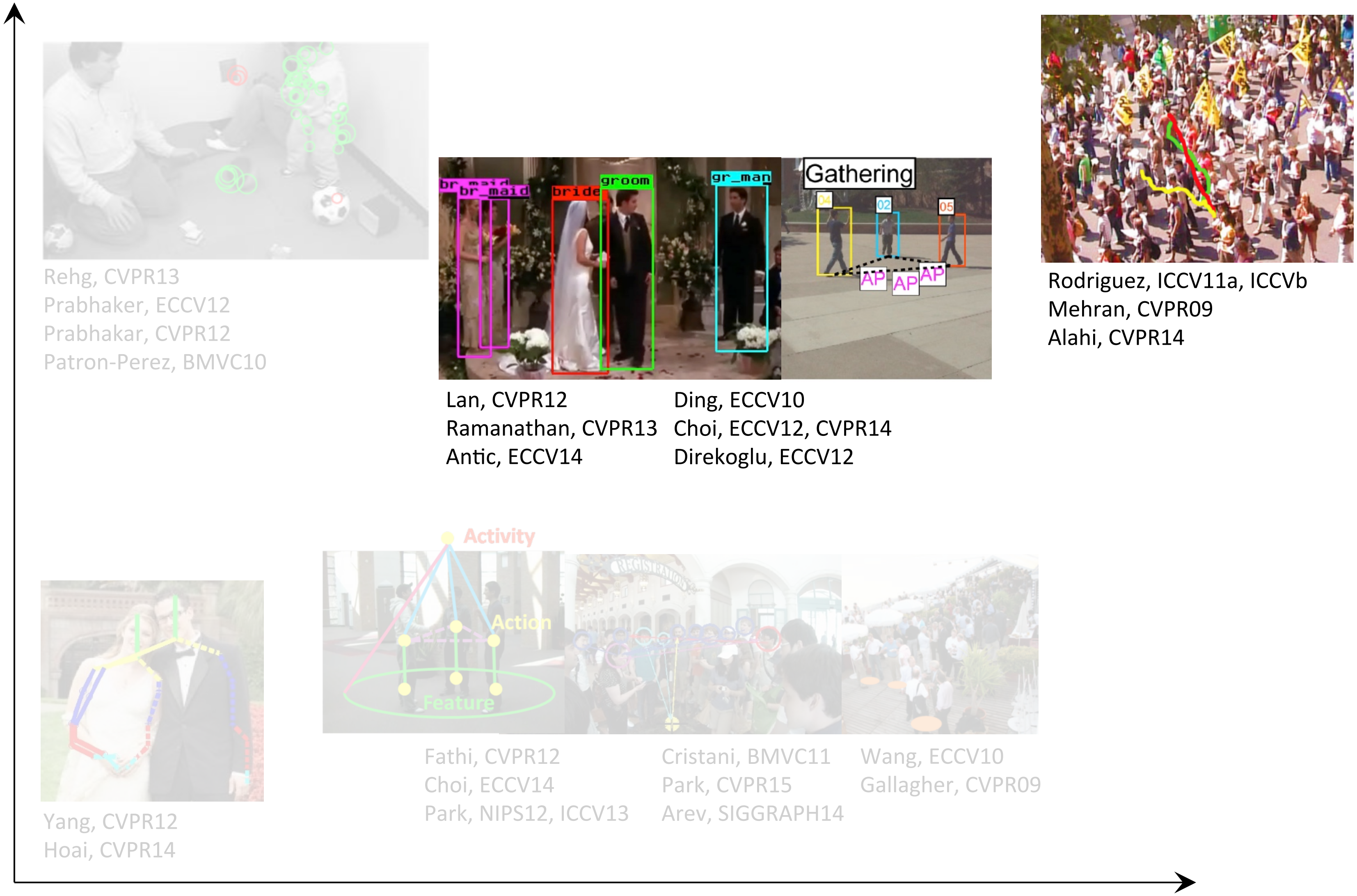


Dynamic scene

Static scene

Rehg, CVPR13
Prabhaker, ECCV12
Prabhakar, CVPR12
Patron-Perez, BMVC10

Lan, CVPR12
Ramanathan, CVPR13
Antic, ECCV14

Ding, ECCV10
Choi, ECCV12, CVPR14
Direkoglu, ECCV12

Rodriguez, ICCV11a, ICCVb
Mehran, CVPR09
Alahi, CVPR14

Yang, CVPR12
Hoai, CVPR14

Fathi, CVPR12
Choi, ECCV14
Park, NIPS12, ICCV13

Cristani, BMVC11
Park, CVPR15
Arev, SIGGRAPH14

Wang, ECCV10
Gallagher, CVPR09

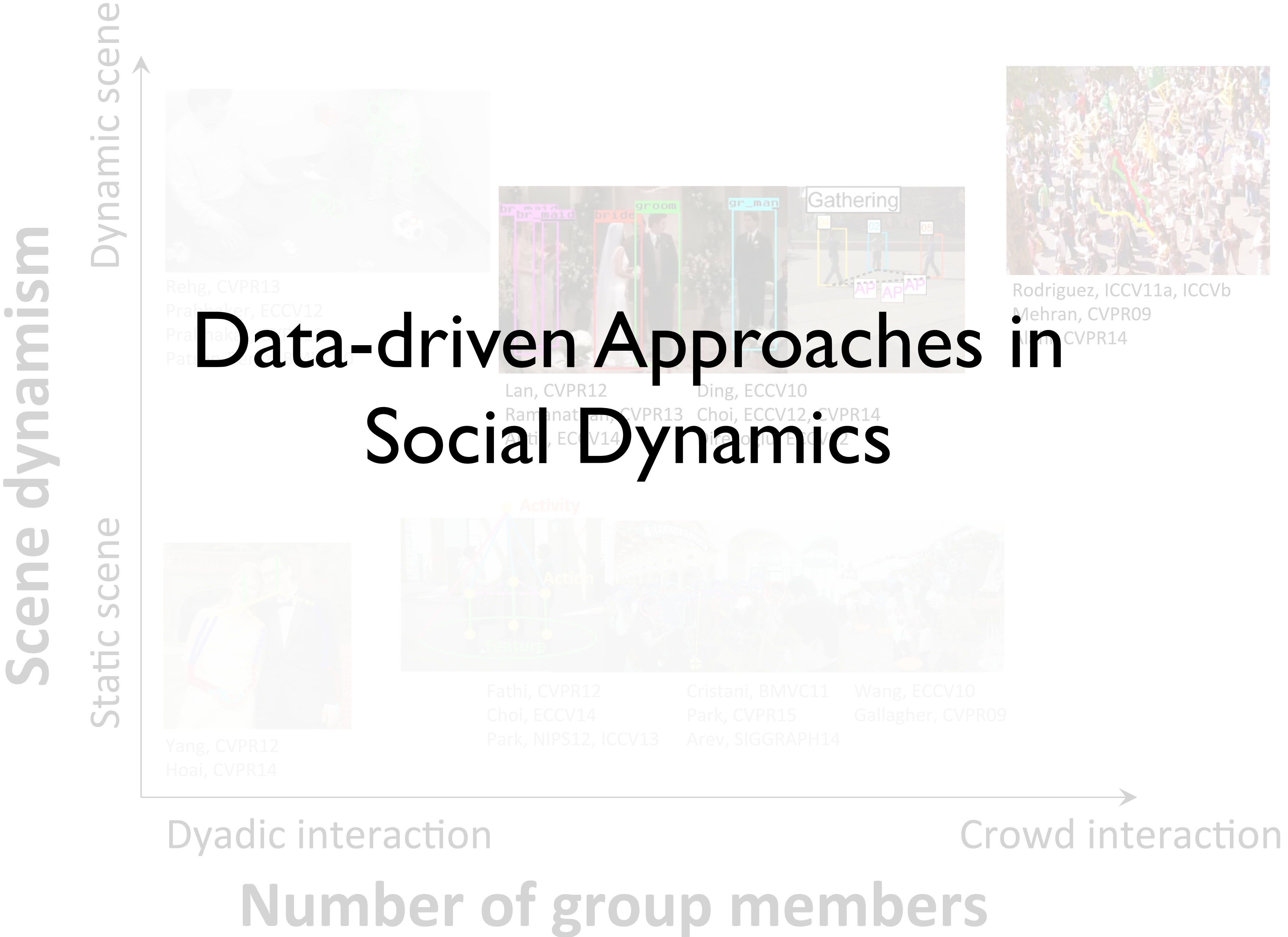
Dyadic interaction

Crowd interaction

Scene dynamism

Number of group members

Data-driven Approaches in Social Dynamics



Scene dynamism

Dynamic scene

Static scene

Rehg, CVPR13
Prabhaker, ECCV12
Prabhakar, CVPR12
Patron-Perez, BMVC10

Lan, CVPR12
Ramanathan, CVPR13
Antic, ECCV14

Ding, ECCV10
Choi, ECCV12, CVPR14
Direkoglu, ECCV12

Rodriguez, ICCV11a, ICCVb
Mehran, CVPR09
Alahi, CVPR14

Yang, CVPR12
Hoai, CVPR14

Fathi, CVPR12
Choi, ECCV14
Park, NIPS12, ICCV13

Cristani, BMVC11
Park, CVPR15
Arev, SIGGRAPH14

Wang, ECCV10
Gallagher, CVPR09

Dyadic interaction

Crowd interaction

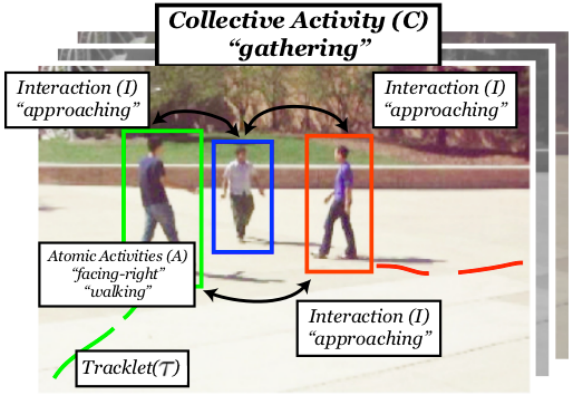
Number of group members

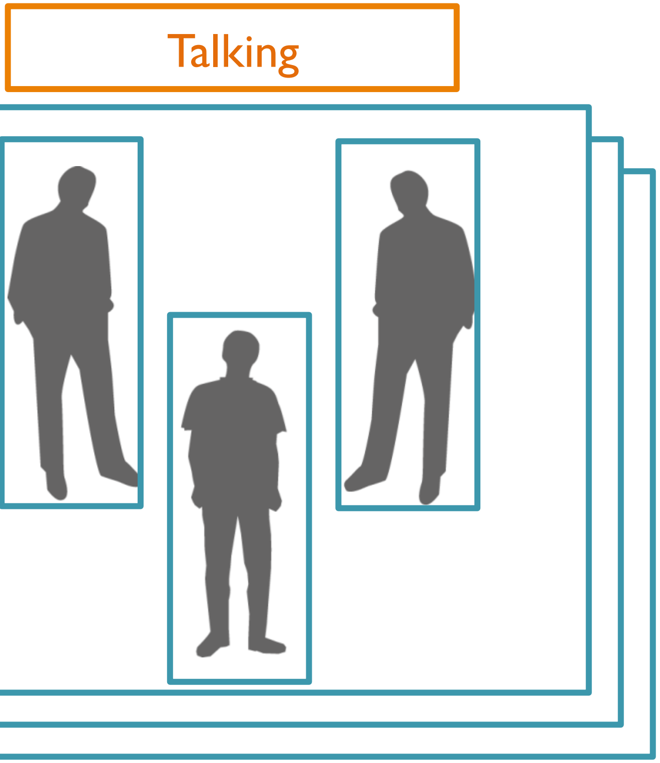
Group Activities in Videos

[Choi VSWS09, CVPR11, ECCV12, Lan WSGA09, NIPS10, Khamis CVPR12, ECCV12]

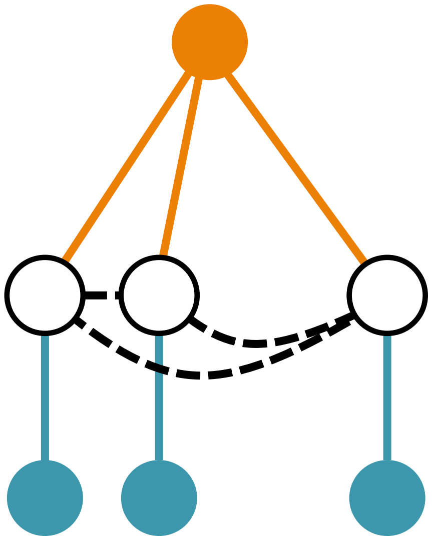
Input: video and box tracks

Output: group activity labels in time





Talking



group activities

interactions/
atomic actions

video/boxes

Collective Activities

Definition:

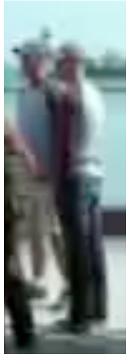
Activities that are defined or reinforced by the existence of a coherent behavior of a group of individuals in time and space.

