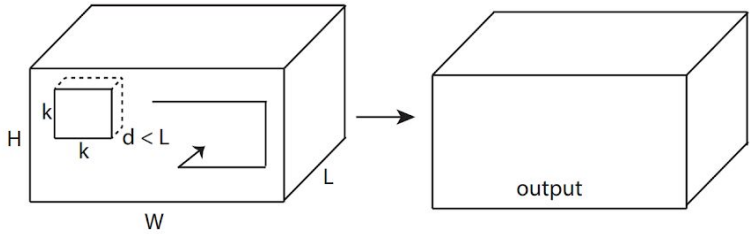
2D vs 3D Convolution

Previous work: 2D convolutions collapse temporal information



(a) 2D convolution (b) 2D convolution on multiple frames

Proposal: 3D convolution → learning features that encode temporal information

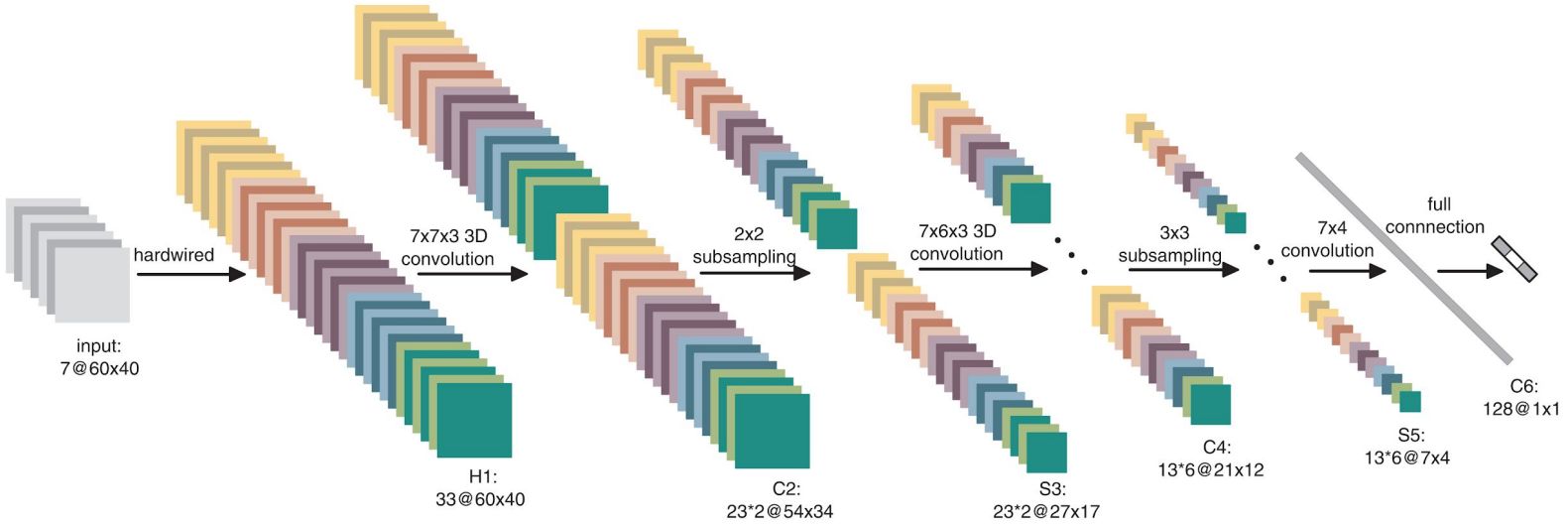


(c) 3D convolution

3D Convolutional Neural Networks for Human Action Recognition (pdf)

