Frontiers of
Human Activity Analysis

J. K. Aggarwal
Michael S. Ryoo
Kris M. Kitani

THE UNIVERSITY OF TEXAS AT AUSTIN
UT ECE
ELECTRICAL & COMPUTER ENGINEERING

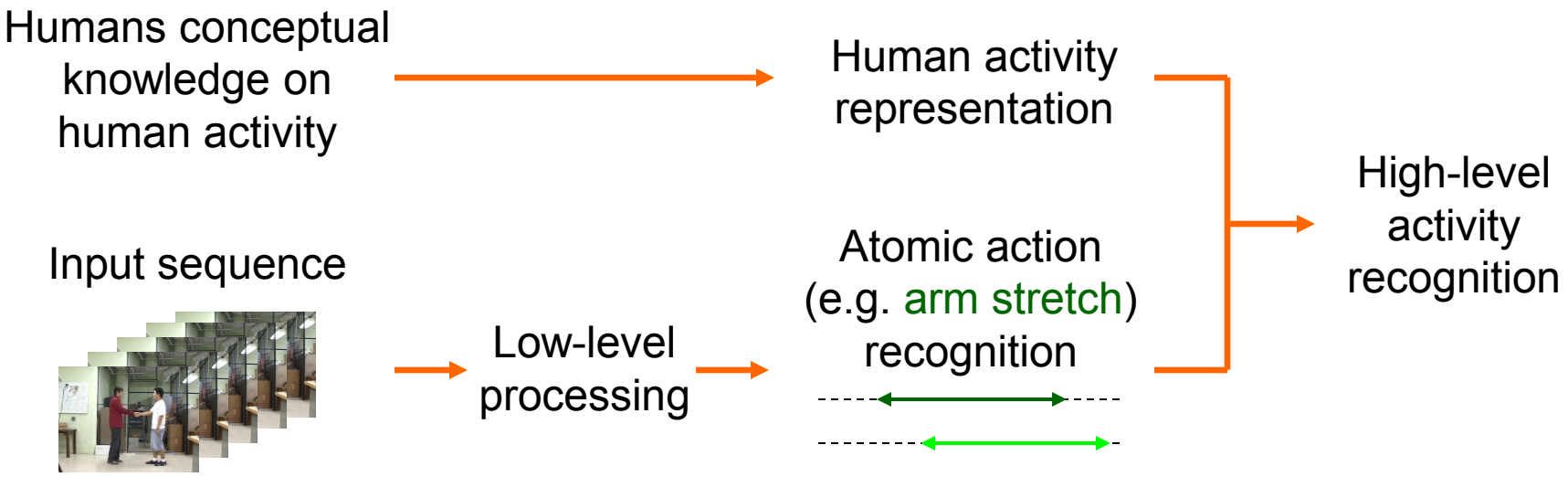
ETRI

Carnegie Mellon
THE ROBOTICS INSTITUTE

Description-based approaches

Approach paradigm

- Description-based approach
  - We **represent** the structure of the activities, and **recognize** activities using semantic matching.
    - **Hand shake** = "*two persons* do **shake-action** (**stretches**, **stays stretched**, **withdraw**) simultaneously, while **touching**".
    - Recognition by finding observations satisfying the definition.



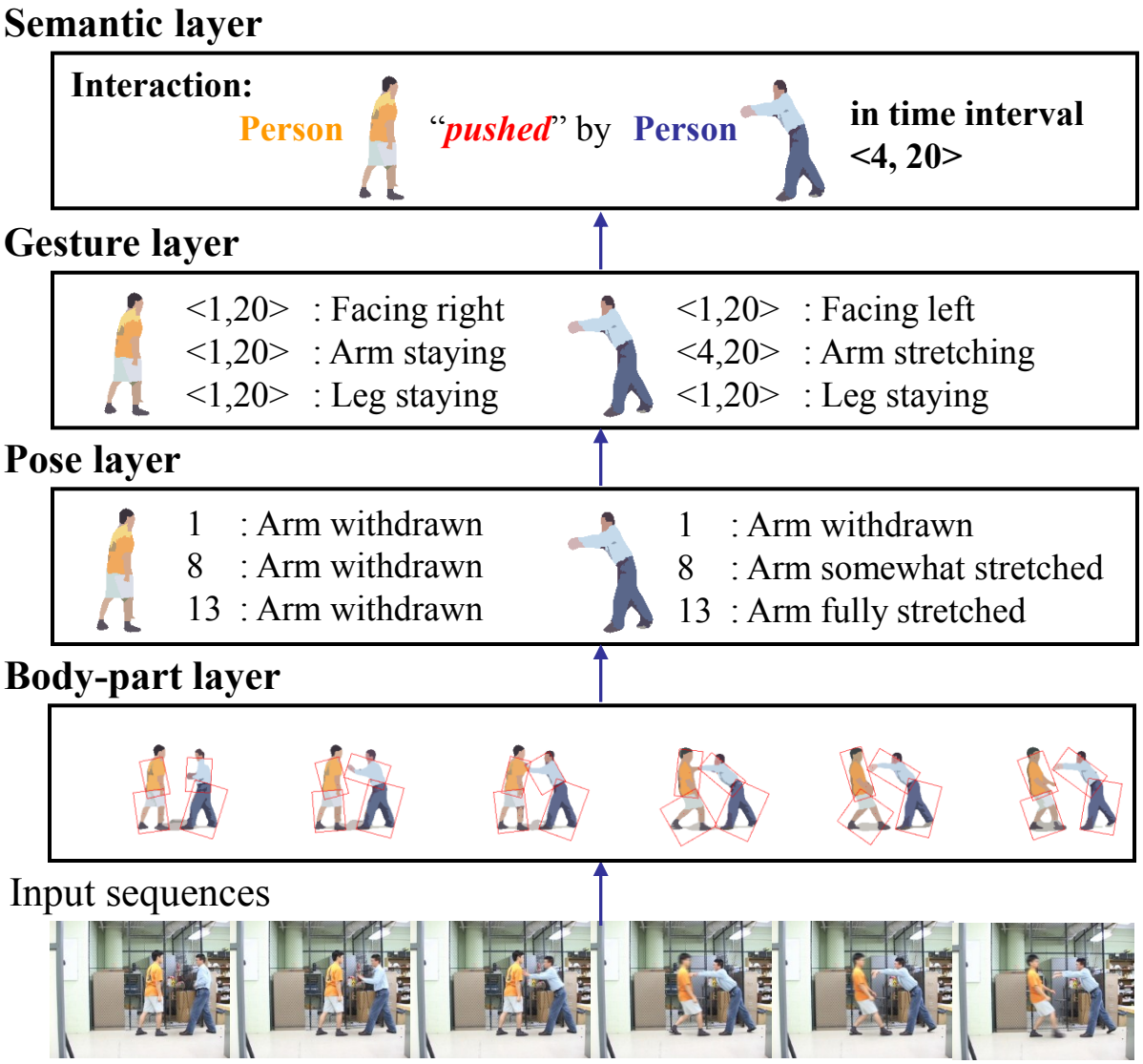
Comparisons

Approaches	Levels of hierarchy	Complex temporal relations	Complex logical con-catenations	Recognition of recursive activities	Handle imperfect low-levels
Statistical	limited (depends on data amount)				√
Syntactic	unlimited			√	√
Siskind 2001	unlimited	a sub-event participates only once	√		
Hongeng et al. 2004	limited (3-levels)	√	√		
Vu et al. 2003	unlimited	√	conjunctions only		
Ryoo and Aggarwal 2009	unlimited	√	√	√	√
Gupta et al. 2009	limited (2-levels)	√	network form only		√

Description-based approaches

Human interactions

Recognition of human interactions

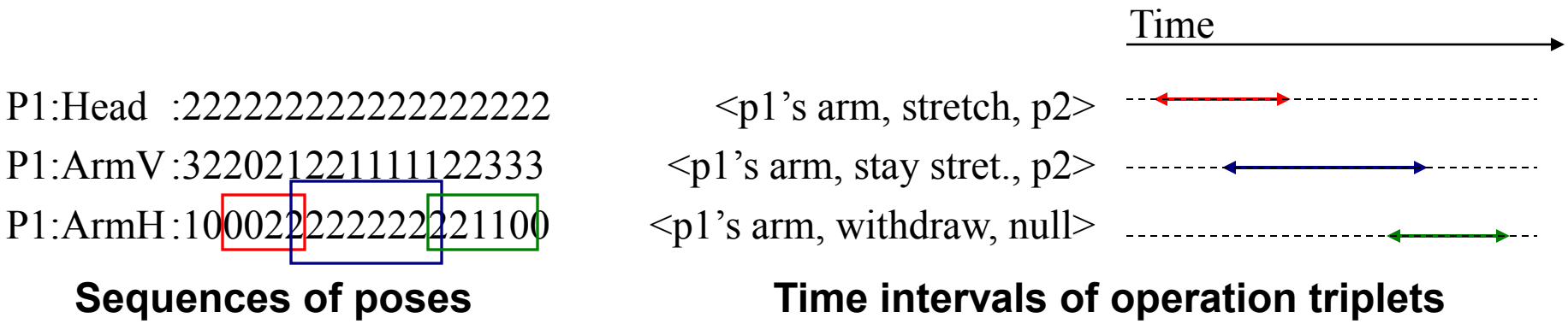


- Interaction
- Gesture
  - Elementary movement of a person
- Pose
  - Abstract status of a body part.
- Body-part feature.
  - Numerical status of a body part.

Ryoo and Aggarwal, CVPR 2006

Atomic actions

- Operation triplets <agent, motion, target>
  - Gesture together with subject and object information.
  - Unit human activity.
  - Computed based on gestures.
  - Ex> person1 stretches his/her arm
    → **<p1's arm, stretch, null>**
- Time intervals
  - Ex> Time intervals detected for Pointing action



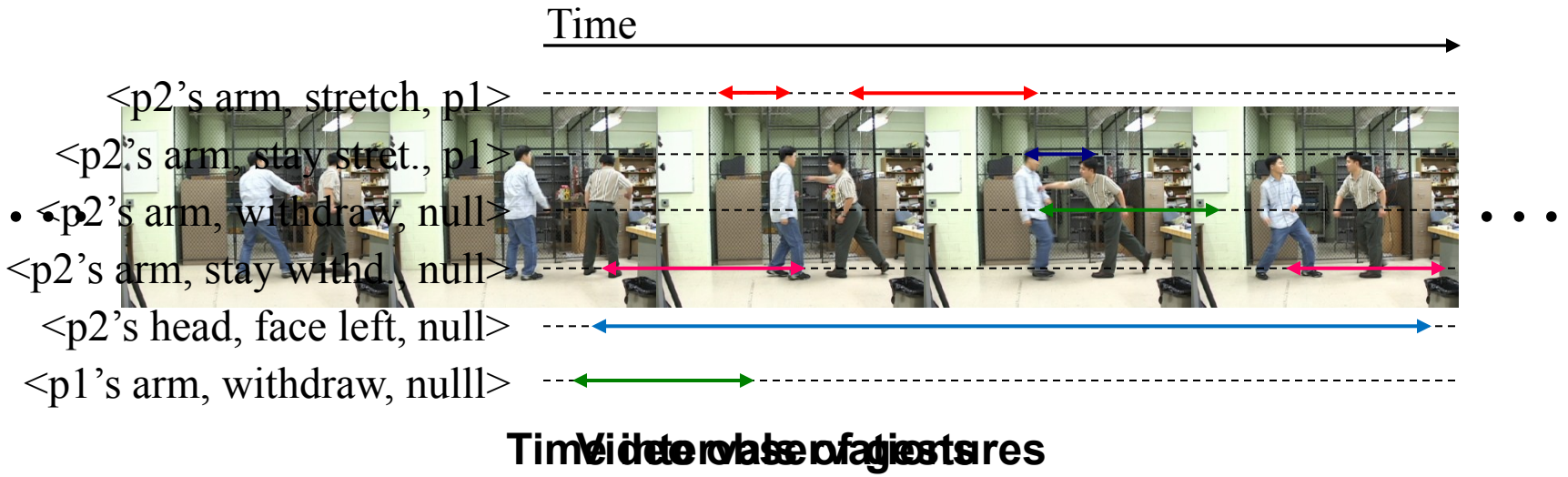
Sequences of poses **Time intervals of operation triplets**

Semantic layer recognition

Machine understandable representation of Punching

Punching is a sequence of hand stretch, withdraw, Pa

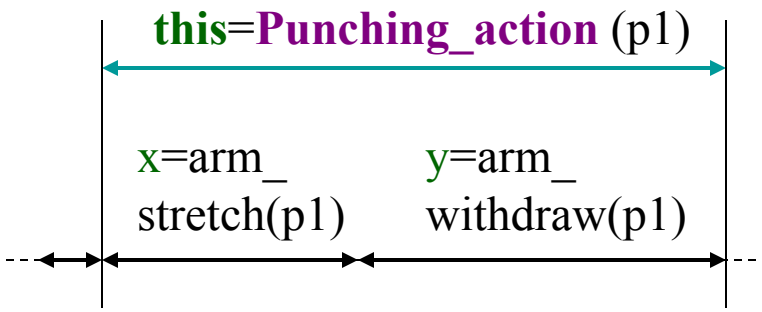
Similar?



