Emerging Topics in Human Activity Recognition

Michael S. Ryoo	**NASA-JPL**
Ivan Laptev	**INRIA**
Greg Mori	**Simon Fraser University**
Sangmin Oh	**Kitware**

CVPR tutorial on 2014/06/23

Introduction

Introduction

Computer Vision – Video Understanding

- Labeling of **events** by **humans** in a given video

Person 1 – *teases* **P2**, *runs away*

Person 3 – *kicks* **P1**

Person 4 – *stops fighting*

- Development of automated perception algorithms

Why video analysis?

Data:

~2.5 Billion new images / month

BBC Motion Gallery

ina

TV-channels recorded since 60’s

flickr

~5K image uploads every min.

YouTube Broadcast Yourself

>34K hours of video upload every day

~30M surveillance cameras in US
=> ~700K video hours/day

And even more with future wearable devices

Why video analysis?

Applications:

First appearance of N. Sarkozy on TV

Sociology research: Influence of character smoking in movies

Education: How do I make a pizza?

Where is my cat?

Predicting crowd behavior
Counting people

Motion capture and animation

Why video analysis?

Applications:

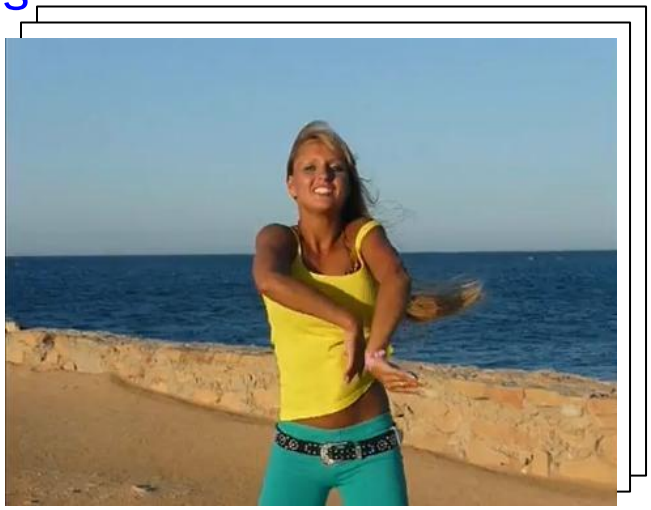
Unconstrained video search

Why human activities?

How many person-pixels are in the video?

