Lecture 6

Human Action Recognition

Abid Ali

Mohammed Guermal

✉ abid.ali@inria.fr, mohammed.guermal@inria.fr

Deep Learning for Computer Vision

GUERMAL Mohammed
3 year PhD student: Human activity recognition for human-robot interraction
Supervisor: BREMOND Francois
Publications : https://arxiv.org/abs/2204.09468

ALI Abid (abid.ali@inria.fr)
3 year PhD student: Act4Autism: Action detection for diagnosis of Autism in Children
Supervisors: BREMOND Francois, Thummler Susanne
Publications : [Abid Ali - Google Scholar]

Outline

- Introduction

- Different Modalities
  - RGB
  - Optical Flow
  - 3D Poses

- Deep Networks for Action Recognition
  - Two-stream network
  - LRCN
  - 3D ConvNets (I3D)

Section 1

Introduction

Video analysis

Large amount of videos are accessible

Why human actions?

How many person-pixels are in the video?

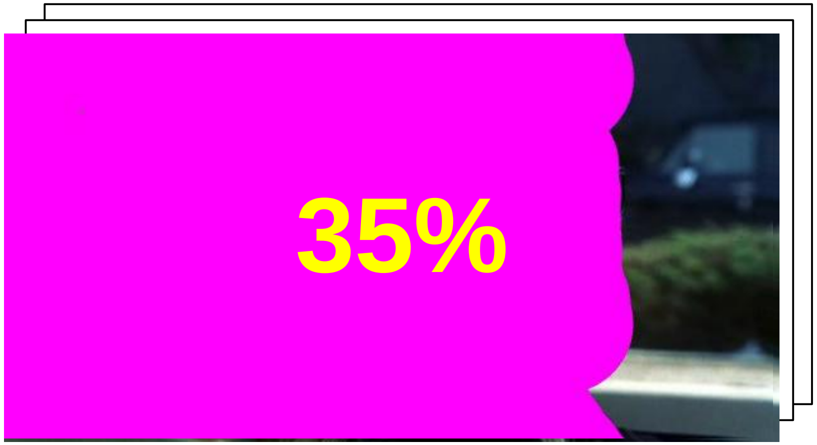
Movies

TV

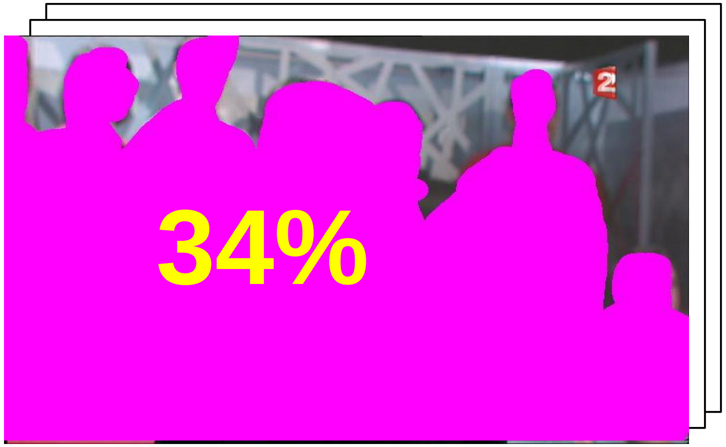
YouTube

Why human actions?

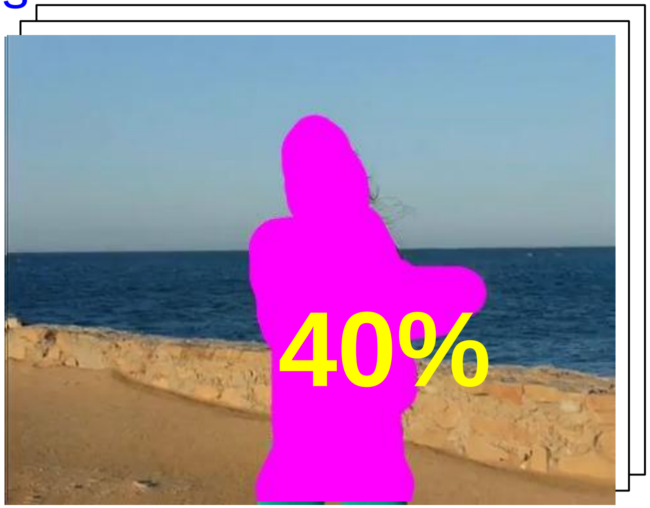
How many person-pixels are in the video?



Movies



TV



YouTube

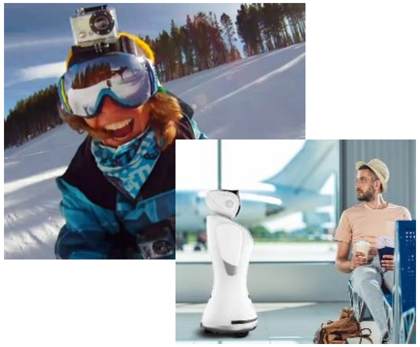
Plenty of videos are relevant to human being.

Data source

User videos/Media

Monitoring cameras

Robotics/
wearable cameras



~300 hours /minute

Streaming videos 24/7

Streaming videos
to be analyzed in real-time

- Recommendation systems
- Advertising

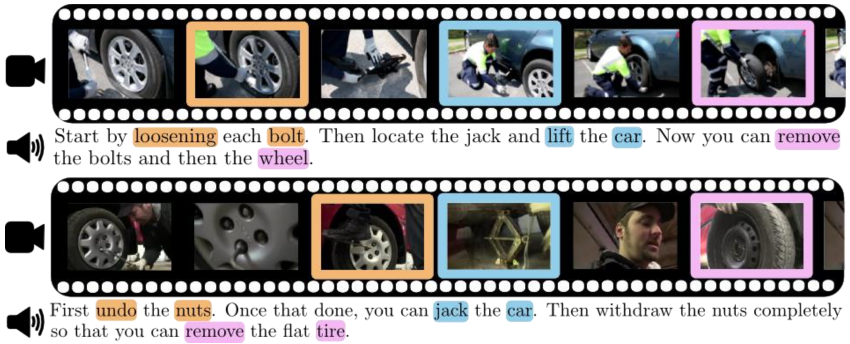
- Surveillance system
- Patient/elderly monitoring

- Life logging
- Robot operations and actions

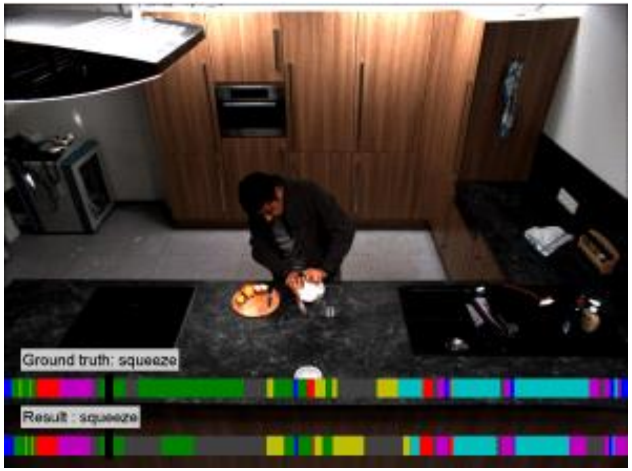
Categories of Action Recognition Data

Sports

Instruction videos



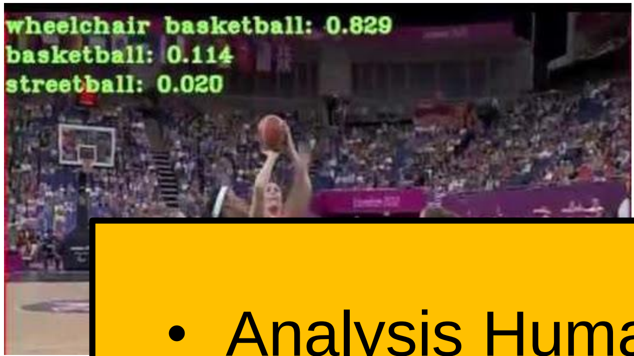
Cooking



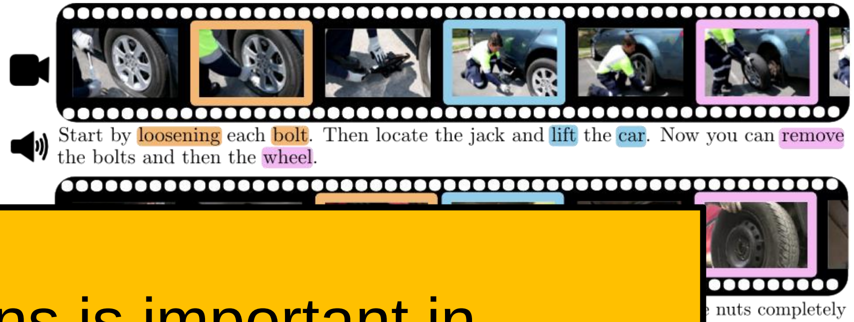
Ego-centric

Categories of Action Recognition Data

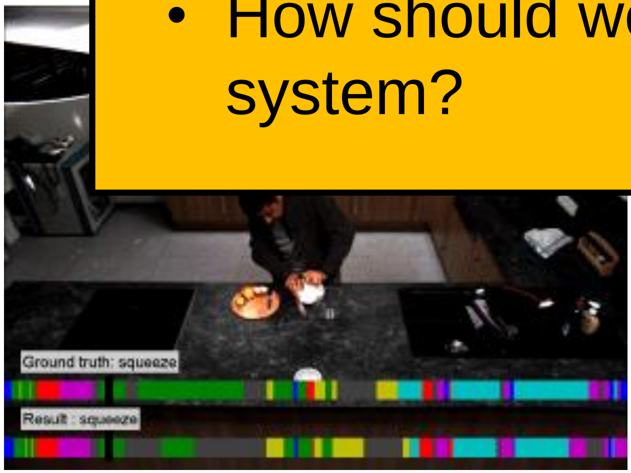
Sports

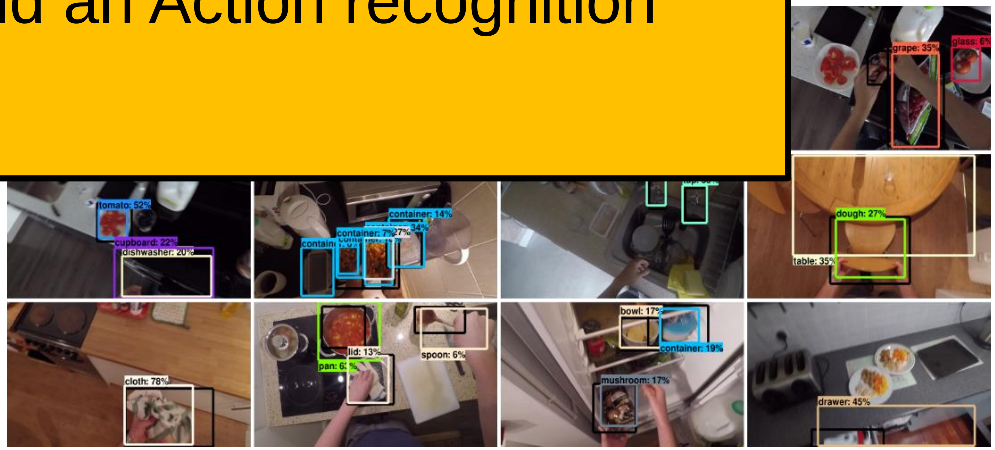


Instruction videos



- Analysis Human Actions is important in Real-life, there are plenty of applications.
- How should we build an Action recognition system?





