# Emerging Topics in Human Activity Recognition

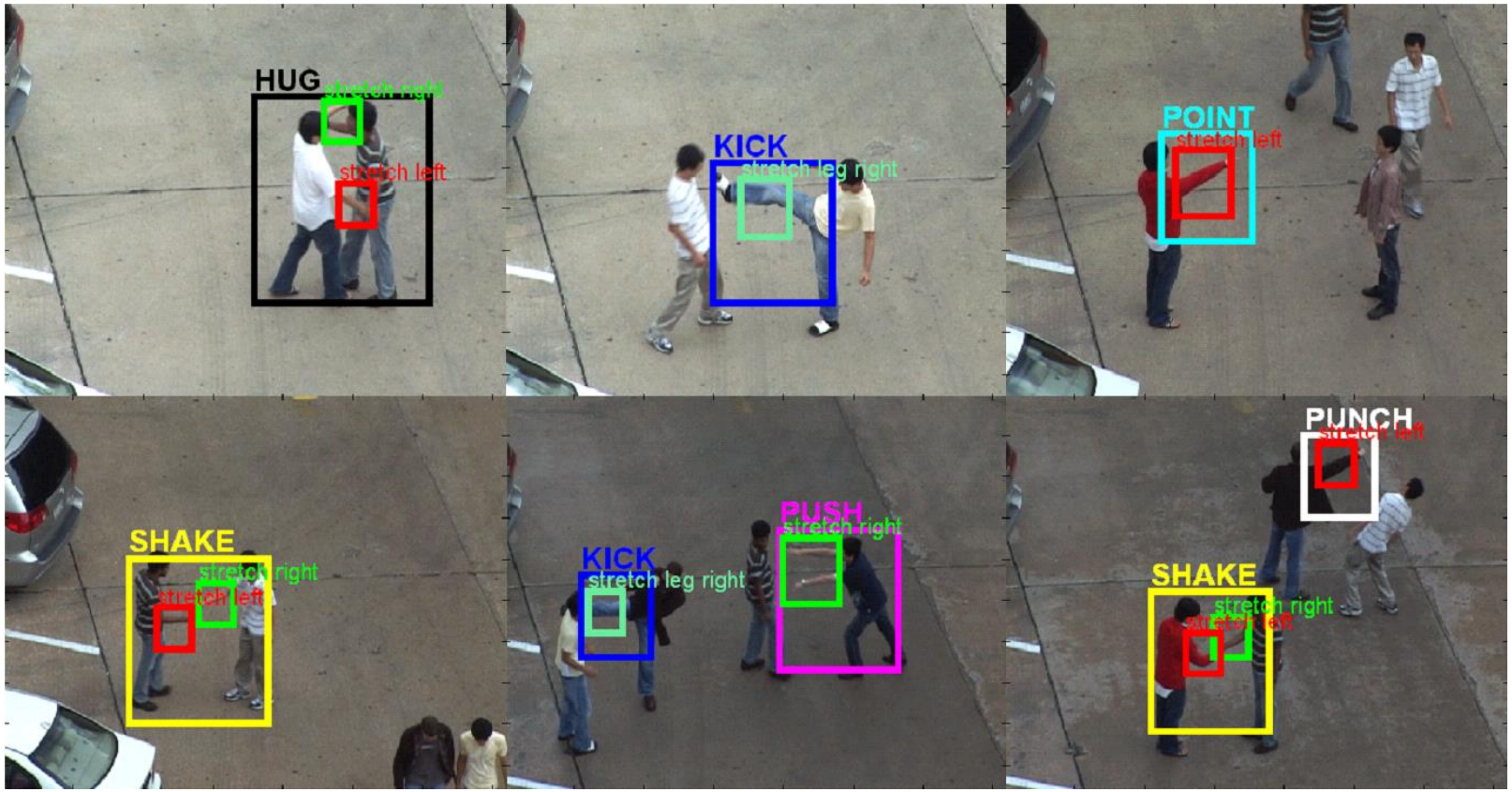

**Michael Ryoo** NASA Jet Propulsion Laboratory
**Ivan Laptev** INRIA
**Greg Mori** Simon Fraser University
**Sangmin Oh** Kitware


**CVPR tutorial on 2014/06/23**

# Human activity recognition

**Detecting interaction-level human activities**

- Multiple actors, pedestrians, weather changes, …




Ryoo and Aggarwal, ICCV 2009

# Group activity recognition

**Group stealing**

- Real video from a CCTV in Malaysia
  - A group of thieves steal a laptop by distracting the owner


Thieves

Owners

Laptop

Ryoo and Aggarwal, IJCV 2011

# First Person Activity Recognition

**Michael Ryoo**
**NASA Jet Propulsion Laboratory**

**CVPR tutorial on 2014/06/23**

# First-person video understanding

**Objective**

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

**P1** and **P2** are *fighting*

**P2** *punches* **P1**

**P1** *blocks*

**P2** *punches* **P1**

**P2** *punches* **P1**

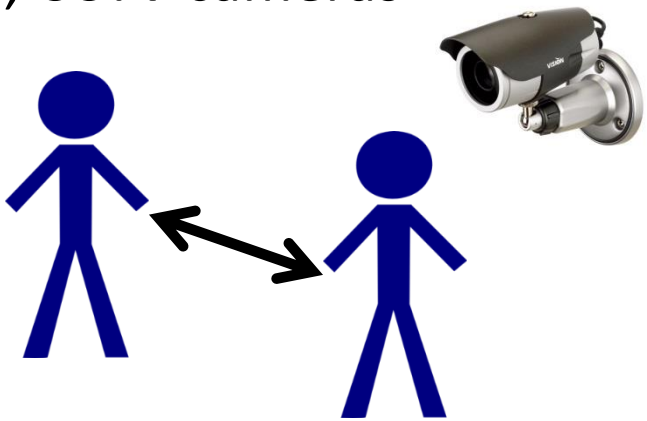
**P1** *kicks* **P2**

- Videos are taken from the actor's own viewpoint

# 1st person vs. 3rd person

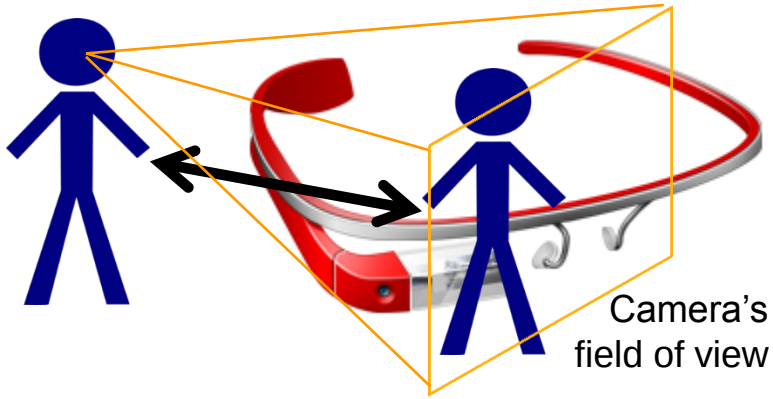
## 3rd person viewpoint

- E.g., CCTV cameras



## 1st person viewpoint

- E.g., wearable cameras

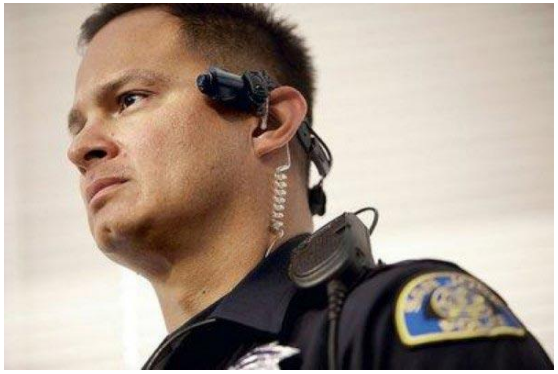



# Why first-person?

## Wearable cameras and computers

Personal assistant

Sports play analysis



Enforcement / rescue

## Robots

- Any video observation from a robot is a first-person video
  - US Army – Big Dog, LS3
  - Particularly interested in reactions to the robot

# Topics in first-person recognition

## Ego-action recognition

- Sports videos
  - Skiing, riding a bike, …
  - [Kitani et al. 2011]


## Object-oriented analysis

- Objects in front of me
  - [Fathi et al. 2011]
  - [Pirsiavash and Ramanan 2012]

## First-person interaction recognition

- Human-human, human-robot
- What are they doing to me?

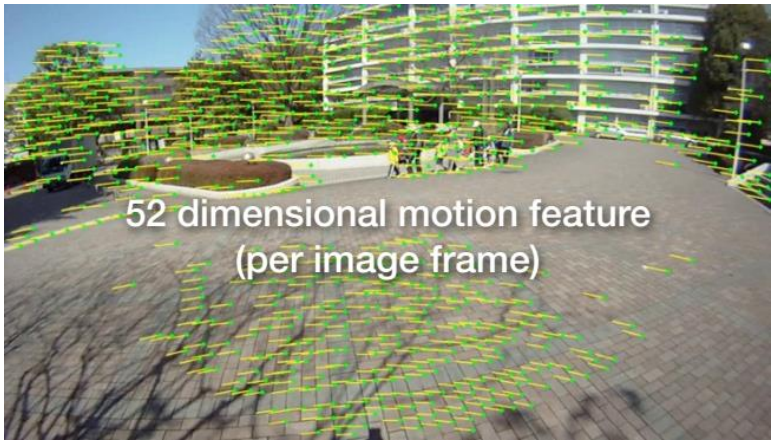
# Ego-action recognition

**[Kitani et al., CVPR 2011]**

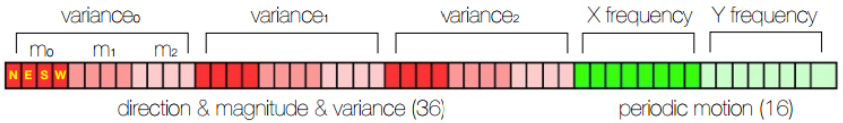
- Ego-actions of a person wearing a camera
- Sports videos

# Ego-action recognition (cont'd)

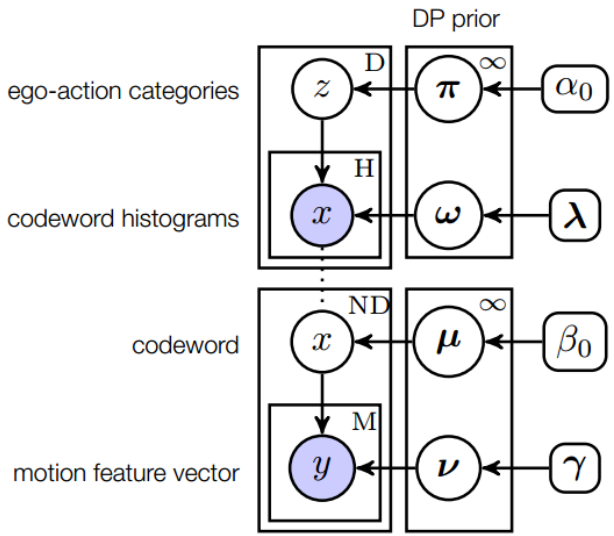
## Optical flow-based global motion descriptors




1. Track corner points
2. Compute dominant motion plane
3. Bin each flow vector by orientation, magnitude and global variance
4. Concatenate DFT frequency magnitudes




## Stacked Dirichlet Process Mixture Model




codeword

codeword assignment

motion histogram

$$\hat{h} = \arg\max_h \left\{ p(x_{ndh} | \boldsymbol{y}_{nd}, \mathbf{Y}^{-nd}, \mathbf{X}^{-nd}) \right\}$$

past observations

past assignments

$$\propto \arg\max_h \left\{ p(x_{ndh} | \mathbf{X}^{-nd}) p(\boldsymbol{y}_{nd} | \mathbf{Y}^{-nd}, \mathbf{X}^{nd}) \right\}$$