Time Video raw features

Human activity representation

- Semantics
  - Knowledge on the structure of an activity.
    - **Punching** is a *sequence* of hand **stretch** and **withdrawal**.
  - Time intervals
  - Allen's temporal predicates
- Syntax
  - Rules to construct formal representation.
  - Organizes a set of vocabularies to describe the activities' structure.
  - Context-free grammar



CFG

```
Punching_action(i) = (
   list(    def('x', Arm_Stretch(i)),
            def('y', Arm_Withdraw(i)) ),
   and(     meets('x', 'y'),
            and( starts('x', 'this'),
                 finishes('y', 'this')) )  );
```

Conceptual/verbal description

Machine-understandable language

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” “)”
| $TemporalPredicate$“(” $var(a)$“,” $var(b)$“)”

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**’)))
);

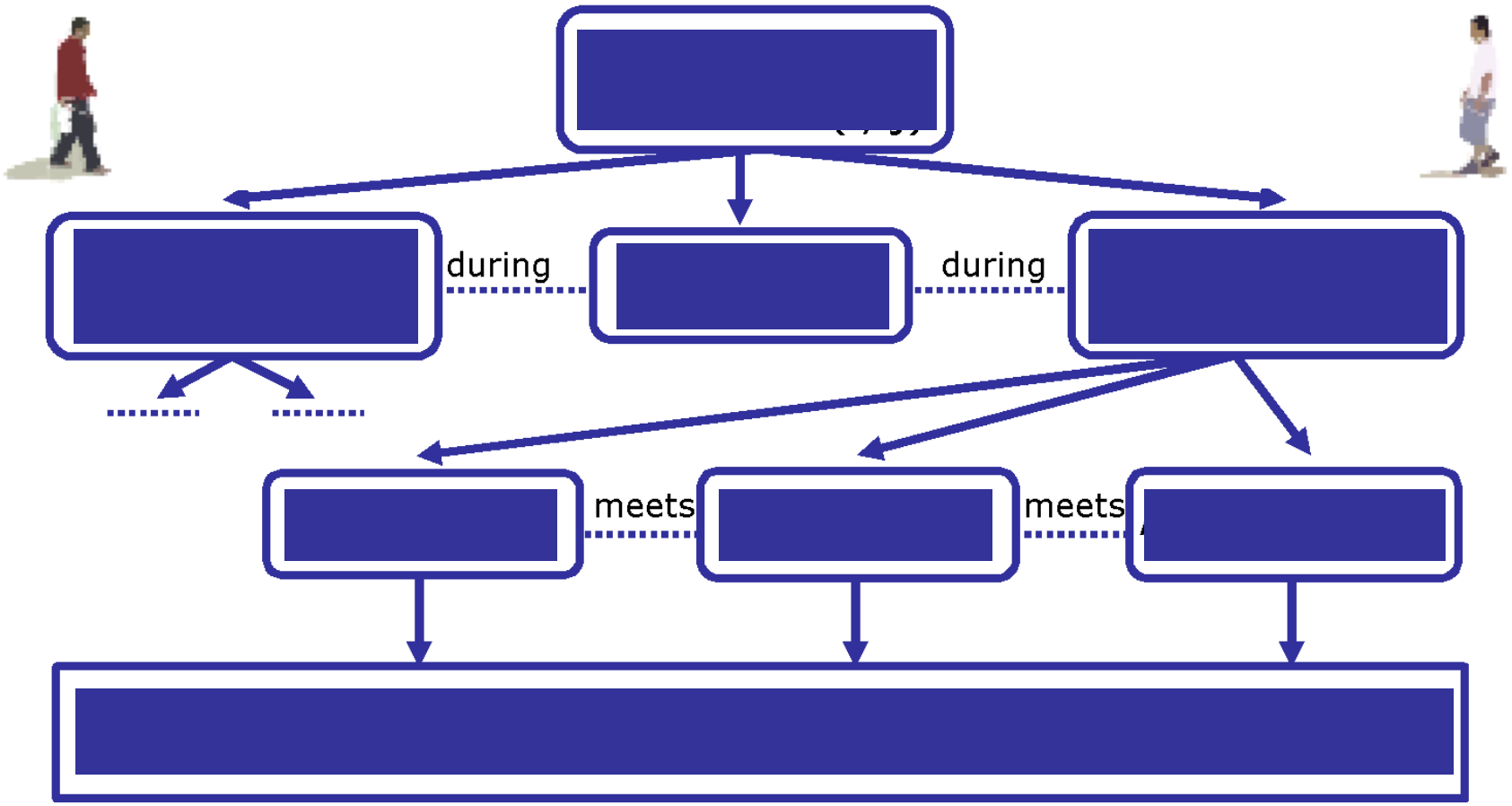
ShakeHandsActions(i) = (
list(def(‘x’, Arm_Stretch(i)),
list(def(‘y’, Arm_Stay_Stretched(i)),
def(‘z’, Arm_withdraw(i)))),
and(and(meets(‘x’, ‘y’), meets(‘y’, ‘z’)),
and(starts(‘x’, ‘**this**’), finishes(‘z’, ‘**this**’)))
);

TouchingInteraction(i, j)=(null, touch(i, j, 0));

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

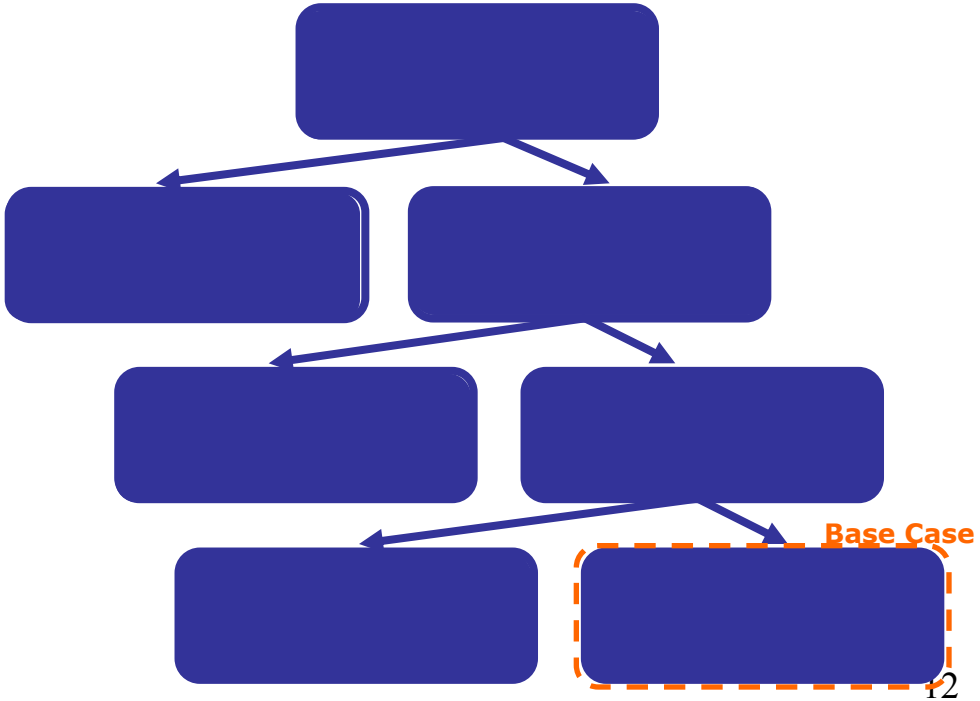
Hierarchical recognition algorithm

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



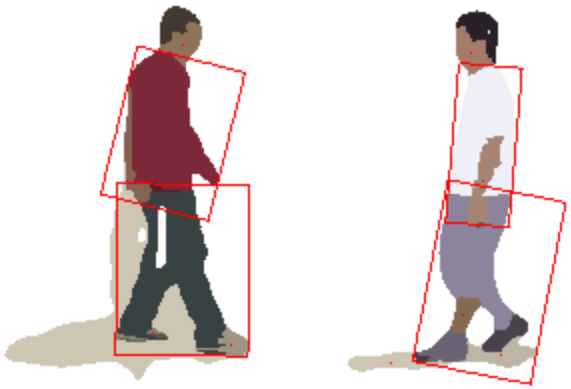
Continued and recursive activities

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



Experiments - Simple interactions

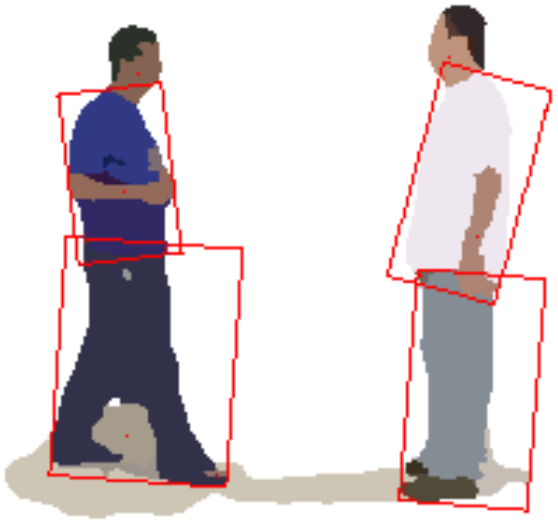
- Recognized 8 types of simple interactions, which were recognized in Park and Aggarwal, 2004
  - (approach, depart, point, shake-hands, hug, punch, kick, and push)
  - A videos of a **sequence** of interactions are taken. (**continuous** executions)
  - Interactions are described in more detailed and formal way, resulting better recognition accuracy.



Example Experiment - Fighting

Input video:

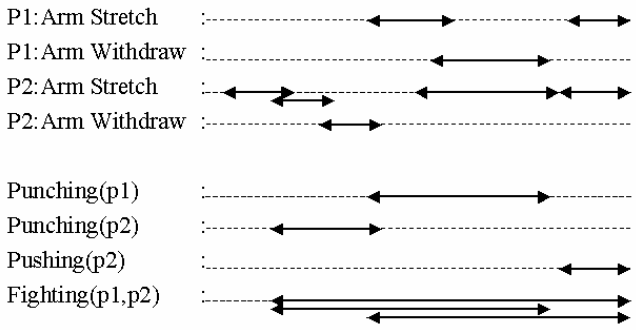
Processed video:



Poses:

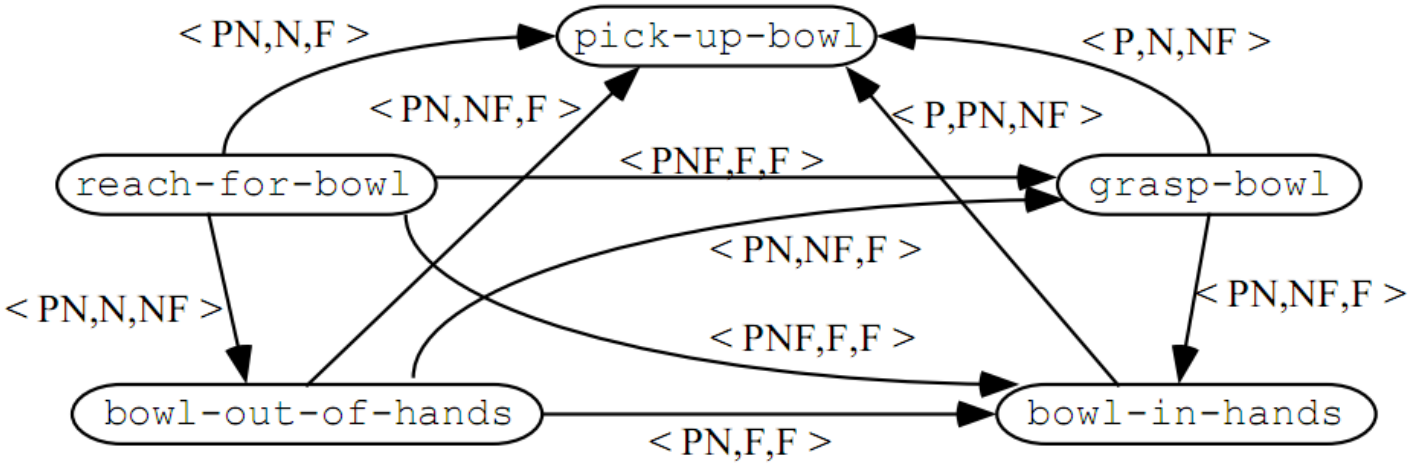
	Time
P1:ArmV	:3112222222131121000012112331133103
P1:ArmH	:1000000011111112222212111110001222
P2:ArmV	:3332100000121122222120111110000022
P2:ArmH	:0001121222211000000111111112211222