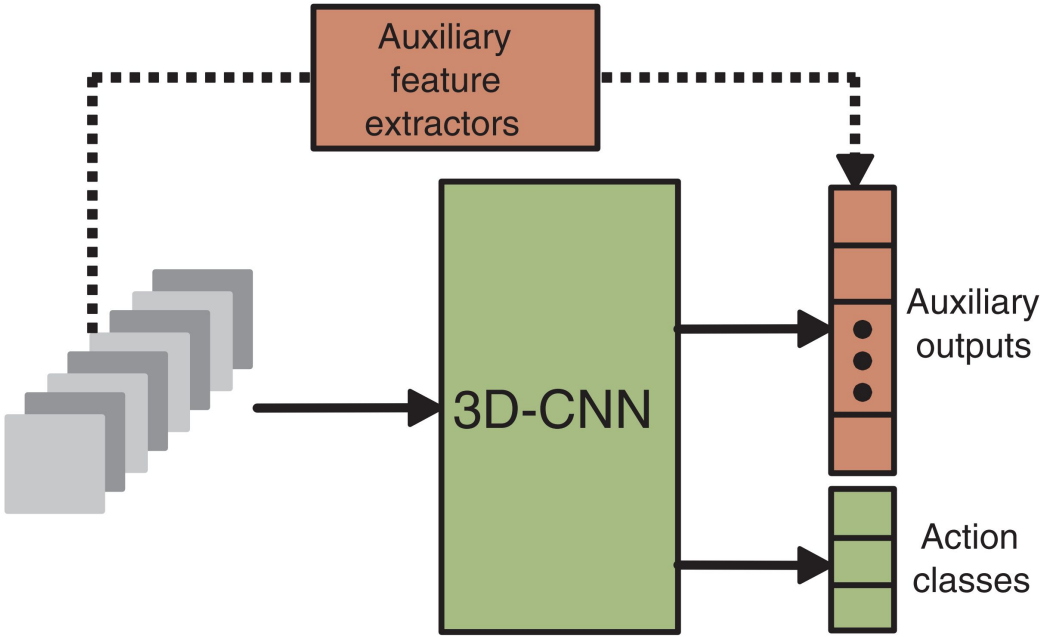
Multiple channels as input:

1) gray, 2) gradient x, 3) gradient y, 4) optical flow x, 5) optical flow y



3D Convolutional Neural Networks for Human Action Recognition (pdf)

Handcrafted long-term features: information beyond the 7 frames + regularization



Learning Spatiotemporal Features with 3D Convolutional Networks (pdf)

Improve over the previous 3D conv model

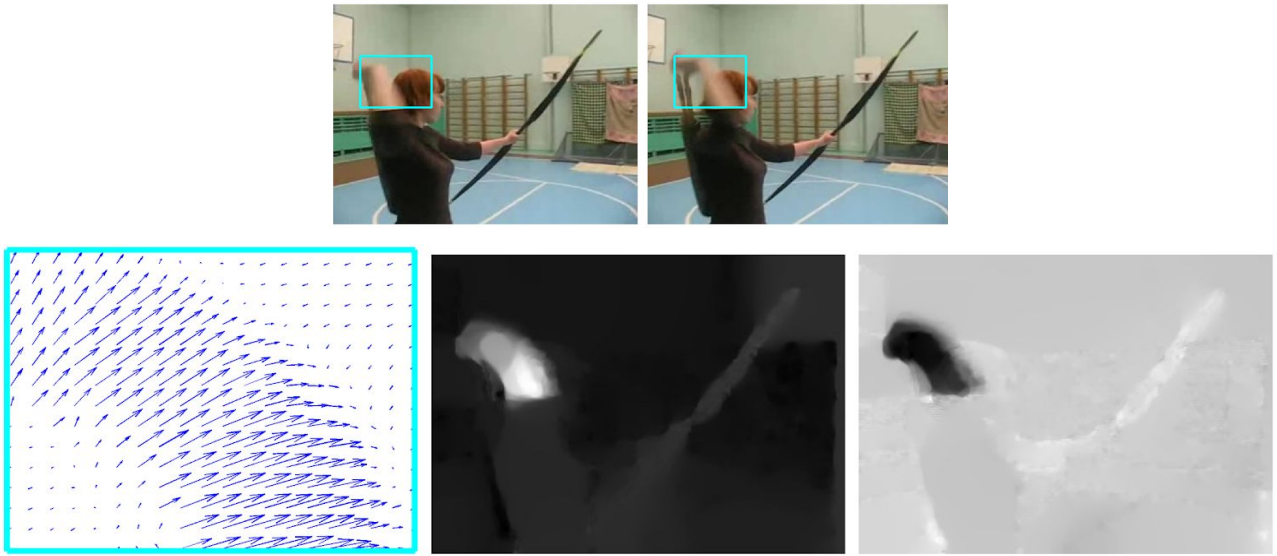
- 3 x 3 x 3 homogeneous kernels
- End-to-end: no human detection preprocessing required
- Compact features; new SOTA on several benchmarks