Waiting

Queuing

Talking

Individual Appearance



- Individual appearance/motion does not provide an important social signal to recognize collective activities.

Crowd Context

- Contextual relationship among people is the key signal to recognize collective activities.

Challenges

- Large intra-class variation.
  - View point variation.
  - Number of group participants.
- Multiple groups in the scene.
- Activities changes over time.

Key Modules

- Crowd Context and the Representation.
  - Individual posture representation.
  - Encoding the context with posture representation.

- Exploit spatial-temporal correlations.
  - Utilize the structure in group activities.

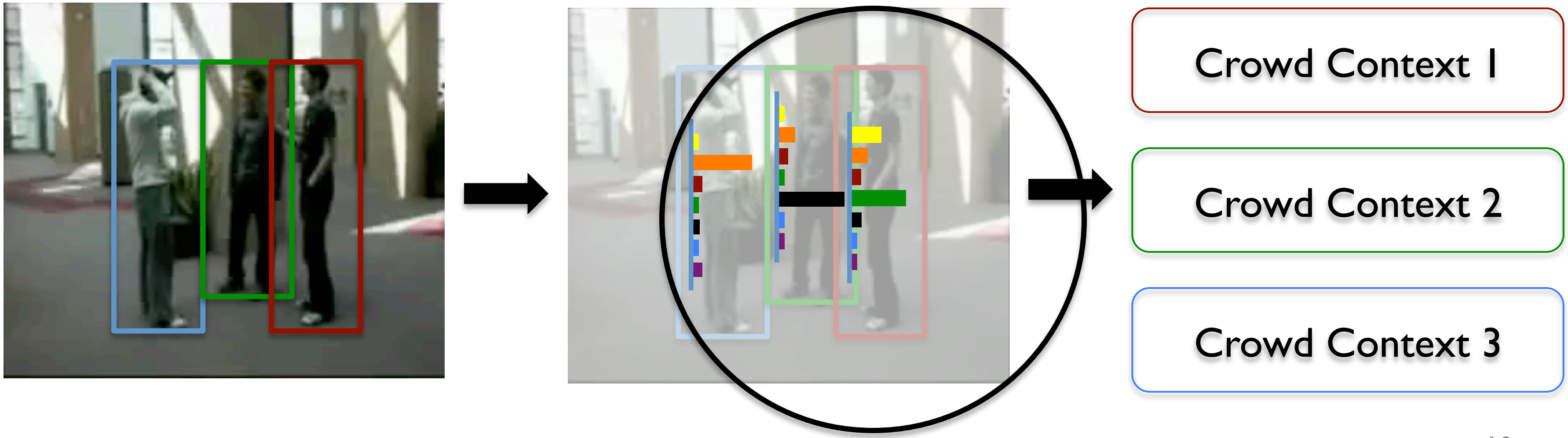
Key Modules

- Crowd Context and the Representation.
  - Individual posture representation.
  - Encoding the context with posture representation.

- Exploit spatial-temporal correlations.
  - Utilize the structure in group activities.

Crowd Context: How-to

- Given trajectories (boxes over time) of people.
- Represent each individual box with a posture description:
  - Combination of view point and velocity.
  - Finite set of activity labels.
  - A bag of mid-level discriminative parts.
- Encode the context using a spatio-temporal descriptor.



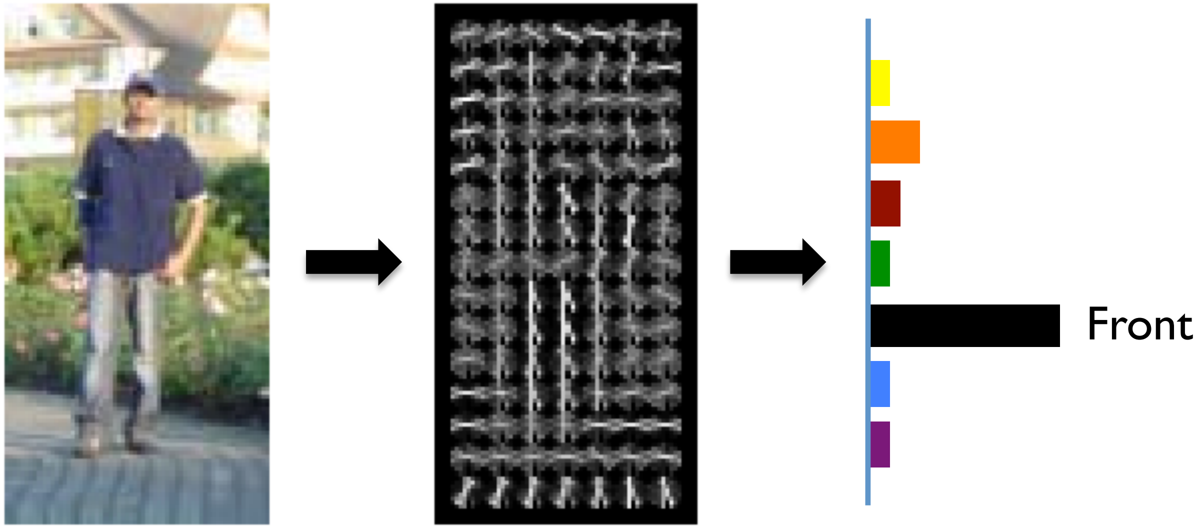
Crowd Context 1

Crowd Context 2

Crowd Context 3

Individual Posture Representation

- Extract a feature (e.g. HoG) in a bounding box.
- Represent the box in a posture space (e.g. SVM).



Dalal and Triggs, CVPR 2005

Simple and well-studied.

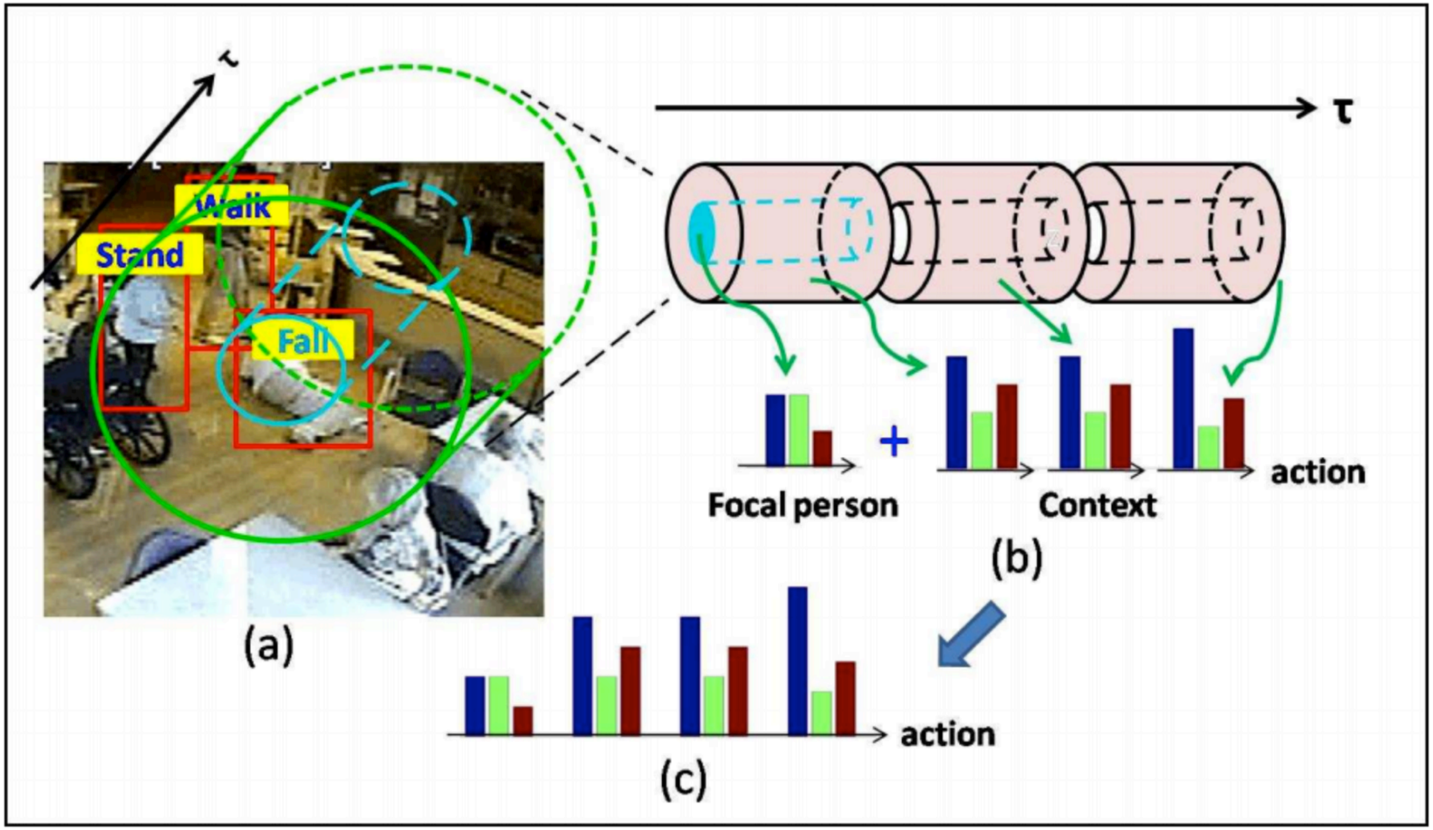
Require the definition and annotation of the posture space.

Encoding the Context

- Given a person of interest (anchor), aggregate the posture information of the others around the anchor person.
- Common ideas:
  - Define spatio-temporal support regions.
  - Pull the features in the space.

[Lan WSGA09]