An example

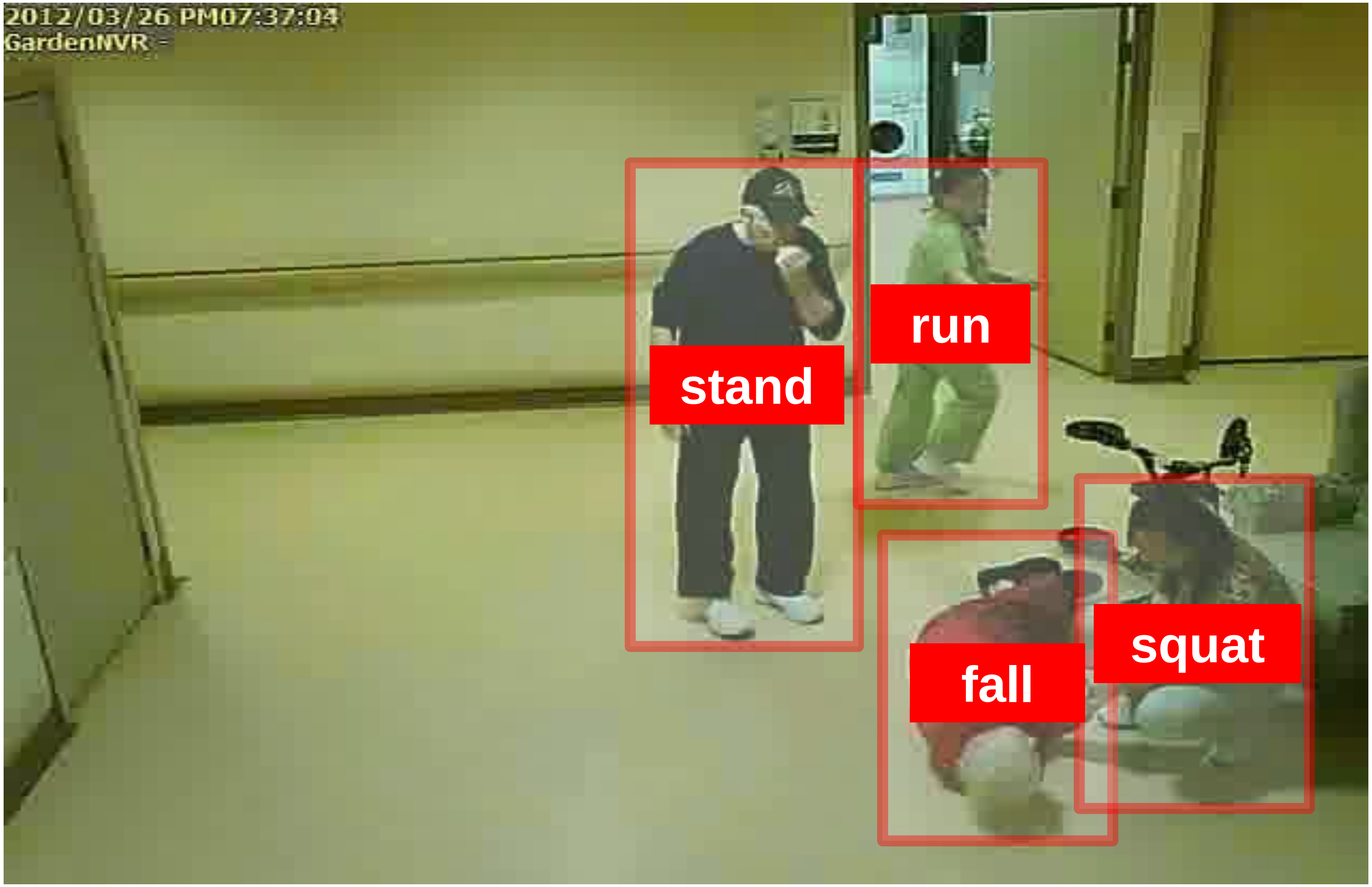
- What does action recognition involve?

An example

- Object Detection: Are they Human?

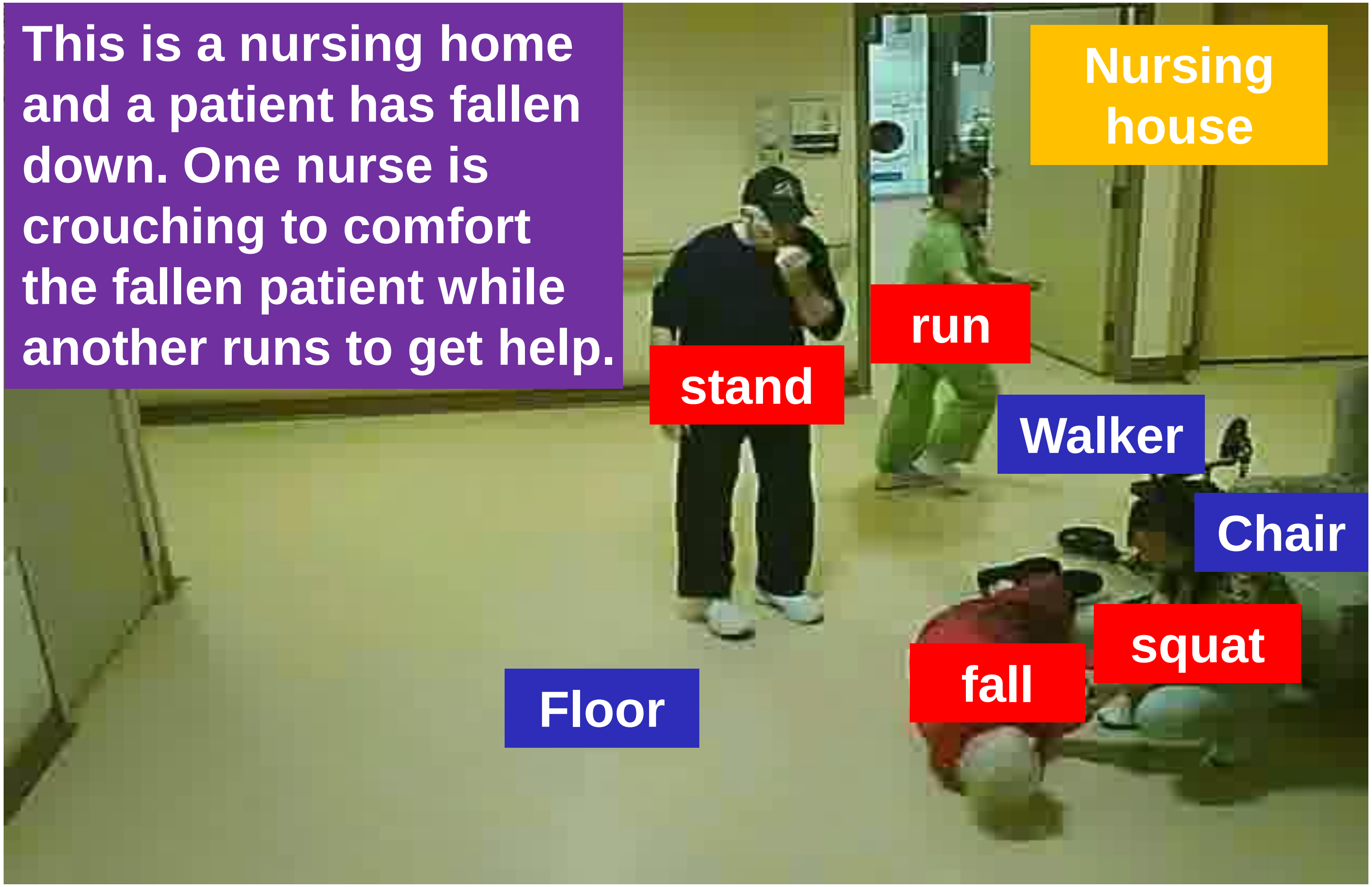
An example

- Action Recognition: What are they doing?



An example

- Full semantic understanding

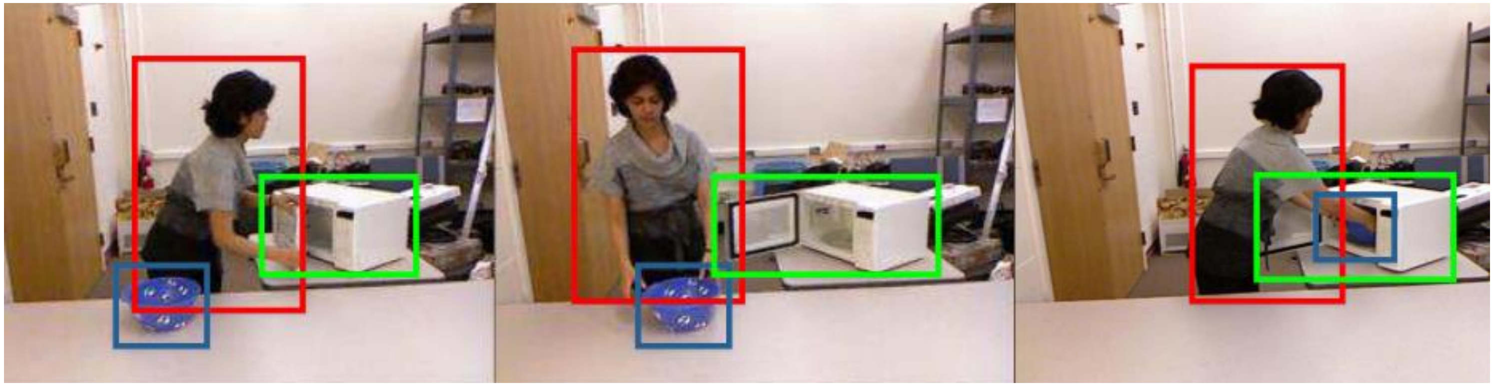


Action Recognition

- Classification of Videos into Pre-defined Action Categories

Complexity of Structure

- Different levels of structure complexity (temporal/spatial)



Opening ⟶ Reaching ⟶ Placing

Use Microwave

Complexity of structure in human actions

Semantic levels of human actions

There are levels of actions

- The ultimate goal is to make computers recognize all of them reliably.

“Activity” can be used for complex action

gestures

actions

human-object interactions

interactions

group activities

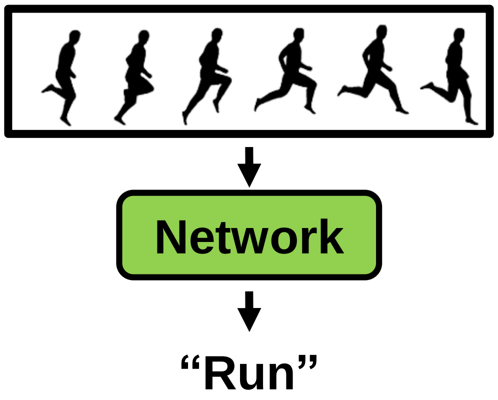
Levels of human actions

Action Recognition

A video classification task

Input: A clipped video (a sequence of frames)

Output: An action label



Actions of Daily Living (ADL)

- Actions of our boring everyday life: getting up, getting dressed, putting groceries in fridge, cutting vegetables and so on.



Challenges

- Subtle motion
- High intra-class variance
- Low inter-class variance

Subtle motion

Typing a keyboard **Reading**

- Same background
- Actions with subtle motion

High intra-class variation

Drinking

Drinking

- Same background
- High intra-class variation

Low inter-class variation

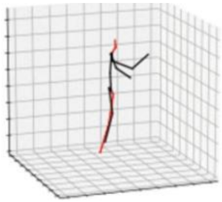
Wearing shoes | **Take off shoes**

- Same background
- Actions with similar appearance

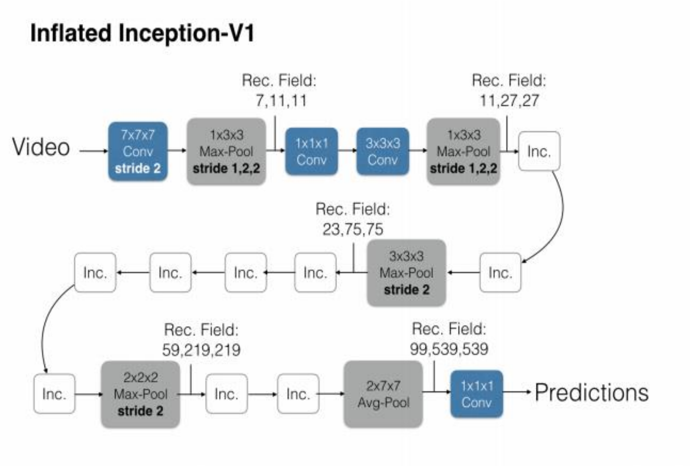
How to tackle these challenges?