Gestures and activities:



Past-Now-Future networks

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

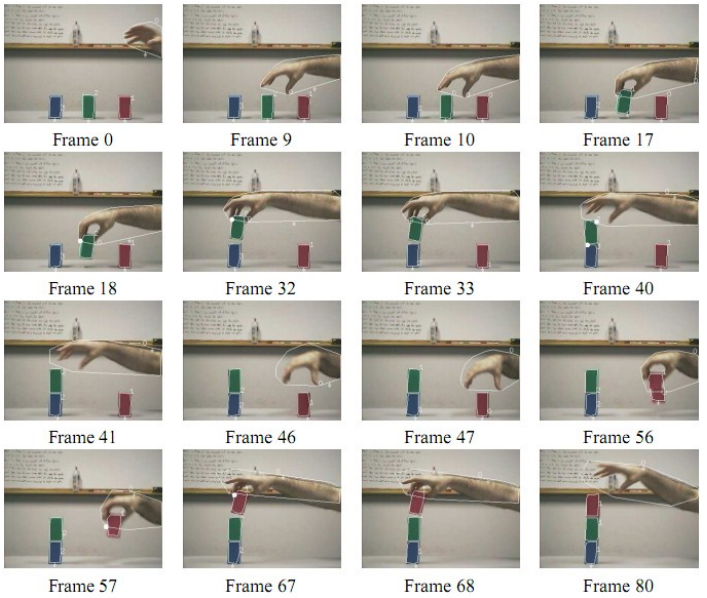


[Pinhanez, C. S. and Bobick, A. F., Human action detection using PNF propagation of temporal constraints. CVPR 1998]

Event logic

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

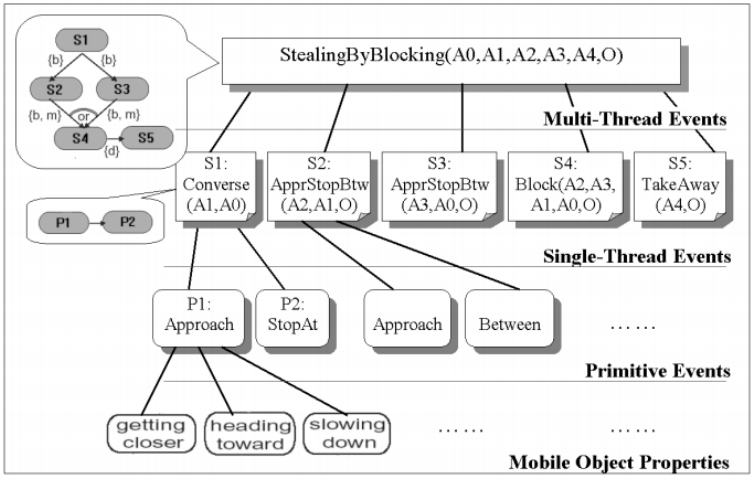
$$\textsc{PickUp}(x,y,z) \triangleq \left\langle \begin{array}{l} \neg\Diamond x = y \wedge \neg\Diamond z = x \wedge \neg\Diamond z = y \wedge \\ \textsc{Supported}(y) \wedge \neg\Diamond \textsc{Attached}(x,z) \wedge \\ \left\{ \begin{array}{l} \left[\begin{array}{l} \neg\Diamond \textsc{Attached}(x,y)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



