# Discriminative Latent Variable Models for Human Action Recognition

**Greg Mori**
School of Computing Science
Simon Fraser University

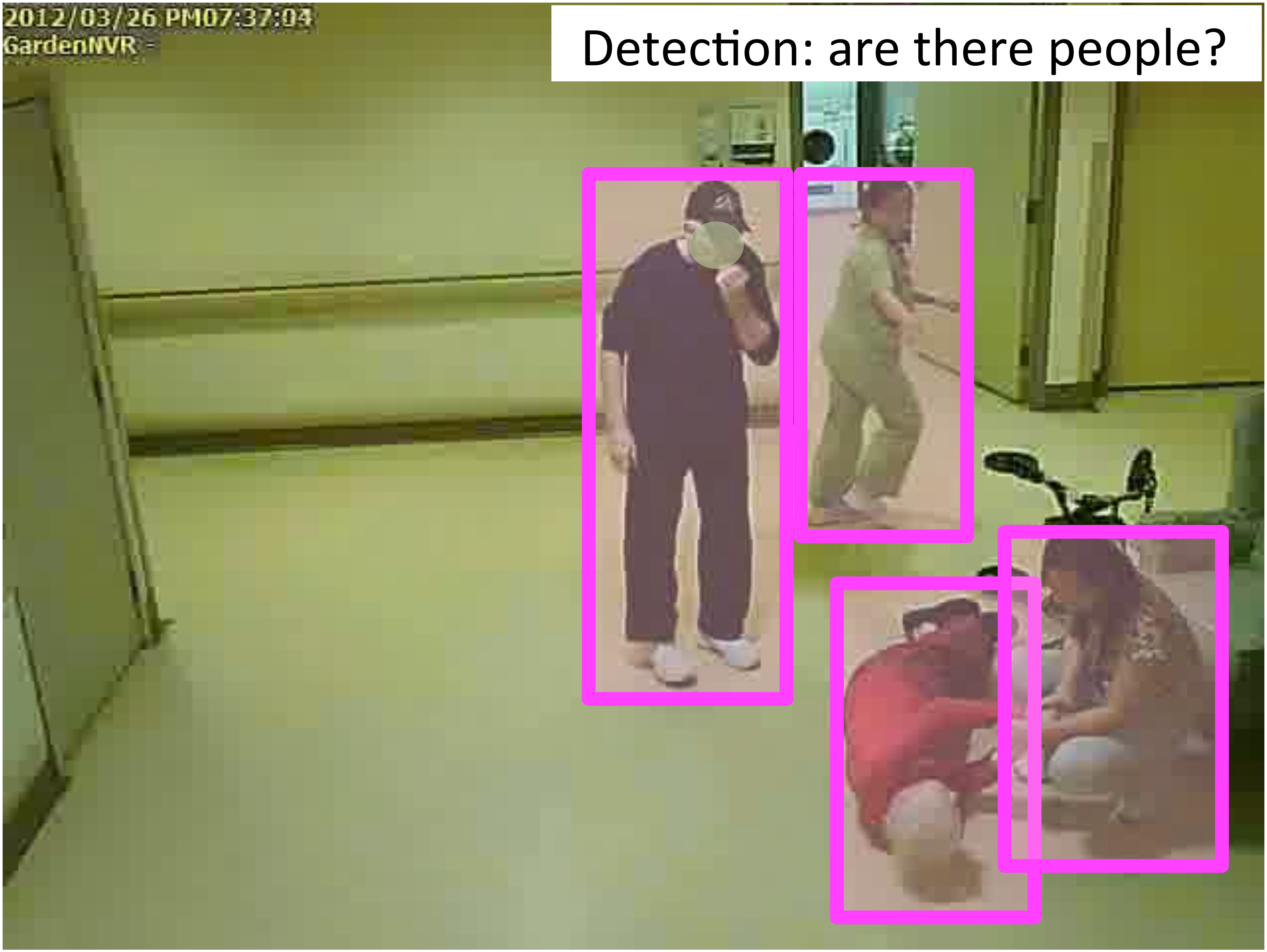
ICCV HACI'13 Workshop
December 8, 2013

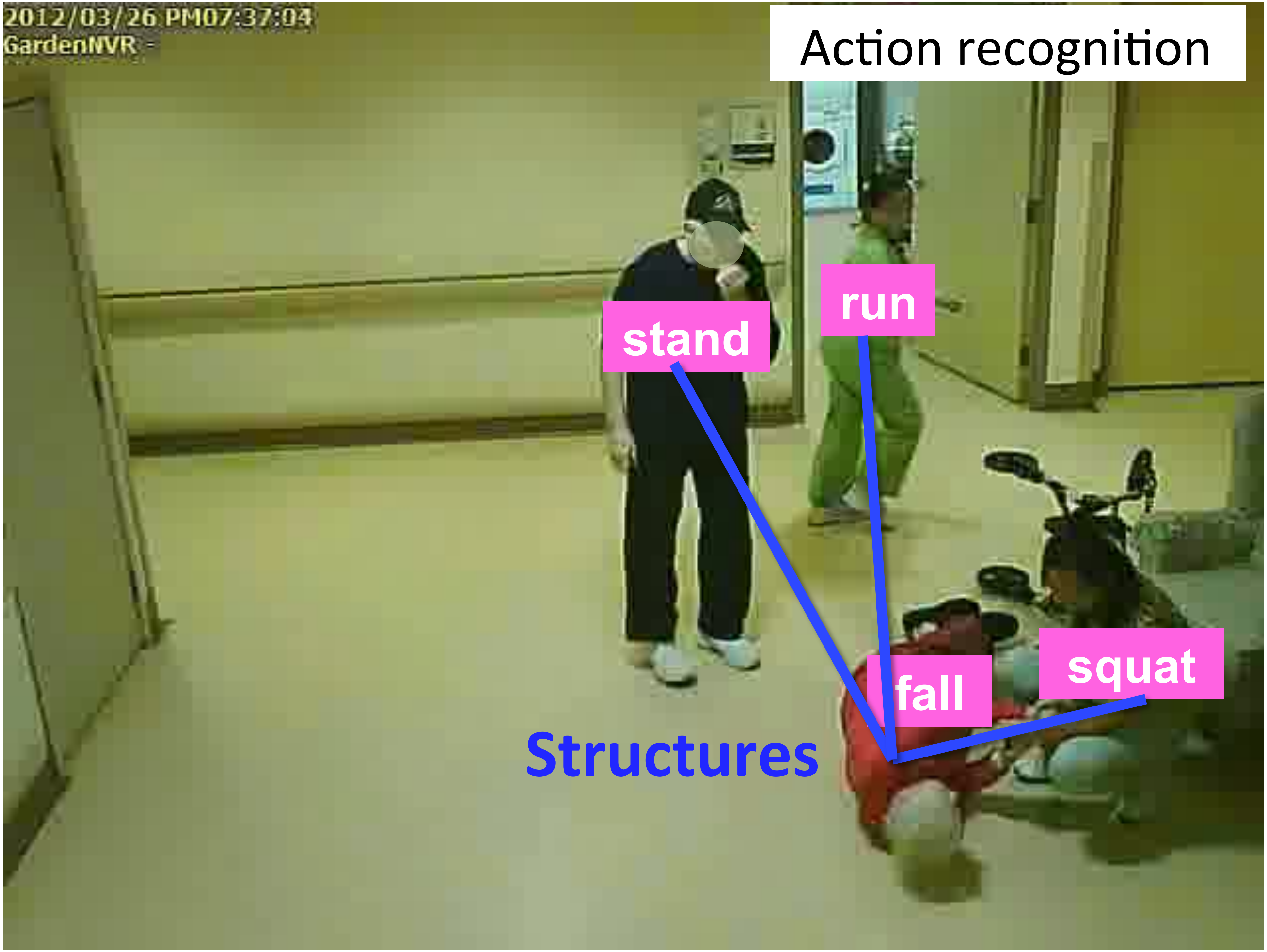
Joint work with: Tian Lan, Weilong Yang, Yang Wang, Steve Robinovitch, Leonid Sigal

SFU

SFU

# What does activity recognition involve?


2012/03/26 PM07:37:04
GardenNVR -
Detection: are there people?


2012/03/26 PM07:37:04
GardenNVR
Action recognition
stand
run
fall
squat
Structures


2012/03/26 PM07:37:04
GardenNVR -
Group activity recognition
help the fallen person


2012/03/26 PM07:37:04
GardenNVR -
Intention/social role
get help
watch
comfort

# Advantages of Modeling Structures

- Analyze levels of detail
  - Body parts vs. whole
  - Actions of individuals
  - Relationships between individuals
  - Overall scene-level understanding

- Provide context for recognition

# Activity landscape

## Actions

Run

1

- Bobick & Davis, 2001
- Efros et al, 2003
- Schuldt et al, 2004
- Alper & Shah, 2005
- Dollar et al, 2005
- Blank et al, 2005
- Niebles et al, 2006
- Laptev et al, 2008
- Wang & Mori, 2008
- Rodriguez et al, 2008
- Wang & Mori, 2009
- Liu et al, 2009
- Marszalek et al, 2009
- ……

## Human interactions

Point

2

- Oliver et al, 1998
- Park & Aggarwal, 2004
- Ryoo & Aggarwal, 2006
- Ryoo & Aggarwal, 2009
- Yuan et al, 2010
- Vahdat et al, 2011
- Patron-Perez et al, 2012

## Group activities

Talk

5-10

- Cupillard et al, 2002
- Moore & Essa, 2002
- Vaswani et al, 2003
- Khan & Shah, 2003
- Zhang et al, 2006
- Mehran et al, 2009
- Gupta et al, 2009
- Choi & Savarese, 2009
- Lan et al, 2010
- Ryoo & Aggarwal, 2010
- Choi & Savarese, 2011
- Amer & Todorovic, 2011
- ……

## Events

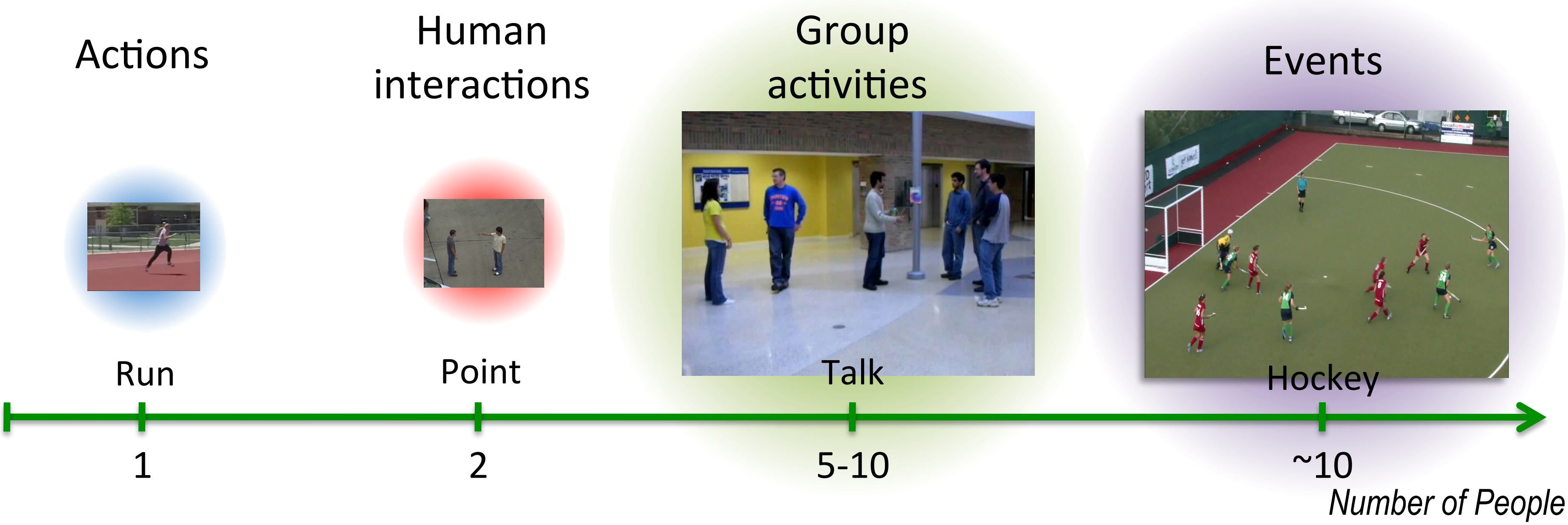
Hockey

~10

*Number of People*

- Intille & Bobick, 2001
- Medioni et al, 2001
- Loy et al, 2010
- Lan et al, 2012
- Amer et al, 2012

# Activity landscape




- Performed by multiple people
- Rich human-human interactions
- Events may consist of multiple group activities, and inter-group interactions

# Activity landscape

| Actions | Human interactions | Group activities | Events |
|---|---|---|---|
|  |  |  |  |
| Run | Point | Talk | Hockey |
| 1 | 2 | 5-10 | ~10 |