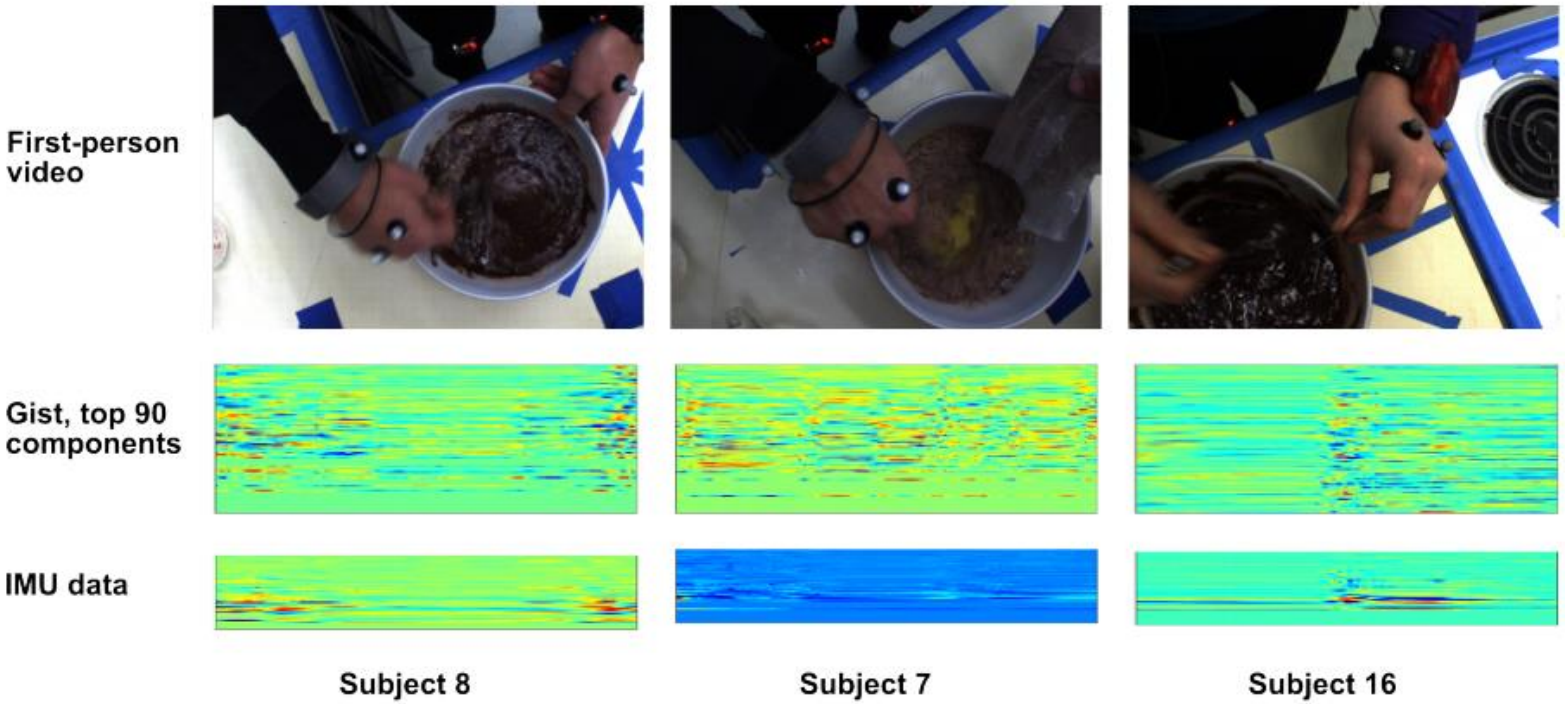
mixture of CRPs

likelihood

# First-person actions of using objects

**[Spriggs, De la Torre, Hebert, CVPRW 2009]**

- First-person videos of cooking actions
  - Vision + IMU
- GIST feature per frame 4*4*8-D (i.e., per-frame scene features)
  - Gaussian mixture model (GMM) + hidden Markov model (HMM)



# Objects in first-person videos

**[Fathi, Ren, Rehg, CVPR 2011] [Fathi, Farhadi, Rehg, ICCV 2011]**

- Foreground segmentation
  - Super-pixels using color-textures and edges
  - MRF using graph-cut

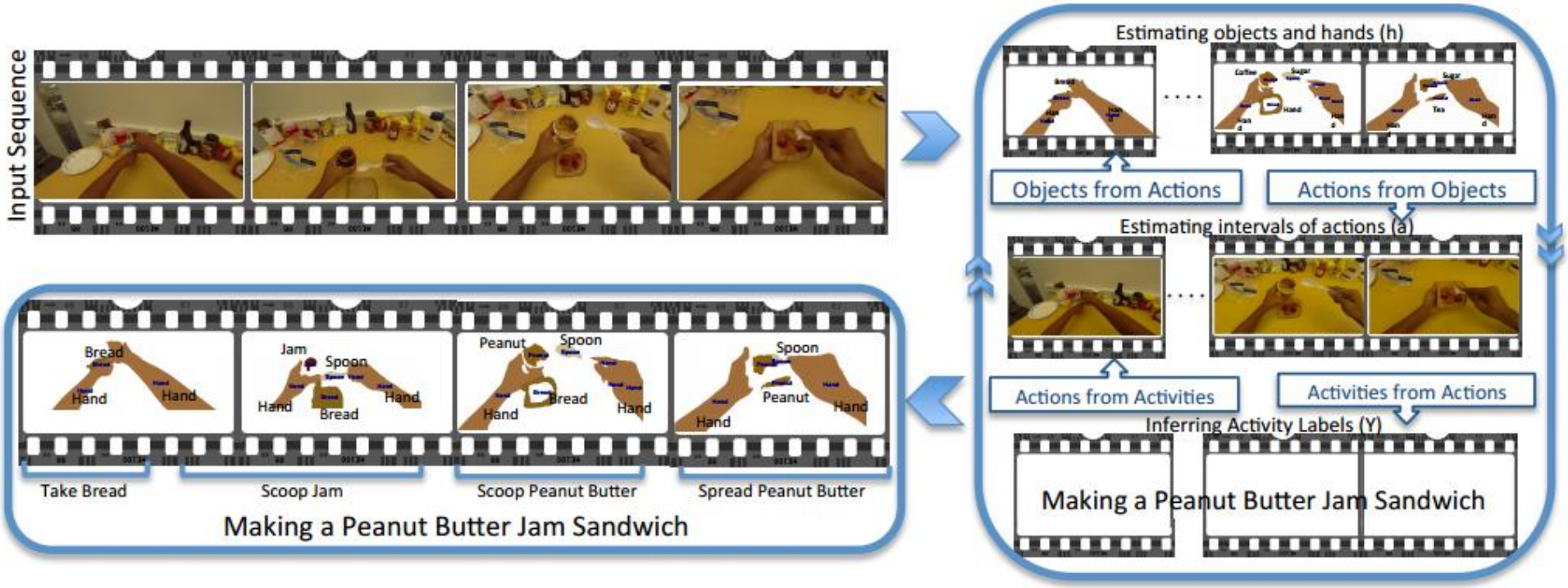
(a) (b) (c) (d) (e)

  - Hand segmentation using color histograms
- Foreground object classification
  - Egocentric activities are highly correlated with appearing objects
  - MIL framework: learning to classify multiple clustered regions

# Objects in first-person videos (cont'd)

**A graphical model to combine object recognition with (ego)action recognition**

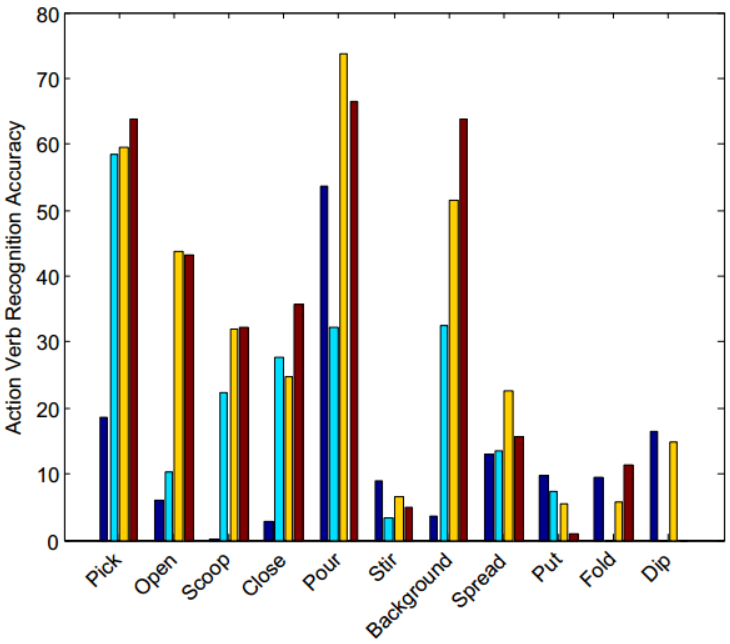
- Low-level: action recognition based on object interactions
  - Object frequency, locations, optical flows, hand pose, hand OF, …
- Activities-from-actions and actions-from-activities



# Objects in first-person videos (cont'd)

**First-person videos of daily activities**

- Hotdog Sandwich, Instant Coffee, Peanut Butter Sandwich, Jam and Peanut Butter Sandwich, Sweet Tea, Coffee and Honey, Cheese Sandwich



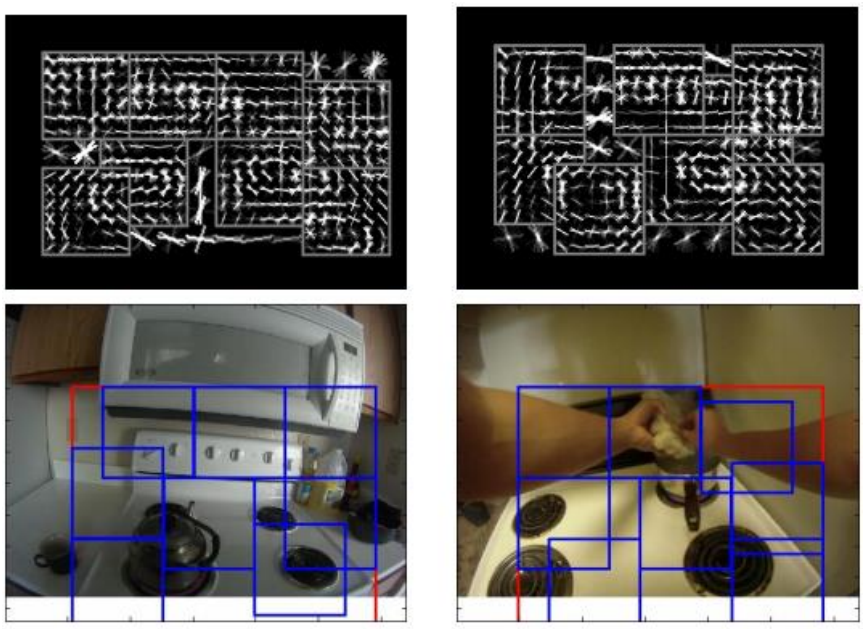

STIP, SIFT, proposed, proposed + activity context

# Objects in first-person videos

**[Pirsiavash and Ramanan, CVPR 2012]**

**Object-centric "Bag-of-objects" activity recognition**

- Composite object models
- Models for 'active' objects
  - Spatial reasoning and skin information



(a) passive stove (b) active stove

# Why are ‘interactions’ important?

**Ego-action: *riding* a motor bike**

**Object: *sheep* in front of me**

**Are these sufficient?**

# First-person interaction recognition

**[Ryoo and Matthies, CVPR 2013]**

# JPL-Interaction dataset

**First-person videos of interaction-level human activities**

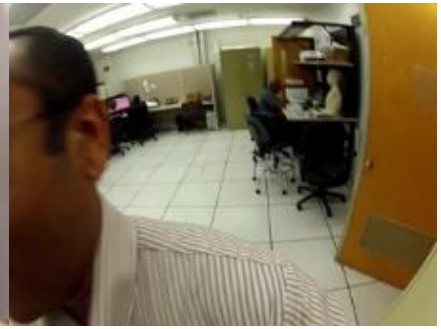
- Taken from a humanoid robot's perspective