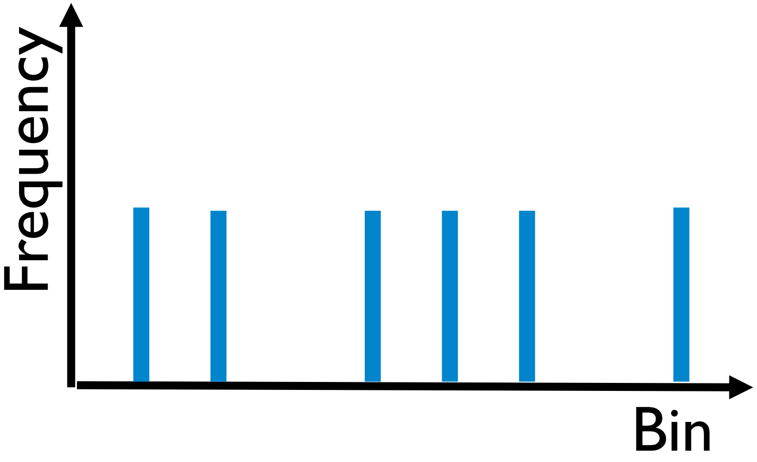
Encoding Context: Action Context

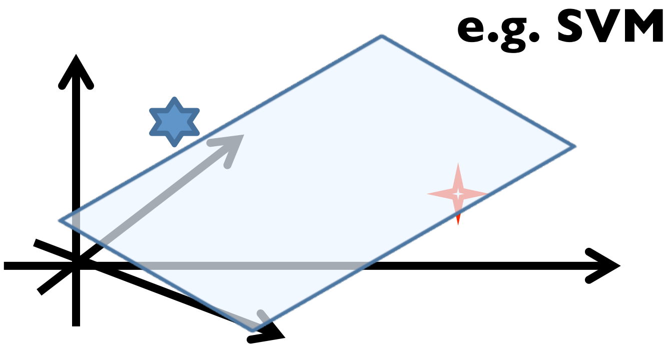


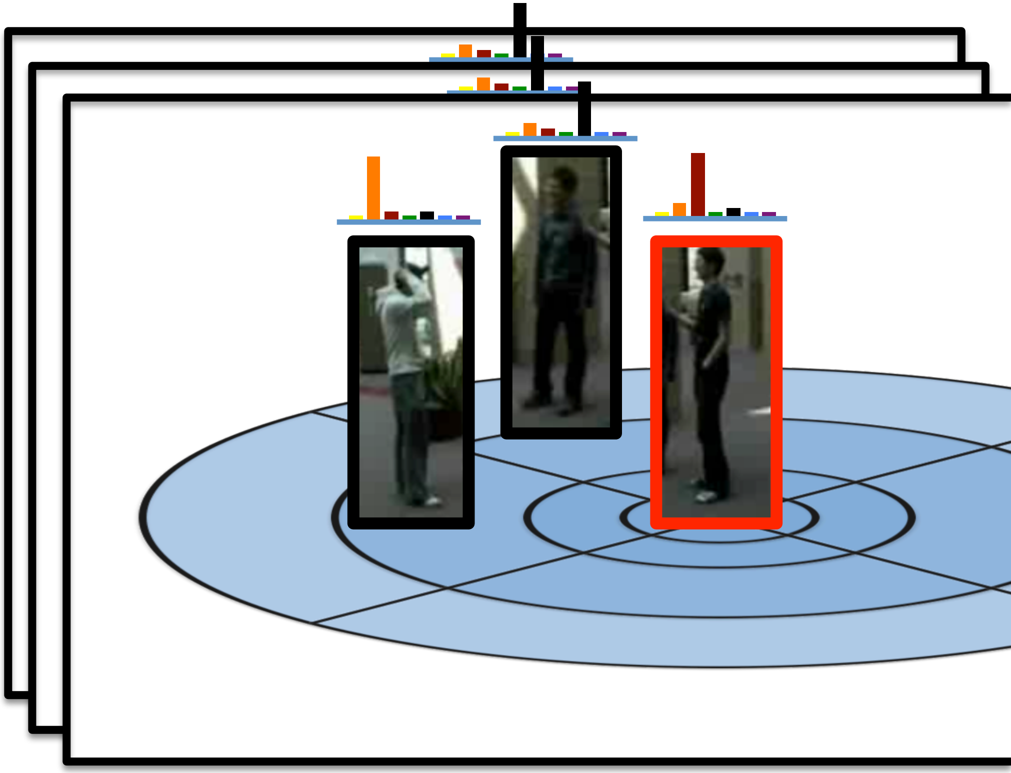
$$C_i = \left[\max_{j \in \mathcal{N}_1(i)} S_{1j}, \ldots, \max_{j \in \mathcal{N}_1(i)} S_{Kj}, \ldots, \max_{j \in \mathcal{N}_M(i)} S_{1j}, \ldots, \max_{j \in \mathcal{N}_M(i)} S_{Kj} \right]$$

[Choi VSWS09]

Encoding Context: Spatio-Temporal Local (STL) Descriptor

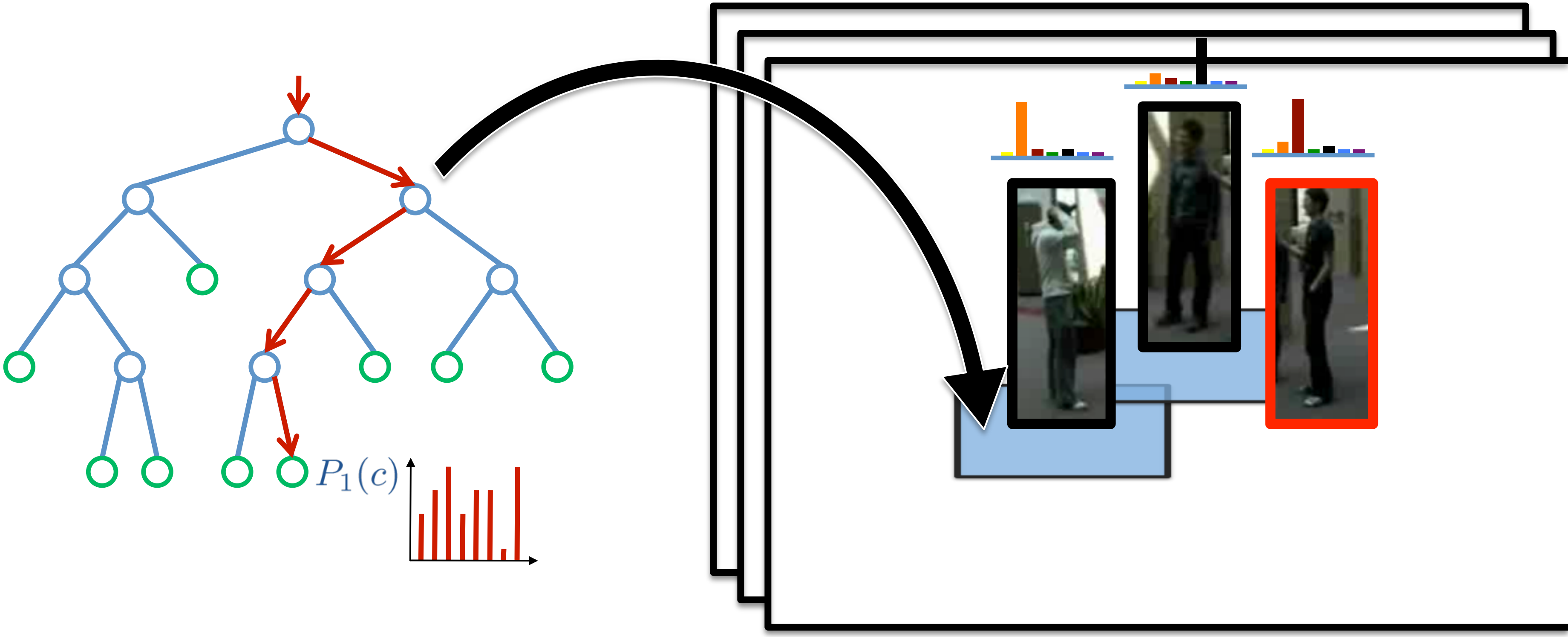






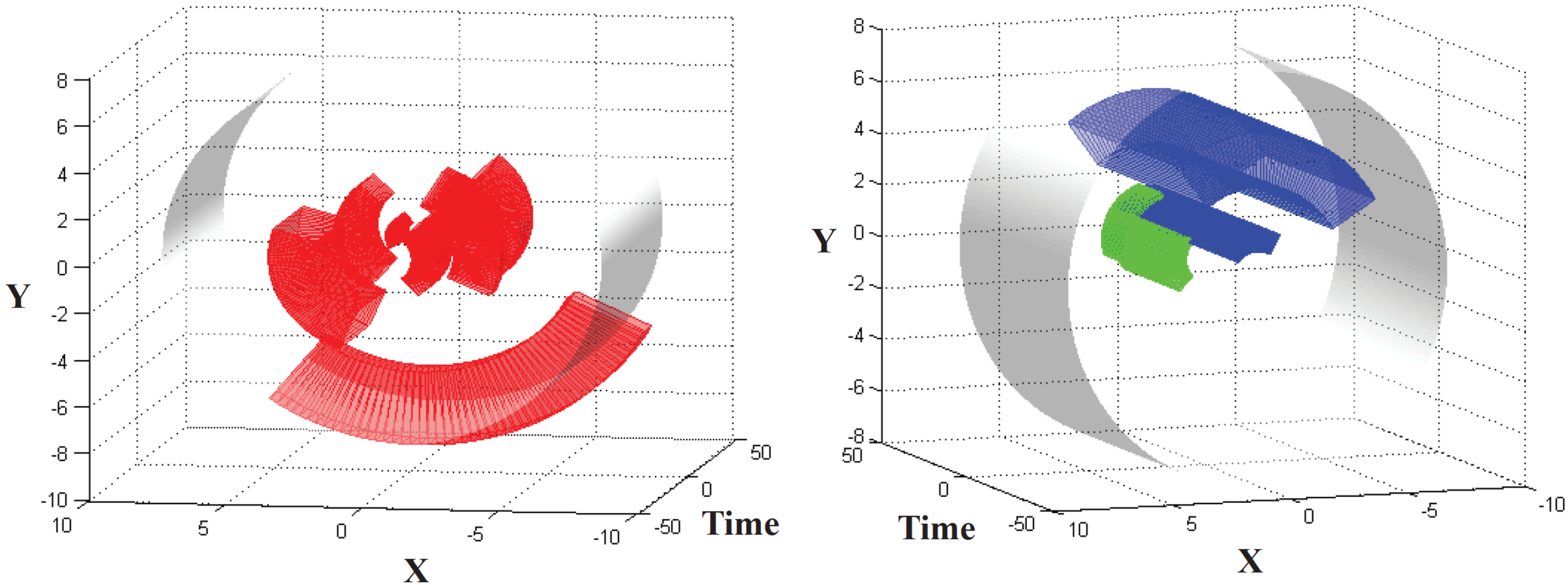
[Choi CVPR11]

Encoding Context: Learning the Contextual Model



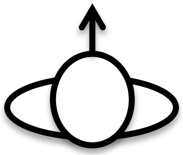
[Choi CVPR11]

Learning the Contextual Model



(a) Waiting

(b) Talking



Anchors are looking upward.

Red: Facing forward,
Blue: Facing down,
Green: Facing right

[Choi CVPR11]

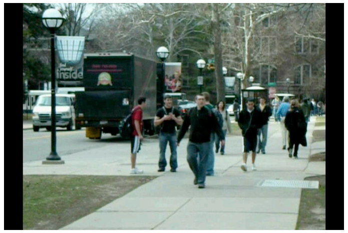
Collective Activity Dataset

Crossing

Waiting

Queuing

Walking



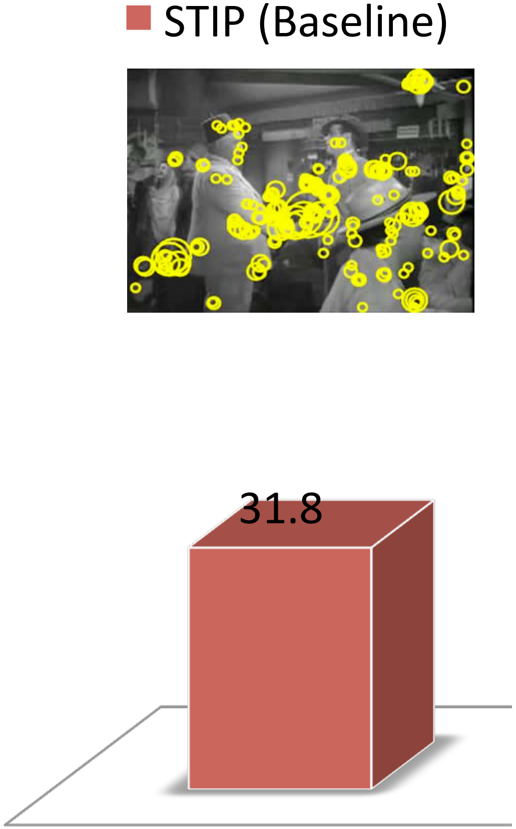
Talking

- 44 videos with multiple people.
- Crossing, Waiting, Queuing, Walking, Talking.
- Leave-One-Video-Out.

