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

Emerging topics and directions

CVPR tutorial on 2014/06/23

Human activity prediction (i.e., early recognition)

[Ryoo, ICCV 2011]

Limitations of conventional paradigm

Most assume *after-the-fact* detection

- Classify after fully observing the video

Stealing happened in an Apple computer store

Even if the system detects crime, it may be too ***late*** to prevent it.

Human activity prediction

Early recognition from initial video streams

- Inference on ongoing/future activities from onsets

Punching

Pushing

Shaking hands

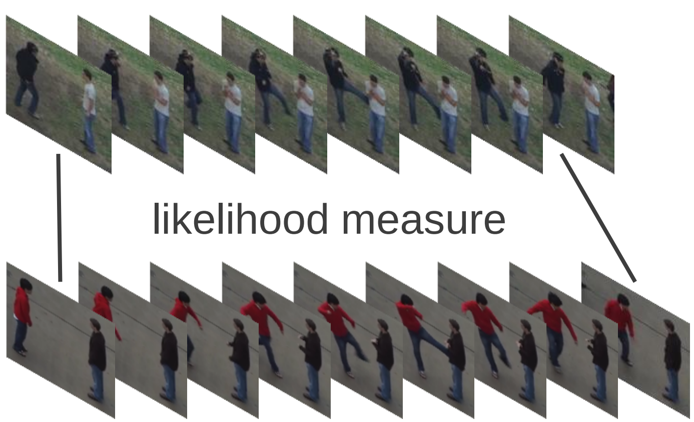
- Particularly important in surveillance scenarios
  - Must identify what it is ***before*** a harmful event occurs.
  - Stealing? Accident? Attack?

Problem formulation

Classification (previous)

- Assumes each video contains an entire activity
  - Activity is always fully progressed, d^*
  - $P(A \mid O, t) = P(A, d^* \mid O)$

Video observation O:

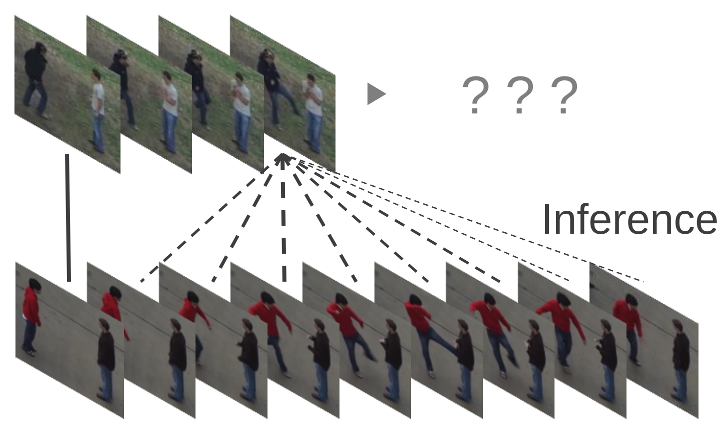


Activity model A:

Activity prediction

- Inference from an initial observation
  - Multiple possible activity progress level d
  - $P(A \mid O, t) = \sum_{d} P(A, d \mid O, t)$

Ongoing observation O:



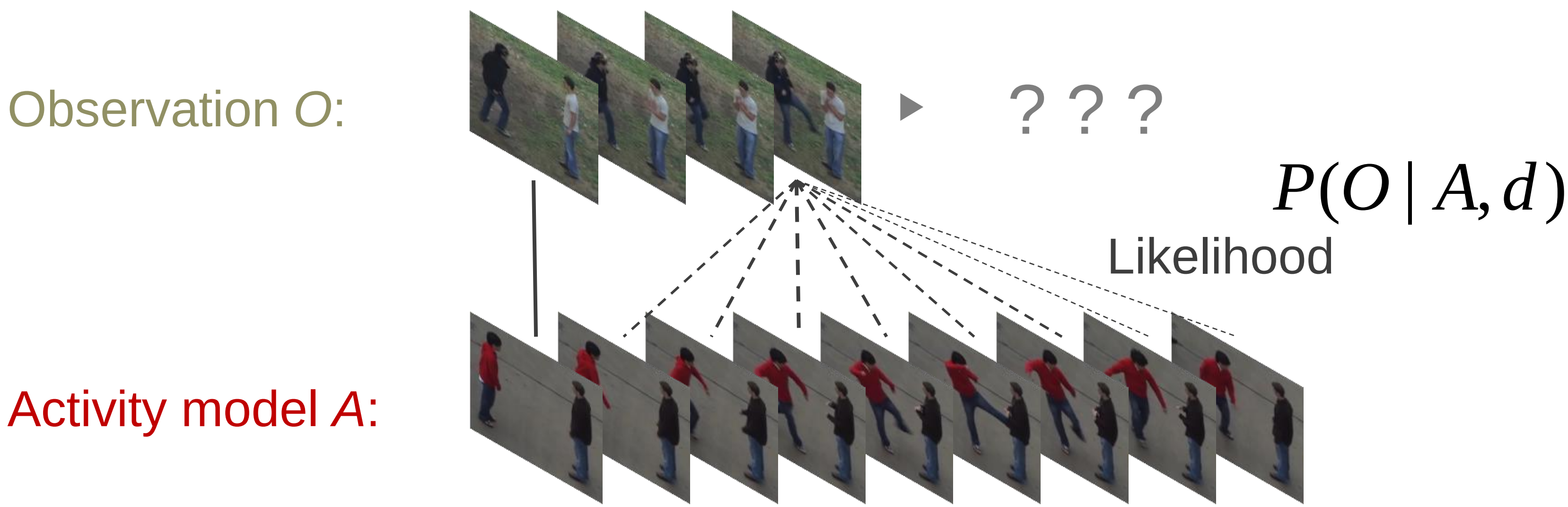
Activity model A:

Activity prediction formulation

Bayesian posterior probability

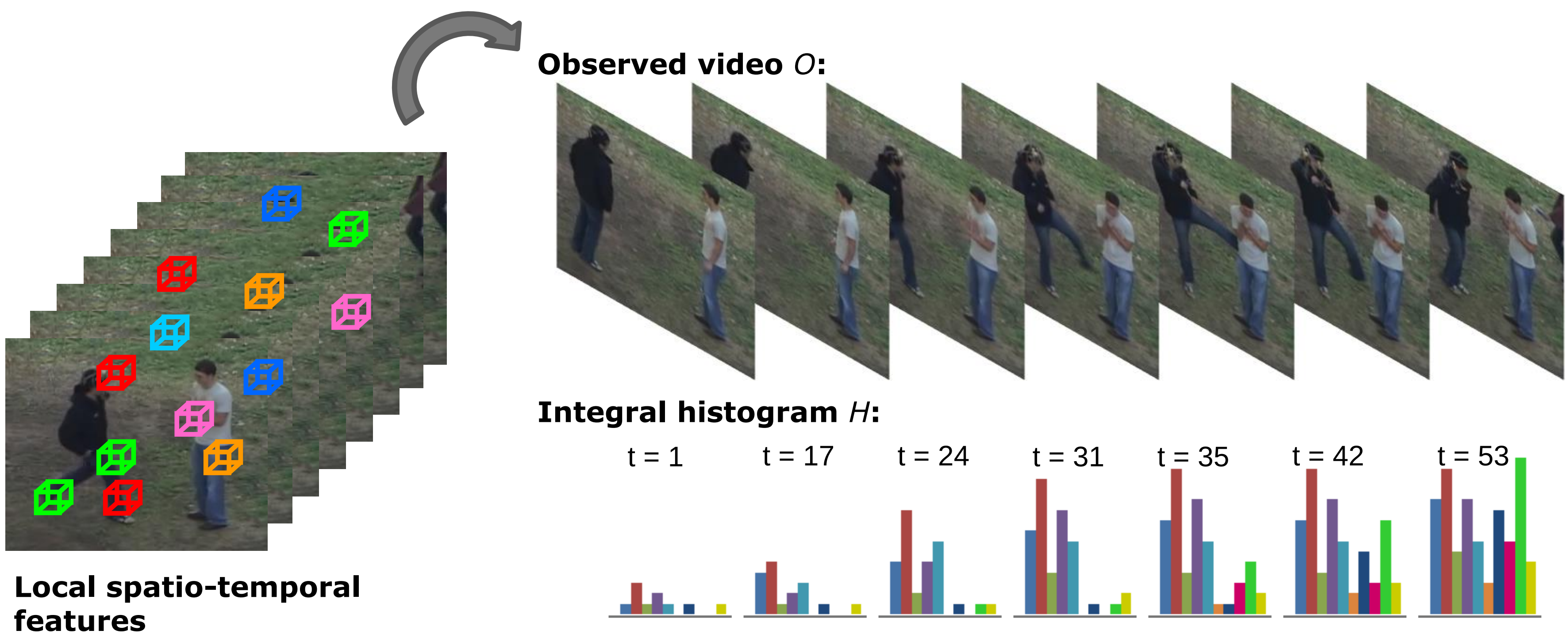
- $$P(A \mid O, t) = \sum_d P(A, d \mid O, t) = \frac{\sum_d P(O \mid A, d) P(t \mid d) P(A, d)}{\sum_i \sum_d P(O \mid A_i, d) P(t \mid d) P(A_i, d)}$$