*Number of People*

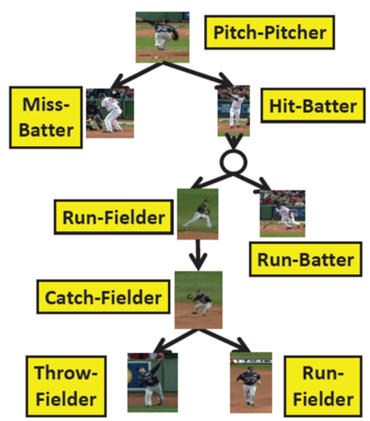
**Possible approaches:**

*Bag of features*

- Statistical methods
- Don't extract semantic descriptions

Laptev et al, 2008
Liu et al, 2009
Tamrakar et al, 2012

*DBN, AND-OR Graph, CRF, Latent SVM*




- Structural methods
- Complex learning / inference

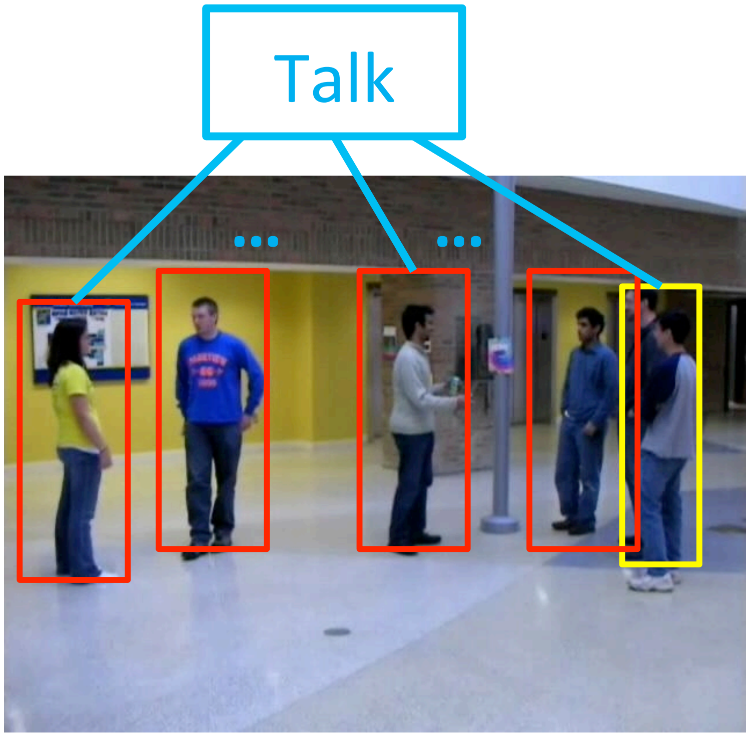
Xiang & Gong, 2006
Gupta et al, 2009
Felzenszwalb et al, 2010
Amer et al, 2012

# Our Proposal - Structured Models

- Models that account for spatial, temporal, relational, or other structures
  - Flexible
  - Richer representation
- This talk: representation and learning of structured models for activity recognition
- These can be applied across the activity landscape, from individual human actions through to group events

# Role of Context in Actions

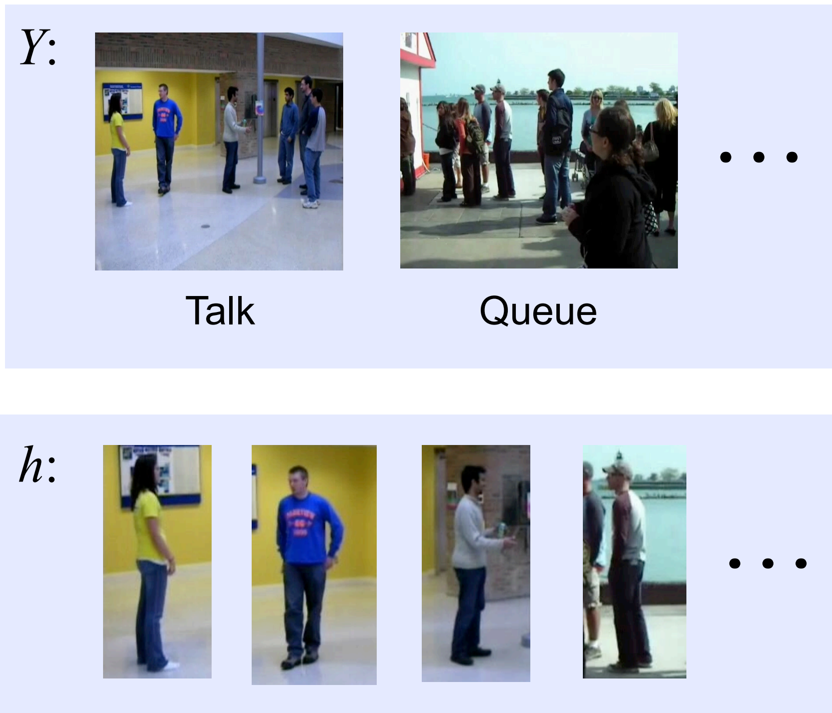
# Group Context



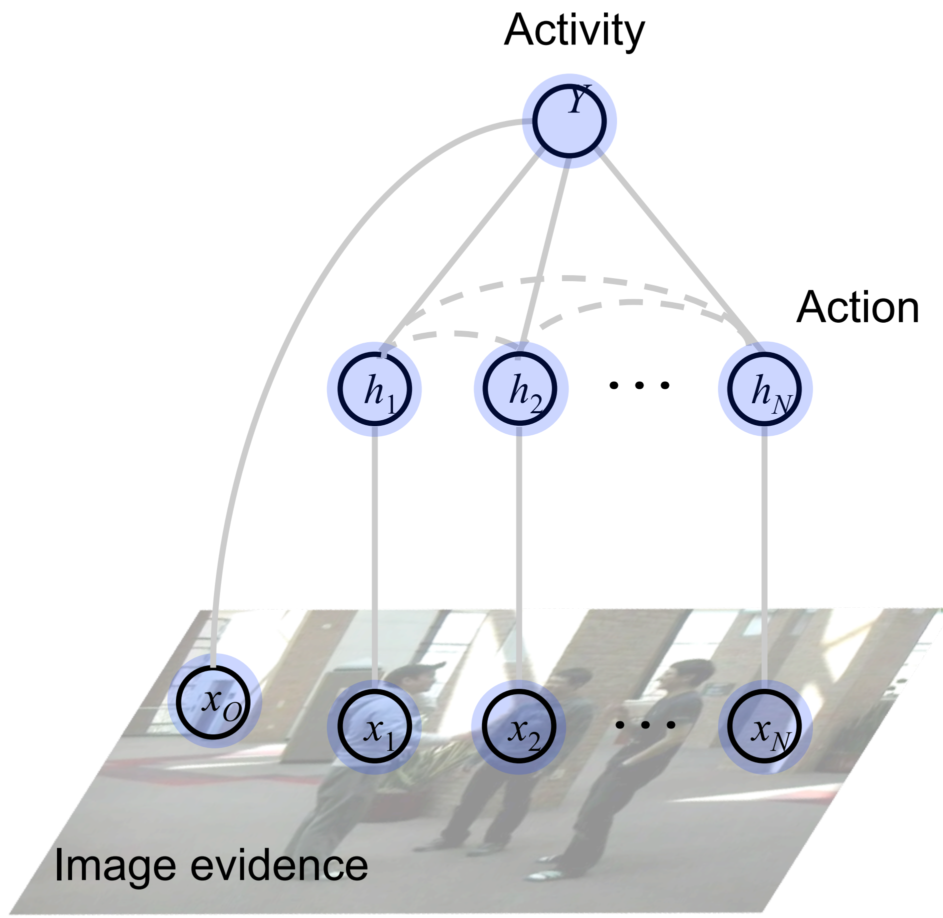

group-person interaction

person-person interaction

# Model of Group Activities



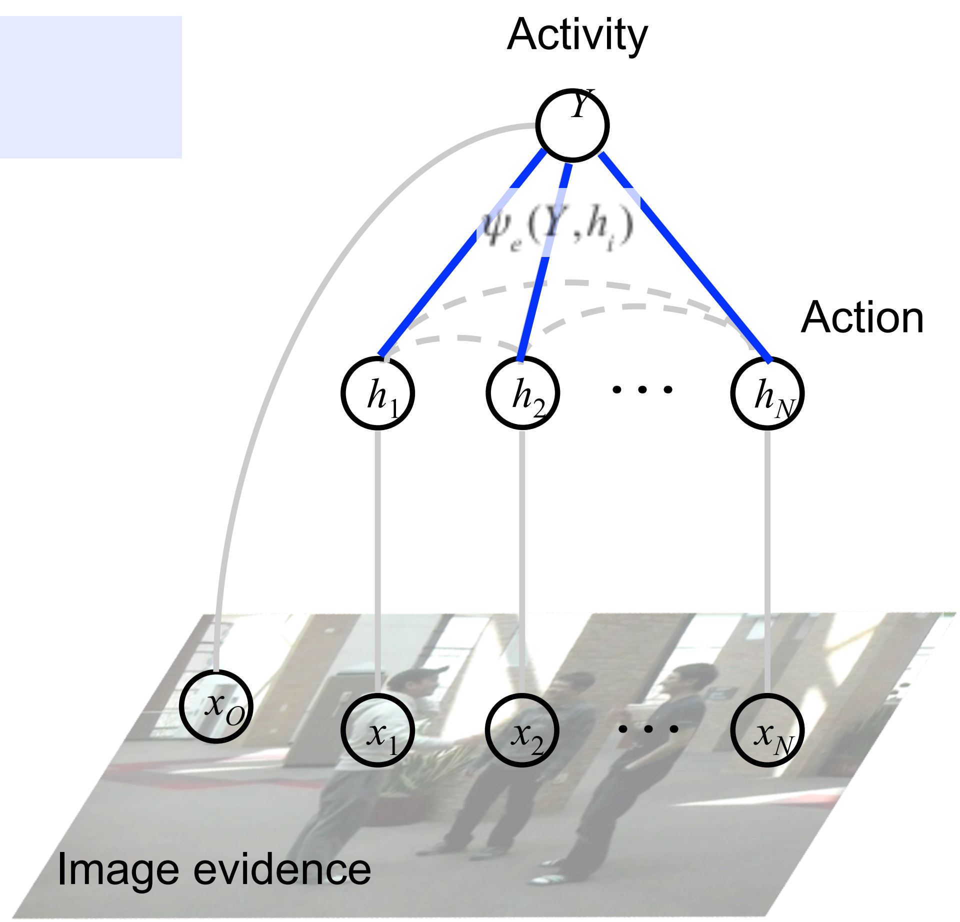

$x$: HOG [Dalal & Triggs, 2005]




Lan et al. NIPS 2010, TPAMI 2012

# Model of Group Activities

- Activity-Action Potential $\psi_e(Y, h_i)$ :
  Co-occurrence between $Y$ and $h_i$




**Markov Random Field**

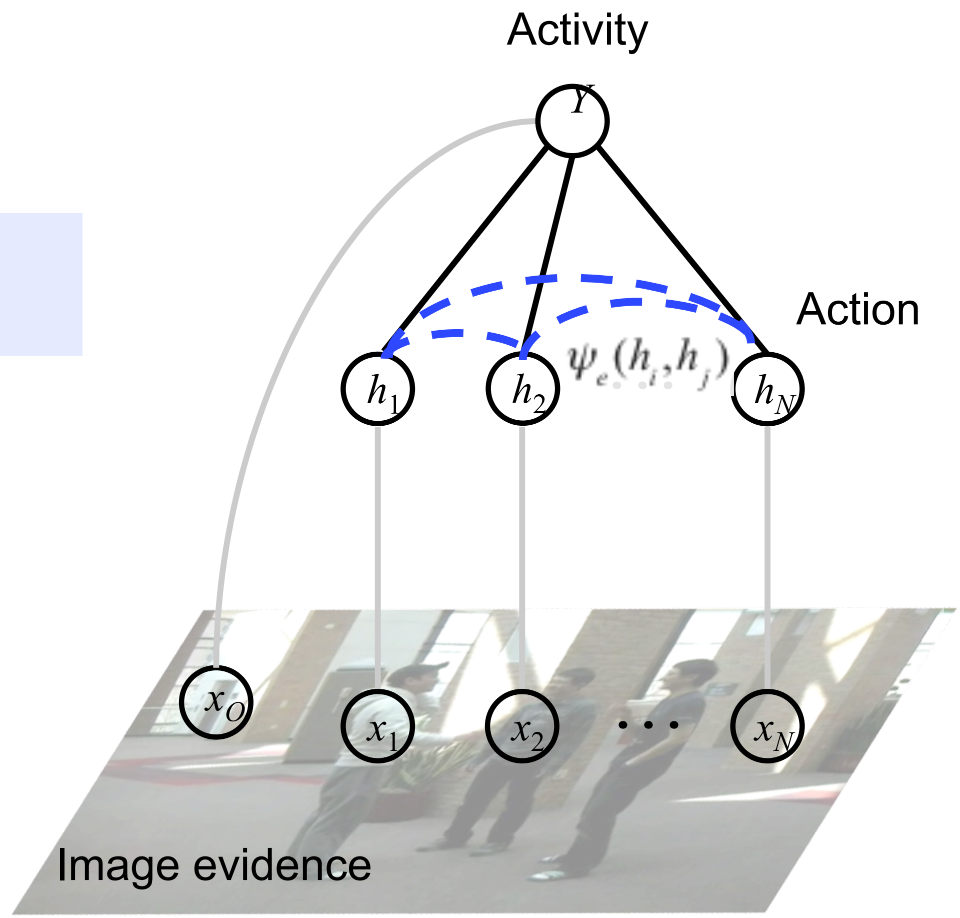
$$\Psi = \sum_{e \in E} w_e \psi_e$$

Clique weight   Clique potential

# Model of Group Activities

- Activity-Action Potential $\psi_e(Y, h_i)$ :
  Co-occurrence between $Y$ and $h_\mathrm{i}$