- Different modalities





- Powerful Network to model the “time”



Section 2

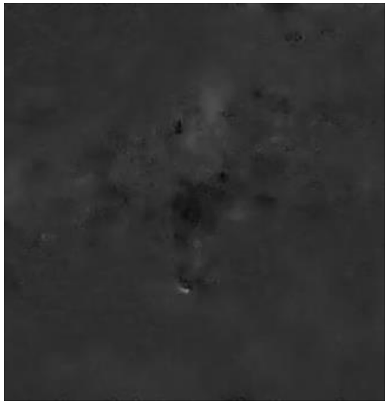
Modalities

Modalities

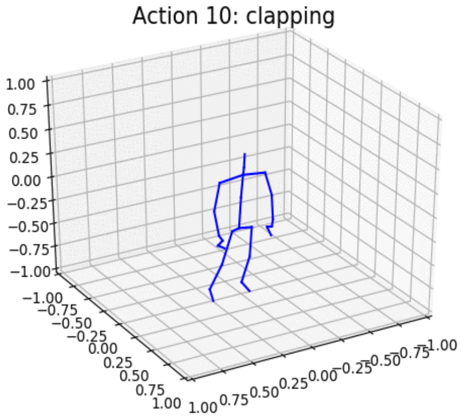
- Different input modalities
- Other: Audio, Depth image…

Clapping

RGB



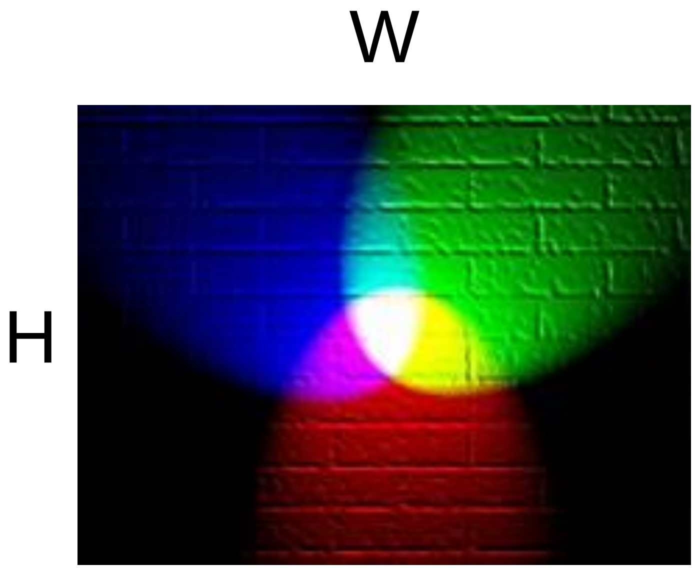
Optical flow

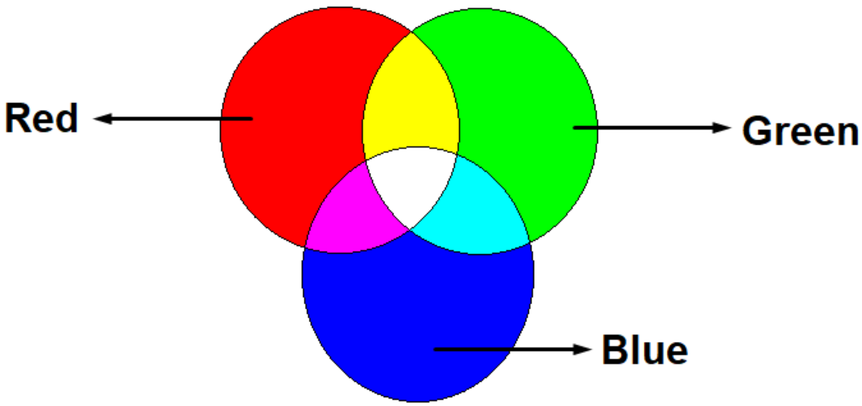


3D poses

RGB

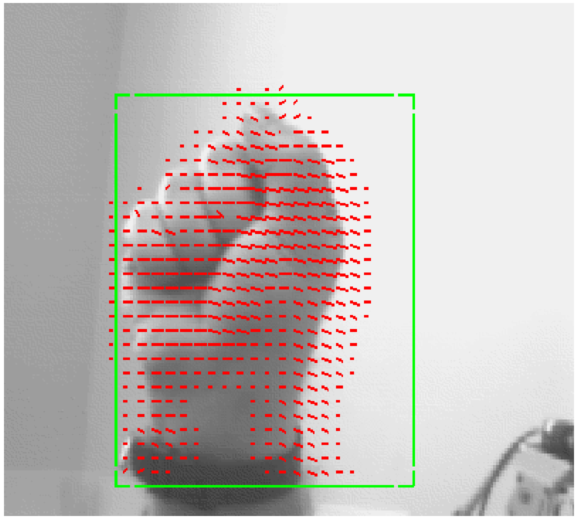
- Tensor: $[H \times W \times 3] \times T$





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).

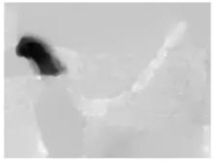
Optical flow

- Speed info

  Tensor: [H × W × 2] × T

RGB

X



Y

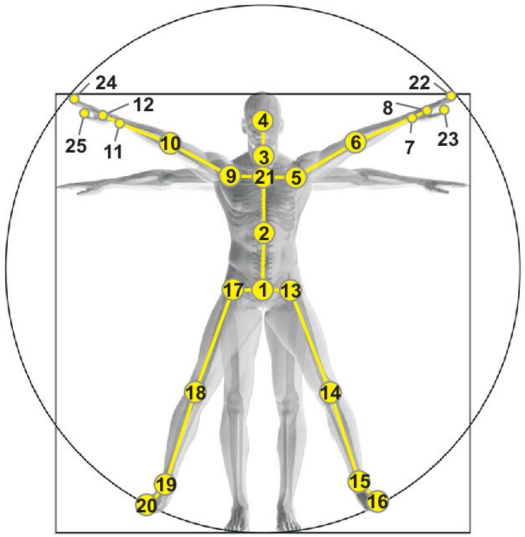
- Channel is 2D Axes
  - 1st (X image: [h,w,0]): Left, right
  - 2nd (Y image: [h,w,1]): Up, down
  - X and Y are Grey images

- Acquisition
  - Flow camera (Unmanned aerial vehicle)
  - Flow estimation algo (TVF1, FlowNet…)

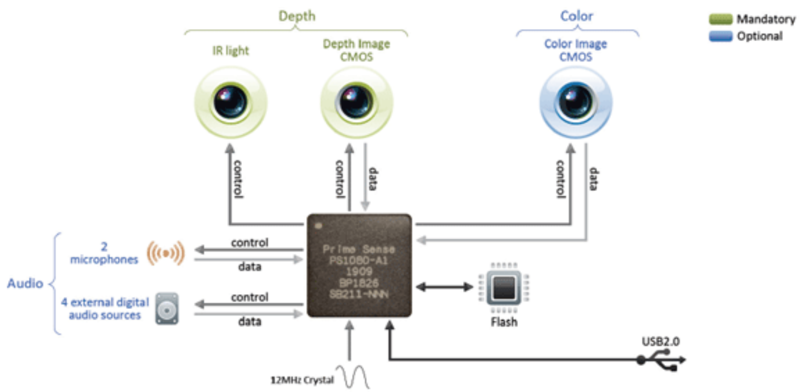
3D Poses/Skeletons

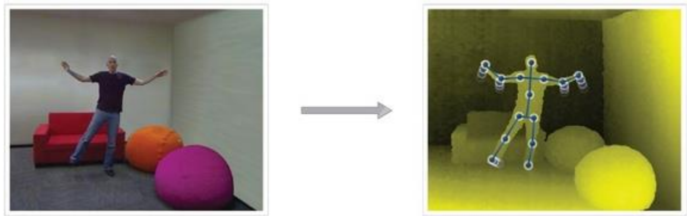
- Location info
- 3D Coordinates of N key joints on Human body

  Tensor: [N × (x, y, z)] × T



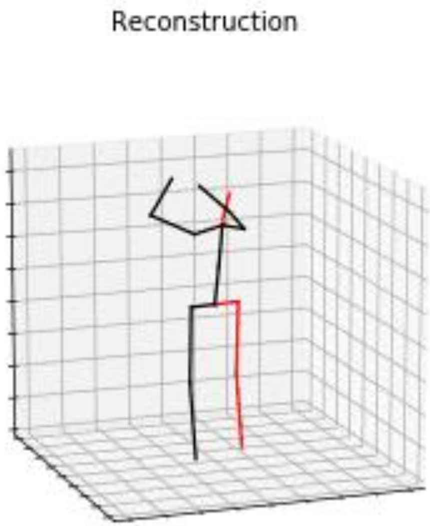
- Acquisition
  - Kinect camera (IR enhanced)
  - Pose estimation algorithm (From RGB images)





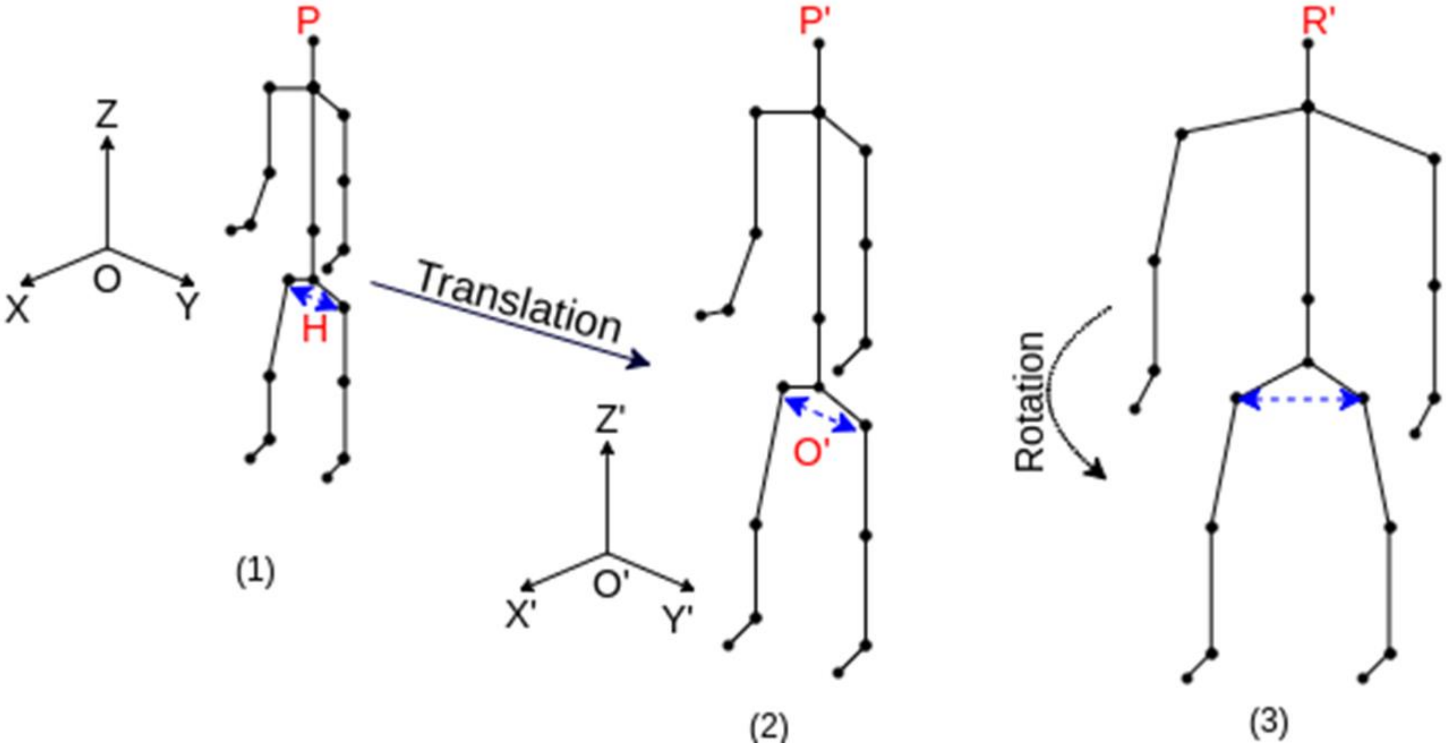
3D Poses

- Pose estimation from RGB (LCRNet+V3D)



3D Poses

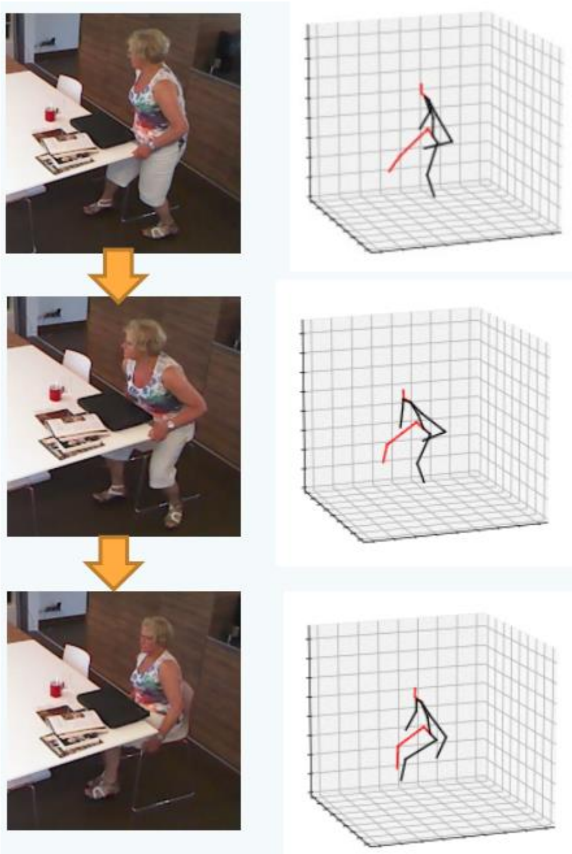
Preprocessing (optional)



- Camera-body translation
- Rotation of bones w.r.t. a line parallel to the hip
- Normalizing the bones

Why?

