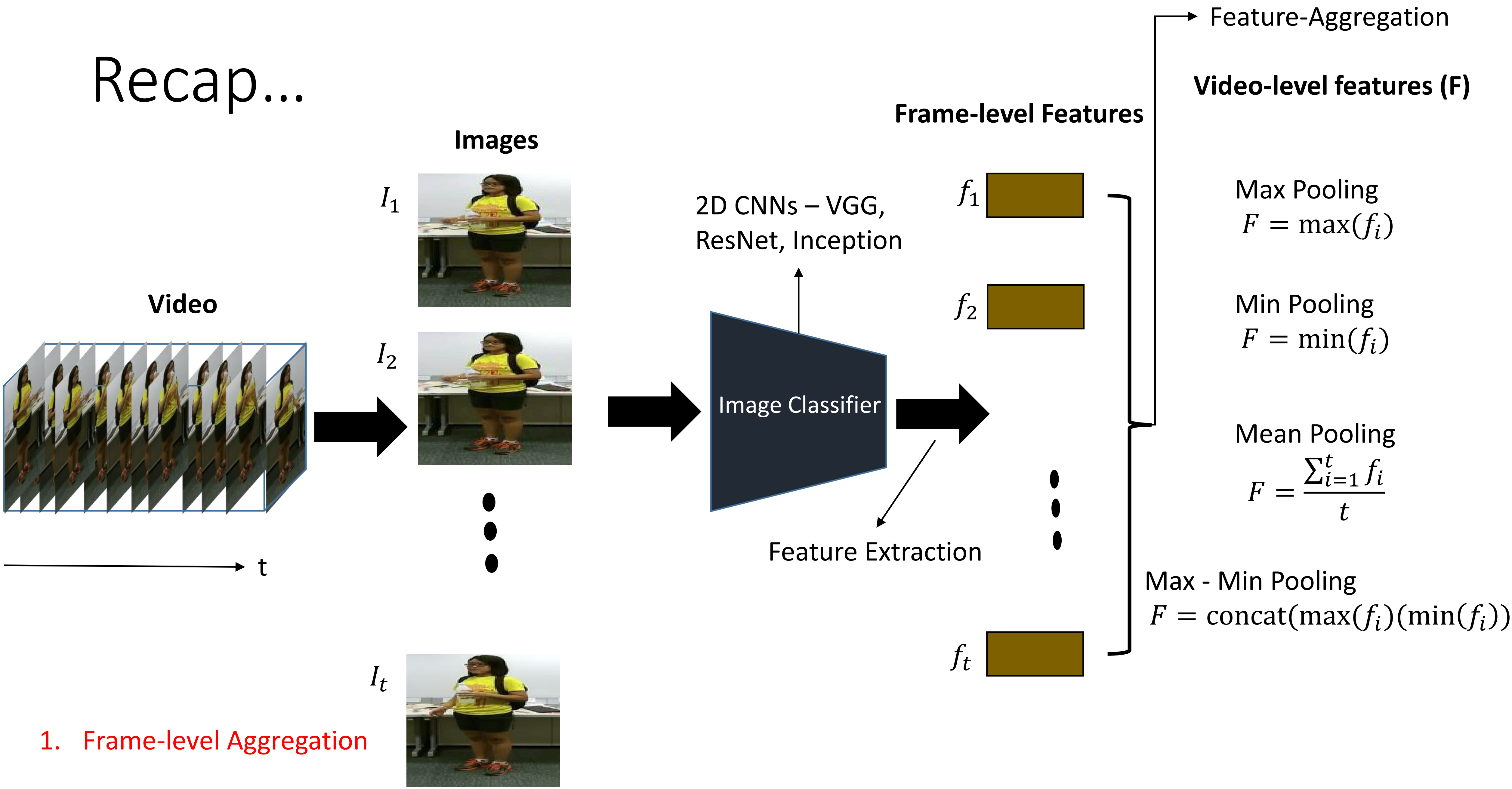
Deep Learning Winter School for Computer Vision

Srijan Das

PhD Scholar

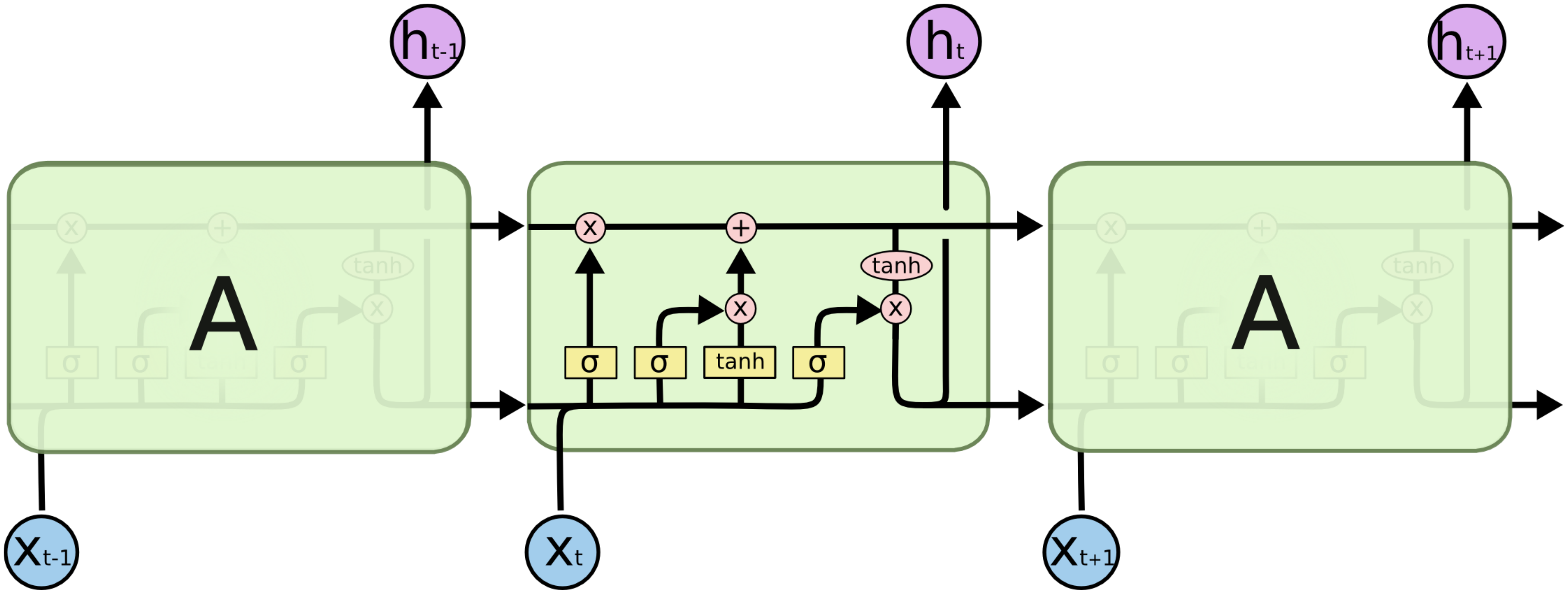
INRIA Sophia Antipolis

Recap…



1. Frame-level Aggregation

Recap...



2. LSTM

Disadvantages (not discussed in last class)

- RNNs/LSTMs can only capture strong temporal evolution of the image level features.

- Not much efficient on small datasets (pre-training is not a good idea as they change the statistics learned by the gates).

