Human Activity Recognition for Visual Surveillance

Michael S. Ryoo, NASA's Jet Propulsion Laboratory
Anthony Hoogs, Kitware, Inc.
Arslan Basharat, Kitware, Inc.
Sangmin Oh, Kitware, Inc.

Jet Propulsion Laboratory
California Institute of Technology

Kitware

Activity recognition

- Goal
  - Labeling of all objects, persons, and their events in a given video

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

Person 3 – *kicks* **P1**

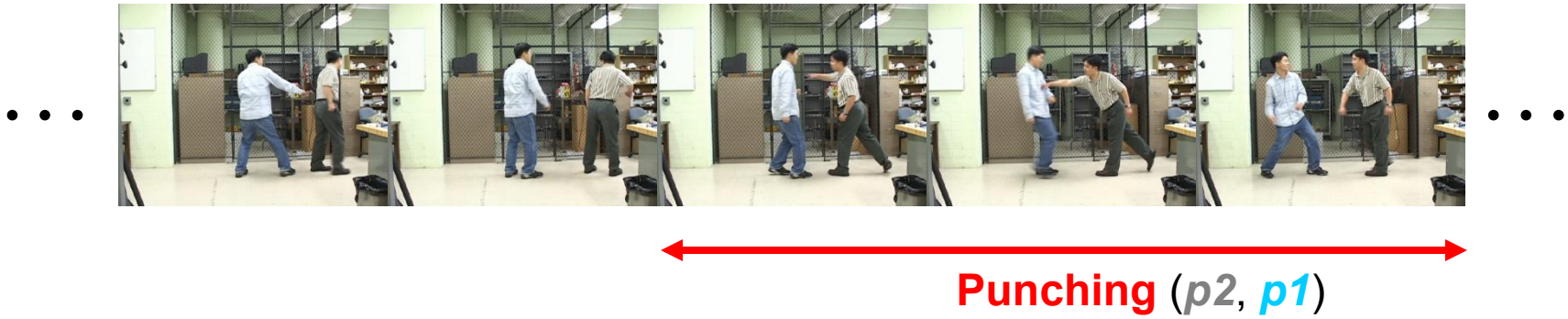
Person 4 – *stops fighting*

  - Develop automated algorithms for the video recognition

What is activity recognition?

- Human activity recognition
  - Automated detection of ongoing events from video data.
    - Computer analysis of inputs from cameras.
  - Human actions, human-human interactions, human-object interactions, group activities.

Input video:



Human activity

- Human activity
  - A collection of human/object movements with a particular semantic meaning
    - i.e., particular structure

- Activity recognition
  - Finding of video segments containing such movements
    - Must ***search*** for video segment that display properties of the movements

Surveillance

- Ubiquitous cameras in public places (e.g. CCTVs).
  - In London, an average person is monitored 300 times / day.
- Goal
  - Monitor suspicious ***activities*** for real-time reactions.
    - ‘Fighting’, ‘stealing’.
  - Currently, surveillance systems are mainly for recording.

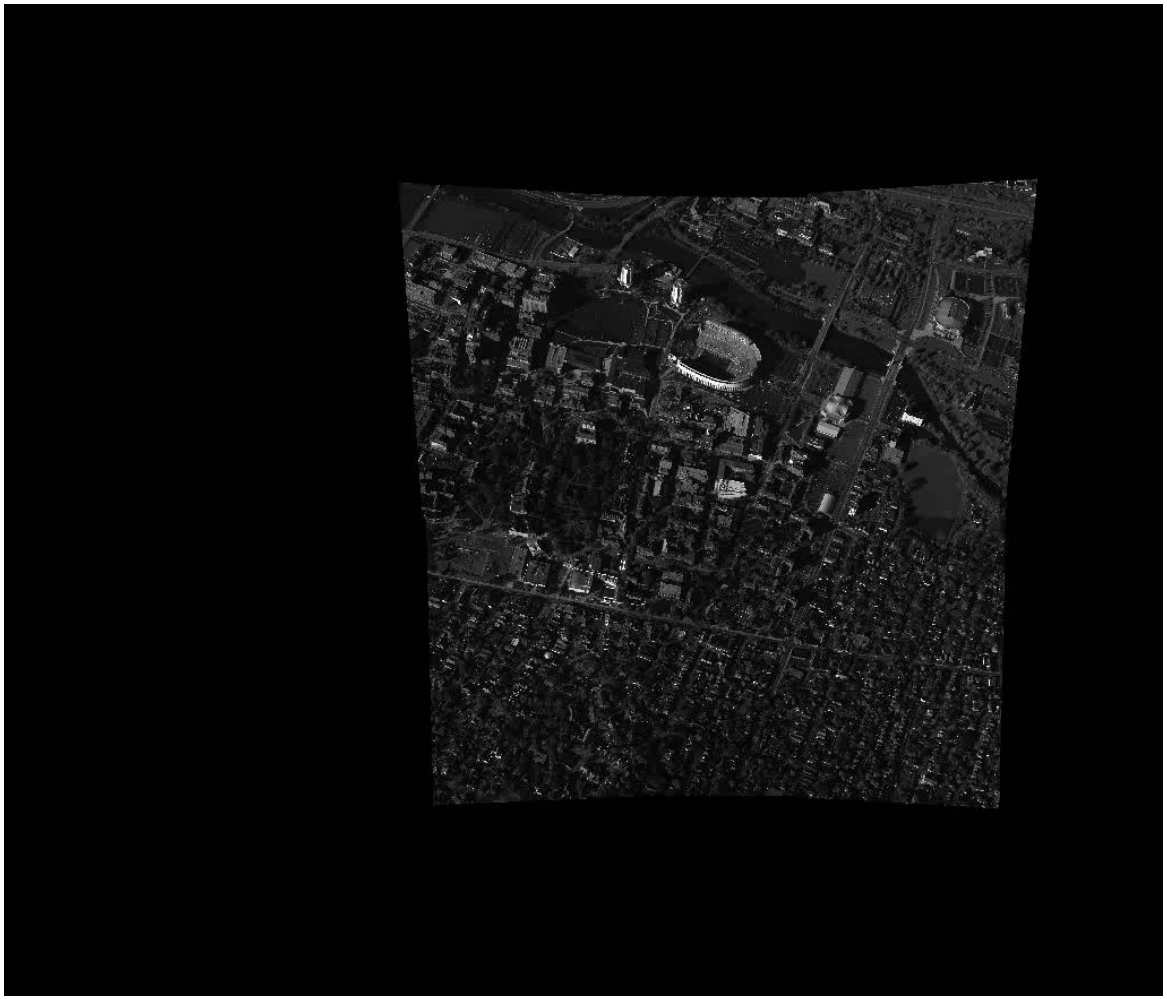
- ***Activity recognition*** is essential for surveillance and other monitoring systems in public places

Activity recognition for near

- Cameras observing multi-human movements
  - Group stealing in an Apple computer store.

Activity recognition for far

- Aerial images
  - Videos recorded from aerial vehicles
  - Moving cameras
  - Body-part level analysis is impossible.



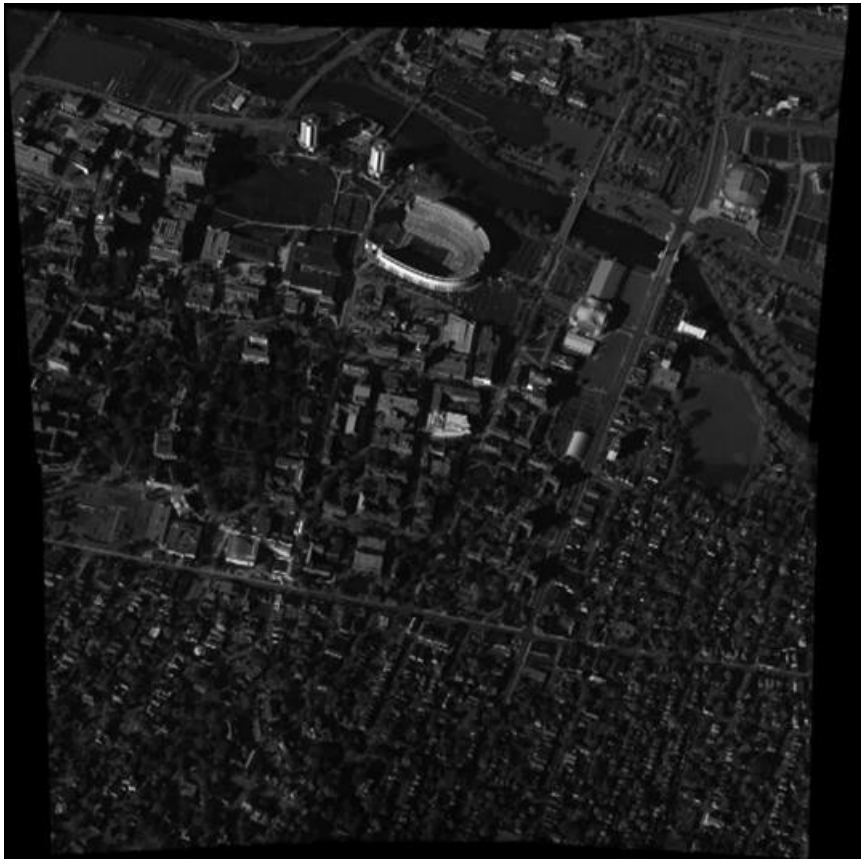
Video FOV granularity & utility

Ground-Level Surveillance

•Alert on anomalies
•Alert on specific events, behaviors
•Track specific objects

Narrow Field-of-View Aerial Video

•Alert on anomalies
•Alert on specific events, behaviors
•Track specific objects



Wide-Area Motion Imagery (WAMI)

•Alert on anomalies
•Alert on specific events, behaviors
•Track specific objects

Group activity recognition from CCTVs

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

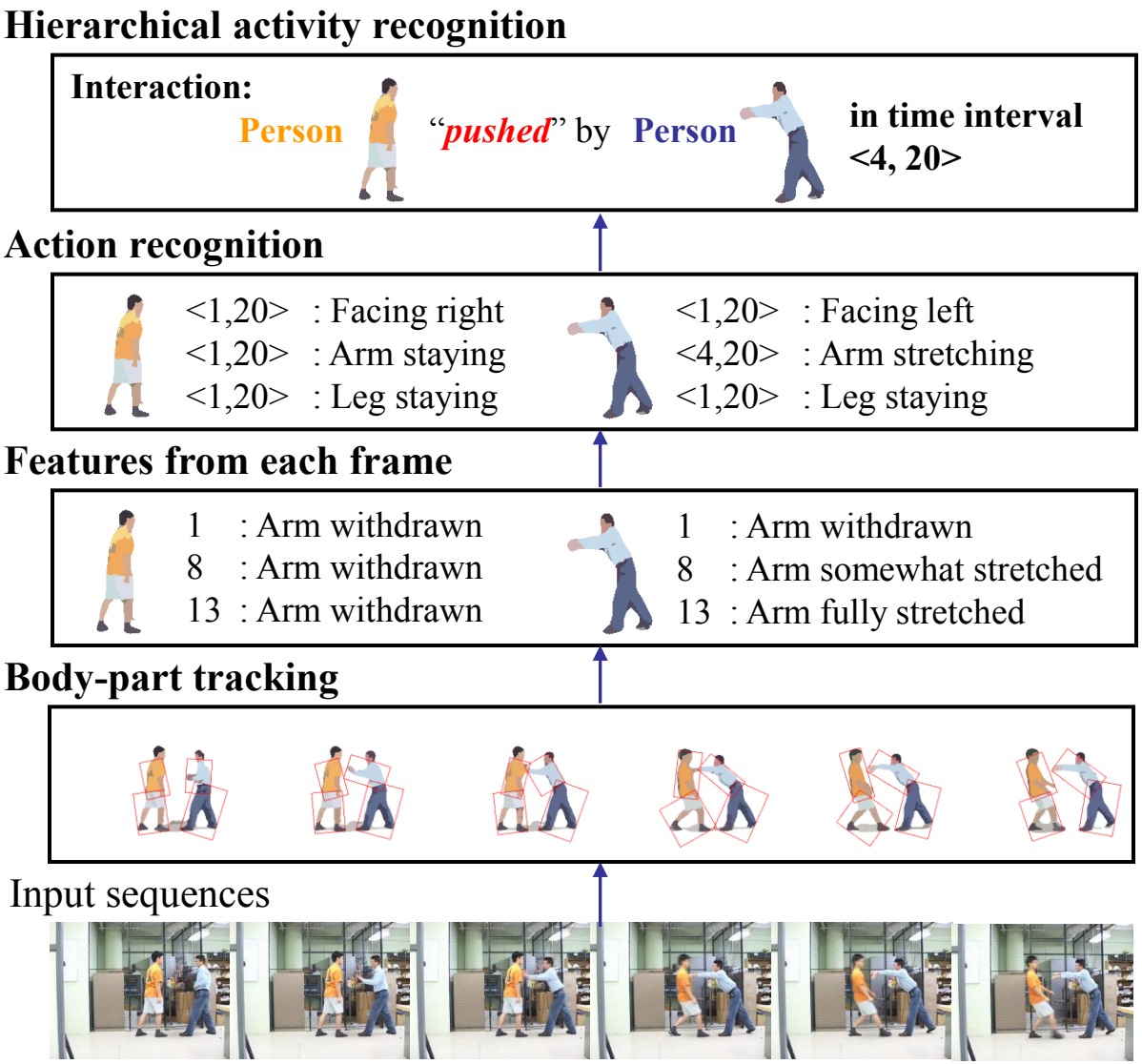
Thieves

Owners

Laptop

Ryoo and Aggarwal, IJCV 2011

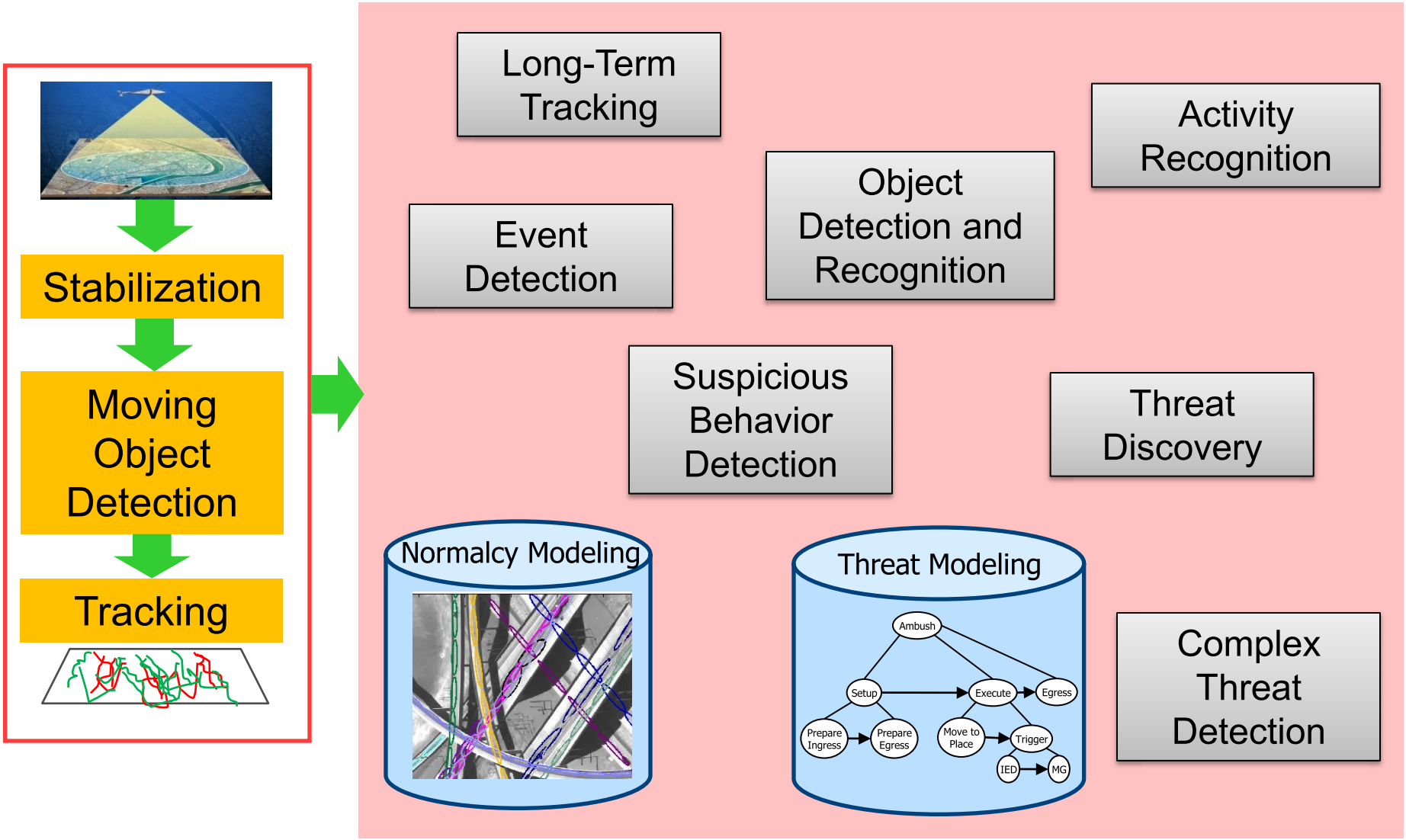
Human activity recognition



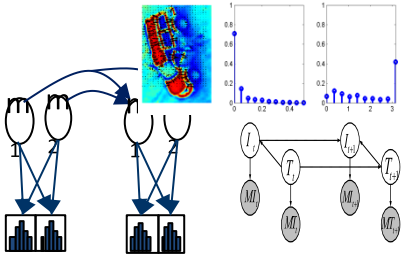
- Interaction
- Action
  - Sequence of features
- Features
  - Numerical status of a body part
- Tracking
  - Estimates locations of human body parts

Ryoo and Aggarwal, CVPR 2006

Persistent Video Analysis

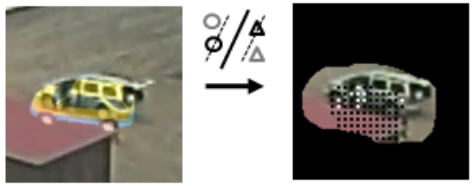


Tools for activity recognition

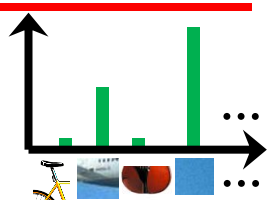


dynamic bayesian networks
[12][13]

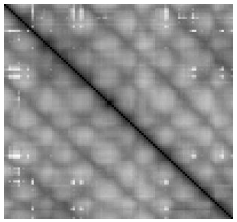
sensor metadata: gsd, pointing angles, etc.



objects interaction modeling
[1][7][8]



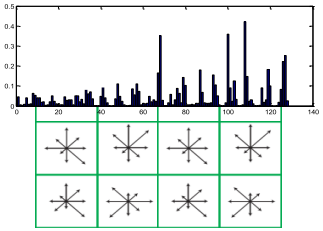
bag of visual words
[4][11][16]



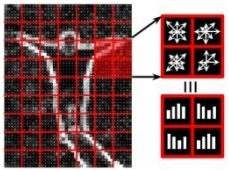
periodicity of self-similarity matrices
[14]

bounding box: area, aspect ratio, etc.

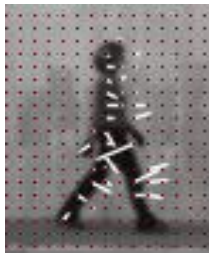
track level: speed, delta heading, curvature, etc.
[1][12][13][8]



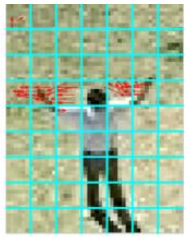
Feature point descriptor: SIFT
[10][11][15]



spatio-temporal histogram of gradients
[6][12]



optical flow
[2][6][9]



histogram of optical flow
[6]

Schedule

- Introduction
  - Background
- Activity recognition from near-field
  - Single-layered approaches
  - Hierarchical approaches
  - Academic datasets
- Activity recognition from far-field
  - Tracking
  - Event recognition
  - Surveillance datasets
- Applications
- Discussions and future direction

Activity Recognition - Background

Levels of human activities

- There are various types of activities
  - The ultimate goal is to make computers recognize all of them reliably.

gestures

actions

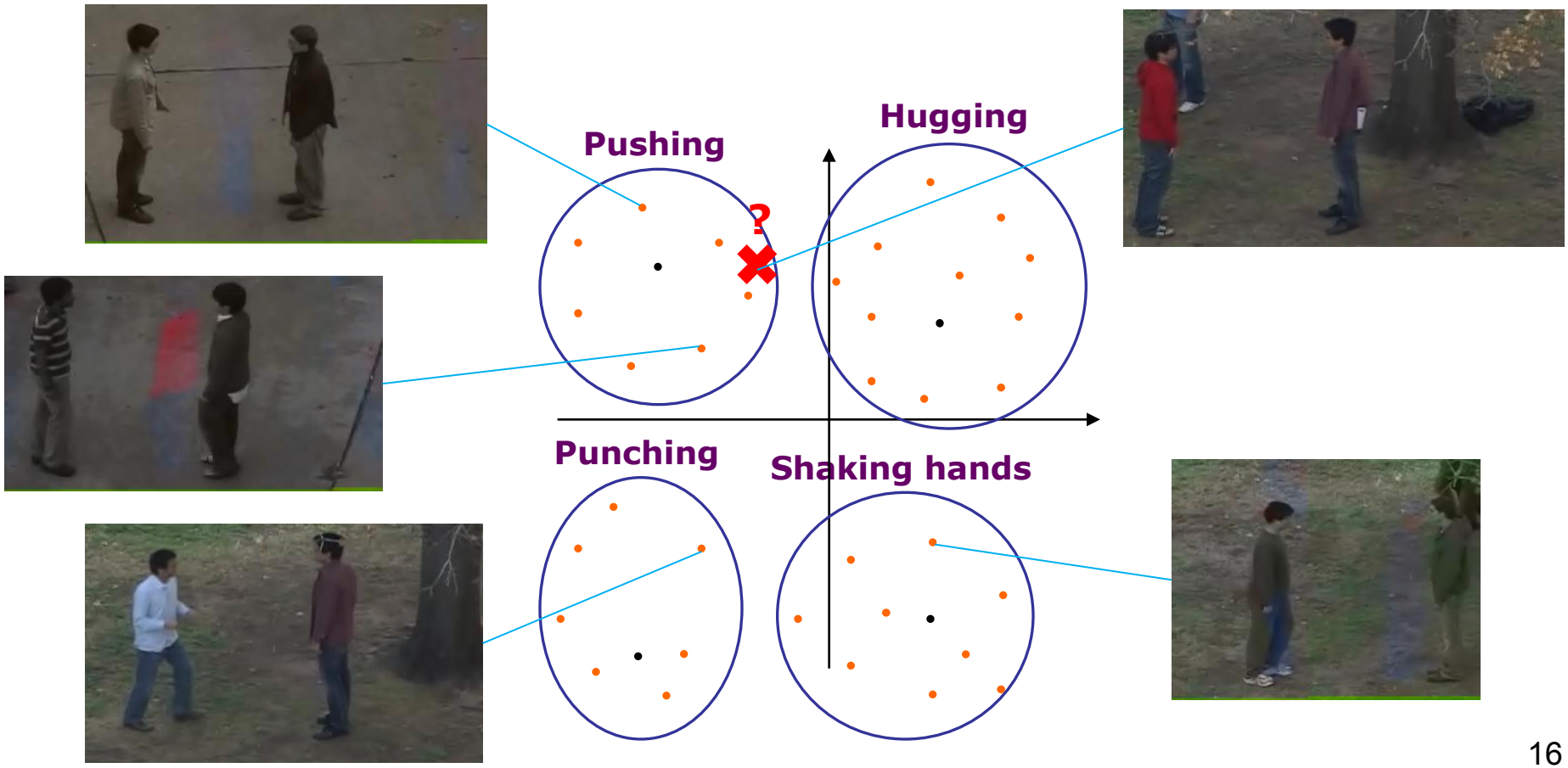
interactions

group activities

Levels of human activities

Activity classification

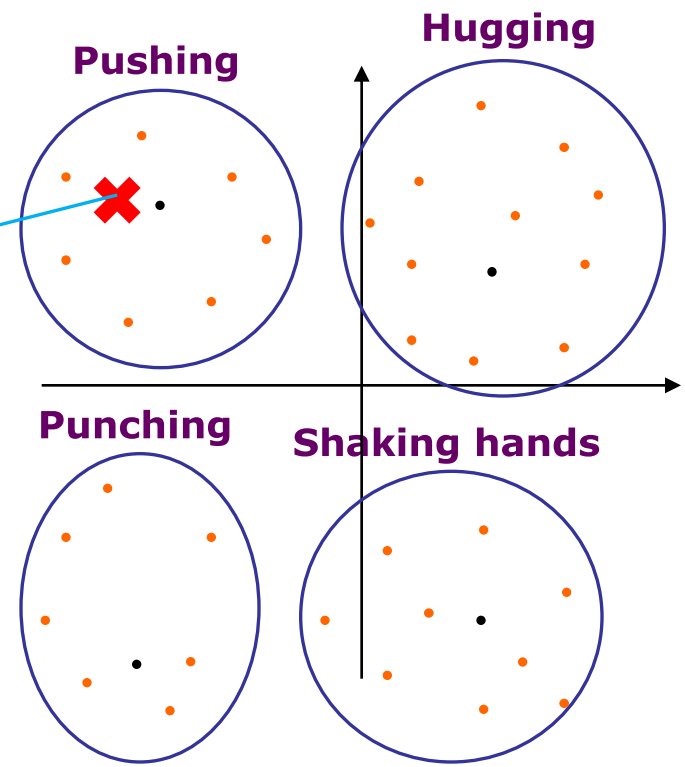
- Activity categorization
  - Input = a video segment containing 1 activity