- Provide complementary information.



Sit down

Wear glasses

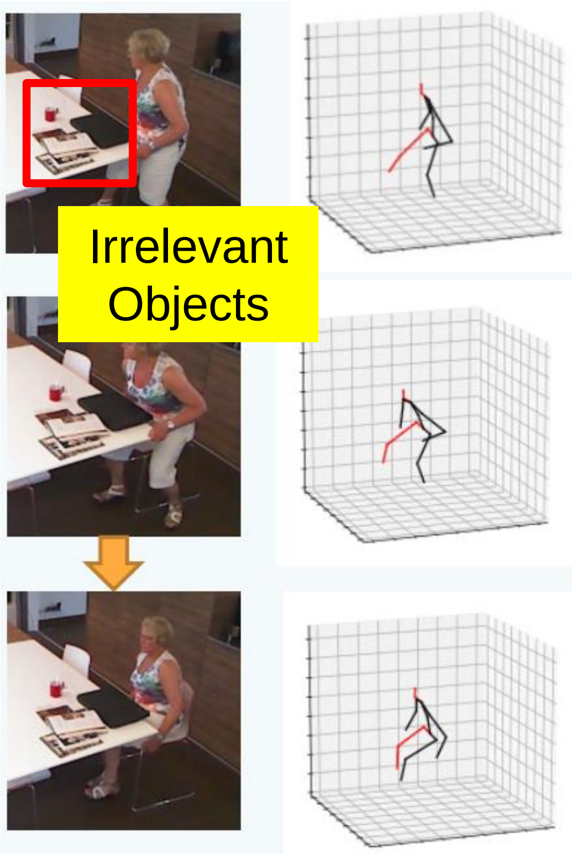
Take off glasses

3D poses

Optical flow

Why?

- Provide complementary information.



Sit down

3D poses

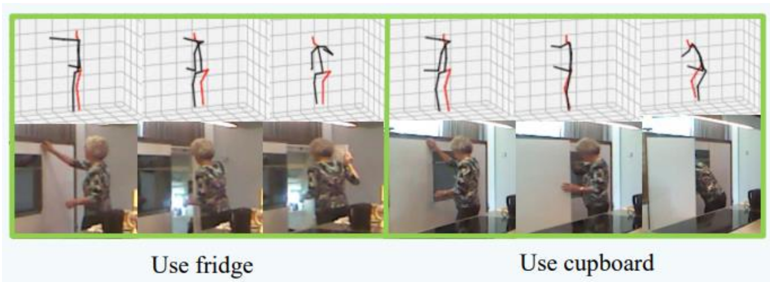
Take off glasses

Optical flow

Drawbacks

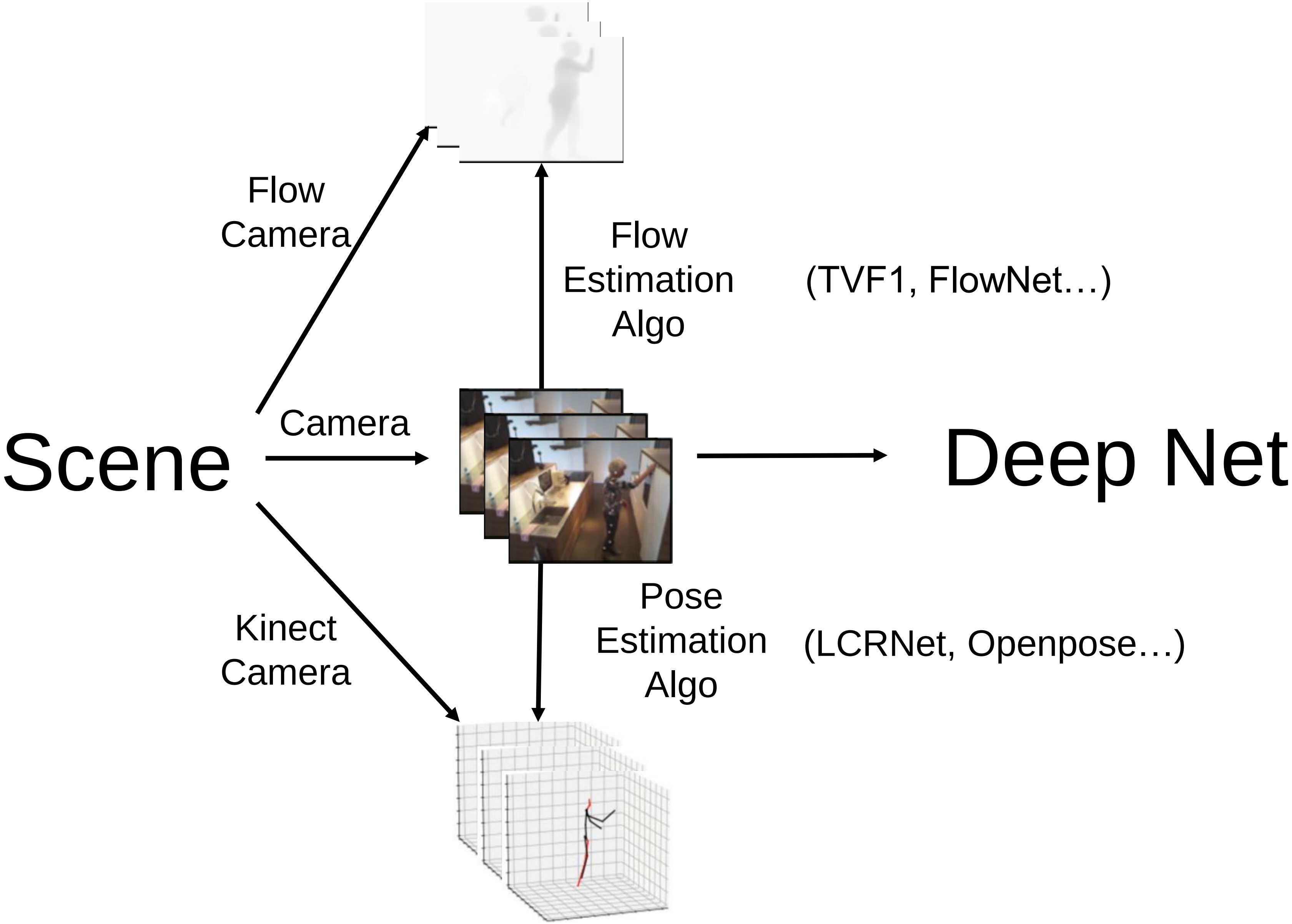
- Optical Flow
  - Time consuming in extracting Flow from RGB
  - Environment information is missing

- 3D Poses
  - Object Information is missing



- RGB
  - Contains the most information, but noisy!

Pipeline

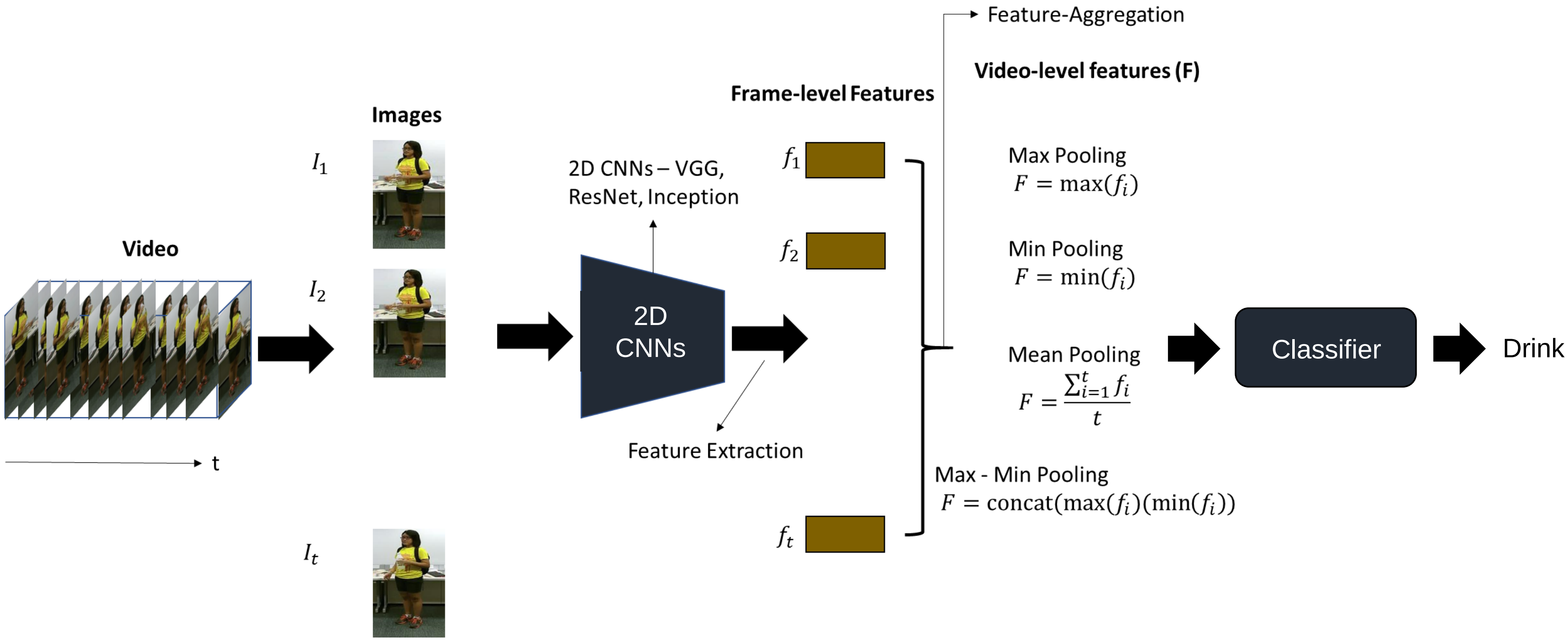

Scene
Flow
Camera
Camera
Kinect
Camera
Flow
Estimation
Algo
(TVF1, FlowNet…)
Deep Net
Pose
Estimation
Algo
(LCRNet, Openpose…)

Section 3

Deep Networks for Action Recognition

Video Classification

Recap…



Temporal modeling is important!

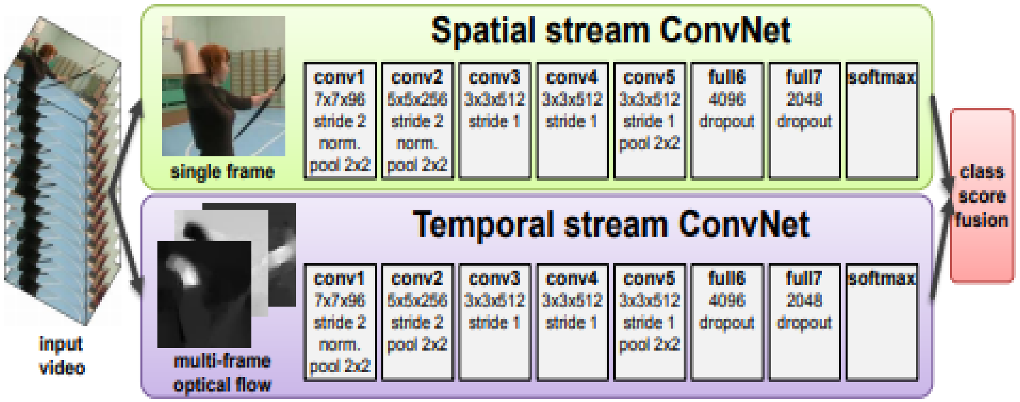
A still from '**Quo Vadis**' (1951). Where is this going? Are these actors about to kiss each other, or have they just done so?

Modeling temporal information is needed!

