Some Announcements

- Homework 4 is announced
  - due : Dec. 10th, 11:59 pm **(night!!!)**
  - The second problem is open-ended. Please start as soon as possible.
- Will send out the practice final exam paper around this Sunday
- Two review lectures after thanksgiving. Please prepare questions that you would ask.

CSE 152: Computer Vision

Hao Su

Lecture 17: Deep Video Understanding

Some of the content are from CMU Vision Class

Agenda

- Motivation
- Some terminologies
- Two-stream representation

Agenda

- **Motivation**
- Some terminologies
- Two-stream representation

Goal of Video Understanding

- Given an input video, obtain an understanding
- Involves:
  - Objects
  - Humans
  - Actions/events

Why do we perform actions?

Agents perform actions to achieve goals

Pour = Transfer milk from bottle to glass

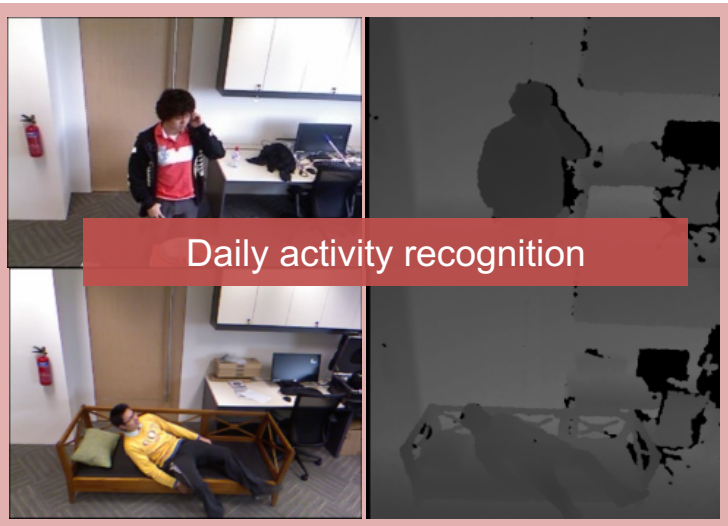
However

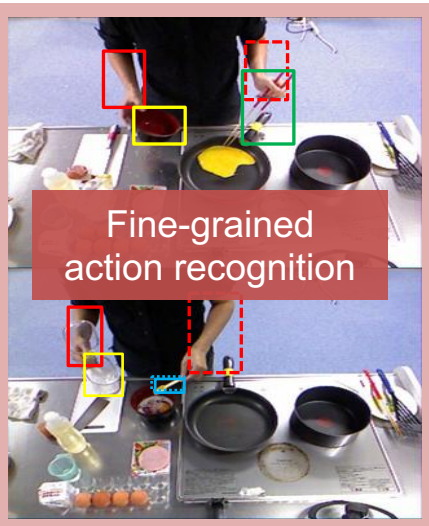
- Most of the state-of-the-art or current literature focuses on modeling actions based on appearances.
- In fact, state-of-the-art comes from just average pooling the static image classification!!

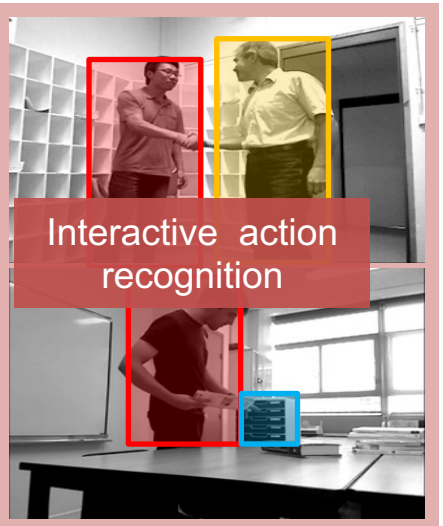
Applications of Action Recognition

- Surveillance footage
- User-interfaces
- Automatic video organization / tagging
- Search-by-video?