Activity detection

- ***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

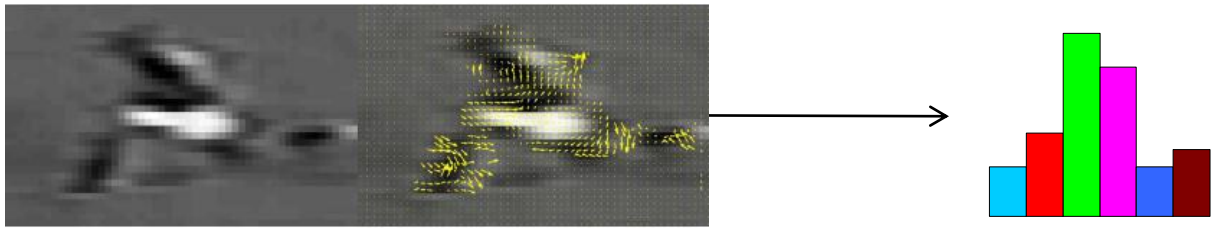
→ **Push/not-push Classifier** → **Yes, a push.** / **No, not a push.**

- Sliding window technique
  - Classify all possible time intervals

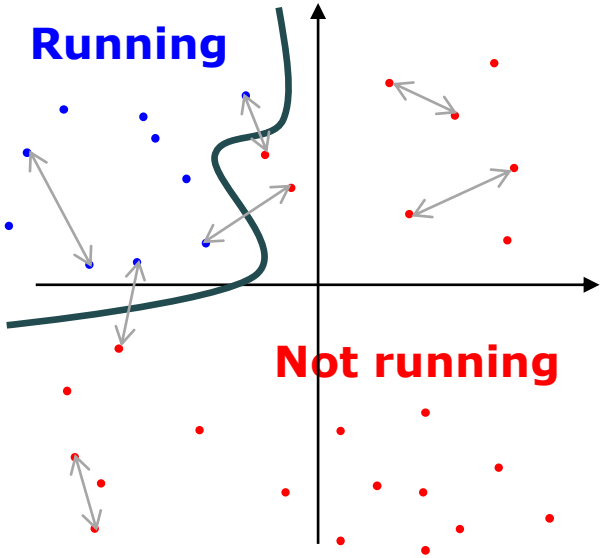
Pushing

Recognition process

- Represent videos in terms of features
  - Captures properties of activity videos

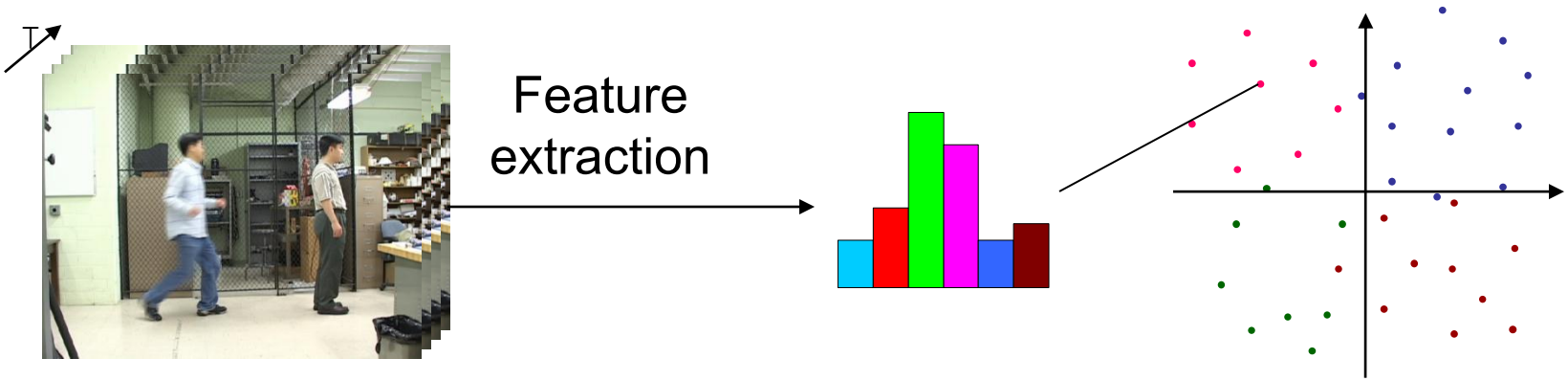


- Recognize activities by comparing video representations
  - Decision boundary

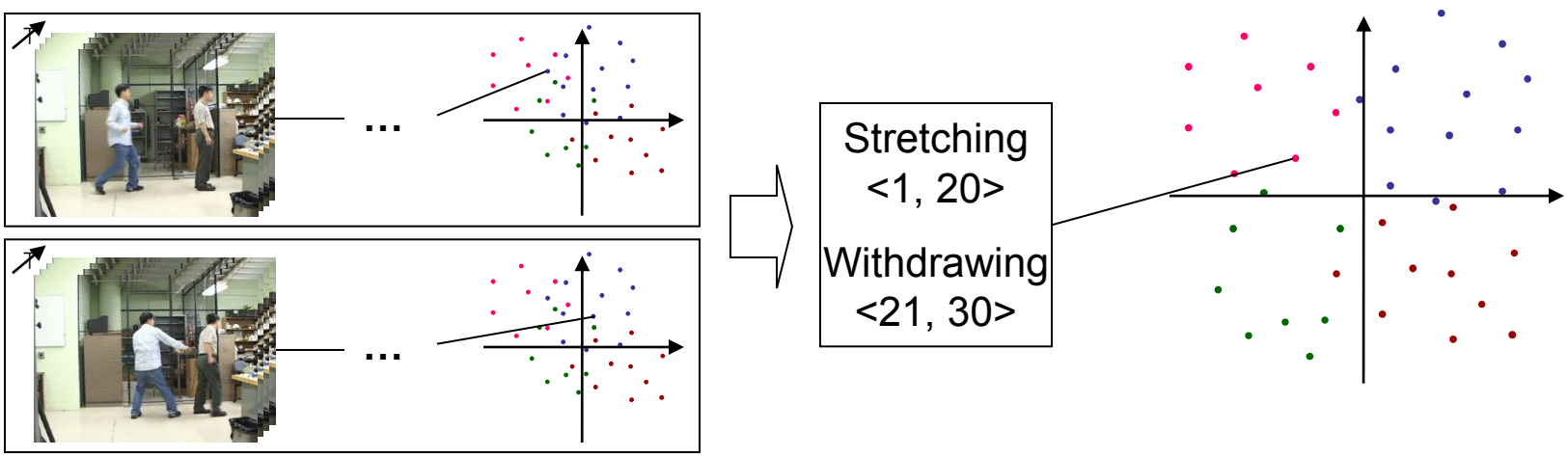


Single layered vs. hierarchical

- Single layered approaches

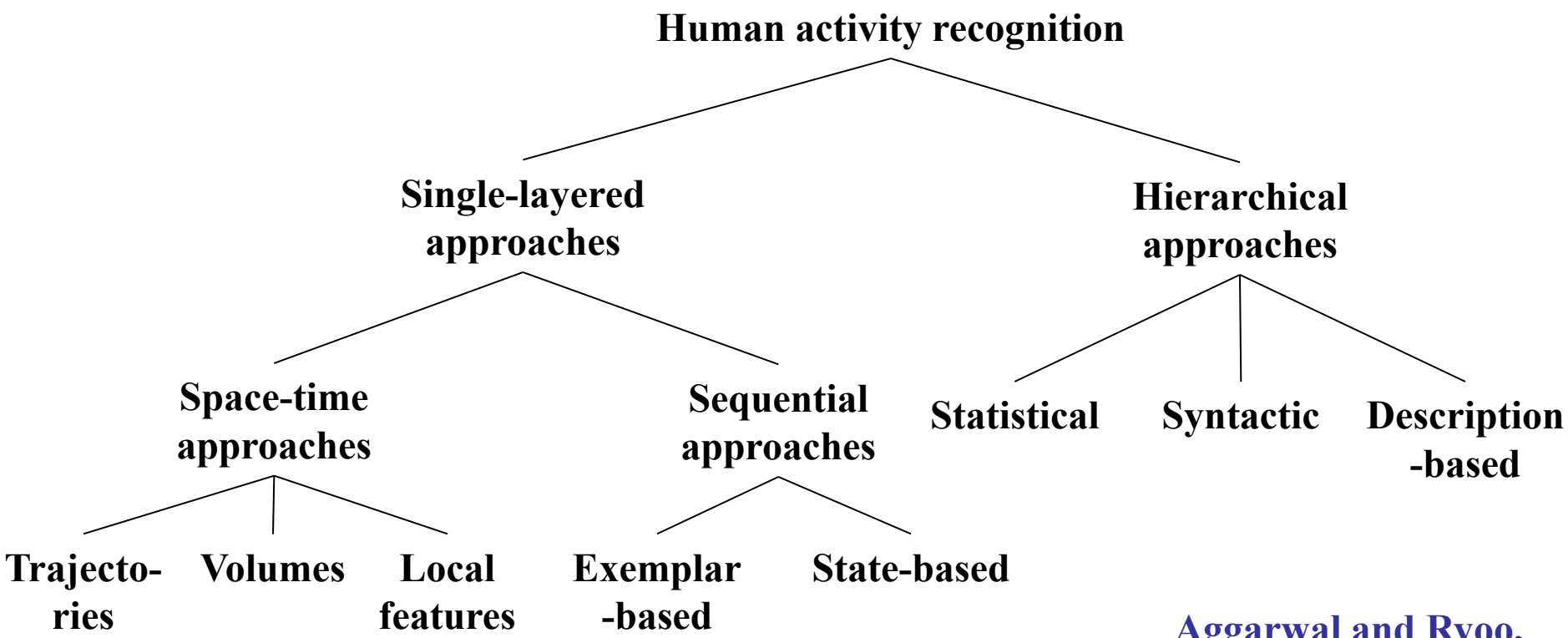


- Hierarchical approaches



Taxonomy

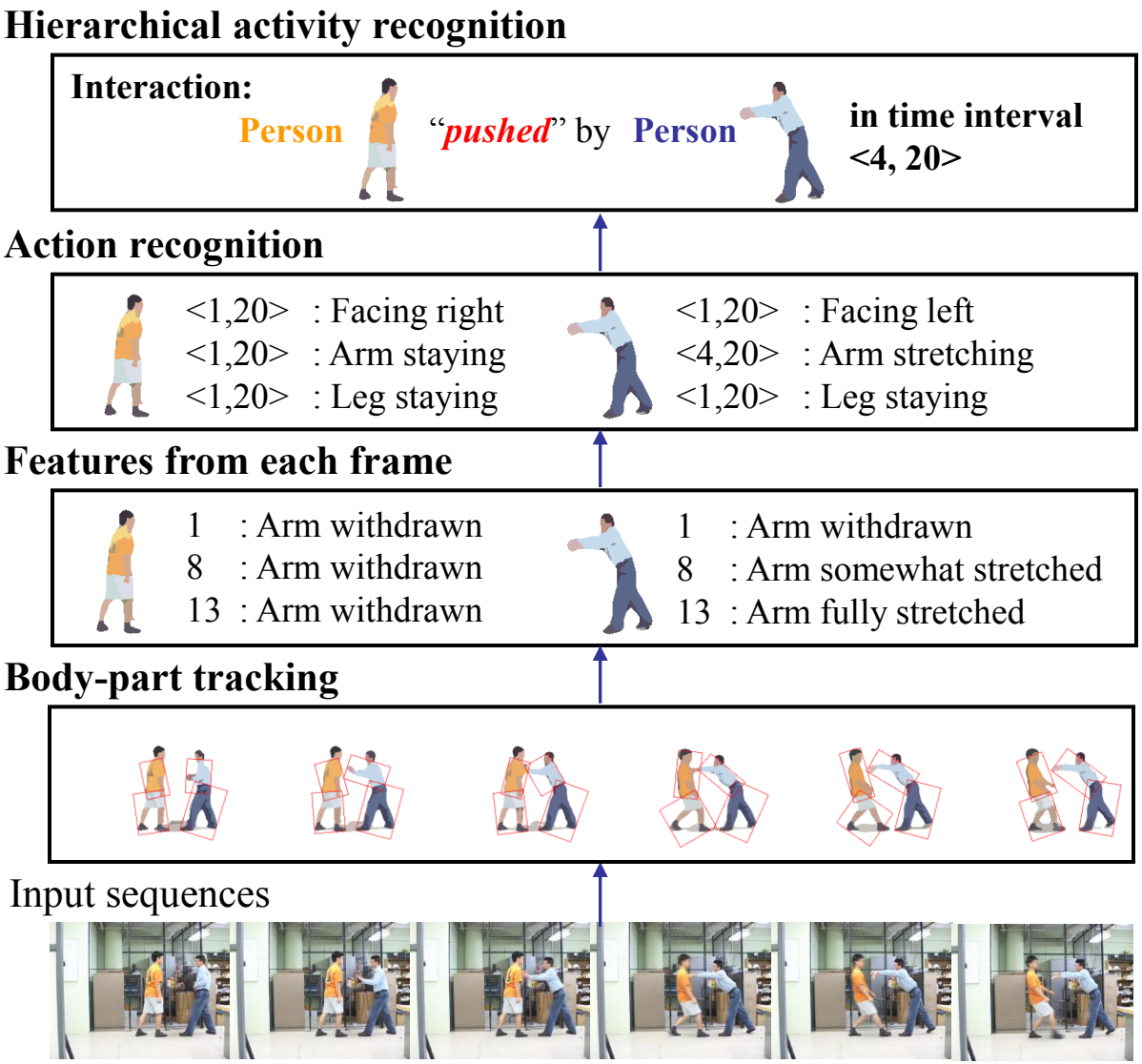
- Approach based taxonomy
  - Recognition approaches can be categorized.



Aggarwal and Ryoo, ACM CSUR 2011

Single layered approaches

Activity recognition



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

Activity recognition – another view

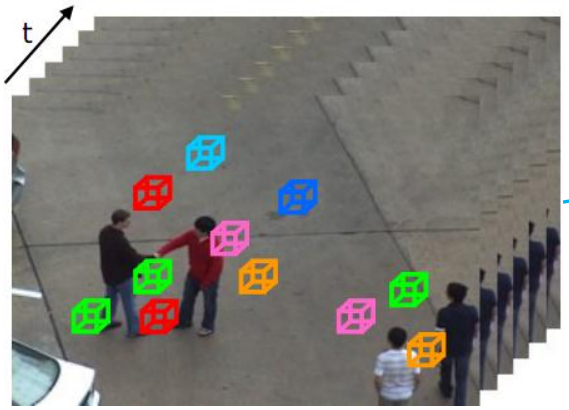
Activity recognition

Activity:

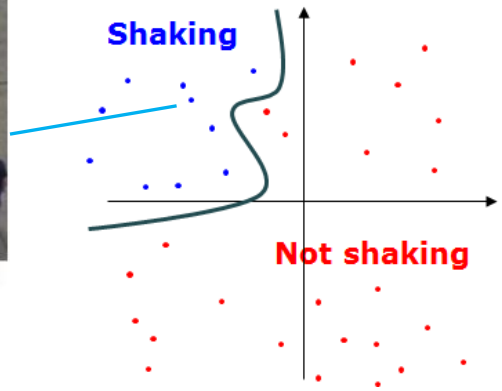
"***Shaking***" occurred

Spatio-temporal features

Videos as 3-D XYT volumes



Decision boundary for histogram of video patches (i.e., space-time features)



Input sequences

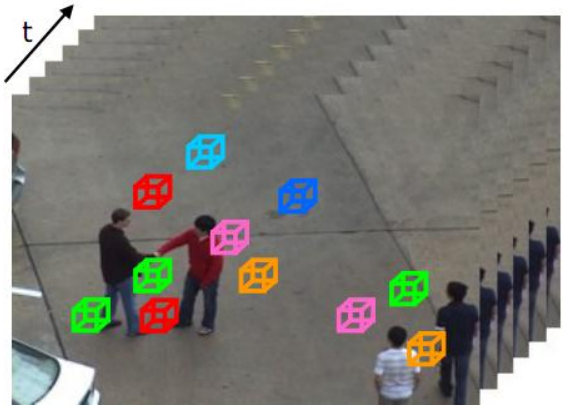
- Activity recognition
  - Decision boundary
- Activity representation
  - A set of local spatio-temporal features
- Features
  - Information in local video patches with salient movements

Two different views

- Activities as human movements
  - Semantic-oriented
  - 3-D body-part estimation
  - Tracking

- Sequence

- Activities as video observations
  - Data-oriented
  - Spatio-temporal features
  - Bag-of-words



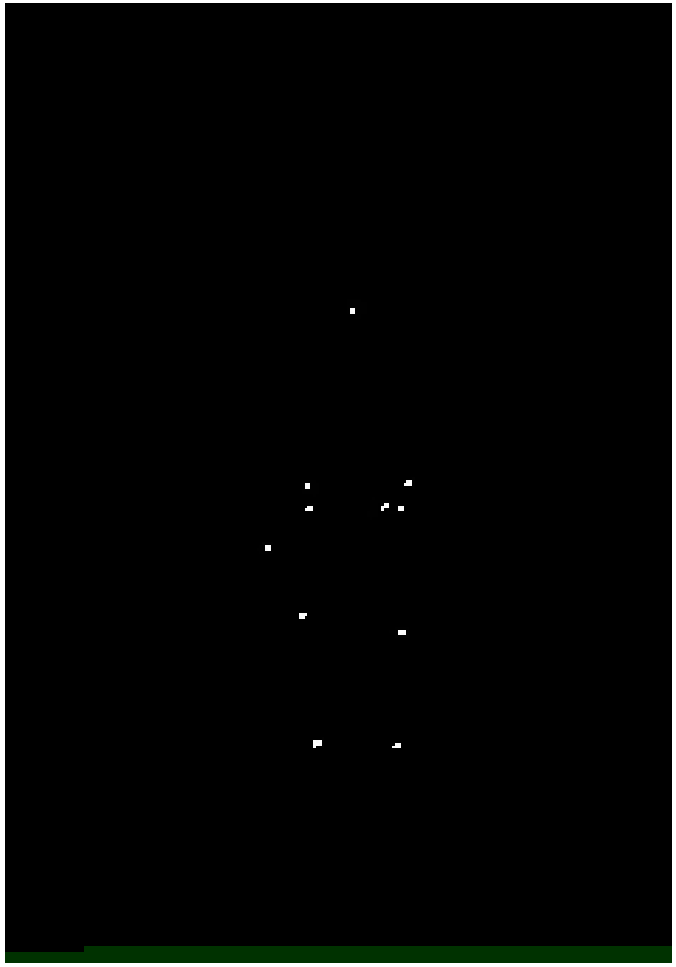
- Space-time distribution

Single layered approaches

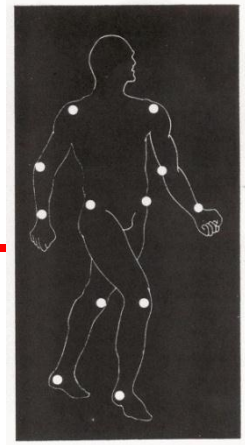
Sequential approaches

Sequential approaches

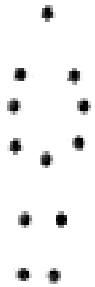
- Motivation
  - An action is a ***sequence*** of body-part states
  - Each frame in an action video describes a particular body-part configuration
    - Example:
      11 points
      body configuration of ‘kicking’



Beginnings of Activity Recognition



- Johansson's experiments (1973) - lights attached to major joints of a person, dressed in black and human recognition of activity.

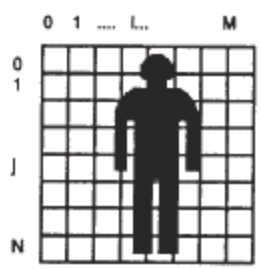


- Representing each rigid body part by two points, and determining the structure of jointed objects under orthographic projection.