Efficient computation of likelihood



Integral histogram

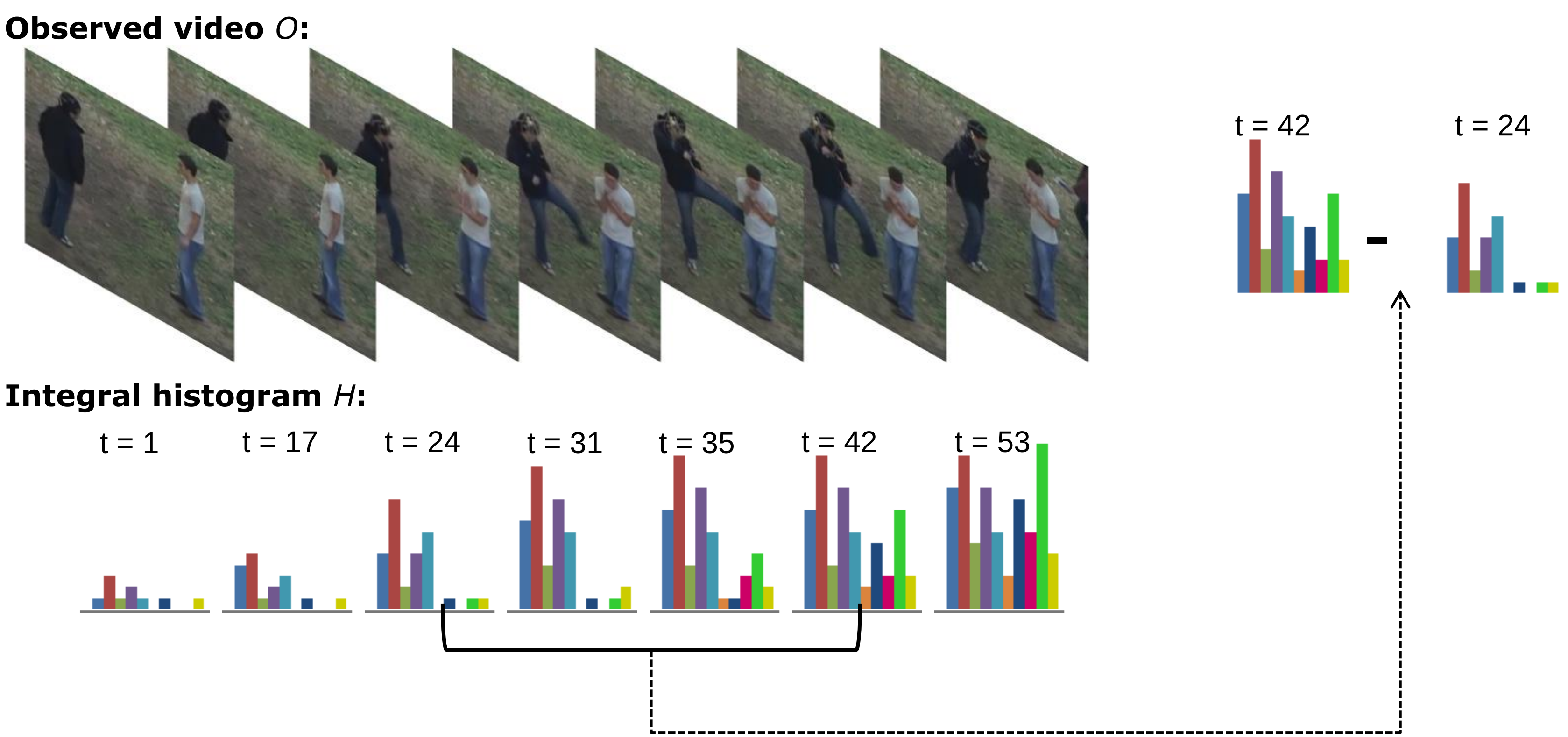
Enables efficient computation of feature histograms for any particular time interval:



Integral histogram

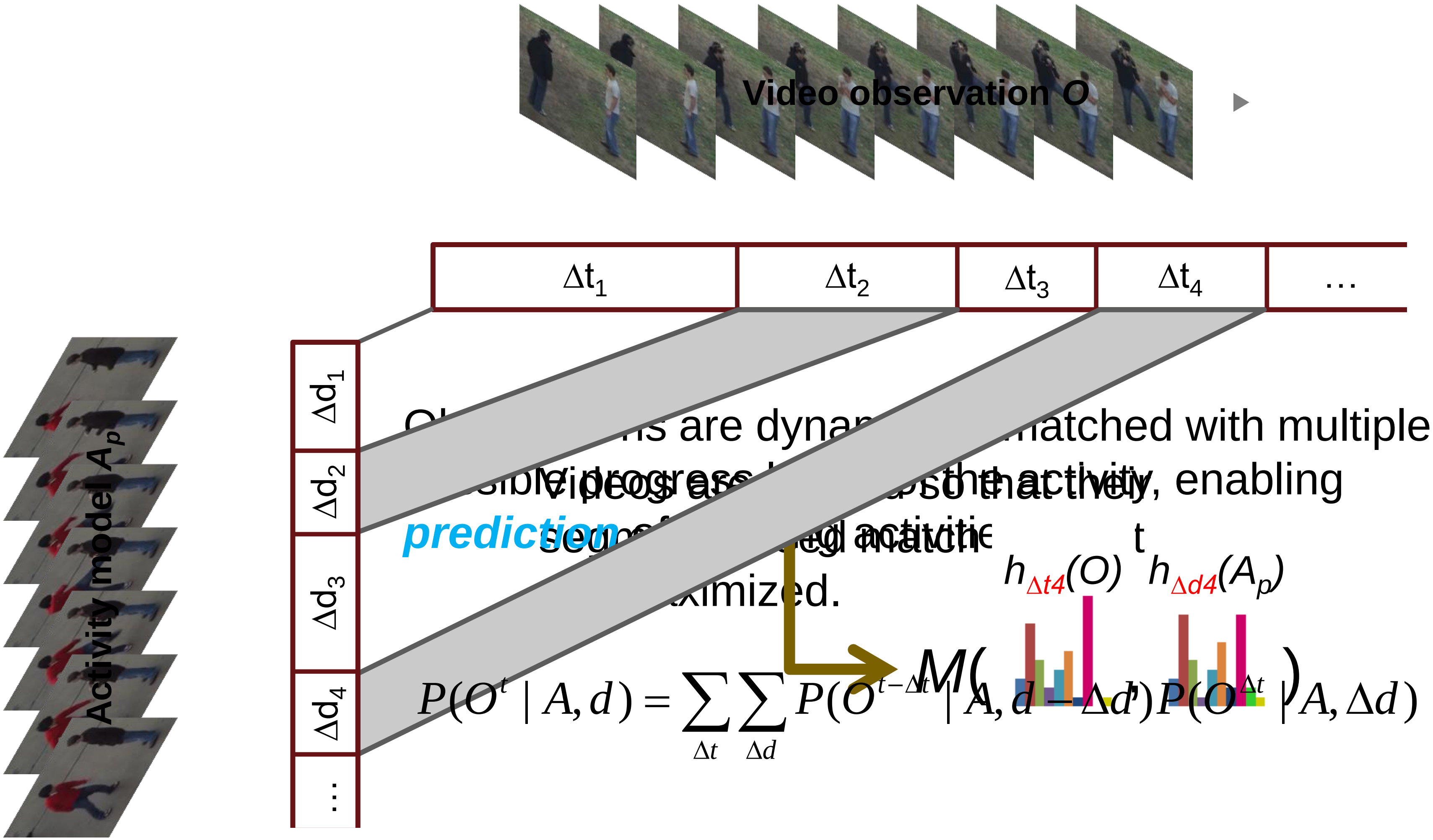
For any time interval [t1, t2]