[Choi CVPR11]

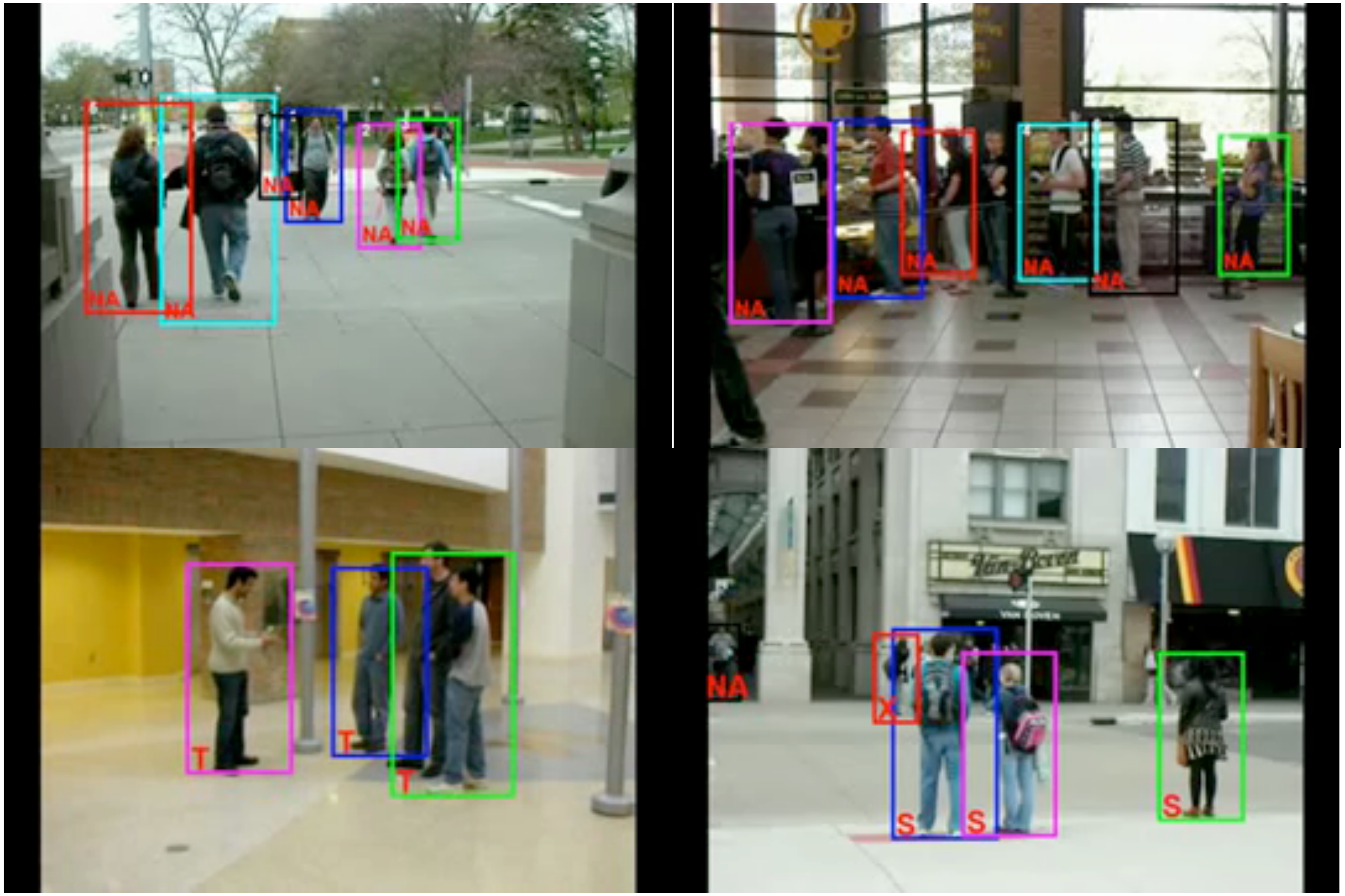
Atomic Activity Feature v.s. Crowd Context

Classification Accuracy



[Choi CVPR11]

Qualitative Examples



X: Crossing, S: Waiting, Q: Queuing, W: Walking, T: Talking

Key Modules

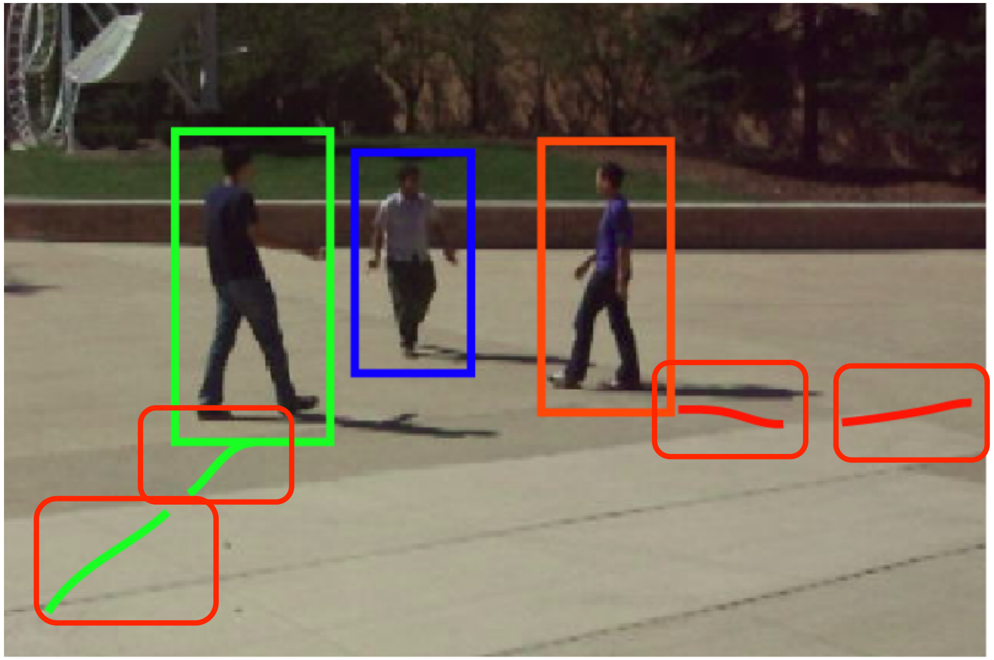
- Crowd Context and the Representation.
  - Individual posture representation.
  - Encoding the context with posture representation.

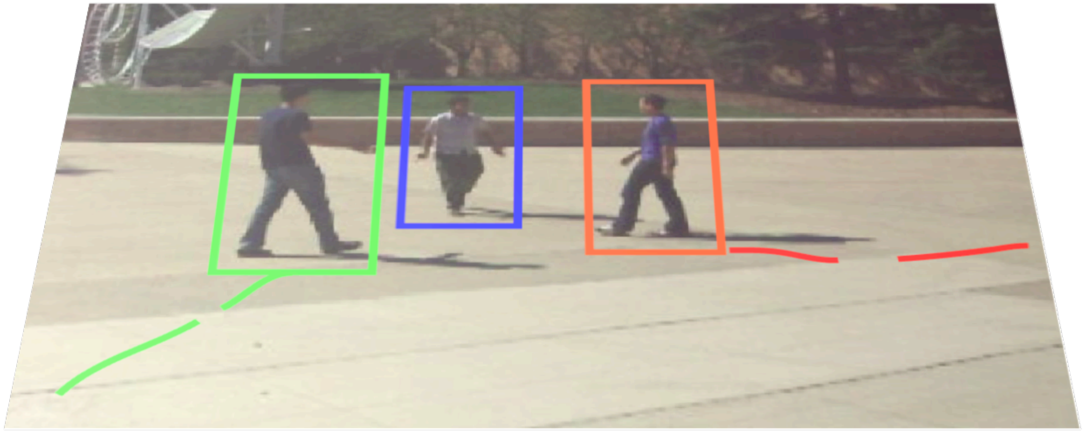
- Exploit spatial-temporal correlations.
  - Utilize the structure in group activities.

[Choi ECCV12]

Hierarchical Activity Model

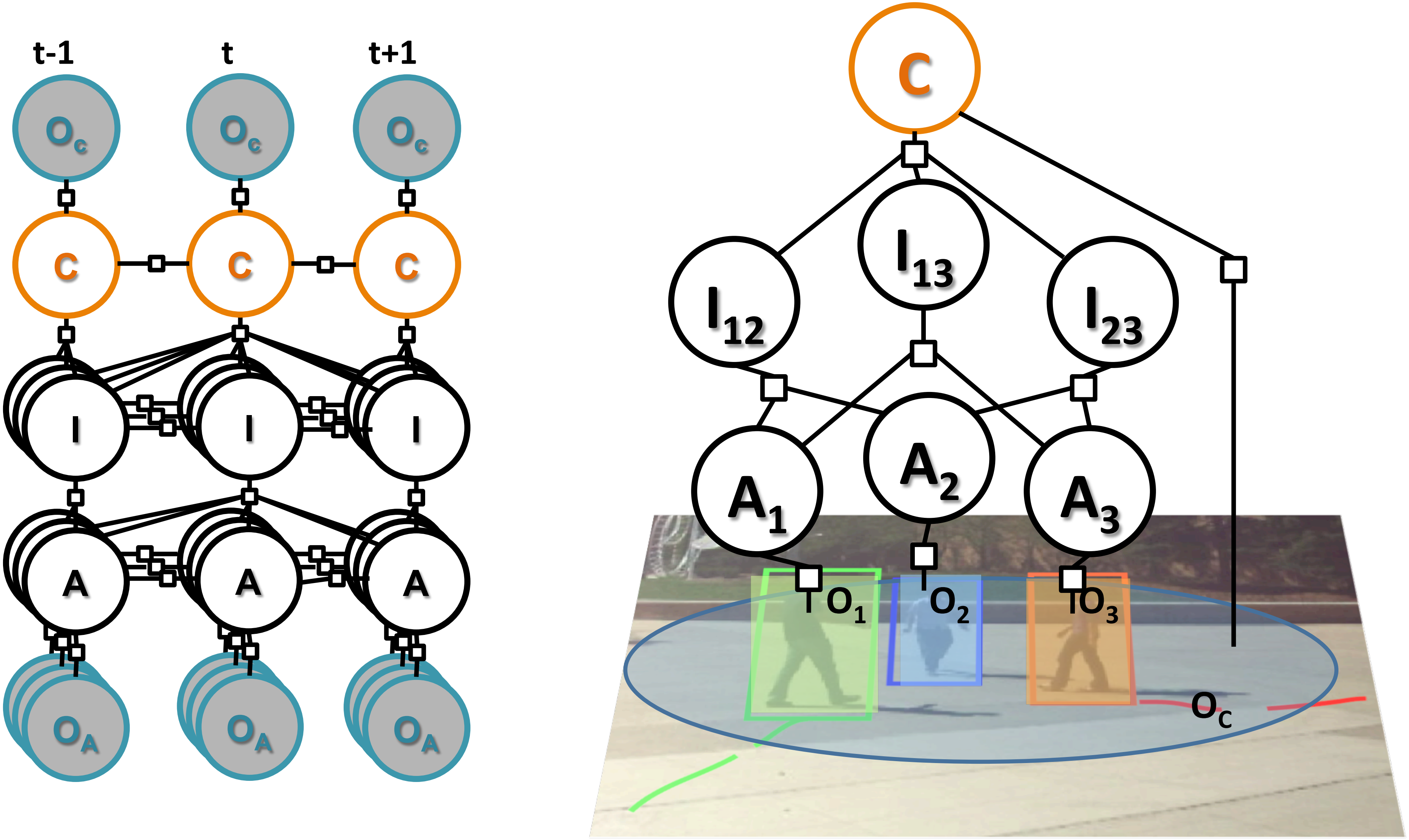
Input: video with tracklets



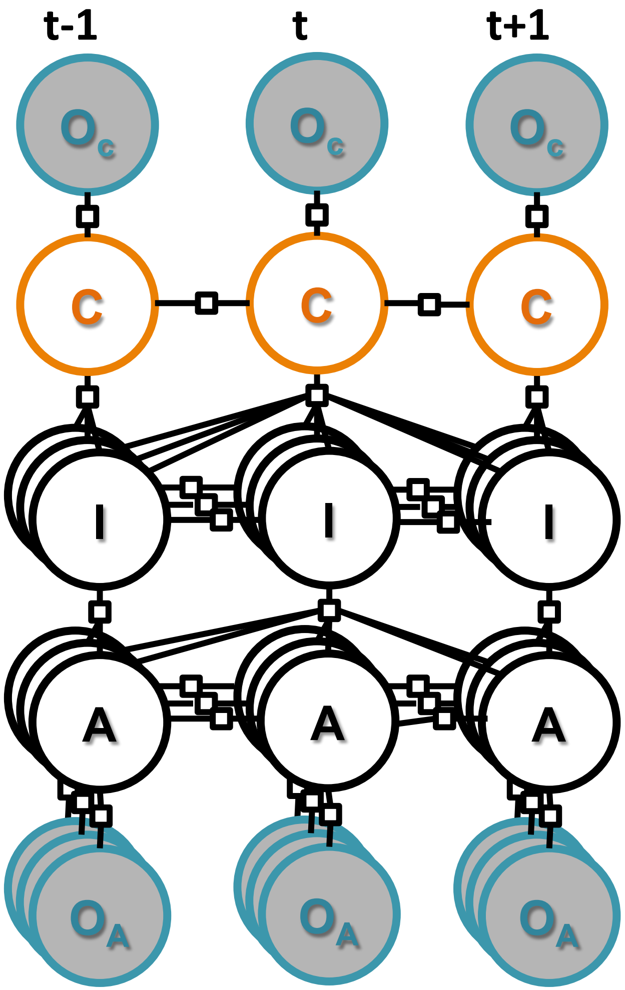


[Choi ECCV12]