# JPL-Interaction dataset (cont'd)

**Seven types of human activities involving physical interactions**

- 4 friendly, 1 neutral, 2 negative interactions

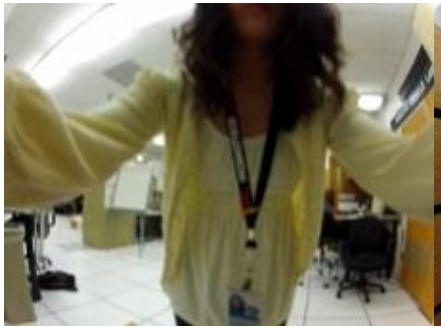
Hand shake



Hug

Point-Converse

Punch



Pet

Wave

Throw

# Two questions

**What features do we need for first-person (interaction) videos?**

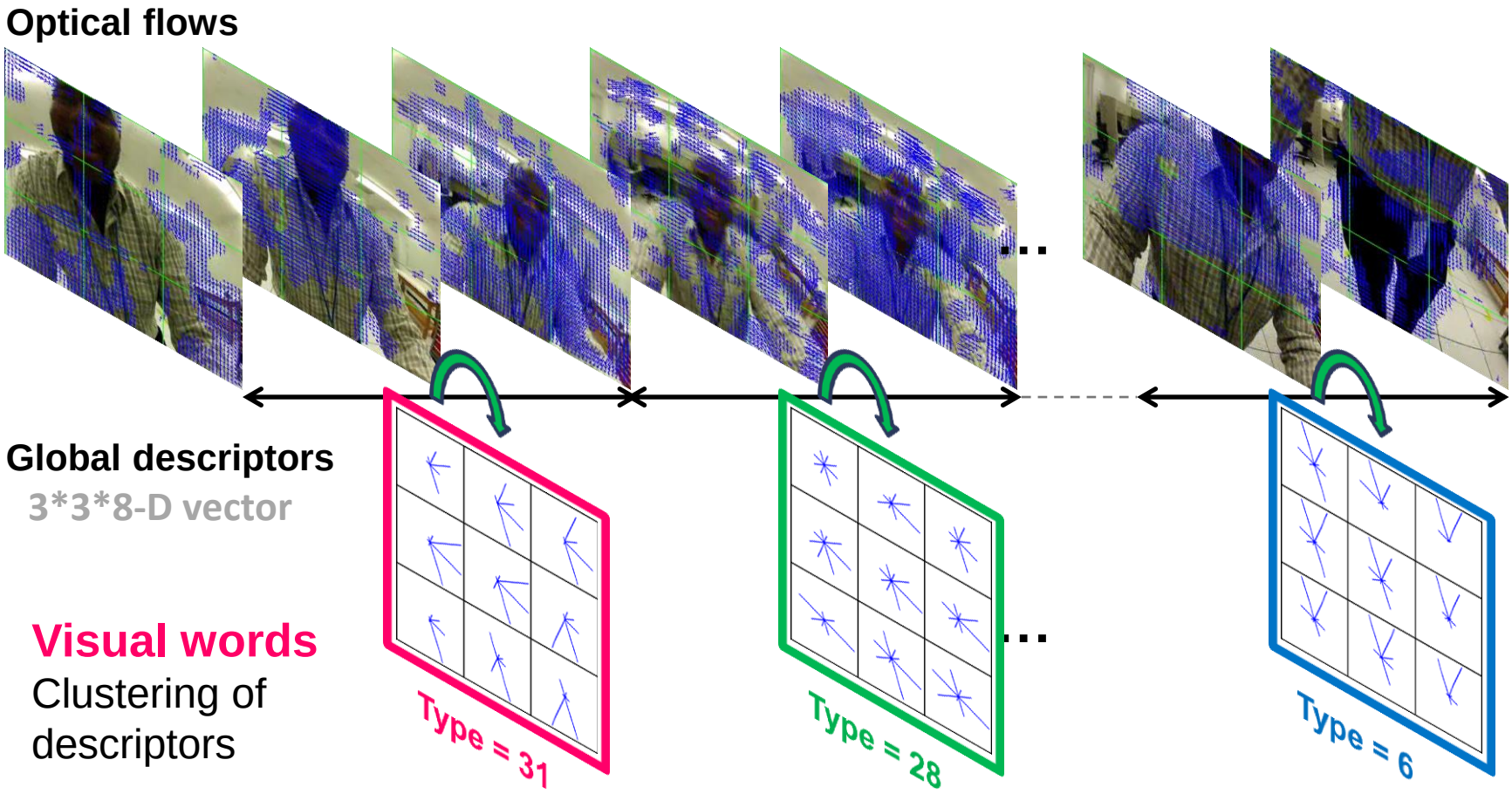
**How important is it to consider activities' temporal structures in first-person videos?**

- Cause-effect relations: punch-collapse

# Features for first-person videos

## Global motion descriptors: ego-motion

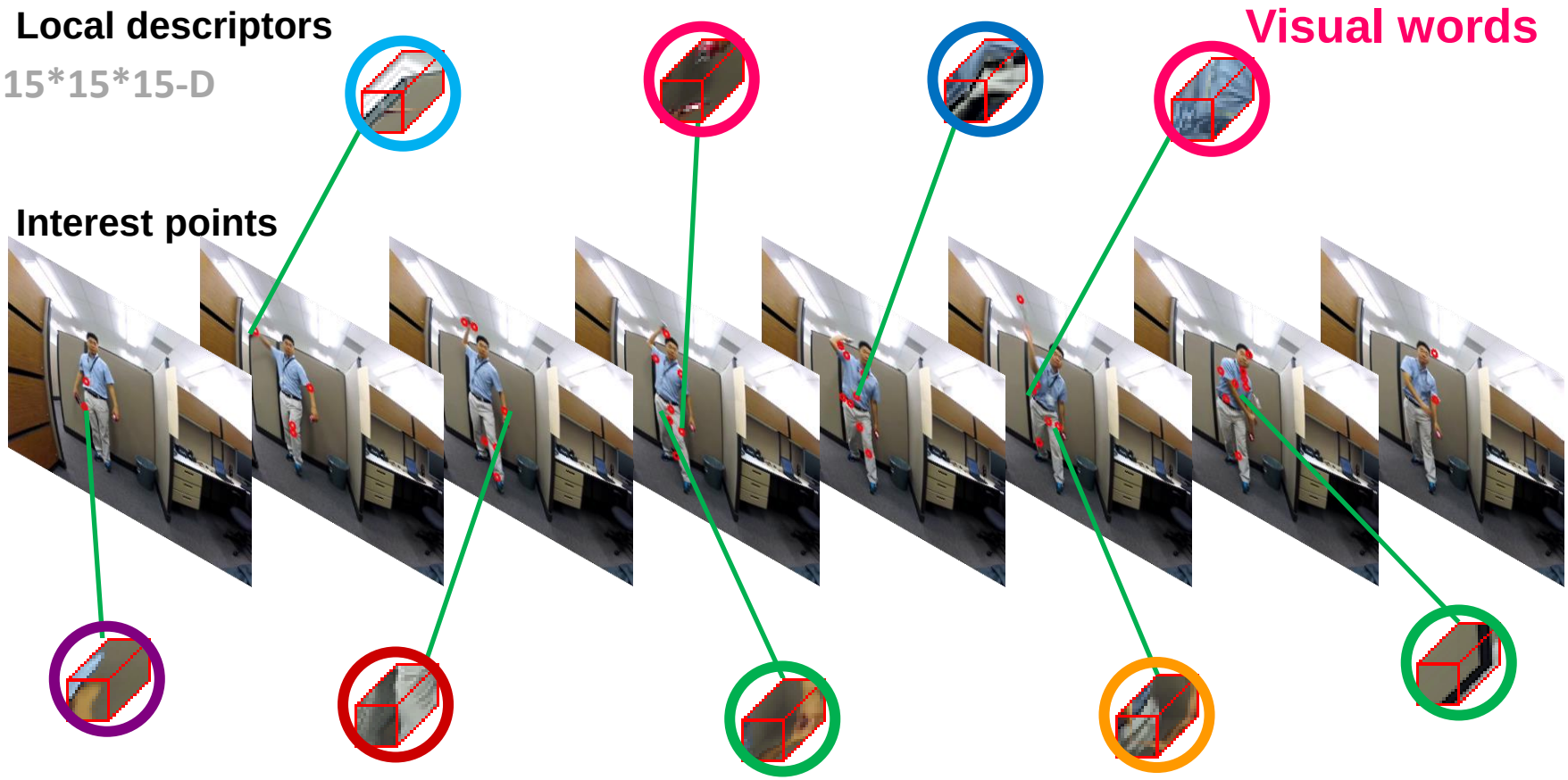
- Camera motion caused by human (e.g., shaking, hugging,…)