Action recognition using HMMs

- Recognition using hidden Markov models
  - Each HMM generates a particular sequence of features.
  - Matching observed features with the model.
    - An action -> a set of sequences of features

- [Yamato et al. CVPR 1992]: Tennis plays

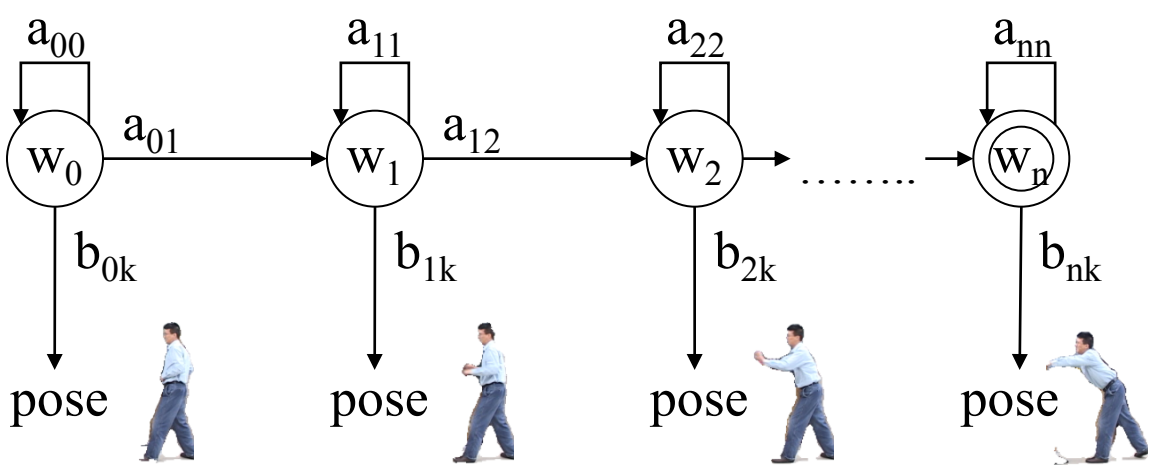


$f=(a_{00},a_{01},\ldots,a_{ij},\ldots a_{MN})$

Symbol sequence	60 61 61 62 62 62 63 63 64 64 65 66 66 66 67 68 68 69 69 70 70 70 71 71

HMMs for actions

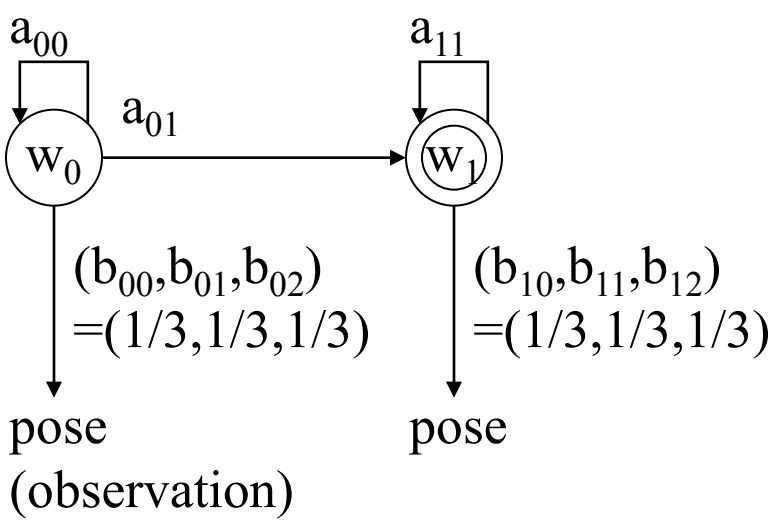
- Human action as a pose sequence
- Each hidden state is trained to generate a particular body posture.
  - Each HMM produces a pose sequence: action



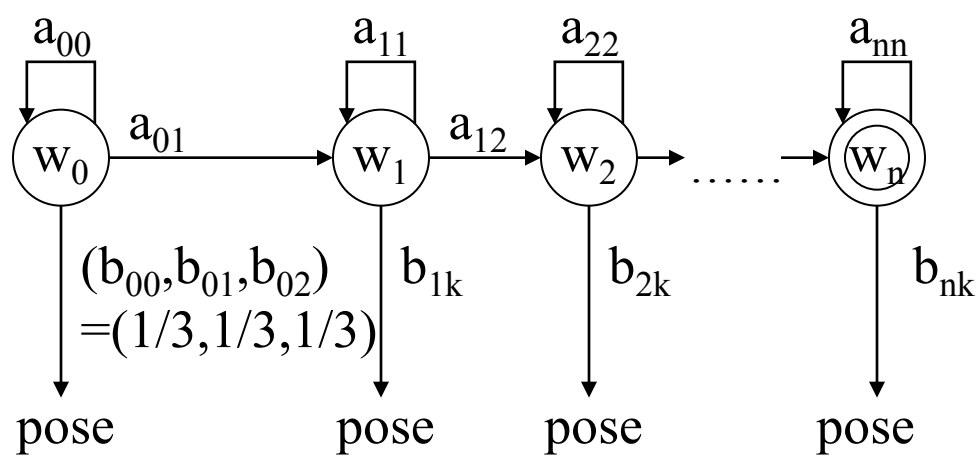
Hidden Markov models

- This is a classic evaluation problem of HMMs.
  - Given observations V^T (a sequence of poses), find the HMM M_i that maximizes $P(V^T|M_i)$: forward algorithm.
  - Transition probabilities a_{ij} and observations probabilities b_{ik} are pre-trained using training data.

Noise HMM

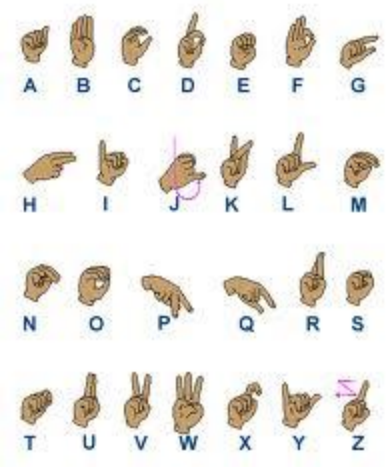


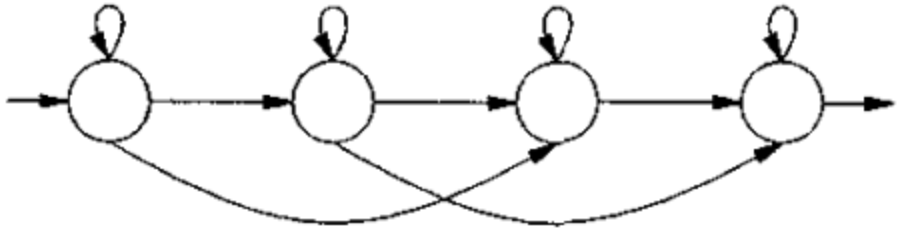
Structure of other basic HMMs



HMMs for hand gestures

- HMMs for gesture recognition
  - American Sign Language (ASL)
  - Sequential HMMs
    - Features from colored globes

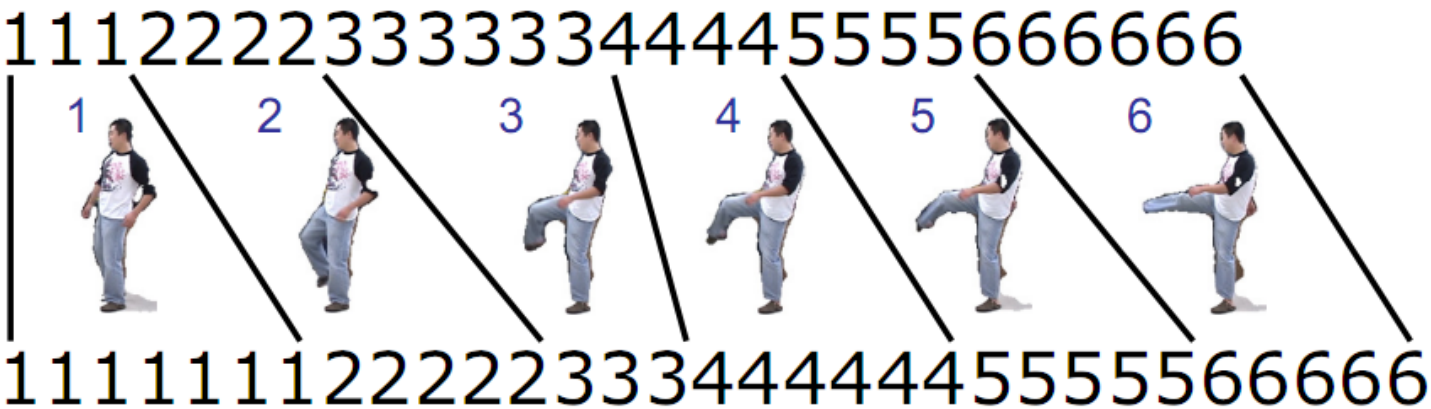




[Starner, T. and Pentland, A., Real-time American Sign Language recognition from video using hidden Markov models. International Symposium on Computer Vision, 1995.]

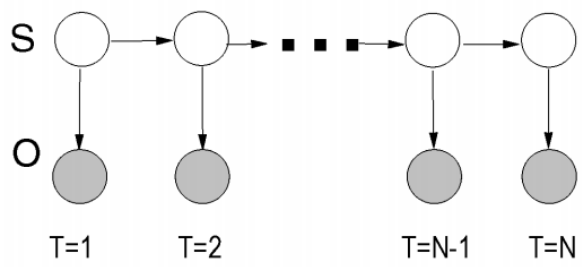
Dynamic time warping

- Dynamic programming algorithm to match two strings (e.g. sequences).
  - [Gavrila and L. Davis, 1995]
  - Each frame generates a symbol (of a feature vector)

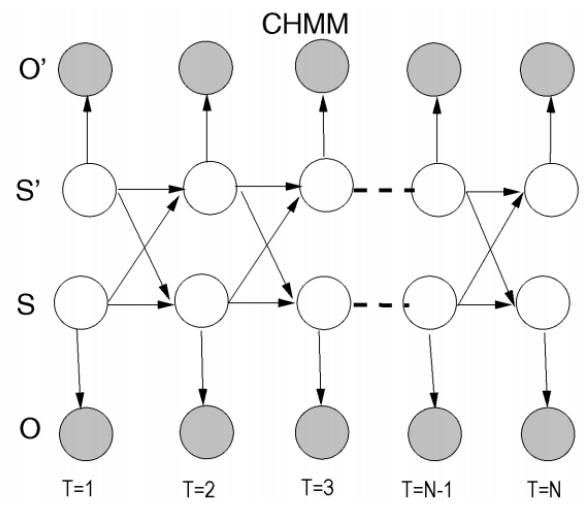


Coupled HMMs

- Pentland CHMMs
  - Human-human interactions
    - Two types of states for two agents
  - Synthetic agents for training HMMs



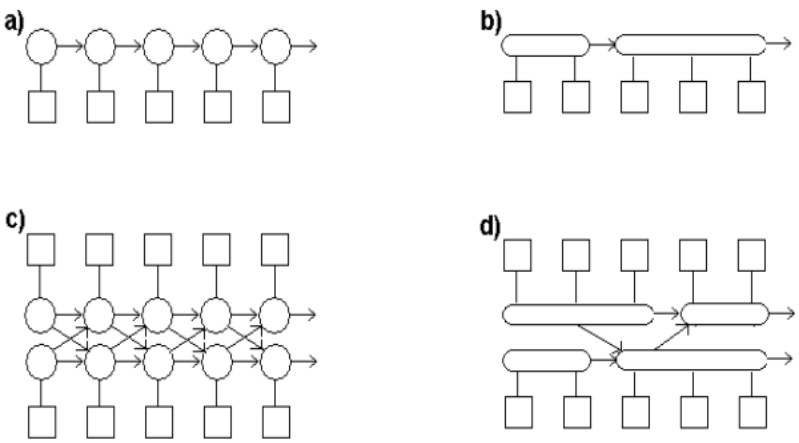
Vs.



[Oliver, N. M., Rosario, B., and Pentland, A. P., A Bayesian computer vision system for modeling human interactions. IEEE T PAMI, 2000.]

HMM Variations

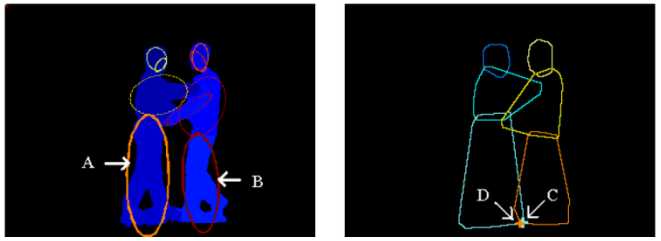
- Coupled hidden semi-Markov models
  - Natarajan and Nevatia 2007
  - Human-human interactions
  - Activities with varying durations
    - Models probabilistic distributions of state durations.

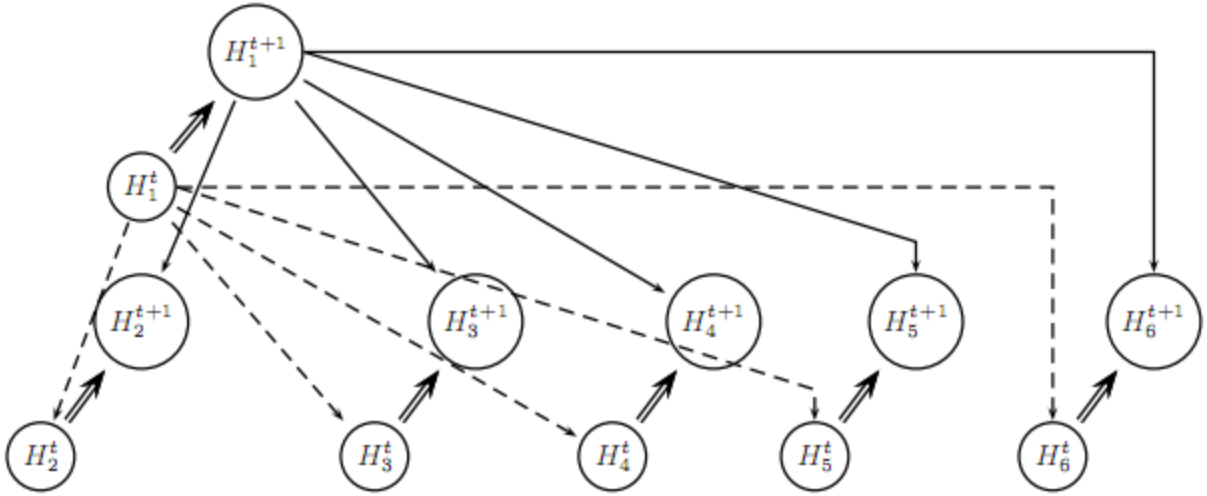


[Natarajan, P. and Nevatia, R., Coupled hidden semi Markov models for activity recognition. WMVC 2007]

Dynamic Bayesian networks

- Diverse variations
  - Dynamic Bayesian networks
    - Body-part analysis

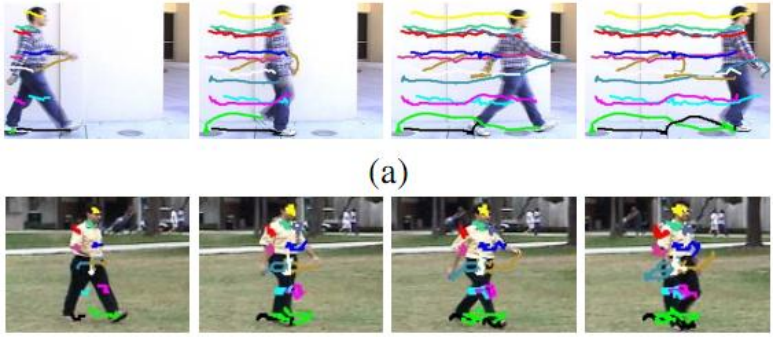




[Park, S. and Aggarwal, J. K., A hierarchical Bayesian network for event recognition of human actions and interactions. Multimedia Systems, 2004]

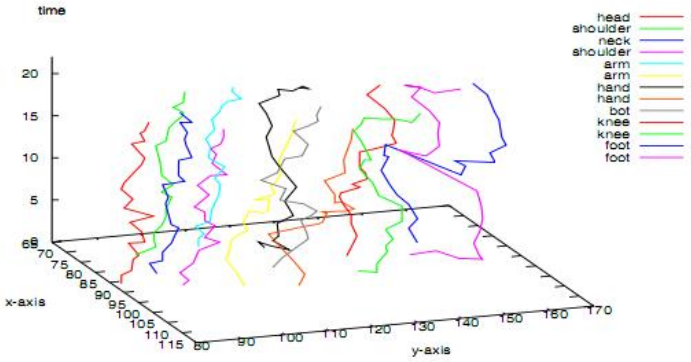
Space-time trajectories

- Trajectory patterns
  - Yilmaz and Shah, 2005 – UCF
  - Joint trajectories in 3-D XYT space.
  - Compared trajectory shapes to classify human actions.



(a)

(b)



(c)

[Yilmaz, A. and Shah, M., Recognizing human actions in videos acquired by uncalibrated moving cameras, ICCV 2005]

Sequential approaches - summary

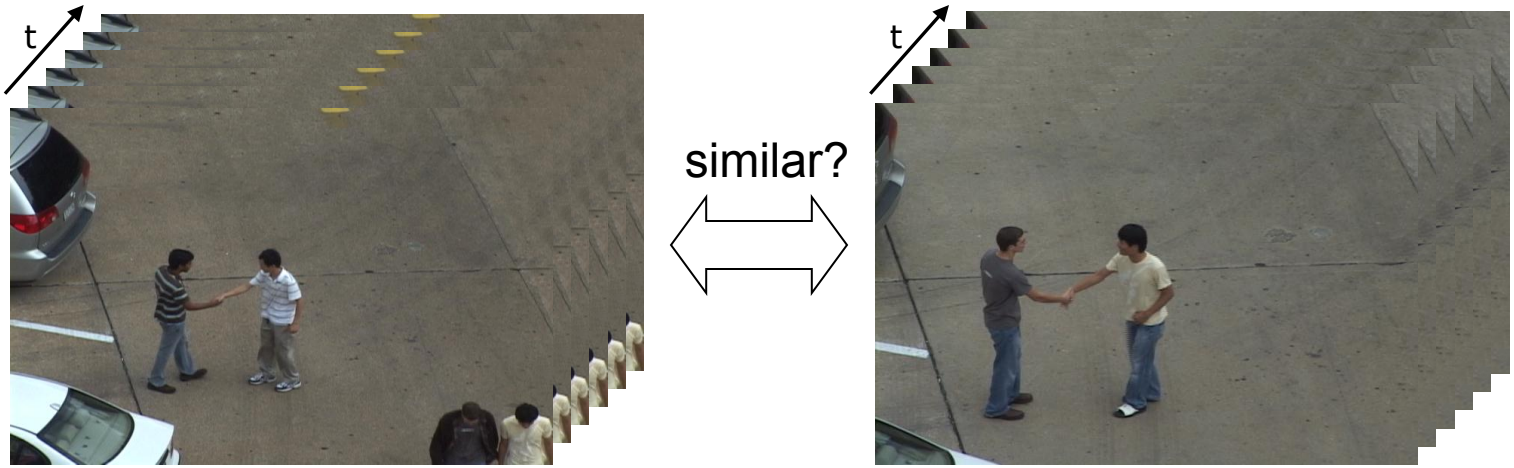
- Designed for modeling sequential dynamics
  - Markov process
  - Motion features are extracted *per frame*
- Limitations
  - Feature extraction
    - Assumes good observation models
  - Complex human activities?
    - Large amount of training data

Single layered approaches

Space-time approaches

Space-time approaches

- Videos as 3-D XYT volumes
- Problem: matching between two volumes
  - Match volumes directly
    - Compare volumes from testing videos with those from training videos.

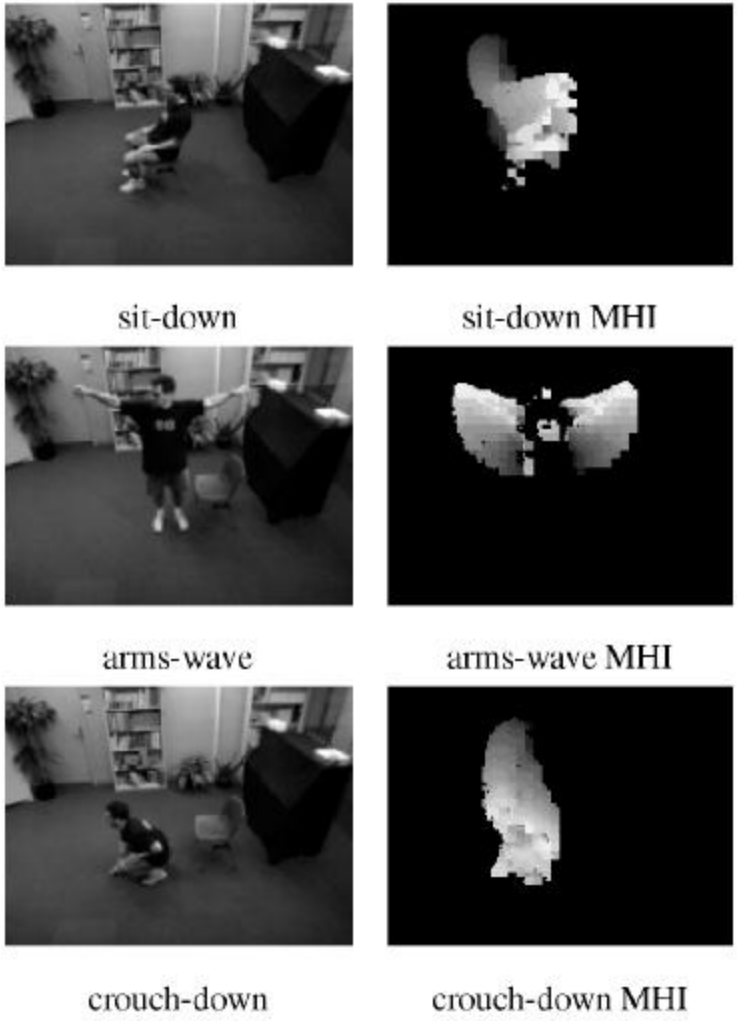


Training video

Testing video

Motion history images

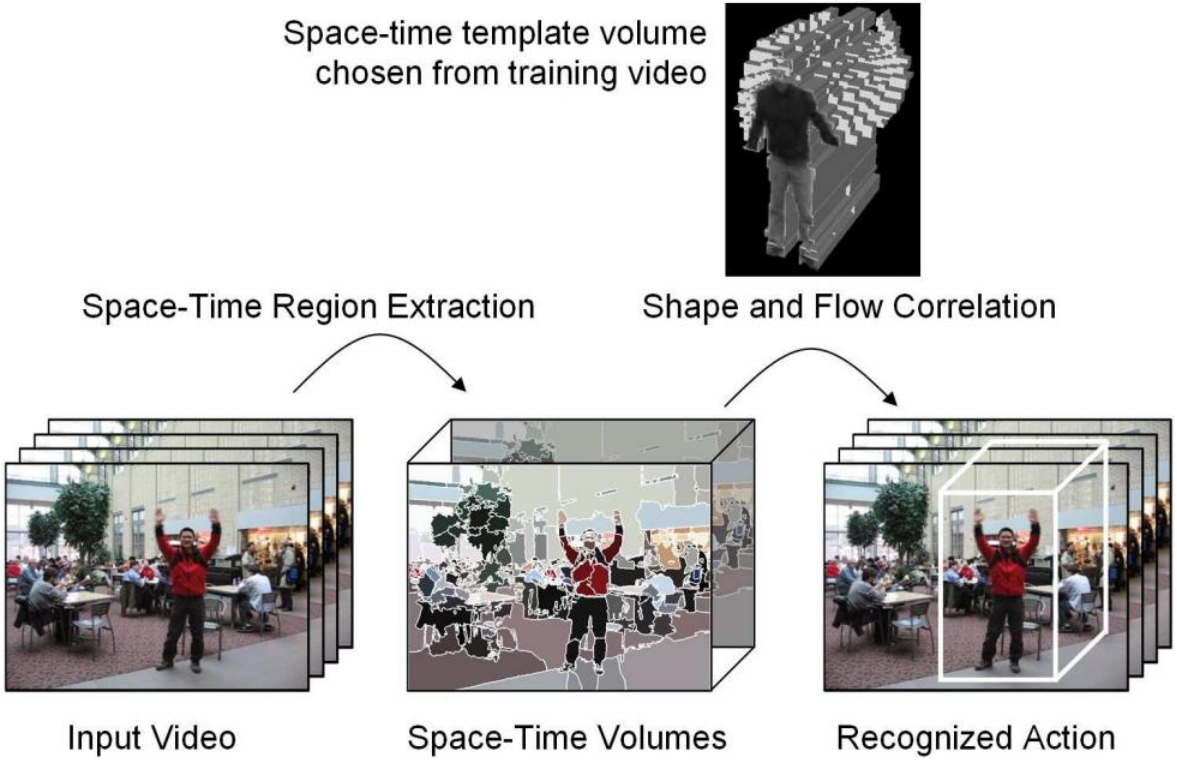
- Matching two volumes
- Bobick and J. Davis, 2001
  - Motion history images (MHIs)
  - Weighted projection of a XYT foreground volume
  - Template matching



[Bobick, A. and Davis, J., The recognition of human movement using temporal templates. IEEE T PAMI 23(3), 2001]

3-D volume matching

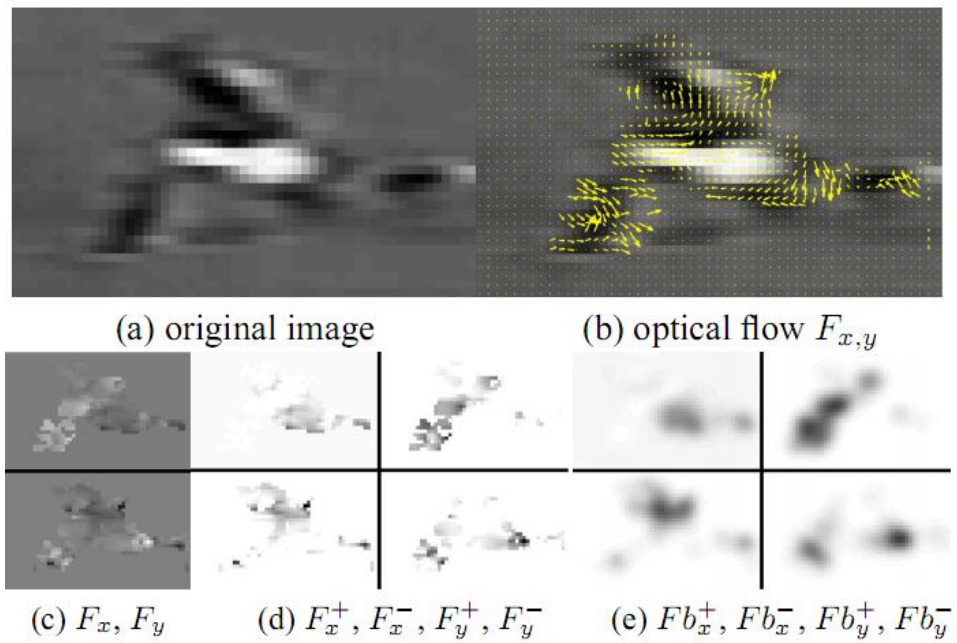
- Ke, Suktankar, Hebert 2007
  - Volume matching based on its segments.
  - Segment matching scores are combined.