- Action-Action Potential $\psi_e(h_i, h_j)$ :
  Co-occurrence between $h_\mathrm{i}$ and $h_\mathrm{j}$




**Markov Random Field**

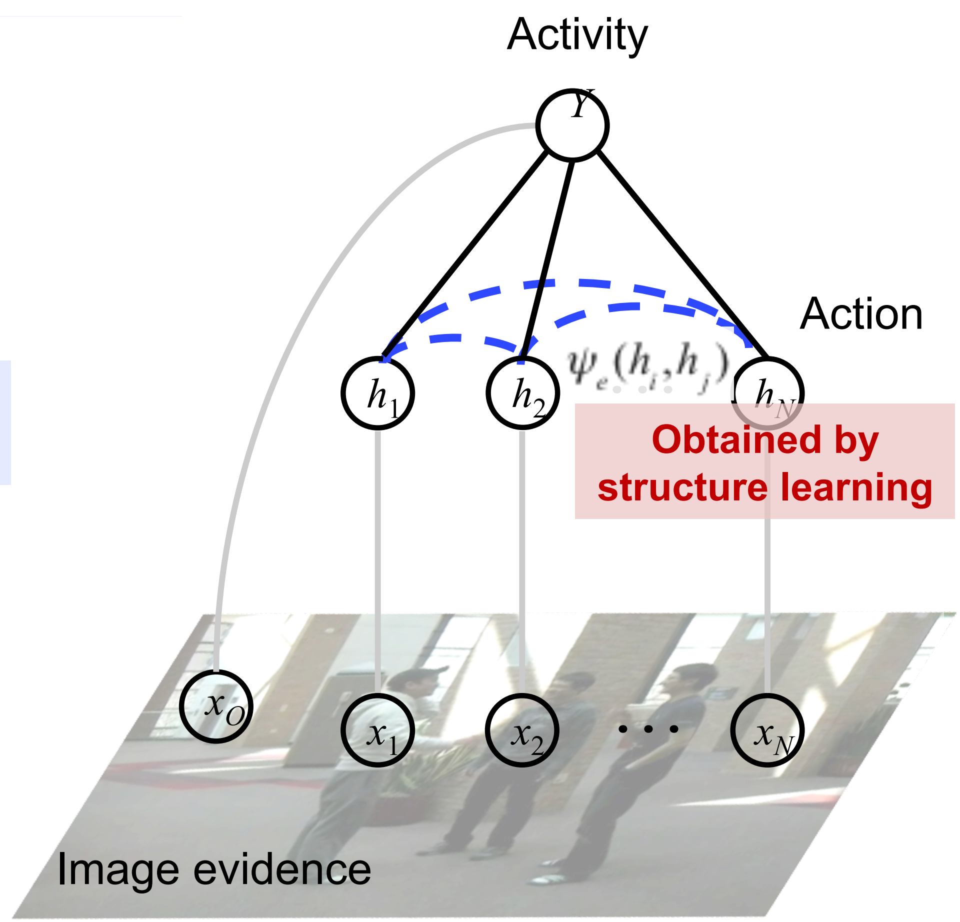
$$\Psi = \sum_{e \in E} w_e \psi_e$$

Clique weight    Clique potential

# Model of Group Activities

- Activity-Action Potential $\psi_e(Y, h_i)$ :
  Co-occurrence between $Y$ and $h_i$

- Action-Action Potential $\psi_e(h_i, h_j)$ :
  Co-occurrence between $h_i$ and $h_j$
  - Learn structural connectivity among the actions.




**Markov Random Field**

$$\Psi = \sum_{e \in E} w_e \psi_e$$

Clique weight　Clique potential

# Model of Group Activities

- Activity-Action Potential $\psi_e(Y, h_i)$ :
  Co-occurrence between $Y$ and $h_i$

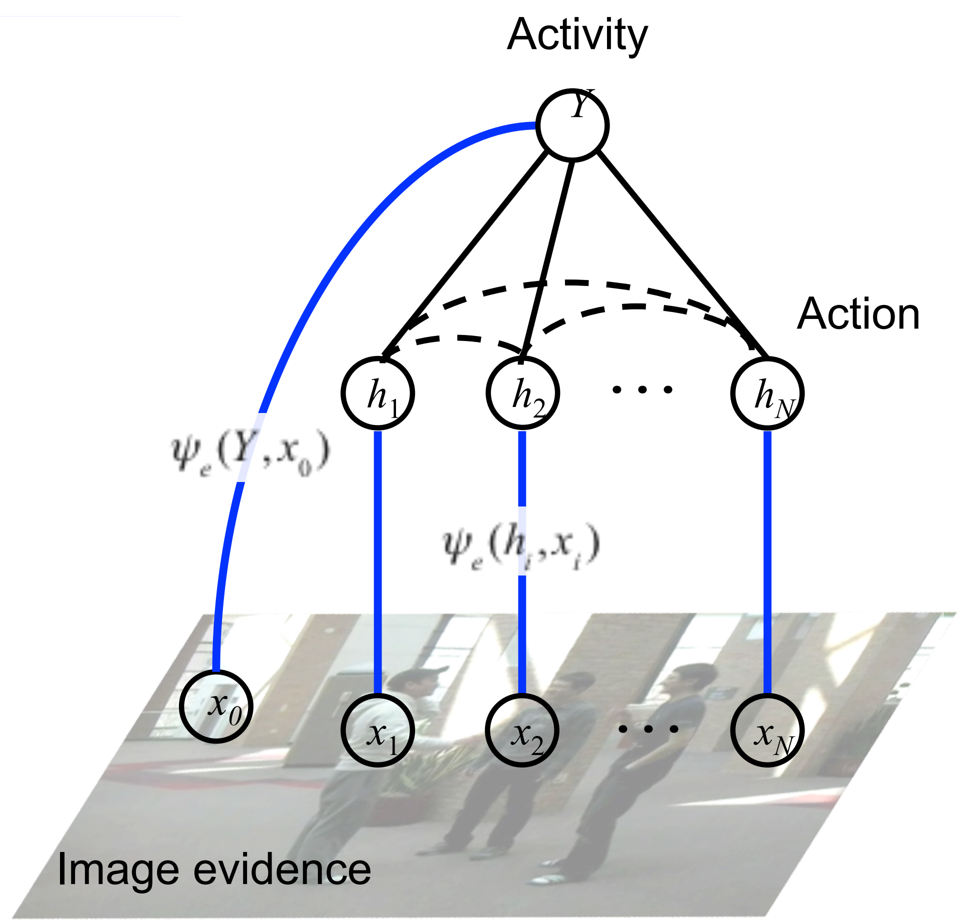
- Action-Action Potential $\psi_e(h_i, h_j)$ :
  Co-occurrence between $h_i$ and $h_j$
  - Learn structural connectivity among the actions.

- $\psi_e(Y, x_0)$ and $\psi_e(h_i, x_i)$ :
  Discriminative action template scores (HOG + SVM).

**Markov Random Field**

$$\Psi = \sum_{e \in E} w_e \psi_e$$

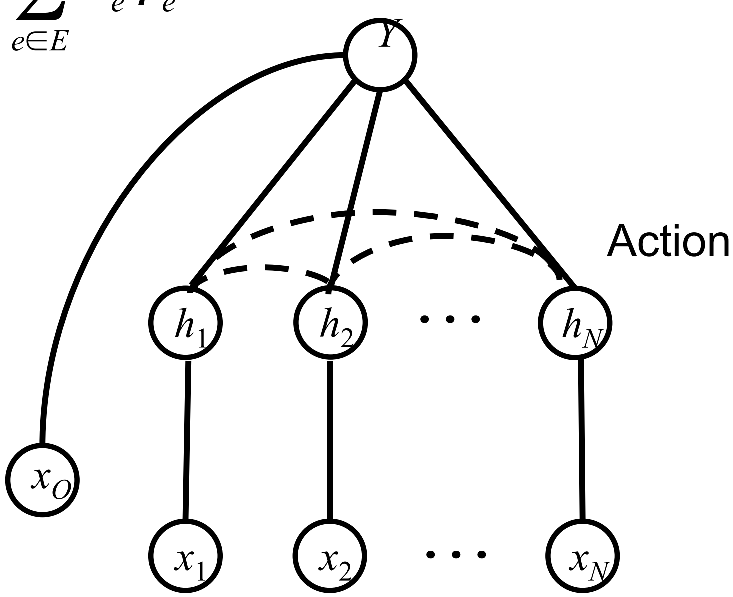
Clique weight    Clique potential



# Model Learning

$$\Psi = \sum_{e \in E} w_e \psi_e$$

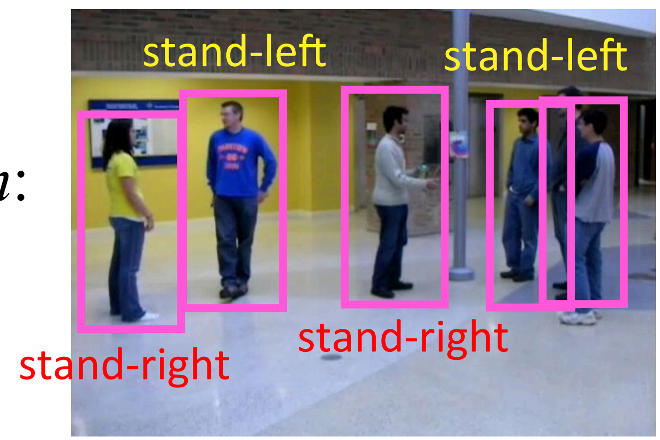
Activity




**Input:**

$Y$: talk

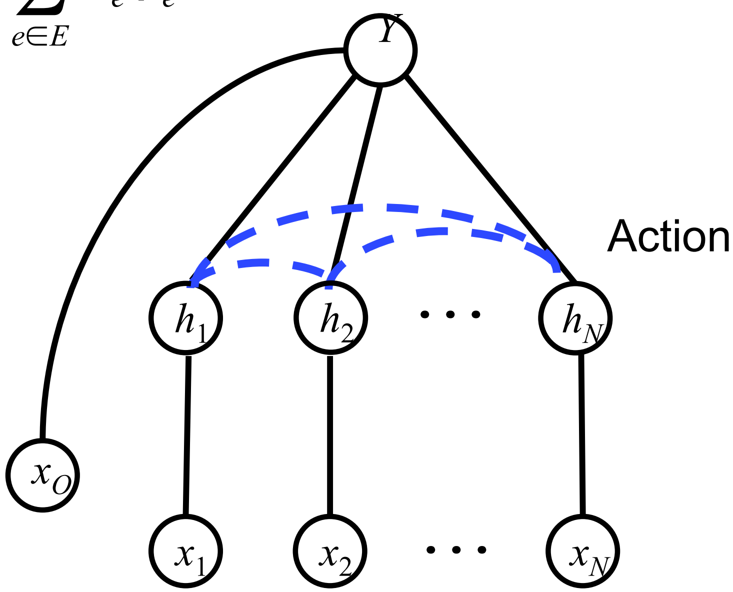
$h$:




**Goals:**

# Model Learning

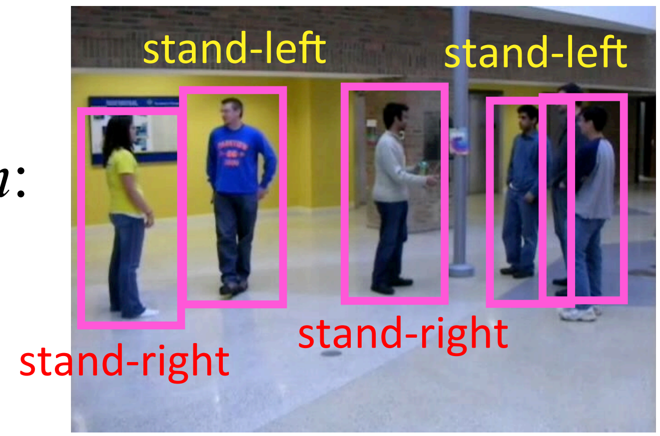
$$\Psi = \sum_{e \in E} w_e \psi_e$$




**Input:**

$Y$: talk

$h$:



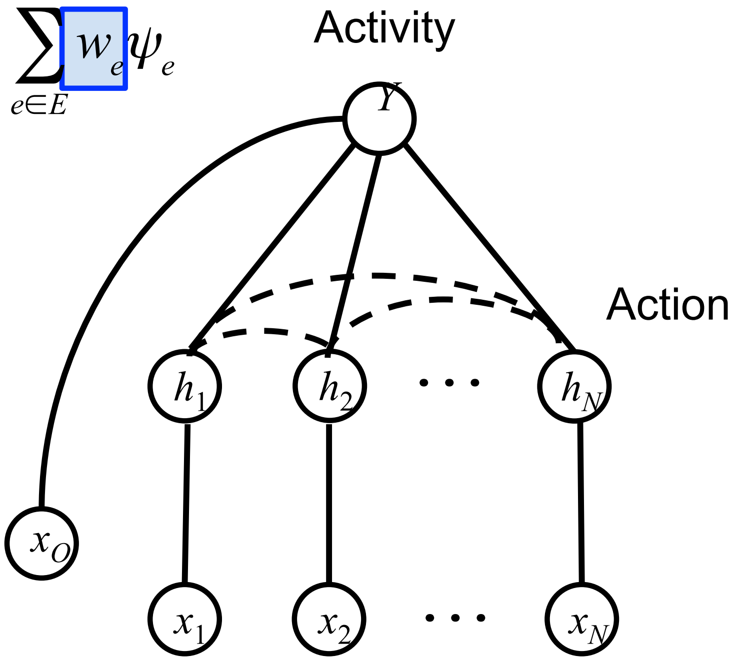

**Goals:**

Structural connectivity (hidden human-human interactions)

Potential weights

# Model Learning

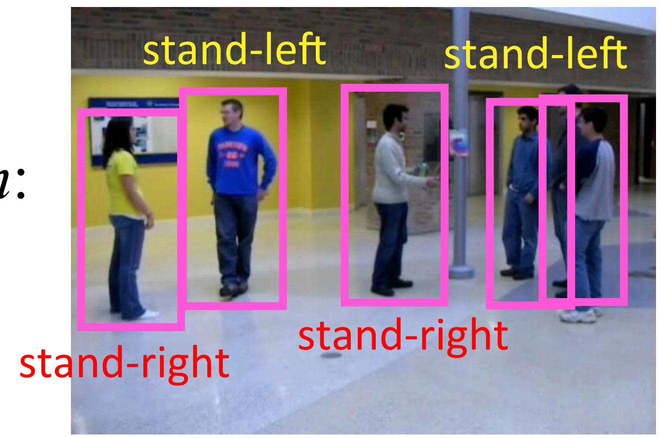
$$\Psi = \sum_{e \in E} \boxed{w_e} \psi_e$$




**Input:**

$Y$: talk

$h$:



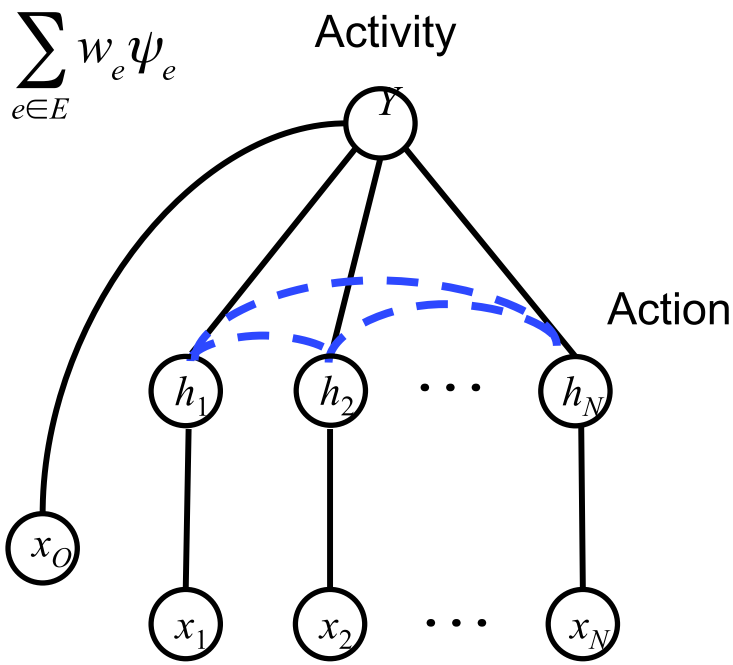

**Goals:**

Structural connectivity (hidden human-human interactions)

**Potential weights**

# Model Learning

$$\Psi = \sum_{e \in E} w_e \psi_e$$

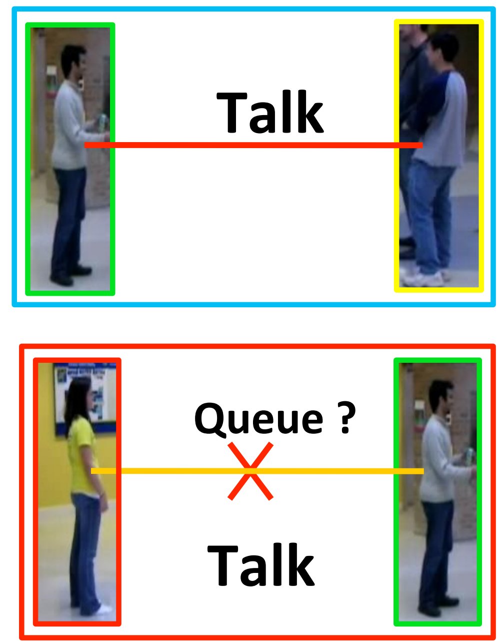



**Goals:**

**Structural connectivity**

Potential weights

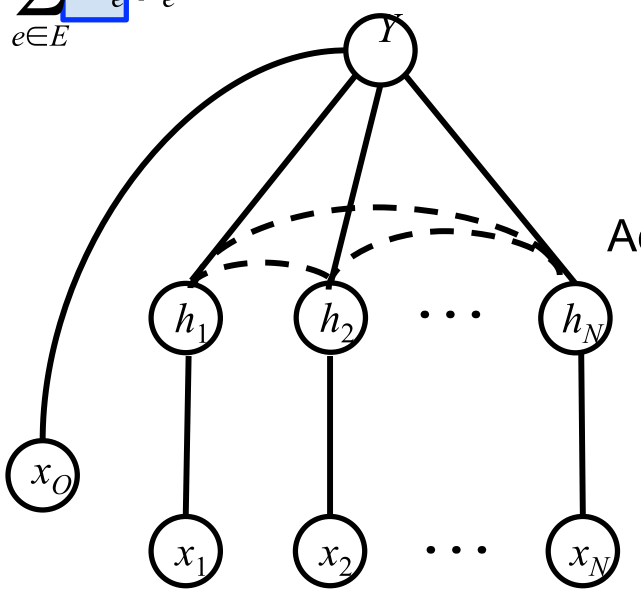
**Approach:**




ILP $\max_{E=\{e\}} \sum_e w_e \psi_e$

# Model Learning

$$\Psi = \sum_{e \in E} \boxed{w_e} \psi_e$$




## Goals:

Structural connectivity

**Potential weights**

## Approach:

Max-margin learning

$$\min_{\mathbf{w},\xi} \frac{1}{2} \sum_r \|\mathbf{w}_r\|_2^2 + \beta \sum_i \xi_i$$

$$\text{s.t. } \forall i, r \text{ where } y(r) \neq y(c_i),$$

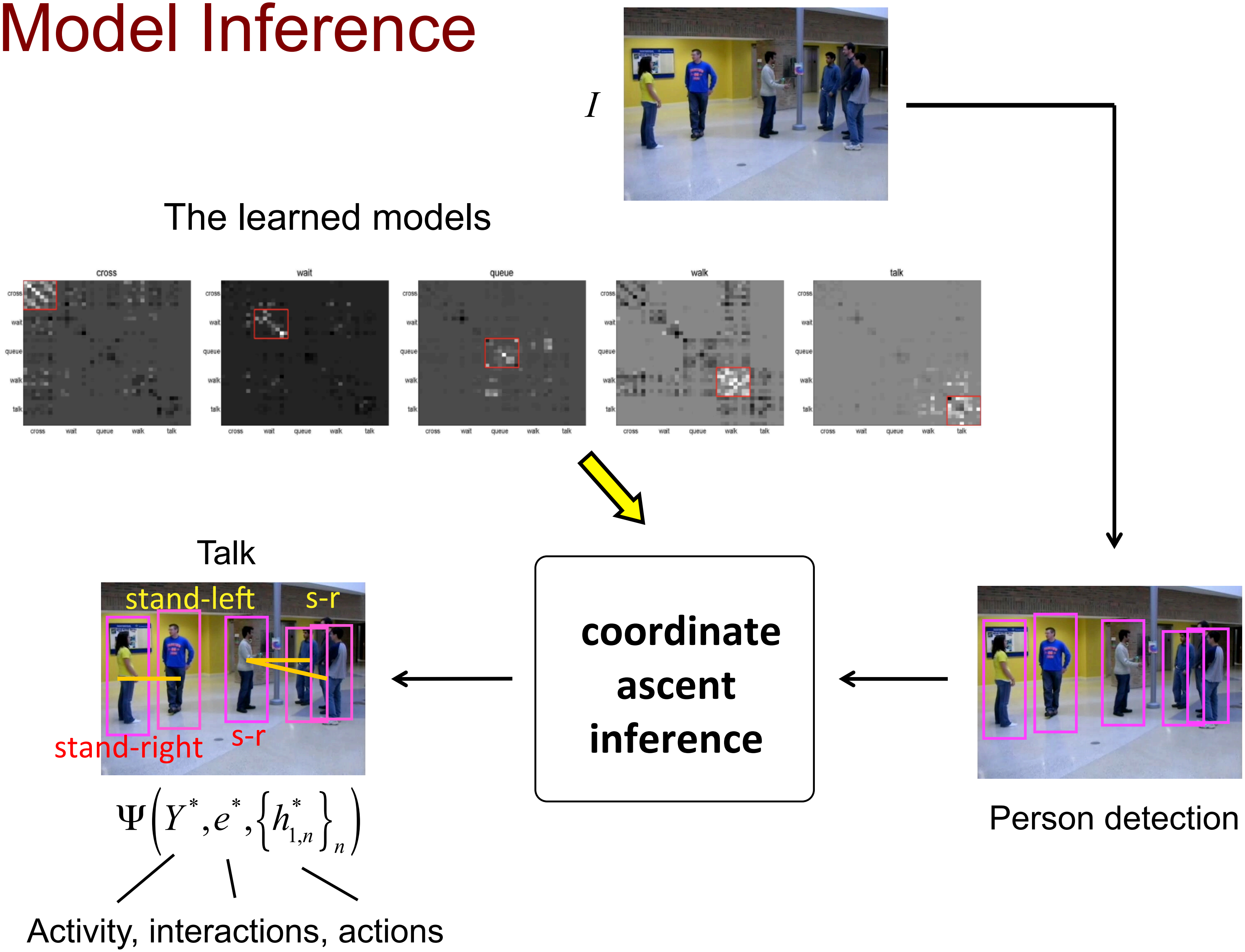
$$\mathbf{w}_{c_i} \cdot \psi_i - \mathbf{w}_r \cdot \psi_i \geq 1 - \xi_i$$