Two-Stream

Video = Appearance + Motion

Complementary information:

- Single frames: static appearance
- Multi-frame: e.g. optical flow: pixel displacement as motion information

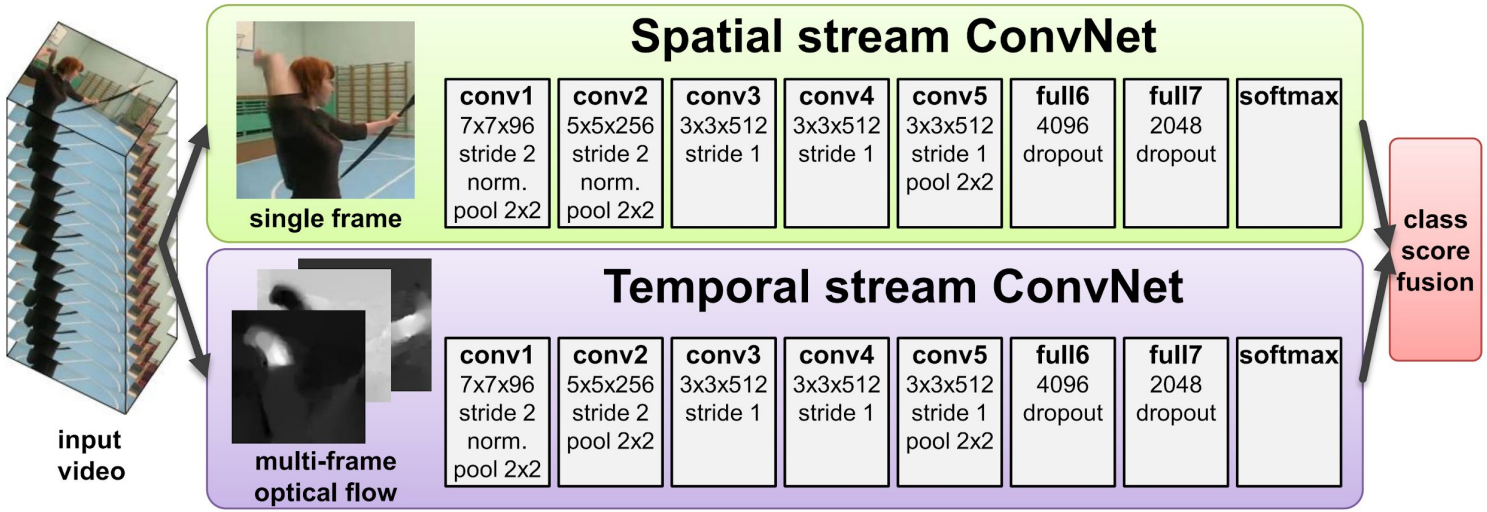


Two-Stream Convolutional Networks for Action Recognition in Videos (pdf)

Previous work: failed because of the difficulty of learning implicit motion

Proposal: separate **motion** (multi-frame) from **static appearance** (single frame)

- Motion: external + camera → mean subtraction to compensate camera motion



Two-Stream Convolutional Networks for Action Recognition in Videos (pdf)