[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))))
```

- Recursive? Uncertainties?

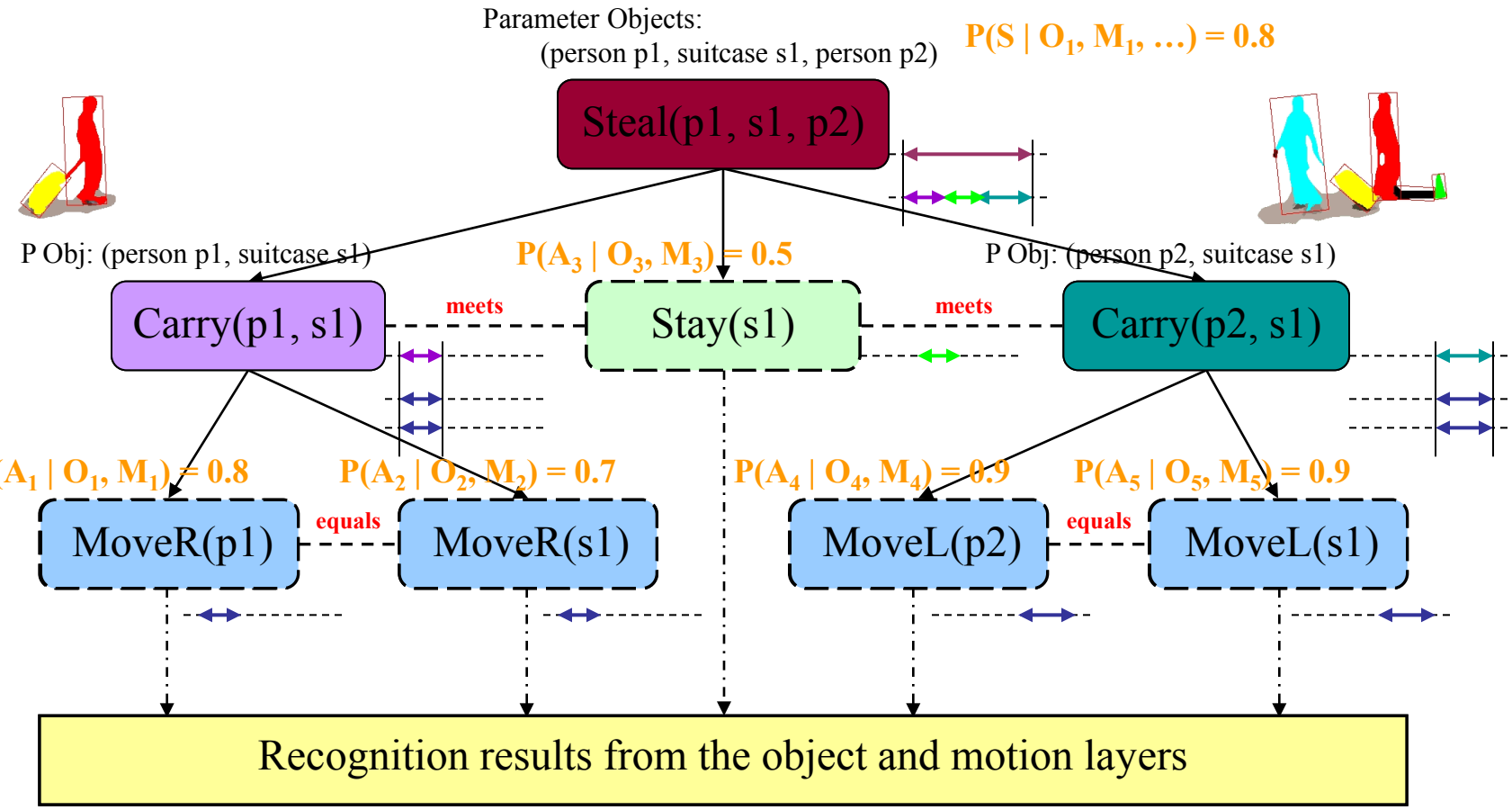
Stochastic approaches

- Limitations of the conventional description-based approaches
  - Uncertainties? – stochastic recognition

- Probabilistic framework needed
  - [Ryoo and Aggarwal, IJCV 2009]
  - [Tran and Davis, ECCV 2008] - MLN

Hierarchical matching algorithm

- Recognition process tree of ‘steal(p1, s1, p2)’



Probabilistic recognition

- Probability of the activity given observation

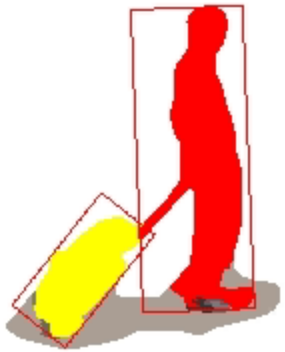
$$
\begin{aligned}
&P(R^{<s,e>}|I^T) \\
&= P(\{R\}|sub(\{R\})) \cdot P(sub(\{R\})|sub(sub(\{R\}))) \\
&\quad \ldots \quad P(sub^d(\{R\})|I^T) \\
&= \prod_{i=0}^{d-1} P(sub^i(\{R\})|sub^{i+1}(\{R\})) \cdot P(sub^d(\{R\})|I^T) \\
&= \prod_{i=0}^{d-1} P(sub^i(\{R\})|sub^{i+1}(\{R\})) \cdot P(a_1,\ldots,a_n|I^T) \\
&= \sum_{g_1} \ldots \sum_{g_n} \left[P(a_1,\ldots,a_n|g_1,\ldots,g_n) \cdot P(g_1,\ldots,g_n|I^T) \right]
\end{aligned}
$$

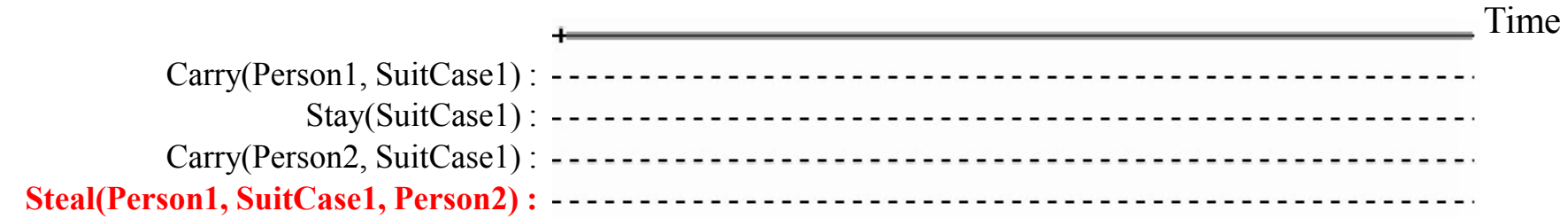
Structural similarity

Gesture detection confidence

Experiments

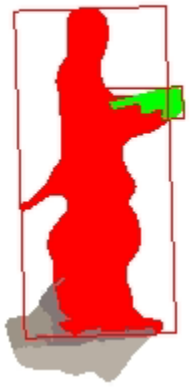
- Recognized following six types of interactions.
  - Each activity was tested with at least 10 sequences.
    - Carrying a box, leaving a box, placing a box into a trash bin.
    - Carrying a suitcase, leaving a suitcase, stealing the suitcase.
  - Object and Motion layer trained with 5 sequences.

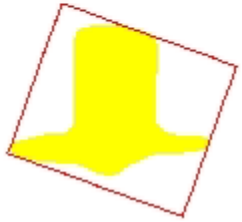


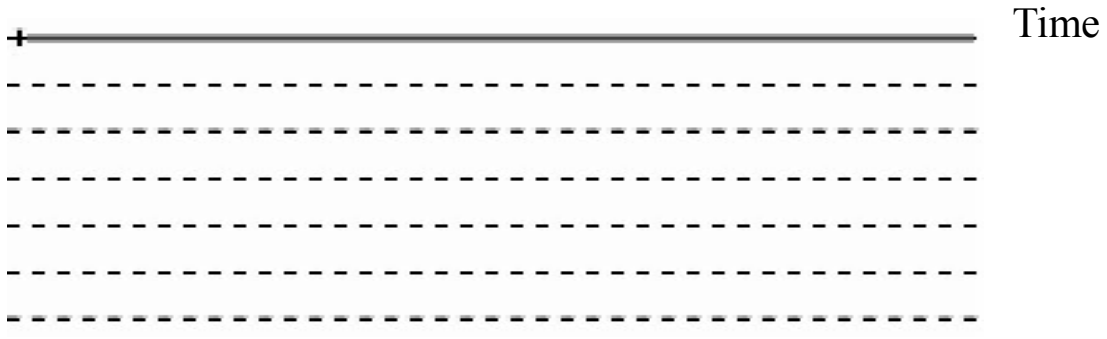


Experiments

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

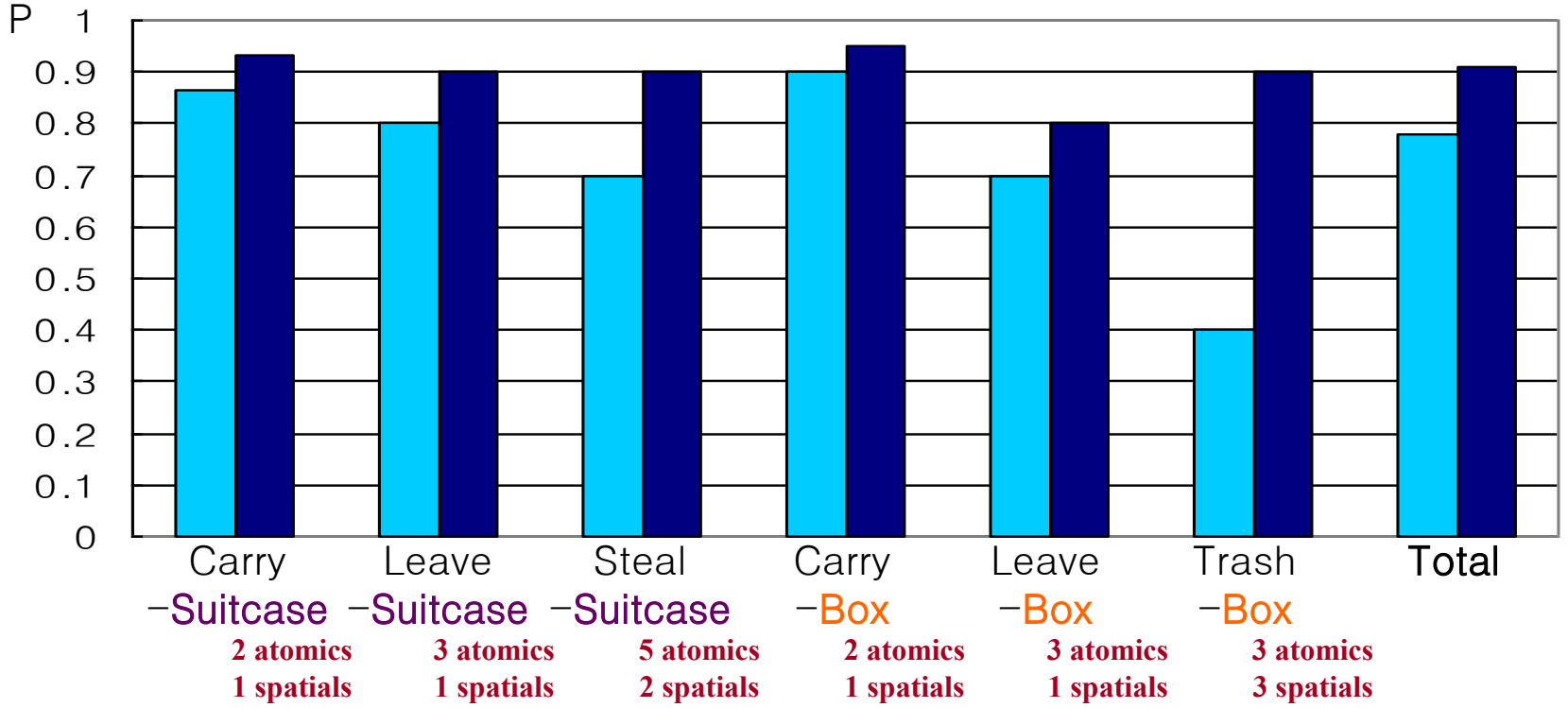






Experiments - Performance

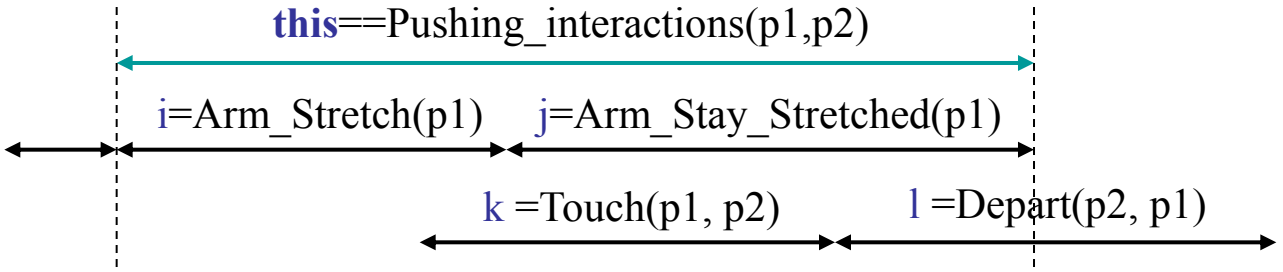
- **Recognition accuracy** (true positives):
  - Compared with a multi-object version of previous works.