$$\forall i, \xi_i \geq 0$$

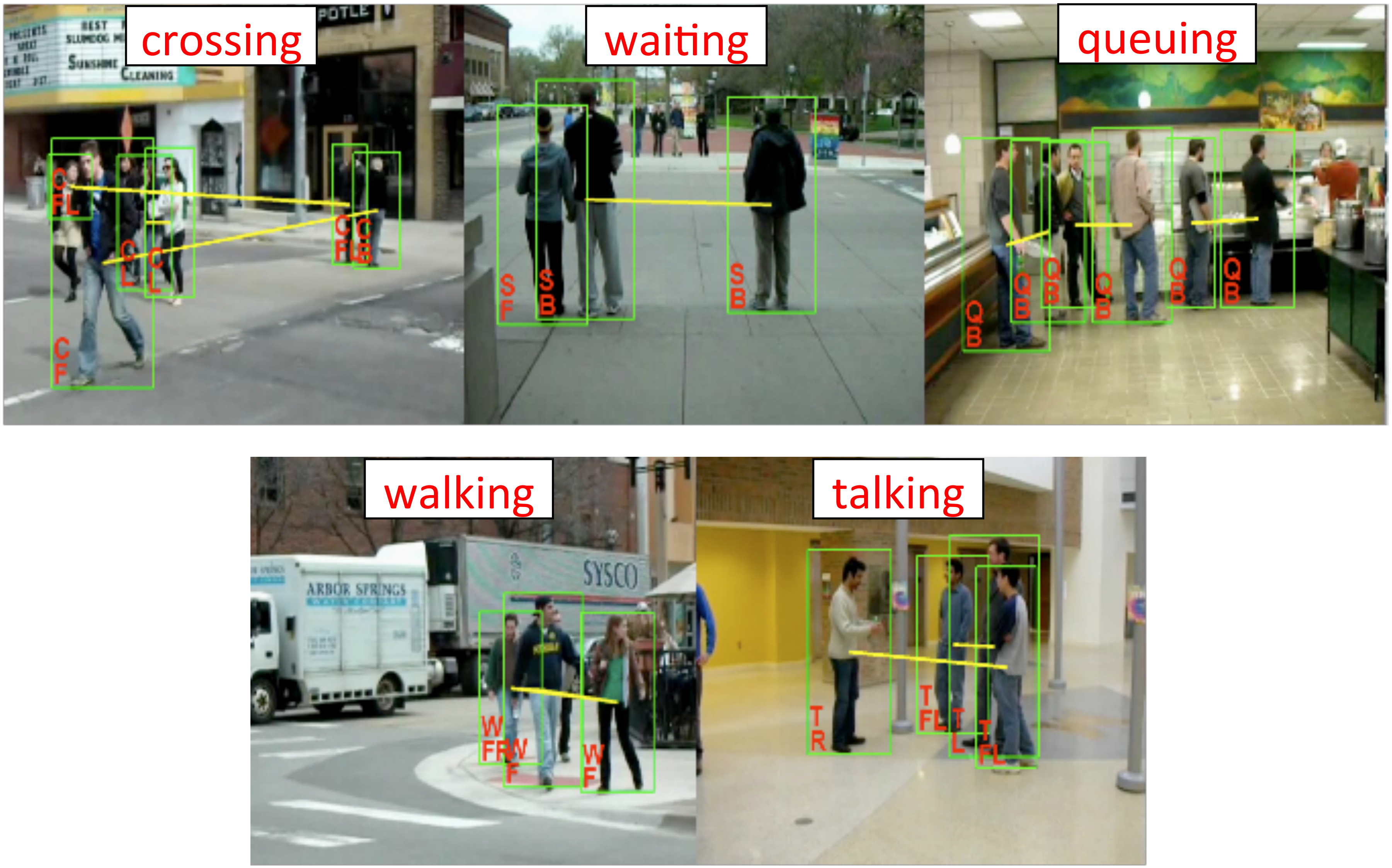
**Notation**

- $\psi_i$ : Potential values of the $i$-th image.
- $\mathbf{w}_r$: Potential weights of the $r$-th activity.
- $y(r)$: $r$-th activity class.
- $\xi_i$: A slack variable for the $i$-th image.

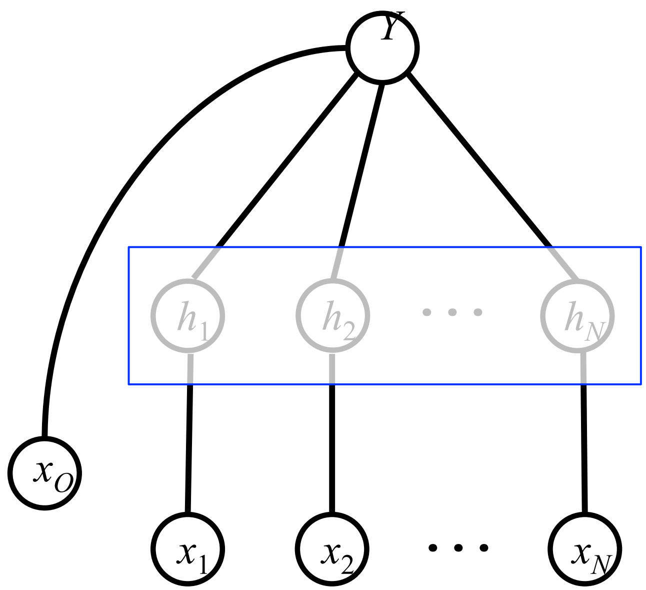
# Model Inference



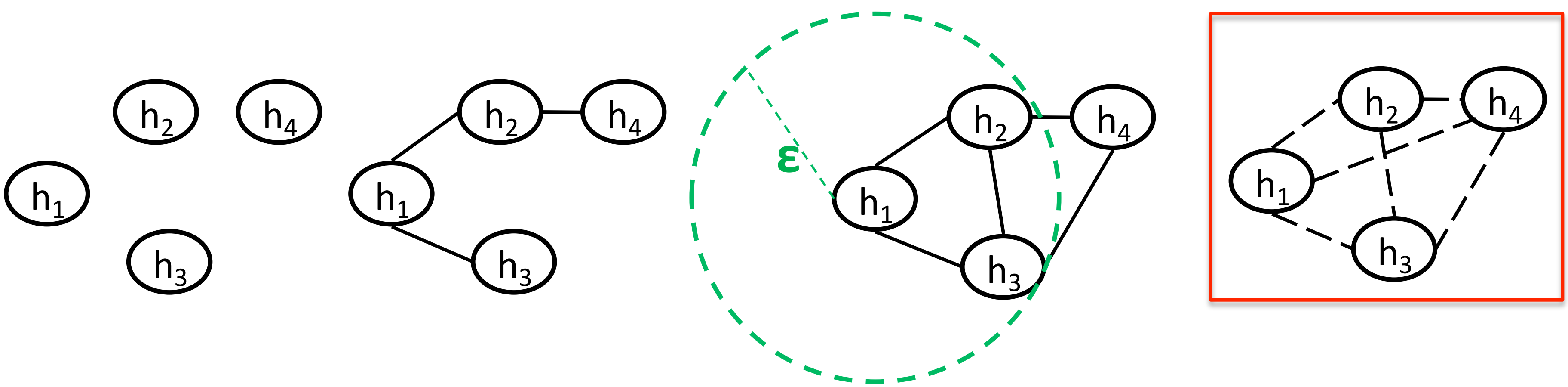
# Visualization of the Results



# Baselines




- SVM
- No connection
- Min-spanning tree
- ε-neighborhood graph



# Results – Collective Activity Dataset

| Method | Overall | Mean per-class |
|---|---|---|
| SVM | 70.9 | 68.6 |
| no connection | 75.9 | 73.7 |
| min-spanning tree | 73.6 | 70.0 |
| ε-neighborhood graph, ε=100 | 74.3 | 72.9 |
| ε-neighborhood graph, ε=200 | 70.4 | 66.2 |
| ε-neighborhood graph, ε=300 | 62.2 | 62.5 |
| complete graph | 62.6 | 58.7 |
| our approach | 79.1 | 77.5 |

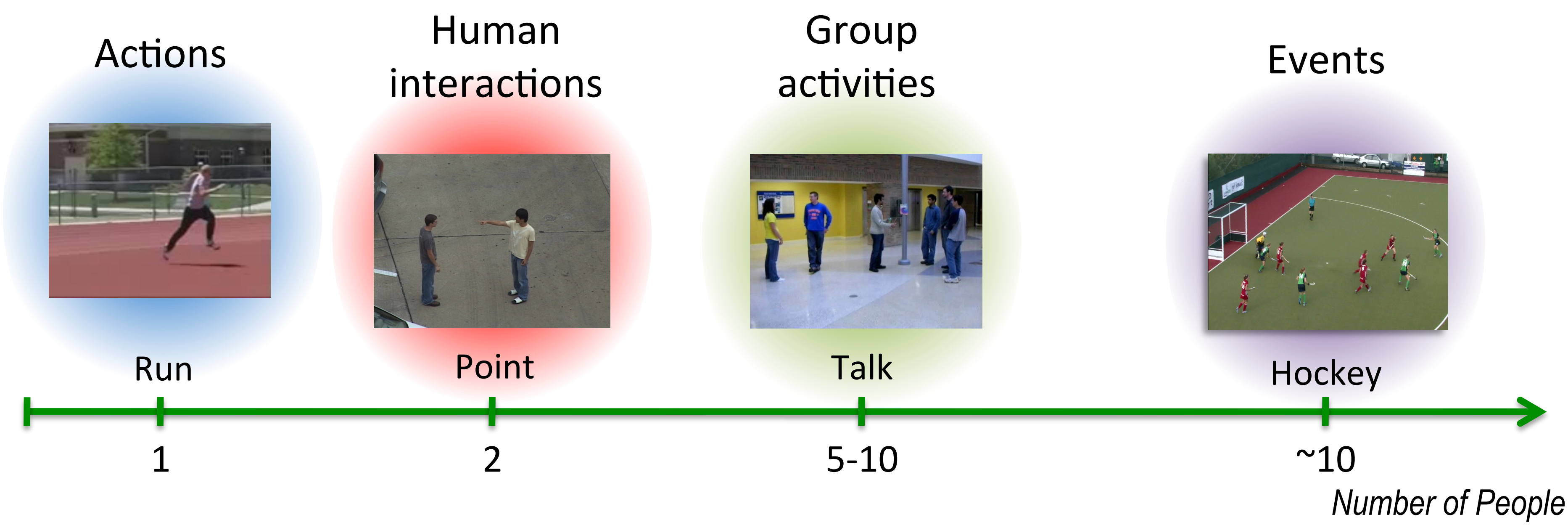
# Nursing Home Data


- 22 short clips of fall + a 30-min non-fall clip, 5 actions, 2 group activities

# Results – Nursing Home Data

| Method | Overall | Mean per-class |
|---|---|---|
| SVM | 48.0 | 52.4 |
| no connection | 54.4 | 56.1 |
| min-spanning tree | 66.9 | 62.3 |
| ε-neighborhood graph, ε=100 | 72.7 | 61.3 |
| ε-neighborhood graph, ε=200 | 67.6 | 61.1 |
| ε-neighborhood graph, ε=300 | 68.6 | 64.2 |
| complete graph | 70.6 | 62.2 |
| our approach | 71.5 | 67.4 |

# Roadmap




• Tian Lan, Leonid Sigal, Greg Mori. Social Roles in Hierarchical Models for Human Activity Recognition. CVPR 2012

# Semantic Descriptions of Videos

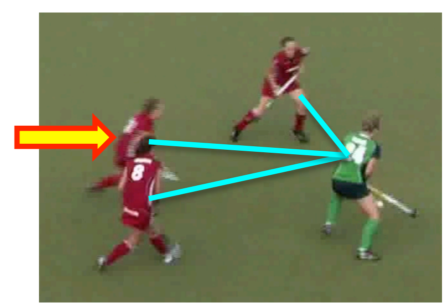
| actions | social roles | event |
|---|---|---|
| walk | attacker | corner hit |
| run | first defenders | free hit |
| jog | man-marking | attack play |
| bend | defend-space | |
| shoot | teammate | |
| dribble | | |
| pass | | |

**Social Roles**

- Mid-level semantics that describe individual/group behaviors in the context of social interactions.