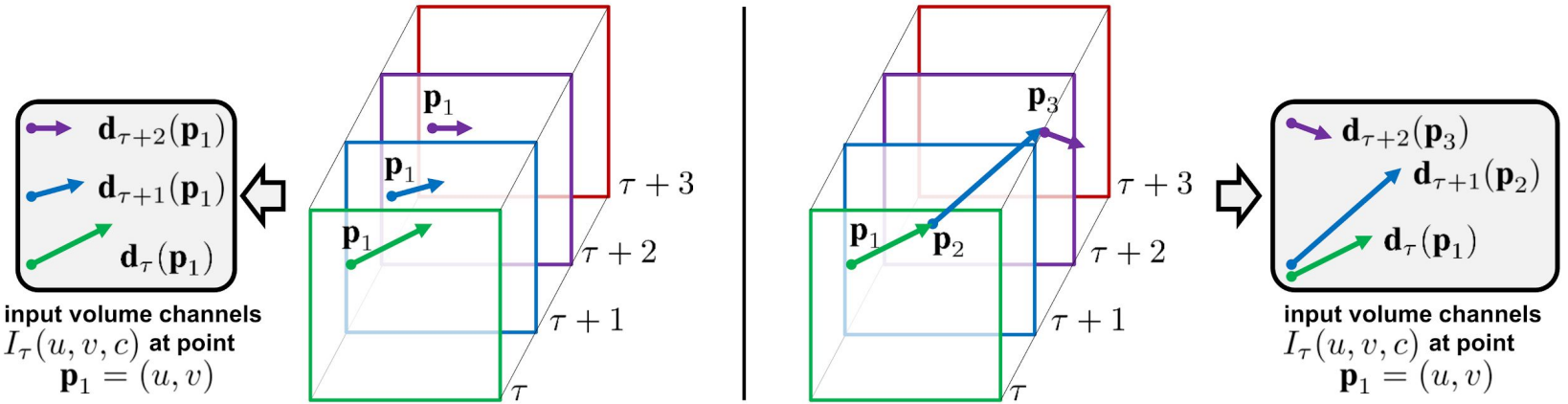
Two types of motion representations:

1) Optical flow

$$I_\tau(u, v, 2k-1) = d^x_{\tau+k-1}(u, v),$$
$$I_\tau(u, v, 2k) = d^y_{\tau+k-1}(u, v)$$

2) Trajectory stacking

$$I_\tau(u, v, 2k-1) = d^x_{\tau+k-1}(\mathbf{p}_k),$$
$$I_\tau(u, v, 2k) = d^y_{\tau+k-1}(\mathbf{p}_k)$$



Convolutional Two-Stream Network Fusion for Video Action Recognition (pdf)

Disadvantages of the previous two-stream network:

- The appearance and motion stream are not aligned
  - Solution: **spatial fusion**
- Lacking modeling of temporal evolution
  - Solution: **temporal fusion**

Convolutional Two-Stream Network Fusion for Video Action Recognition (pdf)

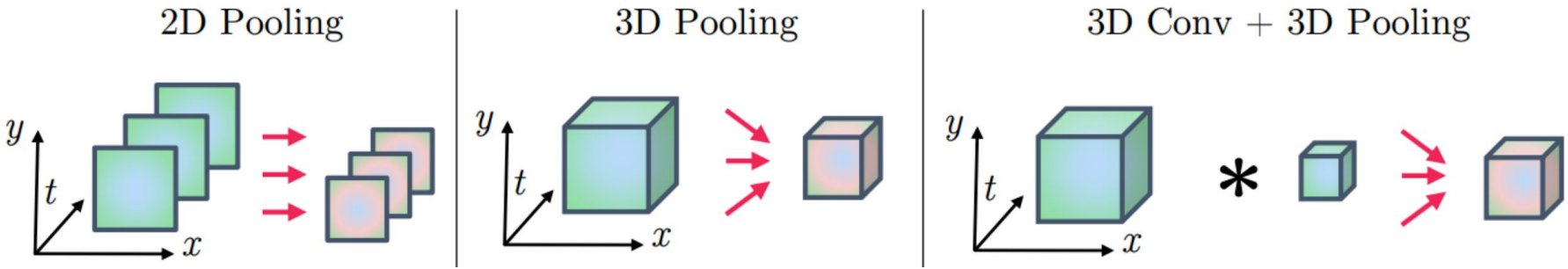
Spatial fusion:

- Spatial correspondence: upsample to the same spatial dimension
- Channel correspondence: **fusion**:
  - **Sum** fusion: $y^{\text{sum}}_{i,j,d} = x^{a}_{i,j,d} + x^{b}_{i,j,d}$
  - **Max** fusion: $y^{\text{max}}_{i,j,d} = \max\{x^{a}_{i,j,d}, x^{b}_{i,j,d}\}$
  - **Concat-conv** fusion: stacking + conv layer for dimension reduction
    - Learned channel correspondence: $\mathbf{y}^{\text{conv}} = \mathbf{y}^{\text{cat}} * \mathbf{f} + b$
  - **Bilinear** fusion: $\mathbf{y}^{\text{bil}} = \sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{x}^{a\top}_{i,j}\mathbf{x}^{b}_{i,j}$