- False positives rates are almost zero for all activities.

Advantages

- Ability to represent and recognize an activity composed of concurrent sub-events.
  - Ex> “touching occurred *during* pushing”



- Ability to represent and recognize ‘recursive activities’
  - Ex> Fighting = Fighting + another negative interaction.
- Less data required for training.
  - ‘Structure of activities’ are encoded based on human knowledge.
- High recognition accuracy?

Description-based approaches

Group activities

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

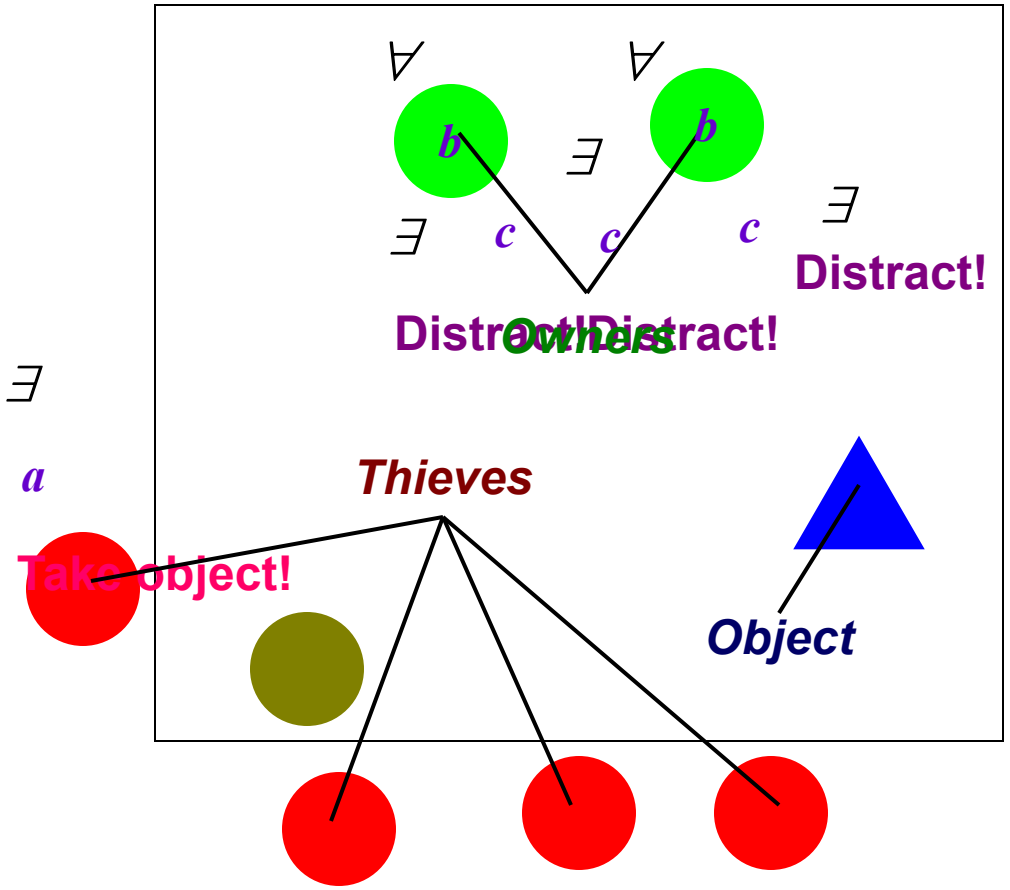
Group Assault (grp vs. grp)

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

Ryoo and Aggarwal, CVPR-SIG 09, IJCV 11

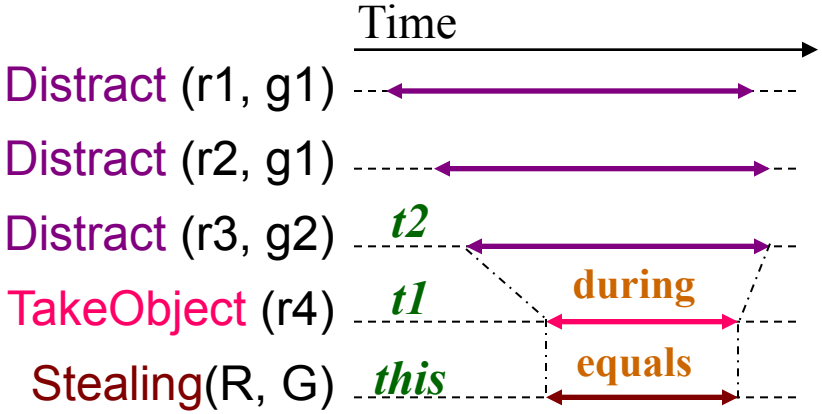
Representation

- Group stealing



Formal representation:

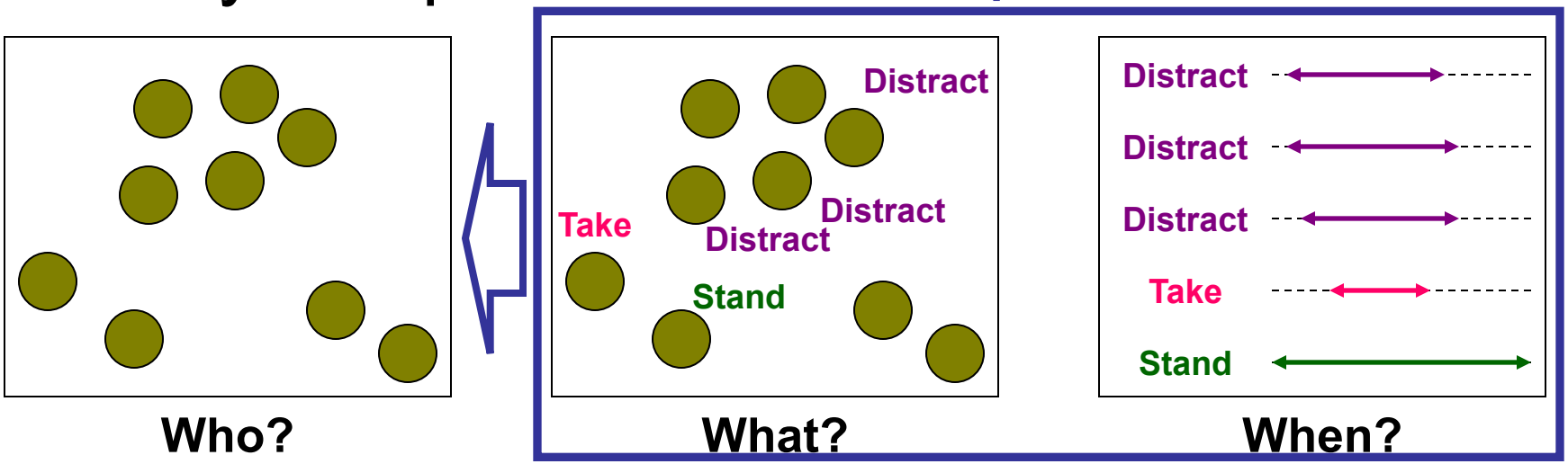
1. ∃ *a* in *Thieves*, ∀ *b* in *Owners*, ∃ *c* in *Thieves*
2. def(*t1*, TakeObject(*a*)), def(*t2*, Distract(*c*, *b*))
3. equals (*t1*, *this*), during (*t1*, *t2*)



Time intervals of activities of individual members

Recognition overview

- 3 key components **Generates a pool of member candidates**



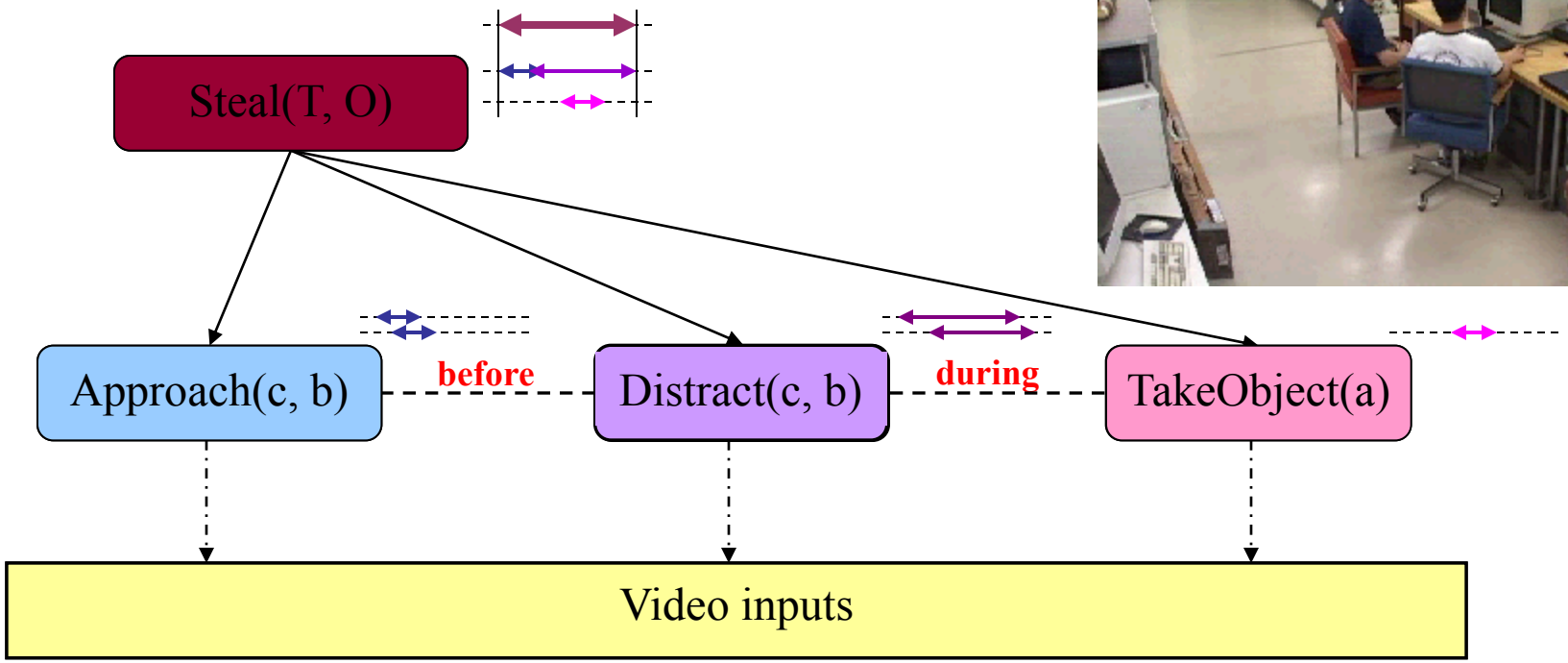
- Recognition: *NP-hard.* **Approximation required.**
  - Obtain a pool of group member candidates with non-zero probability.
    - Not many persons perform sub-events.
  - $M^* = \arg\max_M P(G^t(M) \mid O^M)$

Temporal constraints

- Hierarchical temporal constraint matching.

Member variables:

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

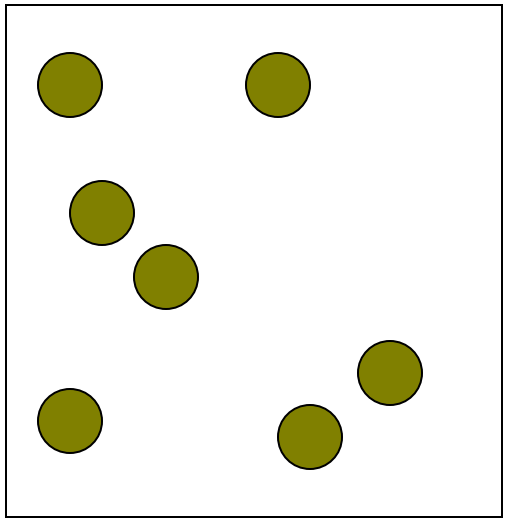


Group candidates

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

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

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