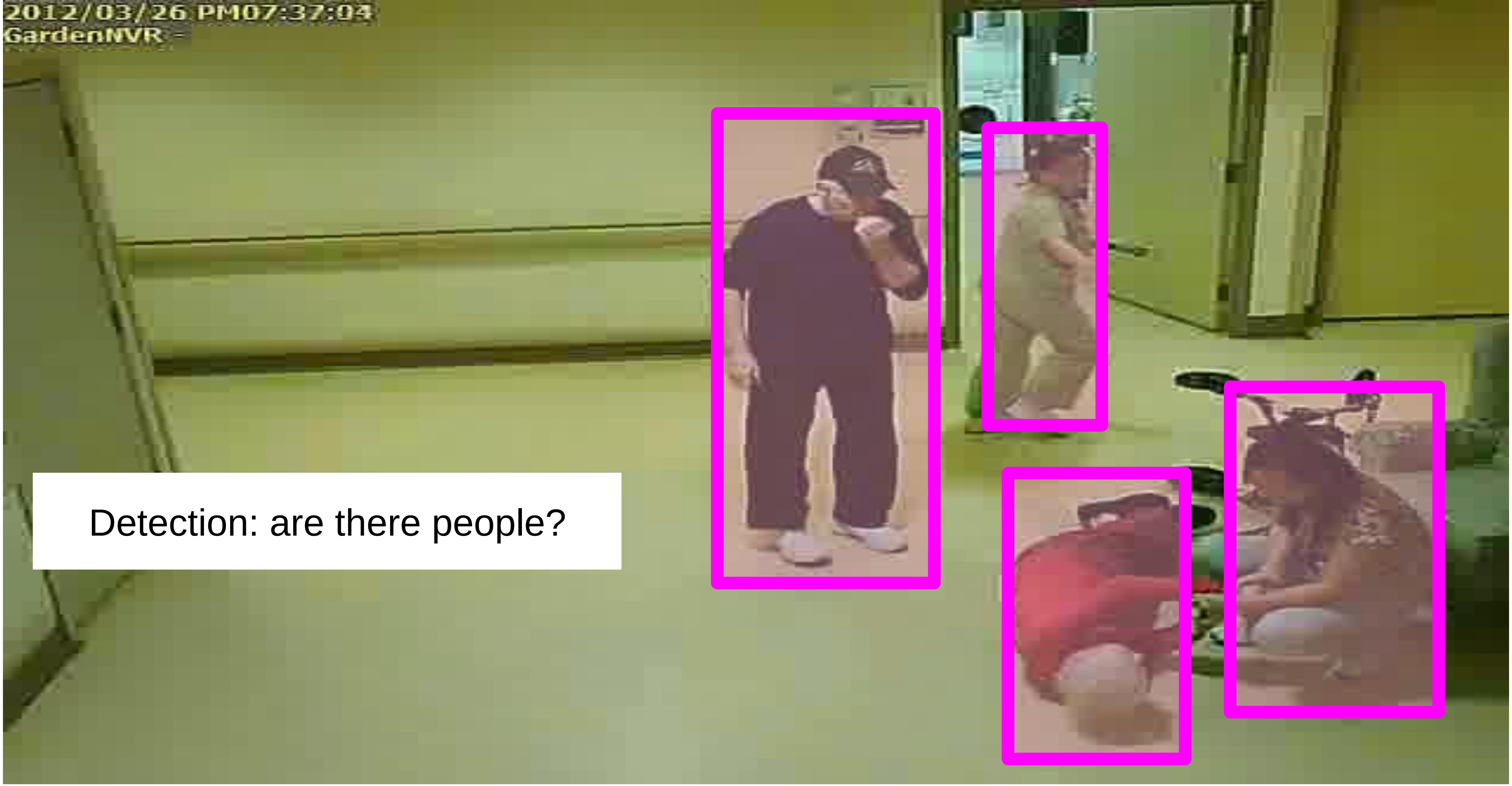
Outline: Action Recognition

- Introduction to Action Recognition

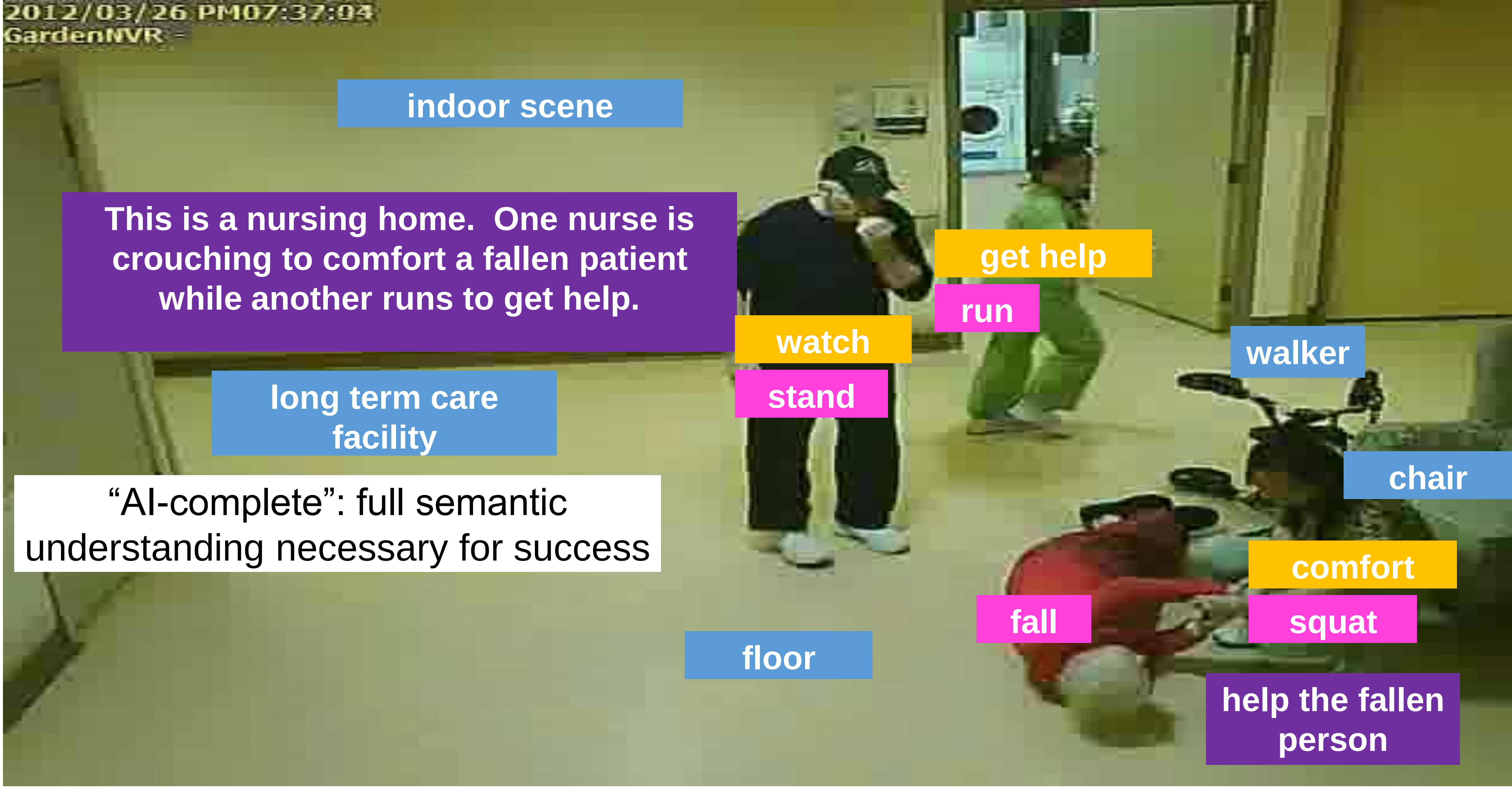
- Different Features for Action classification
  - RGB
  - Optical Flow
  - Skeleton

- Action Recognition Framework
  - Two-streams
  - LRCN
  - 3D ConvNets

What does activity recognition involve?


2012/03/26 PM07:37:04
GardenNVR -
Detection: are there people?

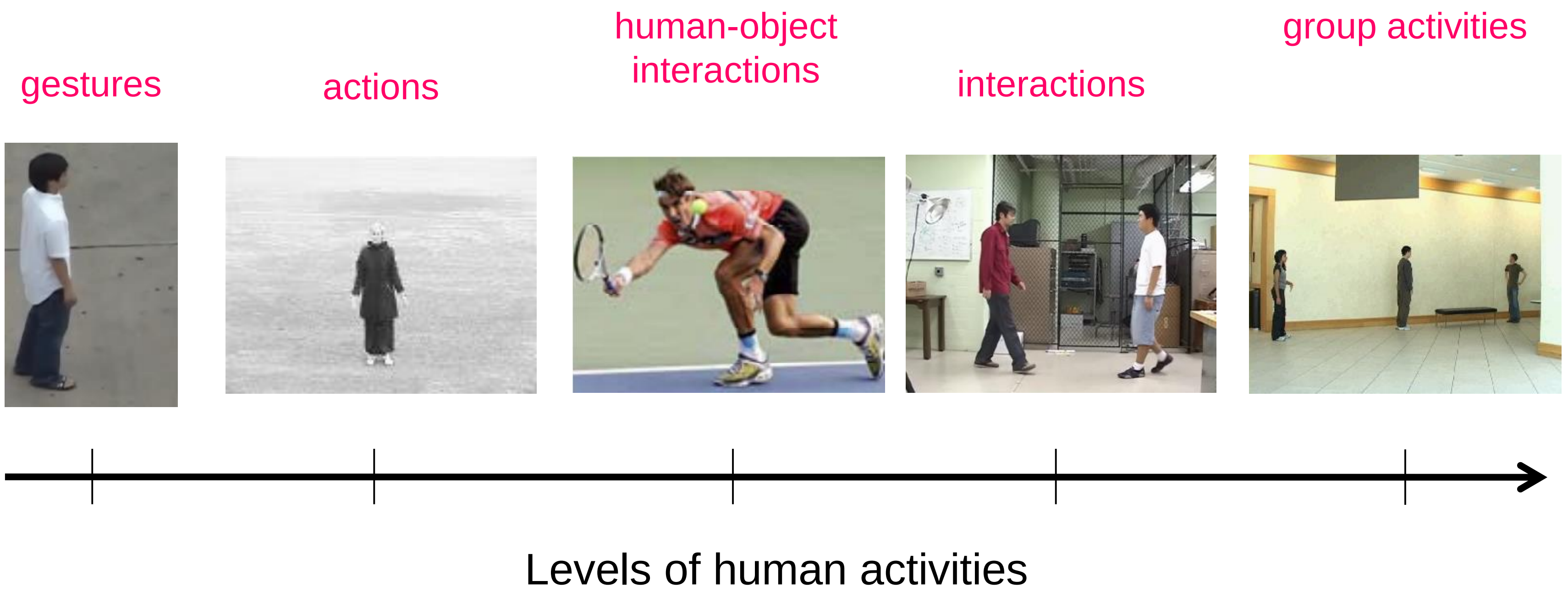
Action recognition: what are they doing?



“AI-complete”: full semantic understanding necessary for success

Human Activity Recognition

- There are various types/levels of activities
  - The ultimate goal is to make computers recognize all of them reliably.



Why is activity recognition important?

User videos

~300 hours of videos per minute

- Video indexing and retrieval

Monitoring cameras

Streaming videos 24/7

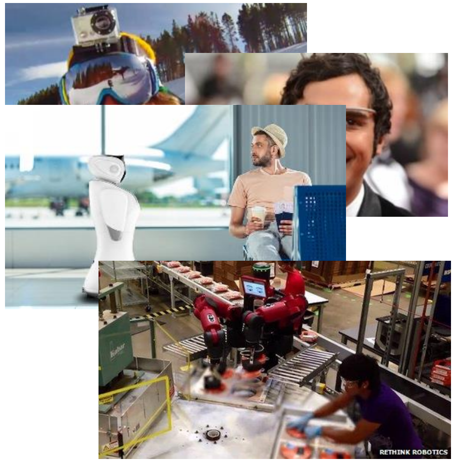
- Surveillance
- Patient/elderly monitoring

Media

Content analysis, experience enrichment

- Recommendation systems
- Advertising
- Sports analytics

Wearables/robots



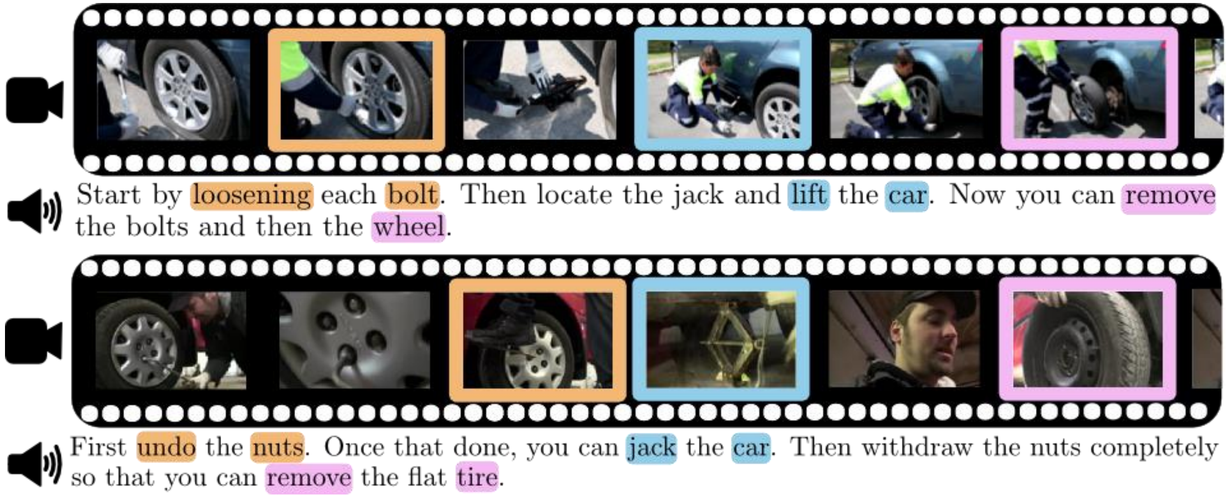
Streaming videos to be analyzed in real-time