Movies

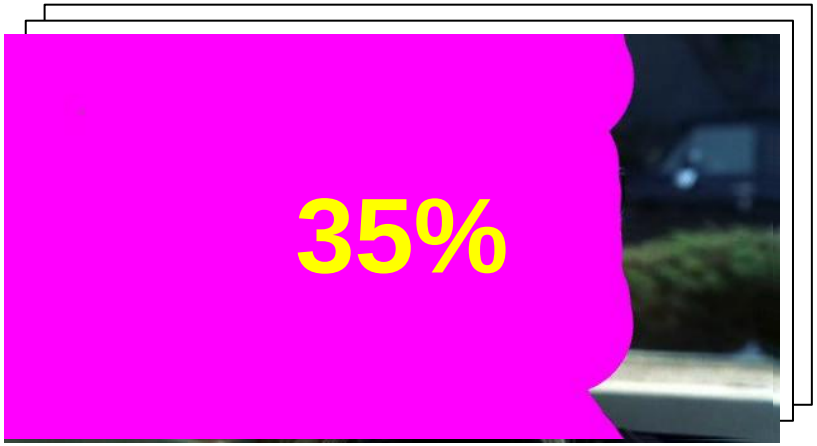
TV



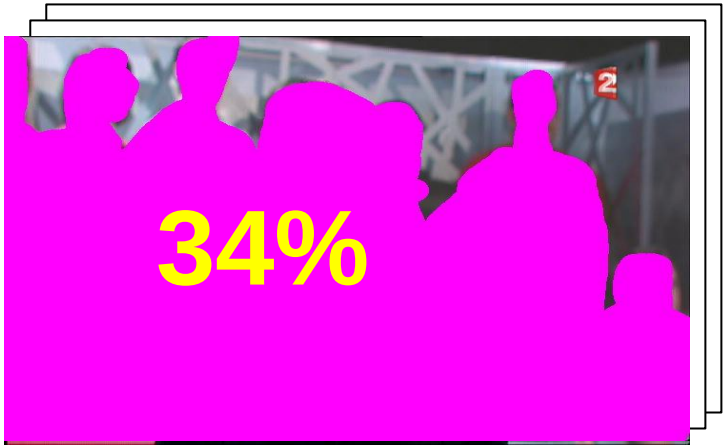
YouTube

Why human activities?

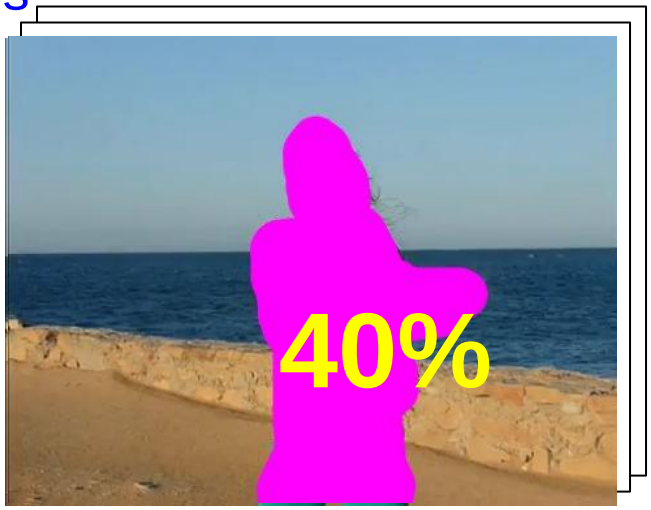
How many person-pixels are in the video?