Hierarchical Activity Model

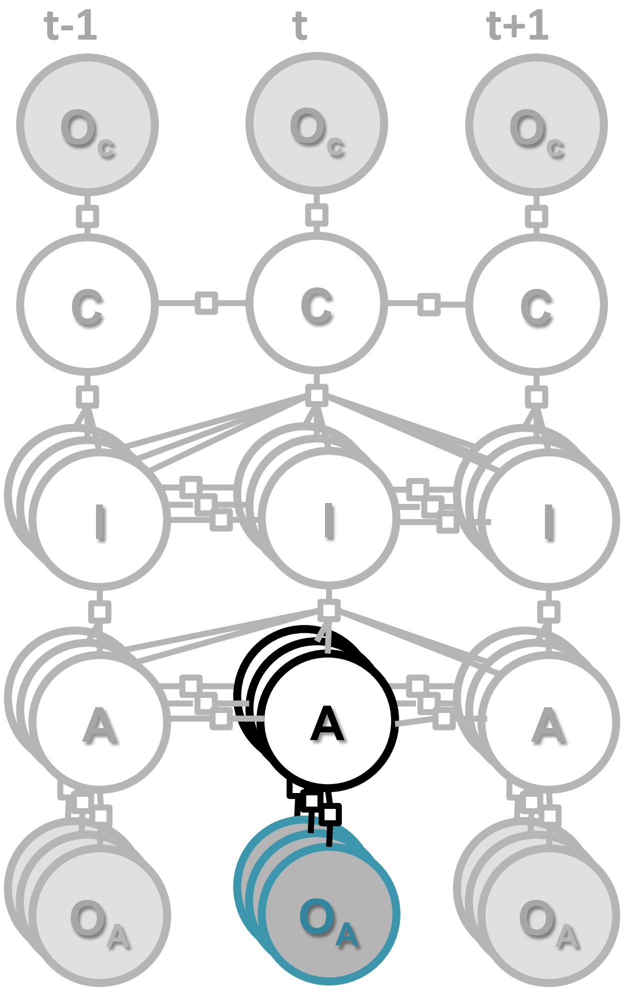


Hierarchical Activity Model



$$\Psi(C,I,A,O,f) = \\ \Psi(A,O)+\Psi(I,A,f)+\Psi(C,I)+\Psi(C,O)+ \\ \Psi(C)+\Psi(I)+\Psi(A)-c^Tf,\ f\in S$$

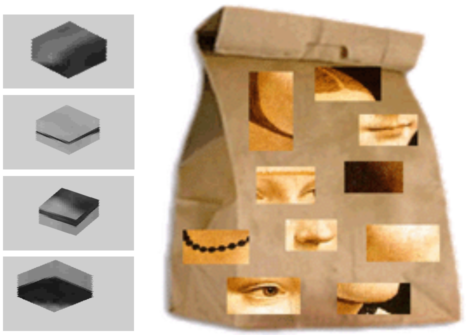
Atomic-Observation Potential



$$\Psi(C,I,A,O,f) = \underline{\Psi(A,O)} + \Psi(I,A,f) + \Psi(C,I) + \Psi(C,O) + \Psi(C) + \Psi(I) + \Psi(A) - c^T f,\ f \in S$$

Atomic Activity Models

- Action: BoW with STIP
- Pose: HoG

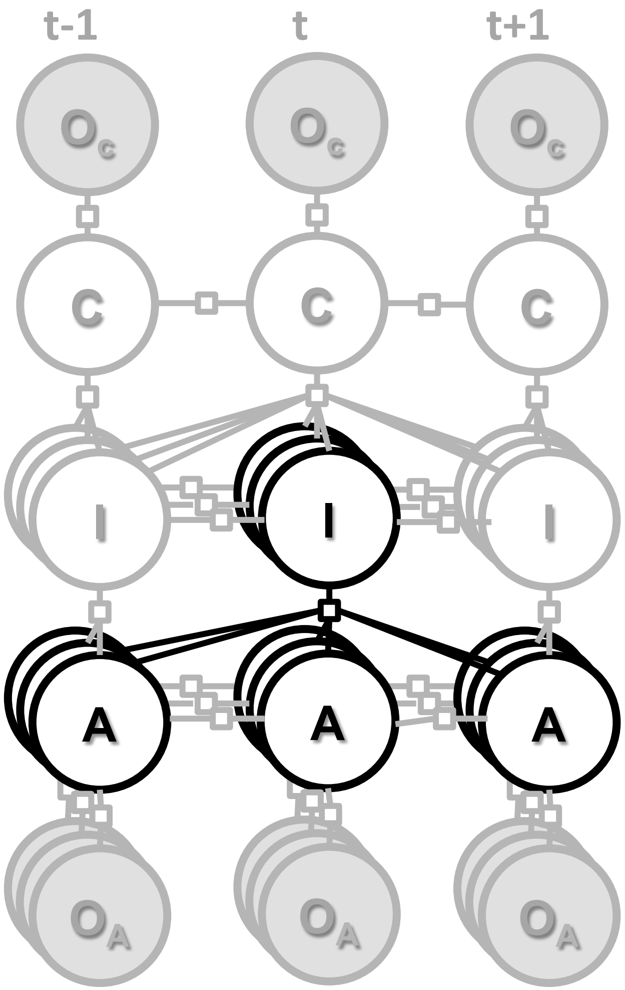


Dollar et al, 06; Niebles et al, 07

Dalal and Triggs, 05

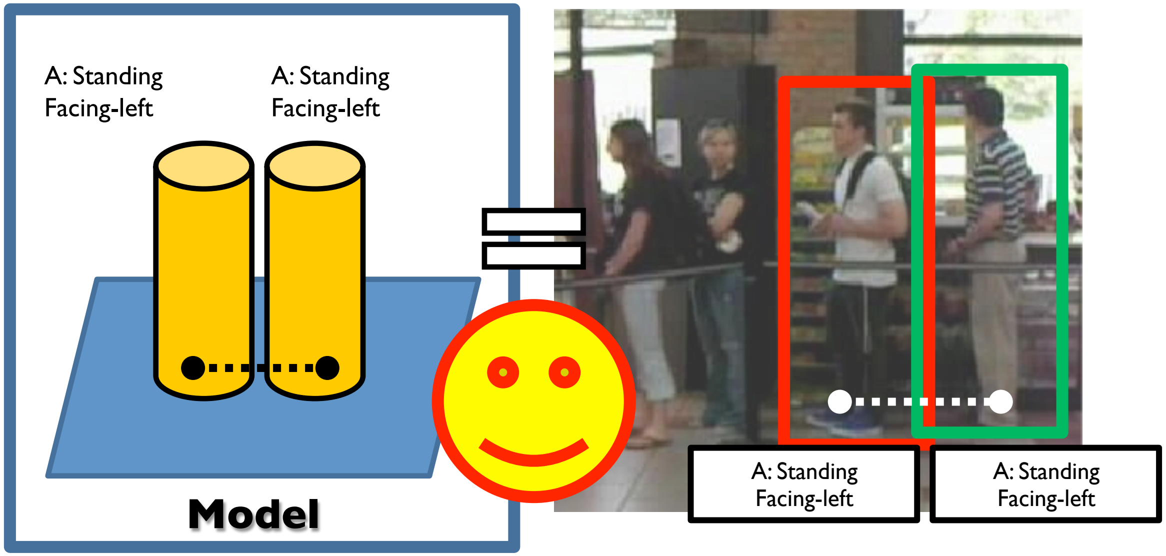
[Choi ECCV12]

Interaction-Atomic Potential

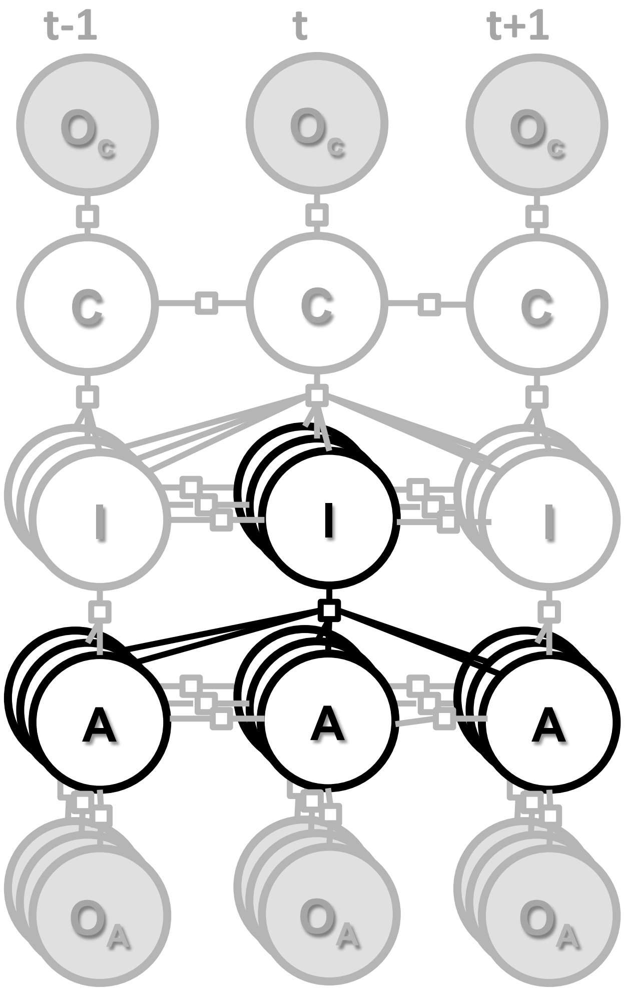


$$\begin{aligned}\Psi(C,I,A,O,f) = \\ &\Psi(A,O)+\underline{\Psi(I,A,f)}+\Psi(C,I)+\Psi(C,O)+ \\ &\Psi(C)+\Psi(I)+\Psi(A)-c^Tf,\ f\in S\end{aligned}$$