- Lifelogging
- Robot operations and actions

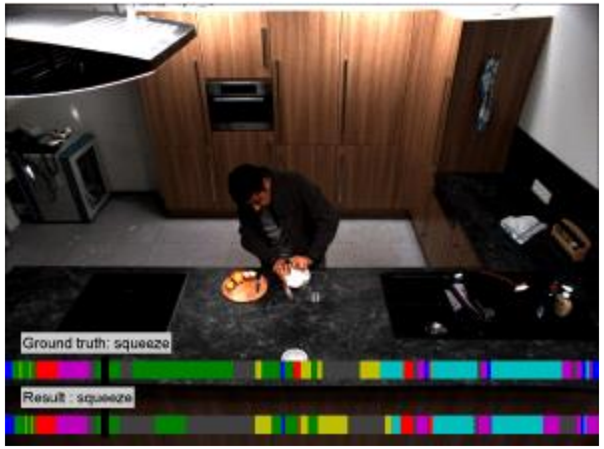
Categories of Action Recognition Data

Sports 1M

Instruction videos



Cooking



Internet (Youtube8M)

Ego-centric

Categories of Action Recognition Data

Activities of Daily Living

Web videos vs Activities of Daily Living (ADL)

Web Videos

ADL

Challenges in ADL

Drinking

Drinking

- Same background
- High intra-class variation

Challenges in ADL

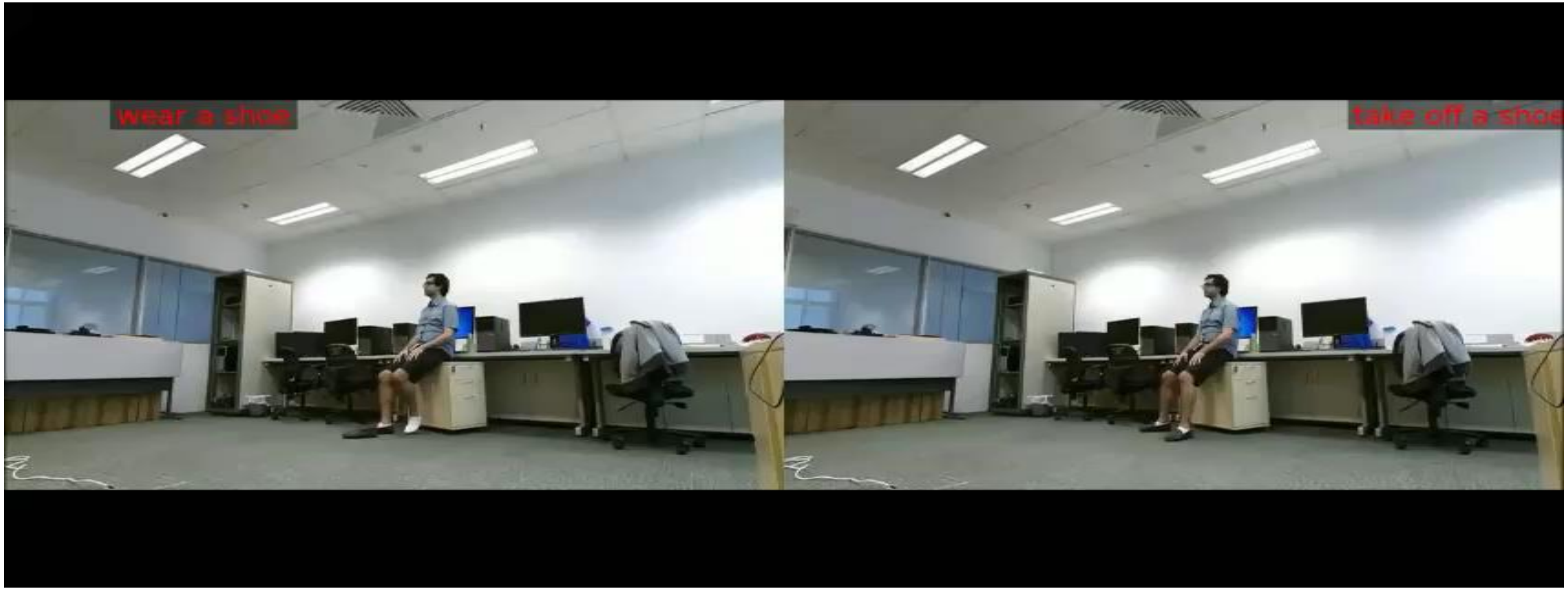
Typing a keyboard **Reading**

- Same background
- Actions with subtle motion

Challenges in ADL

Wear a shoe

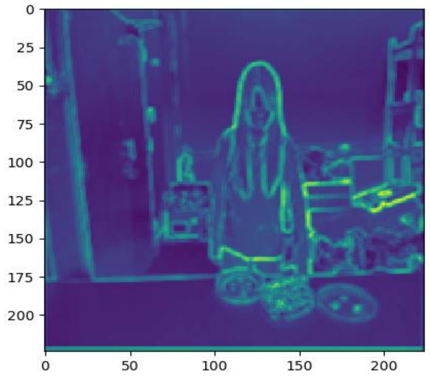
Taking off a shoe



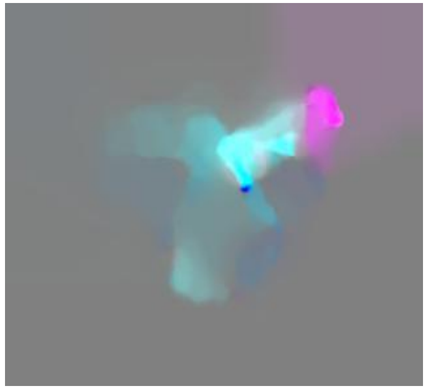
- Same background
- Actions with similar appearance

Different features for modeling Actions

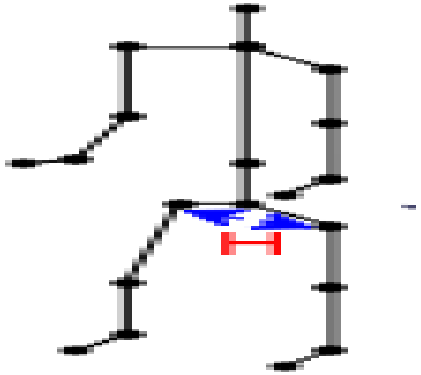
Appearance (RGB)



Optical Flow



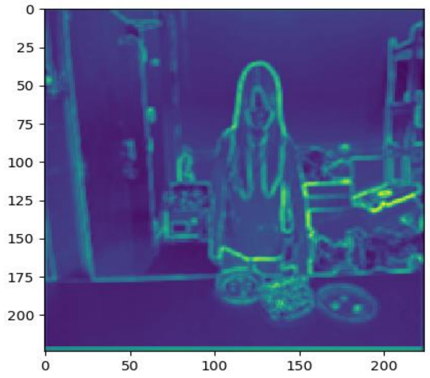
3D Poses



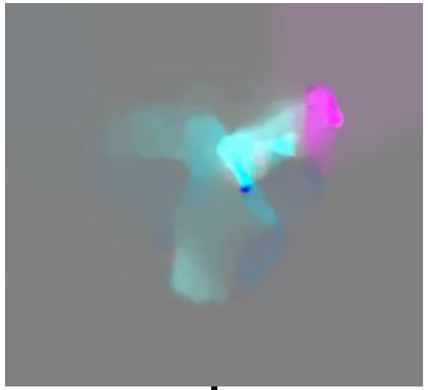
Process RGB images
in standard CNNs, RNNs

Different features for modeling Actions

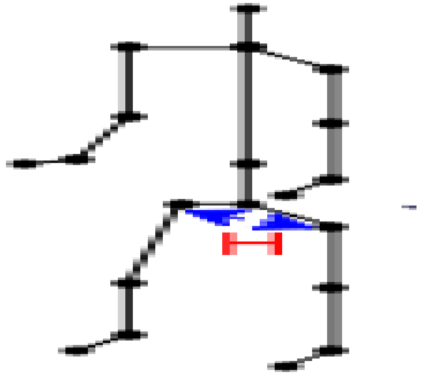
Appearance (RGB)



Optical Flow



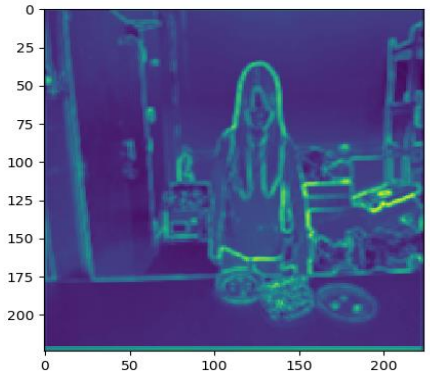
3D Poses



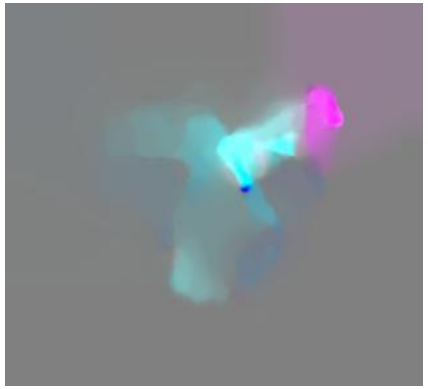
Process color coded optical flow images in standard CNNs.

Different features for modeling Actions

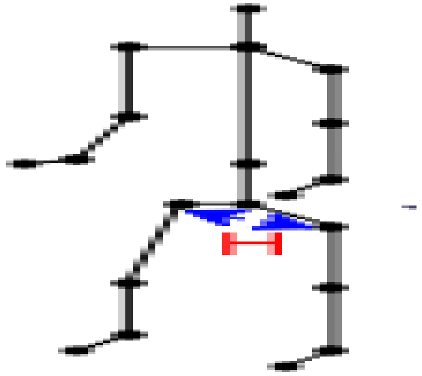
Appearance (RGB)

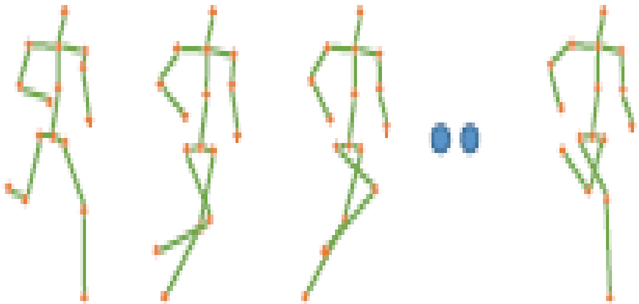


Optical Flow



3D Poses

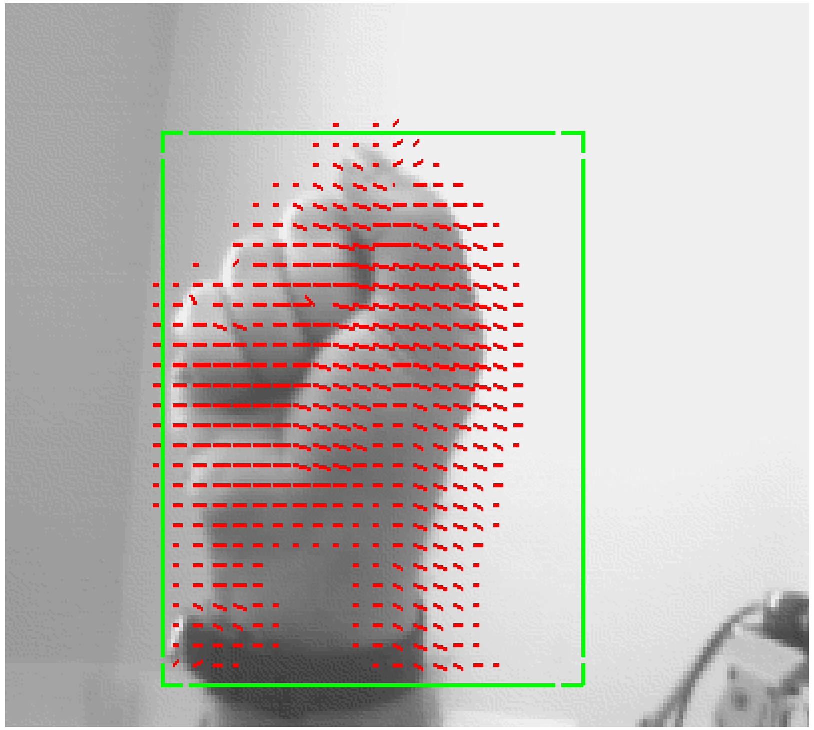




3D poses are highly informative, robust to illumination and view changes.
They are processed by RNNs, CNNs (especially Graph CNNs)

Optical Flow

- Computes the displacement of each pixel compared to the previous frame. (How much does the pixel move?)
- Represented by two displacement vectors (one along x, another along y).