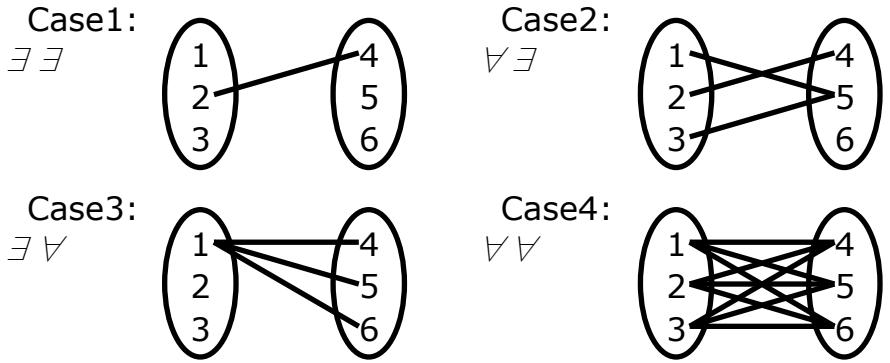
$$\pi_G(M) = \sum_{S_1^{t1},\ldots,S_n^{tn}} \left[\prod_{rel} P(rel \mid S_a, S_b) \cdot \prod_i P(S_i^{ti} \mid G^t(M)) \cdot \prod_i d \cdot e^{-(|K_i - C_i|/|K_i| + |L_i \cap C_i|/|K_i|)} \cdot G^t(M) \right]$$

Structural similarity

- Essential and anti-essential relations: K_i, L_i

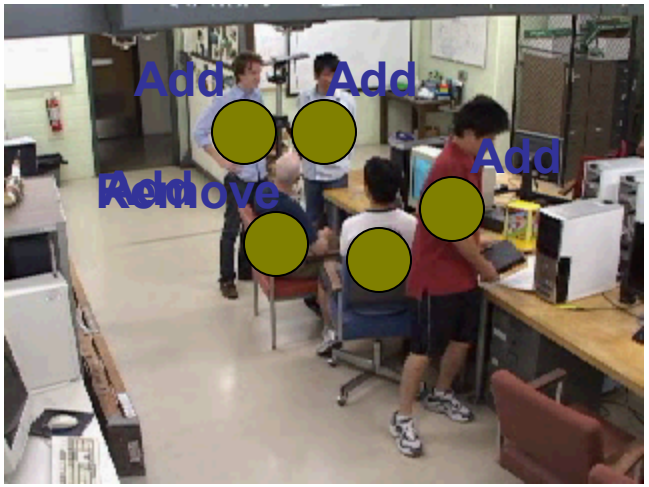
$$|K_i - C_i| = \sum_{k \in K_i - C_i} E[S_i^{ti}(k) \mid O^i]$$

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



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.

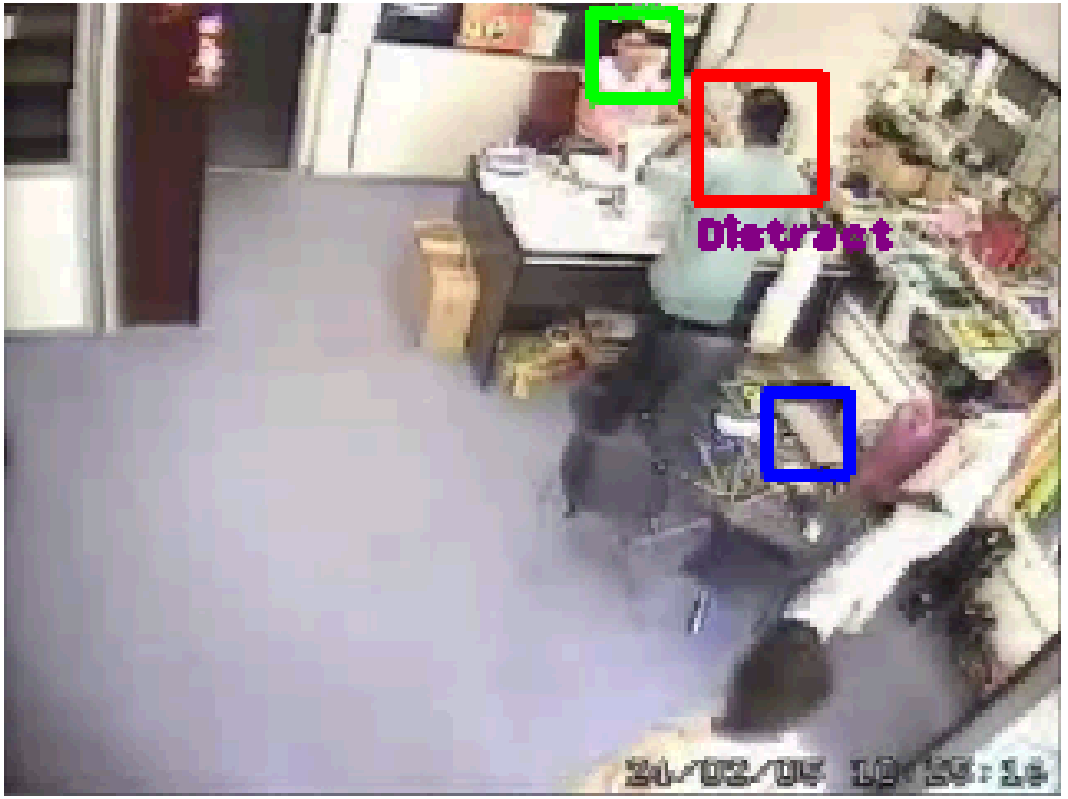




Videos of real human activities

Experiments - stealing

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



Thieves

Owners

Laptop

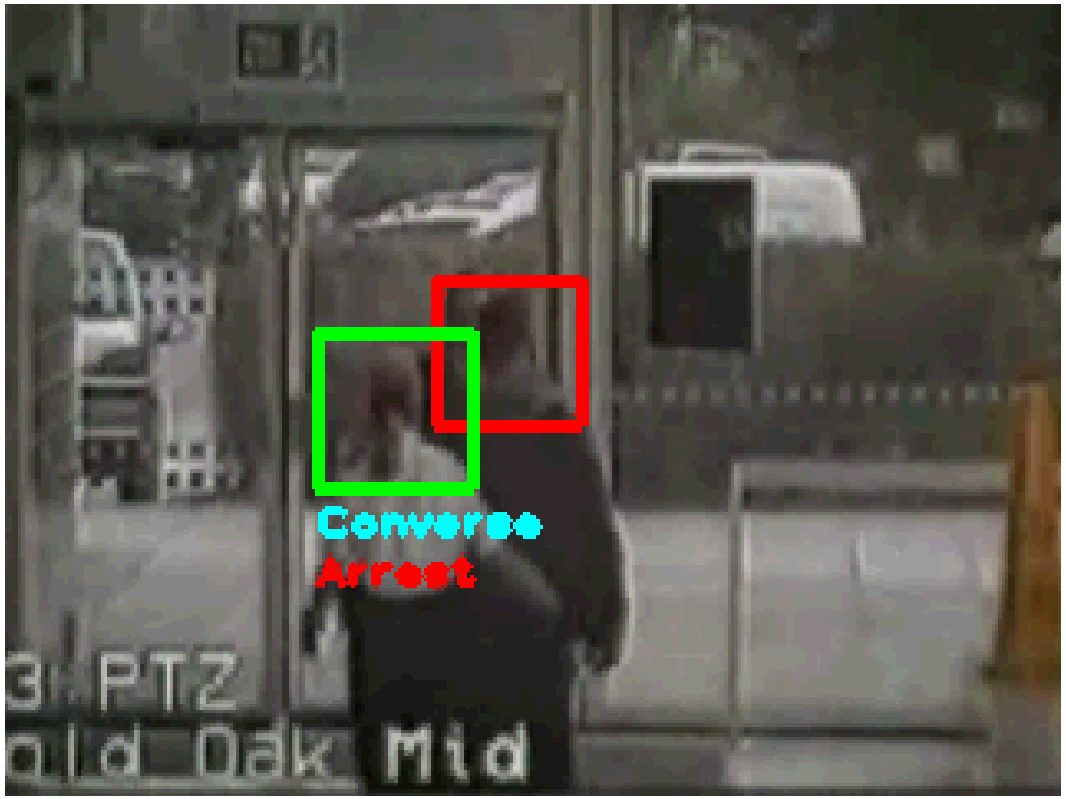
Experiments - arresting

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

Policemen

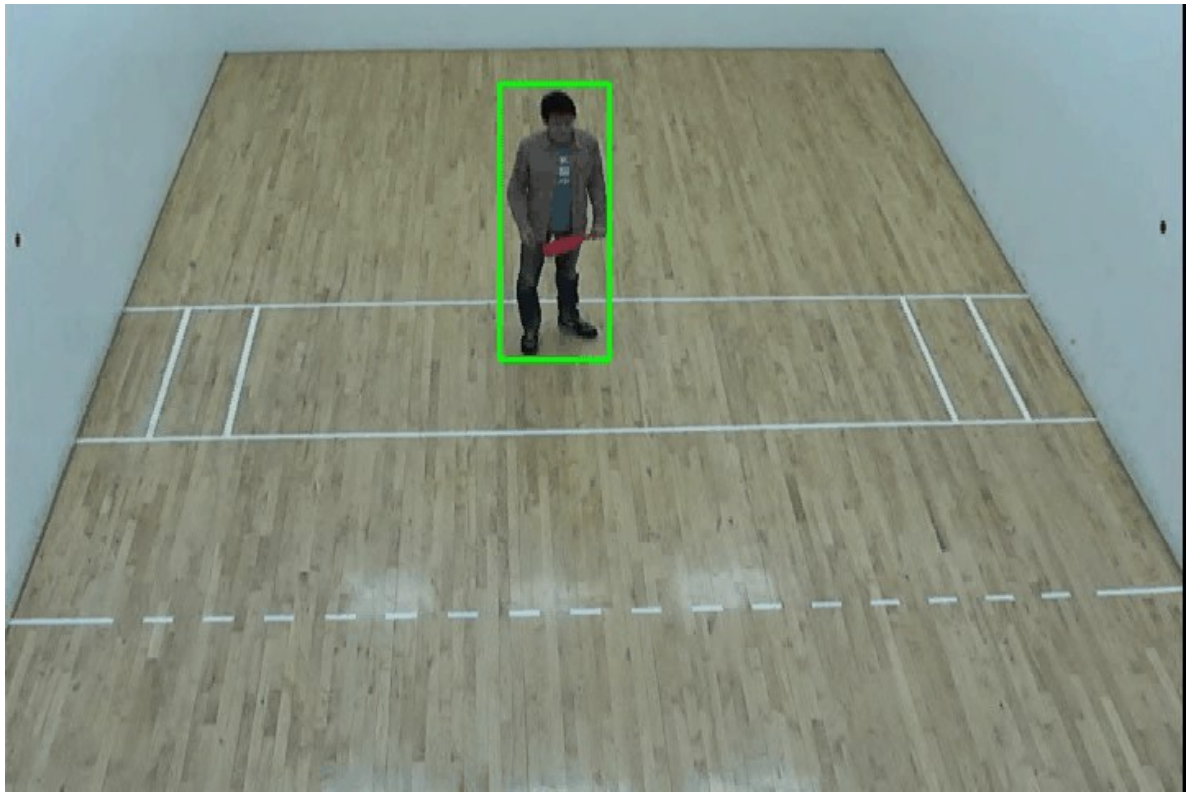
Criminal candidates

Pedestrians



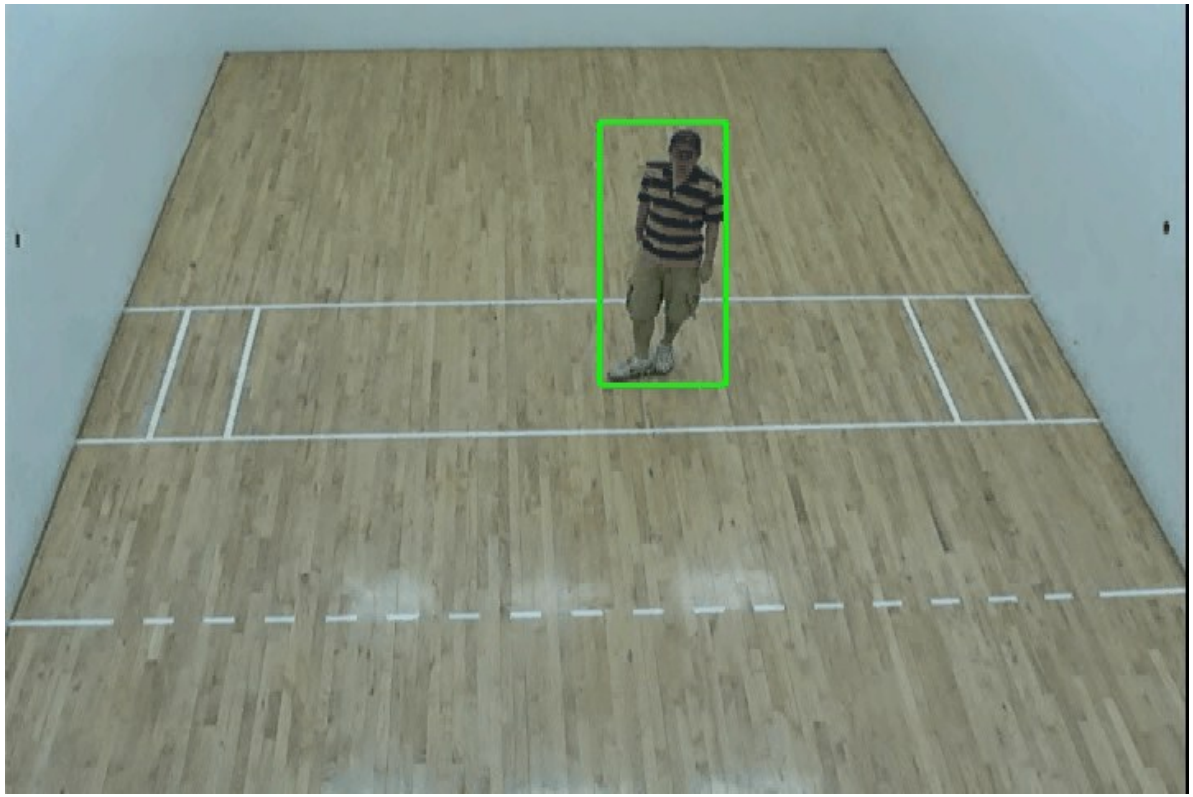
Experiments – group assault

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



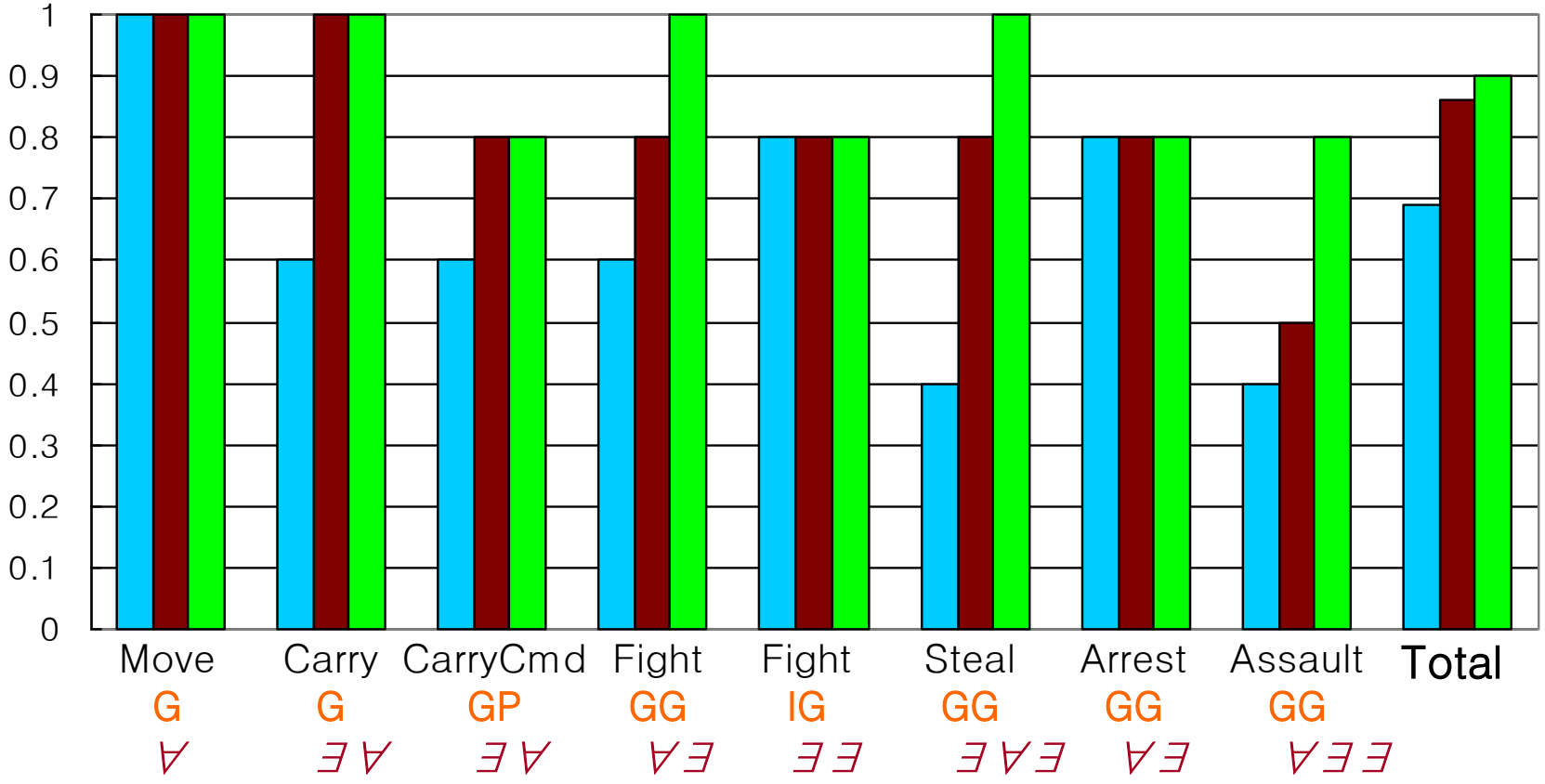
Experiments – group assault

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



Experimental results

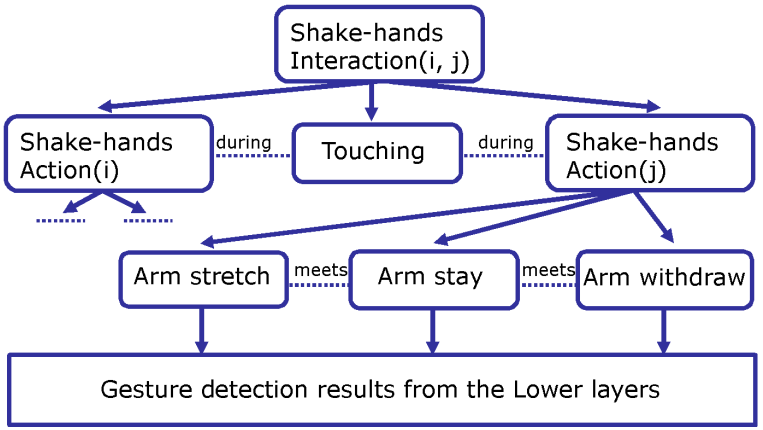
- **Recognition accuracy**
  - **False positive rates are almost 0 because of the detailed representations: previous, deterministic, stochastic.**



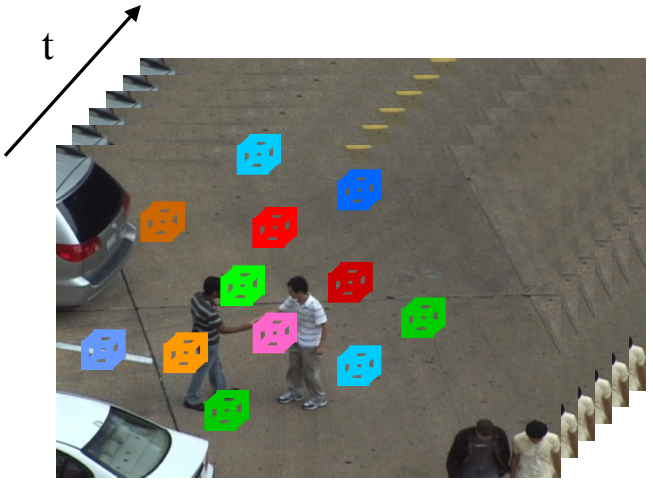
Spatio-Temporal Relationship Match

Description-based vs. Space-time

- Description-based
  - High-level activities
    - Hierarchical
  - Semantic structures
  - Difficult to cope with noise

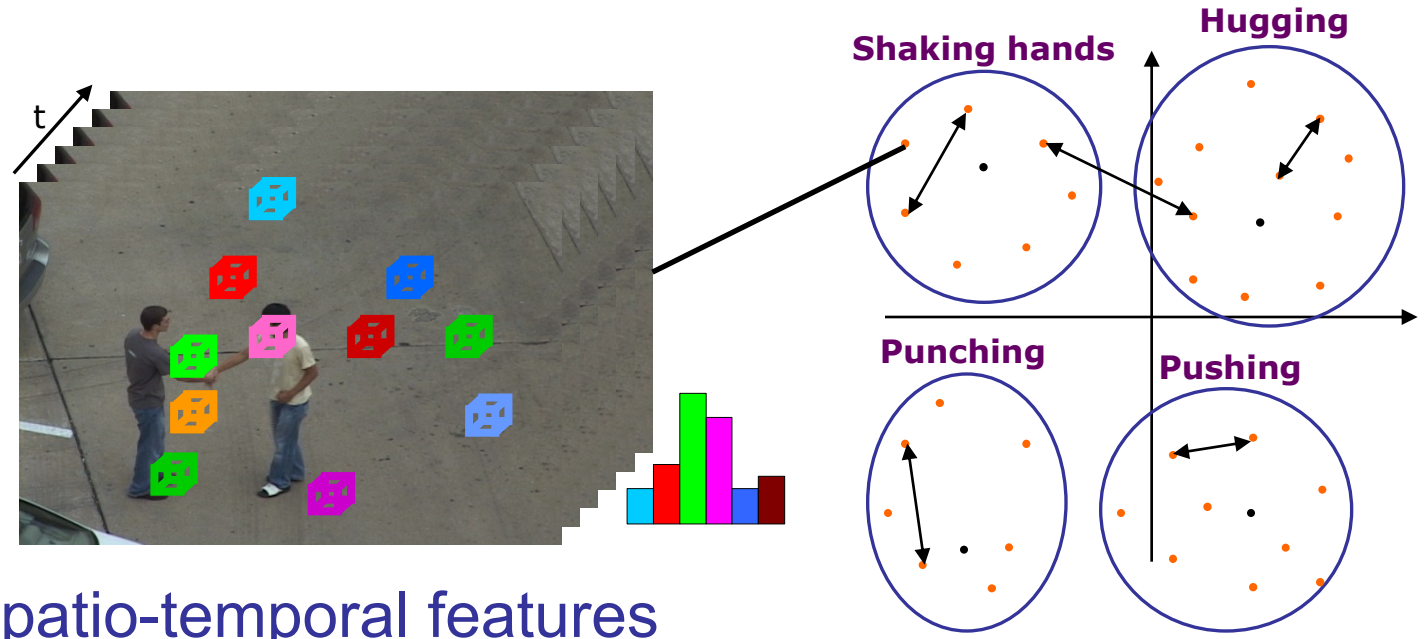


- Space-time
  - Reliable under noise
  - Difficult to model complex activities
  - Miss semantic structures