I: Standing-in-a-line

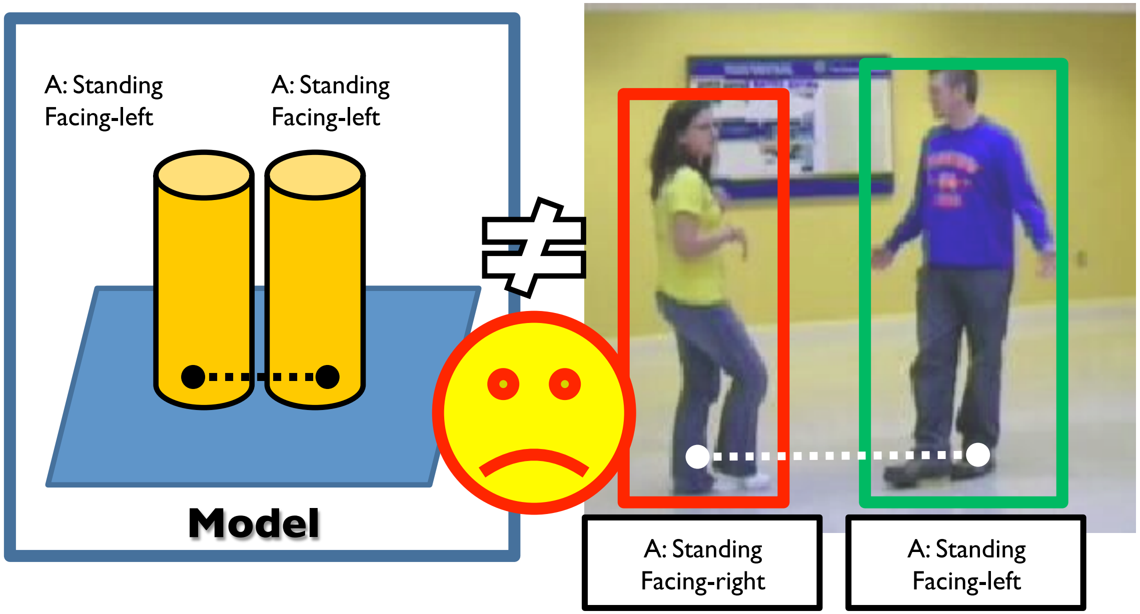


Interaction-Atomic Potential



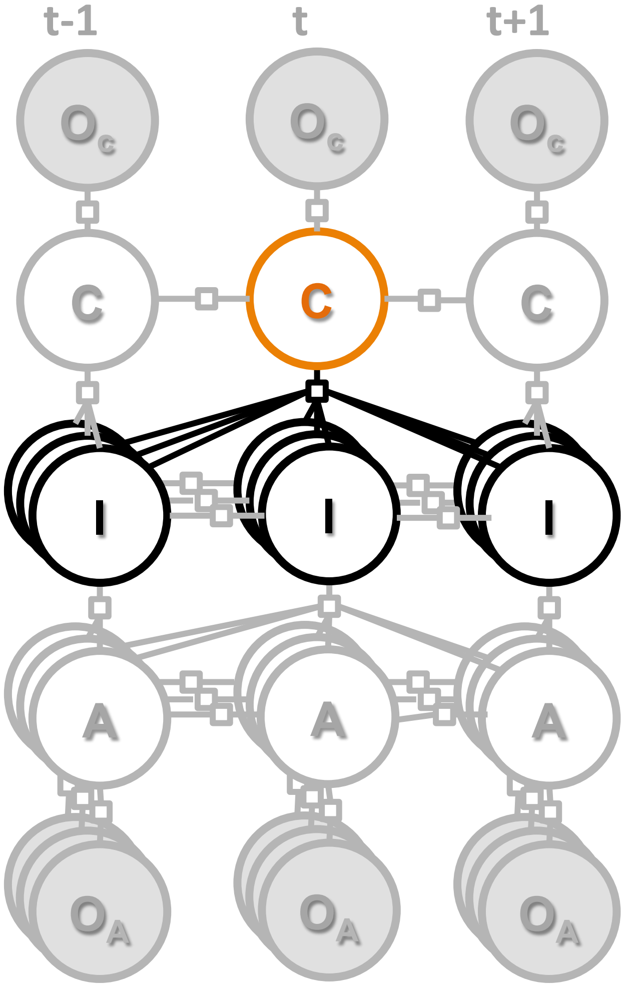
$$\Psi(C,I,A,O,f) = \Psi(A,O) + \underline{\Psi(I,A,f)} + \Psi(C,I) + \Psi(C,O) + \Psi(C) + \Psi(I) + \Psi(A) - c^T f,\ f \in S$$

I: Standing-in-a-line



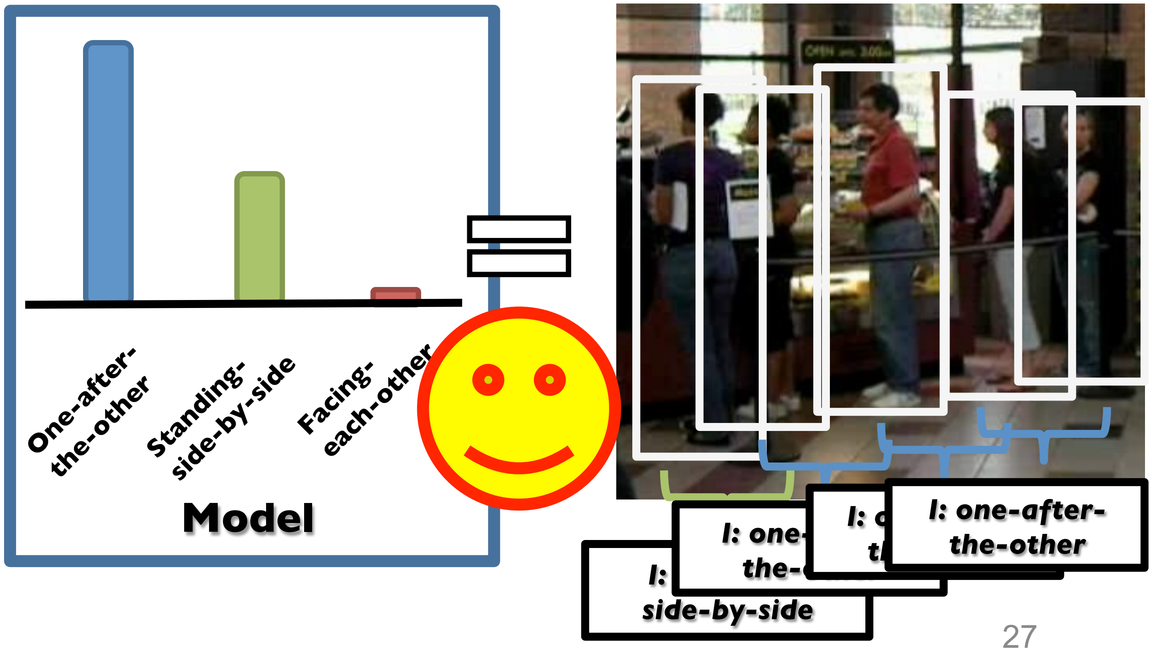
[Choi ECCV12]

Collective-Interaction Potential



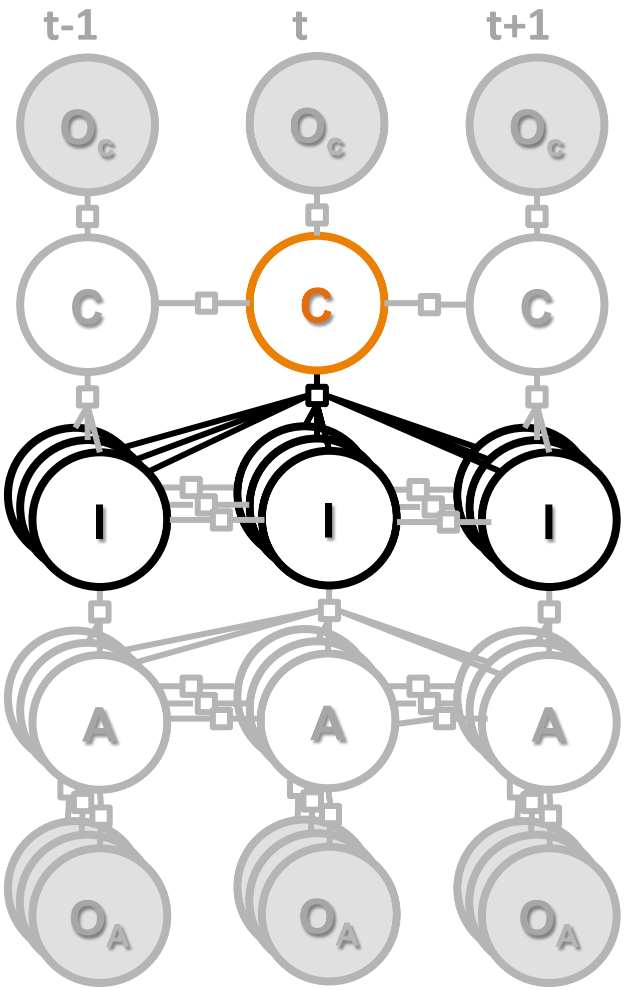
$$
\begin{aligned}
\Psi(C,I,A,O,f) =\ & \Psi(A,O)+\Psi(I,A,f)+\underline{\Psi(C,I)}+\Psi(C,O)+ \\
& \Psi(C)+\Psi(I)+\Psi(A)-c^T f,\ f \in S
\end{aligned}
$$

C: Queuing



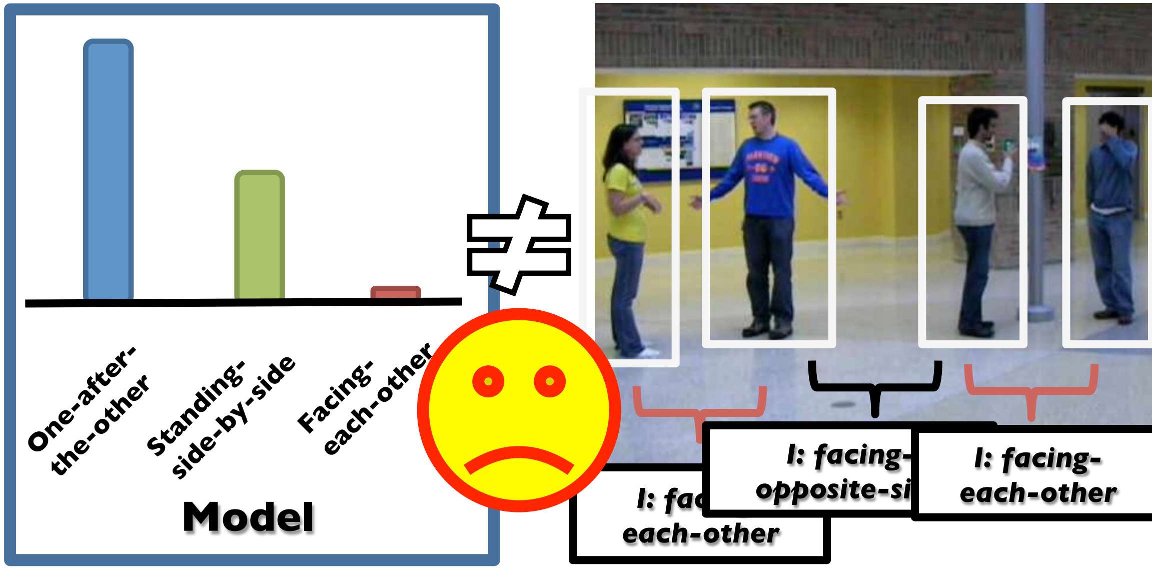
[Choi ECCV12]

Collective-Interaction Potential



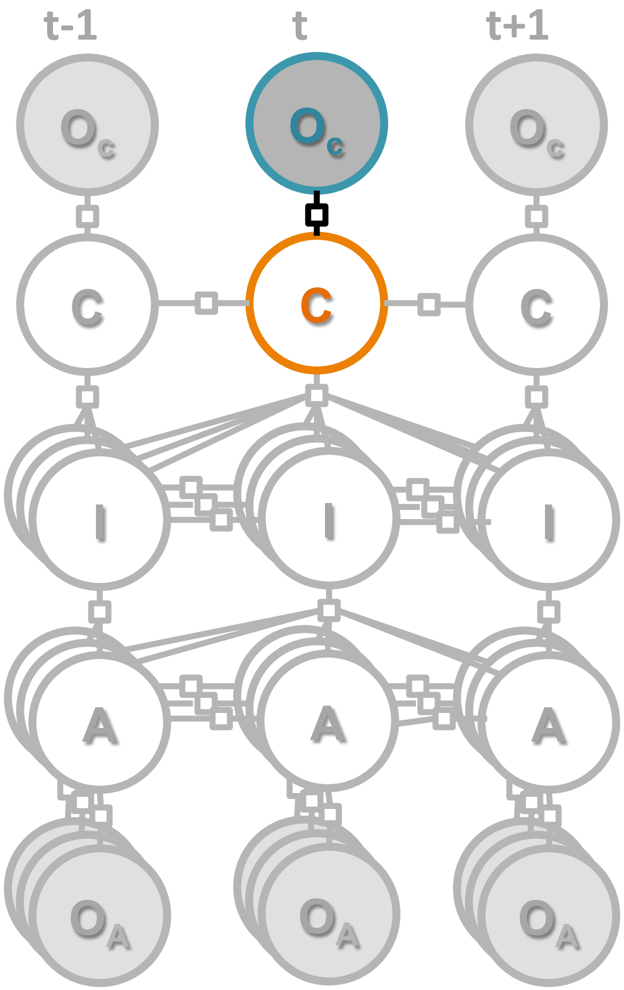
$$\begin{aligned}\Psi(C,I,A,O,f) = \\ &\Psi(A,O)+\Psi(I,A,f)+\underline{\Psi(C,I)}+\Psi(C,O)+ \\ &\Psi(C)+\Psi(I)+\Psi(A)-c^T f,\ f\in S\end{aligned}$$

C: Queuing



[Choi ECCV12]