[Ke, Y., Sukthankar, R., and Hebert, M., Spatio-temporal shape and flow correlation for action recognition. CVPR 2007]

Global features from volumes

- Efros et al. 2003
  - Concatenated optical flow features from 3-D XYT volumes
  - Analyzed soccer plays from low-resolution videos.

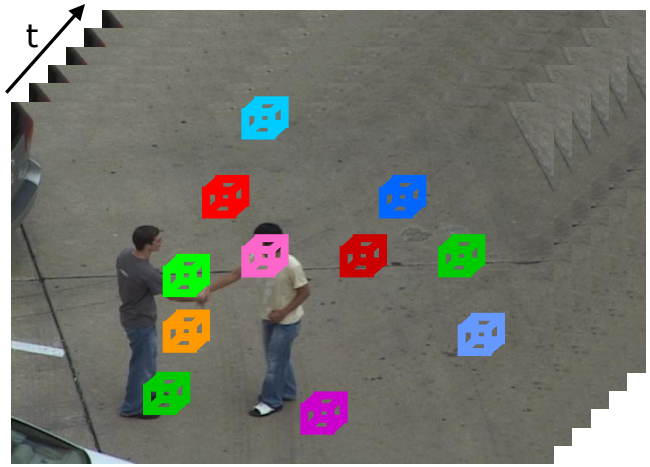


[Efros, A., Berg, A., Mori, G., and Malik, J., Recognizing action at a distance, ICCV 2003]

Sparse features from videos

- Problem: matching between two videos
  - Match volumes directly?
  - Extracts sparse features -
    - Video version of SIFT features

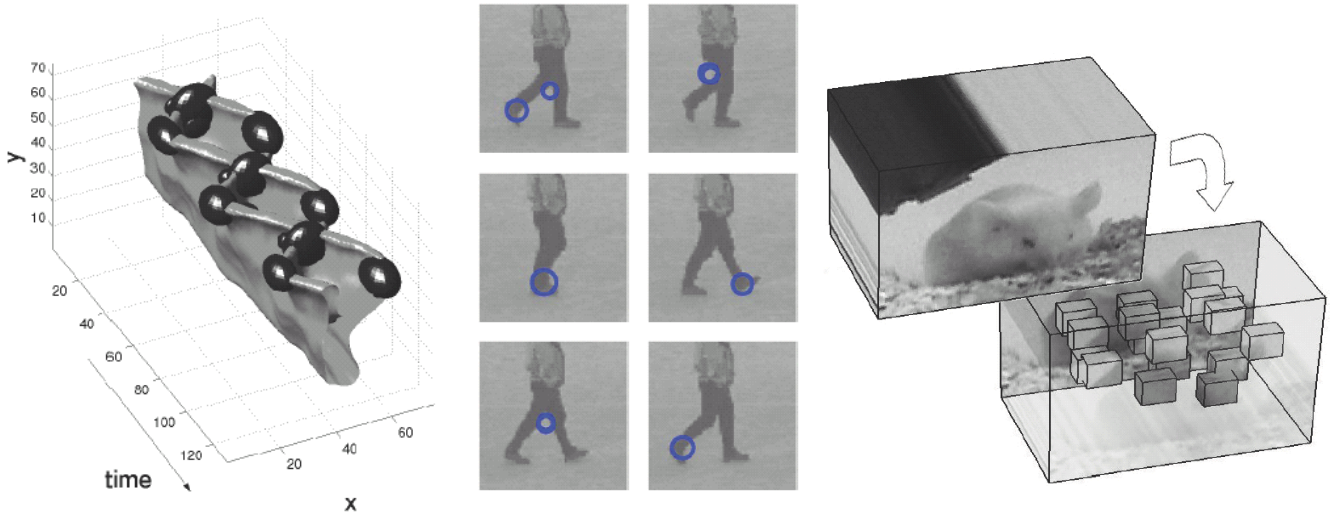
SIFT for images



Features for videos

Sparse features from videos

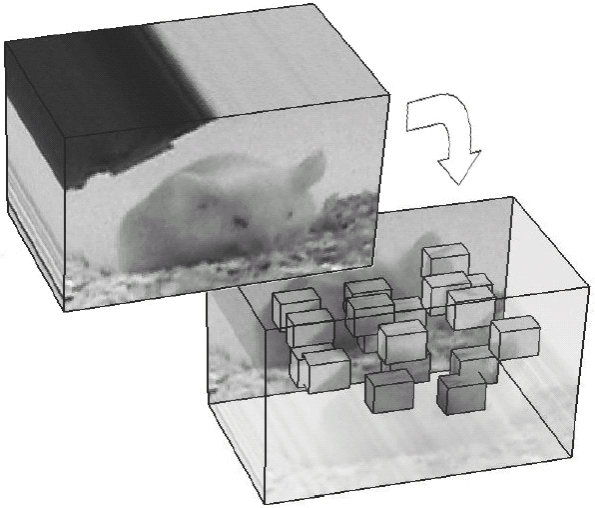
- Spatio-temporal features
  - Reliable under noise, background changes, lighting condition changes, …
  - Laptev 2003, Dollar et al. 2005



(a) (b) (c)

Cuboid features

- Cuboid descriptors
  - Dollar *et al.*, Cuboid, VS-PETS 2005
  - Appearances of local 3-D XYT volumes
    - Raw appearance
    - Gradients
    - Optical flows
  - Captures salient periodic motion.



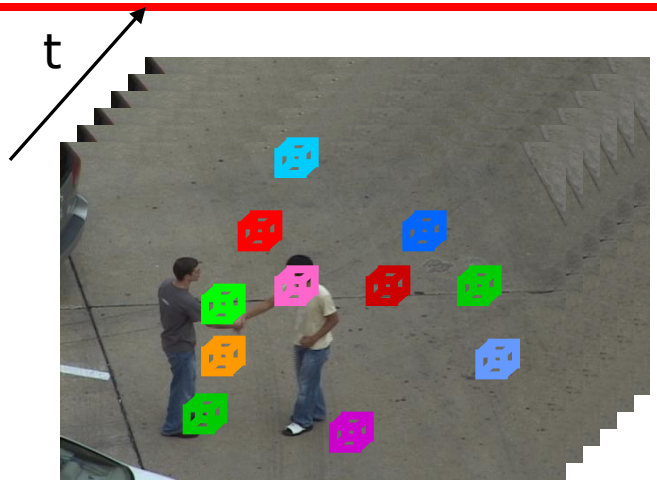
[Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S., Behavior recognition via sparse spatio-temporal features, VS-PETS 2005]

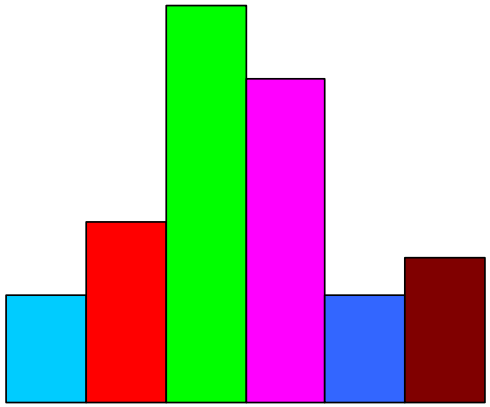
STIP interest point detector

- Laptev and Linderberg 2003
  - Simple periodic actions
    - Spatio-temporal local features + SVMs
  - Introduced the ***KTH dataset***
  - Local descriptor based on Harris corner detector

[Schuldt, C., Laptev, I., and Caputo, B., Recognizing human actions: A local SVM approach, ICPR 2004]

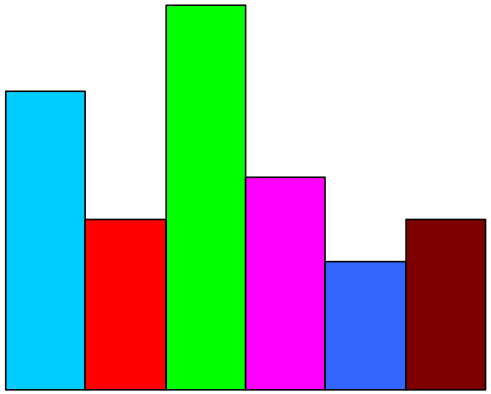
Bag-of-words representation





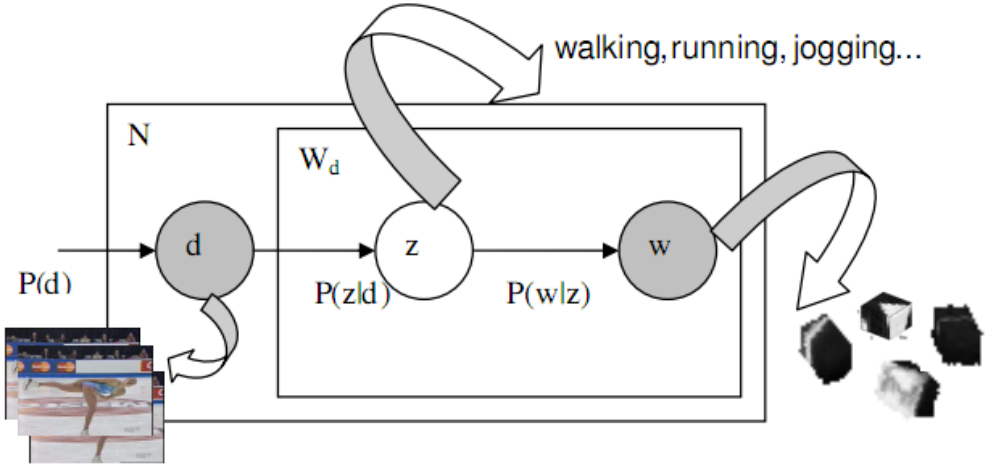
Classify features based on their appearance

Histogram (bag-of-words) similarity



pLSA models for actions

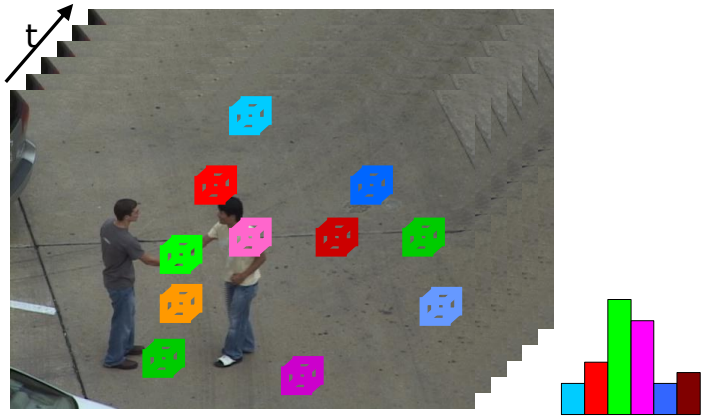
- pLSA from text recognition
  - Probabilistic latent semantic analysis
  - Reasoning the probability of features originated from a particular action video.



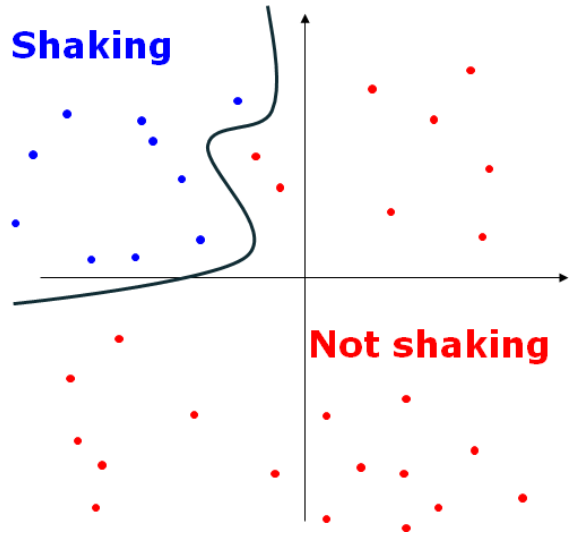
[Niebles, J. C., Wang, H., and Fei-Fei, L., Unsupervised learning of human action categories using spatial-temporal words, BMVC 2006]

Approach overview

- Recognition using local spatio-temporal features
  - Bag-of-words
  - Classifiers
    - e.g. SVMs, pLSA, …



- Extensions
  - Structural considerations
  - Hybrid features
  - Grouping features

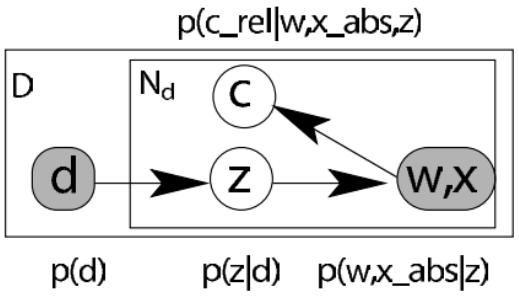


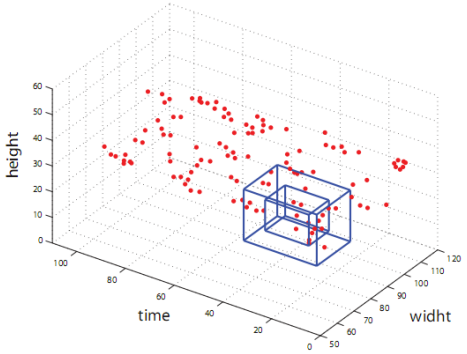
XYT approaches: pros and cons

- Advantages
  - Robust under noise
    - Background changes, camera movements, …
    - YouTube-type videos
- Limitations
  - Bag-of-words
    - Spatio-temporal relations among features are ignored.
  - Not hierarchical
    - Difficult to model complex activities

Structural considerations

- Bag-of-features ignores structure.
- Structures?
  - Wong et al. 2007
    - pLSA-ISM: encodes relative locations of features
  - Savarese et al. 2008
    - Feature correlation: pairwise proximity





[Wong, S.-F., Kim, T.-K., and Cipolla, R., Learning motion categories using both semantic and structural information, CVPR 2007]
[Savarese, S., DelPozo, A., Niebles, J., and Fei-Fei, L., Spatial-temporal correlatons for unsupervised action classification, WMVC 2008]

Problems

- Works well for simple single-human actions
  - Periodic movements
- Localization - sliding windows?

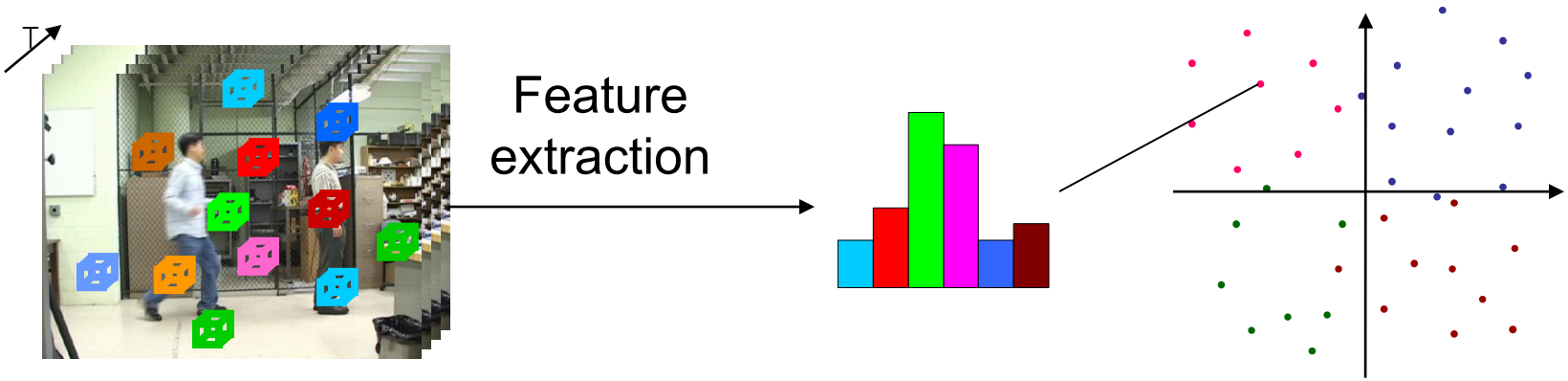
Vs.

Interactions *Pedestrians* *Multiple activities*

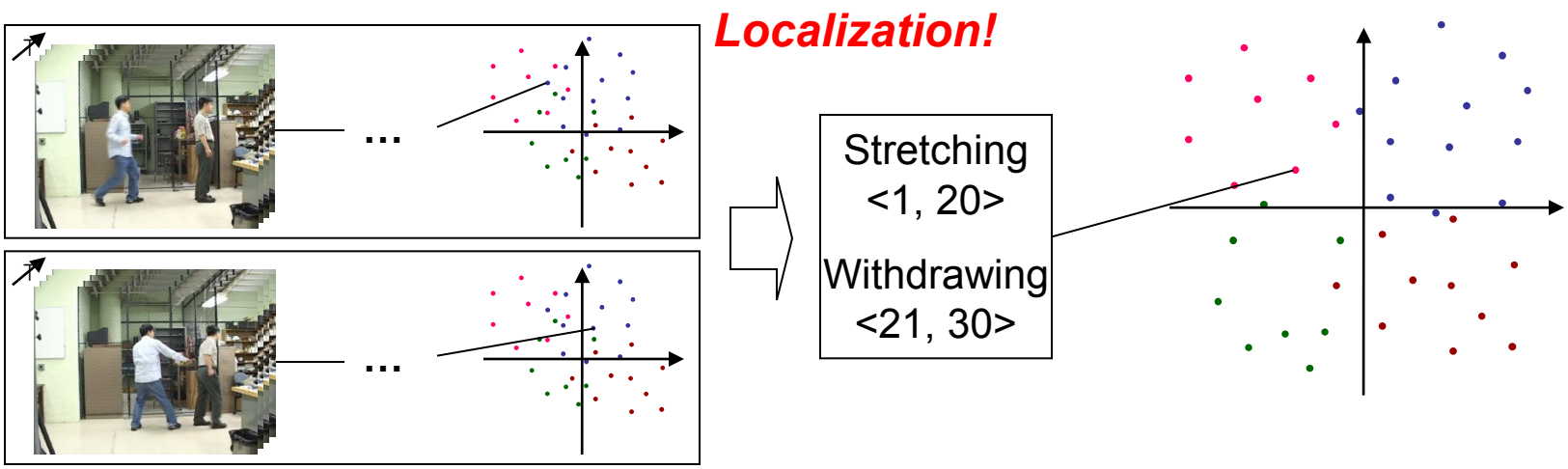
Hierarchical approaches

Hierarchical approach

- Single layered approach



- Hierarchical approach



Hierarchical approaches

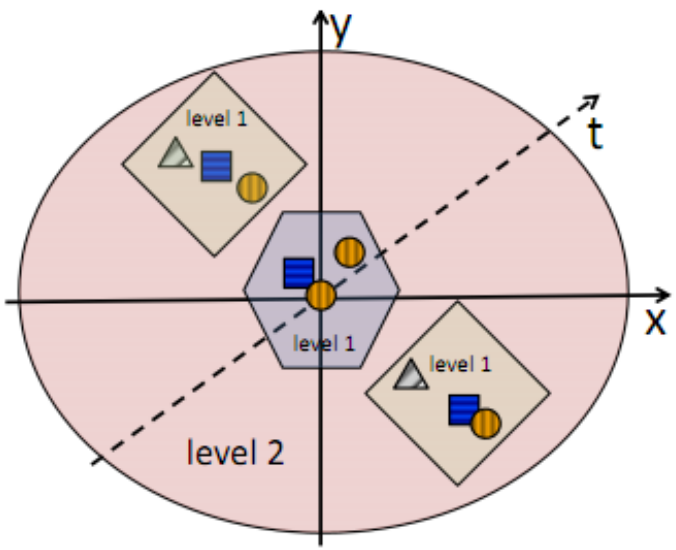
Mid-level recognition

What do we need to make an approach hierarchical?

- Hierarchical approaches
  - Recognition based on sub-events
  - Required information: when and where the sub-events are occurring
    - Detect time intervals
    - Detect bounding boxes
- Bag-of-words of space-time features?
  - Designed mainly for the classification
  - Detection – sliding window technique
    - Many false positives

Grouping features

- Groups a small number of features
  - 2~3 features which appear jointly
  - Spatially/temporally adjacent features
    - grouping



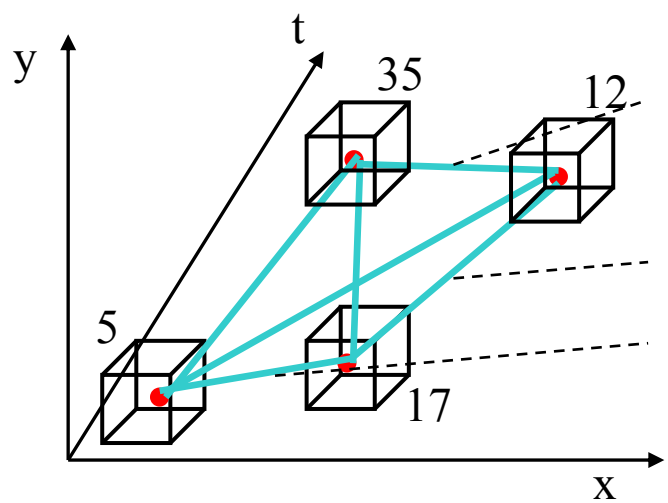
- Multiple levels
  - Multiple persons?
  - Pedestrians?