Example Applications



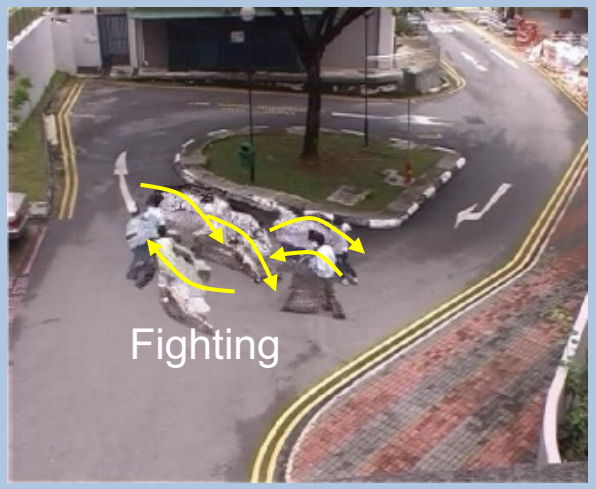


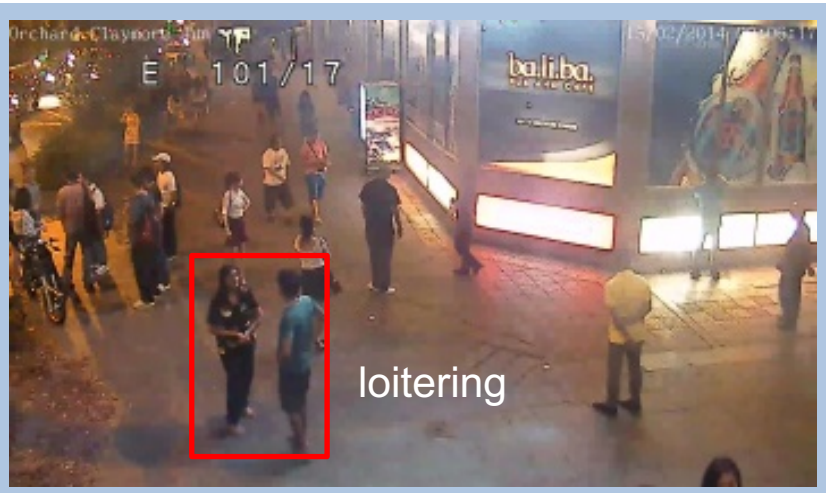


Intelligent Assisted Living and Home Monitoring

credit: Bingbing Ni

Example Applications





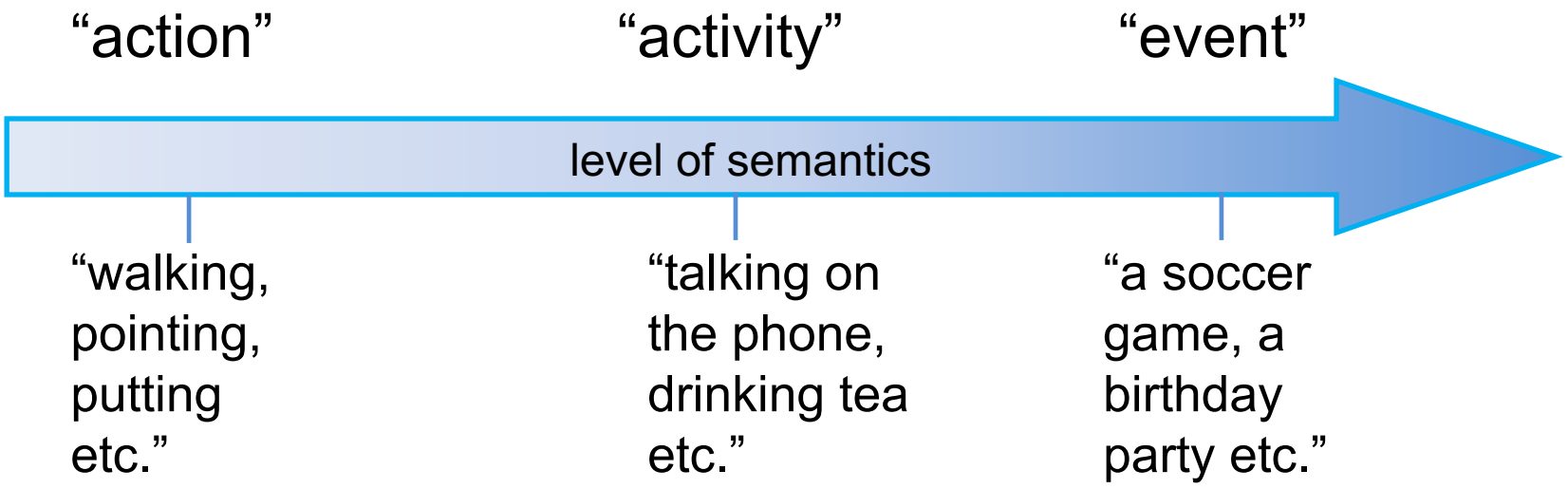
Crowd Behavior/Event Analysis

credit: Bingbing Ni

Demo

Formulation of Action Recognition

- Input: video/image
- Output: the “action label”



Human activities

- Categorized based on their complexity
  - Hierarchy
  - # of participants

Gestures (atomic):

Single body-part movements

Human activities

- Categorized based on their complexity
  - Hierarchy
  - # of participants

Interactions:

Human-human/
Human-object
interactions

Why Action Recognition Is Challenging?