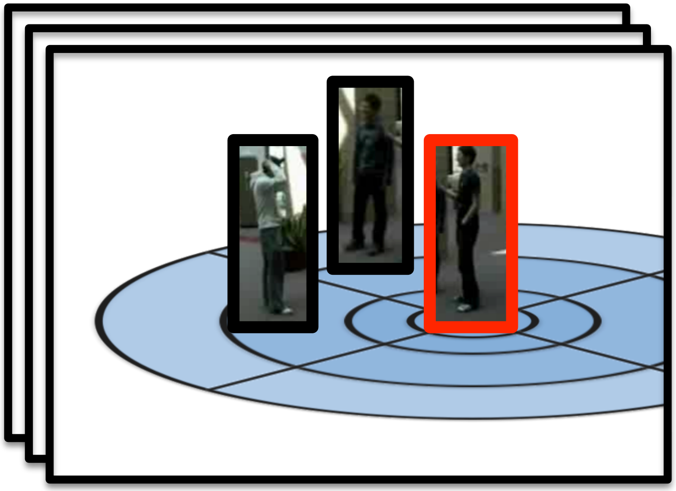
Collective-Observation Potential



$$\begin{aligned}\Psi(C,I,A,O,f) = \\ &\Psi(A,O)+\Psi(I,A,f)+\Psi(C,I)+\underline{\Psi(C,O)+} \\ &\Psi(C)+\Psi(I)+\Psi(A)-c^{T}f,\ f\in S\end{aligned}$$

Collective Activity

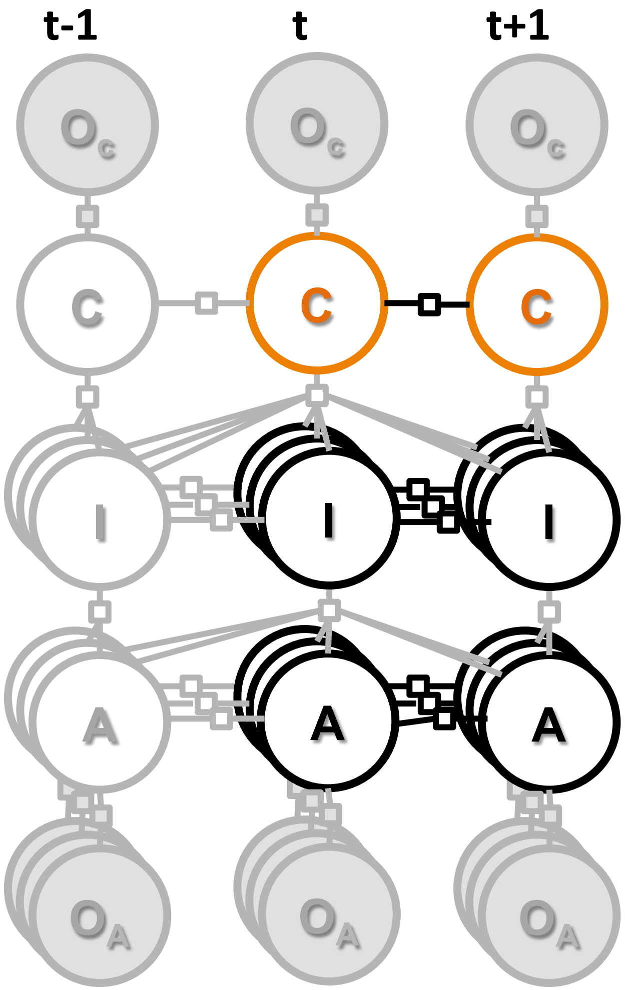
- STL of all targets



Choi et al, 09

[Choi ECCV12]

Activity Transition Potential

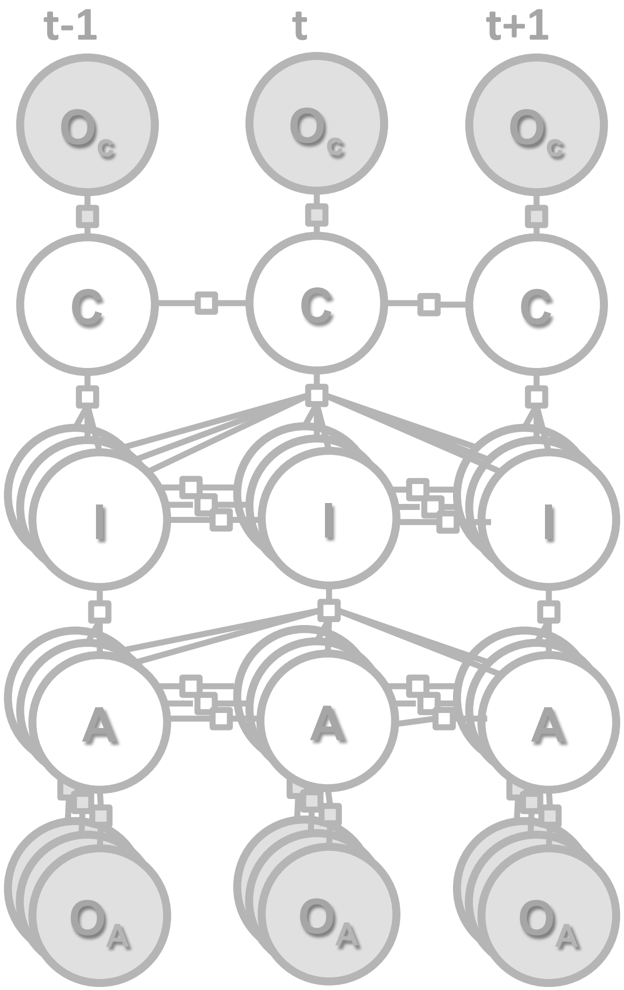


$$\begin{aligned}\Psi(C,I,A,O,f) =\ & \Psi(A,O)+\Psi(I,A,f)+\Psi(C,I)+\Psi(C,O)+ \\ & \underline{\Psi(C)+\Psi(I)+\Psi(A)} - c^T f,\ f \in S\end{aligned}$$

Smooth activity transition

[Choi ECCV12]

Trajectory Estimation



$$\Psi(C, I, A, O, f) = \Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) + \Psi(C) + \Psi(I) + \Psi(A) \underline{- c^T f,\ f \in S}$$

[Choi ECCV12]

Training the Graphical Model

- Model weights can be learned in a Max-Margin framework using Structural SVM.

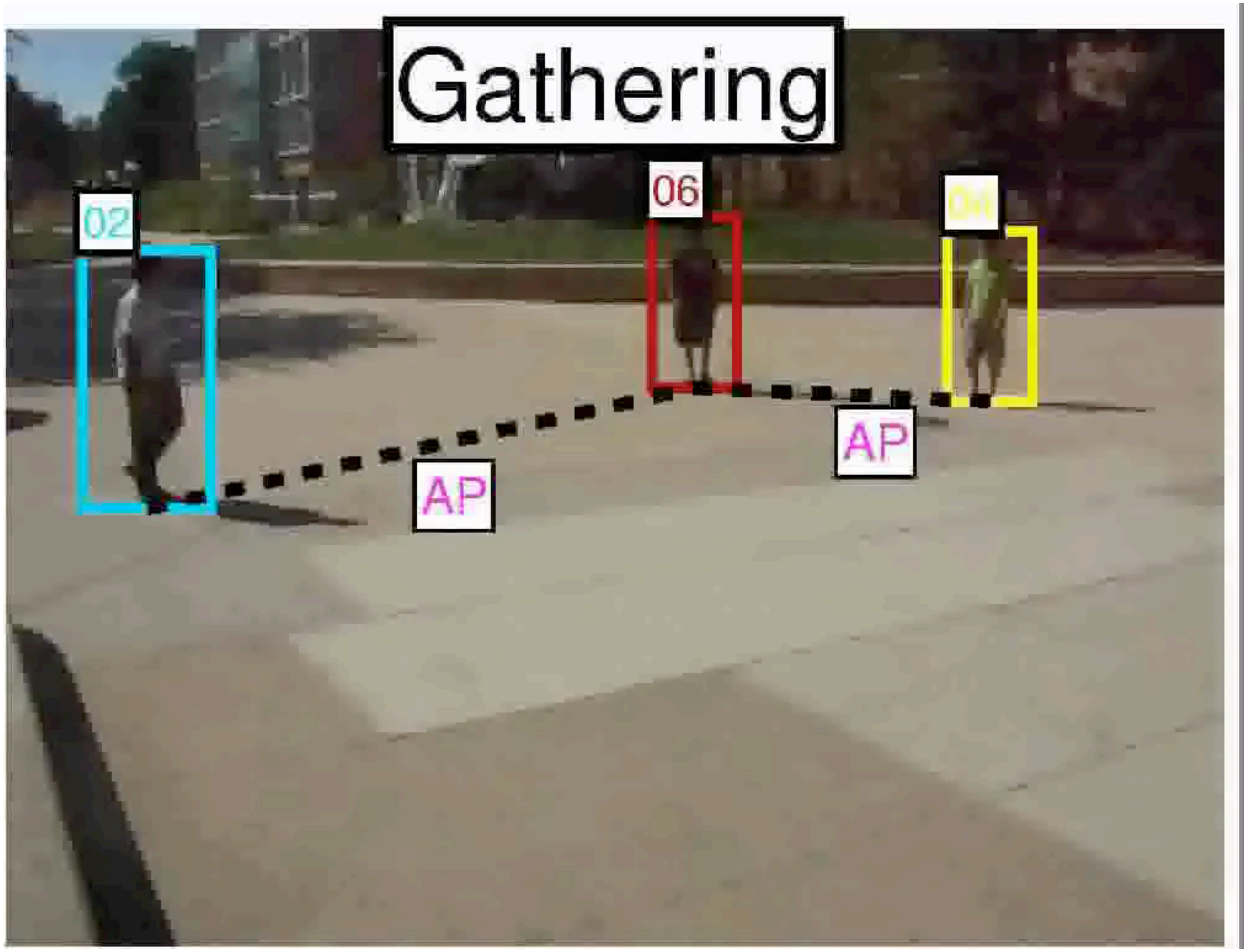
$$\min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i, \ \text{ s.t. } \forall i, \xi_i \geq 0$$

$$\forall i, \ \forall \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}_i : \ \langle \mathbf{w}, \delta\Psi_i(\mathbf{y}) \rangle \geq \triangle(\mathbf{y}_i, \mathbf{y}) - \xi_i$$

Tsochantaridis et al, 2004

[Choi ECCV12]

Example Classification Result



Interaction labels

AP: approaching

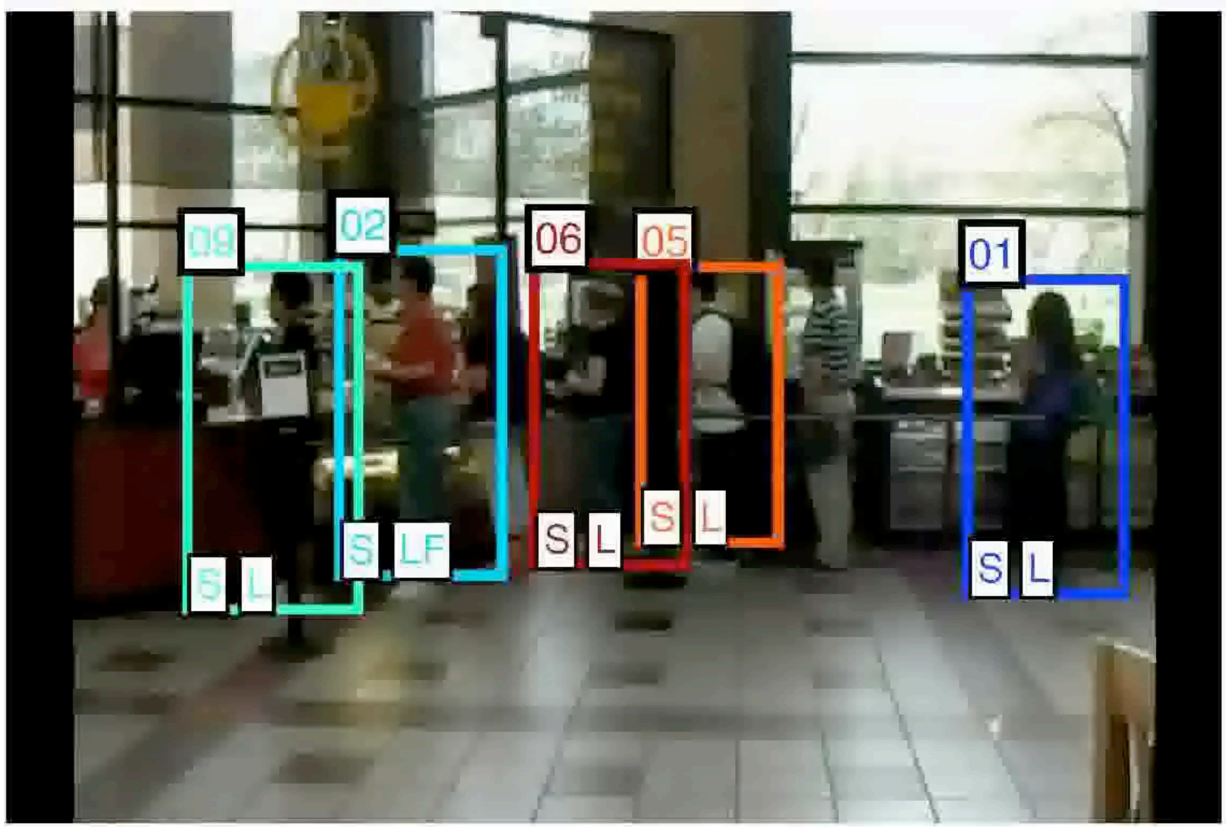
FE: facing-each-other

SR: standing-in-a-row

...