Two-stream Network [NIPS'14]

- **Using multiple modalities as input!**
- **RGB**: One image randomly sampled from the video. (<u>Spatial</u>: encodes object/appearance information)
- **Flow**: *2L* optical flow images from a video. (<u>Temporal</u>: encodes short-term motion)

$224 \times 224 \times 3$



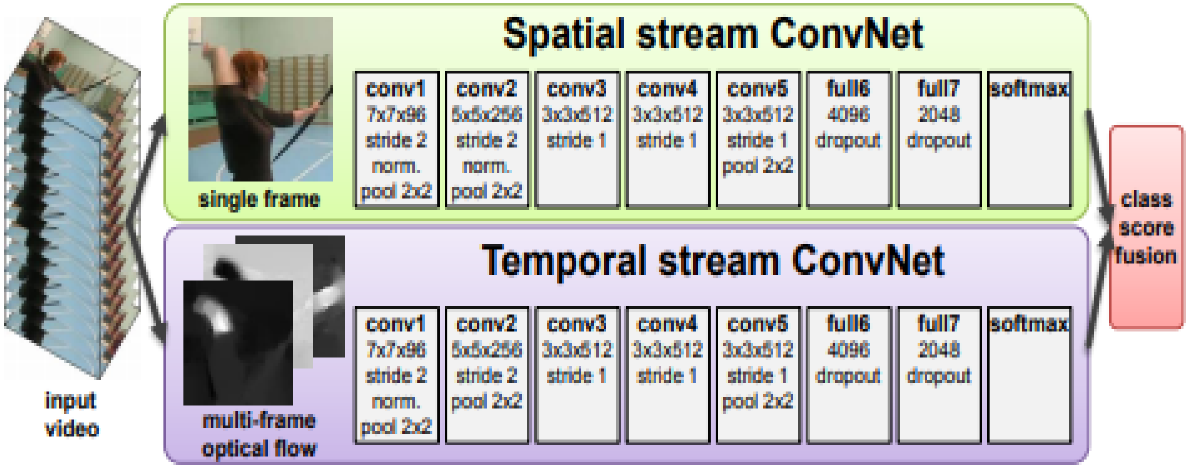
Just Fuse the Score

$224 \times 224 \times 2L$

Two-stream Network [NIPS'14]

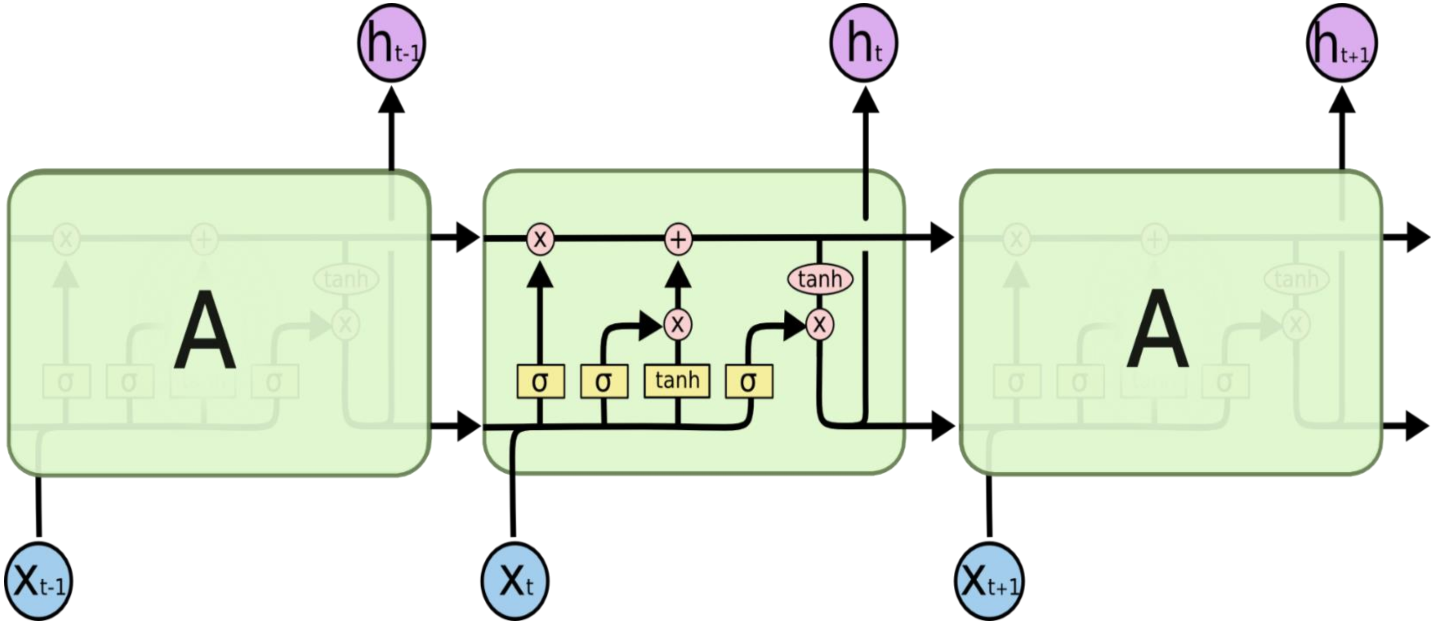
Drawbacks:

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



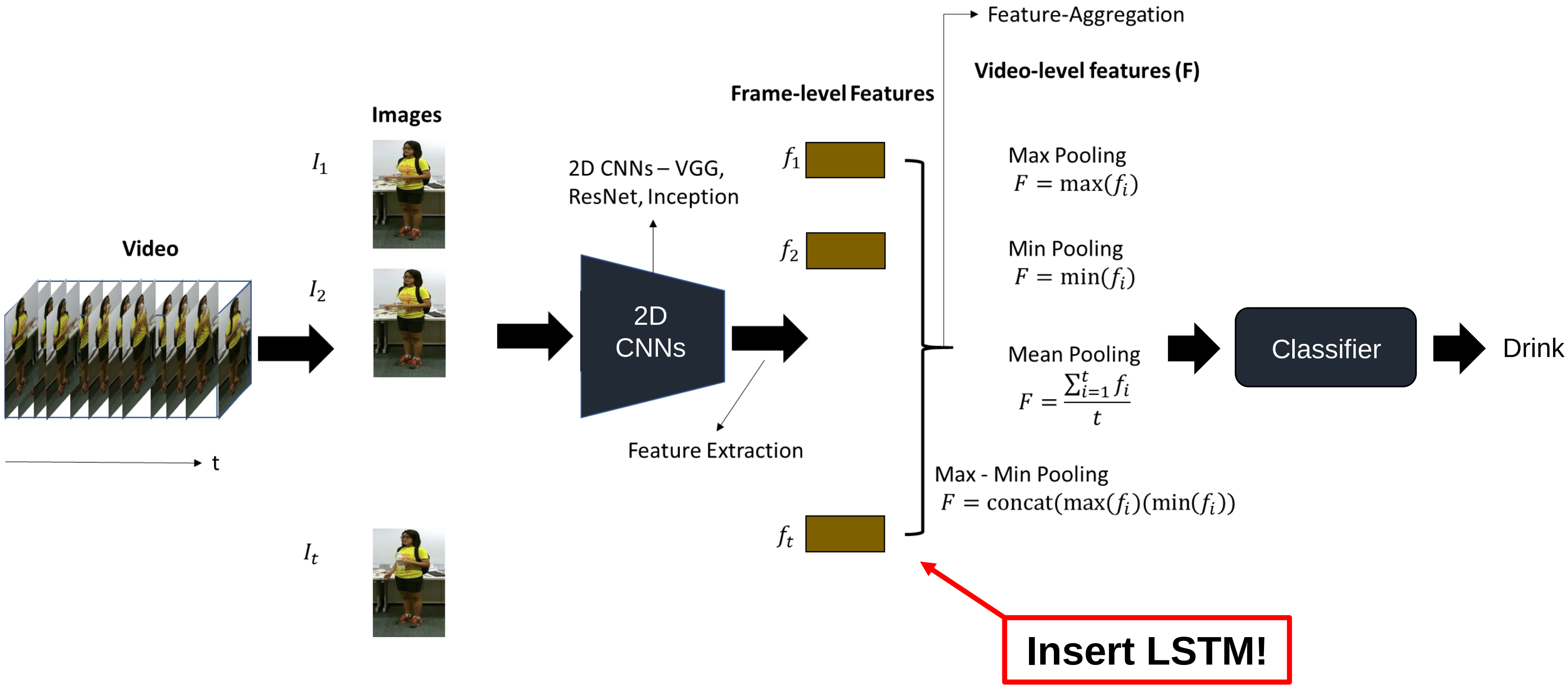
RNN (LSTM)

Recap …



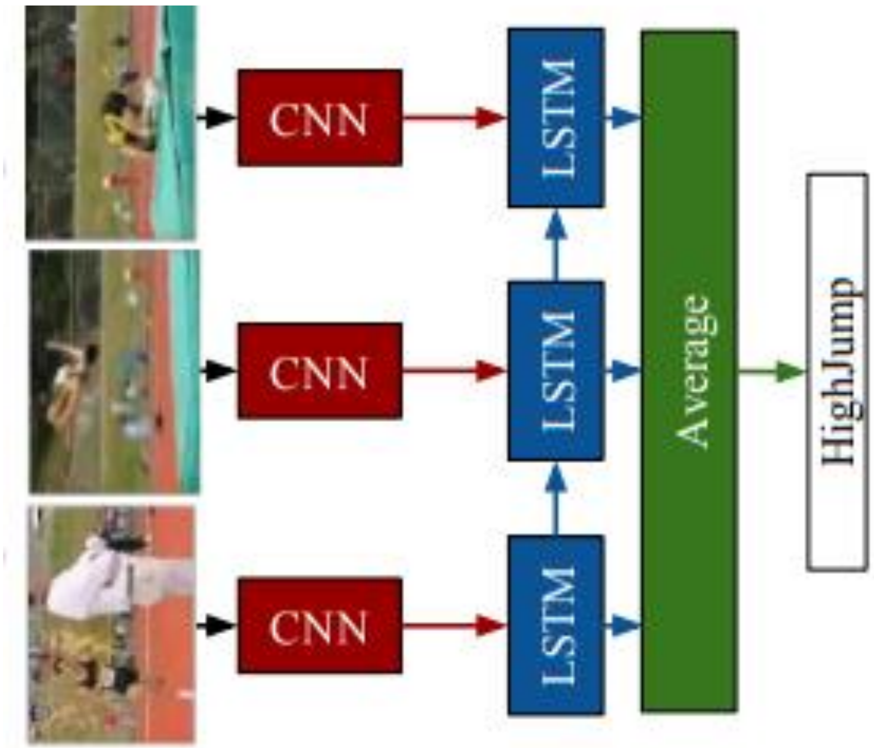
Video Classification

Recap…

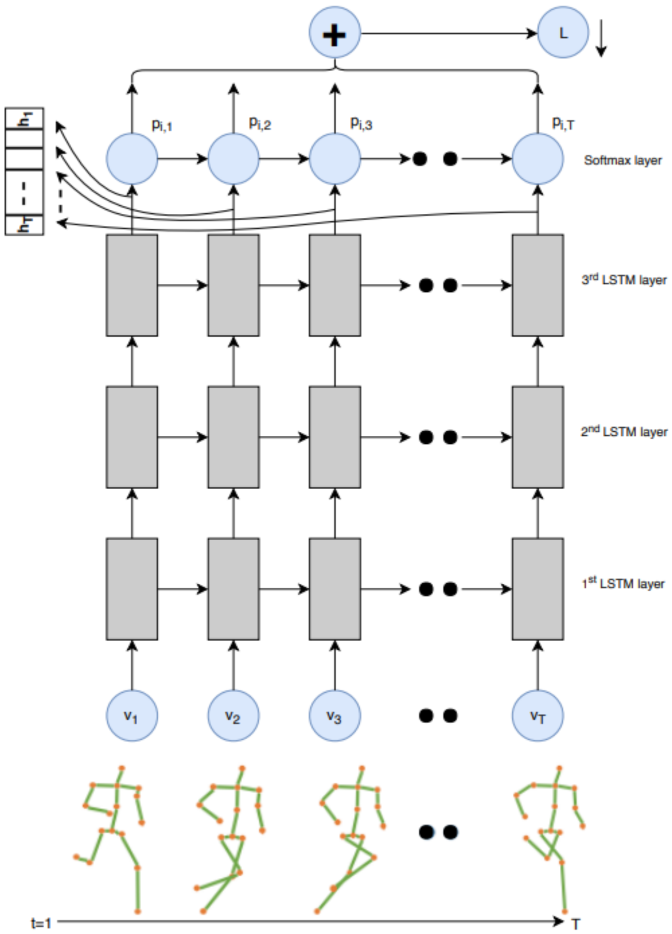


LRCN (2DCNN+RNN)

Sequence of images as input



Pose + LSTM



L
$p_{i,1}$
$p_{i,2}$
$p_{i,3}$
$p_{i,T}$
Softmax layer
3rd LSTM layer
2nd LSTM layer
1st LSTM layer
v_1
v_2
v_3
v_T
t=1
T

Drawback of RNN

- RNNs/LSTMs can only capture strong temporal evolution of the image level features.
- Vanishing gradient issue (Can not remember long term temporal information.)
- Not much efficient on small datasets (pre-training is not a good idea as they change the statistics learned by the gates).