- Different scales (size)
  - People may appear at different scales in different videos yet perform the same action.

- Movement of the camera
  - The camera may be a handheld camera, and the person holding it can cause it to shake.
  - Camera may be mounted on something that moves.


Stabilisation: Off

Why Action Recognition Is Challenging?

- Occlusions
  - Action may not be fully visible

Figure from Ke et al.

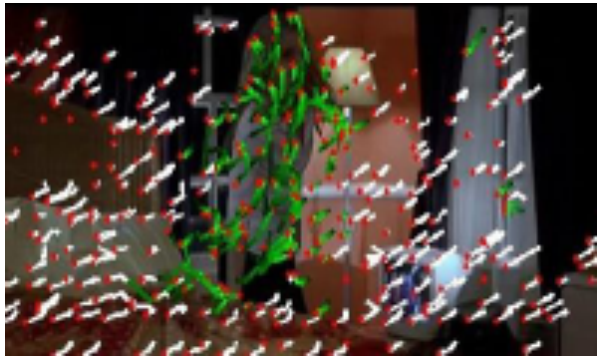
Why Action Recognition Is Challenging?

- Background “clutter”
  - Other objects/humans present in the video frame.
- Human variation
  - Humans are of different sizes/shapes
- Action variation
  - Different people perform different actions in different ways.
- Etc…

Need Good Features for Action Representation

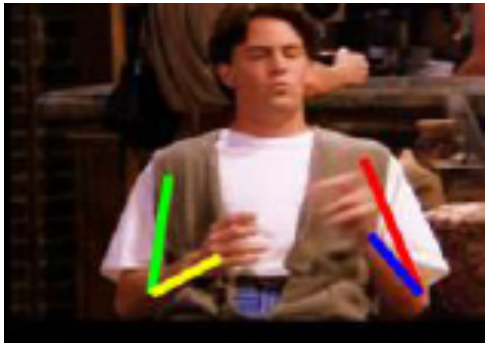
Example features

“space-time interest points”



“dense trajectories”

“motion history images”

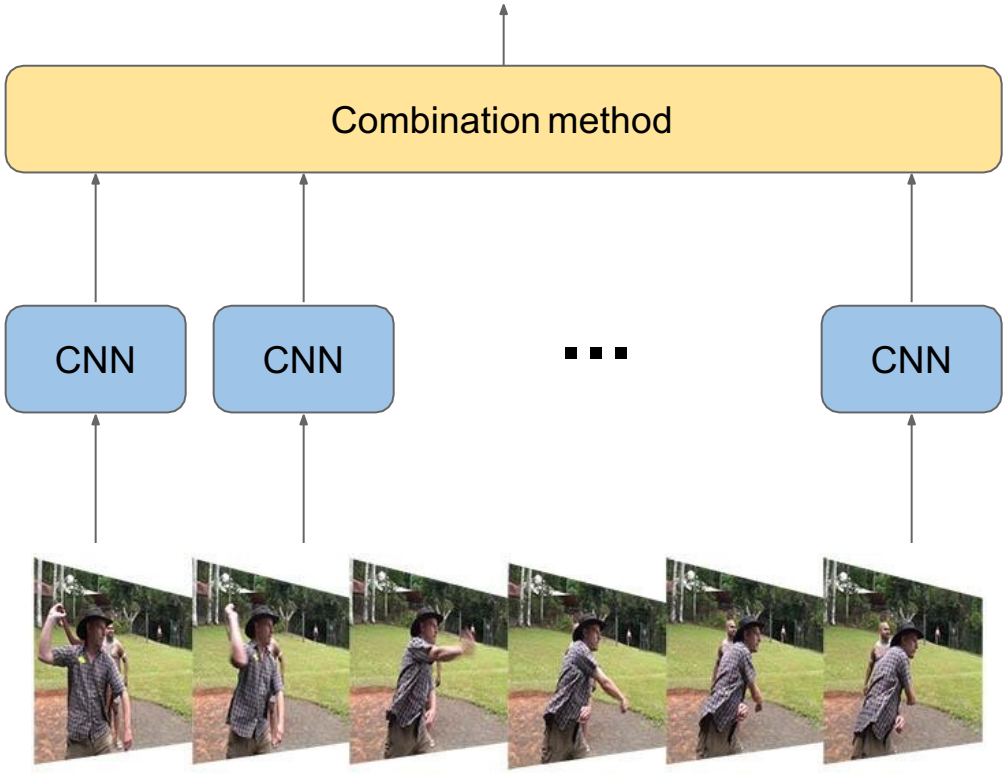


“body joints”