Optical Flow

It is 2D vector field where each vector is a displacement vector showing the movement of points from first frame to second.

Brightness constancy assumption

$$f(x,y,t) = f(x+dx, y+dy, t+dt)$$

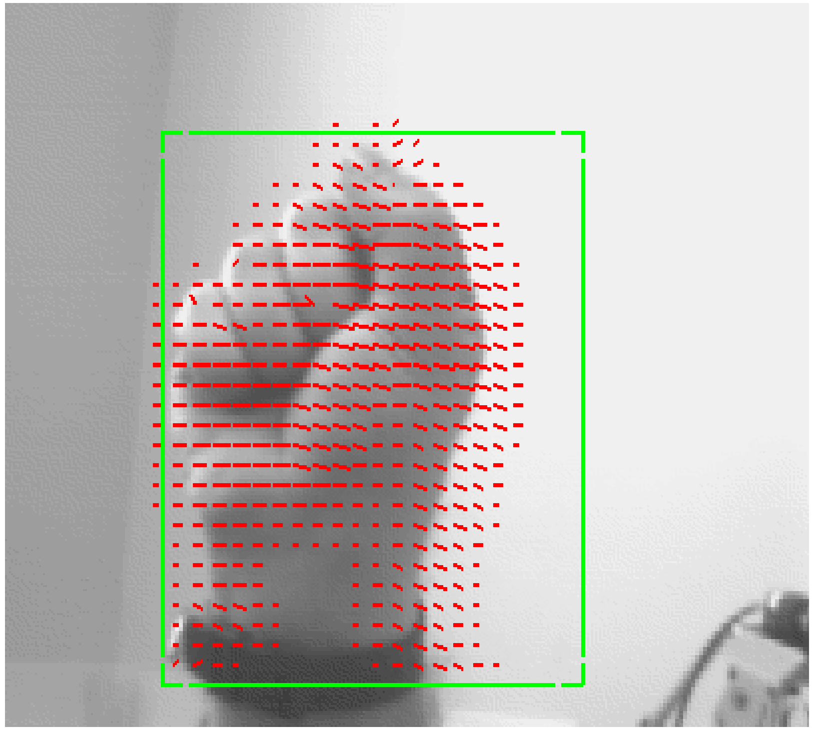
Taylor Series

$$f(x,y,t) = f(x,y,t) + \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial t}dt$$

$$f_x dx + f_y dy + f_t dt = 0$$

$$f_x u + f_y v + f_t = 0$$

Optical Flow equation



We cannot solve this one equation with two unknown variables. So several methods are provided to solve this problem and one of them is Lucas-Kanade.

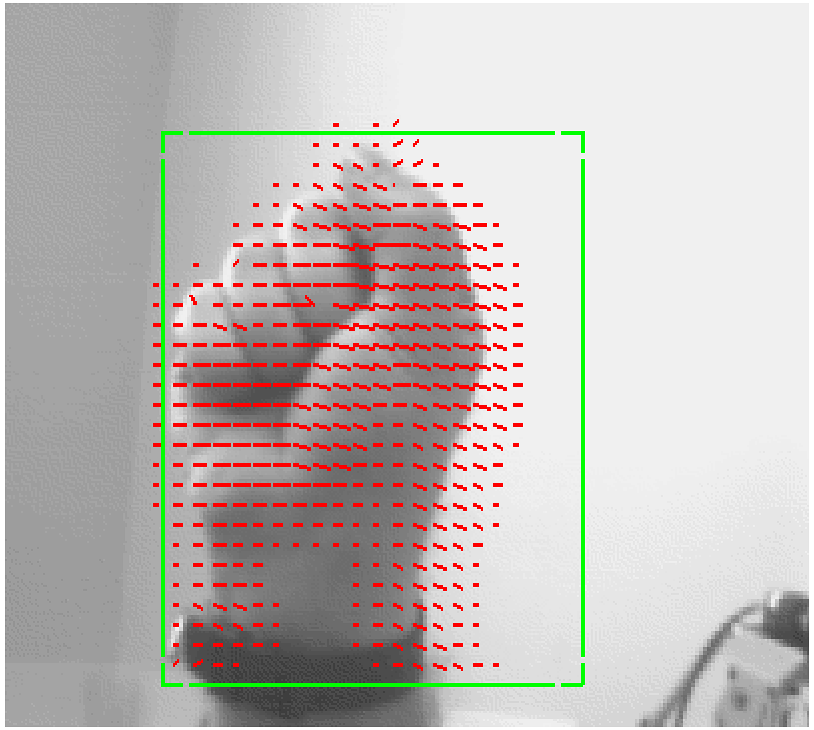
Optical Flow

Color Coded Optical Flow -> We call them flow images.

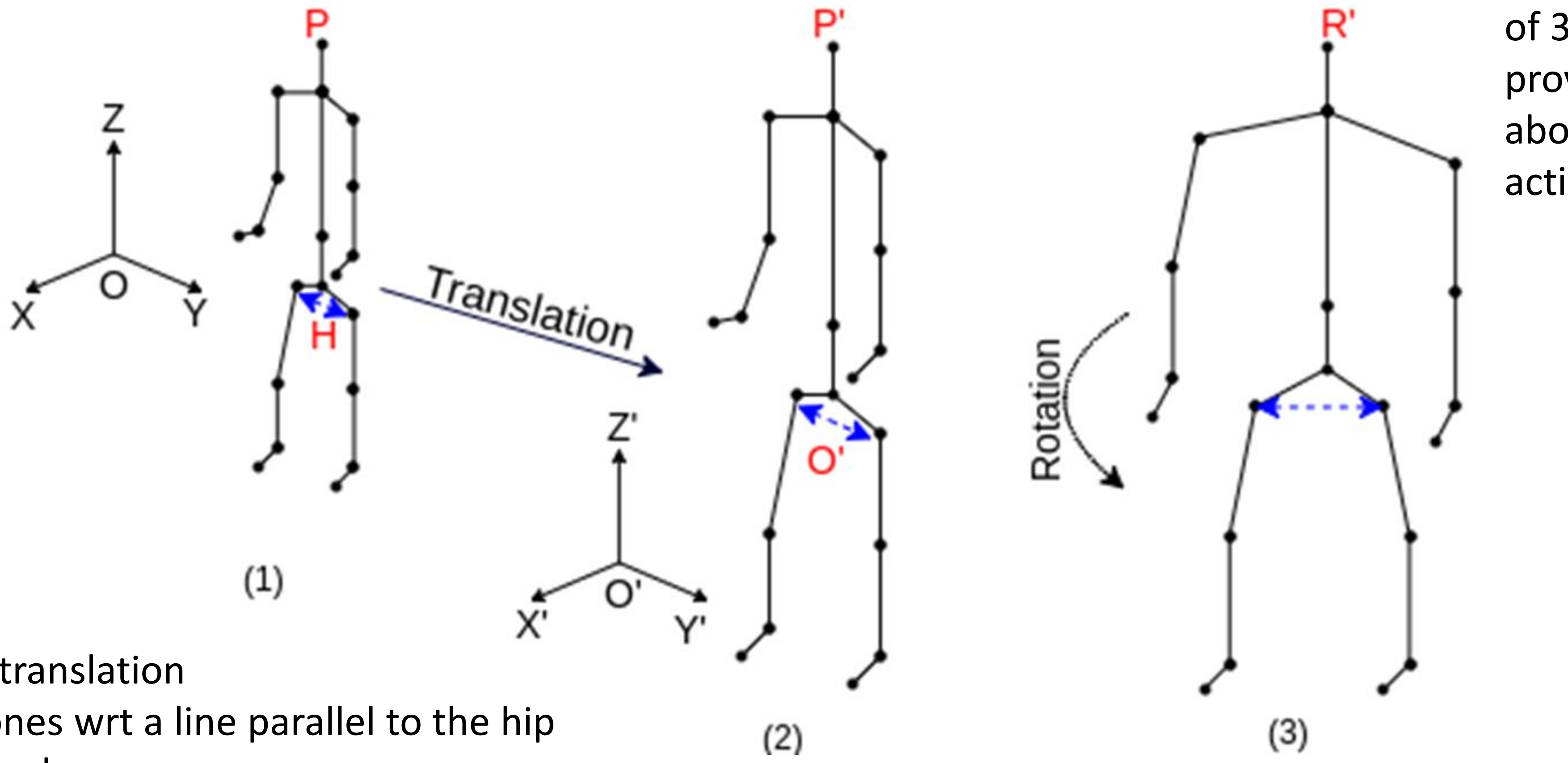
These flow images can be used in 2D CNNs for feature extraction.

Optical Flow

- Is informative for instantaneous motion.
- Thus used in Action classification tasks.