- $$h_{[t1,t2]}(A) = h_{[0,t2]}(A) - h_{[0,t1)}(A)$$



Dynamic bag-of-words

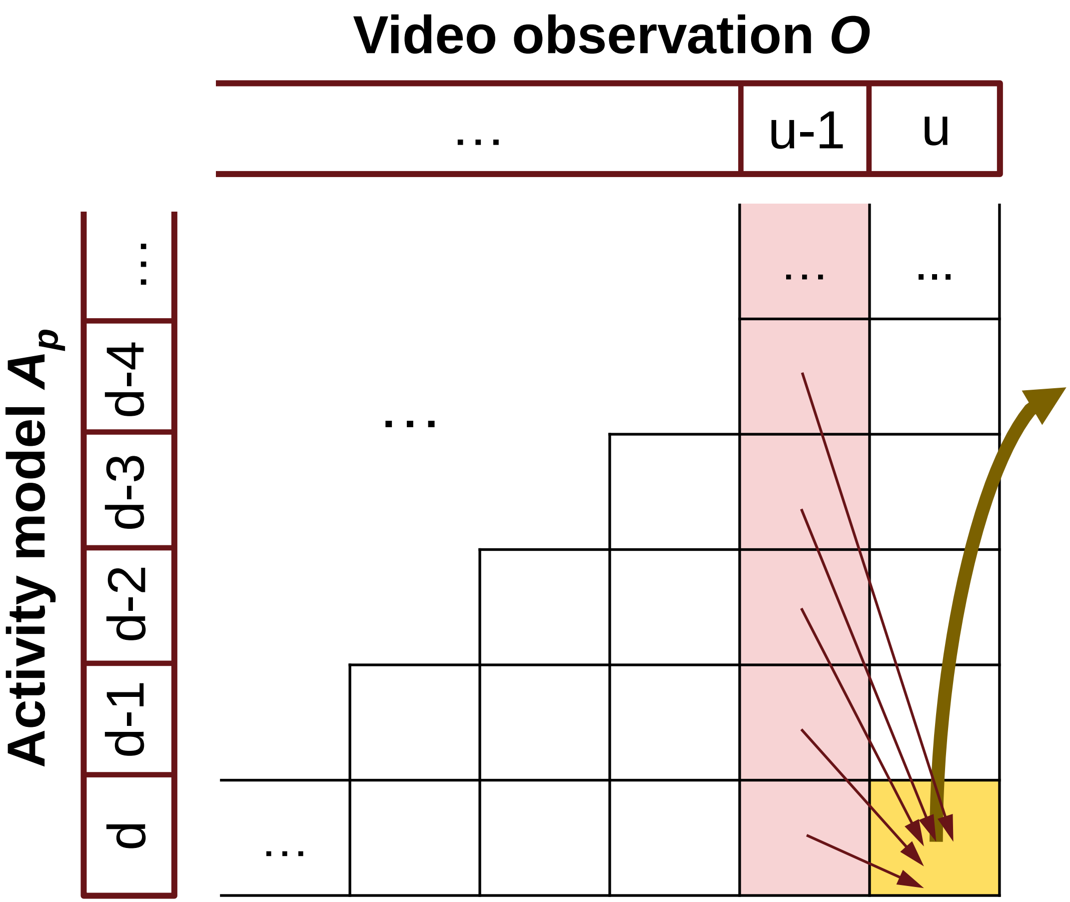
Sequential histogram-based matching



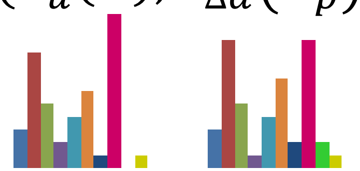
$$P(O^t \mid A, d) = \sum_{\Delta t} \sum_{\Delta d} P(O^{t-\Delta t} \mid A, d-\Delta d) P(O^{\Delta t} \mid A, \Delta d)$$

Dynamic bag-of-words

A dynamic programming-based approximation is designed for efficient computation:



$$F_p'(u, d) = \max_{\Delta d} \left\{ \begin{array}{l} F_p'(u-1, d - \Delta d) \cdot \\ M\left(h_{\overline{u}}(O), h_{\Delta d}\left(A_p\right)\right) \end{array} \right\}$$



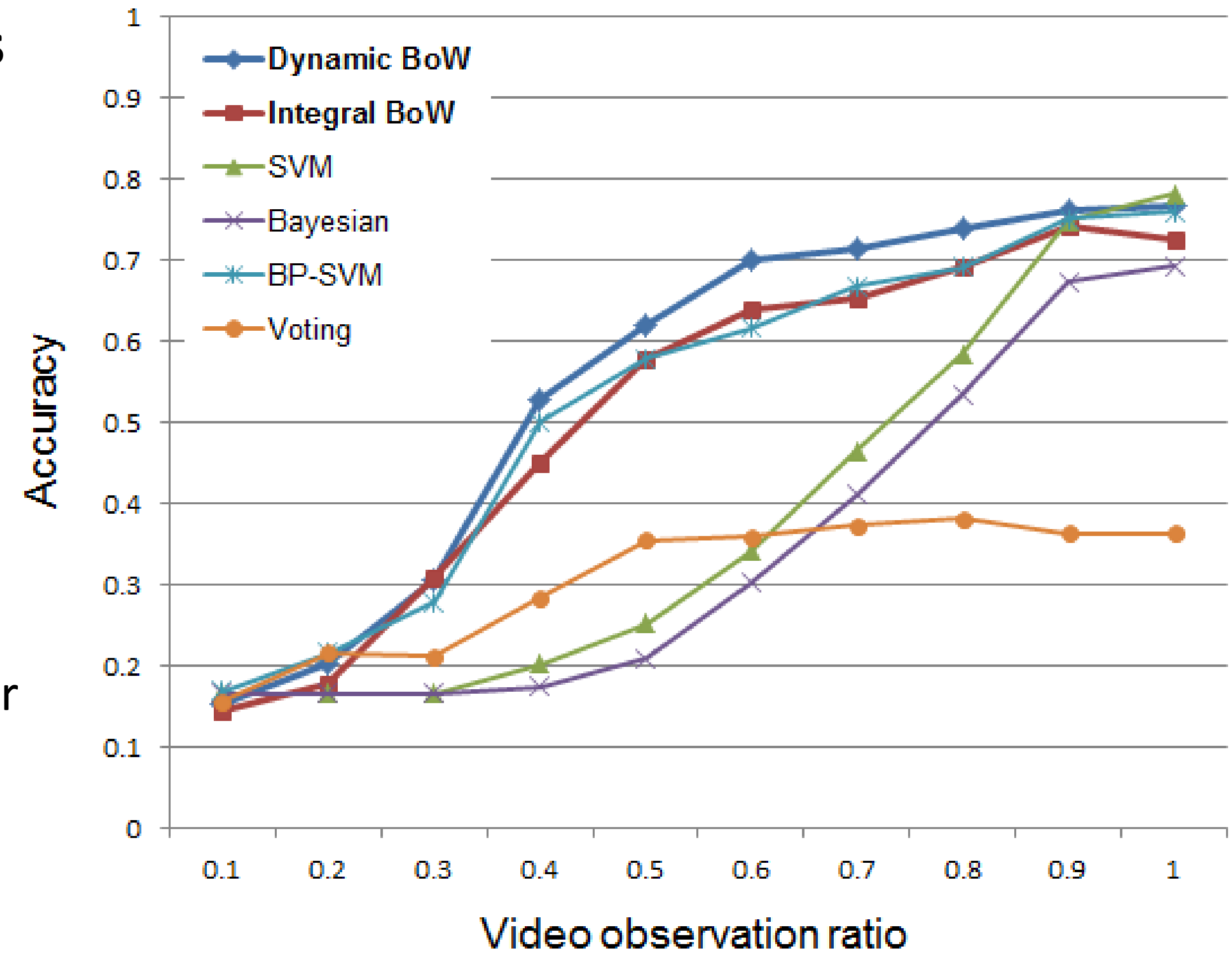
Experimental results

Human activity prediction with UT-Interaction dataset #1 and #2

Human activity prediction

Experimental results

- Human-human interaction
- Our approaches detect activities at much earlier stage.
  - Higher graphs indicate better performance.