Agenda

- Motivation
- **Some terminologies**
- Two-stream representation

Single Frame Models

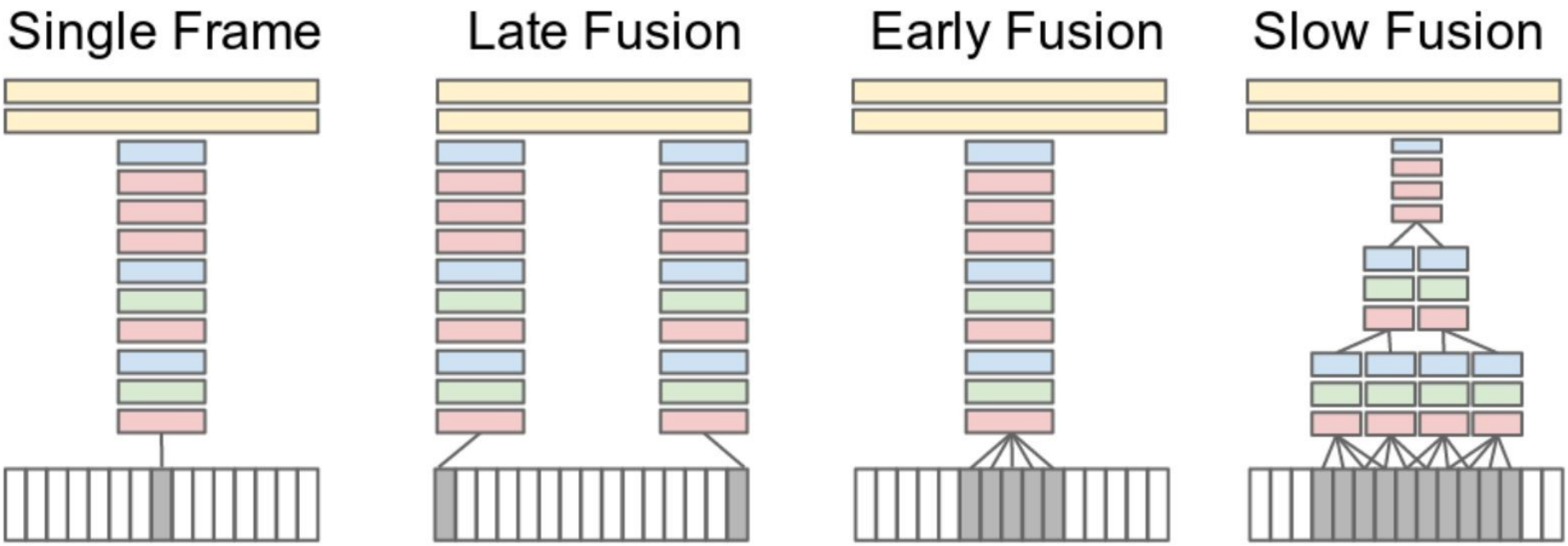


Combination is commonly implemented as a small NN on top of a pooling operation (e.g., max, sum, average).

Problem: pooling is not aware of the temporal order!

Yue-Hei Ng, Joe, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. "Beyond short snippets: Deep networks for video classification." CVPR 2015

Multiple Frames



Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. . Large-scale video classification with convolutional neural networks. CVPR 2014