# Features for first-person videos (2)

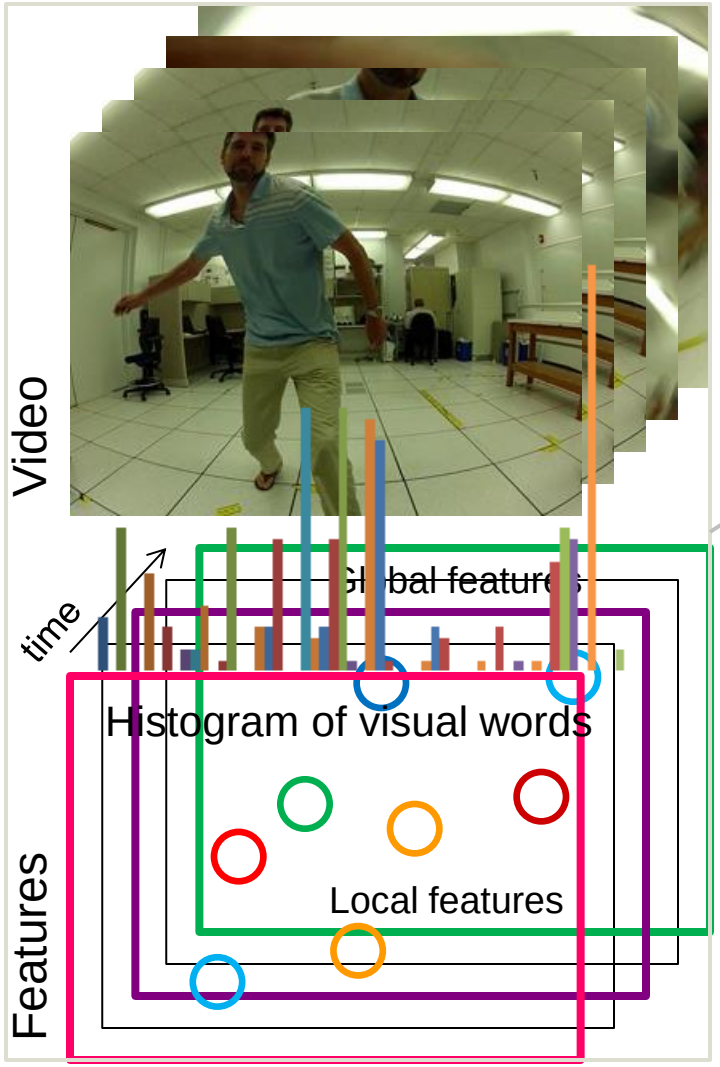
## Local motion descriptors

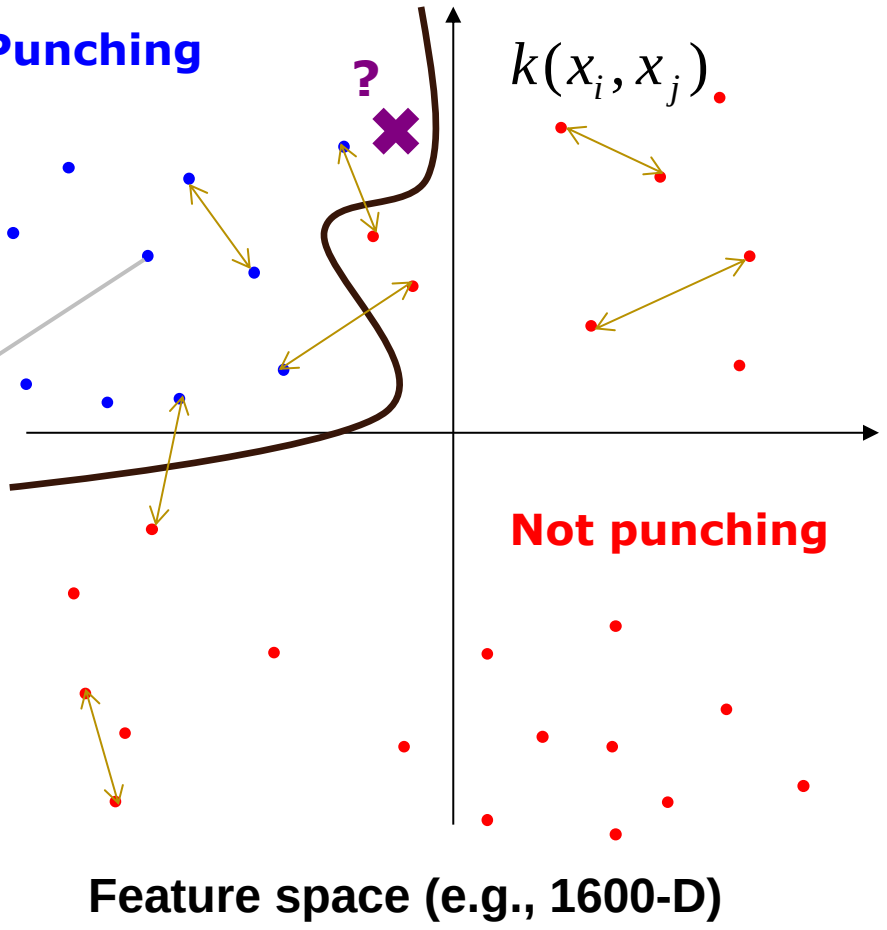
- Movements of the other person (e.g., pointing, throwing,…)



# Basic classification framework

***Bag-of-words*: ignores feature locations**



# Multi-channel kernels

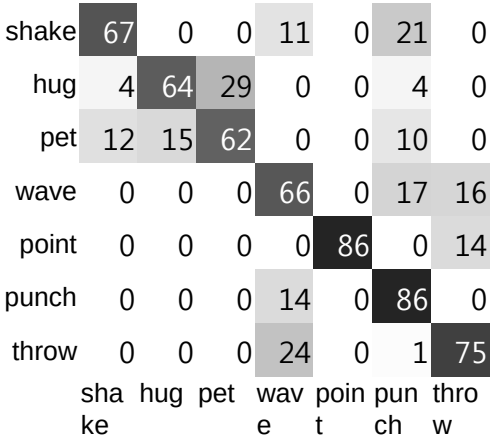
## Multi-channel kernels

- Combines two different motion descriptors
  - Histogram intersection kernel
  - Chi-Square kernel

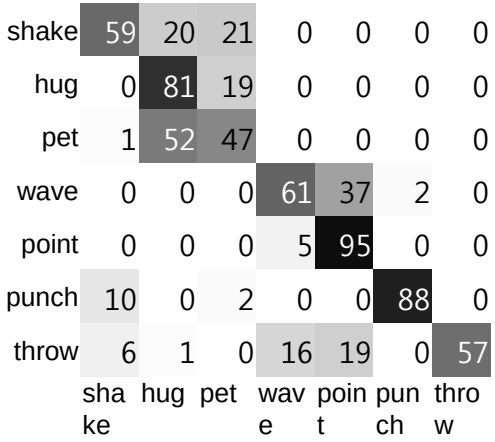
- $K(H_i, H_j) = \exp\left(-\sum_c D_c^h(H_i, H_j)\right)$

$$D_c^h(H_i, H_j) = \sum_{n=1}^{k} \frac{\min(h_{in}, h_{jn})}{\max(h_{in}, h_{jn})}$$

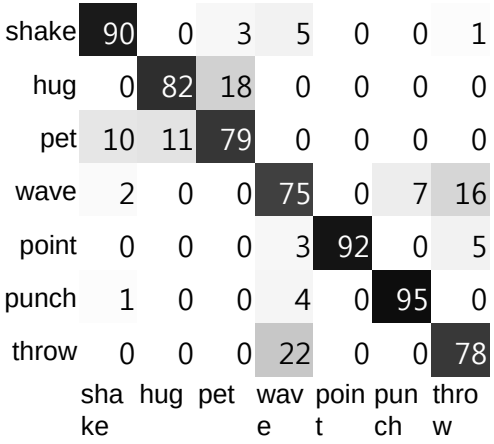
$$D_c^{\chi^2}(H_i, H_j) = \frac{1}{2 \cdot M_c} \sum_{n=1}^{k} \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}}$$



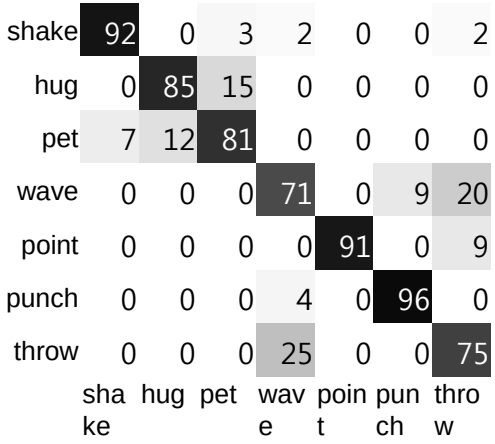

(a) Global descriptors




(b) Local descriptors




(c) Histogram intersection




(d) $\chi^2$ kernel

# Recognition with structure

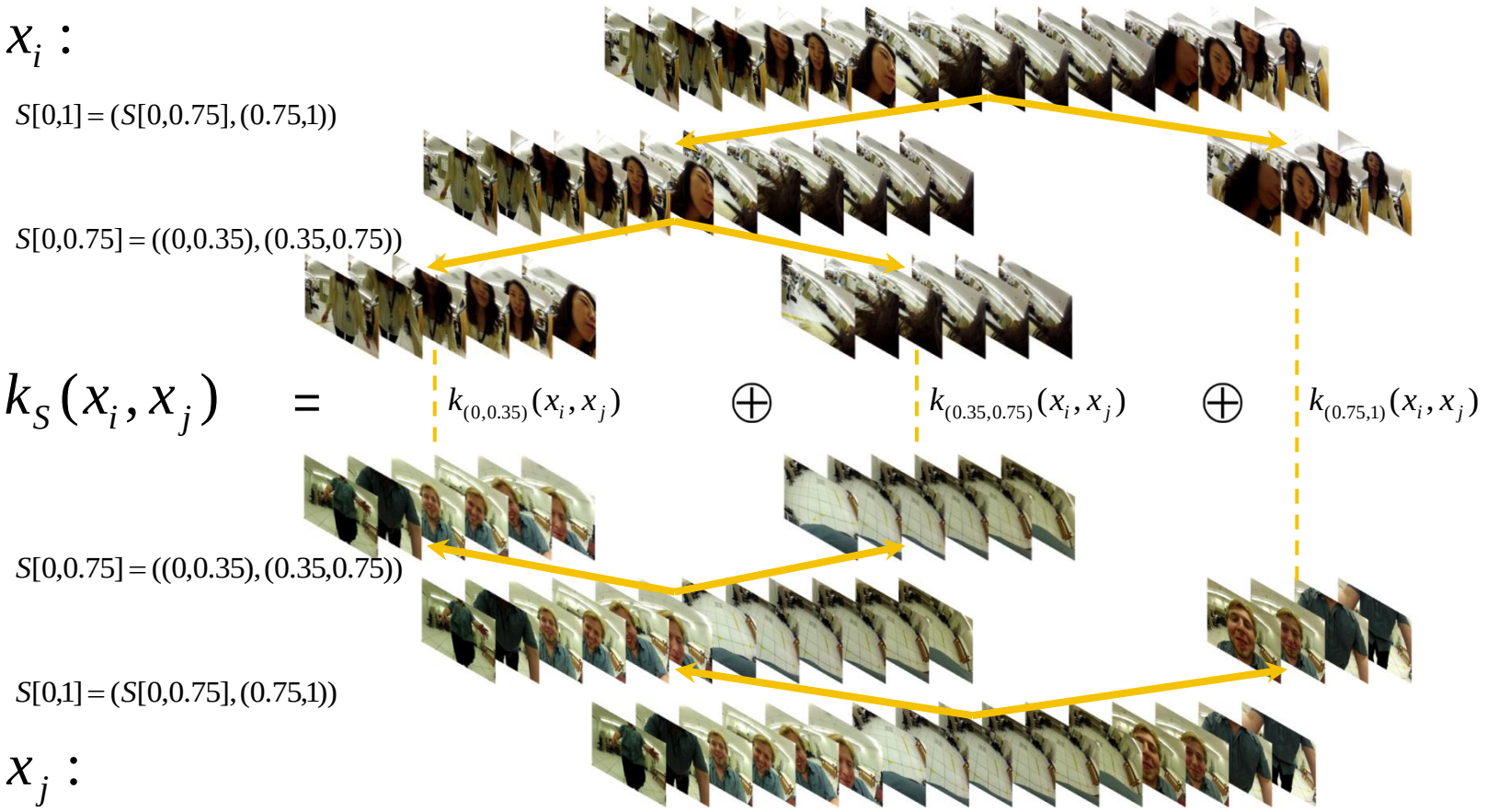
**In first-person videos, clear cause-effect relations are observed (e.g., punching).**

- Problems of ***bag-of-words***
  - Histogram-based representation
  - Ignores spatial/temporal locations of detected features
- Temporal structure
  - Spatio-temporal relationship match?
    - [Ryoo et al., ICCV 2009]
    - Limitation: only pairwise relations
  - A new kernel function considering 'structure' is needed.
    - What is the best 'sequential structure'?

# Structure match kernel

## Measure similarity between two videos

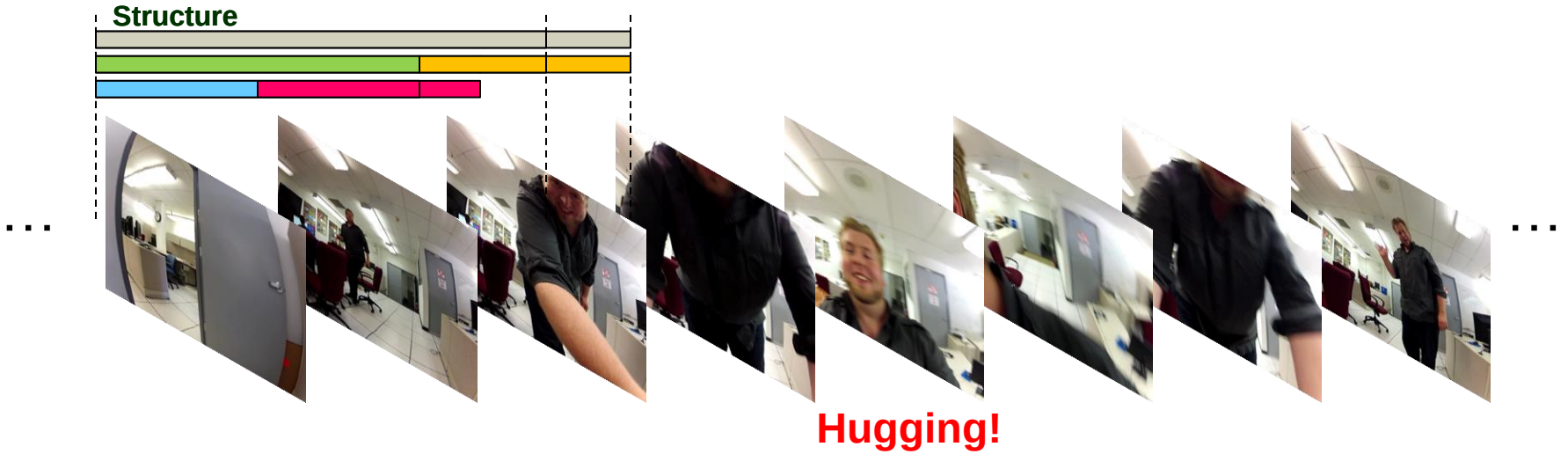
- E.g., S = (((0, 0.35), (0.35, 0.75)), (0.75, 1))