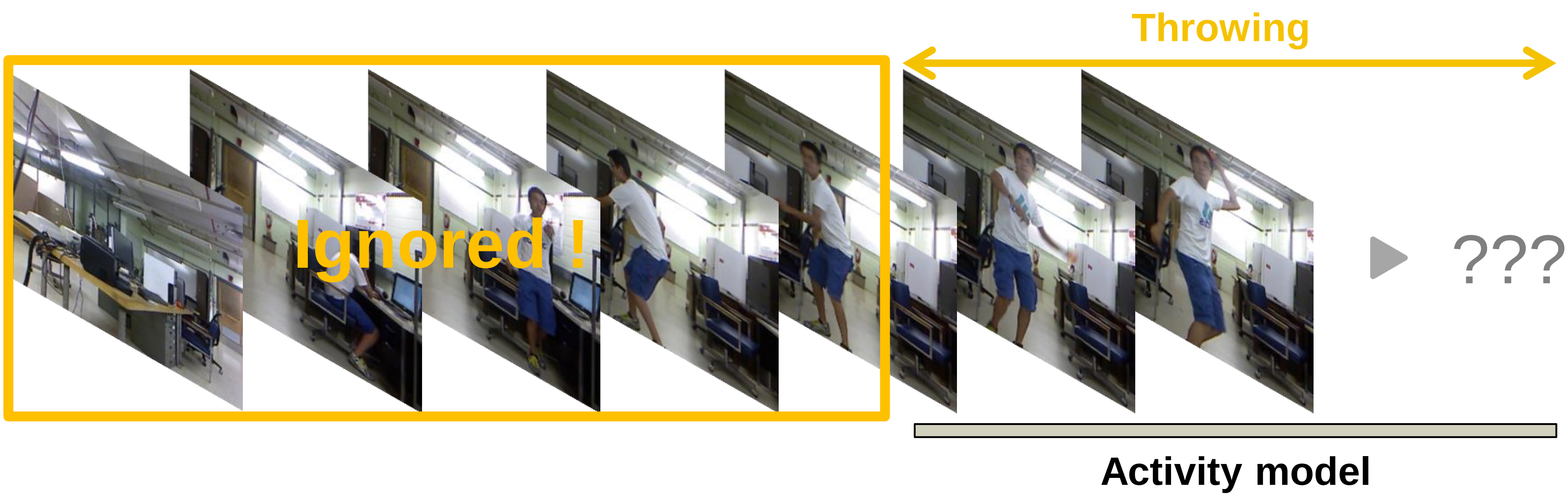


First-person activity prediction

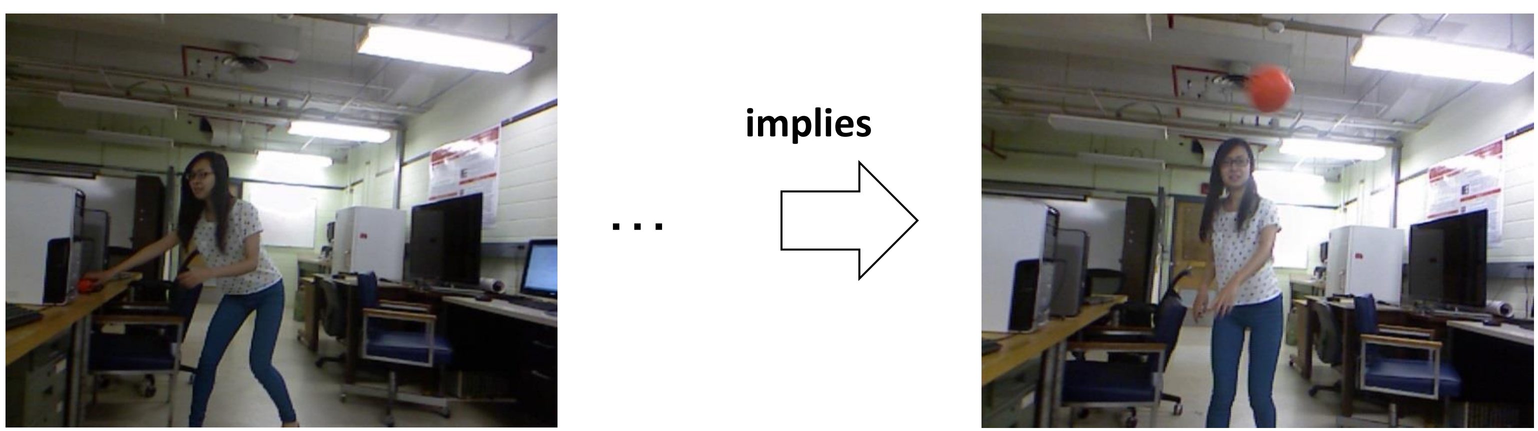
[Ryoo et al., arXiv 2014]

Limitations

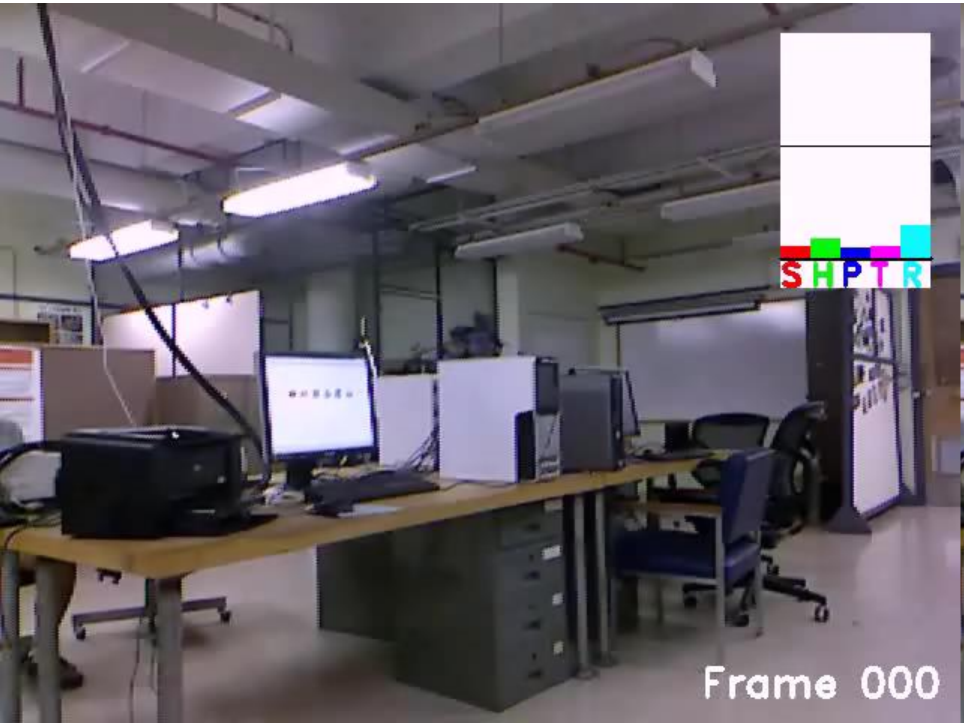
Early recognition from continuous videos?