Agenda

- Motivation
- Some terminologies
- **Two-stream representation**

Task: Video-clip Classification

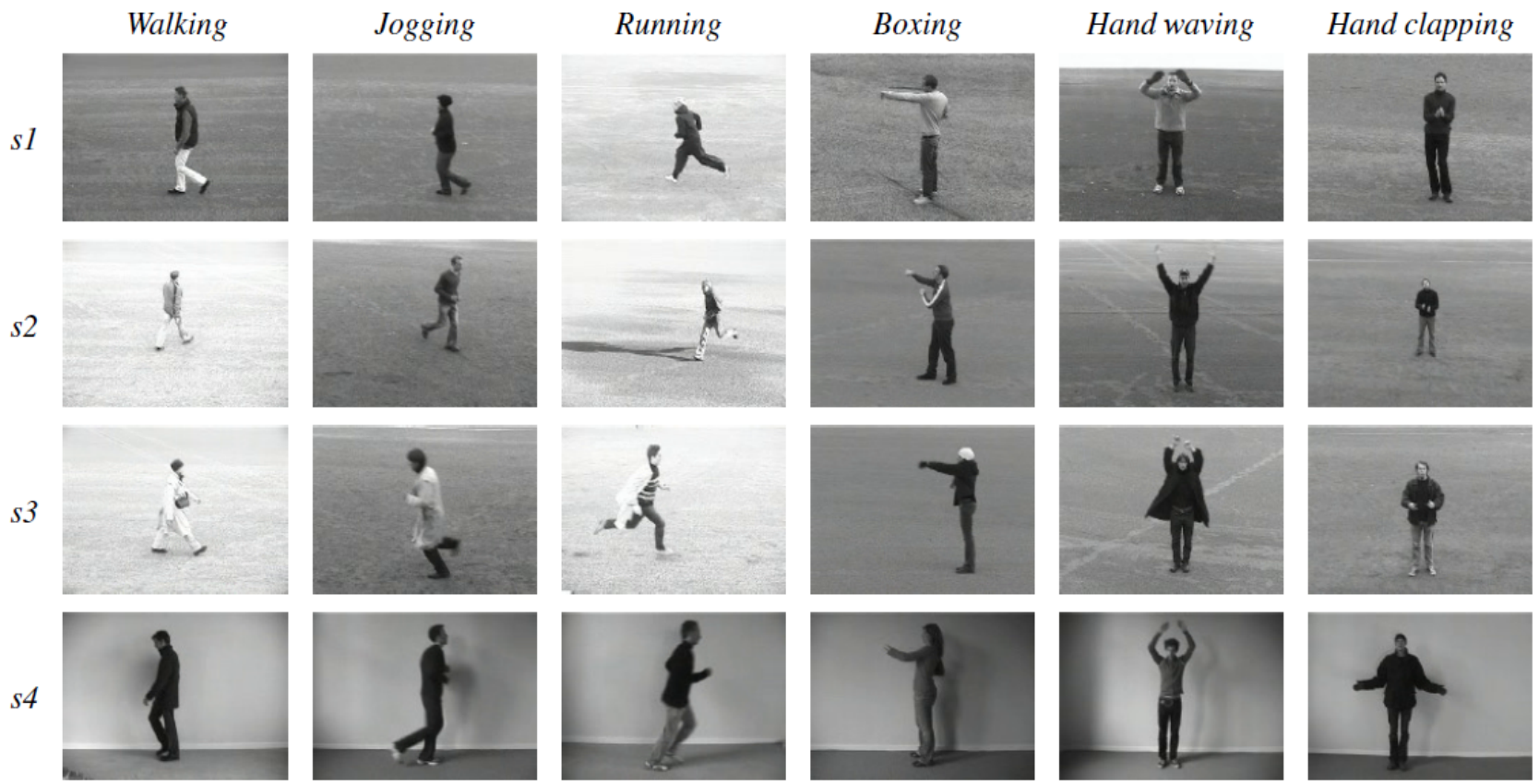
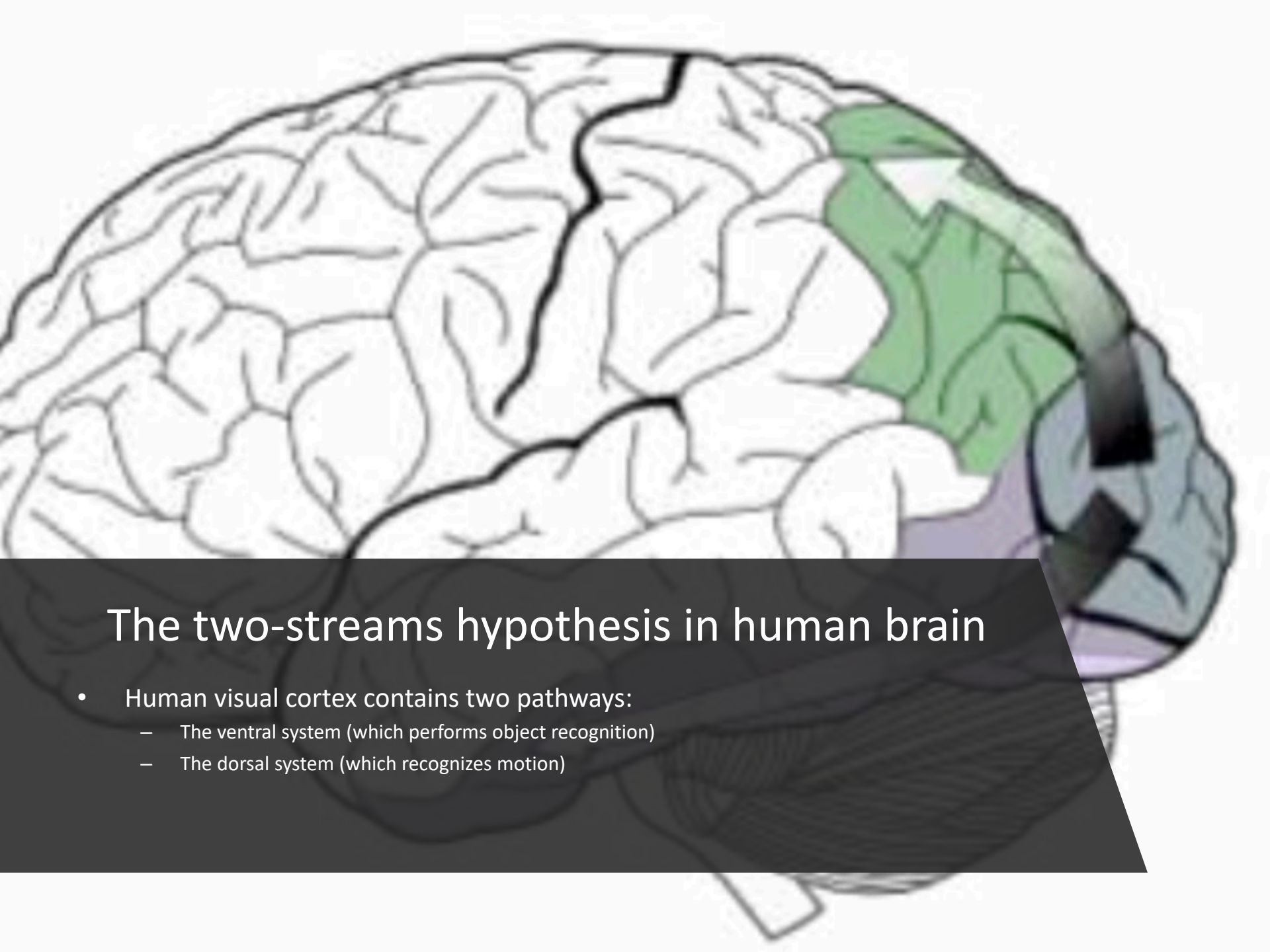