# Recognition from continuous videos

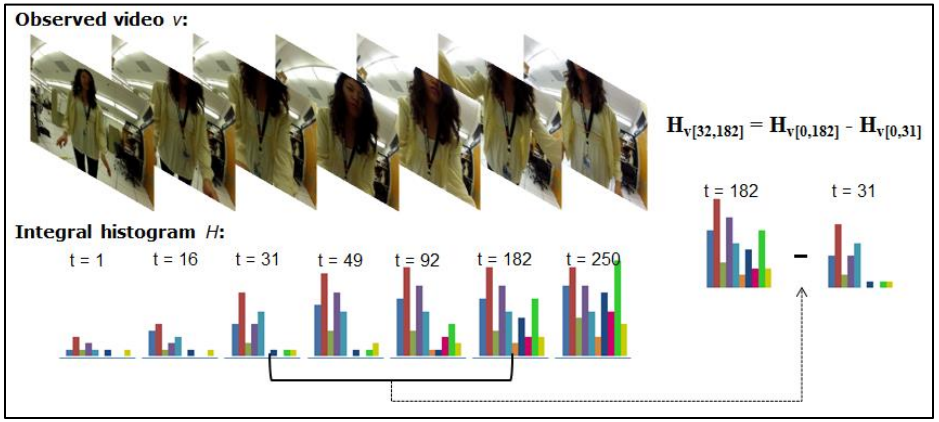
## Activity detection with structure match kernel

- Evaluation of all possible time intervals




- Integral histograms
  - For any video interval v[t1,t2], its histogram representation is:

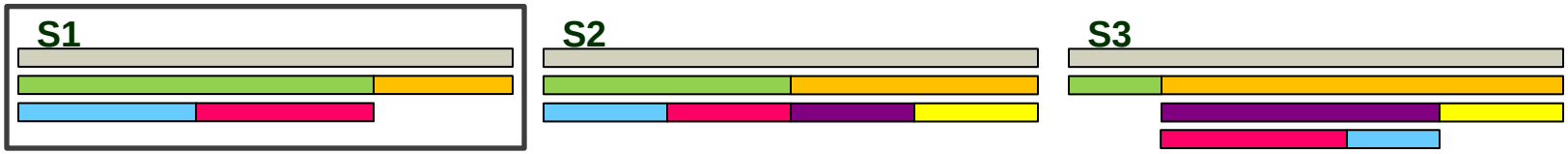
$$H_{v[t1,t2]} = H_{v[0,t2]} - H_{v[0,t1-1]}$$



# Activity structure learning

## We have many possible structure candidates *S*

- How do we find the best structure *S** that maximizes the performance?




## Evaluating 'goodness' of each structure *S*

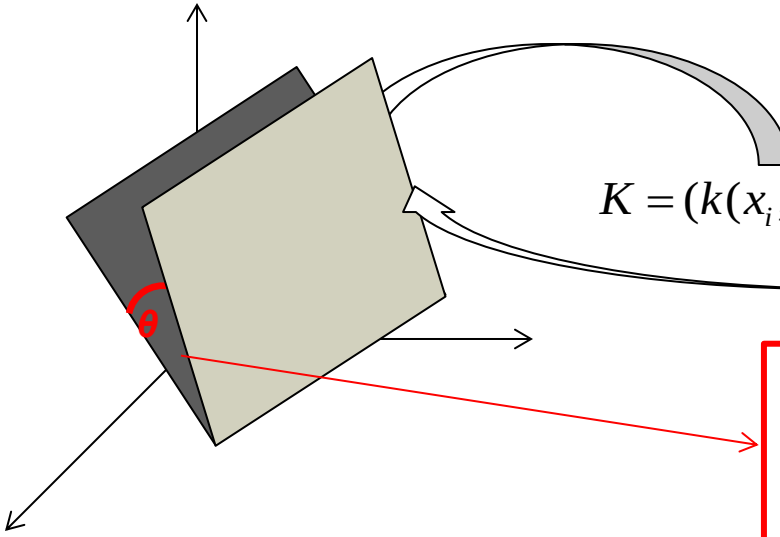
- Let $K_S$ be the Gram matrix describing the kernel distance (i.e., $k_S(x_i, x_j)$), obtained from training videos
- Let *L* be the Gram matrix of the optimal distance function
  - Example:
    four training videos,
    labels are $y_1=y_2!=y_3=y_4$.
    **Are they similar?**

$$K_S = \begin{bmatrix} 0 & 0.3 & 0.7 & 0.9 \\ 0.3 & 0 & 0.6 & 0.7 \\ 0.7 & 0.6 & 0 & 0.2 \\ 0.9 & 0.7 & 0.2 & 0 \end{bmatrix} \qquad L = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

# Finding the optimal structure

## Kernel target alignment

- Measures distance between two gram matrices
- To find $S^*$, we evaluate similarity between each candidate $K_S$ and $L$: $S^* = \arg\max_S A(K_s, L)$




$$L = (l(x_i, x_j))_{i,j=1}^m, \quad l(i,j) = \begin{cases} 0 & y_i = y_j \\ 1 & otherwise \end{cases}$$

$$K = (k(x_i, x_j))_{i,j=1}^m$$

**Kernel target alignment**

$$A(K,L) = \frac{\langle K, L \rangle_F}{\sqrt{\langle K, K \rangle_F \langle L, L \rangle_F}}$$

$$where \quad \langle K_1, K_2 \rangle_F = \sum_{i,j=1}^m k_1(x_i, x_j) k_2(x_i, x_j)$$

Frobenius inner product

# Experimental results

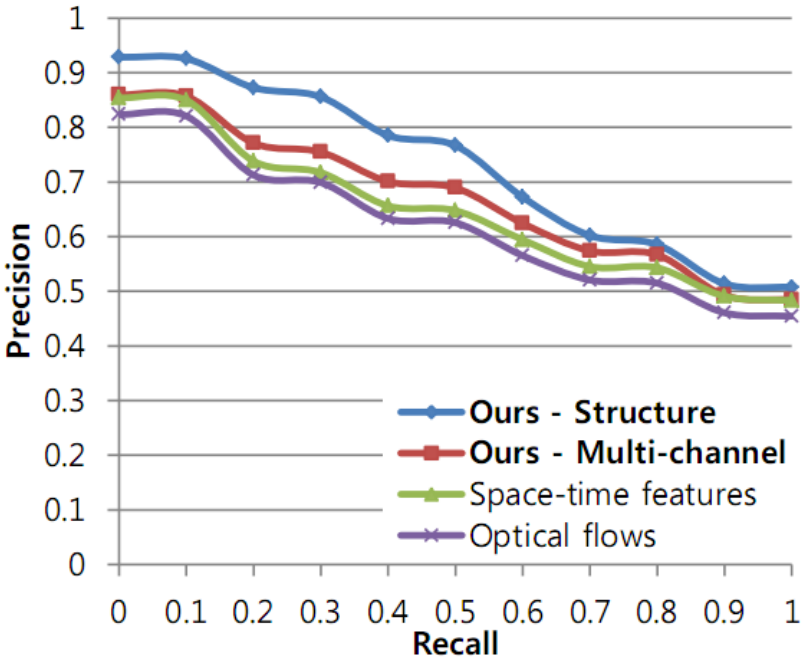
## Classification

- Random half-half training-testing split
  validation (12 sets)
- 7 classes, 84 videos

| Approach | Local feature | Both features |
|---|---|---|
| Histogram intersection | 82.4 % | 84.3 % |
| Chi-Square kernel | 82.4 % | 84.4 % |
| ST-Pyramid match [Choi et al. 2008] | 82.6 % | 86.0 % |
| Dynamic BoW [Ryoo 2011] | 82.8 % | 87.1 % |
| **Structure match** | **83.1 %** | **89.6 %** |

## Detection

- Half-half test/train split
  random validation
- 57 continuous videos
  - 0~3 activities per video



# Intermediate summary

**First-person *interaction* recognition is important.**

- New dataset taken with a humanoid robot model

**What features are necessary?**

- Global/local motion descriptors
- Multi-channel kernels

**Consideration of activities' structures?**

- Structure match kernel

# What's next?

**More features**

- A first-person video contains both global and local motion
  - Multiple types of features are required
- Human postures?
- Hand features?
- Gaze features?
- The more the better

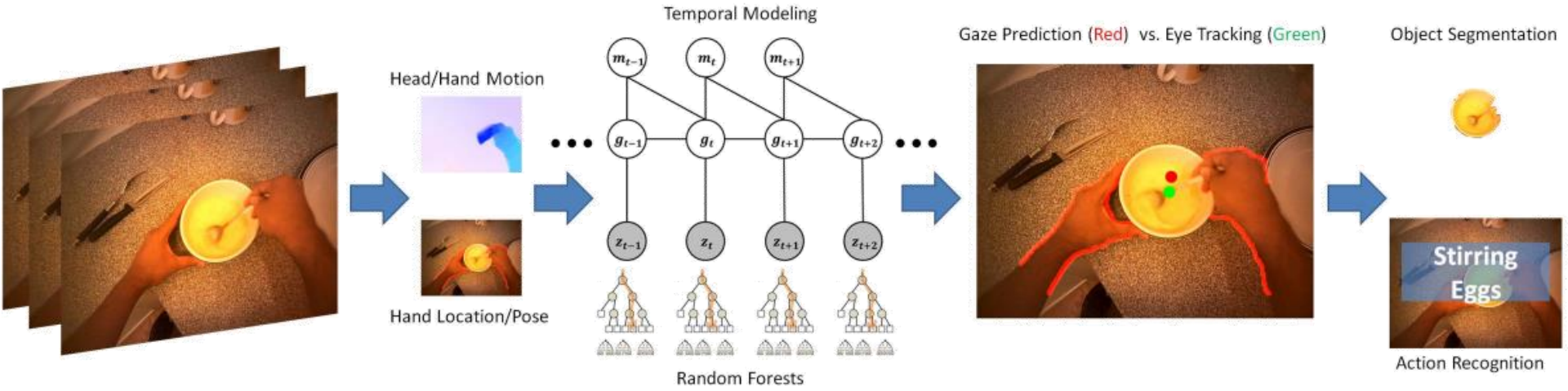
**Various types of videos**

- What's "ego" in "egocentric videos"?
- Humans and robots
- Cars? Animals?