[Kovashka, A. and Grauman, K., Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, CVPR 2010]

Spatio-temporal relation match

- Spatio-temporal formulation
  - Histograms ignores feature locations
  - Many activities share common gestures
    - Pointing vs. punching

- Example: temporal relations



overlaps(12, 35)

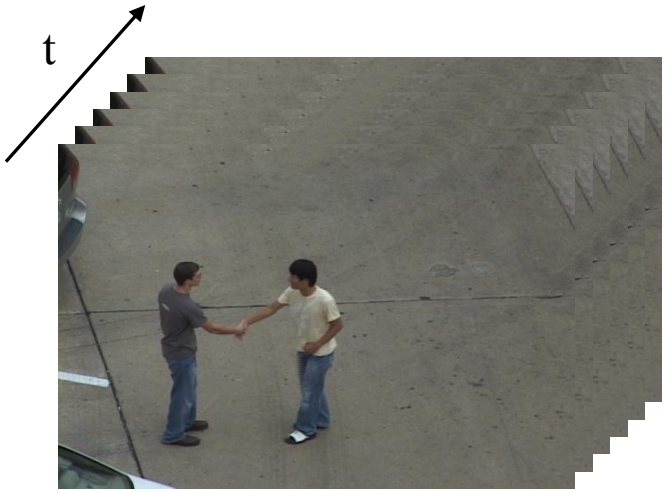
...

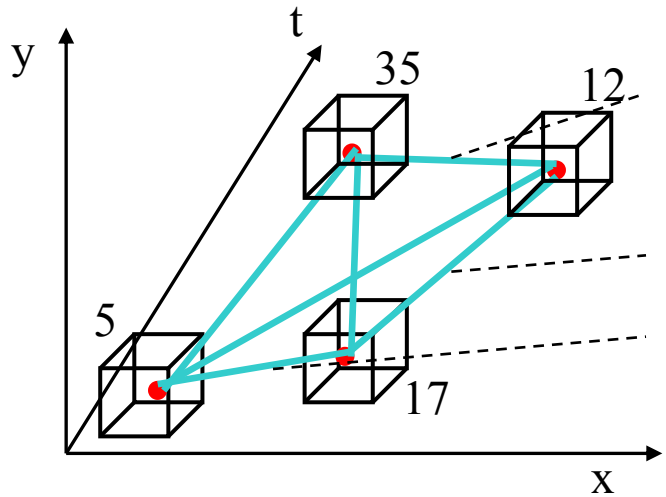
before(17, 12)

overlaps(5, 17)

[Ryoo, M. S. and Aggarwal, J. K., Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities, ICCV 2009]

Spatio-temporal relations (STRs)



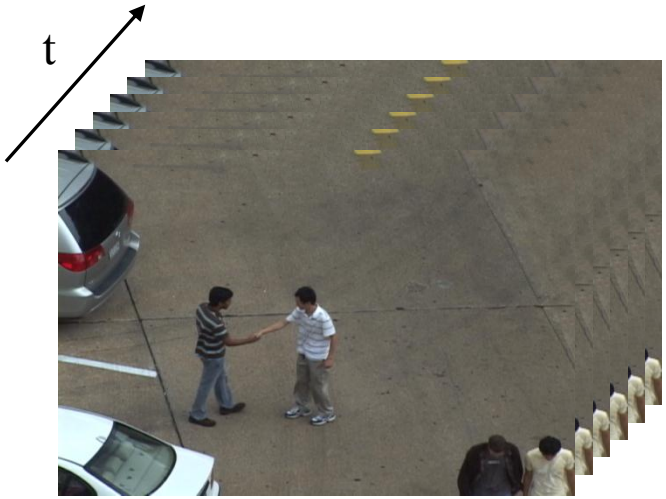


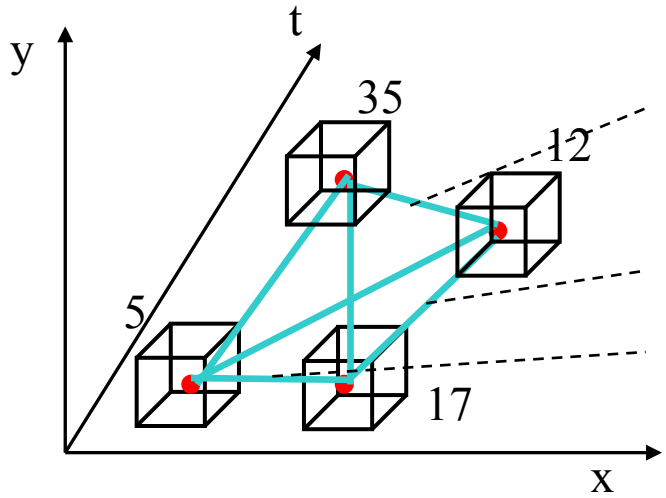
overlaps(12, 35)

...

before(17, 12)

overlaps(5, 17)





overlaps(12, 35)

...

before(17, 12)

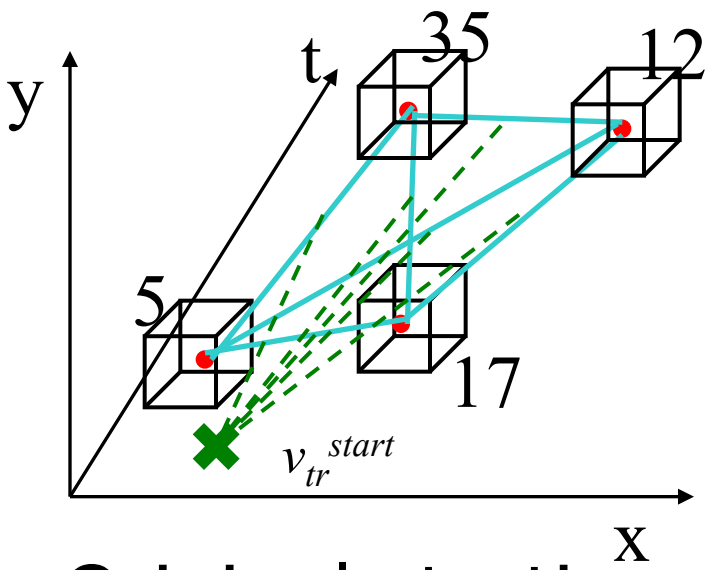
equals(5, 17)

Videos

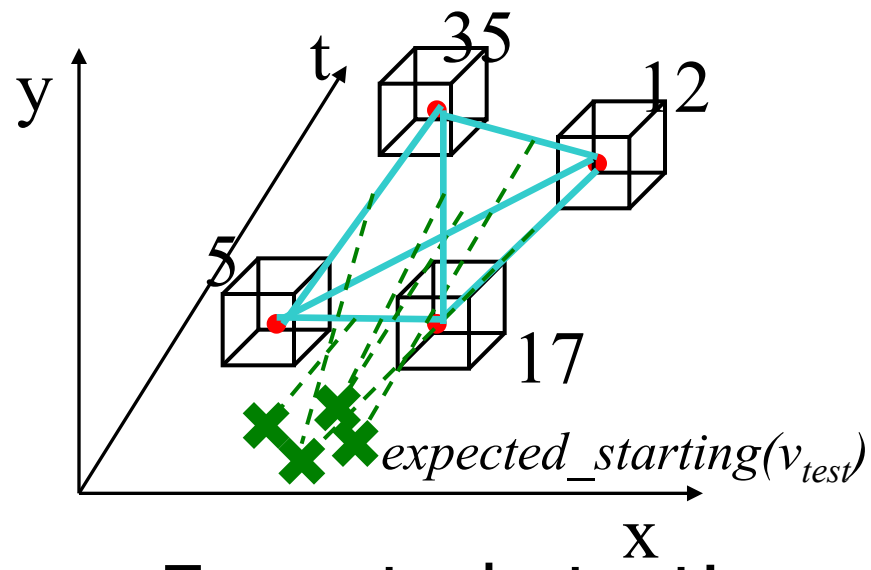
Feature relations

STR-match activity detection

- Must detect starting time and ending time
  - Models starting XYT location of an activity.
  - Each feature pair in a matching training video makes a vote.



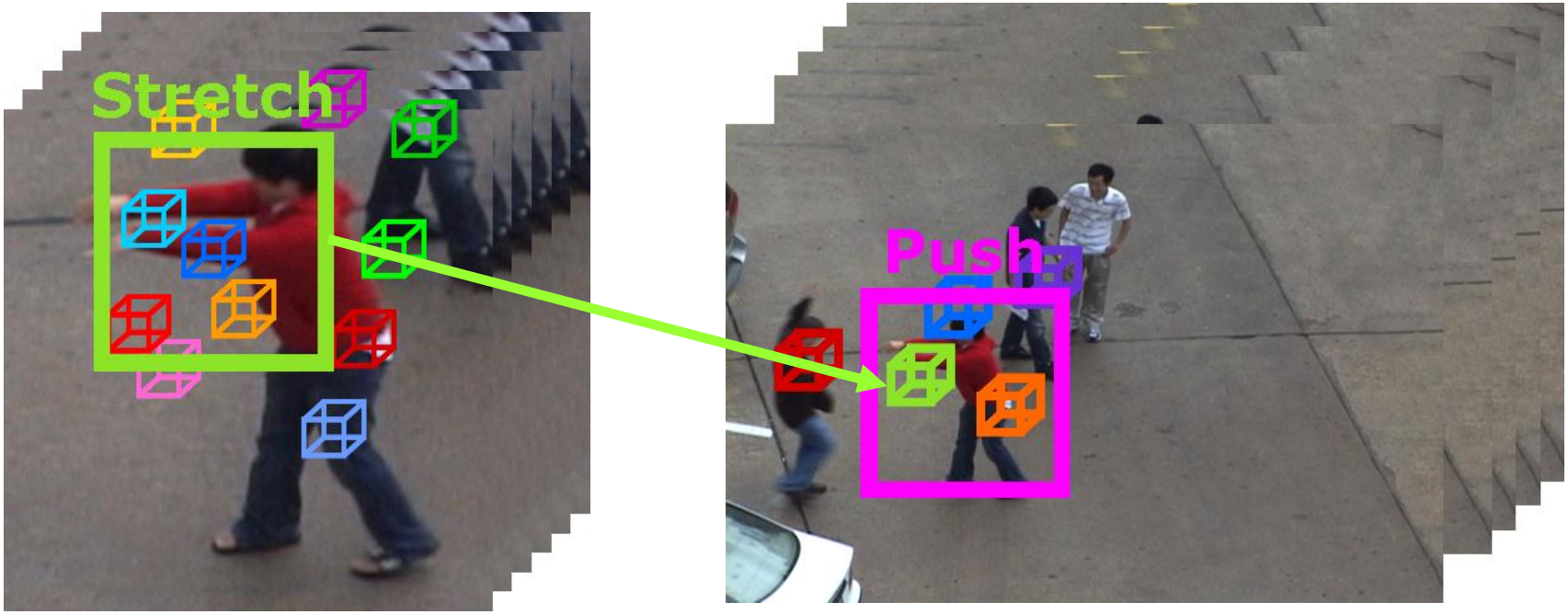
Original starting



Expected starting

Hierarchical recognition

- Atomic action detections as new features
  - Localization ability enables hierarchical recognition



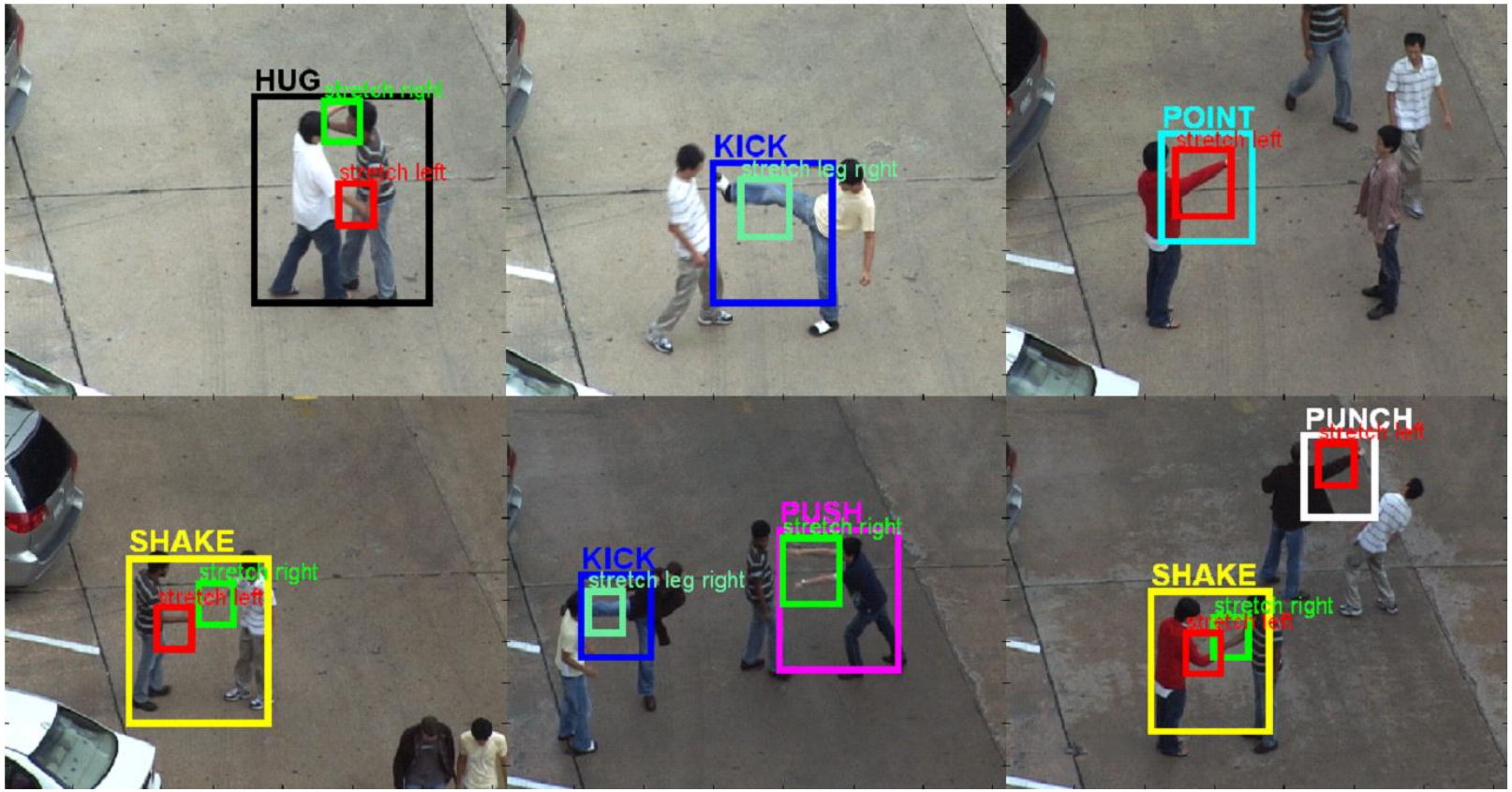
Experiments

- KTH dataset
  - Public dataset composed of simple actions
    - Walking, jogging, running, waving, …

System	Classification accuracy	Performance increase
Laptev et al. '08	91.8 / - %	+2.1%
Ours	**91.1 / 93.8 %**	**+12.6%**
Savarese et al. '08	- / 86.8 %	+5.6%
Niebles et al. '08	- / 81.5 %	+0.3%
Dollar et al. '05	- / 81.2 %	-
Schuldt et al. '04	71.7 / - %	-

Experiments: high-level activities

- High-level human activity detection results
  - Changing backgrounds, lighting conditions, …



Hierarchical approaches

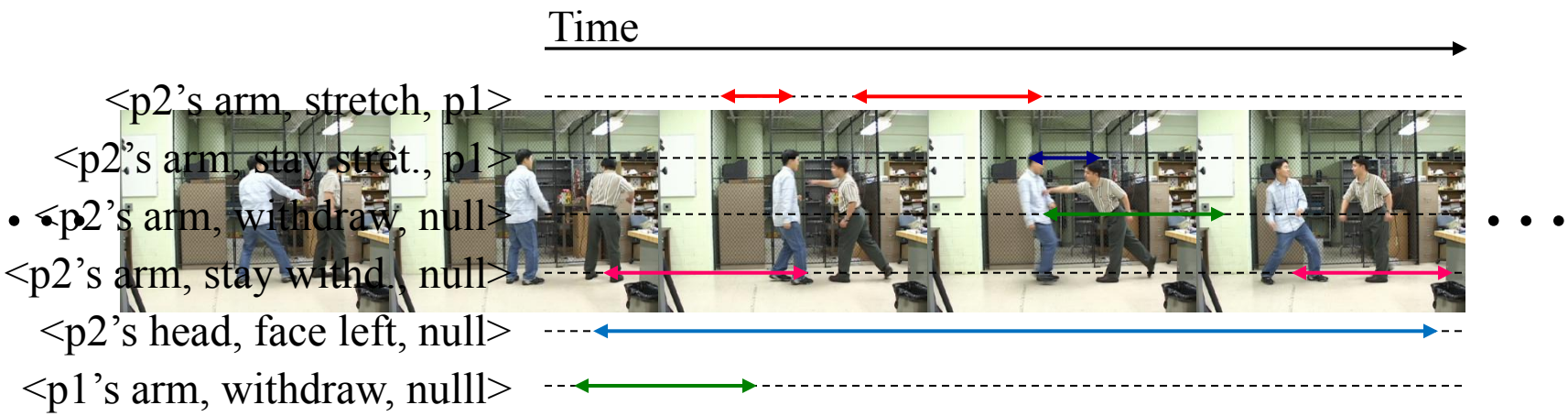
Semantic-level recognition

Semantic-level recognition

Machine-understandable representation of Punching

Punching is a sequence of hand stretch, withdrawal

Similar?

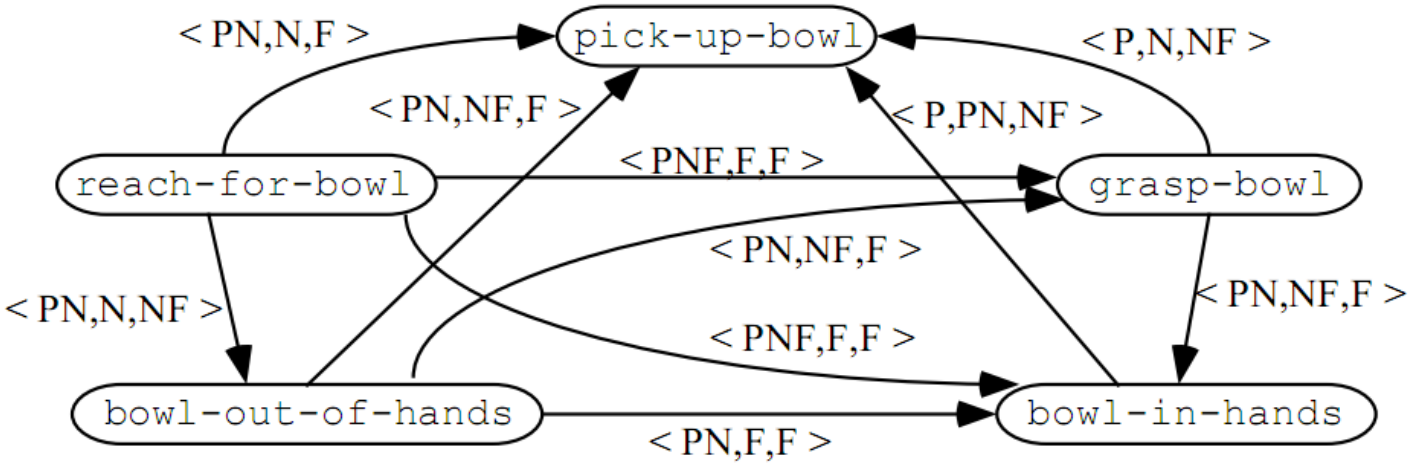


Time interval features

Video observations

Past-Now-Future networks

- Pinhanez and Bobick 1998
  - PNF networks to represent temporal structure of an activity.
  - Kitchen activities:

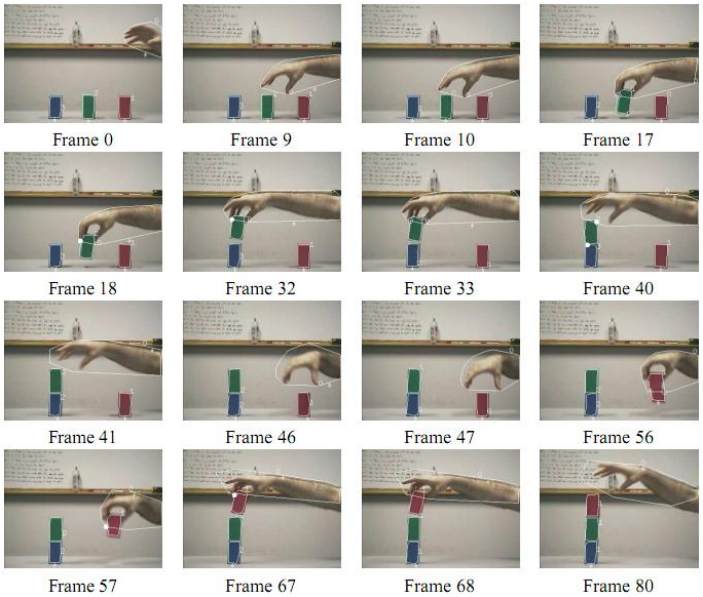


[Pinhanez, C. S. and Bobick, A. F., Human action detection using PNF propagation of temporal constraints. CVPR 1998]

Event logic

- Siskind 2001
  - Logical concatenations of predicates
  - Time intervals?