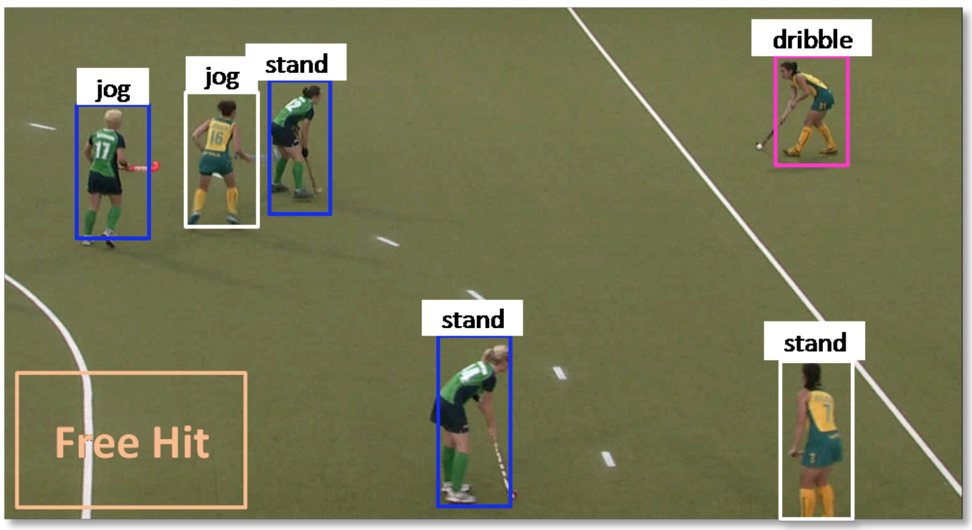
man-marking

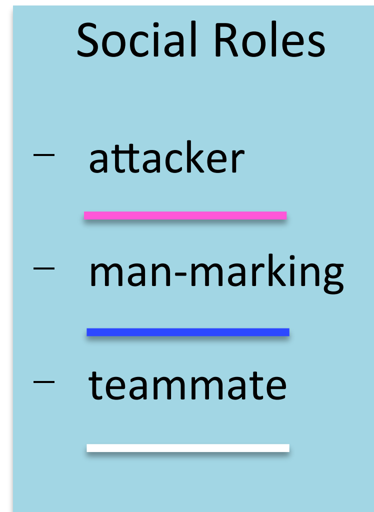
first defenders



# Goal

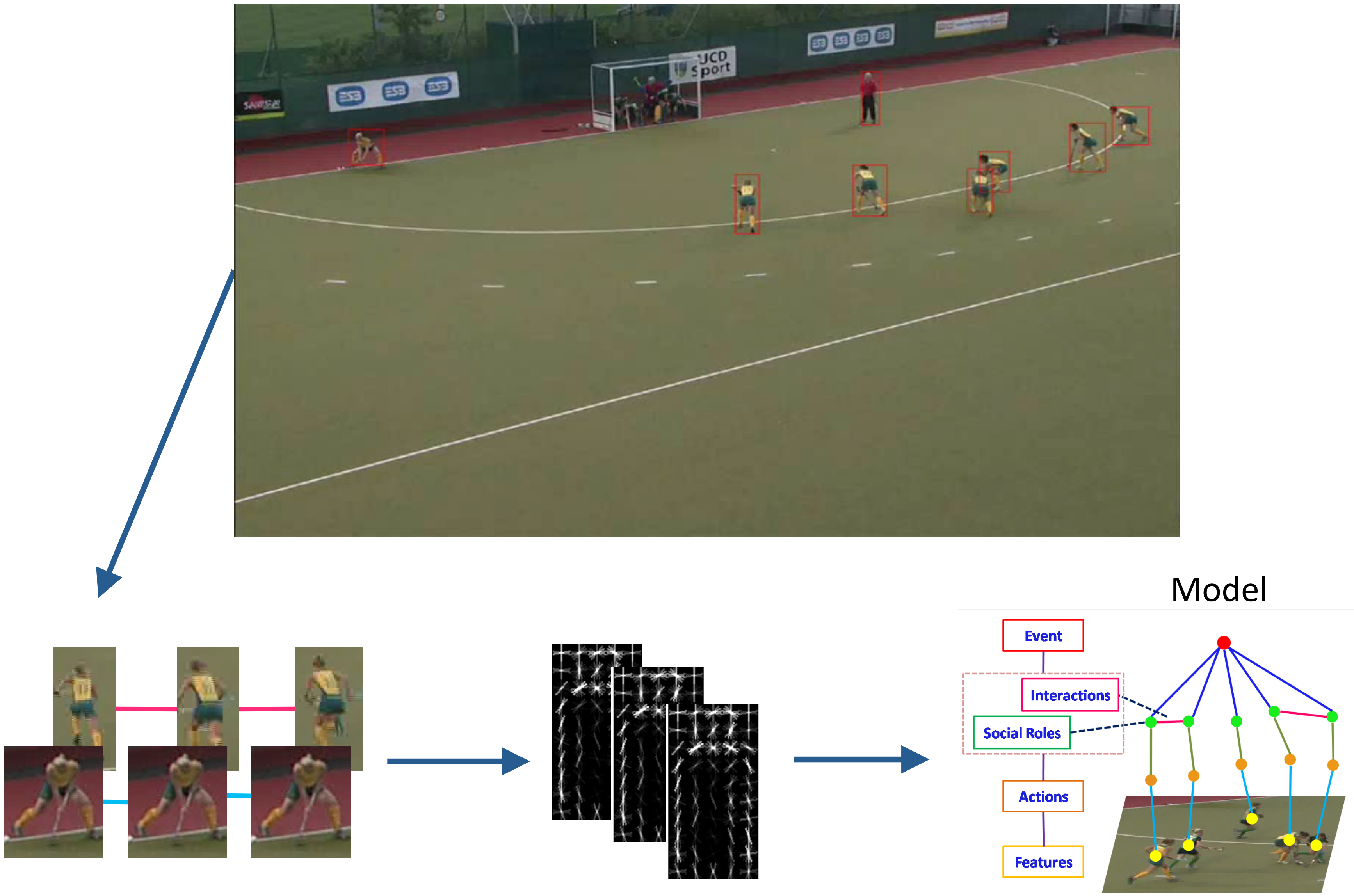
- Label all individuals' actions, social roles and the scene-level events.



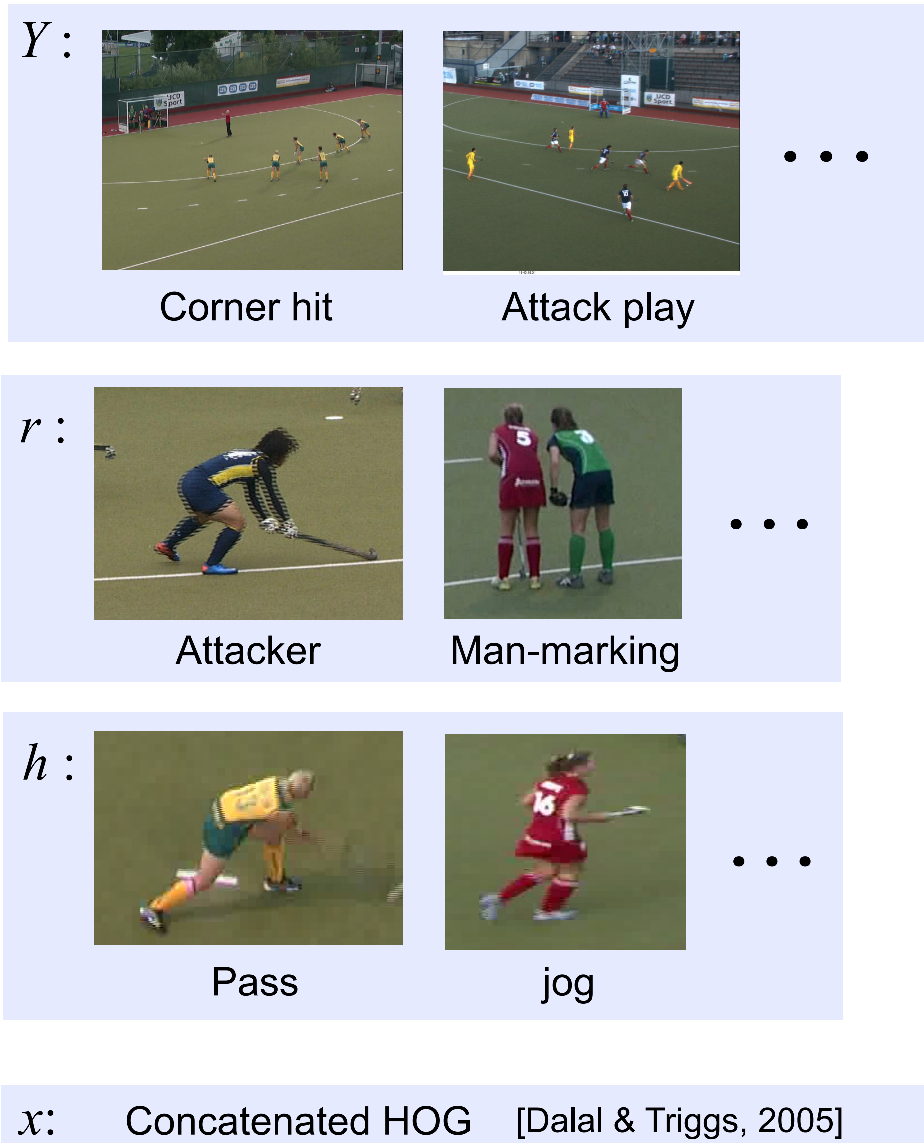



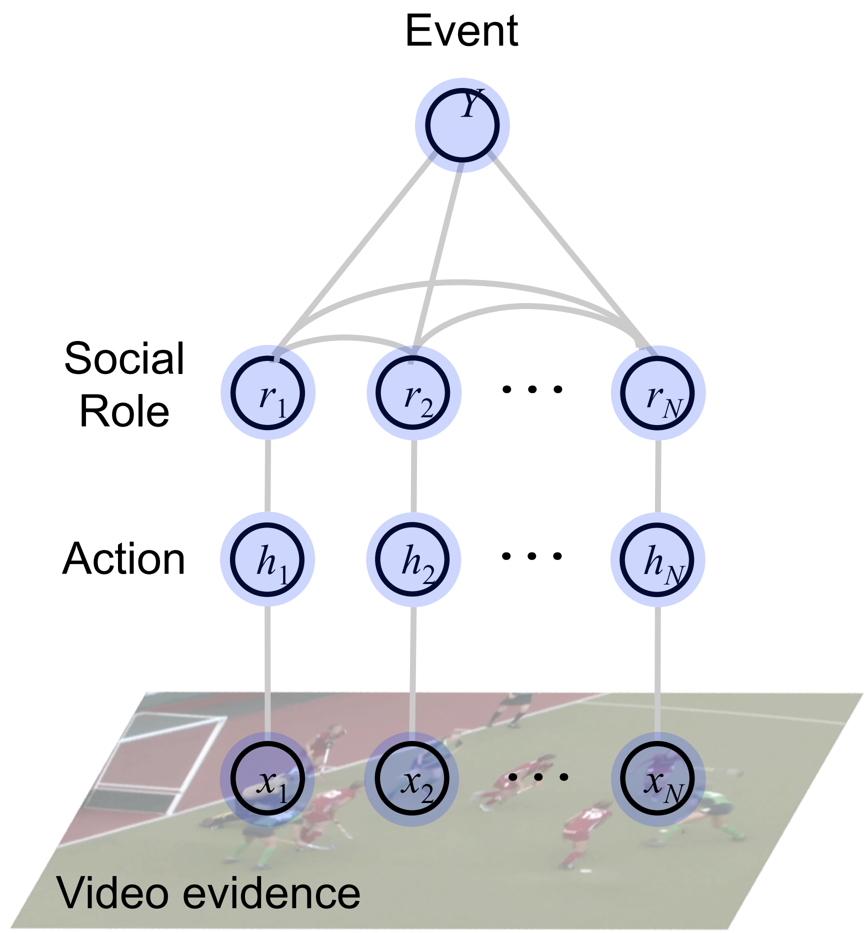

- Search for event/social role/action of interest
  – Who is the attacker? What's the overall game situation?

# System Overview

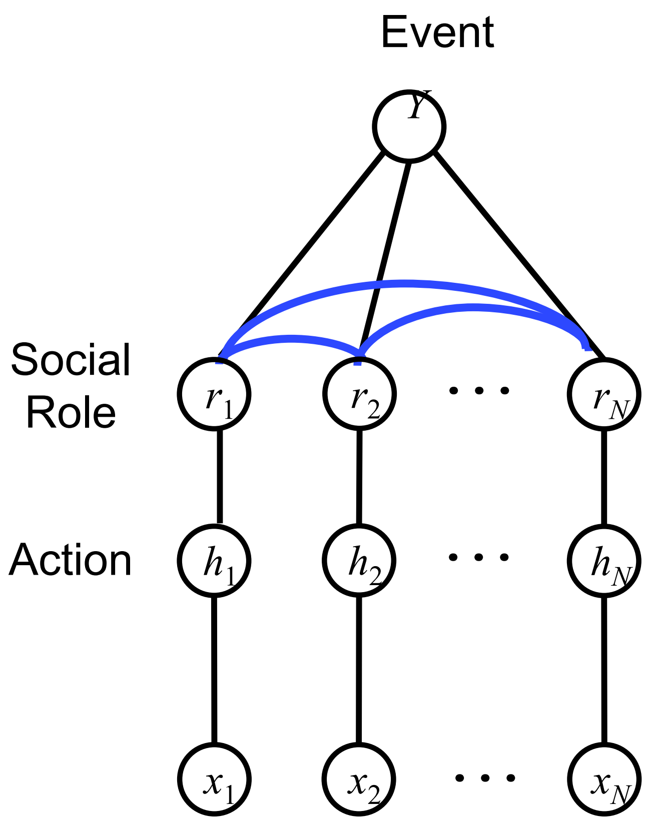


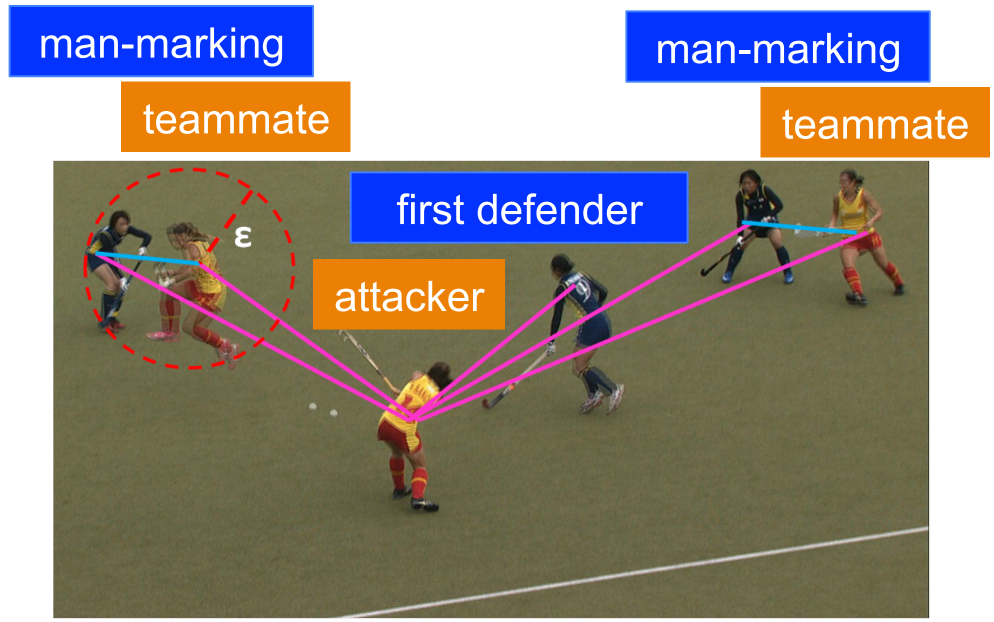
# Activity Hierarchy Model Representation