# Features for first-person activity videos

# Gaze features

**[Li, Fathi, Rehg, ICCV 2013]**




**Gaze prediction based on first-person videos**

- Head motion and hand location
  - Gaze-hand coordination
- Dynamic behavior of gaze

**Benefitting object/ego-action recognition using predicted gaze**

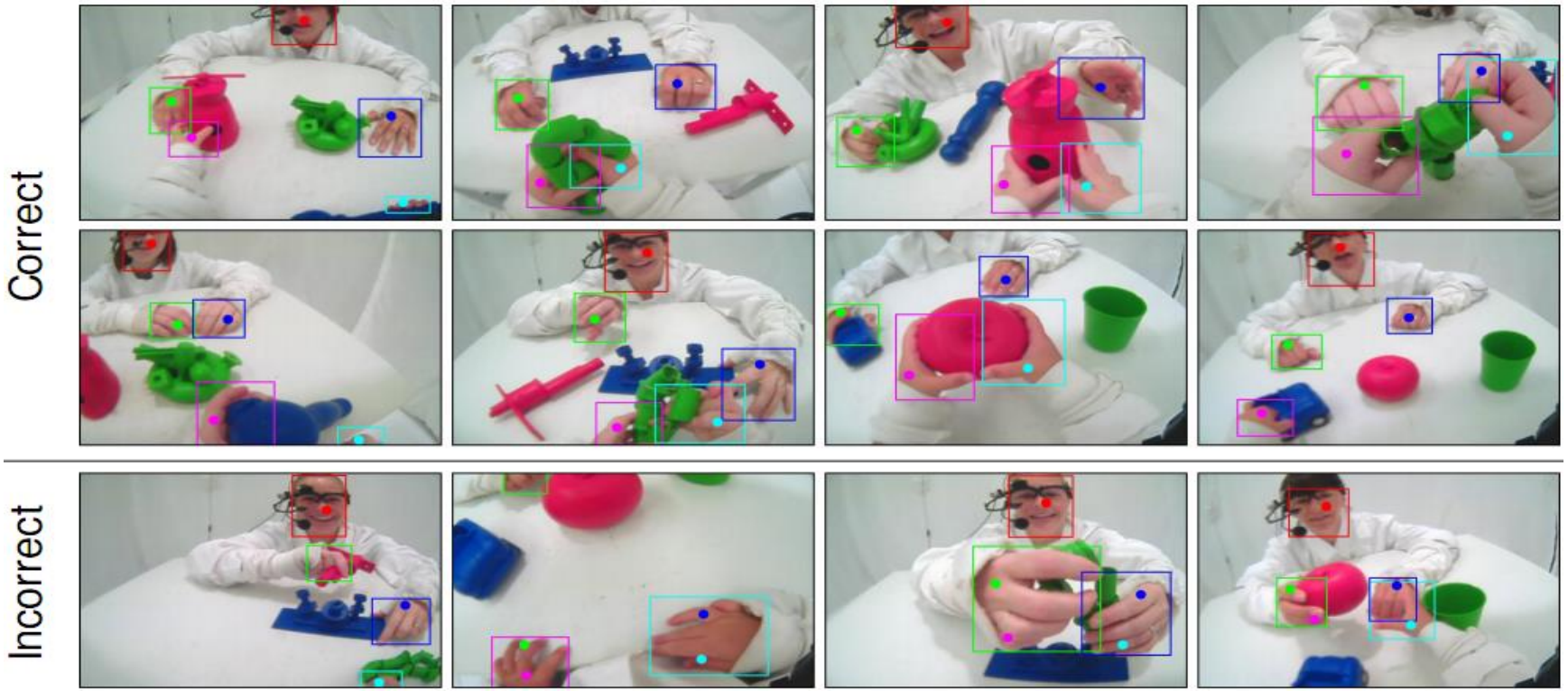
- Foreground object segmentation

# Hand features

**[Bambach, Lee, Crandall, Franchak, Yu, Egocentric Vision 2014]**

**Hands play important role in understanding first-person videos**

- Hands of the wearer, and hands of the interacting person
  - A graphical model considering Spatial prior, skin model, face model, and arm estimation.



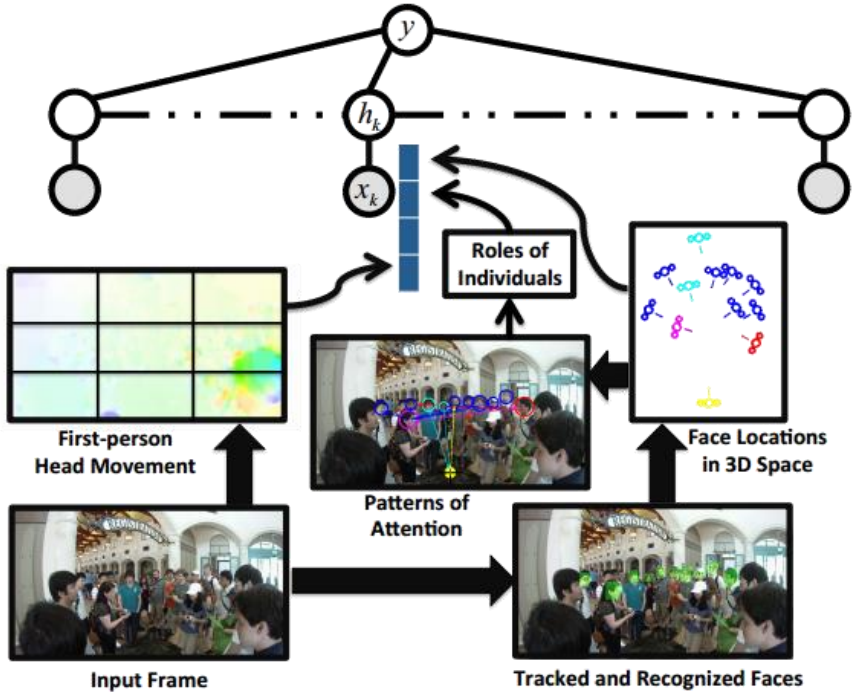
# Various types of first-person videos

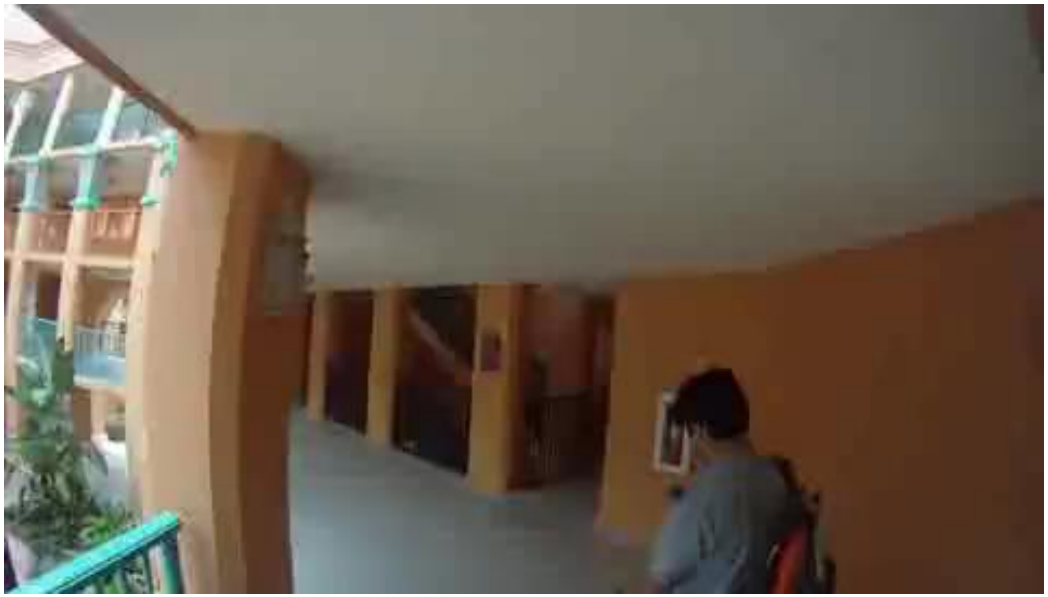
# First-person videos of social interactions

**[Fathi, Hodgins, Rehg, CVPR 2012]**

**Social interactions**

- 6 classes: dialogue, discussion, monologue, walk dialogue, walk discussion, others



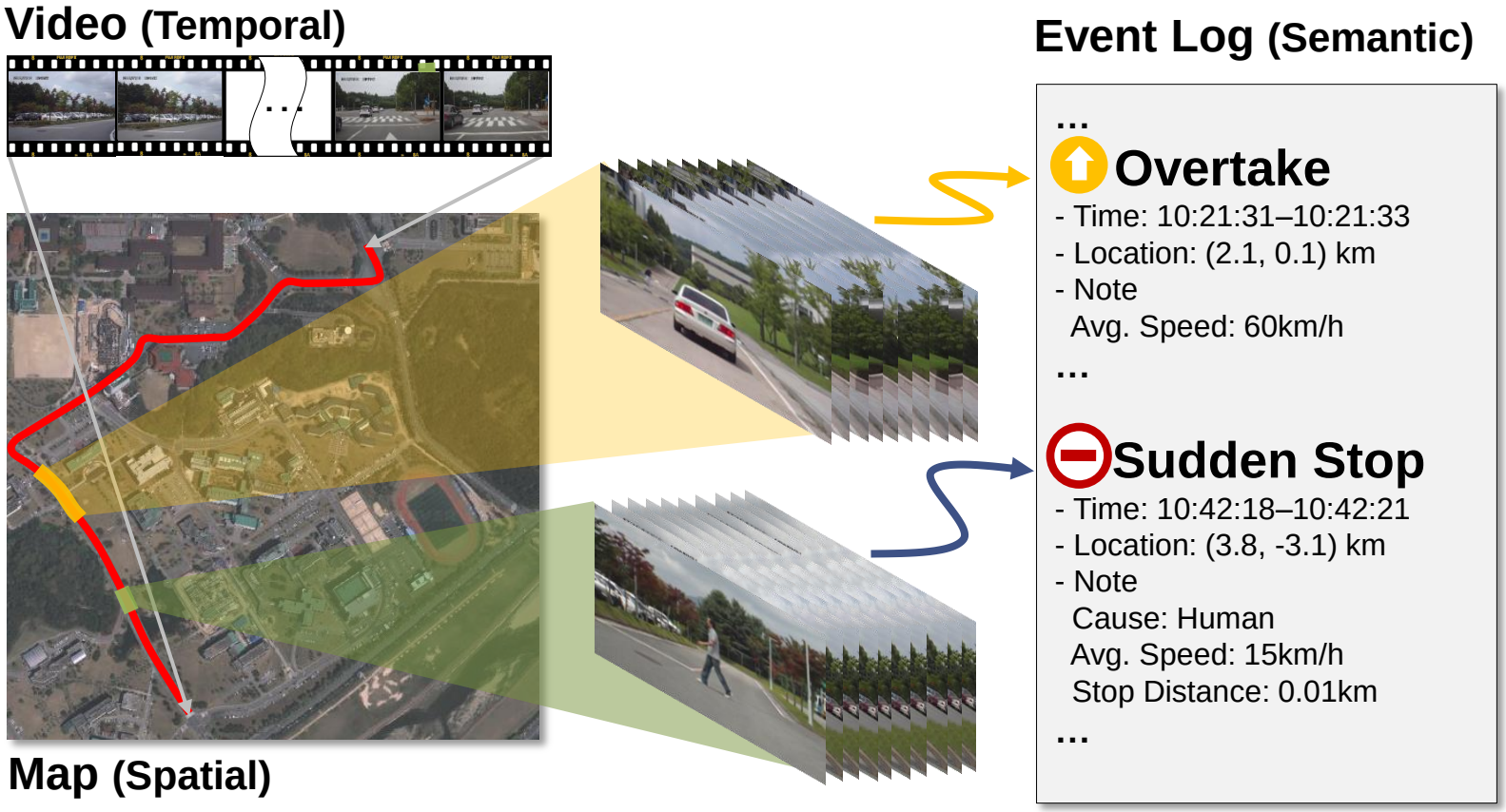
**People attentions**

- Feature: where are these people looking at?