Movies



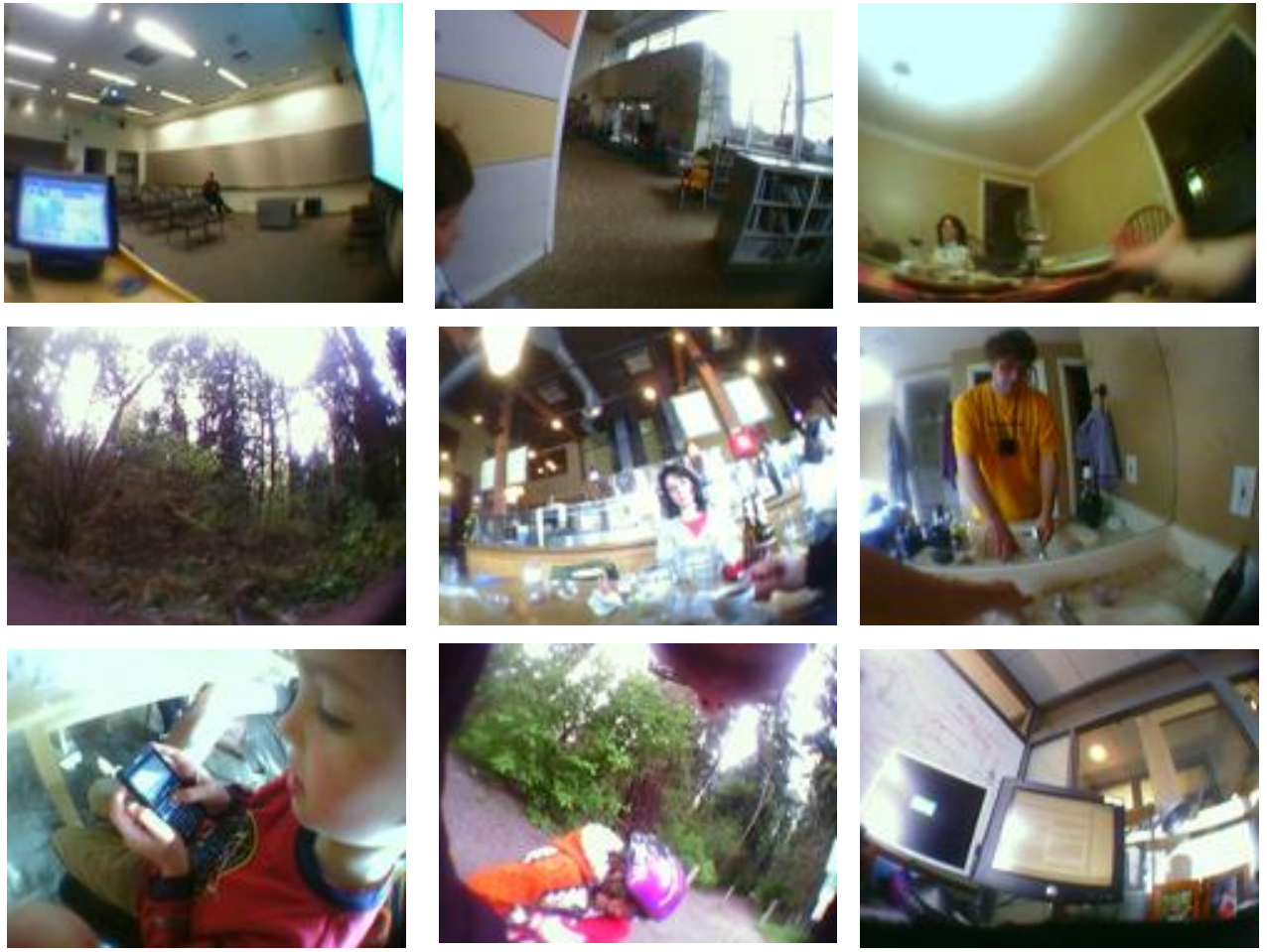
TV



YouTube

How many person pixels in our daily life?

Wearable camera data: Microsoft SenseCam dataset



How many person pixels in our daily life?

Wearable camera data: Microsoft SenseCam dataset

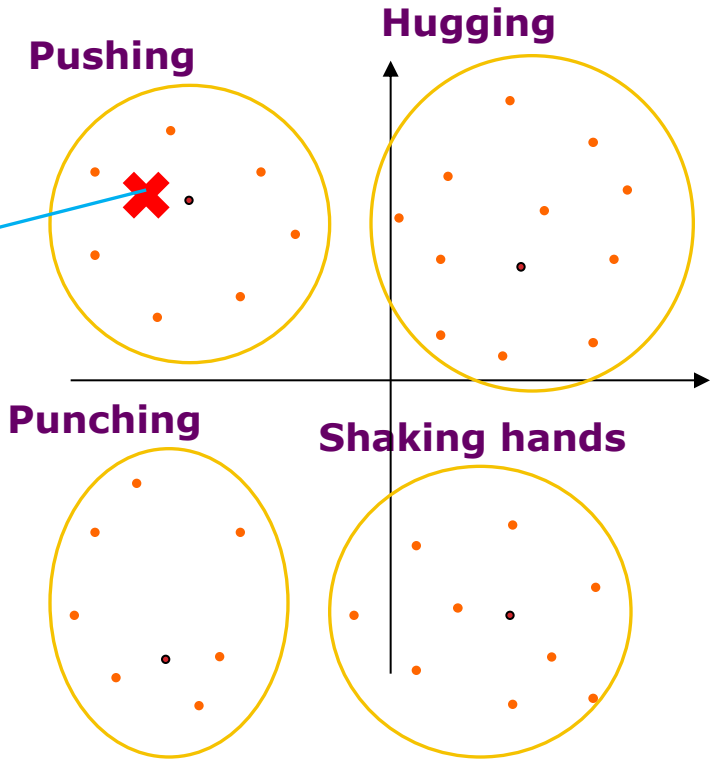
Activity recognition

***Search* for the particular time interval**

- <starting time, ending time>
- Video segment containing the activity

Input:

continuous video stream



Activity detection by classification

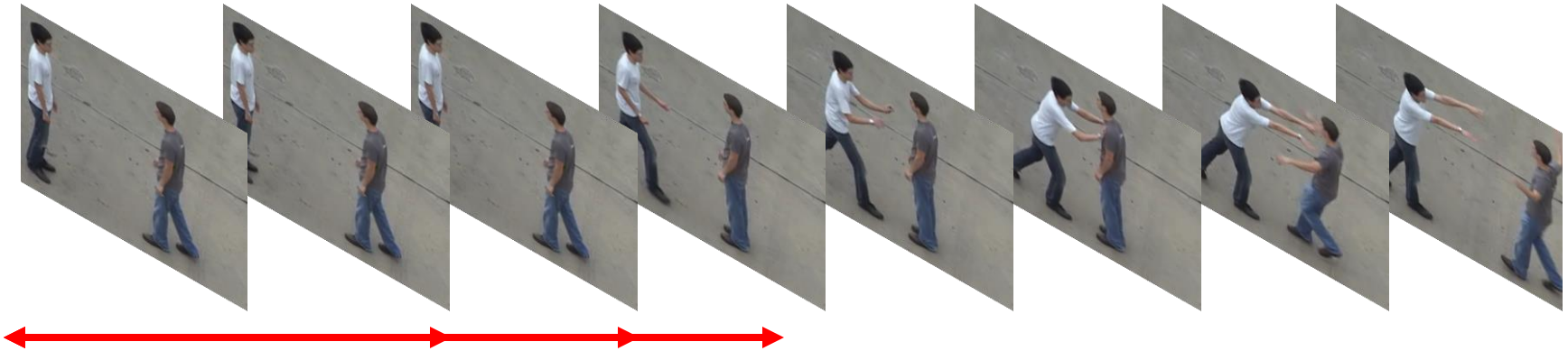
Binary classifier

Push/not-push Classifier

Yes, a push.