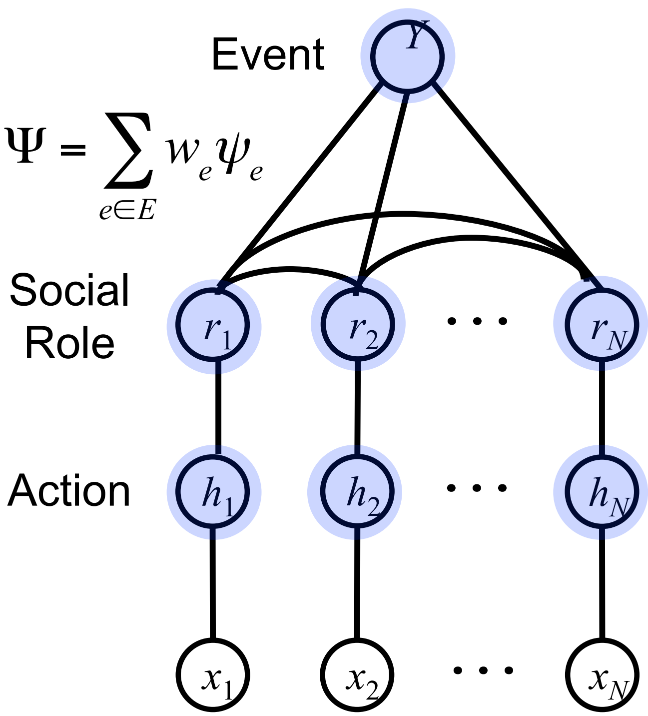
# Activity Hierarchy Model Representation




- Spatial relationships and color among players with different social roles.

# Model Learning




$$\Psi = \sum_{e \in E} w_e \psi_e$$

Query for event: $loss = \Delta(y, y_i)$

$$\Delta(y, y_i) = \begin{cases} 1 & \text{if } y \neq y_i \\ 0 & \text{otherwise} \end{cases}$$

Query for social roles: $loss = \Delta(r, r_i)$

Query for actions: $loss = \Delta(h, h_i)$

Scene labeling: $loss = \Delta(y, y_i) + \Delta(r, r_i) + \Delta(h, h_i)$

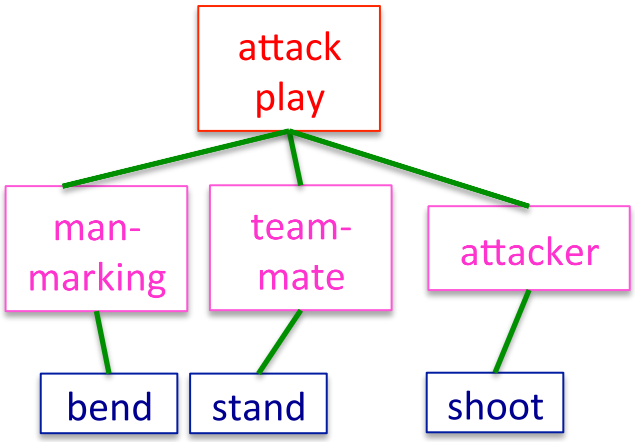
$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|_2^2 + \beta \sum_i \xi_i$$

$$\text{s.t. } \forall i, y, r, h$$

$$\mathbf{w}_{y_i r_i h_i} \cdot \psi_i - \mathbf{w}_{yrh} \cdot \psi_i \geq \boxed{loss} - \xi_i$$

$$\forall i, \xi_i \geq 0$$

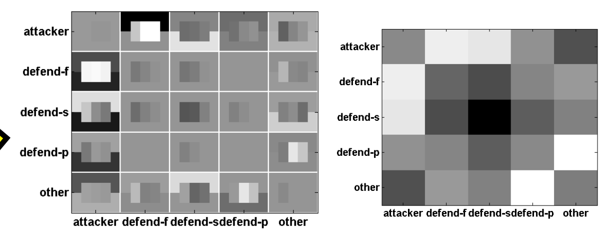
# Model Inference
# - Query

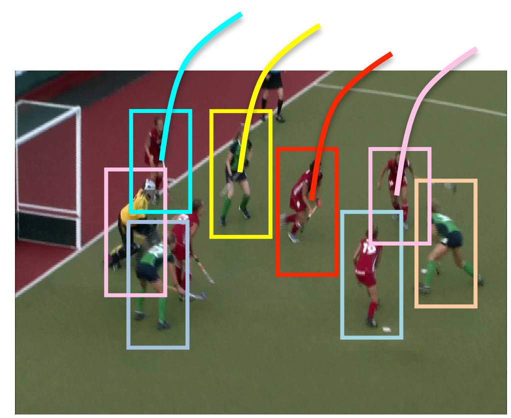



$V$

q: User-specified queries
– e.g. find the attack play

The learned models

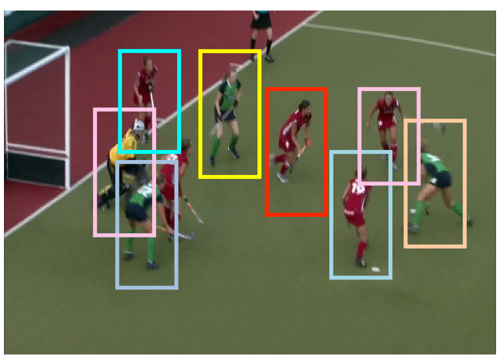


Person detection and tracking

**coordinate ascent inference**

$$\max_{y,r,h\backslash q} \sum_e w_e \psi_e$$



Score: $\Psi\left(Y^*, \left\{r^*_{1,n}\right\}_n, \left\{h^*_{1,n}\right\}_n, q\right)$