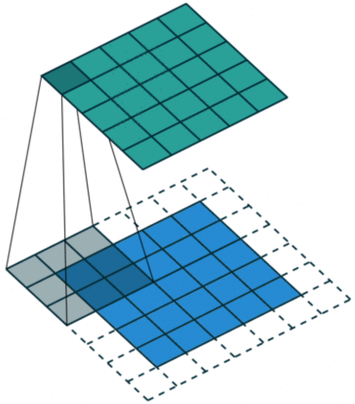
2D Convolution (XY)

Recap…

Input: $[H_{in}, W_{in},\ C]$

Output: $[H_{out},\ W_{out}$, #*Kernel*]

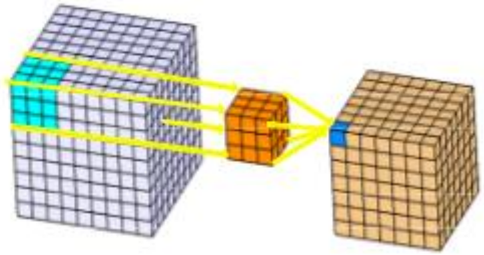
Kernel move in H,W direction

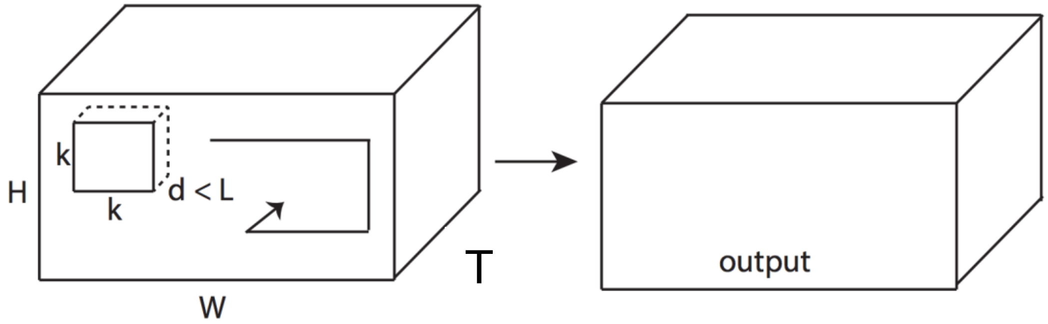


$$H_{out} = \frac{H_{in} + 2 \times padding - dilation \times (kernel\_size - 1) - 1}{stride} + 1$$

$$W_{out} = \frac{W_{in} + 2 \times padding - dilation \times (kernel\_size - 1) - 1}{stride} + 1$$

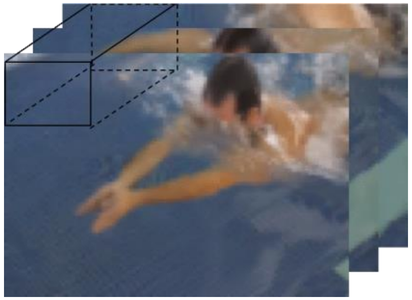
3D Convolution (XYT)





Input: $[H_{in}, W_{in}, T_{in}, C]$

Output: $[H_{out}, W_{out}, T_{out},$ *#Kernel*$]$

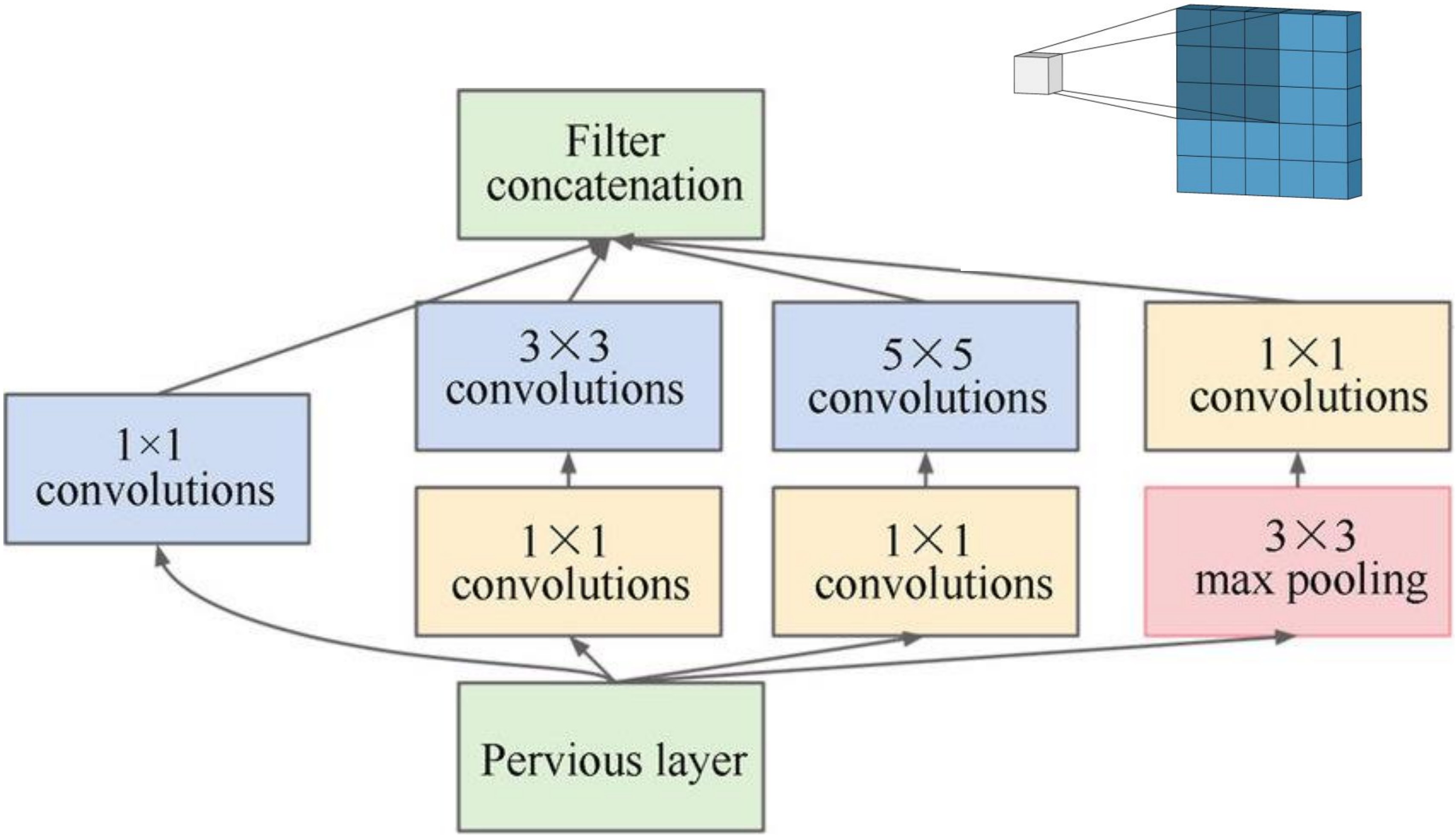


$$H_{out} = \frac{H_{in} + 2 \times padding - dilation \times (kernel\_size - 1) - 1}{stride} + 1$$

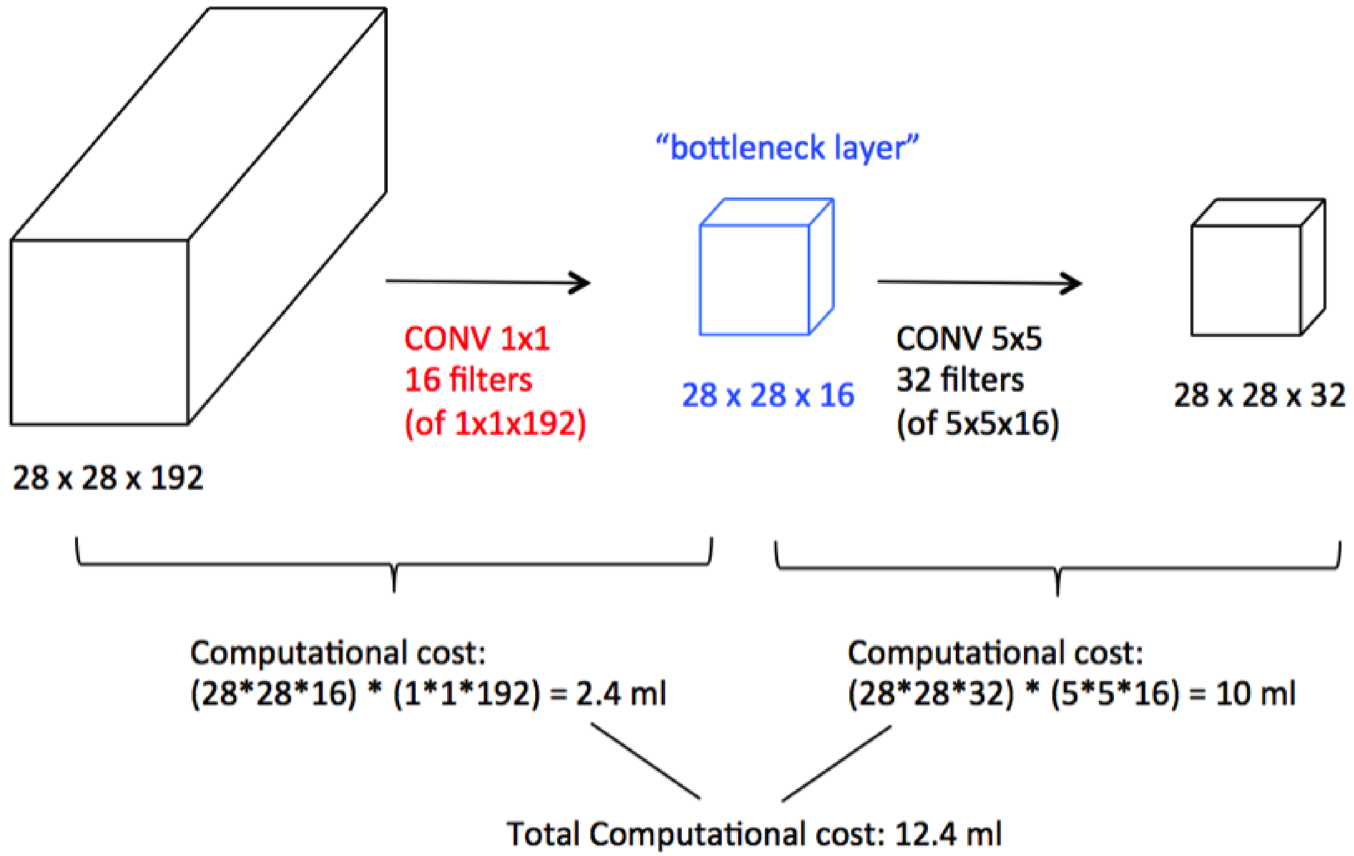
$$W_{out} = \frac{W_{in} + 2 \times padding - dilation \times (kernel\_size - 1) - 1}{stride} + 1$$

$$\boxed{T_{out} = \frac{T_{in} + 2 \times padding - dilation \times (kernel\_size - 1) - 1}{stride} + 1}$$