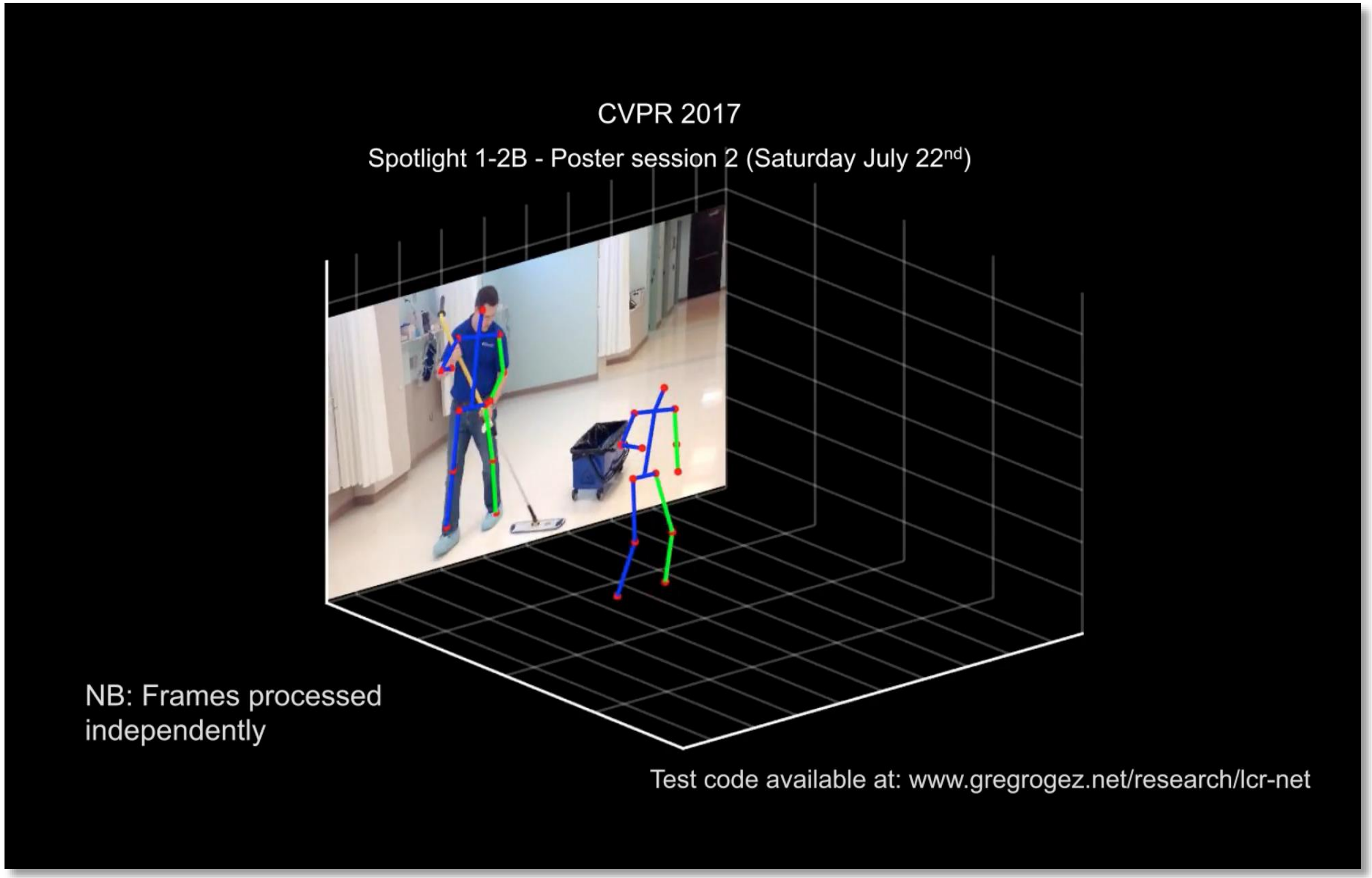
3D Poses (Skeletons)



Temporal evolution of 3D poses can provide inferences about pose related actions.

- Camera-body translation
- Rotation of bones wrt a line parallel to the hip
- Normalizing the bones

3D Poses (from RGB)



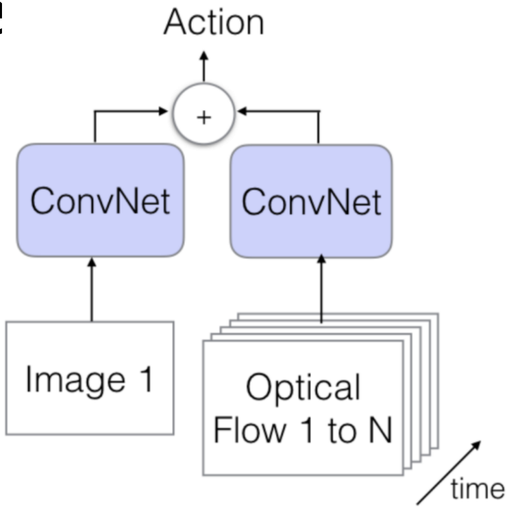
3D Poses

- The temporal evolution of these highly informative 3D poses are often exploited for Action classification (especially in indoor settings).

- The 3D poses can provide strong clue of where (both space and time) an action is happening.

Popular Action Recognition Frameworks

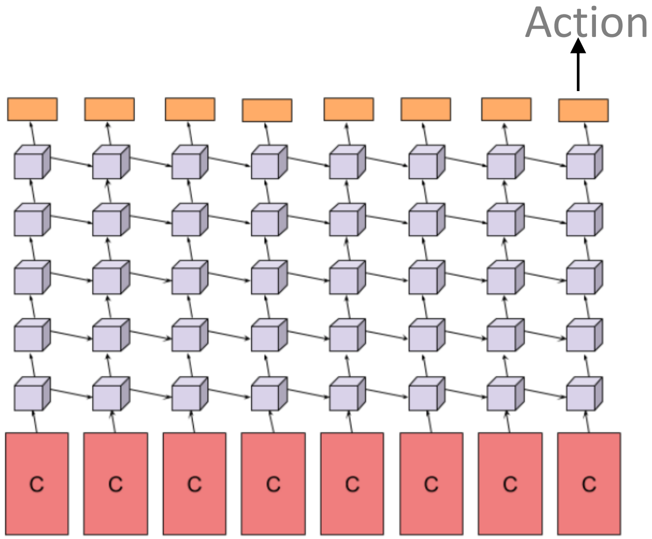
- Input: a fixed number of frames, Output: a class labe'



Two-stream CNNs

- 1 frame **RGB** + 10 frames of **optical flow**

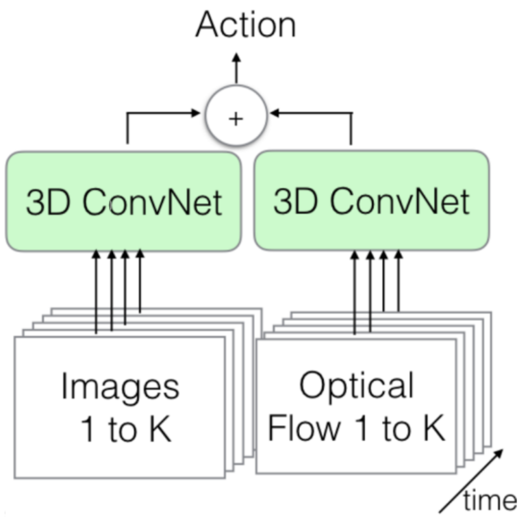
[Carreira and Zisserman, 2017]



Sequential models RNNs

- model 'sequences' of per-frame CNN representations (**RGB/3D Poses**)

[J. Ng et al., 2015]

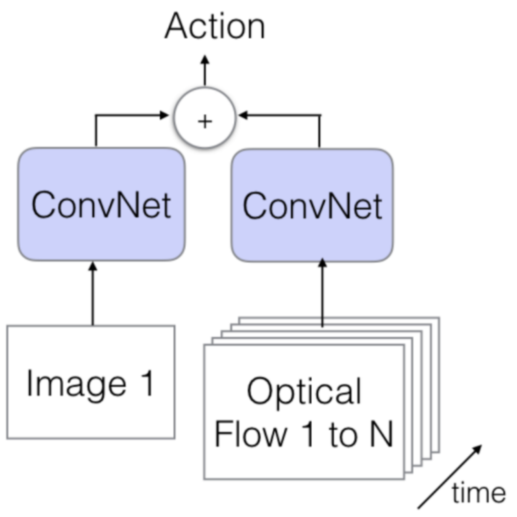


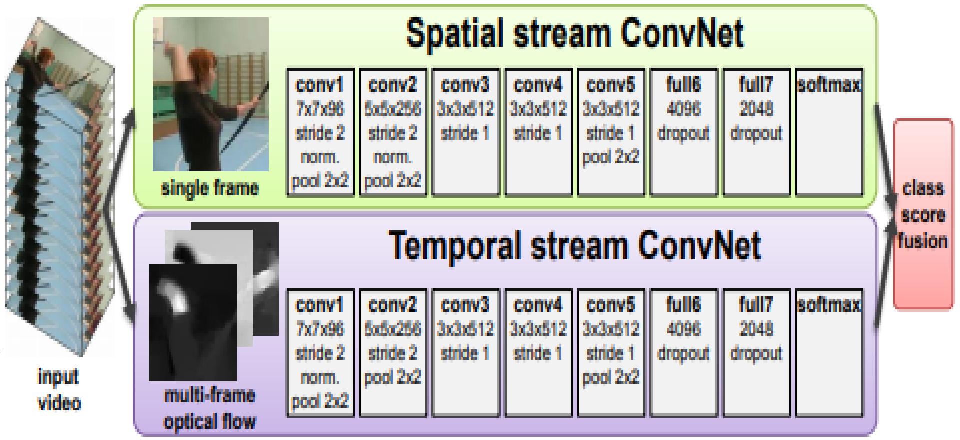
3-D XYT CNNs

- 15~99 frames (**RGB** + **Flow**)
- Facebook C3D, Google I3D

Two stream CNNs

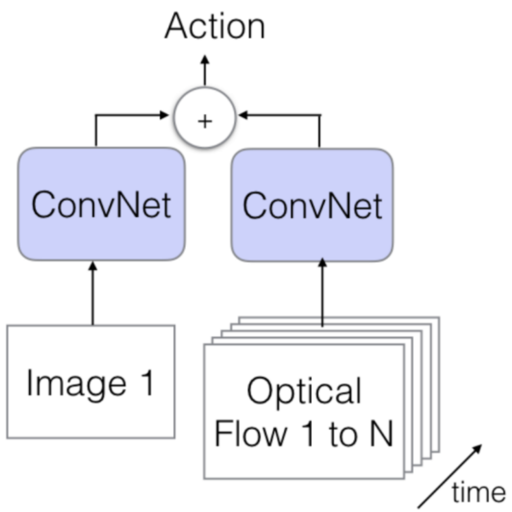
- **Introduction to optical flow ConvNets**
  - the input to this model is formed by stacking optical flow displacement fields between several consecutive frames.
  - stack the flow channels $d_t^{x,y}$ of L consecutive frames to form a total of $2L$ input channels
  - sample a 224 × 224 × 2L sub-volume from a video and pass it to the net as input

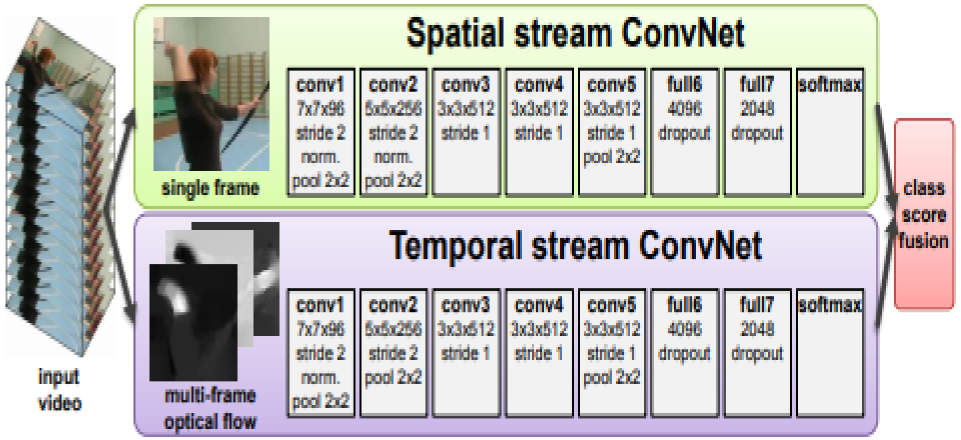




Two stream CNNs

- **Multitask Learning**
  - Input to the Spatial stream ConvNet - One image randomly sampled from the video. (encodes object/appearance information)
  - Input to the Temporal stream ConvNet – *2L* optical flow images from a video. (encodes short-term motion)
  - Both the networks learning together.