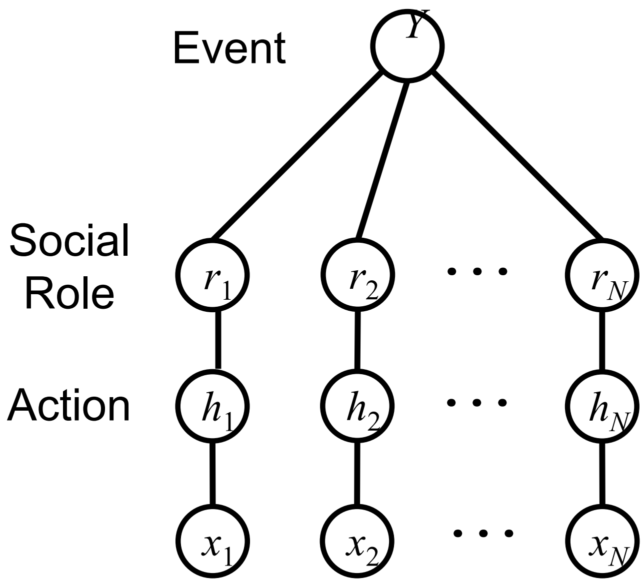
Event, social roles, actions, queries

# ESPN Broadcast Field Hockey Data

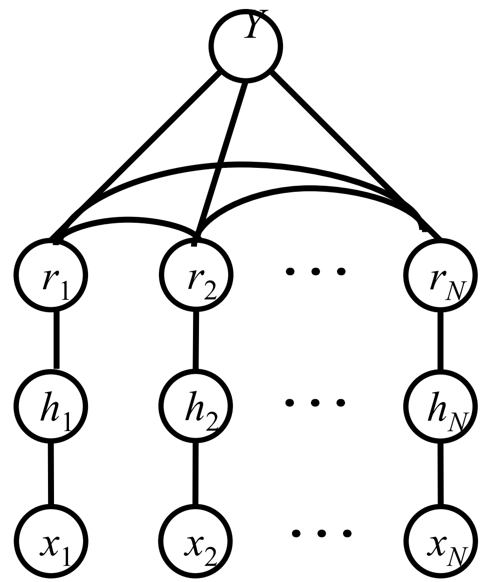

- 58 videos, 11 actions, 5 social roles, 3 scene-level events

# Results – Scene Labeling

Unary

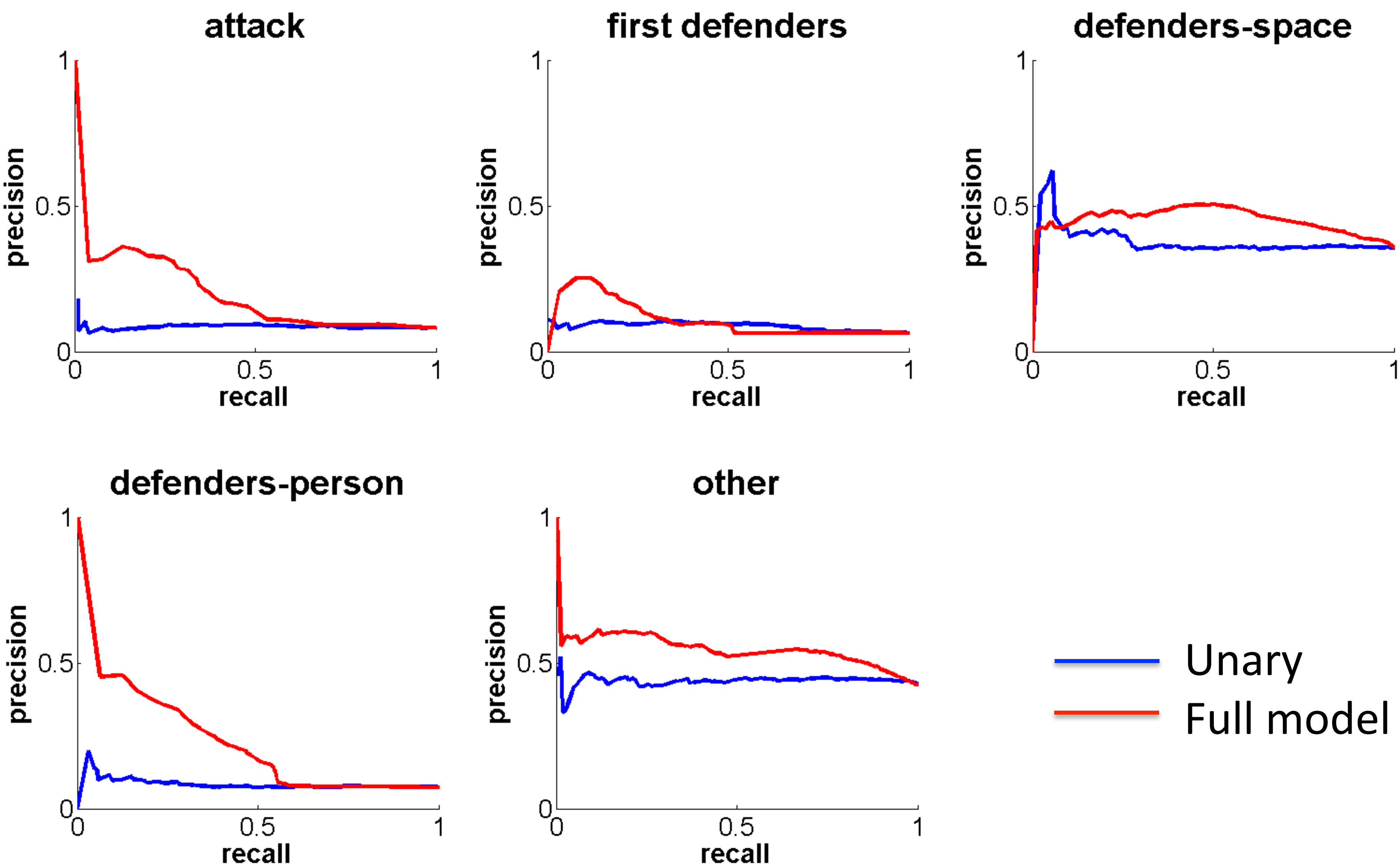


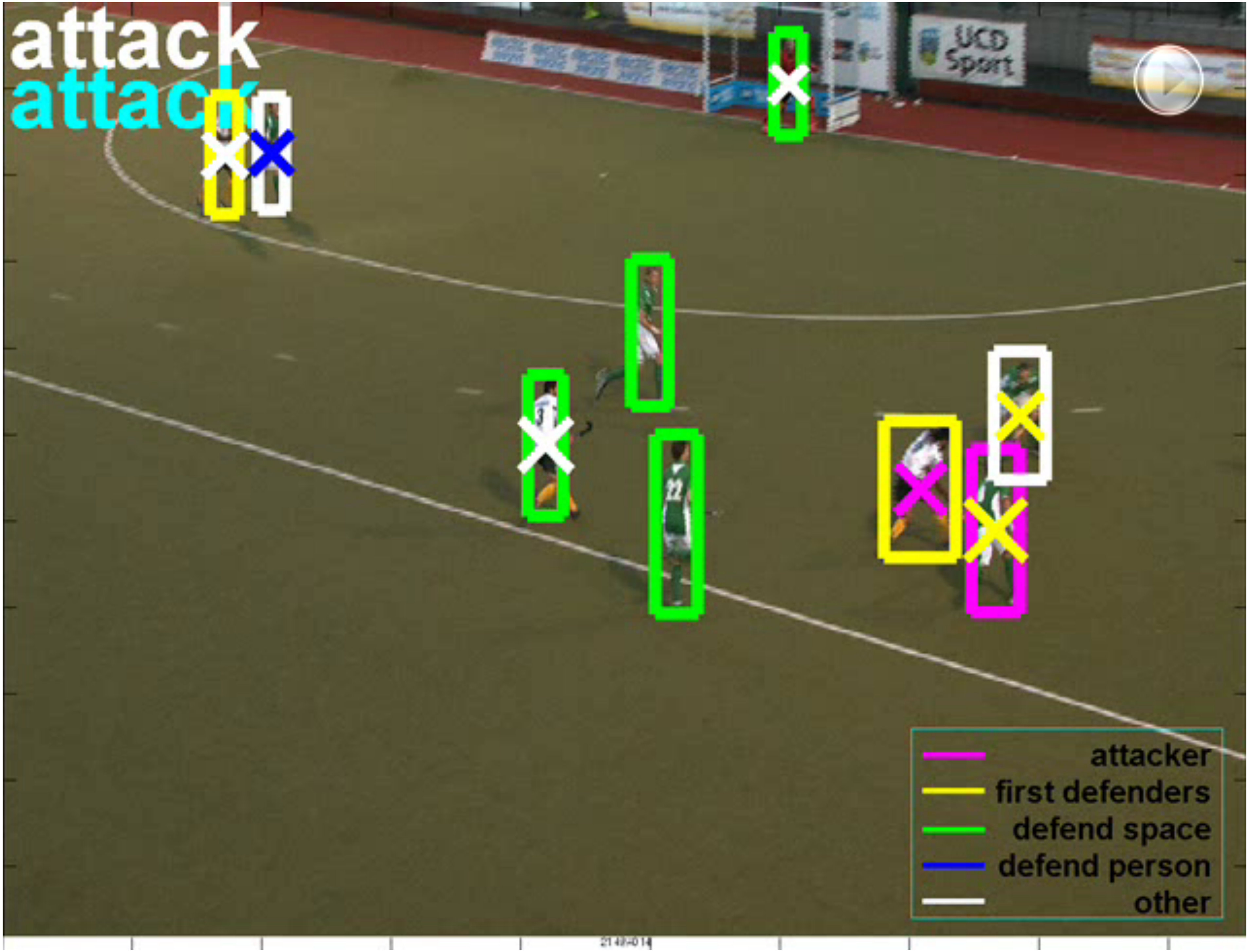

Full model




| Method | Action | Role | Event |
|---|---|---|---|
| unary | 21.5 | 21.7 | 56.9 |
| Full model | 28.8 | 44.0 | 62.8 |
| action model (HOG+SVM) | 26.1 | N/A | N/A |

# Results – Query for Social Roles



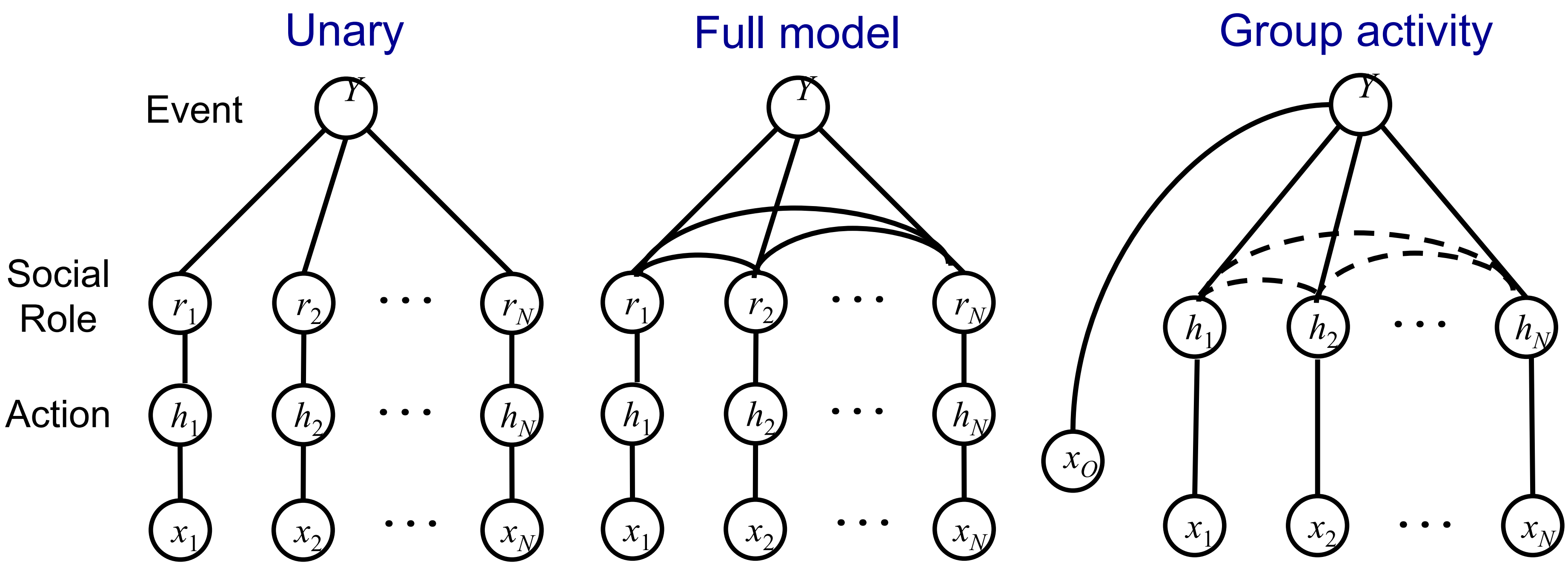


attack
attack
attacker
first defenders
defend space
defend person
other

# Nursing Home Data

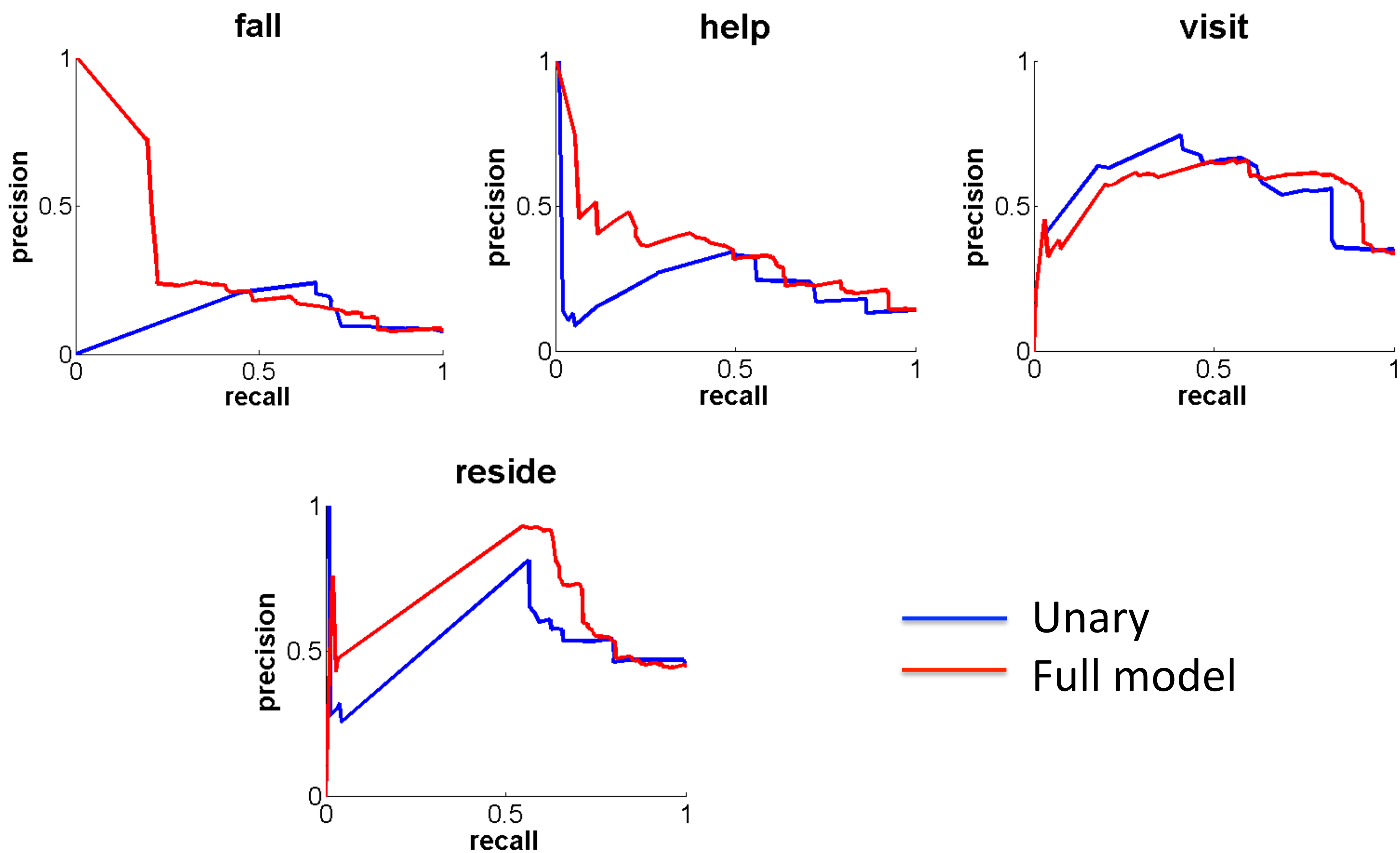

- 22 short clips of fall + a 30-min non-fall video sequence, 5fps, surveillance video
- 5 actions: walk, stand, sit, bend, and fall
- 4 social roles: fall, help, visit and reside
- 2 scene-level events: fall, non-fall

# Results – Scene Labeling (Nursing Home)



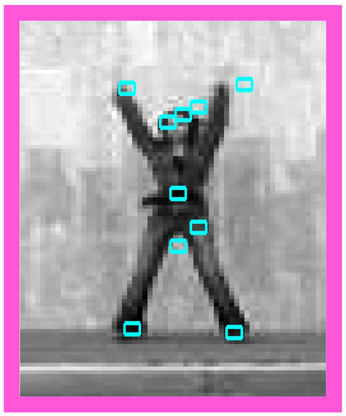

| Method | Action | Role | Event |
|---|---|---|---|
| Unary | 40.9 | 35.0 | 73.2 |
| Full model | 42.0 | 50.1 | 80.5 |
| Action model (HOG+SVM) | 38.7 | N/A | N/A |
| Group activity [Lan et al. PAMI 12] | N/A | N/A | 78.5 |

# Results – Query for Social Roles (Nursing Home)



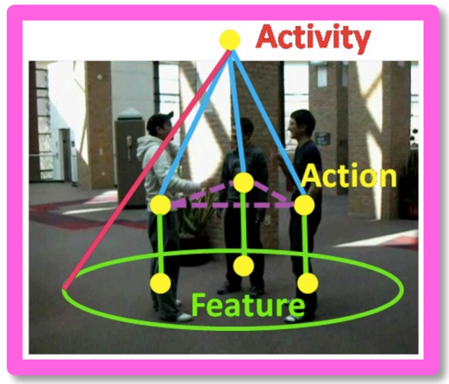
# Conclusion

action recognition



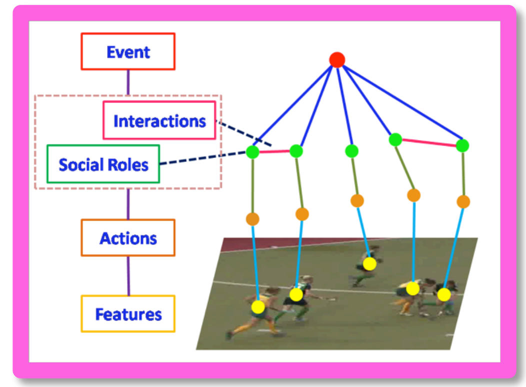
individual

group activity recognition




group

activity hierarchies




scene

Structural Recognition of Human Activities

# Acknowledgements

Tian Lan

SFU