Space-time approaches

- Video classification
  - Each video is represented as a histogram

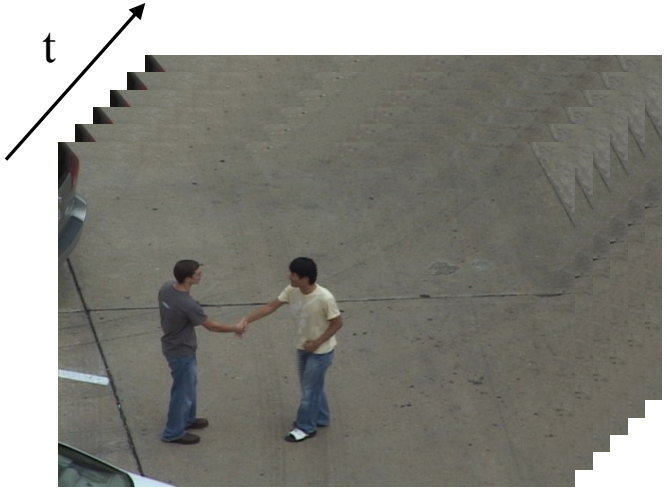


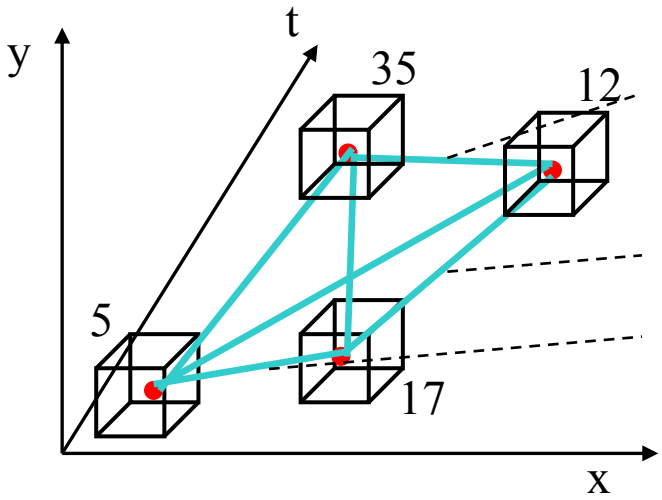
Spatio-temporal features

- Limitation:
  - Unable to model complex activities

Laptev 04,
Dollar et al. 05

Spatio-temporal relations (STRs)



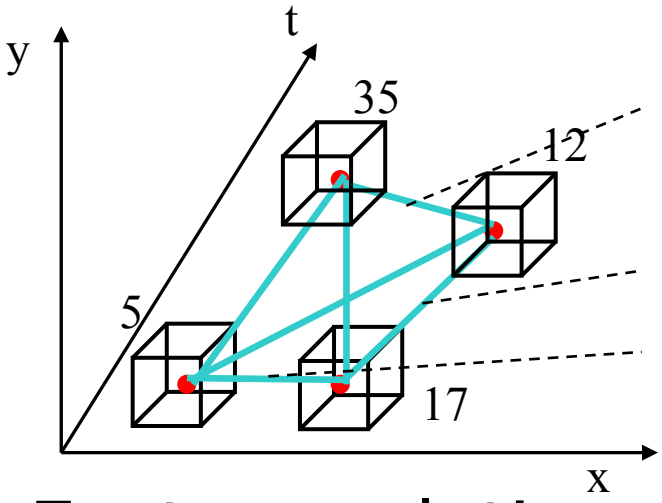


overlaps(12, 35)

...

before(17, 12)

overlaps(5, 17)



overlaps(12, 35)

...

before(17, 12)

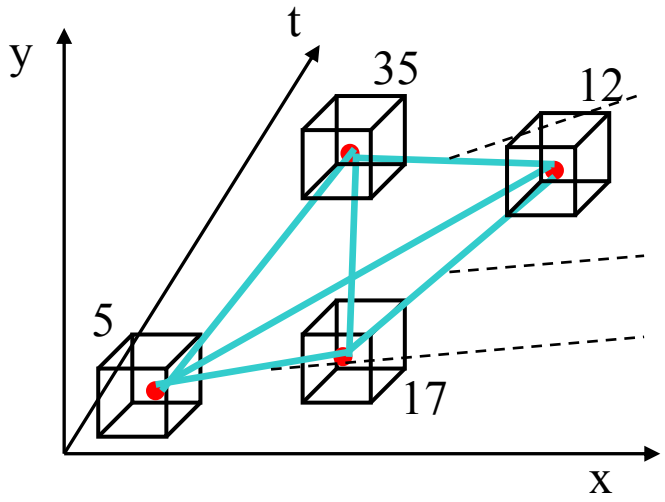
equals(5, 17)

Videos

Feature relations

Ryoo and Aggarwal, ICCV 2009

Histogram of STRs



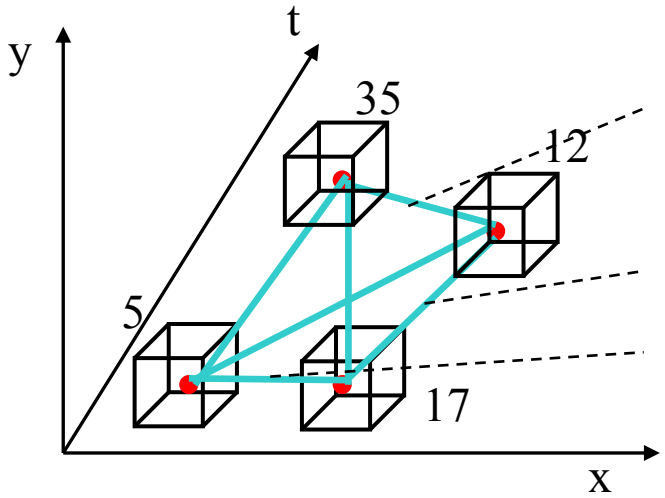
overlaps(12, 35)

...

before(17, 12)

overlaps(5, 17)

5 before 12
5 overlaps 17
5 before 35
12 before-1 17
12 overlaps 35
17 before 35



overlaps(12, 35)

...

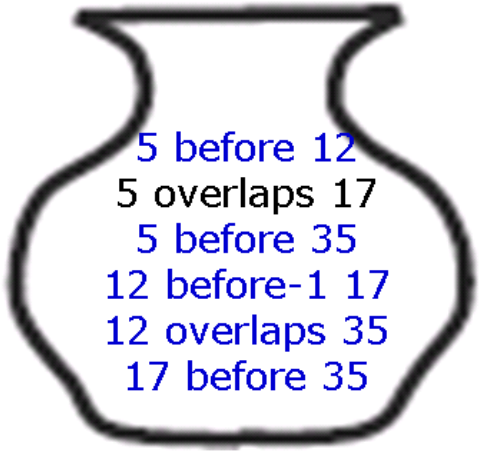
before(17, 12)

equals(5, 17)

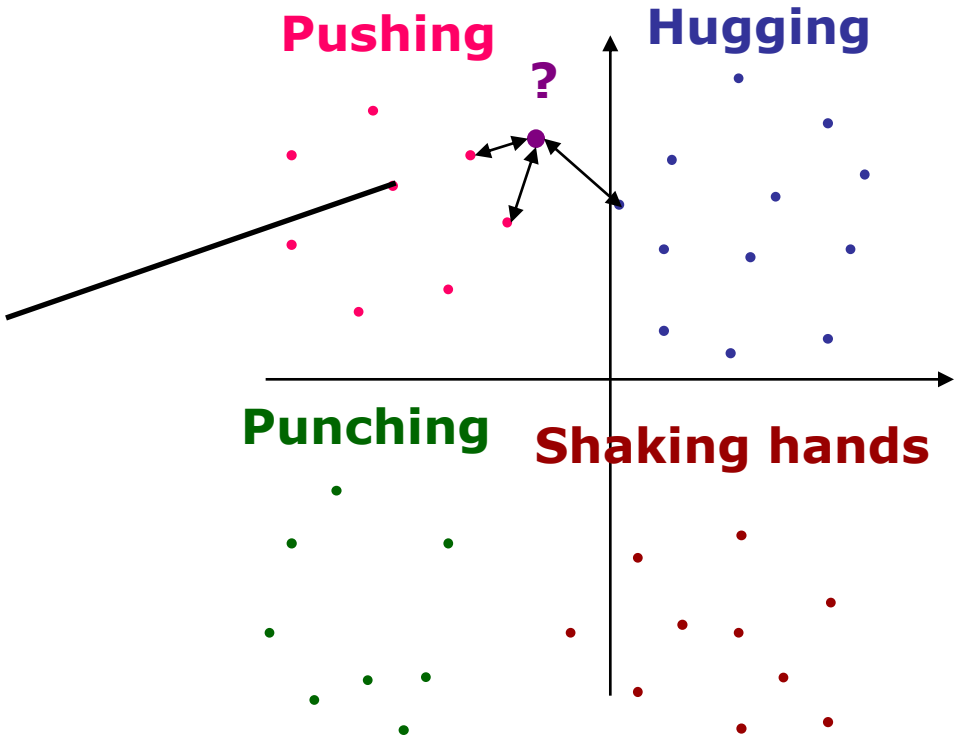
5 before 12
5 equals 17
5 before 35
12 before-1 17
12 overlaps 35
17 before 35

STR-match learning

- Supervised learning
  - Videos with activity labels are provided.



Histogram of Relationships



Feature space

STR equations

- STR match considers distributions of pair-wise relationships among features.
- Histogram construction

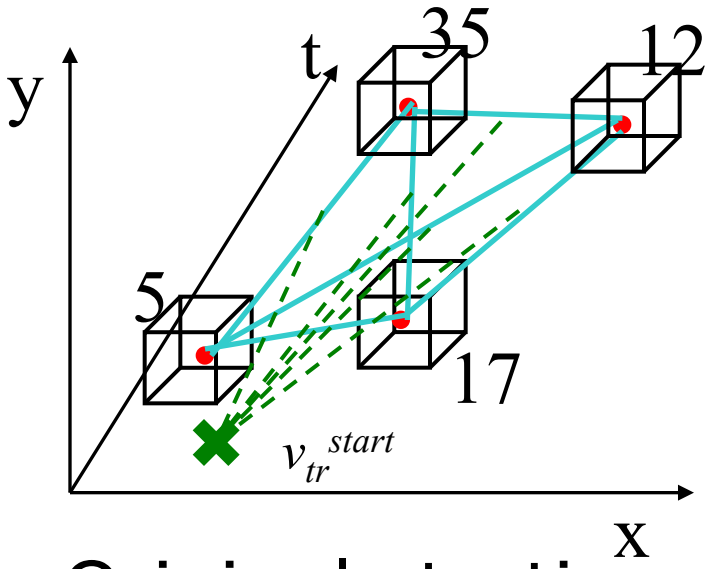
$$T^{trel}_{(i,j)}(v) = \{(f_a, f_b) \mid f_a \in H_i(v) \wedge f_b \in H_j(v) \wedge trel(f_a, f_b) \wedge i < j\}$$

- STR distance:

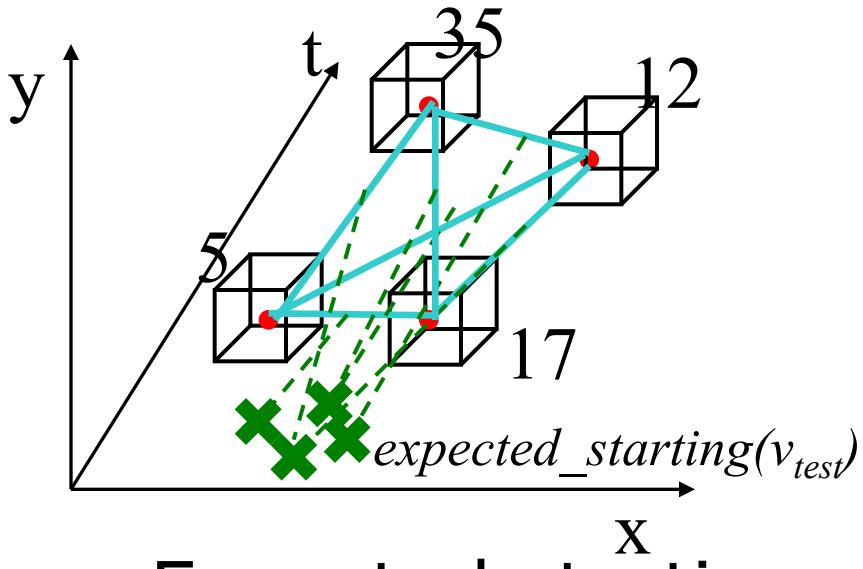
$$K_R(v1, v2) = \sum_{i=1}^{k} \sum_{j=1}^{k} [\sum_{trel} I\left(T^{trel}_{(i,j)}(v1), T^{trel}_{(i,j)}(v2)\right) + \sum_{srel} I\left(S^{srel}_{(i,j)}(v1), S^{srel}_{(i,j)}(v2)\right)]$$

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.



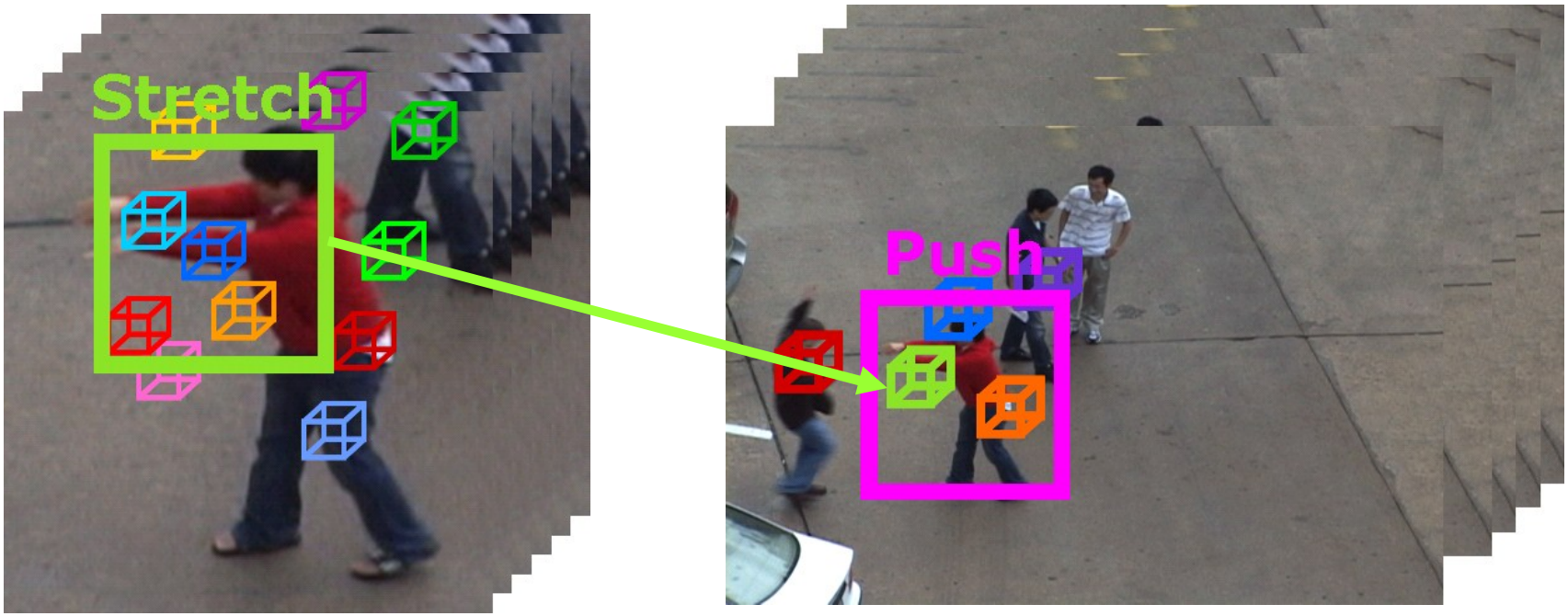
Original starting



Expected starting