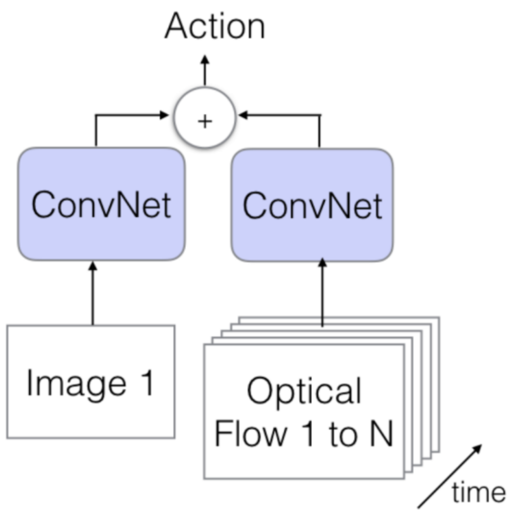


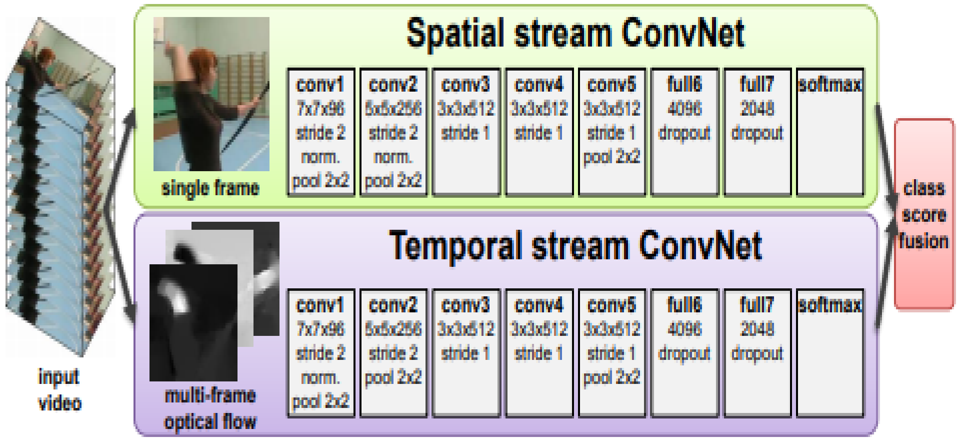
Two stream CNNs

A still from '**Quo Vadis**' (1951). Where is this going? Are these actors about to kiss each other, or have they just done so?

Two stream CNNs

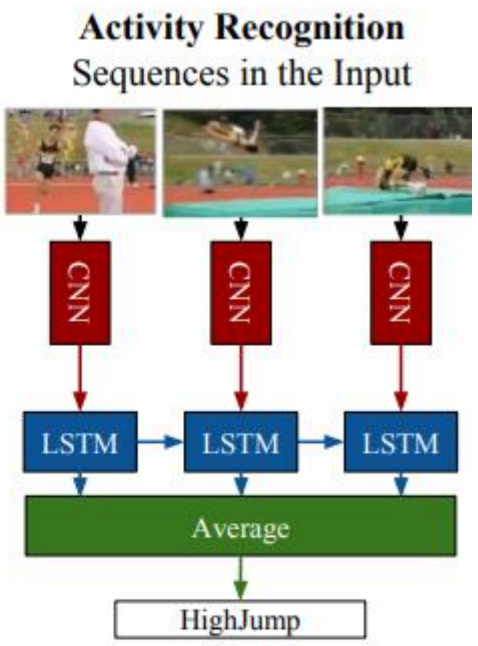
- **Disadvantages**
  - Temporal information is not encoded.
  - Long-term motion is ignored!





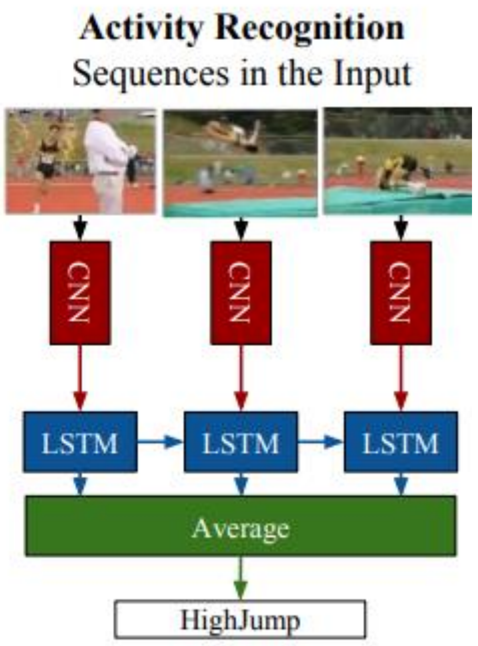
Long-term Recurrent Convolutional Networks for Action Recognition

- Obvious solution is, using sequential networks to model time.
- Uniformly sample images from the video, extract their CNN features and feed to LSTM.
- The Loss is computed from the average error at each time step of the LSTM.



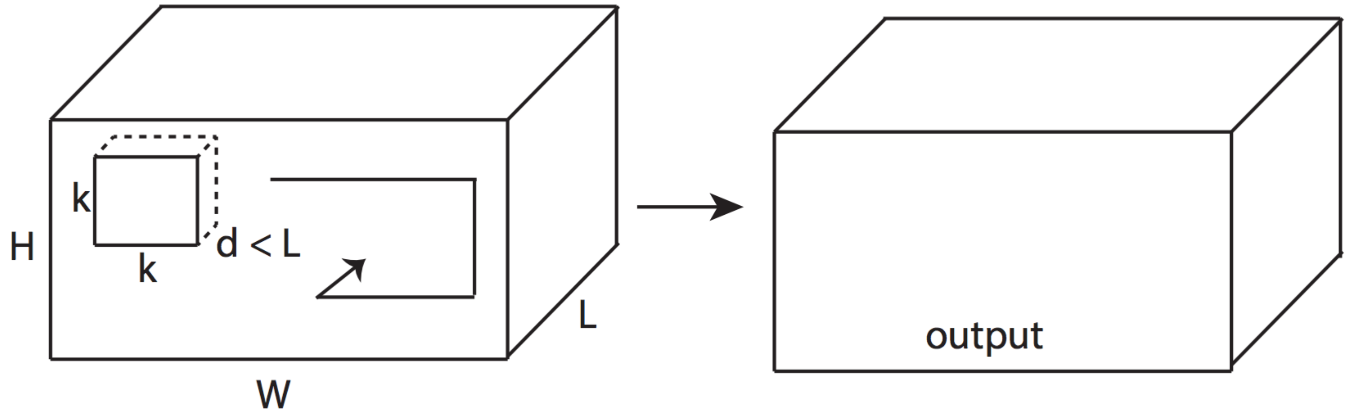
Long-term Recurrent Convolutional Networks for Action Recognition

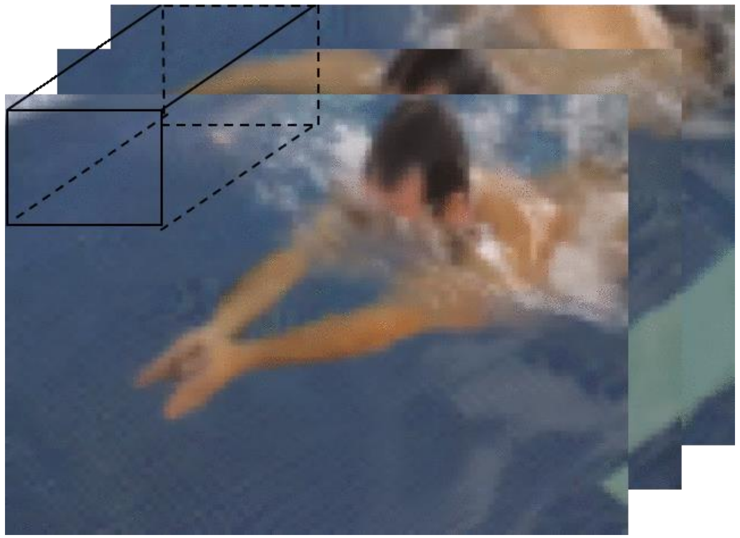
- Disadvantages
  - Doesn’t work for actions with subtle changes in the scene.
  - Spatial and temporal operations are dissociated disabling the model to extract intrinsic spatio-temporal patterns.