# First-person driving videos

**[Ryoo et al., CVIU 2013]**

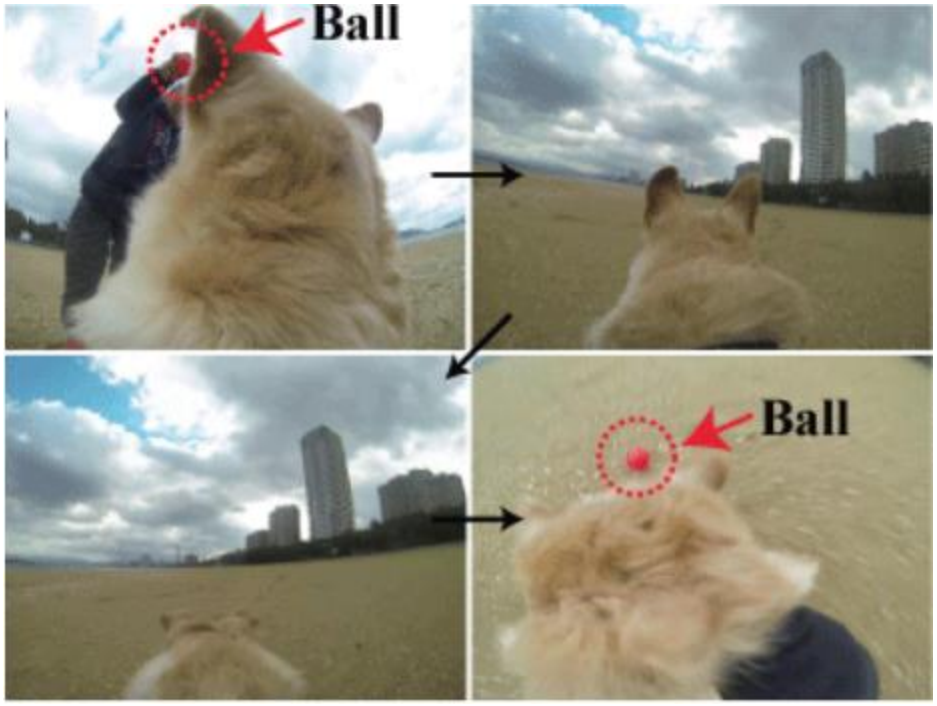
- Videos taken from a vehicle-mounted camera



# First-person animal videos

**[Iwashita, Takamine, Kurazume, Ryoo, ICPR 2014]**

- Activities from a point of view of a ‘dog’
  - Pet/wildlife animal monitoring
  - Animal experiments – extremely heavy ego-motion




# First-person animal videos

**[Iwashita, Takamine, Kurazume, Ryoo, ICPR 2014]**

- Activities from a point of view of a ‘dog’
  - Pet/wildlife animal monitoring
  - Animal experiments – extremely heavy ego-motion

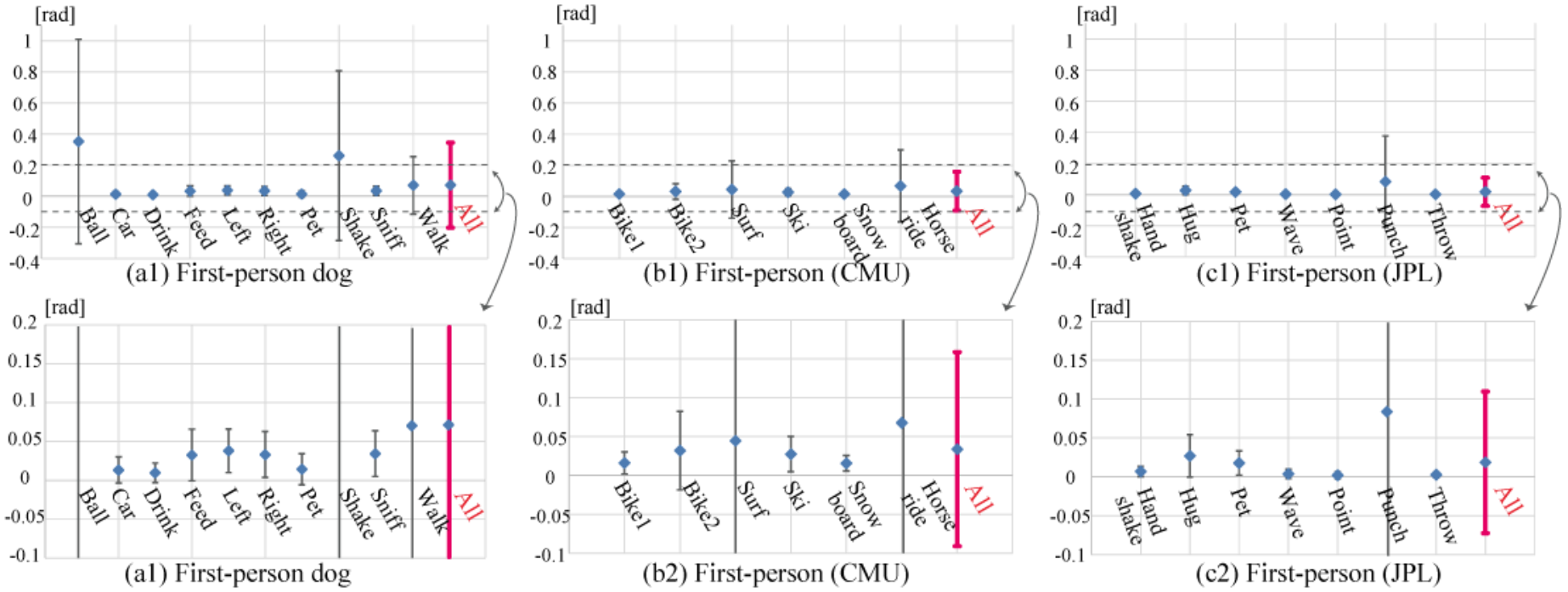
**“DogCentric Activity Dataset”**

4 different dogs, 10 activities

# First-Person Animal Activity Recognition

## Characteristics of the dataset

- Mixture of heavy ego-motion and low ego-motion activities
- Extreme amount of ego-motion in certain activities
- High motion variance




Mean and s.d. of **rotation angle** between frames

# Experiments with baseline approaches

## A repeated random sub-sampling validation

1. Randomly select half video sequence as training dataset
2. Use the rest of sequences for the testing
3. Repeat 1~2 for 100 times

| | Dog A | Dog B | Dog C | Dog D | Total (category) |
|---|---|---|---|---|---|
| Ball play | 6 | 5 | 3 | 0 | 14 |
| Car | 7 | 1 | 14 | 4 | 26 |
| Drink | 5 | 2 | 2 | 1 | 10 |
| Feed | 7 | 3 | 8 | 7 | 25 |
| Turn head (left) | 8 | 4 | 3 | 6 | 21 |
| Turn head (right) | 6 | 3 | 4 | 5 | 18 |
| Pet | 8 | 4 | 8 | 5 | 25 |
| Body shake | 9 | 2 | 3 | 5 | 19 |
| Sniff | 8 | 7 | 7 | 5 | 27 |
| Walk | 7 | 4 | 7 | 7 | 25 |
| Total (dog) | 71 | 35 | 59 | 45 | **210** |

| | Classification accuracy [%] |
|---|---|
| Linear kernel | 52.6 |
| RBF kernel | 54.2 |
| $\chi^2$kernel | 60.5 |
| Histogram intersection | 57.3 |

# DogCentric Activity Dataset

## A dataset to test heavy ego-motion

| | Ball | Car | Drink | Feed | Left | Right | Pet | Shake | Sniff | Walk |
|---|---|---|---|---|---|---|---|---|---|---|
| Ball | 19.4 | 0.0 | 0.0 | 4.0 | 0.0 | 4.0 | 2.9 | 17.7 | 23.1 | 28.9 |
| Car | 0.0 | 74.7 | 5.3 | 10.5 | 1.3 | 0.0 | 8.2 | 0.0 | 0.0 | 0.0 |
| Drink | 0.4 | 34.4 | 25.6 | 12.4 | 3.2 | 0.0 | 12.0 | 0.0 | 8.0 | 4.0 |
| Feed | 0.0 | 23.0 | 0.5 | 22.8 | 12.7 | 2.2 | 20.3 | 0.0 | 12.2 | 6.3 |
| Left | 0.0 | 21.6 | 0.0 | 22.8 | 23.4 | 11.0 | 2.6 | 0.0 | 5.8 | 12.8 |
| Right | 0.0 | 31.1 | 0.2 | 18.2 | 9.6 | 20.2 | 8.0 | 0.4 | 8.2 | 4.0 |
| Pet | 0.0 | 19.3 | 0.8 | 26.0 | 3.7 | 1.7 | 39.7 | 0.0 | 8.8 | 0.0 |
| Shake | 0.2 | 0.7 | 0.0 | 5.6 | 5.1 | 14.4 | 3.3 | 65.6 | 0.7 | 4.4 |
| Sniff | 0.0 | 0.5 | 0.0 | 9.2 | 7.7 | 5.5 | 9.8 | 2.8 | 59.5 | 4.9 |
| Walk | 0.0 | 0.5 | 0.0 | 5.0 | 18.5 | 8.0 | 3.0 | 15.5 | 3.0 | 46.5 |

(a) Optical flow

| | Ball | Car | Drink | Feed | Left | Right | Pet | Shake | Sniff | Walk |
|---|---|---|---|---|---|---|---|---|---|---|
| Ball | 27.7 | 1.1 | 2.0 | 17.1 | 2.6 | 2.9 | 5.7 | 18.0 | 15.4 | 7.4 |
| Car | 3.5 | 51.7 | 0.3 | 3.3 | 11.7 | 14.5 | 1.8 | 0.5 | 5.7 | 7.0 |
| Drink | 6.8 | 6.4 | 1.2 | 7.2 | 5.6 | 2.8 | 32.0 | 2.0 | 24.4 | 11.6 |
| Feed | 2.0 | 7.3 | 0.2 | 38.2 | 16.3 | 10.3 | 9.3 | 8.8 | 0.5 | 7.0 |
| Left | 1.0 | 2.8 | 0.2 | 7.4 | 25.4 | 21.0 | 18.2 | 11.8 | 4.0 | 8.2 |
| Right | 0.0 | 2.4 | 0.0 | 2.9 | 20.0 | 49.1 | 8.2 | 4.0 | 3.8 | 9.6 |
| Pet | 4.0 | 3.0 | 2.2 | 12.8 | 14.7 | 7.8 | 23.5 | 11.2 | 8.5 | 12.3 |
| Shake | 2.4 | 3.6 | 0.0 | 12.4 | 15.1 | 8.9 | 15.3 | 30.2 | 2.4 | 9.6 |
| Sniff | 6.6 | 4.9 | 2.8 | 5.2 | 6.6 | 6.2 | 11.1 | 4.6 | 43.1 | 8.9 |
| Walk | 6.0 | 19.0 | 1.0 | 1.5 | 15.0 | 15.5 | 3.0 | 1.5 | 8.5 | 29.0 |

(b) LBP

| | Ball | Car | Drink | Feed | Left | Right | Pet | Shake | Sniff | Walk |
|---|---|---|---|---|---|---|---|---|---|---|
| Ball | 43.4 | 0.0 | 0.0 | 3.4 | 3.7 | 1.7 | 5.4 | 14.3 | 14.3 | 13.7 |
| Car | 0.0 | 84.8 | 0.8 | 1.3 | 3.0 | 3.0 | 7.0 | 0.0 | 0.0 | 0.0 |
| Drink | 2.4 | 11.2 | 42.0 | 3.6 | 2.0 | 16.8 | 18.8 | 0.4 | 1.2 | 1.6 |
| Feed | 0.7 | 3.0 | 1.2 | 34.0 | 21.7 | 11.2 | 13.2 | 0.0 | 7.8 | 7.3 |
| Left | 0.0 | 4.4 | 0.0 | 12.4 | 41.4 | 21.2 | 5.6 | 2.0 | 2.6 | 10.4 |
| Right | 0.2 | 2.7 | 1.3 | 4.0 | 25.3 | 33.8 | 15.3 | 3.3 | 8.4 | 5.6 |
| Pet | 0.0 | 14.0 | 3.2 | 9.7 | 10.5 | 4.2 | 50.5 | 0.5 | 7.3 | 0.2 |
| Shake | 0.7 | 0.0 | 0.0 | 0.2 | 7.8 | 9.6 | 0.0 | 73.3 | 3.3 | 5.1 |
| Sniff | 0.5 | 0.0 | 0.8 | 3.7 | 4.3 | 4.2 | 2.0 | 3.4 | 68.2 | 13.1 |
| Walk | 0.5 | 0.0 | 1.5 | 0.5 | 2.0 | 3.5 | 0.0 | 7.5 | 1.5 | 83.0 |