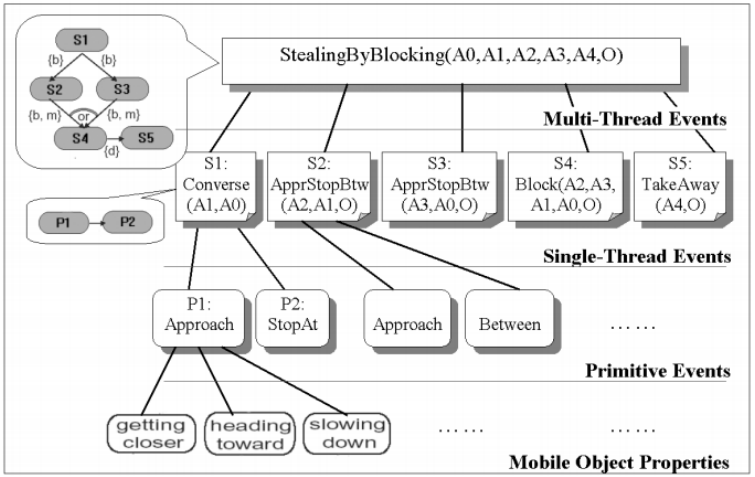
$$
\textsc{PickUp}(x,y,z) \triangleq \left(\begin{array}{l}
\neg\Diamond x = y \wedge \neg\Diamond z = x \wedge \neg\Diamond z = y \wedge \\
\textsc{Supported}(y) \wedge \neg\Diamond\textsc{Attached}(x,z) \wedge \\
\left\{ \begin{array}{l}
\left[\begin{array}{l}
\neg\Diamond\textsc{Attached}(x,y) \wedge \neg\Diamond\textsc{Supports}(x,y) \wedge \\
\textsc{Supports}(z,y) \wedge \\
\neg\Diamond\textsc{Supported}(x) \wedge \neg\Diamond\textsc{Attached}(y,z) \wedge \\
\neg\Diamond\textsc{Supports}(y,x) \wedge \neg\Diamond\textsc{Supports}(y,z) \wedge \\
\neg\Diamond\textsc{Supports}(x,z) \wedge \neg\Diamond\textsc{Supports}(z,x)
\end{array} \right] ; \\
[\textsc{Attached}(x,y) \vee \textsc{Attached}(y,z)] ; \\
\left[\begin{array}{l}
\textsc{Attached}(x,y) \wedge \textsc{Supports}(x,y) \wedge \\
\neg\Diamond\textsc{Supports}(z,y) \wedge \\
\neg\Diamond\textsc{Supported}(x) \wedge \neg\Diamond\textsc{Attached}(y,z) \wedge \\
\neg\Diamond\textsc{Supports}(y,x) \wedge \neg\Diamond\textsc{Supports}(y,z) \wedge \\
\neg\Diamond\textsc{Supports}(x,z) \wedge \neg\Diamond\textsc{Supports}(z,x)
\end{array} \right]
\end{array} \right\}
\end{array} \right)
$$



[Siskind, J. M., Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. Journal of Artificial Intelligence Research (JAIR) 15, 2001]

Representation languages

- Nevatia, Zhao, and Hongeng 2003
  - VERL - language



- Vu, Bremond, Thonnat 2003
  - Similar to Nevatia et al. 2003

```
Scenario(Bank_attack,
  Characters((cashier:Person), (robber:Person))
  SubScenarios(
    (cas_at_pos,inside_zone,cashier,"Back_Counter")
    (rob_enters, changes_zone, robber,
            "Entrance_zone", "Infront_Counter")
    (cas_at_safe, inside_zone, cashier, "Safe")
    (rob_at_safe, inside_zone, robber, "Safe") )
  ForbiddenSubScenarios(
    (any_in_branch, inside_zone, any_p,"Branch"))
  Constraints(
    Temporal ((rob_enters during cas_at_pos)
              (rob_enters before cas_at_safe)
              (cas_at_pos before cas_at_safe)
              (rob_enters before rob_at_safe)
              (rob_at_safe during cas_at_safe))
    Atemporal((cashier ≠ robber))
    Forbidden((any_p ≠ cashier) (any_p ≠ robber)
              (any_in_branch during rob_at_safe))))
```

Hierarchical activity representation

- Representation of the ‘shake-hands’ interaction

Shake hands interaction

CFG Syntax

```
InteractionDefine(i, j)
  →  InteractionName(i, j) " = " InteractionExp(i, j) ";"
InteractionName(i, j)
  →  name("person(i)", "person(j)")

Interaction(i, j)
  →  InteractionExp(i, j)
  |  InteractionName(i, j)
InteractionExp(i, j)
  →  "("InteractionDefs(i, j, var)", "
                InteractionRelationship(i, j, var)")"

InteractionDefs(i, j, var)
  →  "list" "(" "def" "("c", "Interaction(i, j)")" ", "
                InteractionDefs(i, j, var − c)")"
  |  "list" "(" "def" "("c", "Action(i or j)")" ", "
                InteractionDefs(i, j, var − c)")"
  |  "def" "("c", "Interaction(i, j)")"
  |  "def" "("c", "Action(i or j)")"
  |  "null"
InteractionRelationship(var)
  →  LogicalPredicate("
                InteractionRelationship(var)", "
                InteractionRelationship(var)")"
  |  TemporalPredicate("this", "var(a)")"
  |  TemporalPredicate("var(a)", "this")"
  |  TemporalPredicate("var(a)", "var(b)")"
```

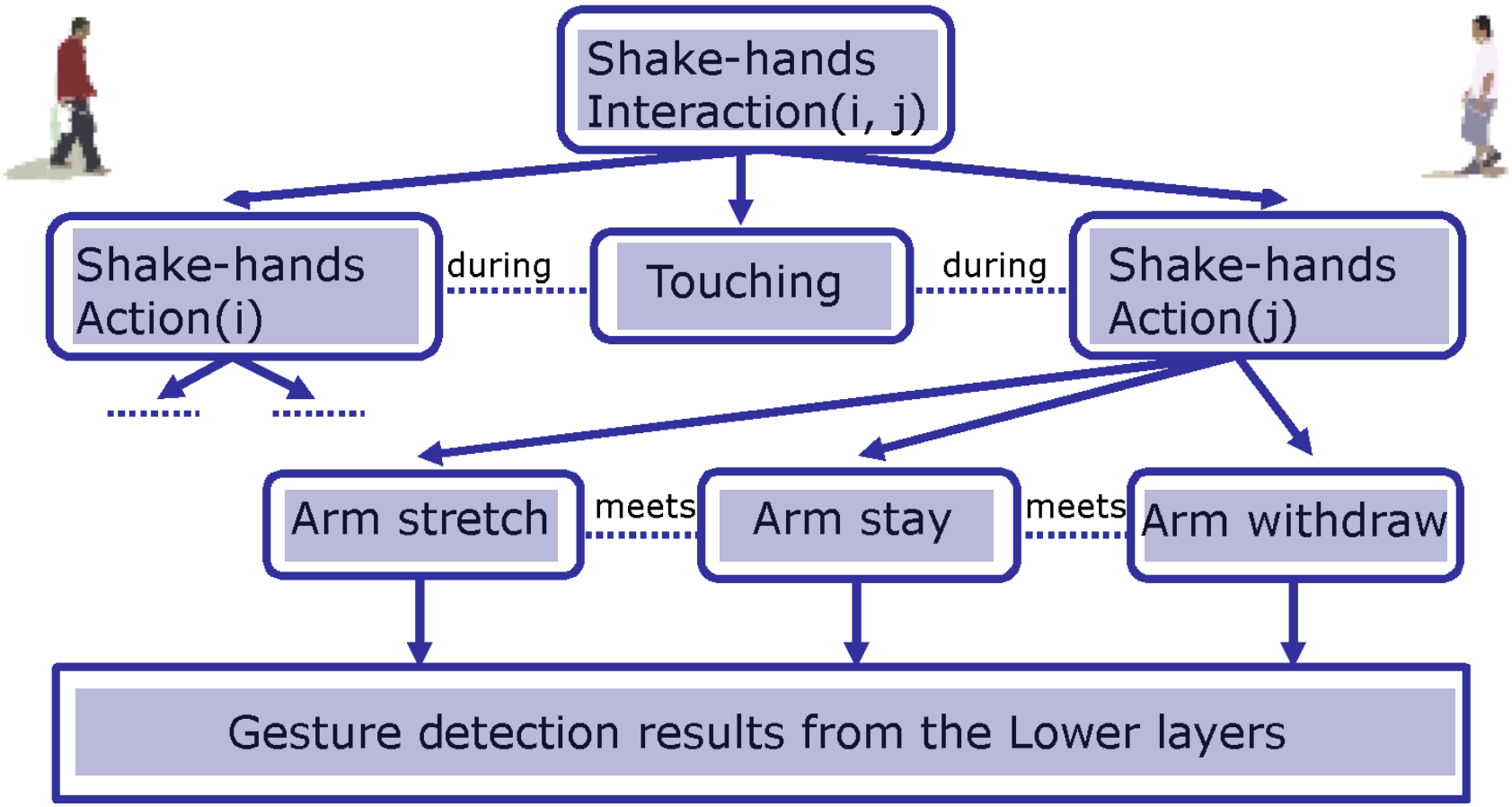
Hand shake = *two persons* do **shake-action** (**stretches**, **stays stretched**, **withdraw**) **simultaneously while touching**”

```
ShakeHandsInteractions(i, j) = (
    list( def(‘x’, ShakeHandsAction(i)),
          list( def(‘y’, ShakeHandsAction(j)),
                def(‘z’, TouchingInteraction(i, j))) ),
    and( and( during(‘z’, ‘x’), during(‘z’, ‘y’)),
         and( starts(‘z’, ‘this’), finishes(‘z’, ‘this’)))
);
ShakeHandsActions(i) = (
    list( def(‘x’, Arm_Stretch(i)),
          list( def(‘y’, Arm_Stay_Stretched(i)),
                def(‘z’, Arm_withdraw(i))) ),
    and( and( meets(‘x’, ‘y’), meets(‘y’, ‘z’)),
         and( starts(‘x’, ‘this’), finishes(‘z’, ‘this’)))
);
TouchingInteraction(i, j)=(null, touch(i, j, 0));
```

[Ryoo, M. S. and Aggarwal, J. K., Recognition of Composite Human Activities through Context-Free Grammar based Representation. CVPR 2006]

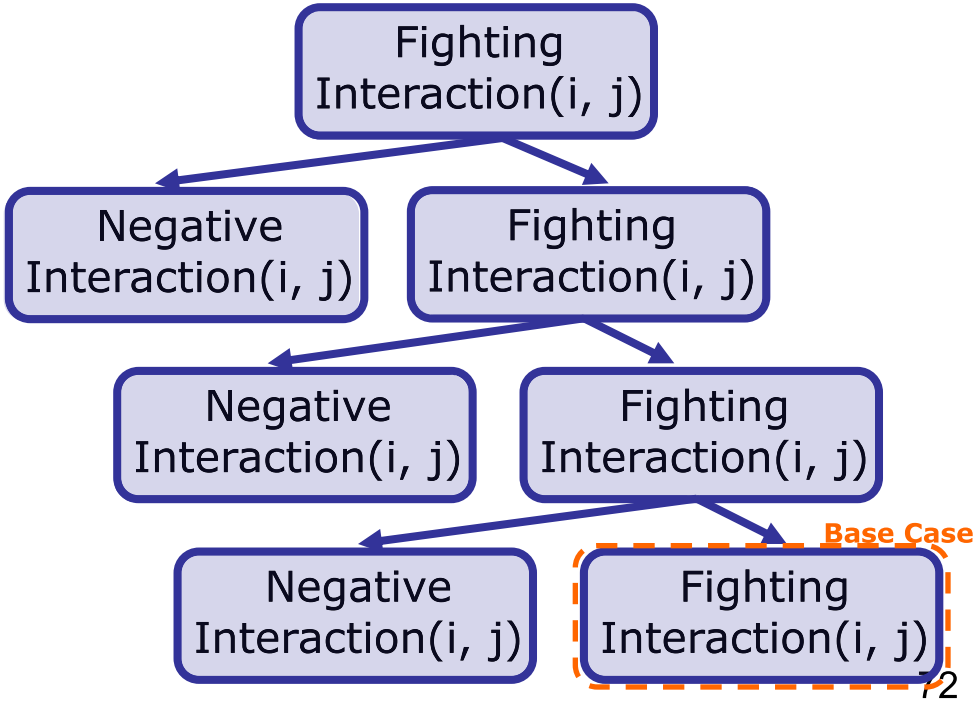
Hierarchical recognition algorithm

- Recognition process of the ‘Shake-hands’ interaction.



Continued and recursive activities

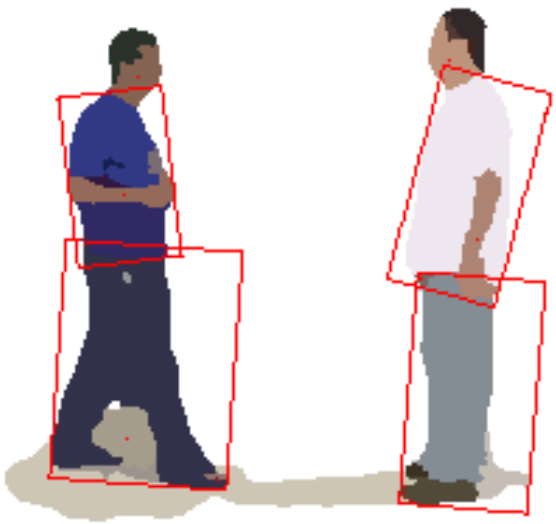
- Interaction ‘fighting’
  - Composed of multiple negative interactions
    - Punching + kicking + pushing + punching + …
  - Iterative approach is taken.



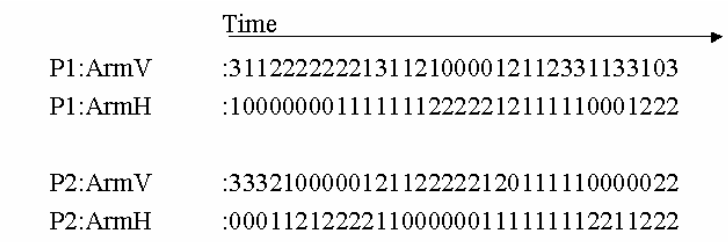
Example Experiment - Fighting

Input video:

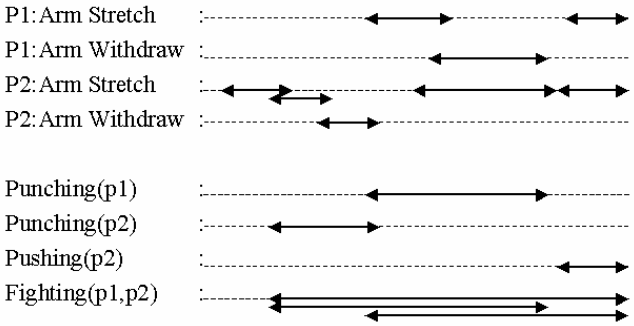
Processed video:



Poses:

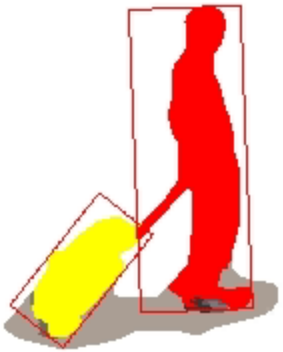


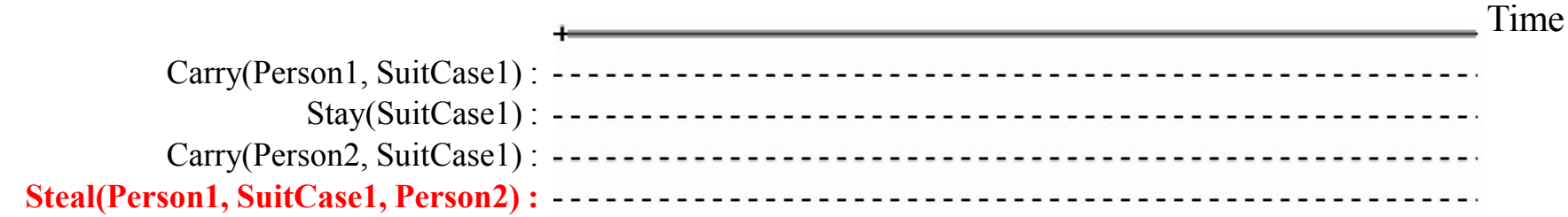
Gestures and activities:



Experiments

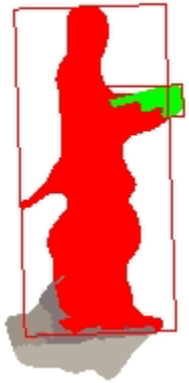
- Recognized following six types of interactions.
  - Each activity was tested with at least 10 sequences.
    - Carrying a box, leaving a box, placing a box into a trash bin.
    - Carrying a suitcase, leaving a suitcase, stealing the suitcase.
  - Object and Motion layer trained with 5 sequences.

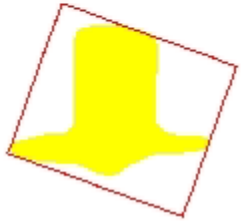


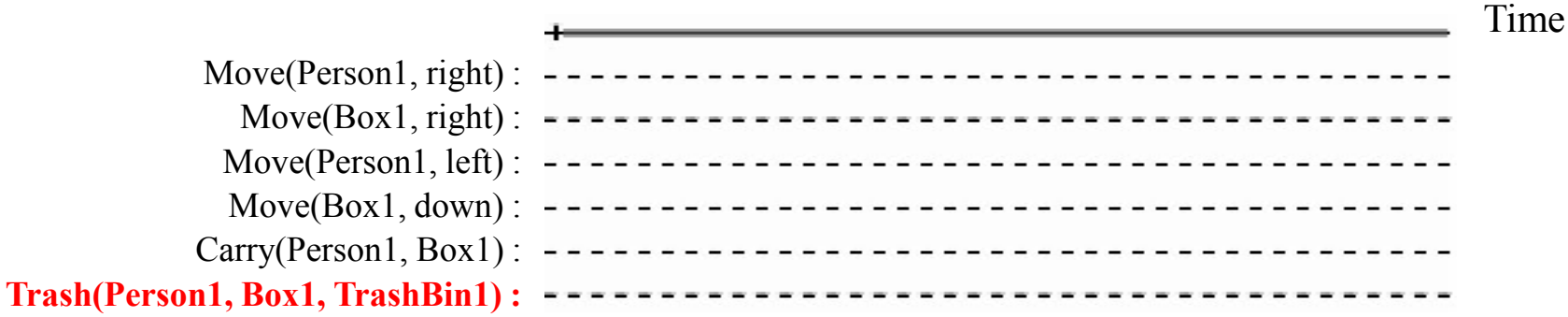


Experiments

- Example
  - a person placing a box into a trash bin

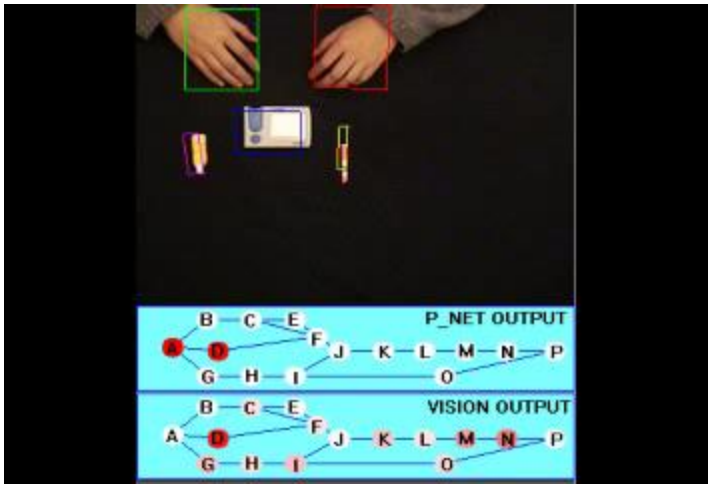


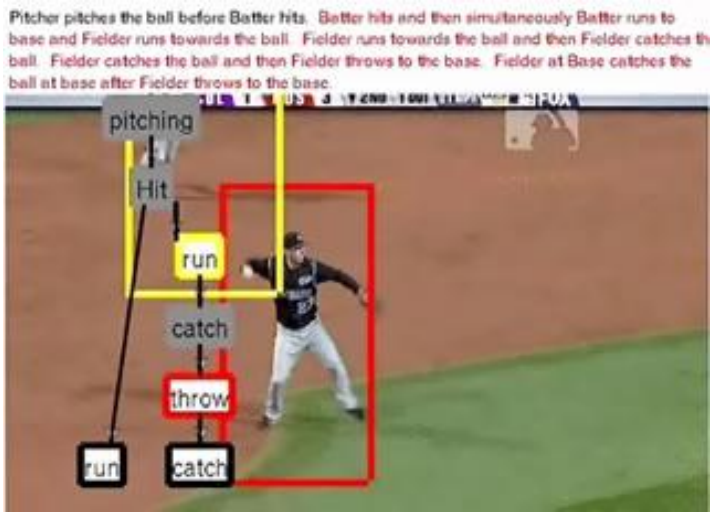




Hierarchical approaches

- Other examples:
  - Graph-based representation
    - P-Nets from Gatech
  - AND-OR tree representation
    - [Gupta et al., 2009]





Hierarchical approaches

Group activity recognition

Group activity

- Events performed by ***groups***
  - Various types of complex activities
    - Group-person interaction
    - Group-group interaction
- Uncertain nature
  - ***Varying*** # of participants
  - ***Dynamic*** spatial relation

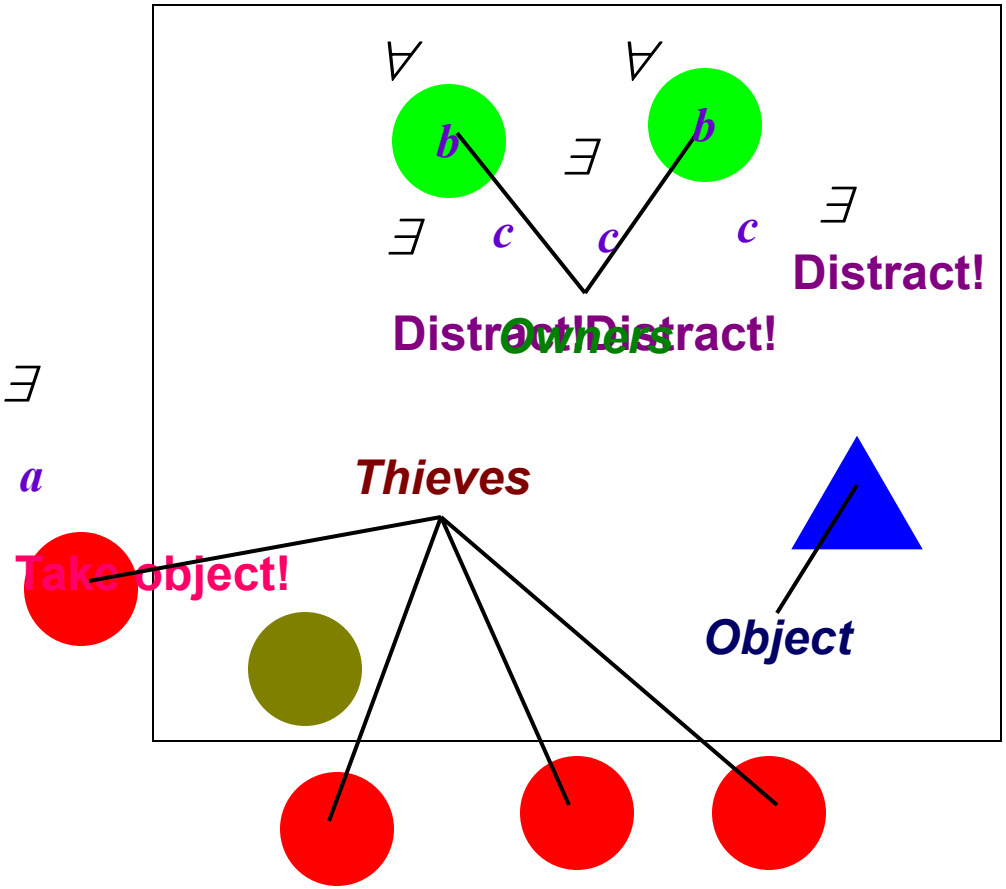
Group Activity (grp vs. grp)

A person with red shirts is **taking** the laptop on the table while the others are **talking**

[Ryoo, M. S. and Aggarwal, J. K., Stochastic Representation and Recognition of High-level Group Activities, IJCV 2011]

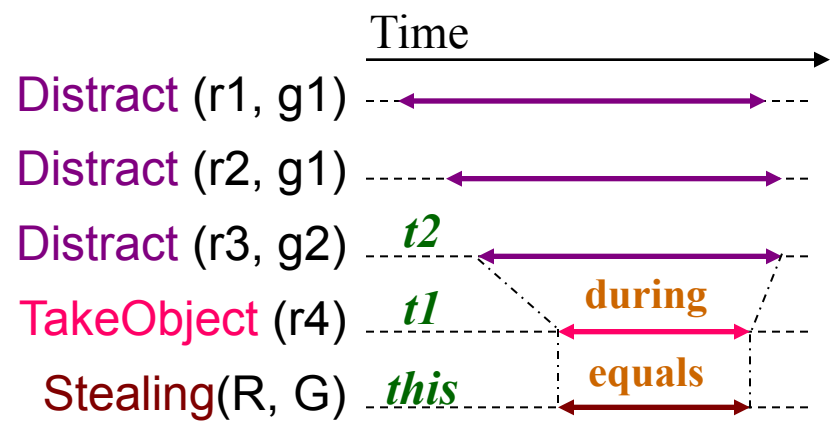
Representation

- Group stealing



Formal representation:

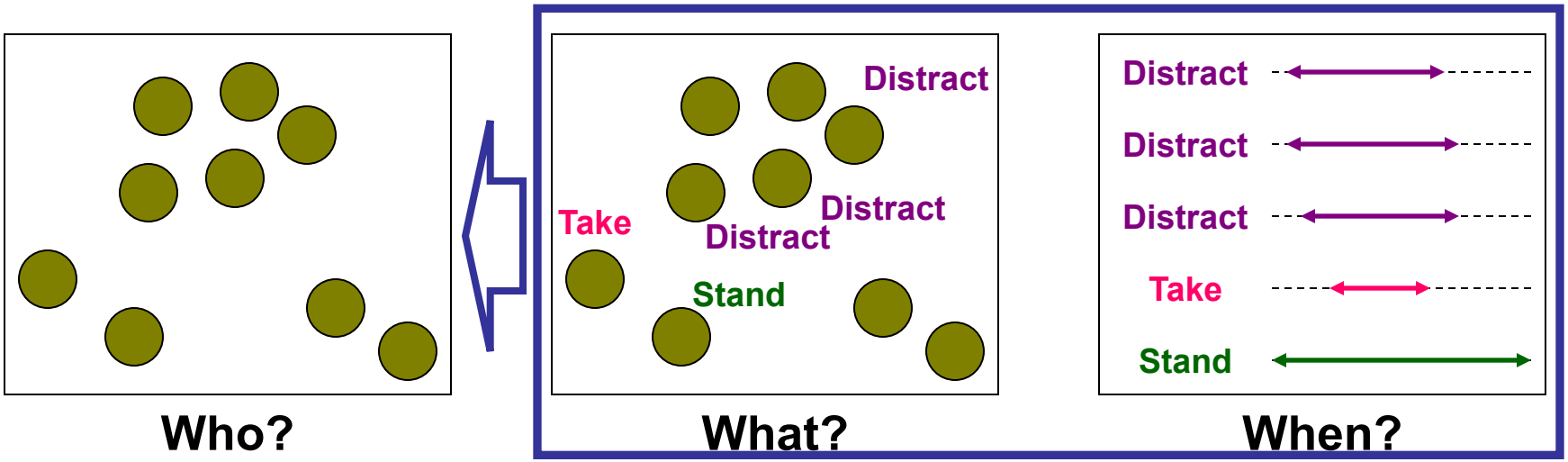
1. Member variables: ∀ *b* in *Owners*, ∃ *c* in *Thieves*
2. Time intervals: def(*t1*, TakeObject(*a*)), def(*t2*, Distract(*c*, *b*))
3. Predicates: equals (*t1*, *this*), during (*t1*, *t2*)



Time intervals of activities of individual members

Recognition overview

- 3 key components



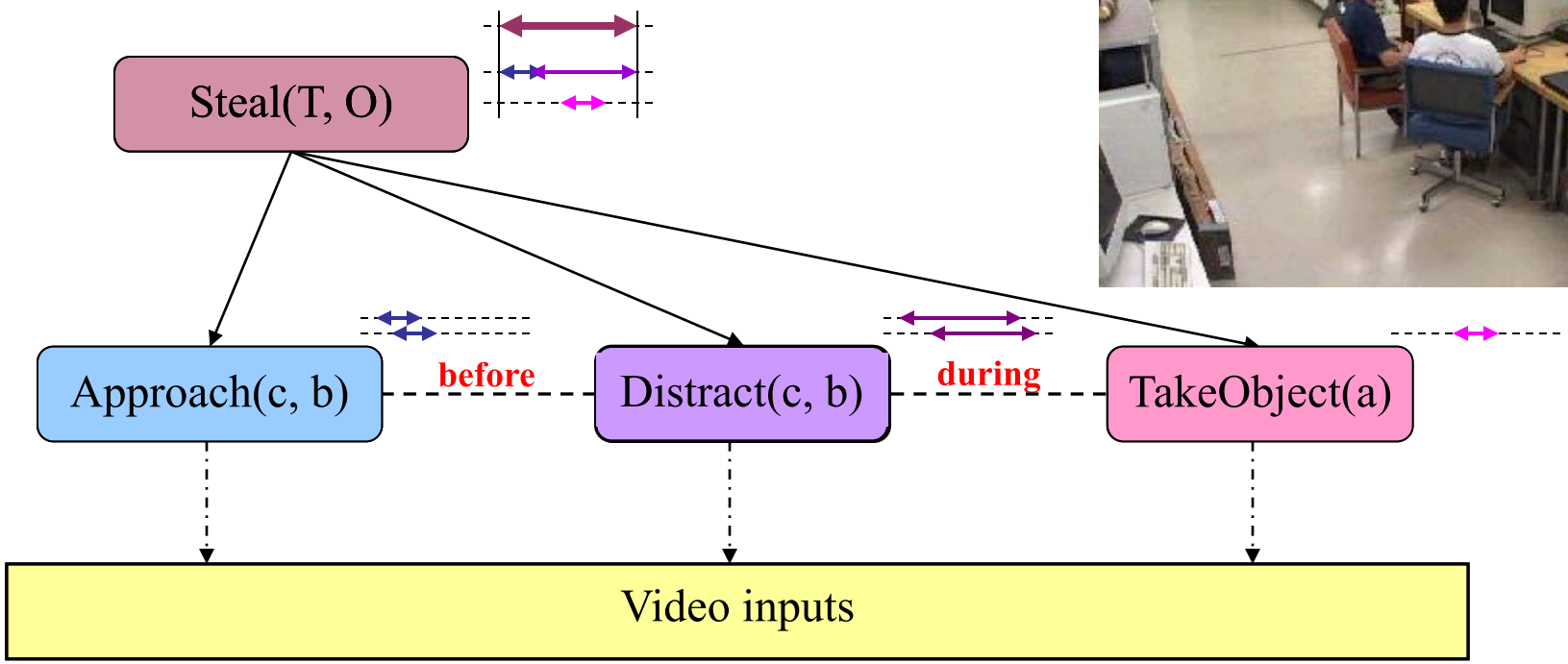
- Recognition: *NP-hard.* **Approximation required.**
  - Obtain a pool of group member candidates with non-zero probability.
    - Not many persons perform sub-events.
  - $M^* = \arg\max_M P(G^t(M) \mid O^M)$

Temporal constraints

- Hierarchical temporal constraint matching.

Member variables:

$\exists$ *a* in *Thieves*, $\forall$ *b* in *Owners*, $\exists$ *c* in *Thieves*

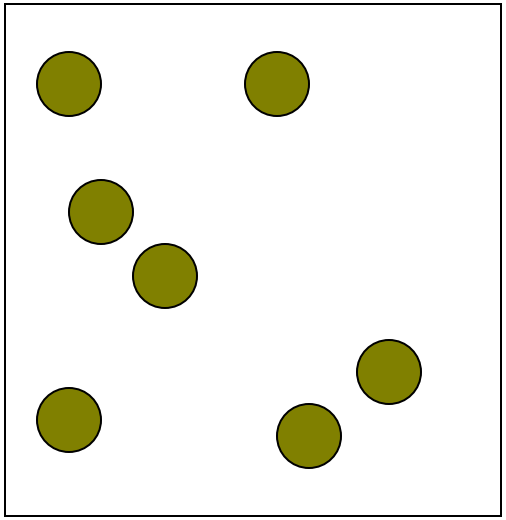


Group candidates

- Among possible groupings,
  - Find a set of group members:

    $$M = \{m_1, m_2, \ldots, m_{|M|}\}$$

    which maximizes the overall probability.



- Bayesian formulation

$$P(G^t \mid O) = \max_M P(G^t(M) \mid O^M)$$

$$= \max_M \frac{\pi_G(M)}{\pi_G(M) + \pi_{\neg G}(M)}$$

where $\pi_G(M) = P(O_M \mid G^t(M)) \cdot P(G^t(M))$

Bayesian formulation

- C_i: persons performing i^{th} sub-event.

$$P(O_M \mid G^t(M)) = \sum_{S_1^{t1},\ldots,S_n^{tn}} P(O_M \mid M, Q, S_1^{t1},\ldots,S_n^{tn}) \cdot P(S_1^{t1},\ldots,S_n^{tn} \mid G^t(M))$$

Represented relations Represented sub-events

$$\pi_G(M) = \sum_{S_1^{t1},\ldots,S_n^{tn}} \left[\prod_{rel} P(rel \mid S_a, S_b) \cdot \prod_i P(S_i^{ti} \mid G^t(M)) \cdot \prod_i d \cdot e^{-(|K_i - C_i|/|K_i| + |L_i \cap C_i|/|K_i|)} \cdot G^t(M) \right]$$

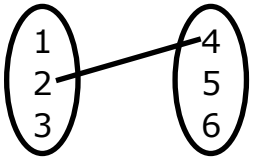
Structural similarity

- Essential and anti-essential relations: K_i, L_i

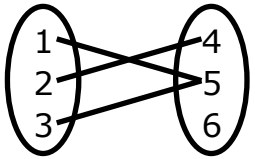
$$|K_i - C_i| = \sum_{k \in K_i - C_i} E[S_i^{ti}(k) \mid O^i]$$

$$|L_i \cap C_i| = \sum_{l \in L_i \cap C_i} E[S_i^{ti}(l) \mid O^i]$$

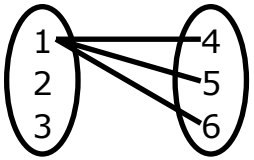
Case1: $\exists\,\exists$



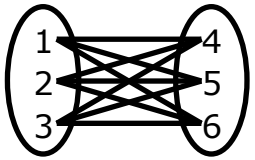
Case2: $\forall\,\exists$



Case3: $\exists\,\forall$



Case4: $\forall\,\forall$



Markov chain Monte Carlo

- MCMC-based probability estimation.
  - Provides a set of samples from the distribution.
  - Models the probability distribution.
- Metropolis-Hastings algorithm
  - $P(M_{t-1}, M') = \min(1, a)$
  - $a = \dfrac{\pi_G(M') \cdot q(M', M)}{\pi_G(M) \cdot q(M, M')}$
- Actions:
  - Add: $M' = M_{t-1} \cup \{m\}$ where $m \in C_i$
  - Remove: $M' = M_{t-1} - \{m\}$

Experimental setting

- We have tested 45 sequences of 8 activities.
  - 320*240 with 10 fps
- CCTV videos download from *YouTube*.
  - Group **stealing** in Malaysia and group **arresting** in UK.
- Videos that we have taken with 10 participants in various environments.
  - A group of people **carrying** a large object.
  - A group of people **assaulting** a person.

Videos of real human activities

Experiments - stealing

- Group stealing
  - One of thieves steals a laptop, while the other thieves are distracting the shop owner.

Thieves

Owners

Laptop

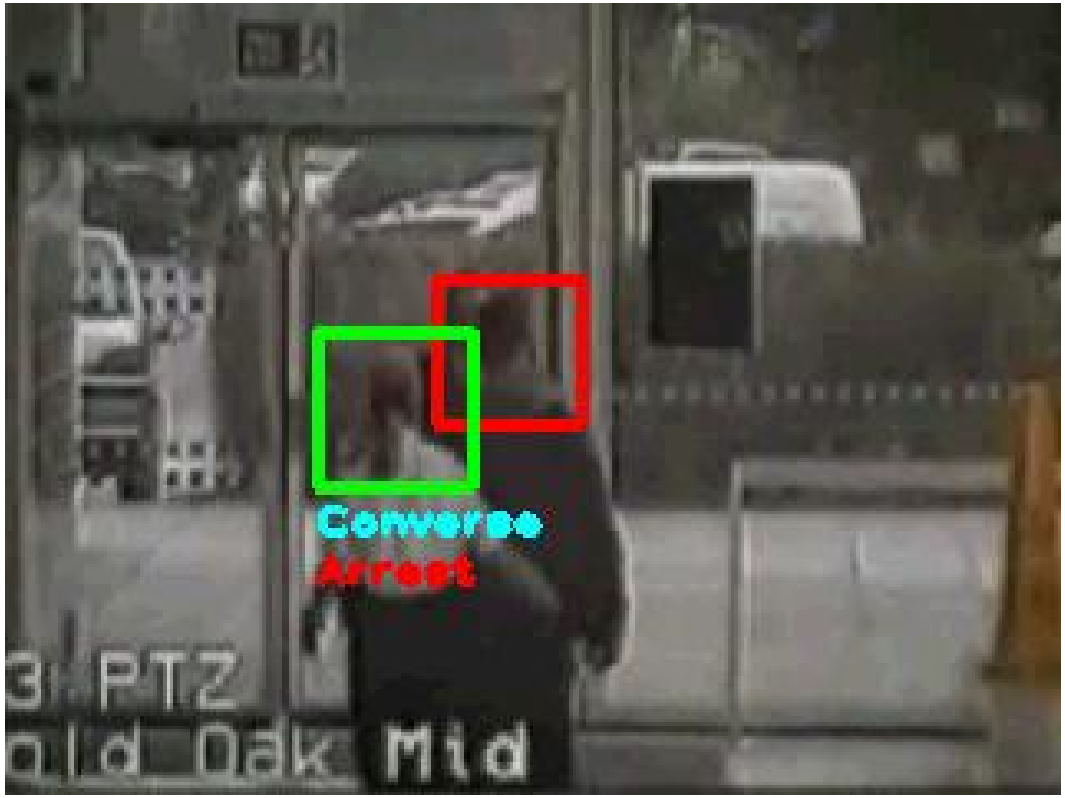
Experiments - arresting

- Group arresting
  - A group of policemen arresting a group of suspicious persons.
  - Color histogram

Policemen

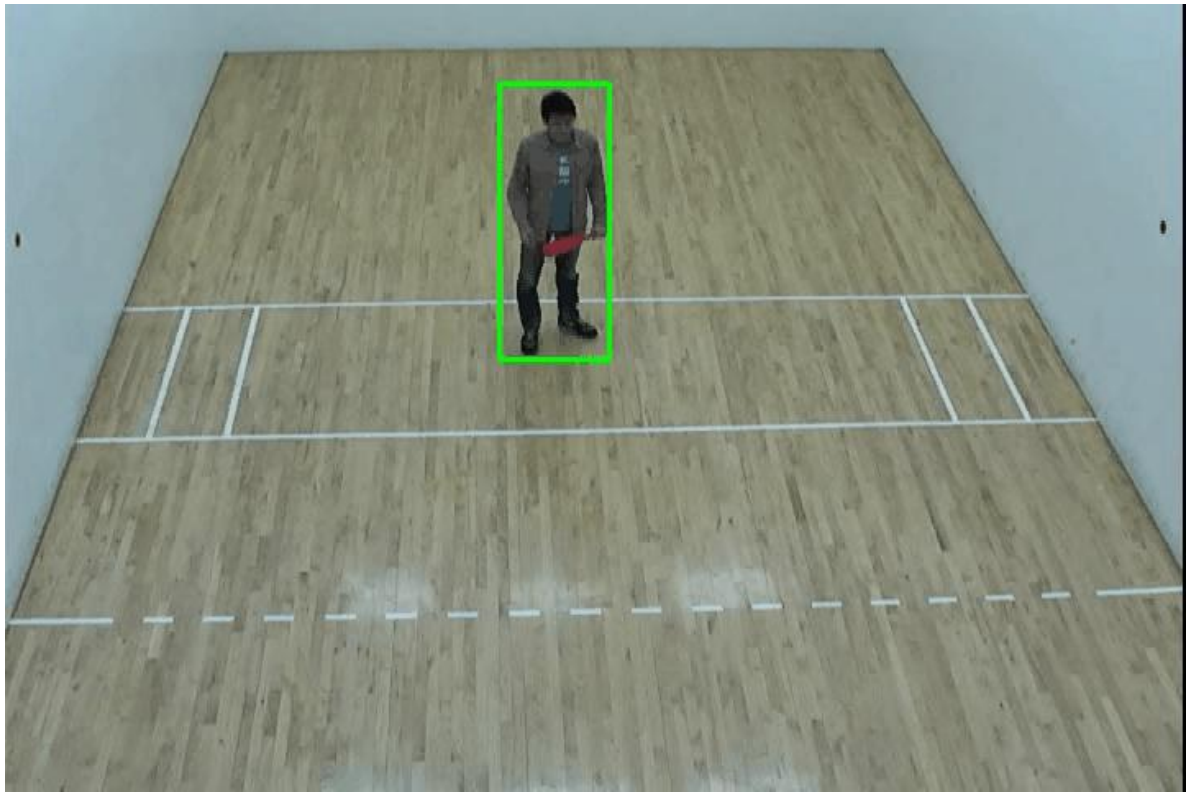
Criminal candidates

Pedestrians



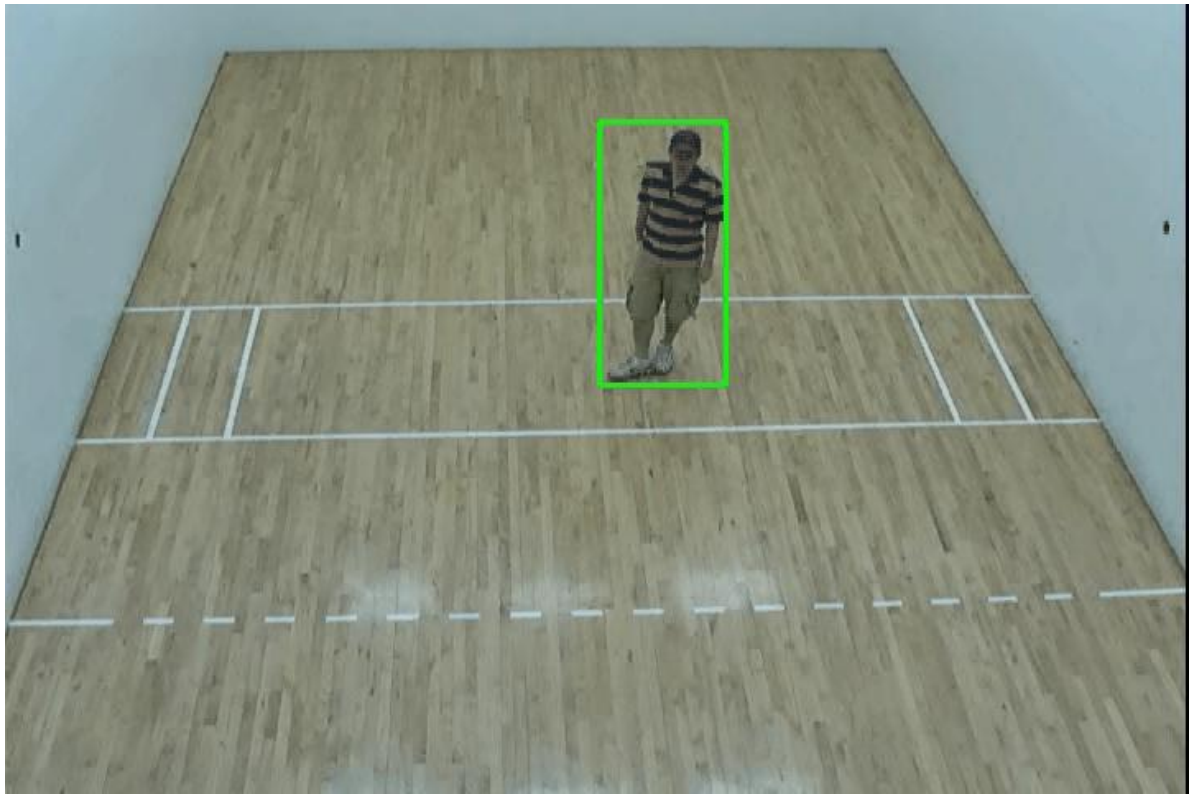
Experiments – group assault

- Highly stochastic
  - There may be (and may not be) attackers whose guarding the area, or just watching.
  - 10 videos.



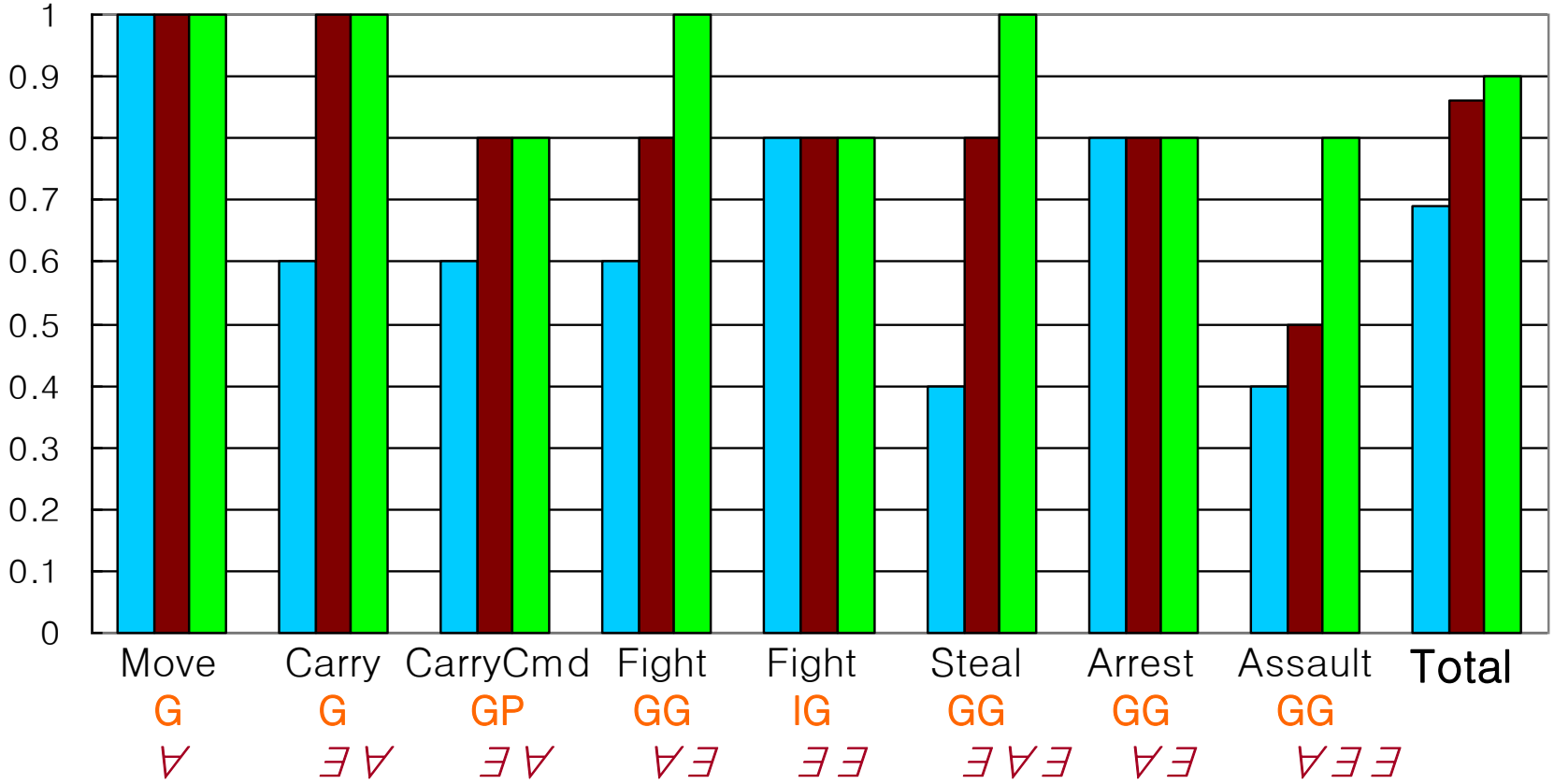
Experiments – group assault

- Highly stochastic
  - There may be (and may not be) attackers whose guarding the area, or just watching.
  - 10 videos.