3D CNNs (XYT) for Action Recognition

- Facebook C3D [Tran et al., 2015]
  - Spatio-temporal filters for short video segments (e.g., 15 frames) – **coupling** space and time

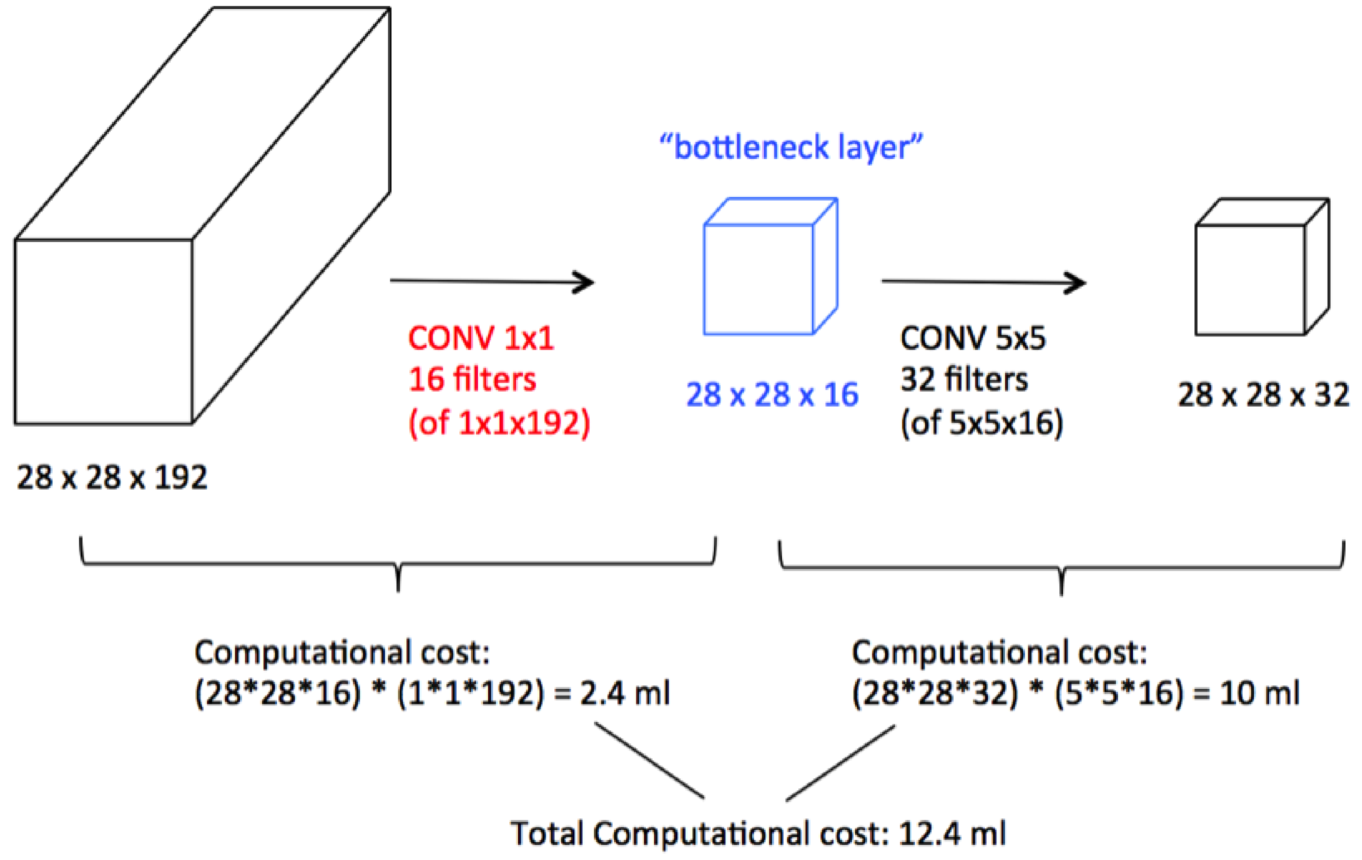




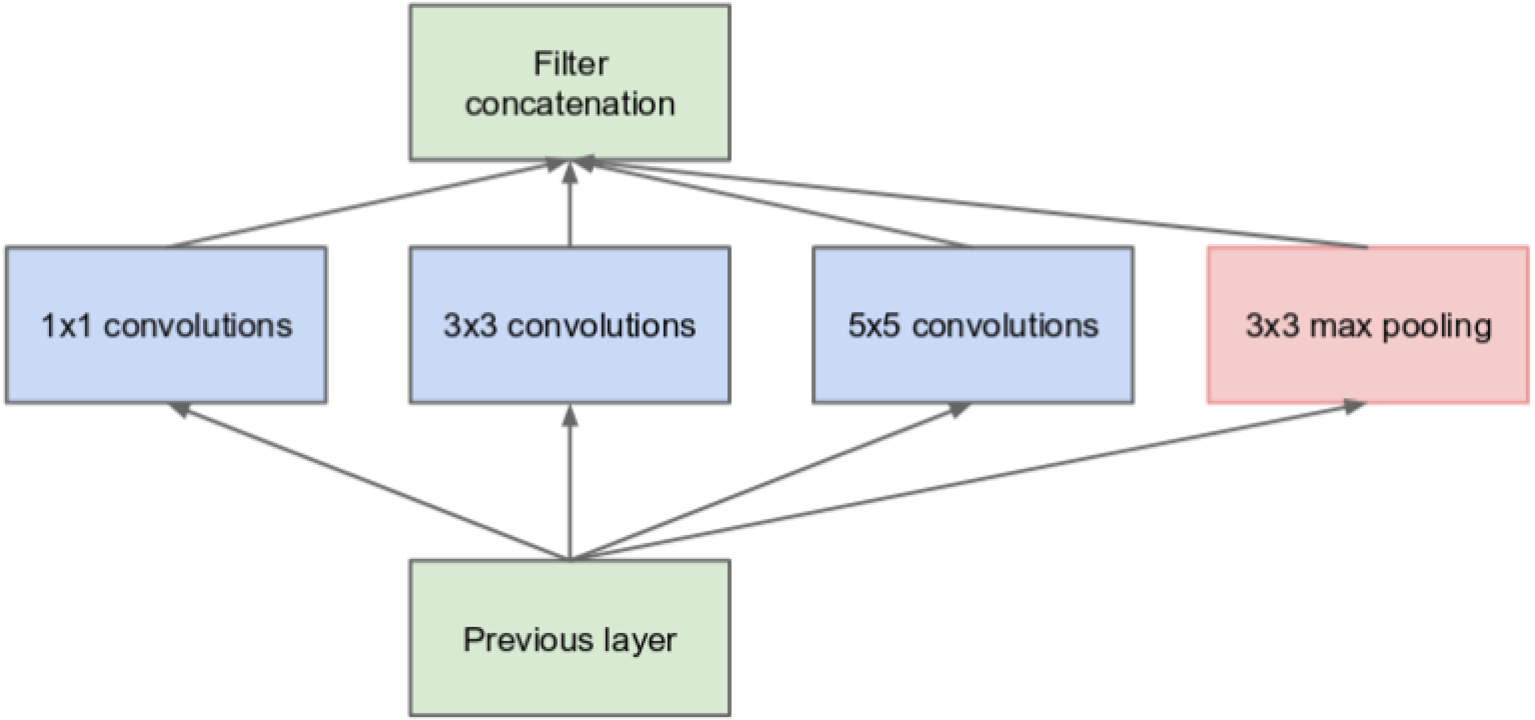
- Google I3D [Careirra et al., 2017]
  - Extended by inflation from Spatial domain

Pseudo-I3D, ResNet3D, Non-Local Network, Slow-fast Network,

Recap: Network in Networks (Bottleneck)