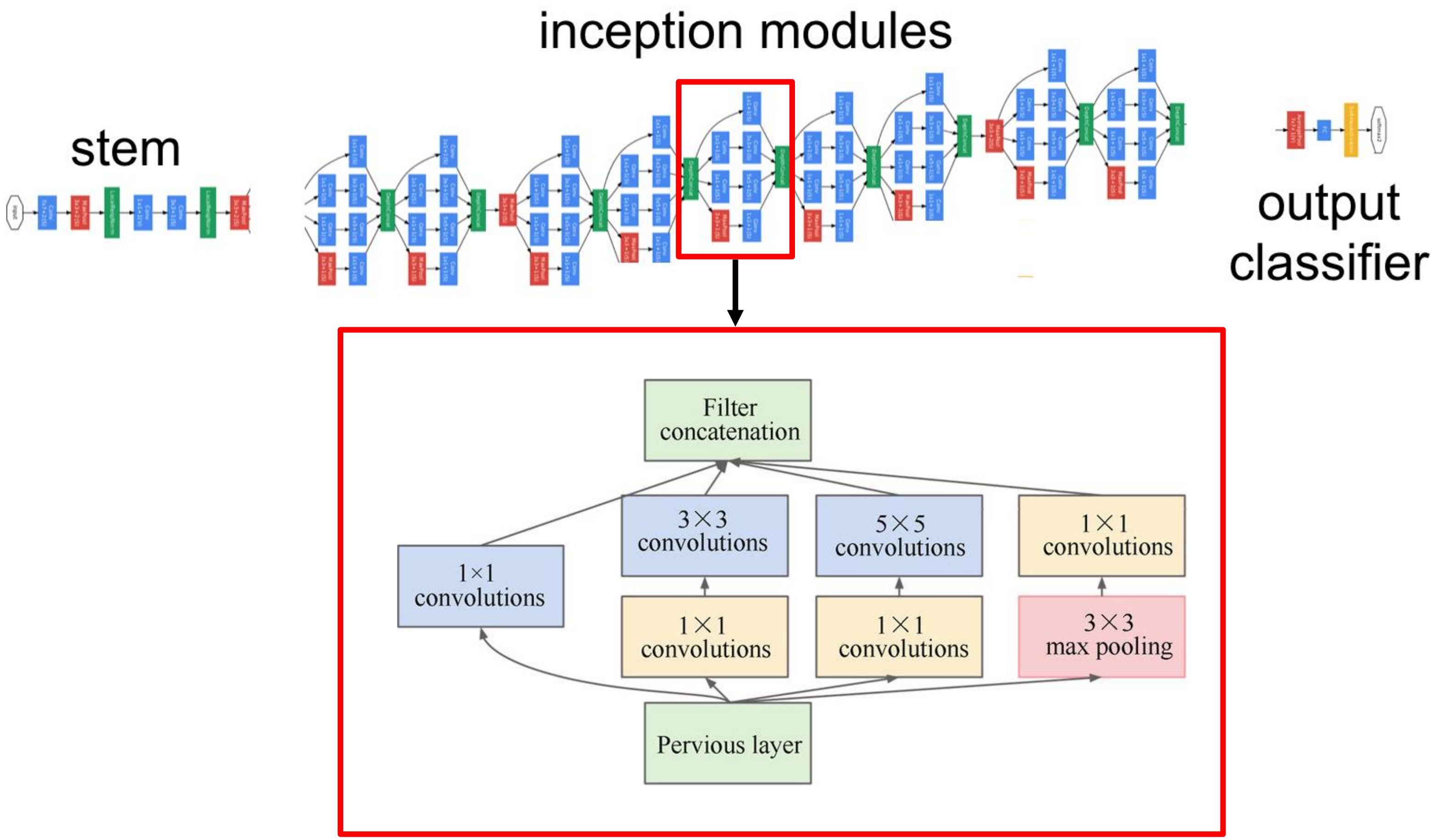
Inception Module



Bottleneck layer

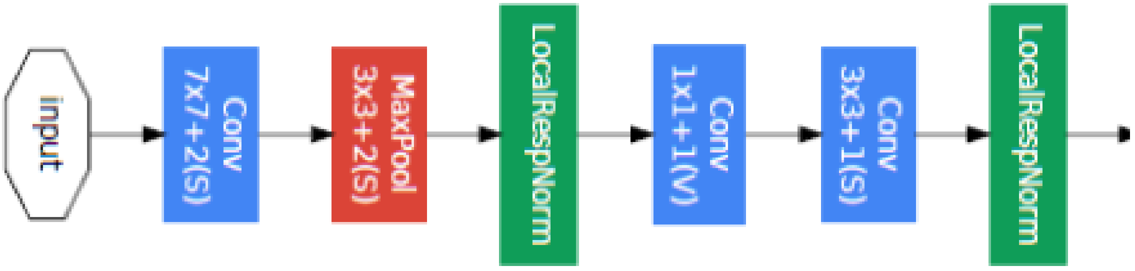


GoogleNet


inception modules
stem
output
classifier
Filter
concatenation
3×3
convolutions
5×5
convolutions
1×1
convolutions
1×1
convolutions
1×1
convolutions
1×1
convolutions
3×3
max pooling
Pervious layer

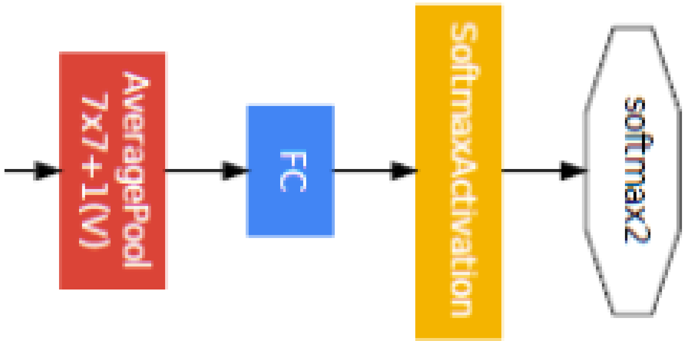
Stem

Stem has some preliminary convolutions.



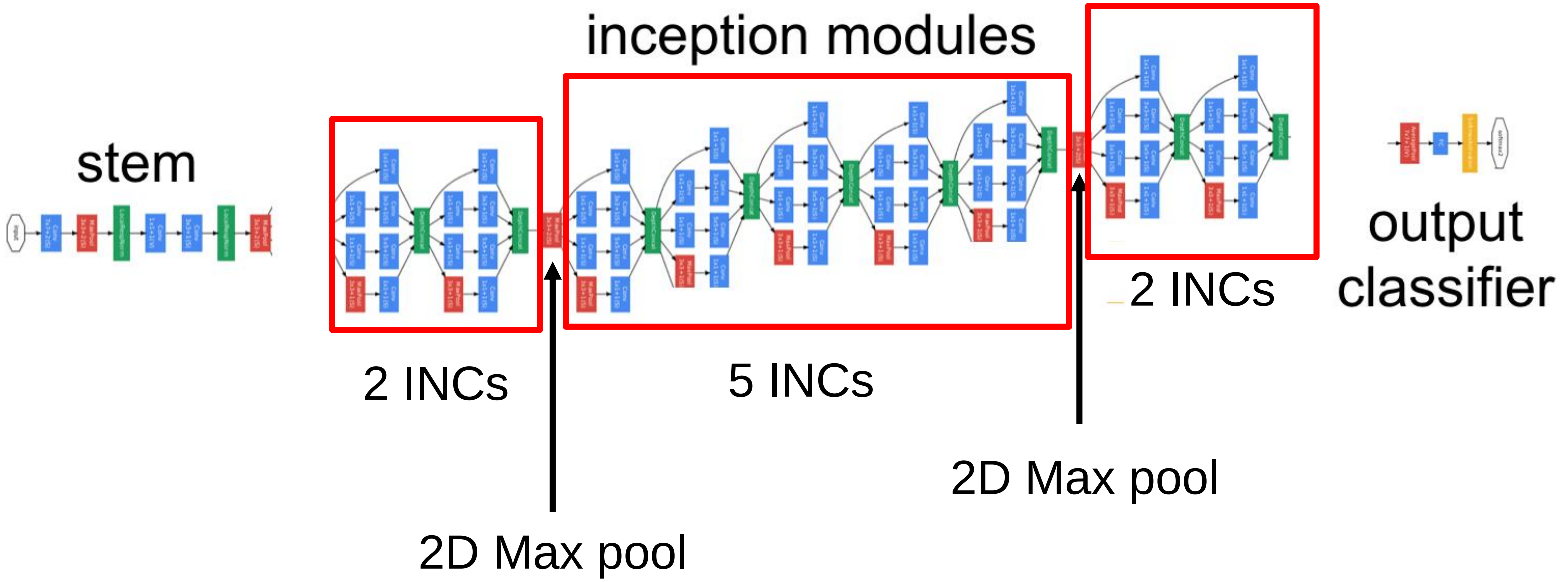
Classifier

- Project the channel size into the #classes
- Softmax Activation



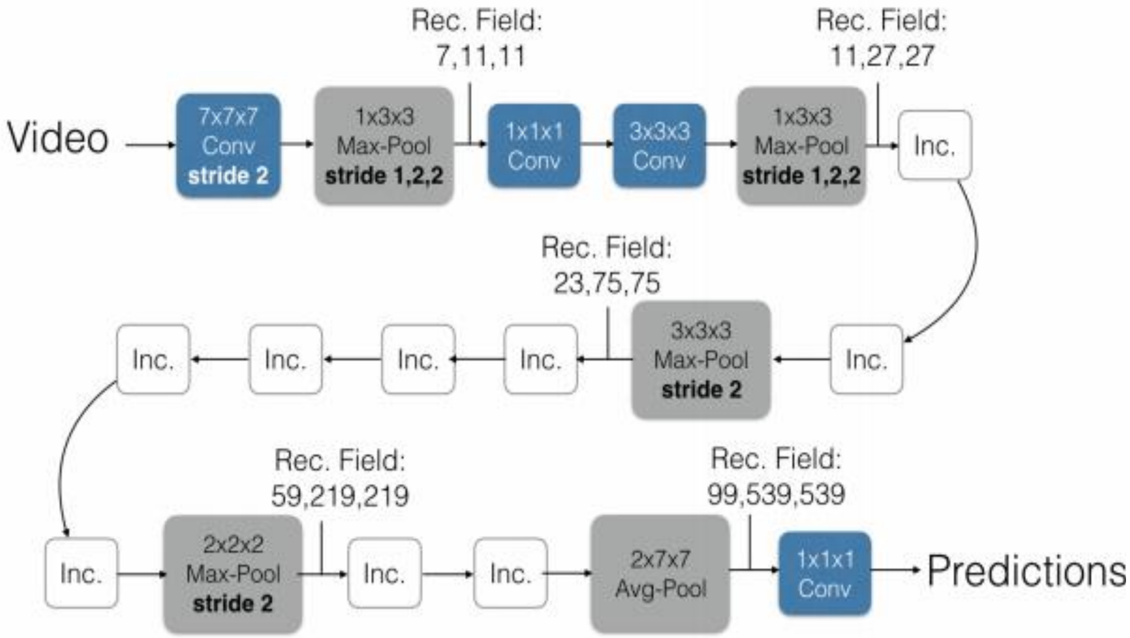
Inception Module (GoogleNet)

Recep…

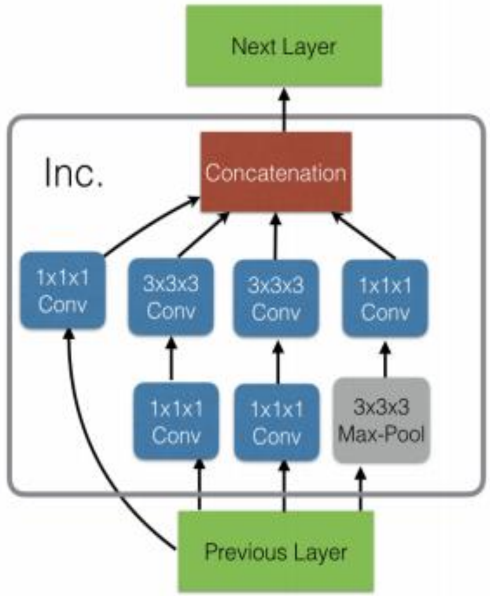


I3D Network [CVPR'17]

I3D

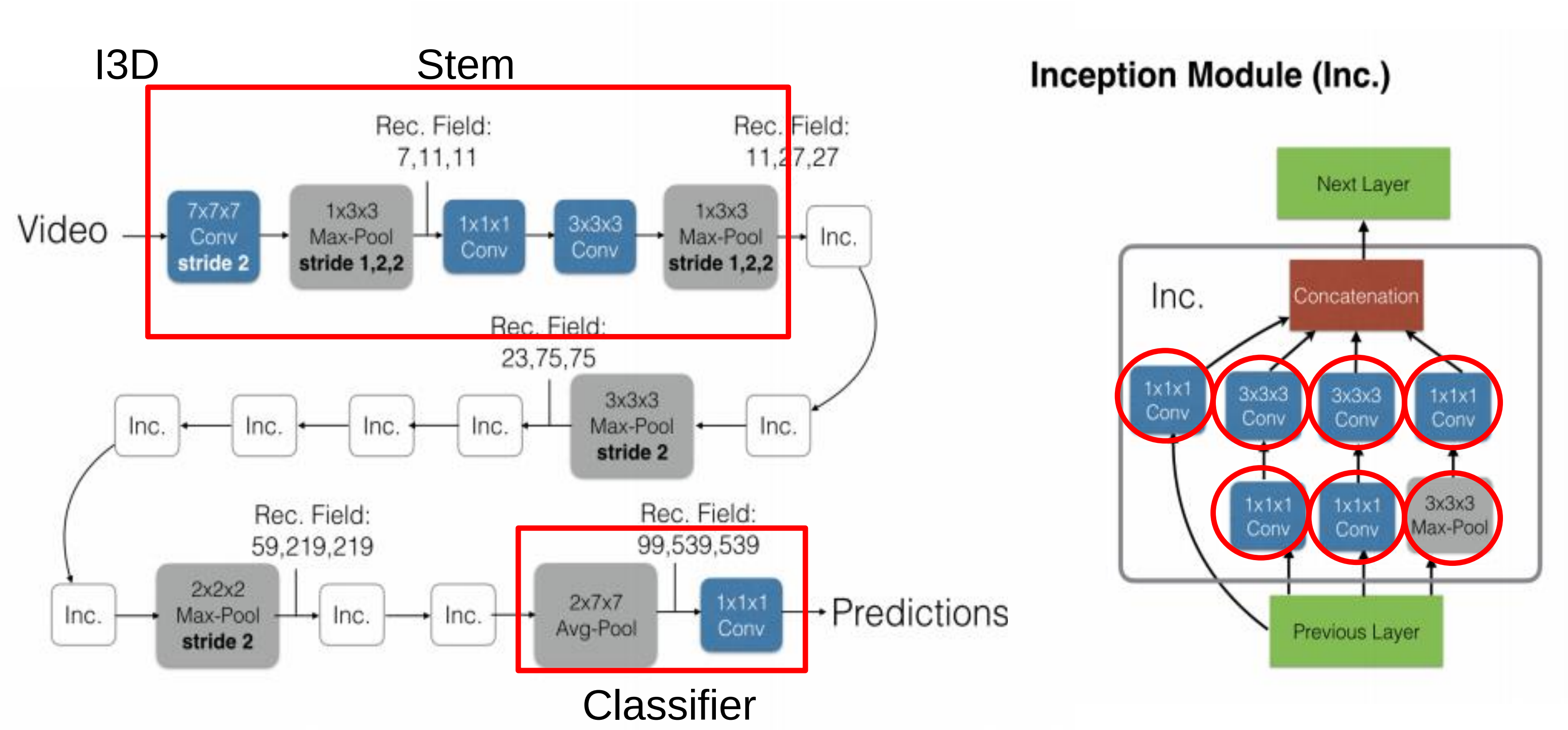


Inception Module (Inc.)