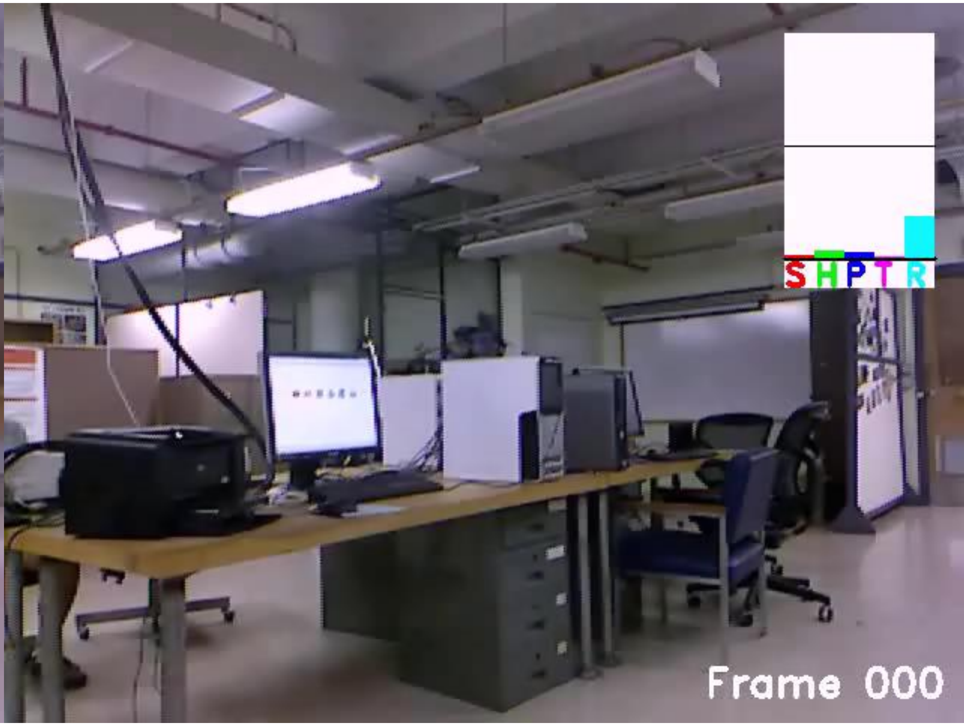
We need to utilize pre-activity videos (onsets)



Video comparison



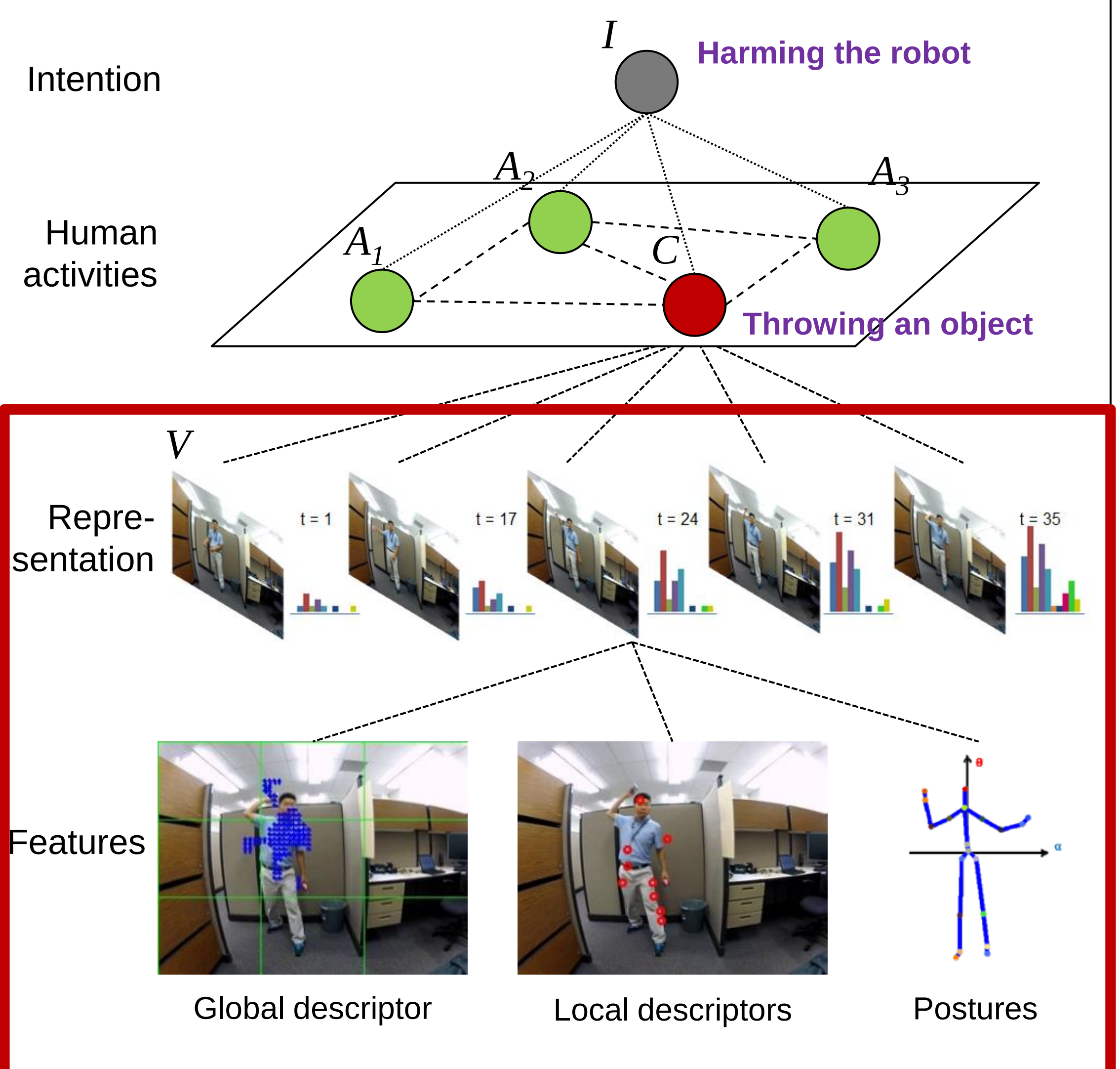
Ours (early detection)



SVM (RBF + [1,9,14,16])