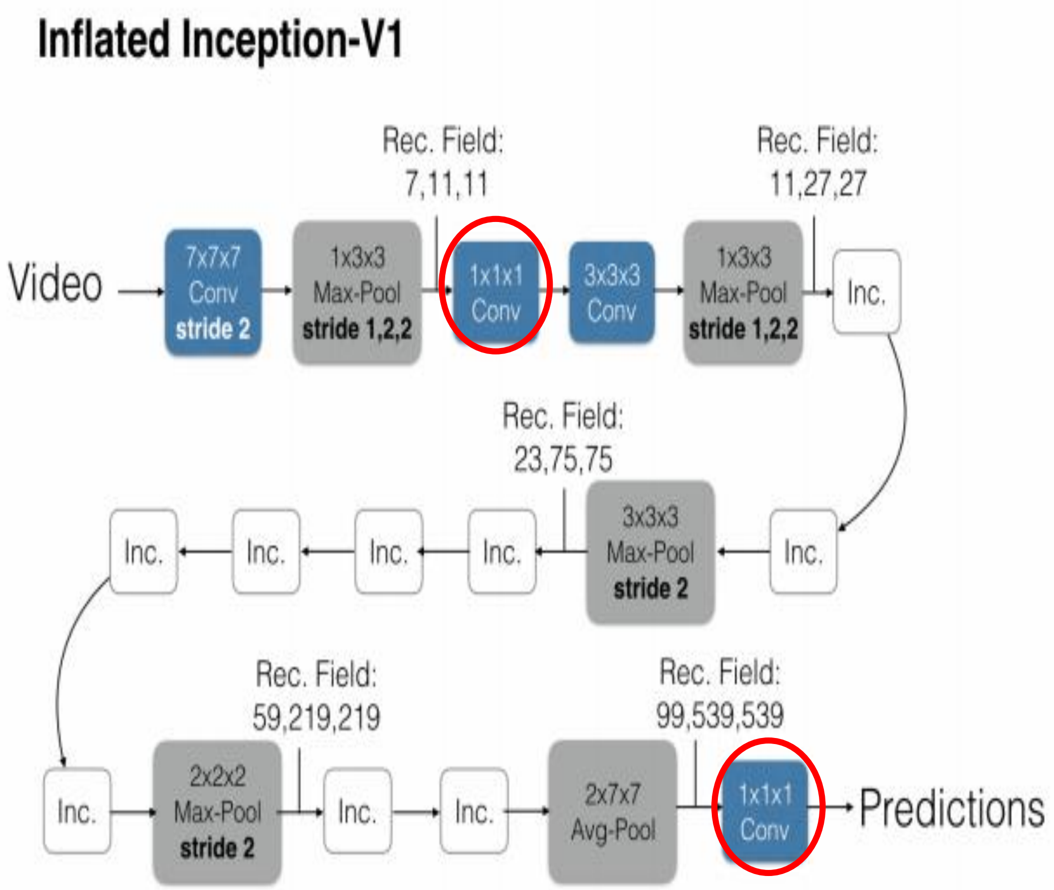
Recap: Inception

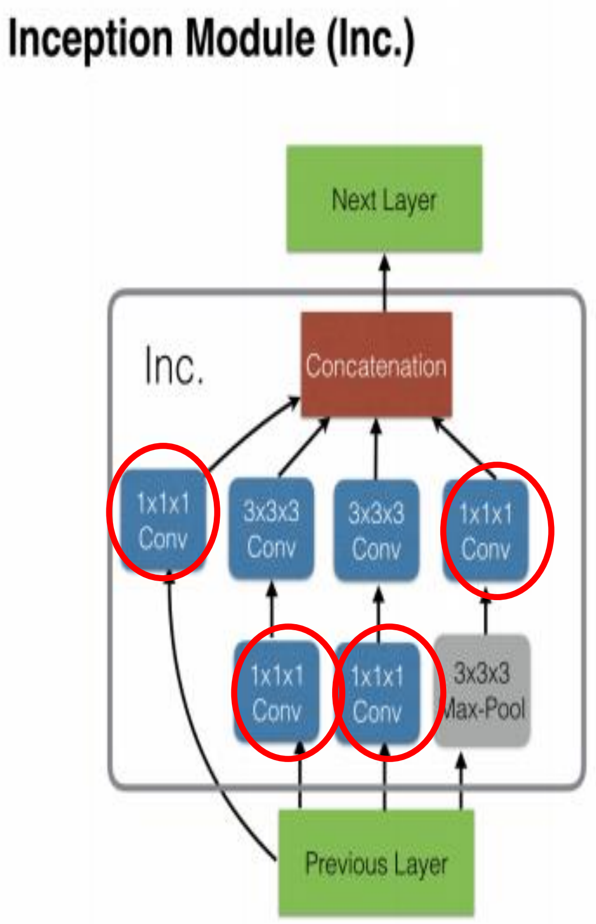


(a) Inception module, naïve version

I3D

- Inflation
- Bottleneck
- Concept of inception

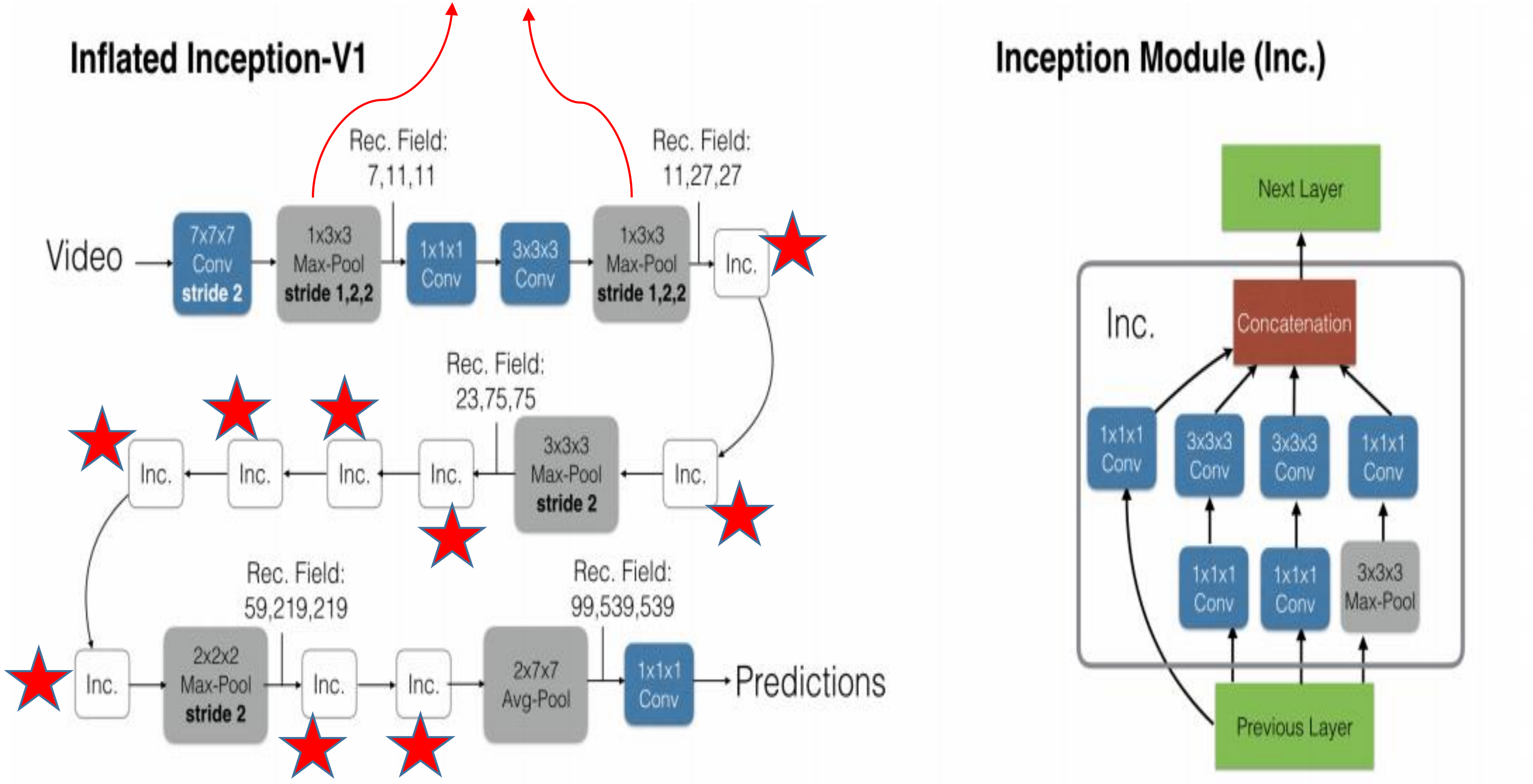




I3D

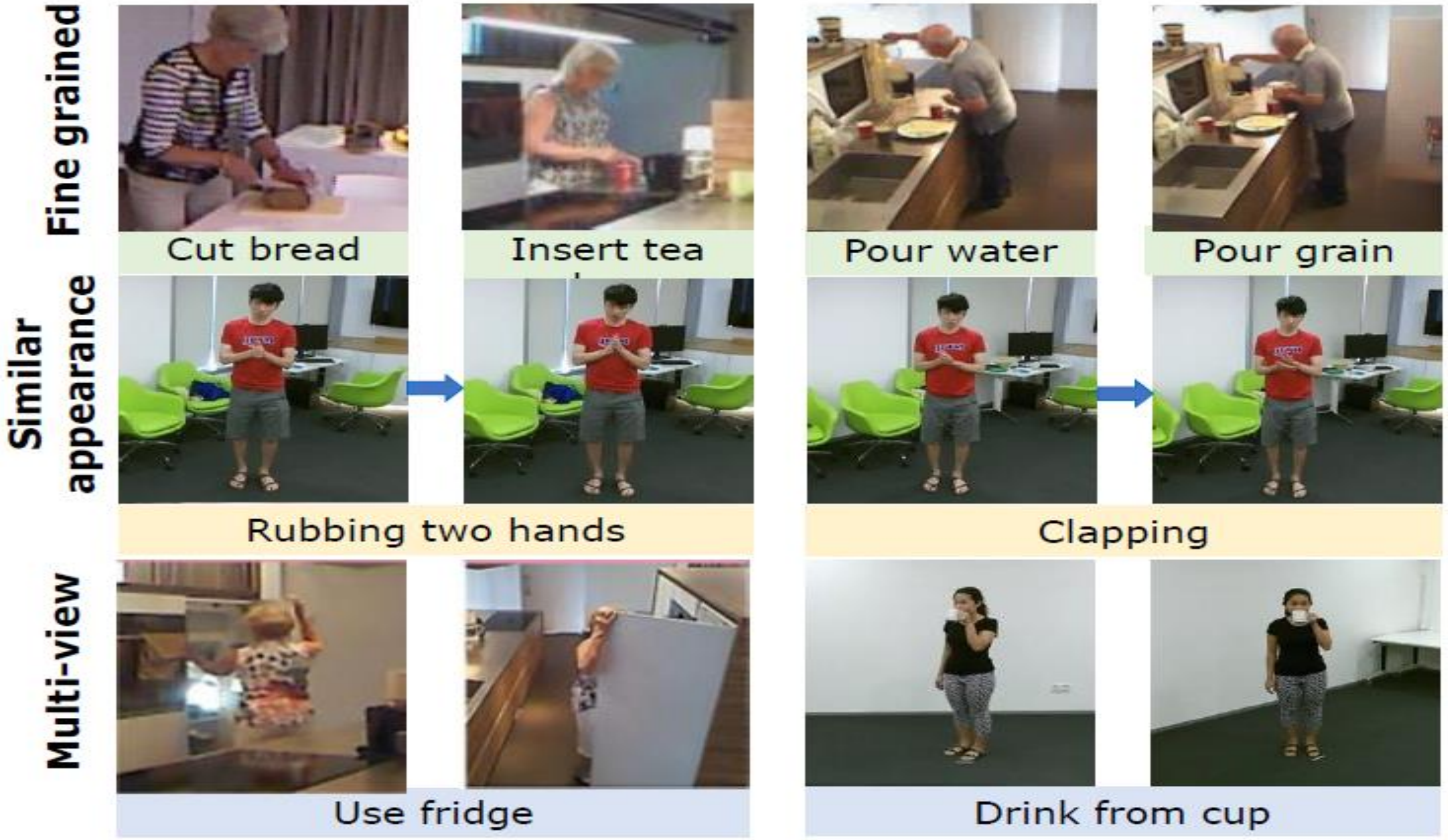
- Inflation
- Bottleneck
- Concept of inception

Handling space-time together with asymmetric operations



Limitations of 3D CNNs

- Rigid spatio-temporal kernels limiting them to capture subtle motion
- No specific operations to help disambiguate similarity in actions.
- 3D (XYT) CNNs are not view-adaptive.



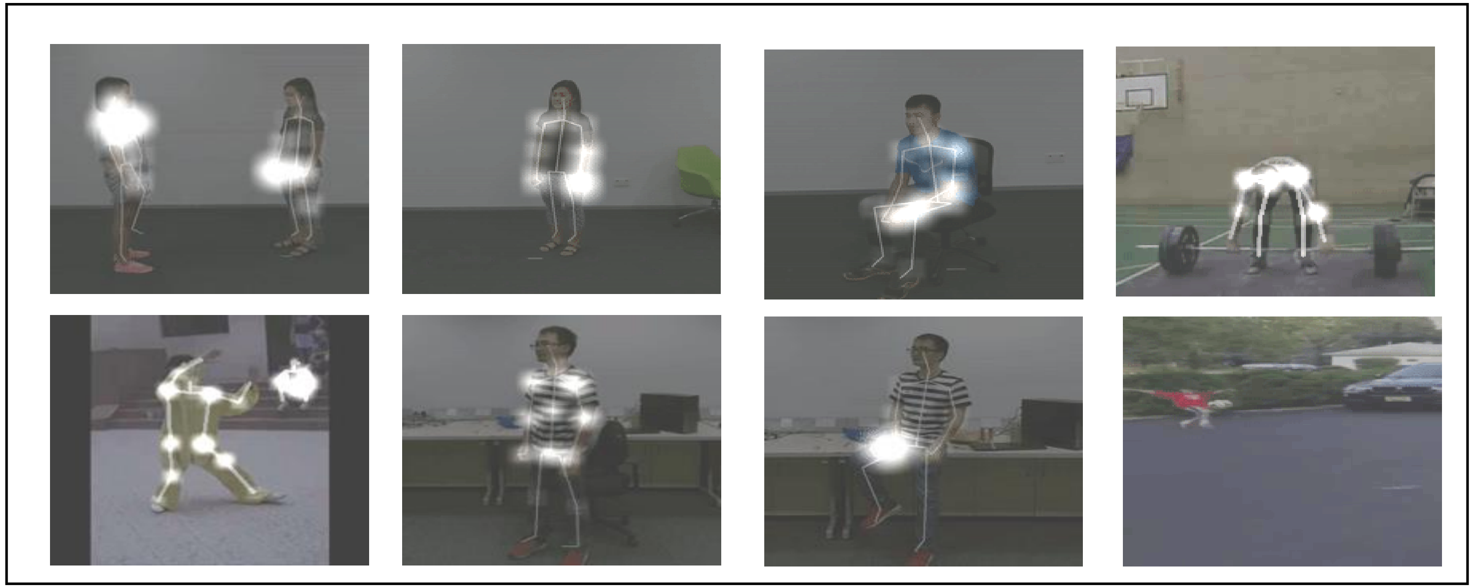
References

- Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, CVPR 2017
- UCF computer vision video Lectures 2012 (Instructor: Mubarak Shah)
- CVPR Tutorial, Human Activity Recognition (M. Ryoo, I. Laptev)

e-mail: srijan.das@inria.fr

Next Week ….

Attention!!!



Attention!!!