Hierarchical recognition

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



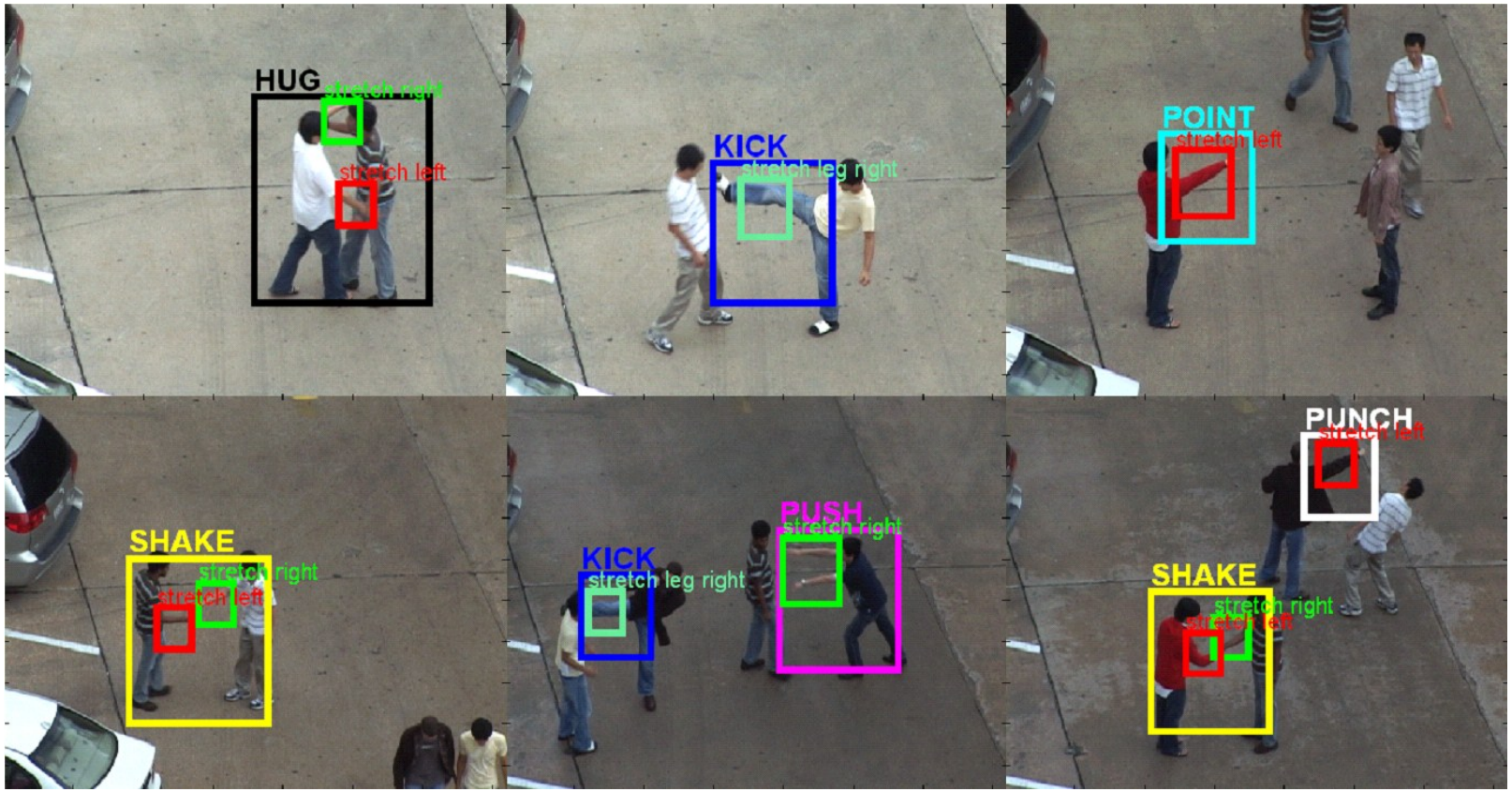
Experiments

- 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, …



STR-match summary

- Detection from continuous videos
  - Localization using voting-based method
- Noisy observations
  - Different backgrounds/lightings
  - Uncertainties
- Human-human interactions
  - Hierarchical recognition
- Future work
  - Hierarchy learning algorithm

Description-based: References

- Allen, J. F. and Ferguson, G., Actions and events in interval temporal logic. Journal of Logic and Computation 1994.
- Pinhanez, C. S. and Bobick, A. F., Human action detection using PNF propagation of temporal constraints. CVPR 1998.
- 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.
- Nevatia, R., Zhao, T., and Hongeng, S., Hierarchical language-based representation of events in video streams. In IEEE Workshop on Event Mining 2003.
- Vu, V.-T., Bremond, F., and Thonnat, M., Automatic video interpretation: A novel algorithm for temporal scenario recognition. IJCAI 2003.
- Ryoo, M. S. and Aggarwal, J. K., Recognition of composite human activities through context-free grammar based representation. CVPR 2006.
- Ryoo, M. S. and Aggarwal, J. K., Hierarchical recognition of human activities interacting with objects. CVPR-SLAM 2007.
- Tran, S. D. and Davis, L. S., Event modeling and recognition using markov logic networks. ECCV 2008.
- Ryoo, M. S. and Aggarwal, J. K., Semantic representation and recognition of continued and recursive human activities. IJCV, 2009.
- Ryoo, M. S. and Aggarwal, J. K., Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. ICCV 2009.
- Ryoo, M. S. and Aggarwal, J. K., Stochastic representation and recognition of high-level group activities. IJCV, 2010.