Figure from Schuldt et al.

KTH Action Classification Dataset

- Video dataset with a few thousand instances.
  - 25 people each:
    - perform 6 different actions
      - Walking, jogging, running, boxing, hand waving, clapping
    - in 4 different scenarios
      - Outdoors, outdoors w/scale variation, outdoors w/different clothes, indoors
    - (several times)
- Backgrounds are mostly free of clutter.
- Only one person performing a single action per video.

credit: Benzaid



The two-streams hypothesis in human brain

- Human visual cortex contains two pathways:
  - The ventral system (which performs object recognition)
  - The dorsal system (which recognizes motion)

Two-stream video recognition

- The spatial part, in the form of individual frame appearance, carries information about scenes and objects depicted in the video.
- The temporal part, in the form of motion across the frames, conveys the movement of the observer (the camera) and the objects.

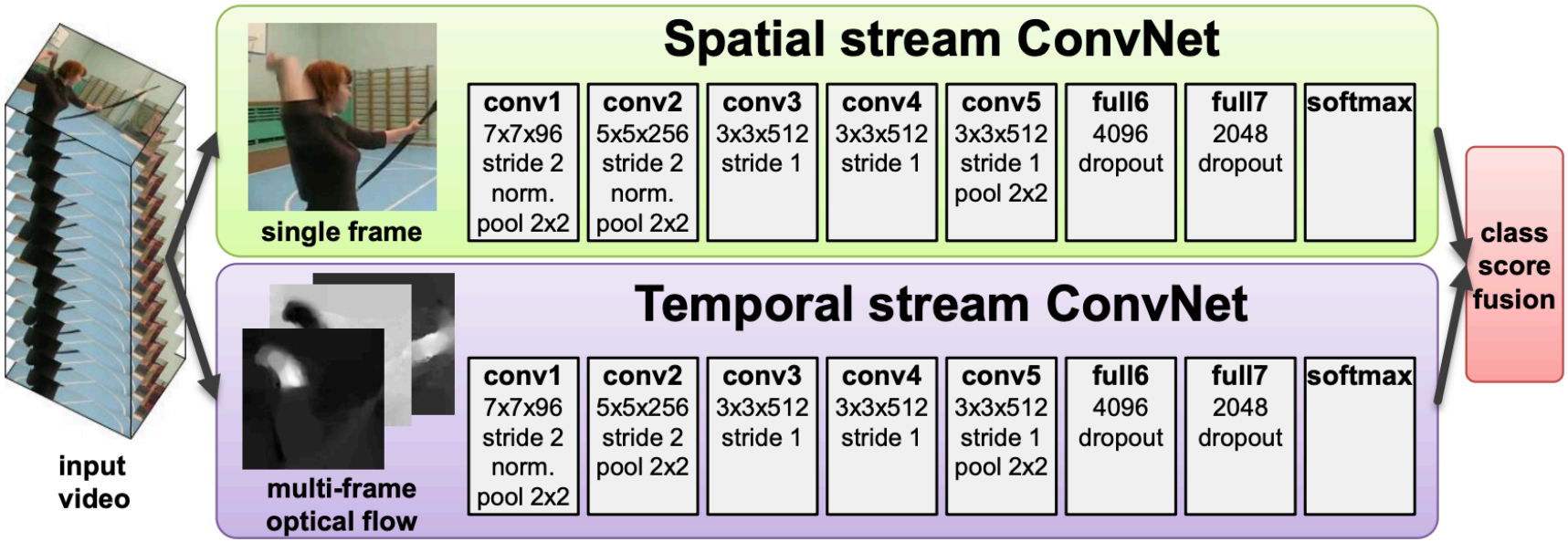
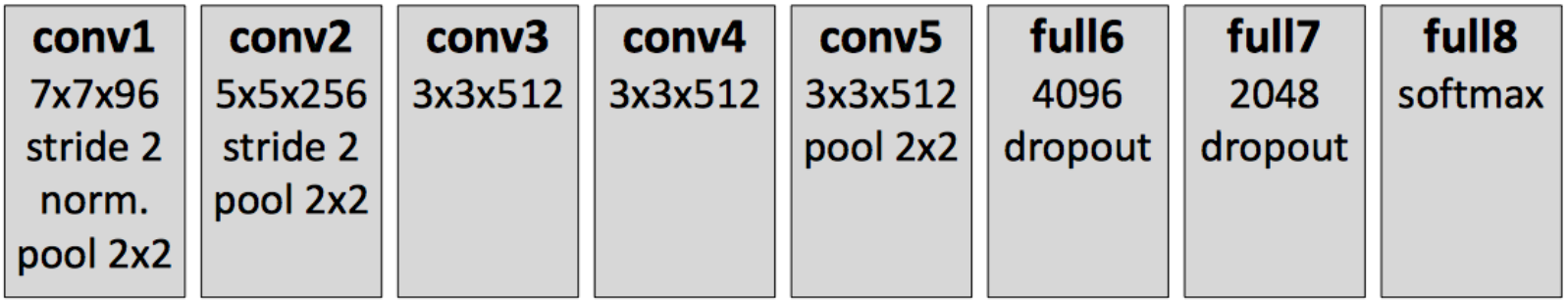


Figure 1: **Two-stream architecture for video classification.**

Two-stream architecture for video recognition

- Each stream is implemented using a deep ConvNet, which are combined by late fusion.

Convnet Layer Conviguration



- 8 weight layers (5 convolutional and 3 fully-connected)
- used for both spatial & temporal streams

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

Spatial Stream

- Predict action from still images – image classification
- Operates on individual video frames
- The static appearance by itself is a useful clue, due to some actions are strongly associated with particular objects

- Since a spatial ConvNet is essentially an image classification architecture
- Build upon the recent advances in large-scale image recognition methods
- Pre-train the network on a large image classification dataset, such as the ImageNet challenge dataset

Temporal Stream

- Optical flow
- Input of the ConvNet model is tracking optical flow displacement fields between several consecutive frames
- This input describes the motion between video frames

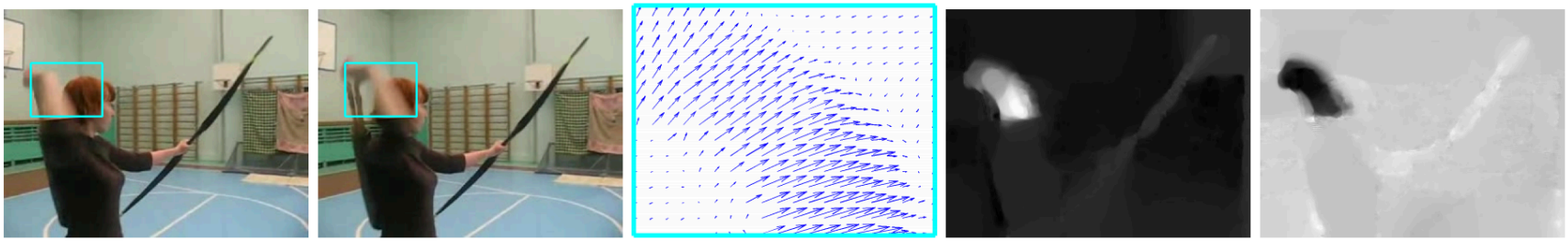


Figure 2: **Optical flow.** (a),(b): a pair of consecutive video frames with the area around a moving hand outlined with a cyan rectangle. (c): a close-up of dense optical flow in the outlined area; (d): horizontal component d^x of the displacement vector field (higher intensity corresponds to positive values, lower intensity to negative values). (e): vertical component d^y. Note how (d) and (e) highlight the moving hand and bow. The input to a ConvNet contains multiple flows (Sect. 3.1).