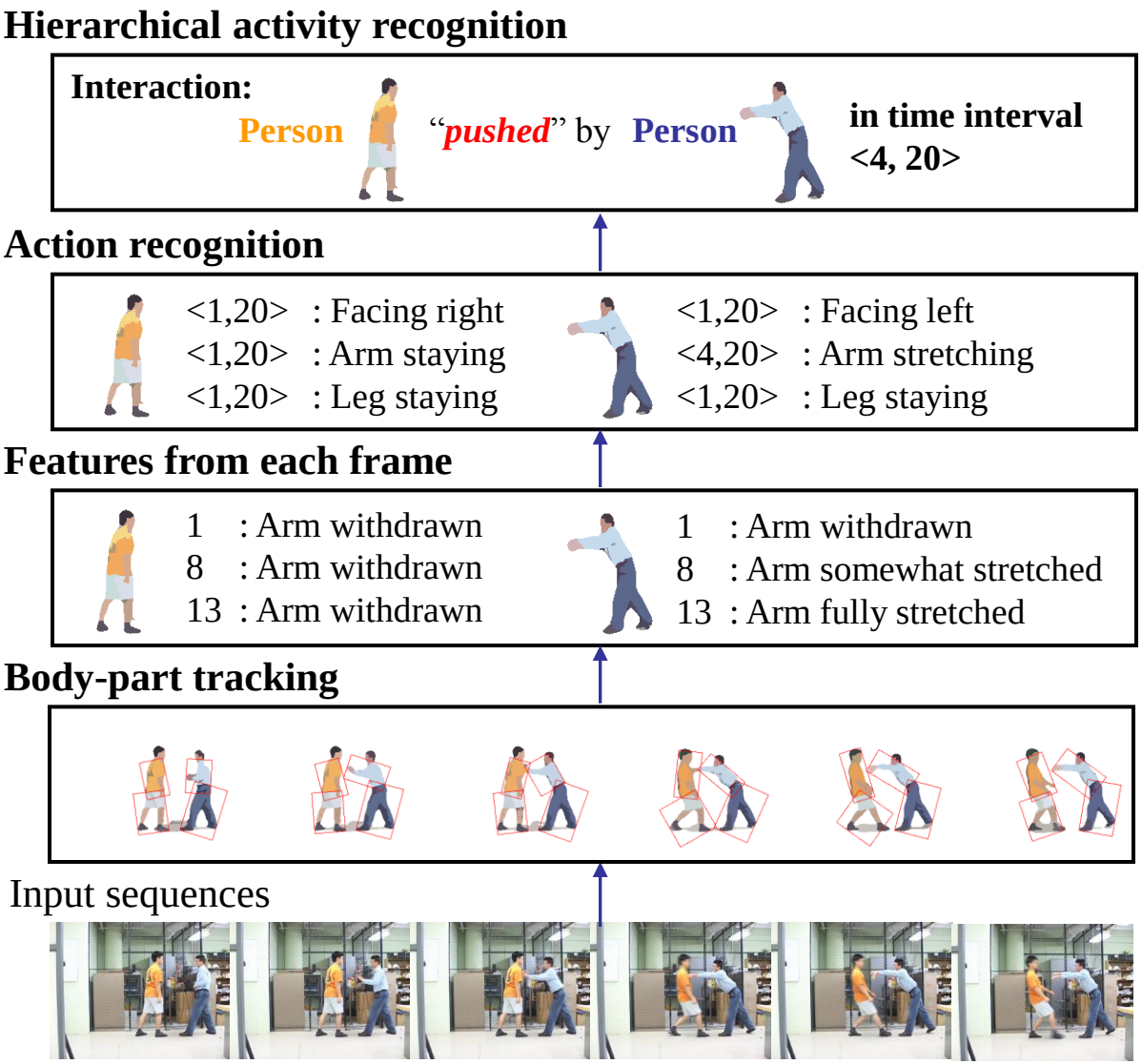
Sliding window technique

- Classify all possible time intervals



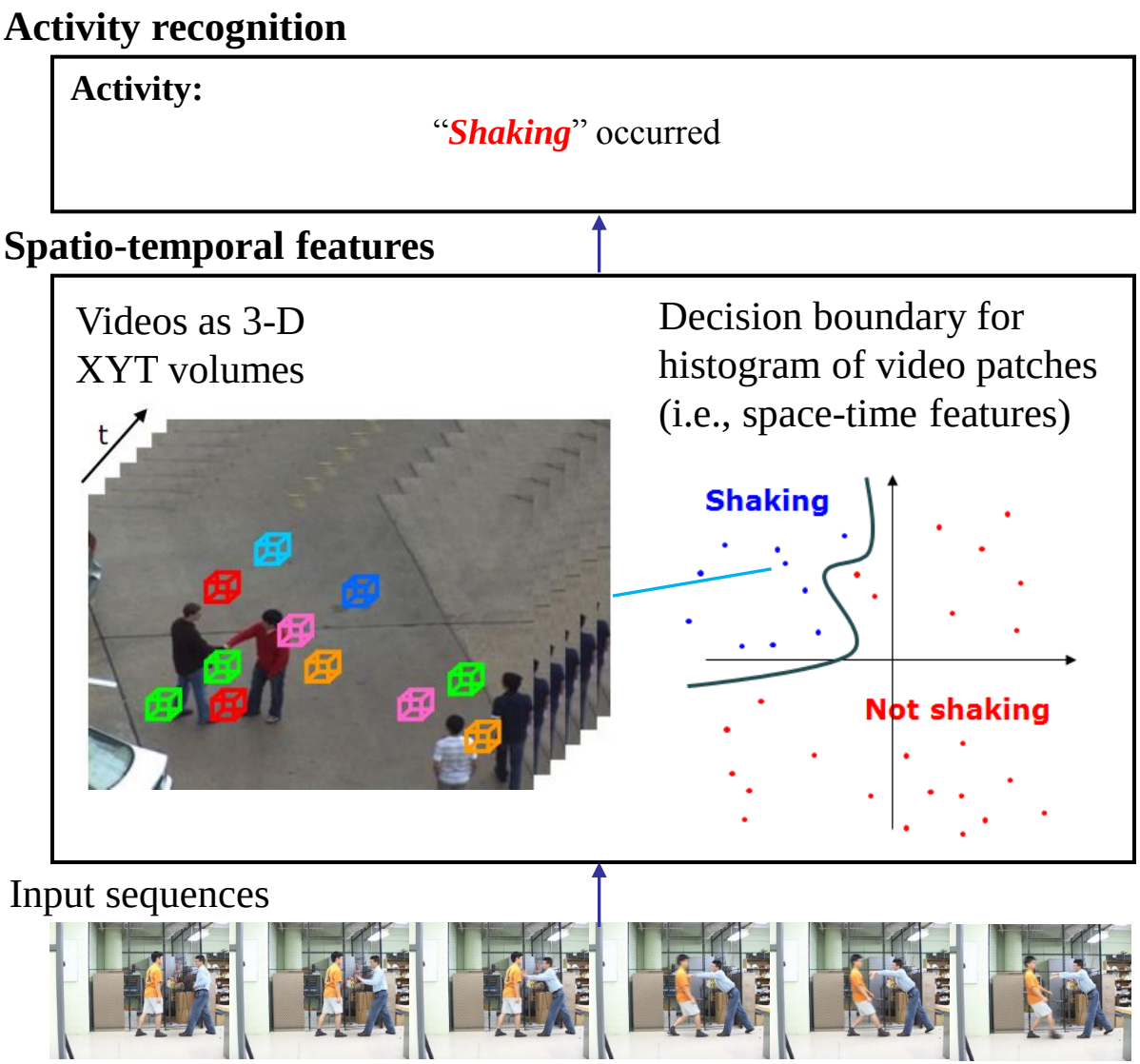
Pushing

Activity Recognition with postures



- **Interaction**
- **Action**
  - Sequence of features
- **Features**
  - Numerical status of a body part
- **Tracking**
  - Estimates locations of human body parts

Activity Recognition with video features



- **Activity recognition**