Convolutional Two-Stream Network Fusion for Video Action Recognition (pdf)

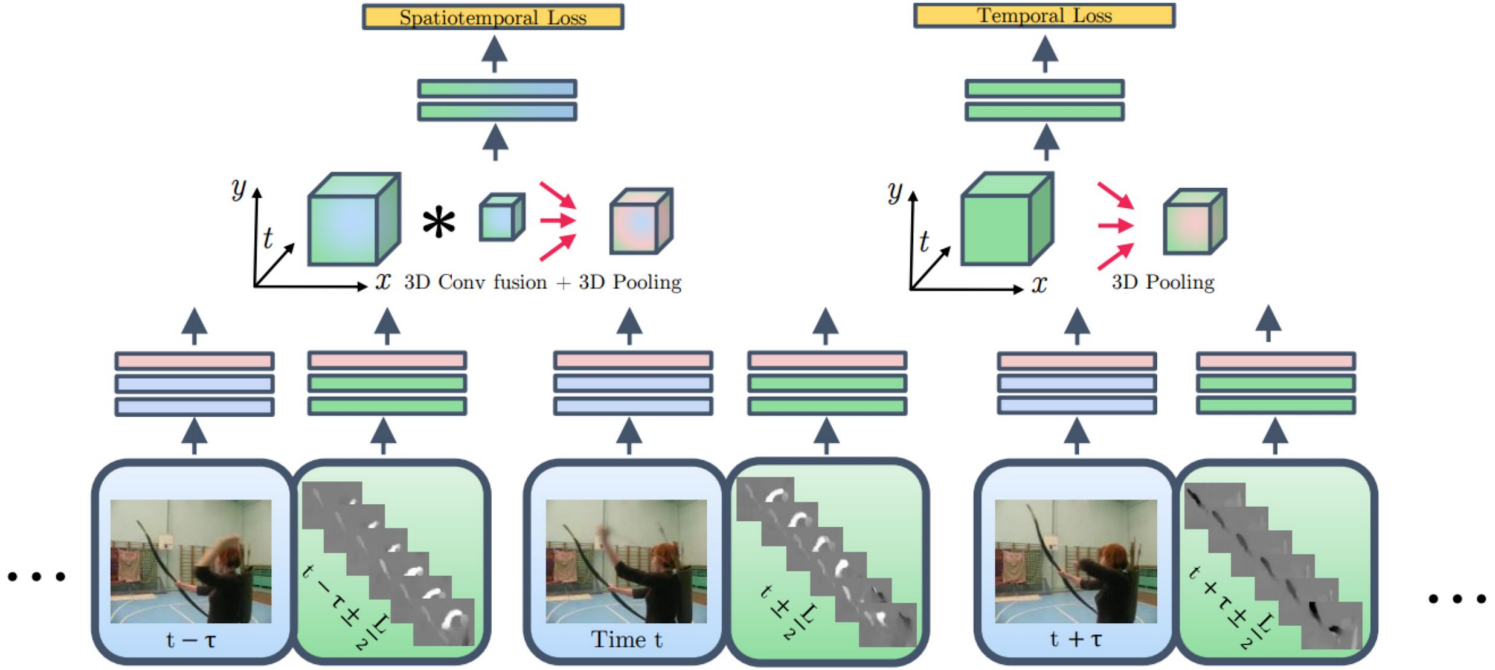
Temporal fusion:

- 3D pooling
- 3D Conv + pooling



Convolutional Two-Stream Network Fusion for Video Action Recognition (pdf)

Multi-scale: local spatiotemporal features + global temporal features



Model Takeaway

The motivations:

- CNN + RNN: video understanding as sequence modeling
- 3D Convolution: embed temporal dimension to CNN
- Two-stream: explicit model of motion

Further Readings

- CNN + RNN

❑ Unsupervised Learning of Video Representations using LSTMs (arXiv)
❑ Long-term Recurrent ConvNets for Visual Recognition and Description (arXiv)

- 3D Convolution

❑ I3D: integration of 2D info
❑ P3D: 3D = 2D + 1D

- Two streams

❑ I3D also uses both modalities

- Others:

❑ Objects2action: Classifying and localizing actions w/o any video example (arXiv)
❑ Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos (arXiv)