Same structure as GoogleNet!

I3D Network [CVPR'17]

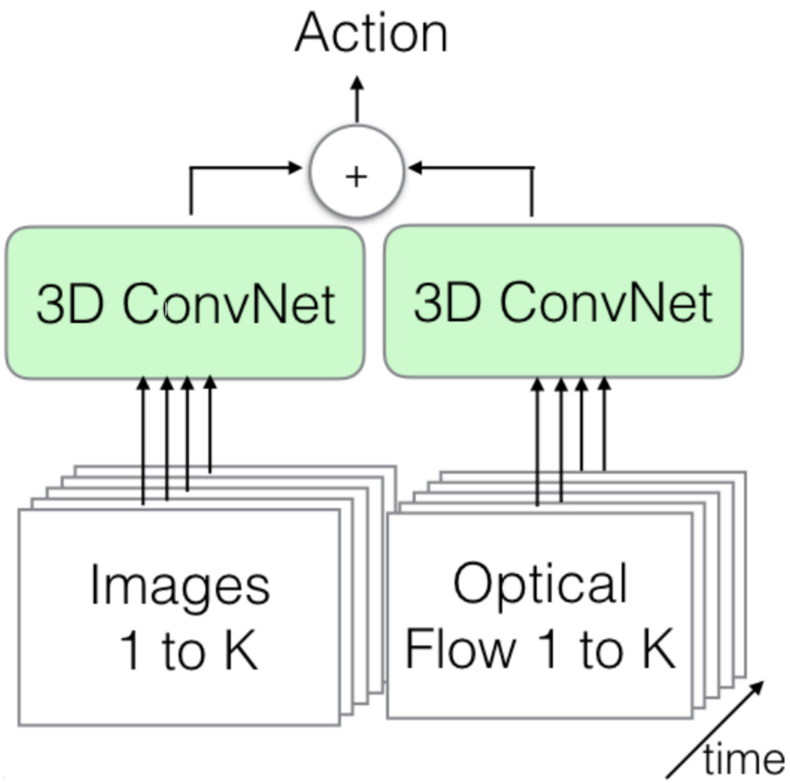


Two-stream structure

Inputs

RGB Stream: $224 \times 224 \times T \times 3$

Flow Stream: $224 \times 224 \times T \times 2$

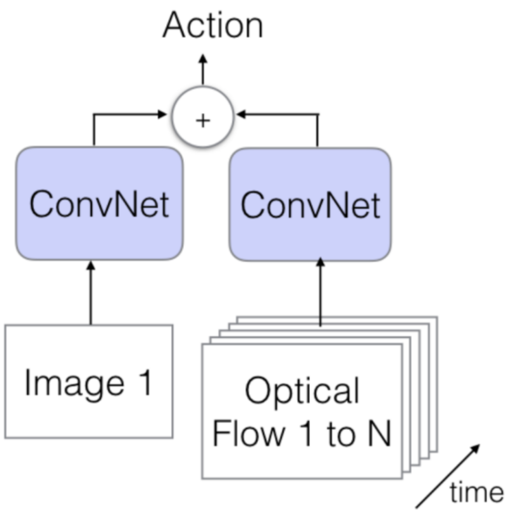


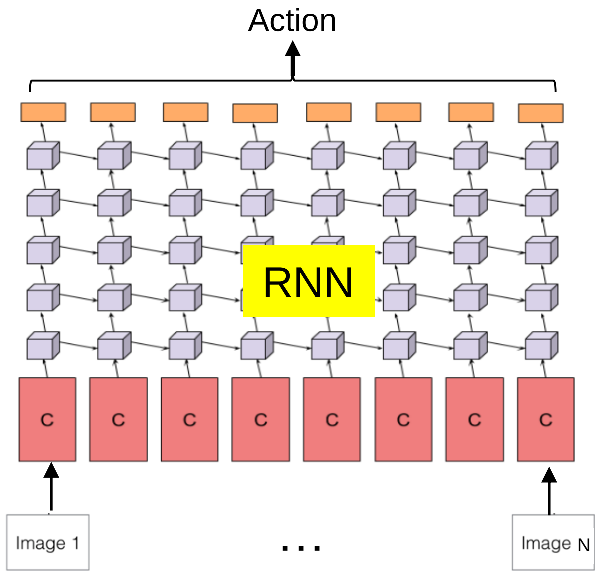
Limitation of 3D CNN

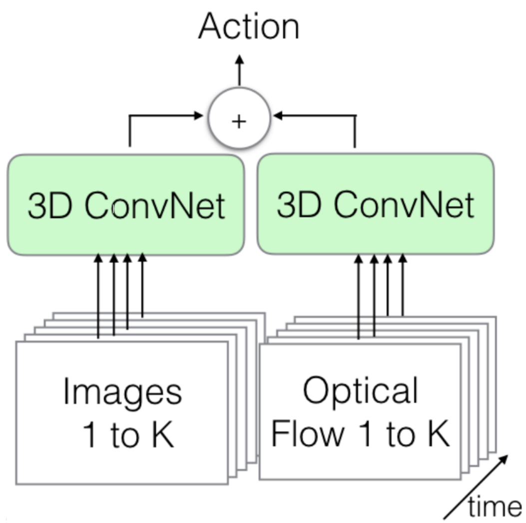
- Rigid Spatio-temporal kernels limiting them to capture subtle motion
- No specific operations to help disambiguate similarity in actions.
- 3D (XYT) CNNs are not view-adaptive.
- Large computational cost.

Summary

Input: A clipped video, Output: A class label







- **Two-stream CNNs**

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

  [Karen and Zisserman, 2014]

- **Sequential models RNNs**

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

  [J. Ng et al., 2015]

- **3-D XYT CNNs**

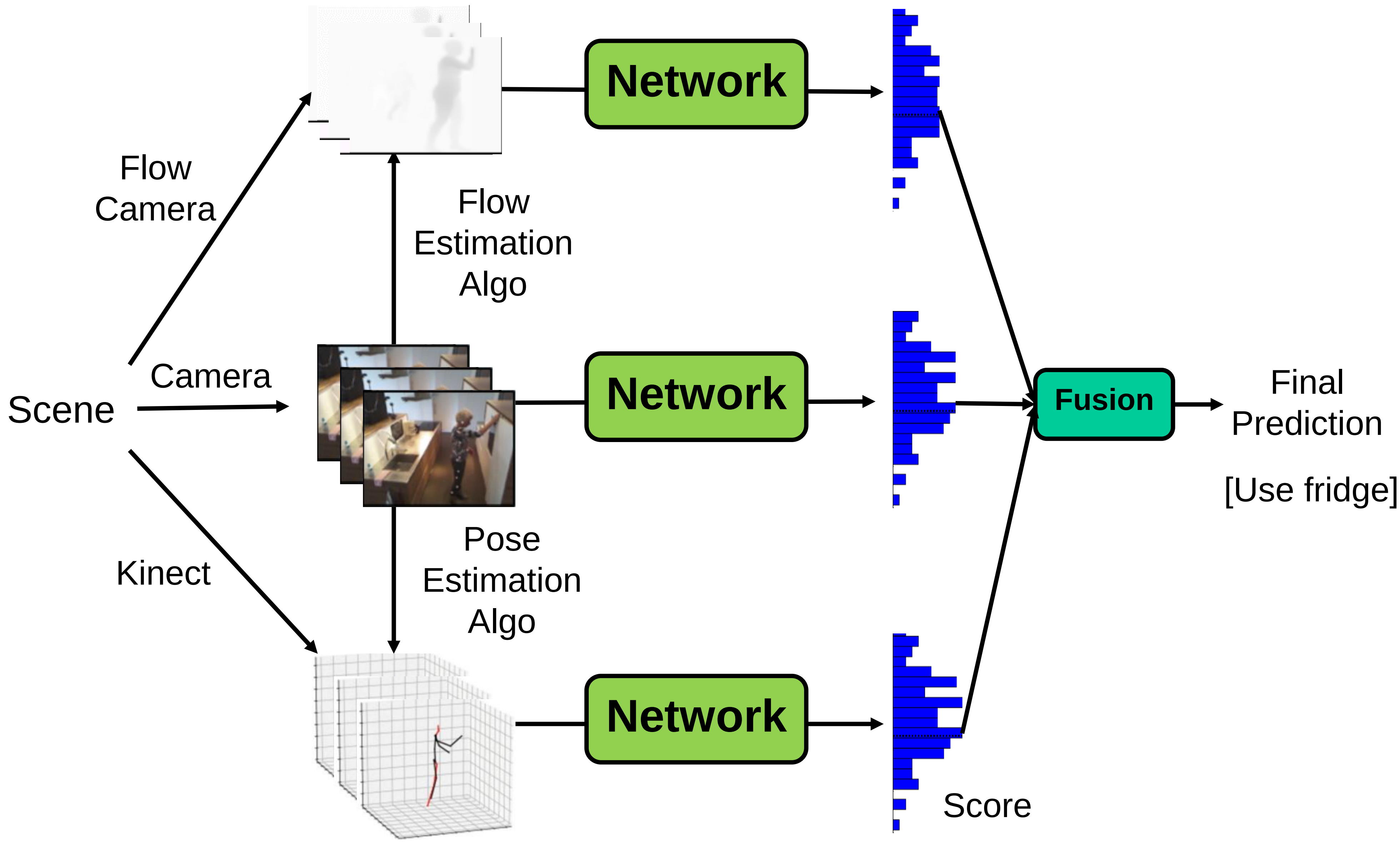
  I3D, C3D…

  10-99 frames

  (**RGB** + **Flow**)

  [Tran et al., 2015]

Total Pipeline



Travaux Pratiques

Practice

Two-Stream Network

- Generate Flow from RGB
- Evaluate a video using Two-stream Network

- https://colab.research.google.com/drive/1C8gPsD_sJlxNj1v4Z5kQDifhkTeTgEmY?usp=sharing

Practice

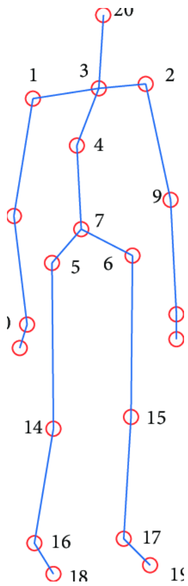
Evaluate a video of UCF-101 using I3D

- https://colab.research.google.com/drive/1M5Hj2tqBL0L2sDDzOPM_U0OzCiotqv29?usp=sharing

Practice (optional)

Train a 3-layer LSTM, inputs are 3D Poses.

- https://colab.research.google.com/drive/1AUVjpLg8_8E0l-up6CiB-4pwbE_BSfkf?usp=sharing



Reference

- UCF computer vision video Lectures 2012 (Instructor: Mubarak Shah)
- CVPR Tutorial, Human Activity Recognition (M. Ryoo, I. Laptev, J. Mori)

Thanks!

E-mail: rui.dai@inria.fr