  - Decision boundary
- **Activity representation**
  - A set of local spatio-temporal features
- **Features**
  - Information in local video patches with salient movements

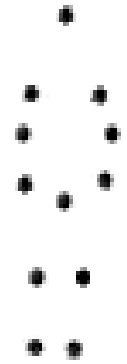
History

Time

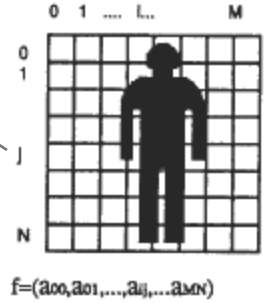
1973

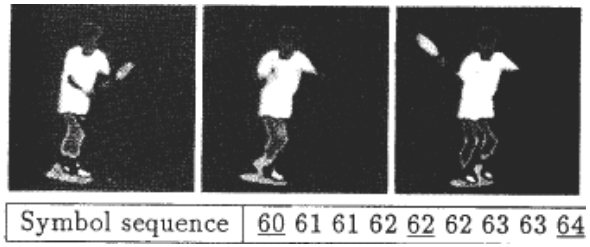
1992

1995

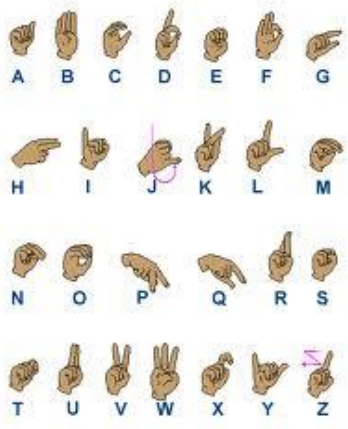


Johansson's experiments [Johansson 1973]