Optical Flow

- Displacement vector field between a pair of consecutive frames
- Each flow – 2 channels: horizontal & vertical components
- Computed using [Brox et al., ECCV 2004]
  - based on generic assumptions of constancy and smoothness
  - pre-computed on GPU (17fps), JPEG-compressed
- Global (camera) motion compensated by mean flow subtraction

Temporal Stream

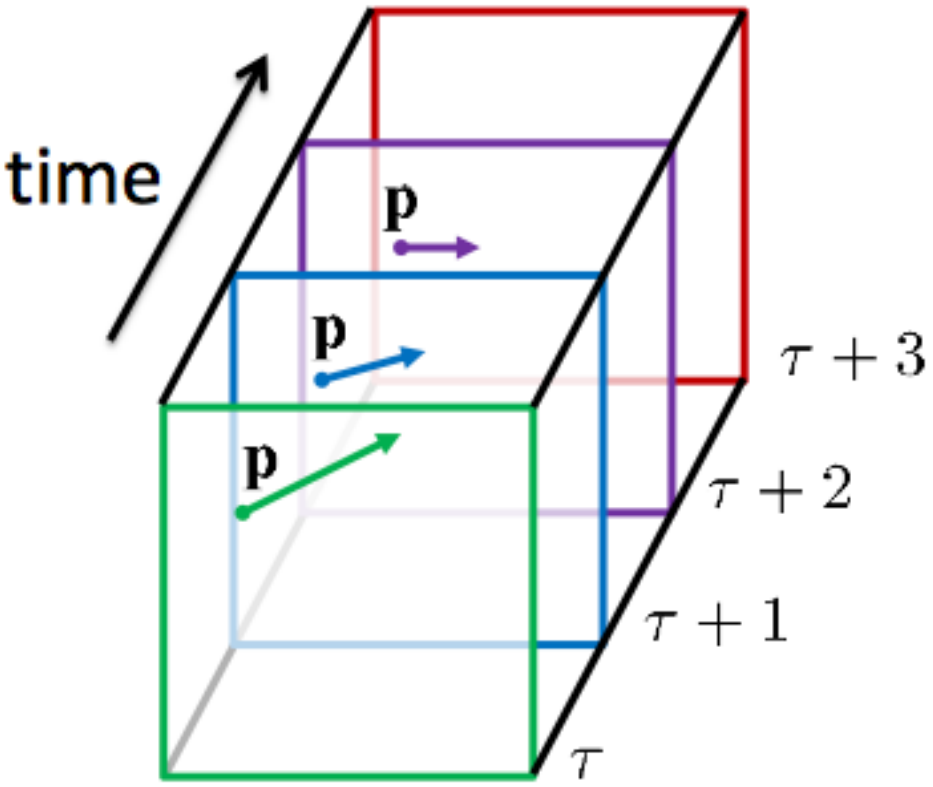
Predicts action from motion

Input

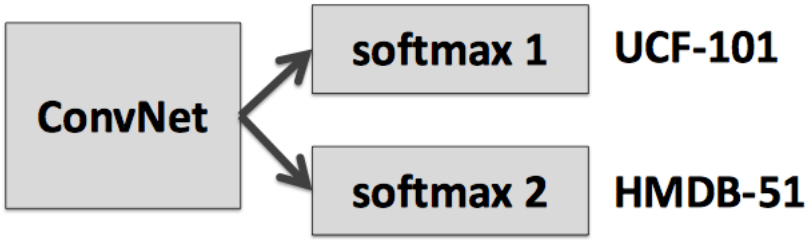
- Explicitly describes motion in video
- Stacked optical flow over several frames

Training

- From scratch with high drop-out (90%)



Evaluation



Video action classification datasets

- UCF-101 (101 class, 13K videos)
- HMDB-51 (51 class, 6.8K videos)

Results

Table 4: **Mean accuracy (over three splits) on UCF-101 and HMDB-51.**

Method	UCF-101	HMDB-51
Improved dense trajectories (IDT) [26, 27]	85.9%	57.2%
IDT with higher-dimensional encodings [20]	**87.9%**	61.1%
IDT with stacked Fisher encoding [21] (based on Deep Fisher Net [23])	-	**66.8%**
Spatio-temporal HMAX network [11, 16]	-	22.8%
"Slow fusion" spatio-temporal ConvNet [14]	65.4%	-
Spatial stream ConvNet	73.0%	40.5%
Temporal stream ConvNet	83.7%	54.6%
Two-stream model (fusion by averaging)	86.9%	58.0%
Two-stream model (fusion by SVM)	**88.0%**	**59.4%**