S = shake hands, H = hug, P = punch, T = throw, R = run away

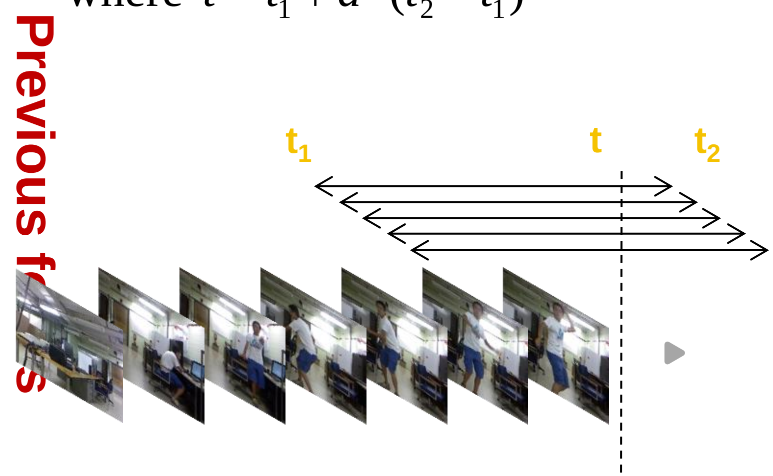
Graphical model formulation



Early detection **of activities with context**

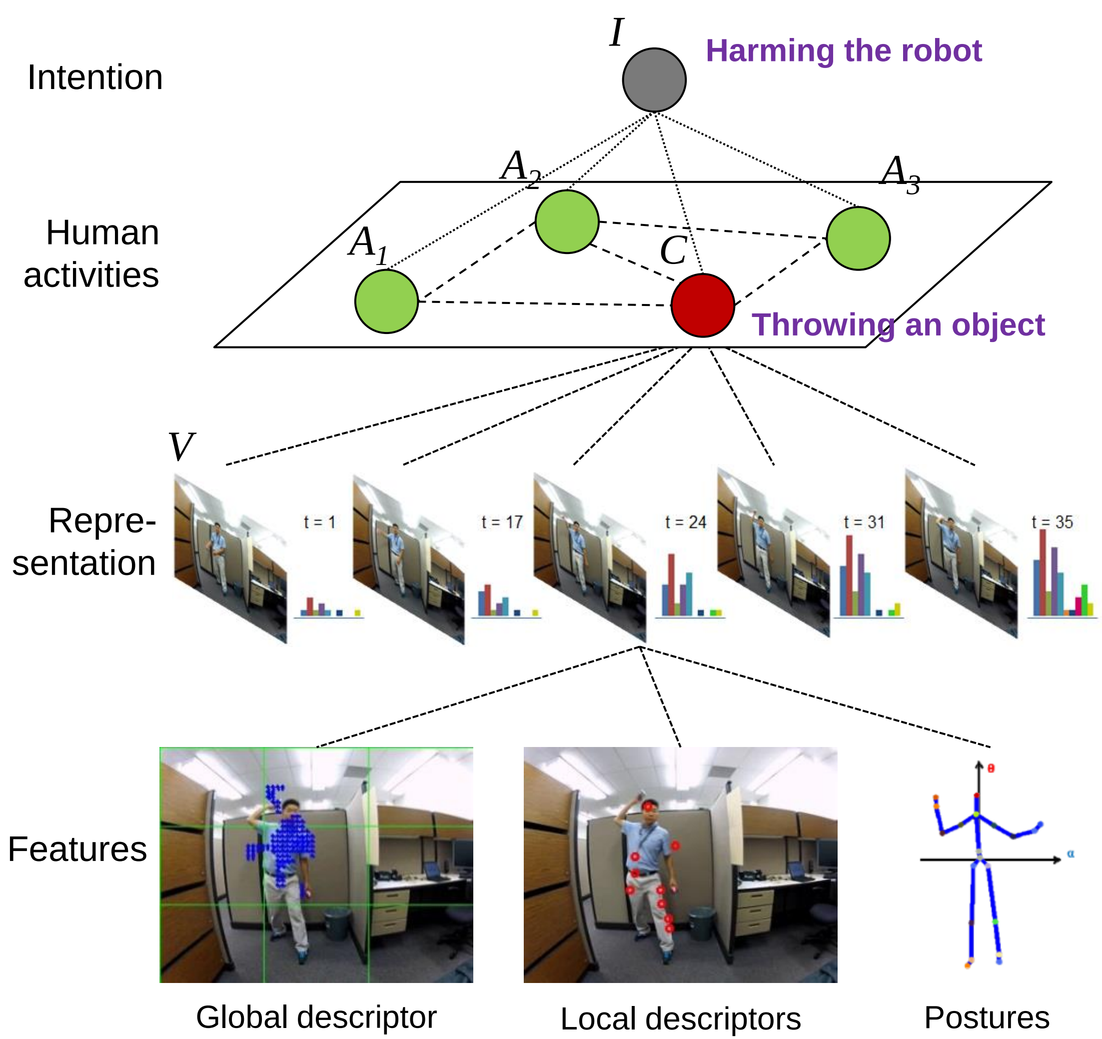
$$P(C^t \mid V,t) = \sum_{d} \sum_{[t_1,t_2]} \boxed{P(C^{[t_1,t_2]}, d \mid V)}$$

where $t = t_1 + d \cdot (t_2 - t_1)$



Multiple possible progress levels

Graphical model formulation



Early detection **of activities with context**

$$P(C^t \mid V, t) = \sum_d \sum_{[t_1,t_2]} \boxed{P(C^{[t_1,t_2]}, d \mid V)}$$

where $t = t_1 + d \cdot (t_2 - t_1)$

Modeling

exponential!

$$P(C^{[t_1,t_2]}, d \mid V) \propto \sum_{(\mathbf{A},I)} F(C^{[t_1,t_2]}, d, \mathbf{A}, I, V)$$

Learning/Inference

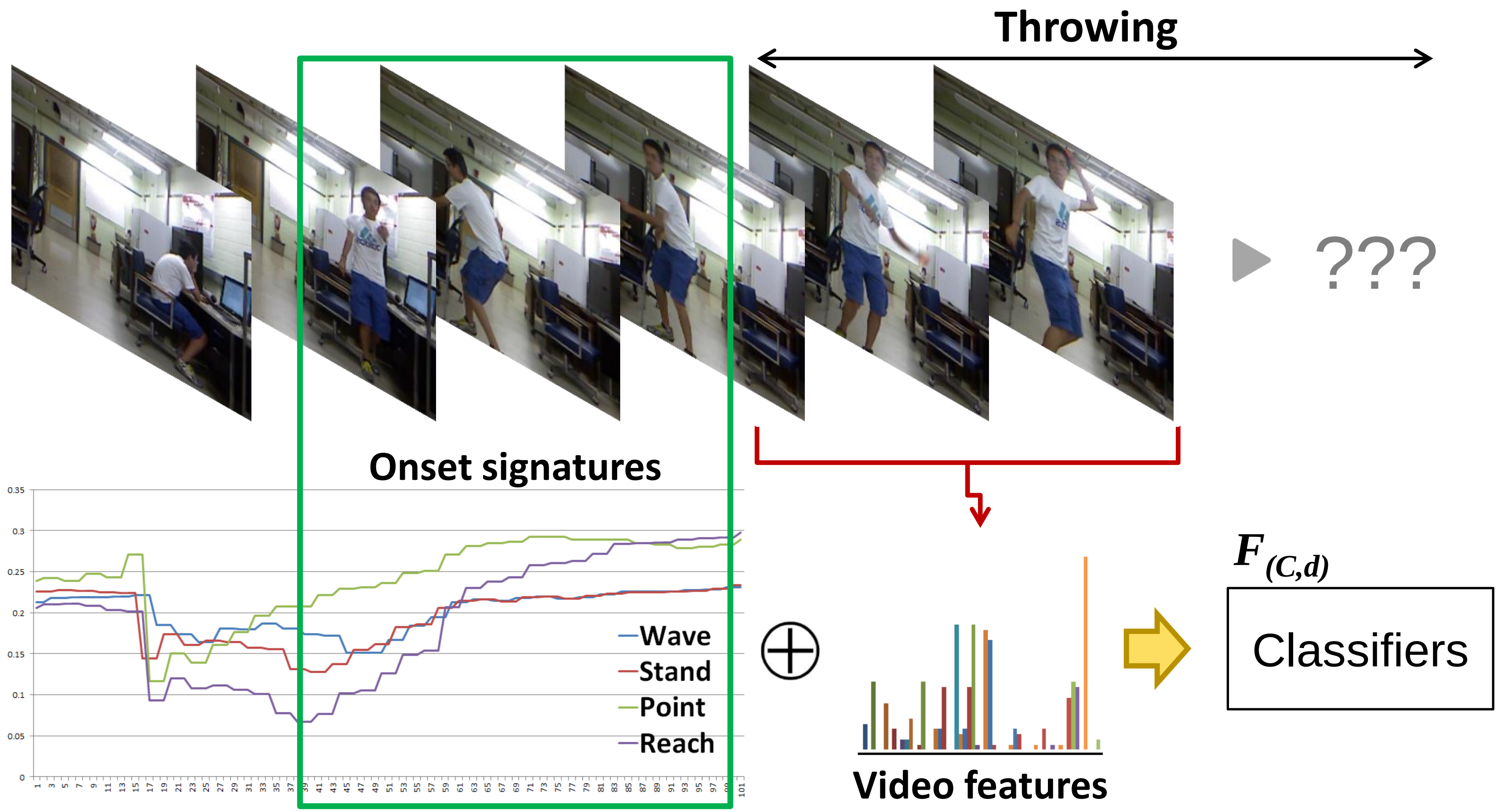
- Markov Chain Monte Carlo (MCMC)?
- Latent SVM?

real-time?