Tennis action recognition [Yamato et al. 1992]



American sign language (ASL) recognition [Starner and Pentland 1995]

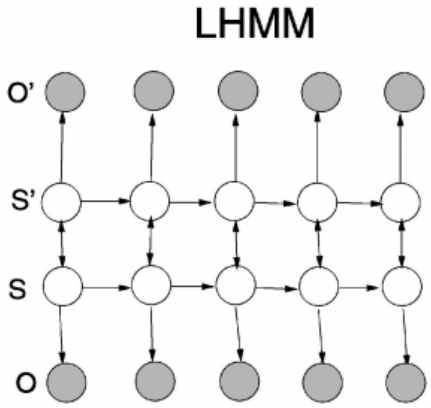
History

Time

1995
1997
2000

3-D posture-based recognition [Gavrila and L. Davis1995]

Motion history images [J. Davis, Bobick 1997]

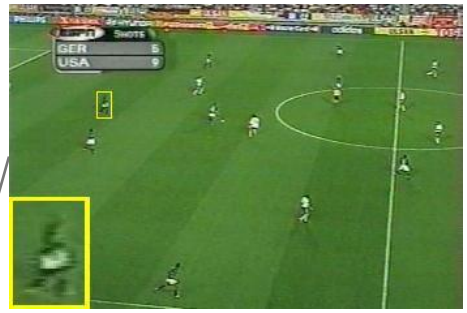


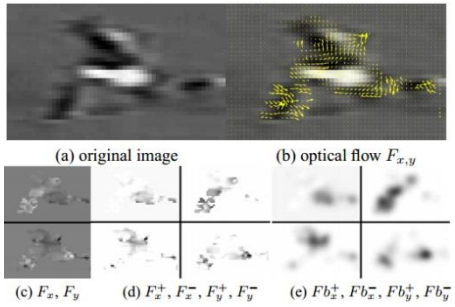
Human-human interactions [Oliver, Rosario, Pentland 2000]

History

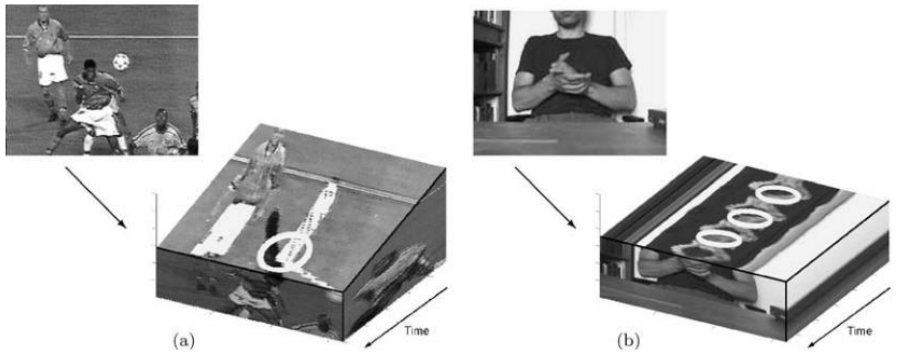
Time

2003
2005
2008
2009





Far-field action recognition [Efros, Berg, Mori, Malik 2003]