Experimental results

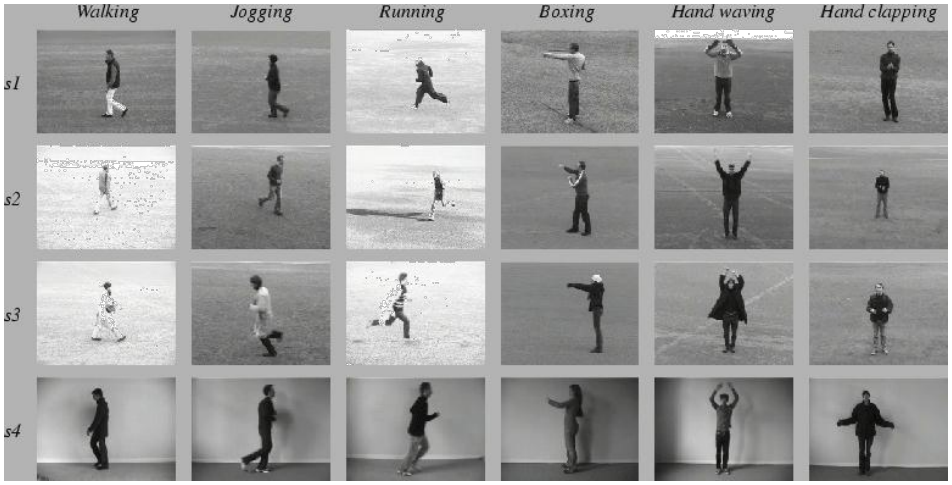
- **Recognition accuracy**
  - **False positive rates are almost 0 because of the detailed representations: previous, deterministic, stochastic.**

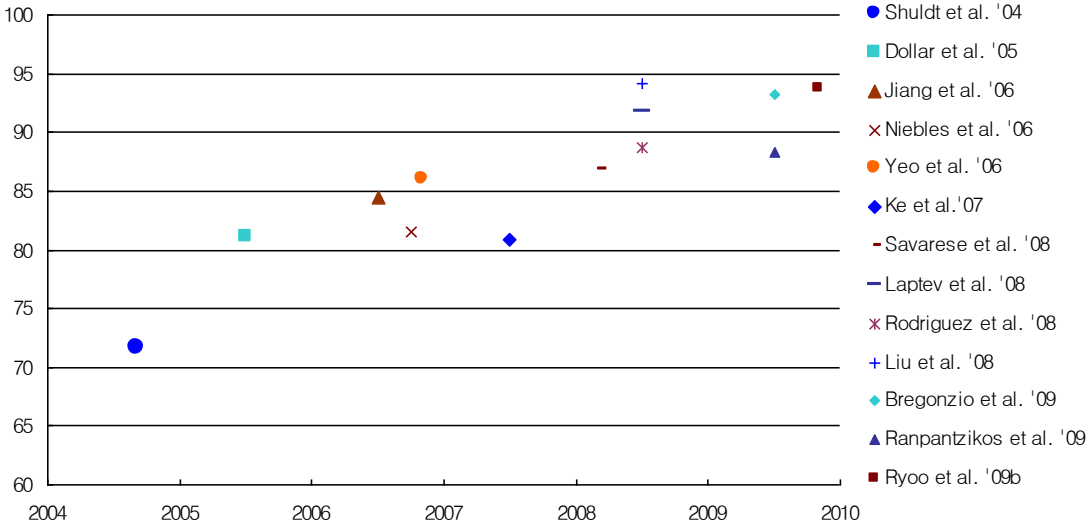


Datasets – academic

KTH dataset

- KTH dataset
  - Single action video classification
    - Single actor – 25 sets
    - One action per video





Weizmann dataset

- Similar to the KTH dataset (single action)
  - Actually, easier than KTH
    - Static camera (KTH has camera motion)
    - 90 videos (KTH has 600 videos)
    - 100% classification accuracy when it was first proposed

[Gorelick, Blank, Shechtman, Irani, and Basri, Actions as Space-Time Shapes, ICCV 2005]

Hollywood action dataset 1 & 2

- Hollywood datasets [Laptev 07,08]
  - Movie scenes – 8 classes of actions
  - Goal: recognition in complex environments
    - Moving cameras
    - Background changes
  - Action classification
    - Segmented videos
    - Atomic movements (e.g. kissing)

[Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B., Learning Realistic Human Actions from Movies, CVPR 2008]

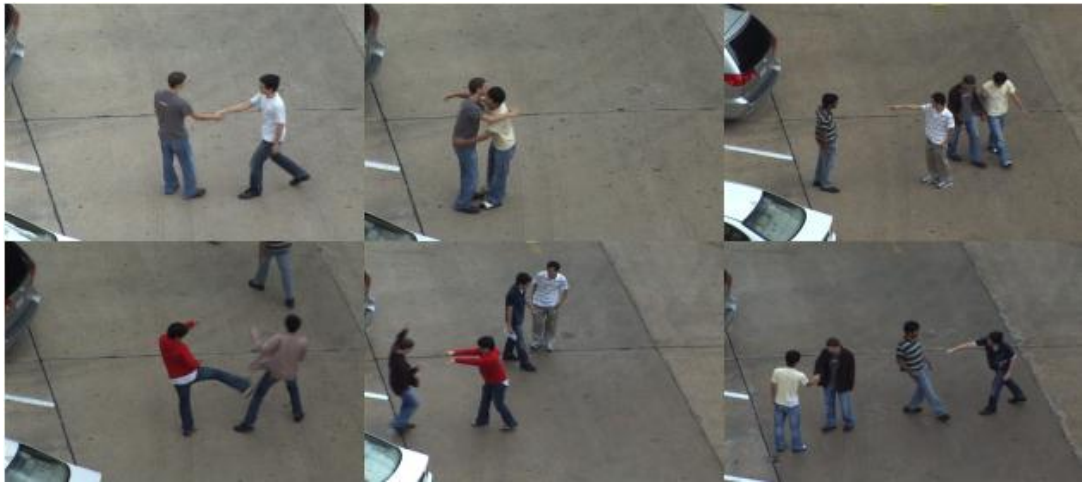
UCF YouTube Action dataset

- Videos downloaded from YouTube
  - 11 action classes
  - Classification task
- Noisy videos
  - Camera motion
  - Various view-points

[Rodriguez, M., Ahmed, J., and Shah, M., Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition, CVPR 2008]

UT-Interaction dataset 1 & 2

- More surveillance-oriented dataset
  - Extension of KTH & Weizmann
  - Two tasks:
    - Classification
    - Detection

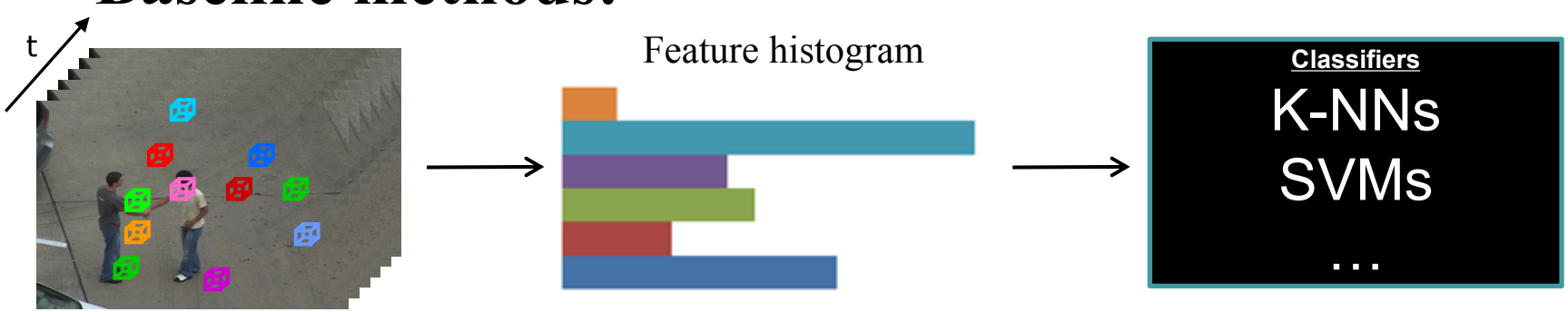


UT-Interaction results – set1

Classification accuracies:

	Shake	Hug	Kick	Point	Punch	Push	Total
Laptev + kNN	0.18	0.49	0.57	0.88	0.73	0.57	0.57
Laptev + Bayes.	0.38	0.72	0.47	0.9	0.5	0.52	0.582
Laptev + SVM	0.49	0.79	0.58	0.8	0.6	0.59	0.642
Latpev + SVM (best)	0.5	0.8	0.7	0.8	0.6	0.7	0.683
Cuboid + kNN	0.56	0.85	0.33	0.93	0.39	0.72	0.63
Cuboid + Bayes.	0.49	0.86	0.72	0.96	0.44	0.53	0.667
Cuboid + SVM	0.72	0.88	0.72	0.92	0.56	0.73	0.755
Cuboid + SVM (best)	0.8	0.9	0.9	1	0.7	0.8	0.85
Team BIWI	0.7	1	1	1	0.7	0.9	0.88

Baseline methods:

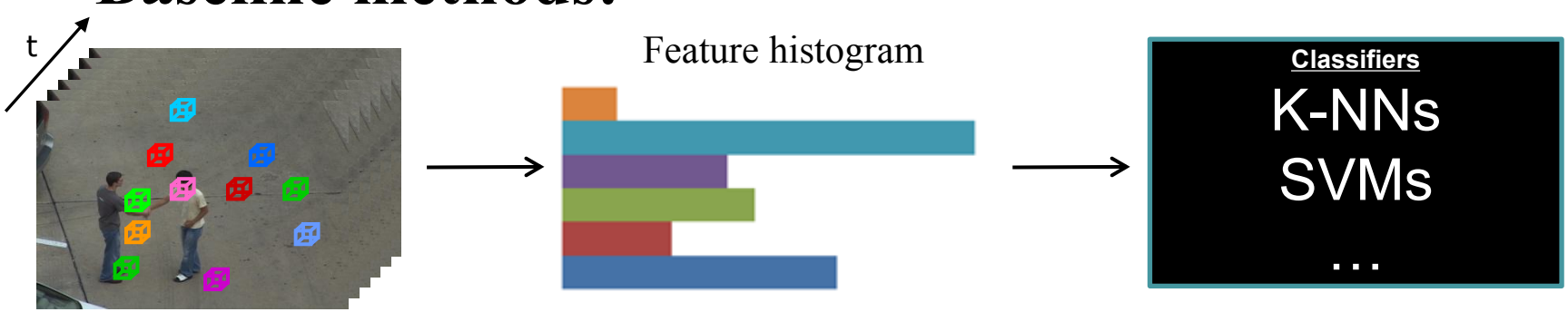


UT-Interaction results – set2

Classification accuracies:

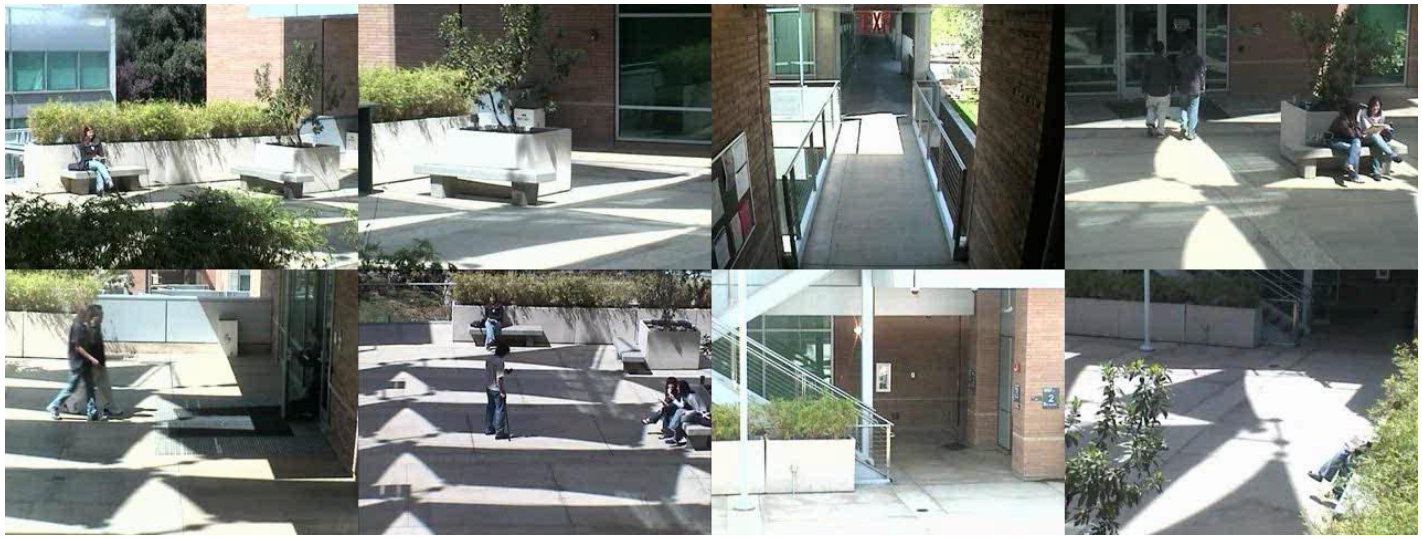
	Shake	Hug	Kick	Point	Punch	Push	Total
Laptev + kNN	0.3	0.38	**0.76**	**0.98**	0.34	0.22	0.497
Laptev + Bayes.	0.36	0.67	0.62	0.9	0.32	**0.4**	0.545
Laptev + SVM	0.49	0.64	0.68	0.9	0.47	**0.4**	0.597
Latpev + SVM (best)	0.5	0.7	**0.8**	0.9	0.5	**0.5**	**0.65**
Cuboid + kNN	**0.65**	**0.75**	0.57	0.9	0.58	0.25	0.617
Cuboid + Bayes.	0.26	0.68	0.72	**0.94**	0.28	0.33	0.535
Cuboid + SVM	**0.61**	**0.75**	0.55	0.9	**0.59**	0.36	0.627
Cuboid + SVM (best)	**0.8**	**0.8**	0.6	0.9	**0.7**	**0.4**	**0.7**
Team BIWI	0.5	**0.9**	**1**	**1**	**0.8**	**0.4**	**0.77**

Baseline methods:



UCR VideoWeb

- Multi-camera dataset (4~8 cameras)
  - Large scale – 2.5 hours
  - Multiple actors – 39 scenes
    - Human-human interactions, group activities…



[Denina, Bhanu, Nguyen, Ding, Kamal, Ravishankar, Roy-Chowdhury, Ivers, and Varda, VideoWeb Dataset for Multi-camera Activities and Non-verbal Communication, Distributed Video Sensor Networks 2010.]

Coffee break!

- This is the end of the first part of the tutorial.
- Are there any questions?

- 2nd half starts at 3:40pm

Future directions

Limitations of previous methods

- Most assume *after-the-fact* detection
  - Classify after fully observing the video

Stealing occurred in an Apple computer store

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

[Ryoo, M. S., Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos, ICCV 2011]

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?

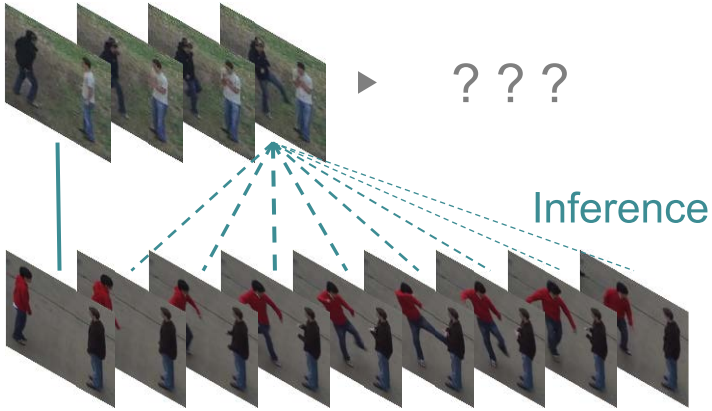
Problem formulation

- Classification (previous)
  - Assumes each video contains an entire activity
    - Activity is always fully progressed, d^*
    - $P(A \mid O,t) = P(A,d^* \mid O)$
- Activity prediction
  - Inference from an initial observation
    - Multiple possible activity progress level d
    - $P(A \mid O,t) = \sum_d P(A,d \mid O,t)$

Video observation O:

Activity model A:

Ongoing observation O:

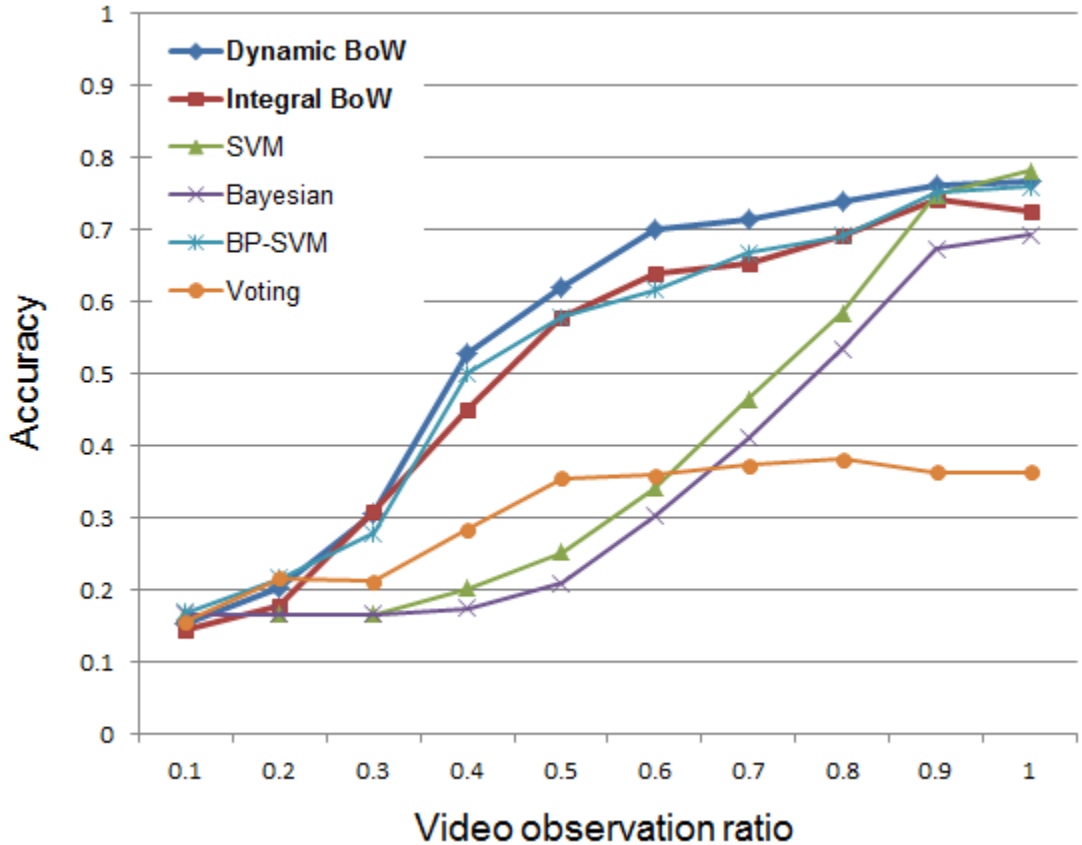


Experimental results

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

Human activity prediction

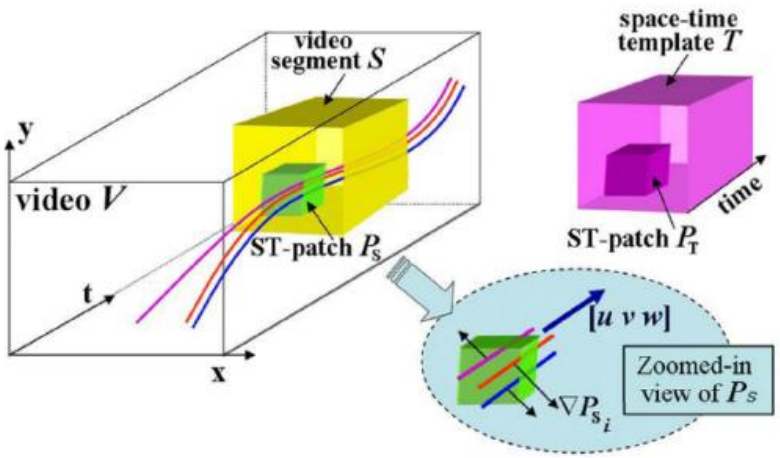
- Experimental results
  - Human-human interaction
  - Our approaches detect activities at much earlier stage.
    - Higher graphs indicate better performance.



Real-time using GP-GPUs

- General purpose graphics processing units (GP-GPUs)
  - Multiple cores running ***thousands*** of threads
  - Parallel processing

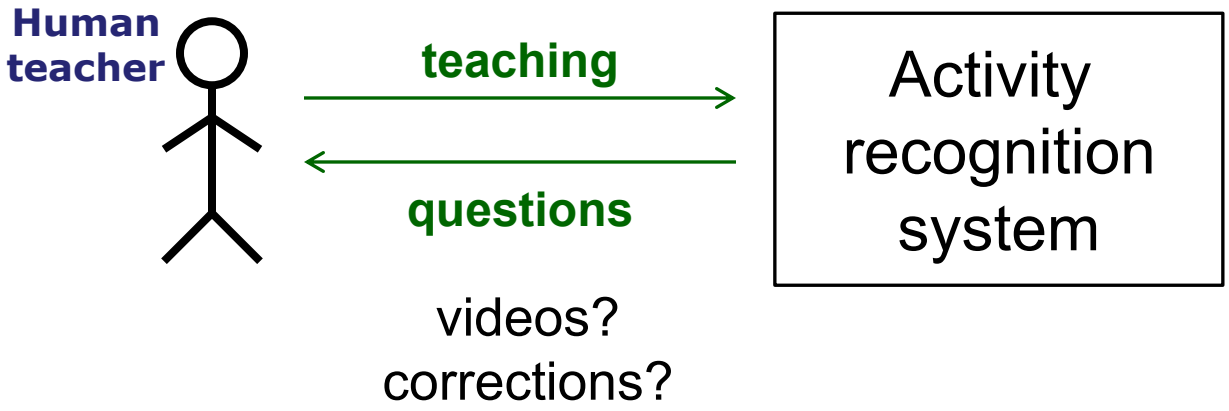
20 times speed-up of video comparison method in Shechtman and Irani 2005:



[Rofouei, M., Moazeni, M., and Sarrafzadeh, M., Fast GPU-based space-time correlation for activity recognition in video sequences, 2008]

Interactive learning

- Interactive learning approach
  - Learning by generating questions



  - Interactive learning
    - Human-in-the-loop
    - Explore decision boundaries actively

Active video composition

- Simulated videos?
  - *Composed videos* from a real video
  - Automatically create **necessary** training videos
  - Structural variations
    - Who stretches his hand first?

Original video:

Composed videos:

Ryoo and Yu, WMVC 2011

- Research works in this tutorial authored by Michael S. Ryoo were done at the University of Texas at Austin, USA and Electrical and Telecommunications Research Institute (ETRI), South Korea, before he joined the Jet Propulsion Laboratory, California Institute of Technology .