Recognition with onsets

Onset signatures

- Efficient abstraction of pre-activity observations



Onset signatures

Onset activities

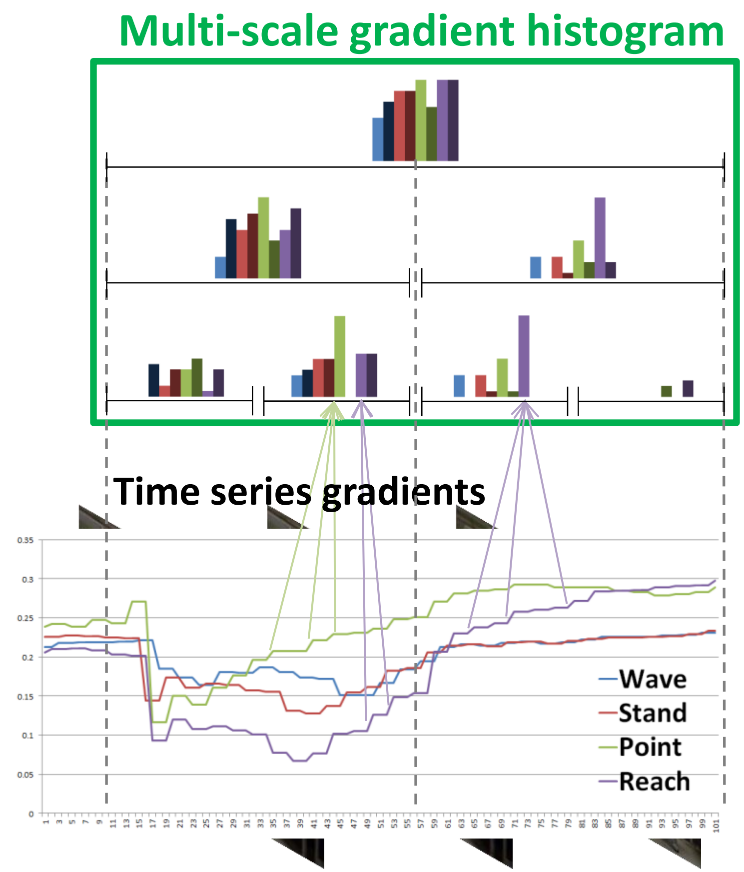
- Subtle human actions commonly observed before other important interactions
  - Waving, standing up, pointing, picking up an object, …

Onset signatures

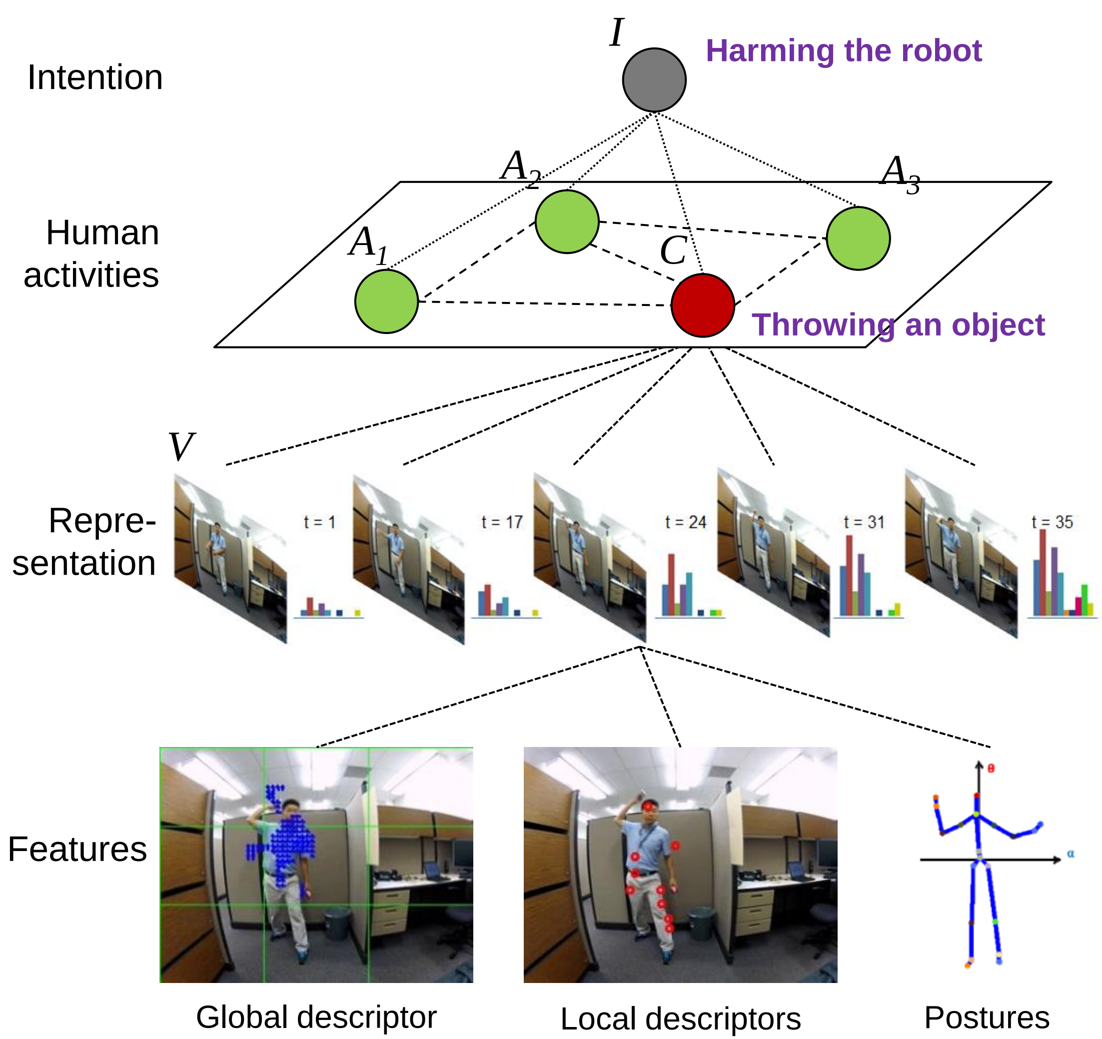
- Weak classifier matching for pre-activity observations
  - Template distance
  - Average precision ~0.1

Multi-scale gradient histograms

- Hierarchical concatenations



Prediction using onsets



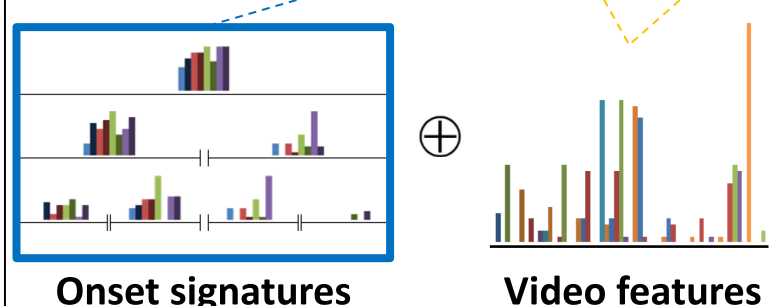
Early detection of activities with context

$$P(C^t \mid V, t) = \sum_{d} \sum_{[t_1,t_2]} \boxed{P(C^{[t_1,t_2]}, d \mid V)}$$