(c) Cuboids

| | Ball | Car | Drink | Feed | Left | Right | Pet | Shake | Sniff | Walk |
|---|---|---|---|---|---|---|---|---|---|---|
| Ball | 40.3 | 0.0 | 0.0 | 7.1 | 2.9 | 0.0 | 0.6 | 16.9 | 12.3 | 20.0 |
| Car | 0.0 | 85.7 | 3.0 | 5.0 | 0.0 | 0.8 | 5.5 | 0.0 | 0.0 | 0.0 |
| Drink | 6.8 | 24.4 | 8.0 | 8.8 | 6.8 | 25.6 | 8.0 | 0.0 | 5.2 | 6.4 |
| Feed | 2.0 | 7.7 | 0.7 | 39.2 | 12.5 | 11.5 | 7.2 | 1.3 | 12.7 | 5.3 |
| Left | 0.0 | 6.4 | 4.6 | 12.6 | 37.0 | 21.6 | 0.6 | 0.4 | 1.6 | 15.2 |
| Right | 0.0 | 6.2 | 2.9 | 7.1 | 9.3 | 45.3 | 8.9 | 0.0 | 17.1 | 3.1 |
| Pet | 0.0 | 6.7 | 2.3 | 16.8 | 3.5 | 6.3 | 31.7 | 8.7 | 22.0 | 2.0 |
| Shake | 2.2 | 0.0 | 0.0 | 4.2 | 0.9 | 1.6 | 2.4 | 86.9 | 0.2 | 1.6 |
| Sniff | 2.0 | 3.2 | 0.2 | 12.8 | 7.4 | 8.0 | 12.2 | 0.0 | 49.5 | 4.8 |
| Walk | 1.5 | 0.0 | 0.0 | 0.5 | 5.5 | 0.0 | 0.0 | 5.0 | 0.0 | 87.5 |

(d) STIP (HOF)

| | Ball | Car | Drink | Feed | Left | Right | Pet | Shake | Sniff | Walk |
|---|---|---|---|---|---|---|---|---|---|---|
| Ball | 48.6 | 0.0 | 0.0 | 3.7 | 2.6 | 0.0 | 0.9 | 18.3 | 13.1 | 12.9 |
| Car | 0.0 | 86.2 | 0.0 | 0.7 | 2.0 | 5.3 | 5.8 | 0.0 | 0.0 | 0.0 |
| Drink | 0.0 | 17.2 | 24.0 | 6.0 | 4.4 | 28.8 | 12.0 | 0.0 | 1.2 | 6.4 |
| Feed | 0.2 | 1.7 | 0.0 | 47.7 | 16.5 | 7.5 | 14.0 | 1.0 | 7.0 | 4.5 |
| Left | 0.0 | 4.6 | 0.0 | 10.8 | 42.2 | 22.0 | 3.0 | 0.6 | 0.0 | 16.8 |
| Right | 0.0 | 6.0 | 0.0 | 3.6 | 13.3 | 49.1 | 15.6 | 0.0 | 8.0 | 4.4 |
| Pet | 0.0 | 7.5 | 0.5 | 15.8 | 9.2 | 4.7 | 53.8 | 0.0 | 8.3 | 0.2 |
| Shake | 1.1 | 0.0 | 0.0 | 1.8 | 4.2 | 0.9 | 0.0 | 89.3 | 1.8 | 0.9 |
| Sniff | 3.2 | 0.5 | 0.0 | 4.8 | 7.4 | 7.5 | 5.1 | 0.0 | 69.1 | 2.5 |
| Walk | 0.0 | 0.0 | 0.0 | 0.0 | 8.0 | 3.0 | 0.0 | 2.0 | 1.0 | 86.0 |

(e) STIP (HOG)

| | Ball | Car | Drink | Feed | Left | Right | Pet | Shake | Sniff | Walk |
|---|---|---|---|---|---|---|---|---|---|---|
| Ball | 67.4 | 0.0 | 0.0 | 0.3 | 0.9 | 0.3 | 0.0 | 16.0 | 7.4 | 7.7 |
| Car | 0.0 | 74.3 | 2.5 | 2.0 | 2.7 | 7.0 | 11.3 | 0.0 | 0.2 | 0.0 |
| Drink | 0.0 | 12.4 | 18.8 | 19.6 | 10.8 | 12.4 | 14.4 | 0.0 | 4.4 | 7.2 |
| Feed | 0.0 | 1.5 | 1.8 | 46.2 | 19.5 | 1.3 | 13.8 | 1.7 | 12.3 | 1.8 |
| Left | 0.0 | 4.0 | 0.2 | 15.4 | 32.2 | 19.4 | 5.2 | 4.8 | 7.0 | 11.8 |
| Right | 0.0 | 7.1 | 1.1 | 5.1 | 24.2 | 35.6 | 16.0 | 0.7 | 7.3 | 2.9 |
| Pet | 1.2 | 10.2 | 0.5 | 18.3 | 9.8 | 2.8 | 44.3 | 1.8 | 7.8 | 3.2 |
| Shake | 2.9 | 0.0 | 0.0 | 0.9 | 13.3 | 0.2 | 3.6 | 53.8 | 11.3 | 14.0 |
| Sniff | 5.8 | 0.0 | 0.3 | 5.2 | 14.6 | 3.7 | 7.1 | 4.8 | 43.1 | 15.4 |
| Walk | 0.0 | 5.5 | 0.0 | 0.5 | 9.0 | 3.0 | 0.5 | 0.0 | 9.0 | 72.5 |

(f) ALL

# Summary

**Wearable cameras will be everywhere**

**A massive amount of 'private' data**

- First-person videos of your daily life (24 hours a day?)
- Manual annotation is difficult
    - Hours of personal data + privacy issues

**Automated recognition of activities from first-person videos**

# First-person activity prediction

**[Ryoo et al., arXiv 2014]**

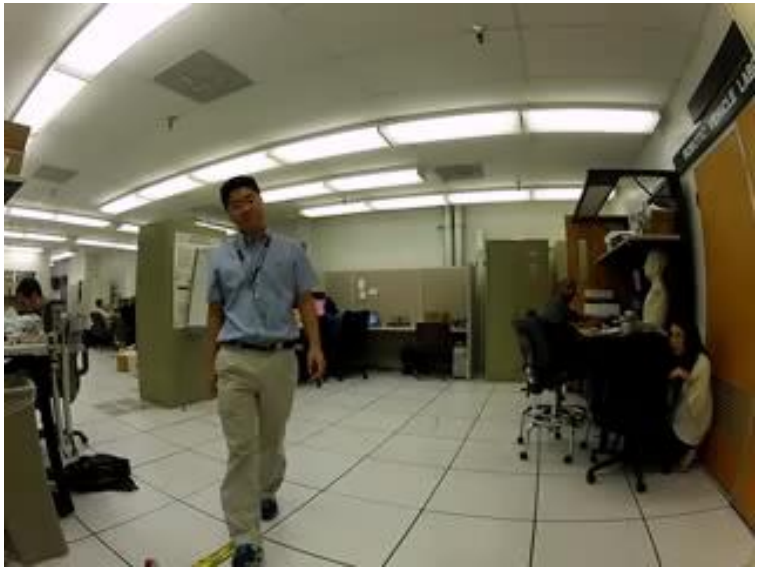
# Are we happy now?


**Early recognition is necessary**

# First-person activity prediction

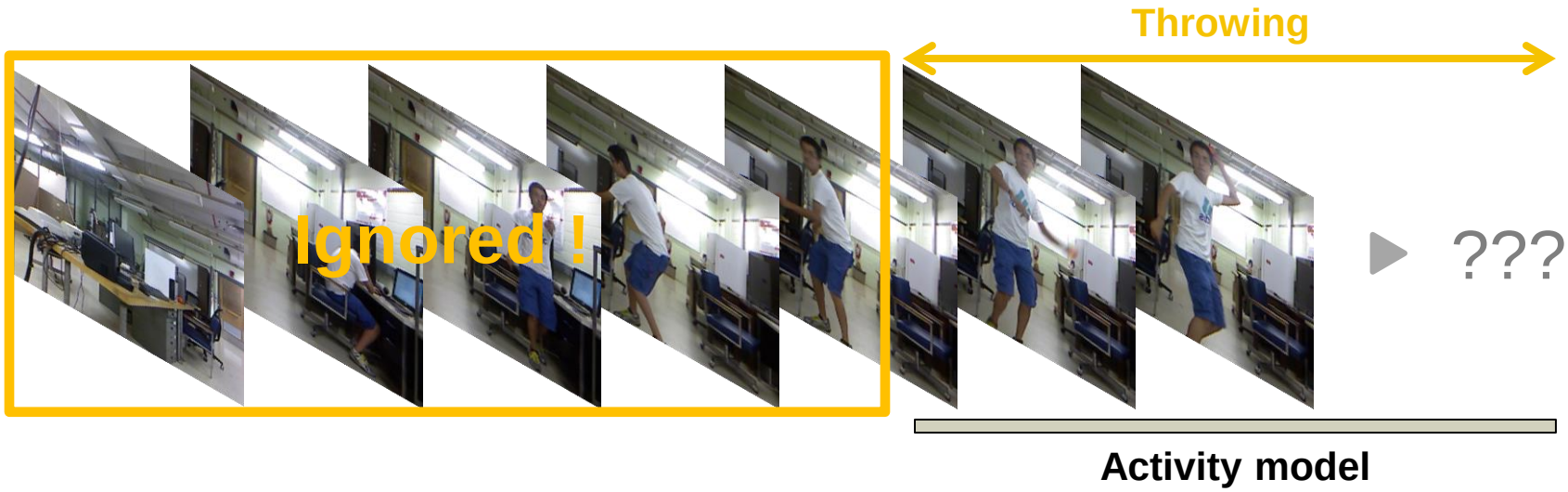
## Early recognition from continuous videos

- After-the-fact detection of activities is less important
  - E.g., recognizing a human assaulting a robot
- Fast real-time robot reaction
  - Avoiding damage

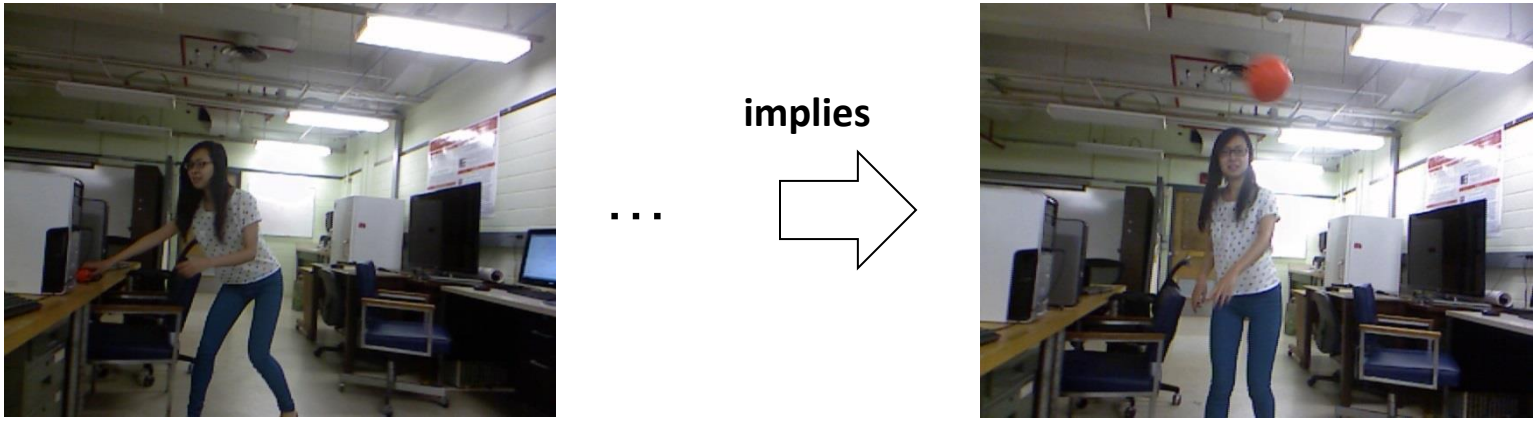


# Limitations

## Early recognition from continuous videos?




## We need to utilize pre-activity videos (onsets)



# To be continued…

**More details on activity ‘prediction’ or ‘early recognition’ will be discussed at the ‘future directions’ section of the tutorial.**