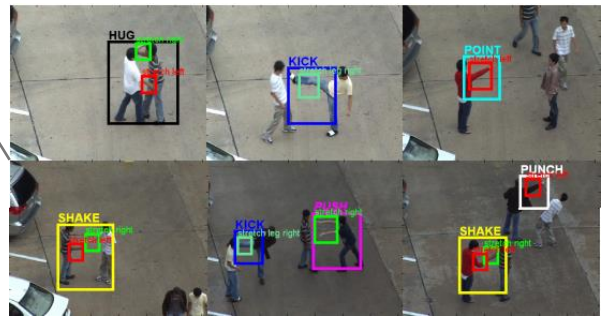


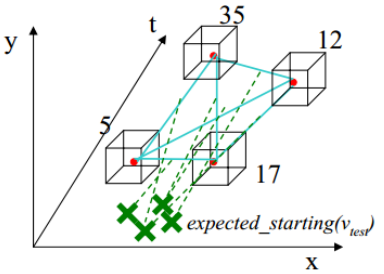
KTH

Local spatio-temporal features [Laptev 2005]

Hollywood

Movies [Laptev 2008]





UT-Interaction

Spatio-temporal relations [Ryoo and Aggarwal 2009]

Dimension 1: type of videos

Different types of videos and their dataset

Surveillance videos

- Static cameras
- Side or top view
- Simple background

Movies and user videos

- Moving cameras
- Side view
- Dynamic

Sports videos

- Video segments
- Side or top view
- Objects/people

First-person videos

- Moving cameras
- 1st-person view
- Very dynamic

Dimension 2: levels of human activities

There are various types of activities

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

gestures

actions

human-object interactions

interactions

group activities

Levels of human activities

Dimension 3: structure in activities

Different levels of structure complexity (temporal/spatial)

Complexity of structure in human activities

Existing works

Based on 'video type' and 'activity level' dimensions

	Actions	**Human-object interactions**	**Human-human interactions**	**Group activities**
Surveillance videos	[Laptev 05]	[Oh et al. 11]	[Ryoo and Aggarwal 09] [Vahdat, Gao, Ranjbar, Mori 11]	[Ryoo and Aggarwal 08,11] [Lan, Wang, Yang, Mori 10]
Movies and user videos	[Laptev 07]	[Marszałek, Laptev, Schmid 09] [Kim, Oh, Vahdat, Cannons, Perera, Mori 13]		
Sports videos	[Efros, Berg, Mori, Malik 2003]	[Rodriguez, Ahmed, Shah 08] [Niebles, Chen, Fei-Fei 10]		[Lan, Sigil, Mori 12]
First-person videos	[Kitani et al. 11]	[Fathi et al. 2011] [Pirsiavash and Ramanan 2012]	[Ryoo and Matthies 13]	

Ivan Laptev
Greg Mori
Michael Ryoo
Sangmin Oh

Existing works

Based on 'video type' and 'activity level' dimensions

	Actions	Human-object interactions	Human-human interactions	Group activities
Surveillance videos	[Laptev 05]	[Oh et al. 11]	[Ryoo and Aggarwal 09] [Vahdat, Gao, Ranjbar, Mori 11]	[Ryoo and Aggarwal 08,11] [Lan, Wang, Yang, Mori 10]
Movies and user videos	[Laptev 07]	[Marszałek, Laptev, Schmid 09] [Kim, Oh, Vahdat, Cannons, Perera, Mori 13]		
Sports videos	[Efros, Berg, Mori, Malik 2003]	[Rodriguez, Ahmed, Shah 08] [Niebles, Chen, Fei-Fei 10]		[Lan, Sigil, Mori 12]
First-person videos	[Kitani et al. 11]	[Fathi et al. 2011] [Pirsiavash and Ramanan 2012]	[Ryoo and Matthies 13]	

Why difficult?

- **Large variations in appearance:** occlusions, non-rigid motion, view-point changes, clothing…

Action *Hugging*:



…

- **Manual collection of training samples is prohibitive:** many action classes, rare occurrence





…

- **Action vocabulary is not well-defined**

Action *Open*:




…

Challenges - variations