where $t = t_1 + d \cdot (t_2 - t_1)$

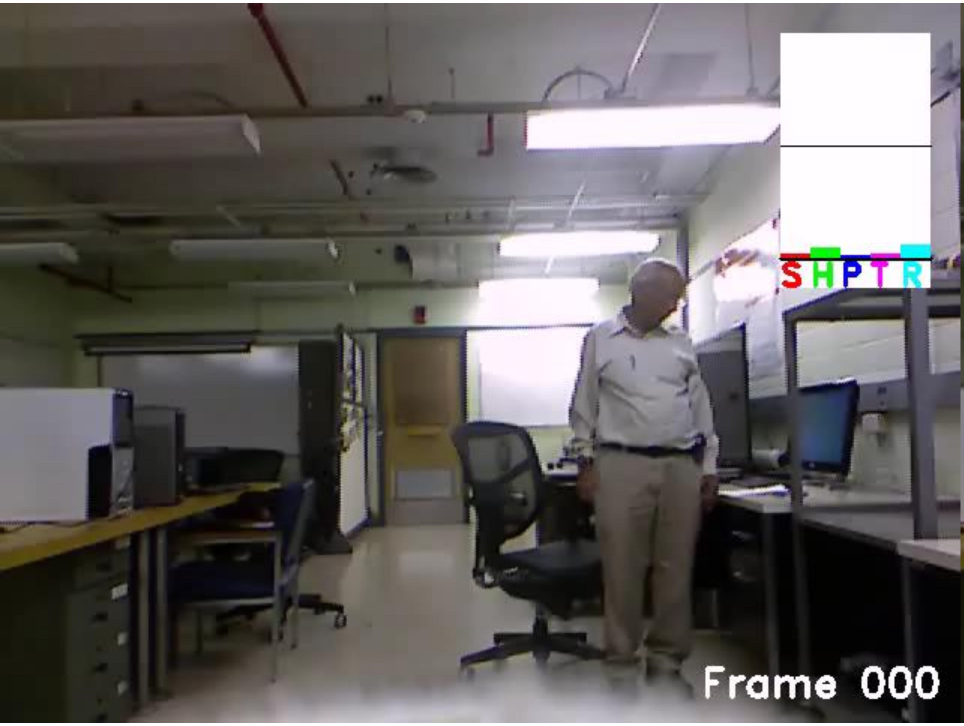
Modeling

$$P(C^{[t_1,t_2]}, d \mid V) \propto \sum_{(\mathbf{A},I)} F(C^{[t_1,t_2]}, d, \mathbf{A}, I, V)$$




S H P T R

Result video comparison

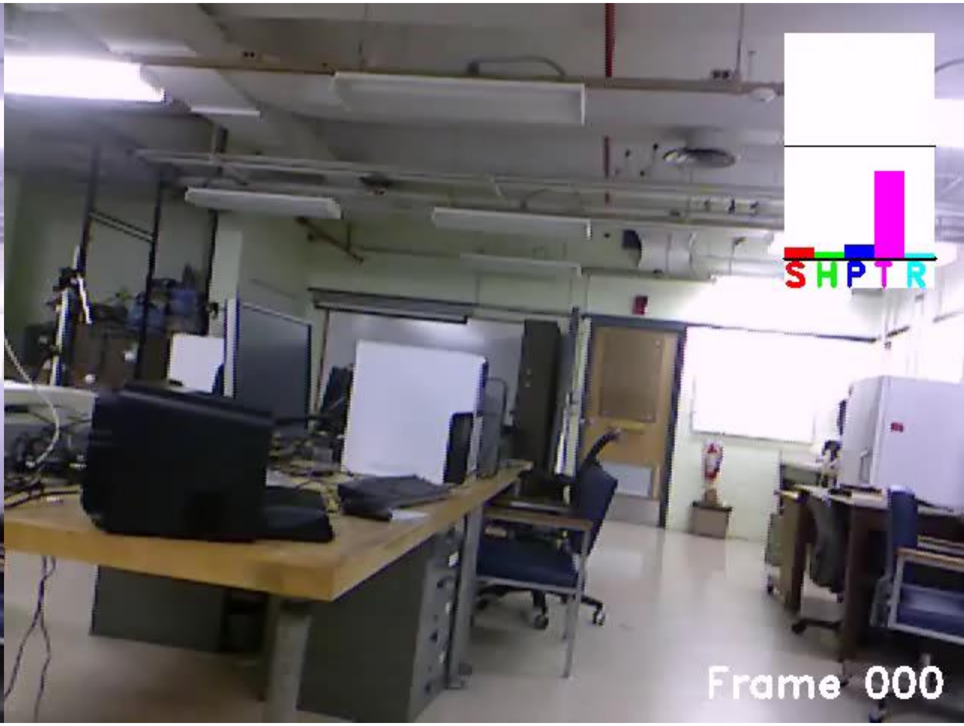


Ours (early detection)

SVM (RBF + [1,9,14,16])

S = shake hands, H = hug, P = punch, T = throw, R = run away

Result video comparison



Ours (early detection) **SVM (RBF + [1,9,14,16])**

S = shake hands, H = hug, P = punch, T = throw, R = run away

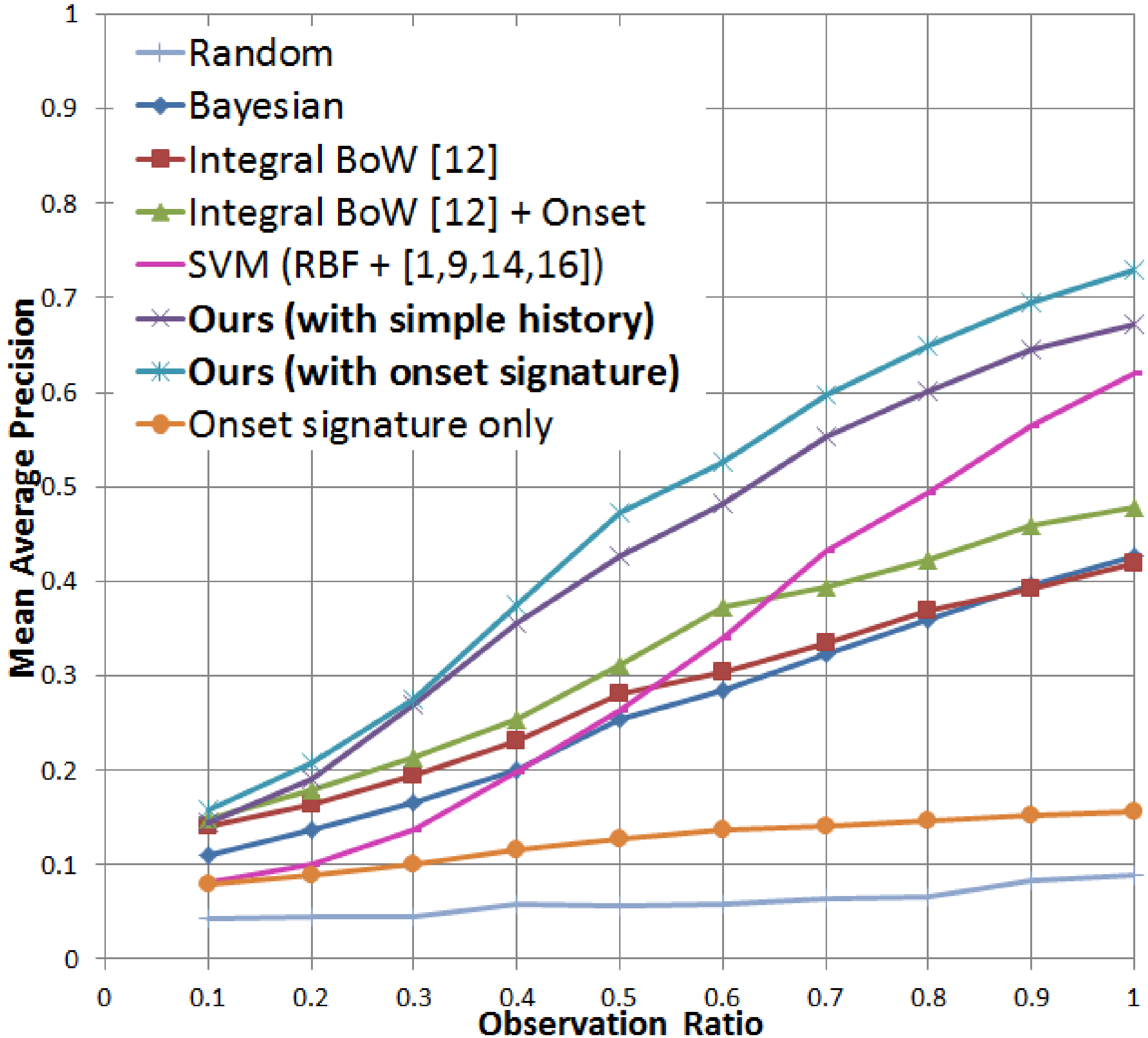
Experimental results

Mean average precision (AP) measure per 'observation ratio'

- Observation ratio 50% implies that the first half of the activity was visible

Dataset

- 5 interactions similar to JPL-Interaction dataset
- 4 onset activities



Action vocabularies