[Choi ECCV12]

Example Classification Result

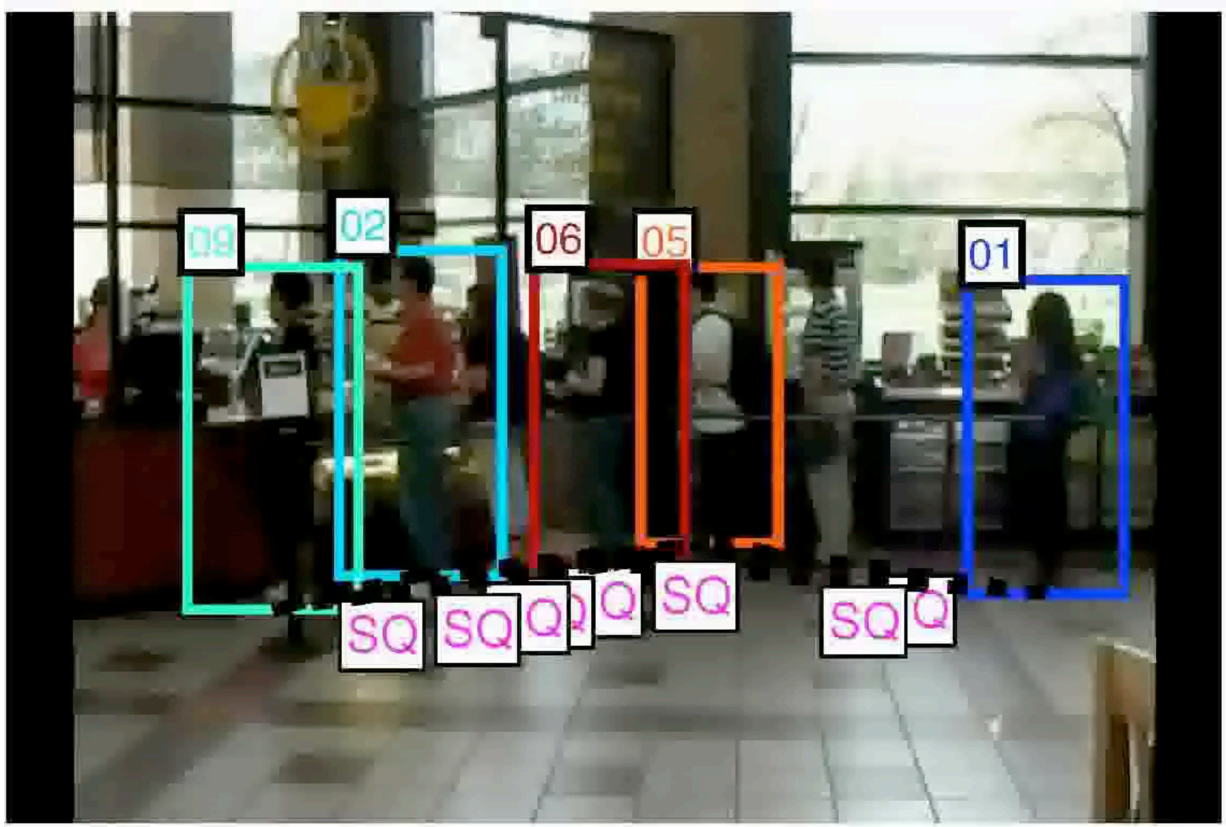


Atomic Activities

Action:
W - walking
S – standing

Pose (8 directions)
L - left
LF– left/front
F – front
RF- right/front
etc.

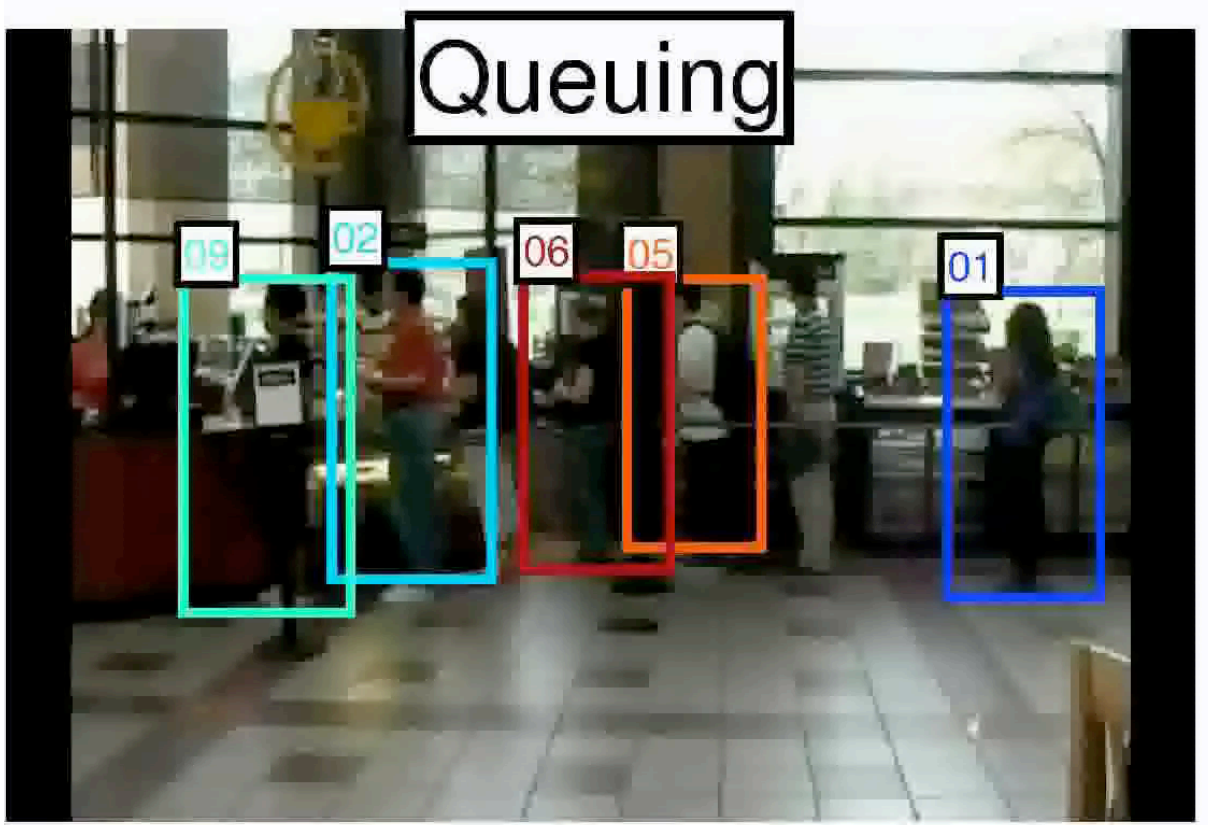
Example Classification Result



Pair-Interactions
- AP: approaching
- .....
- FE: facing-each-other
- SS: standing-side-by-side
- SQ: standing-in-a-queue

[Choi ECCV12]

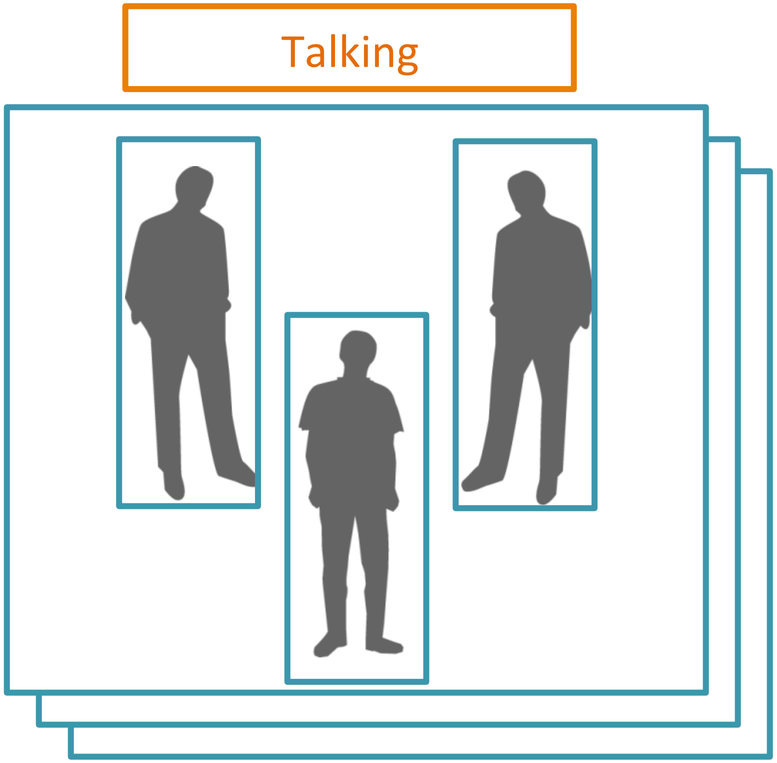
Example Classification Result

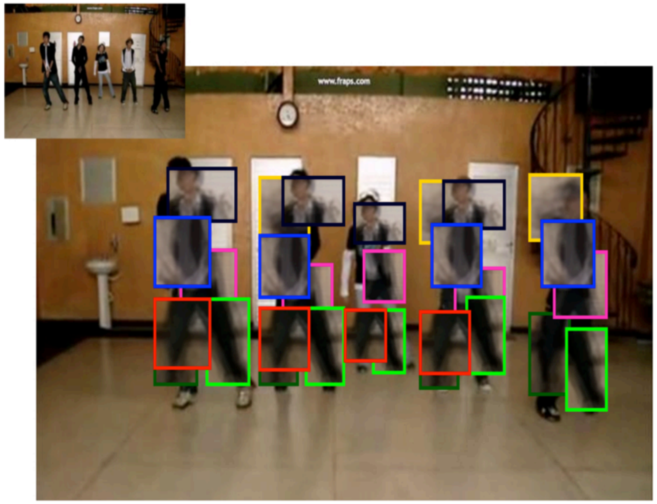


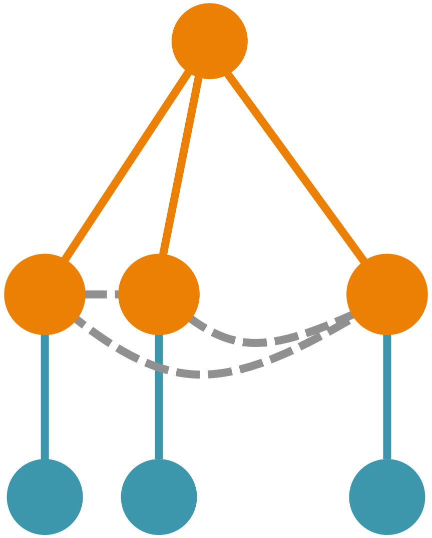
Learning Latent Constituents for Group Activity Recognition (Antic ECCV14)

Input: video and box tracks

Output: group activity labels in time







group activities

latent constituents

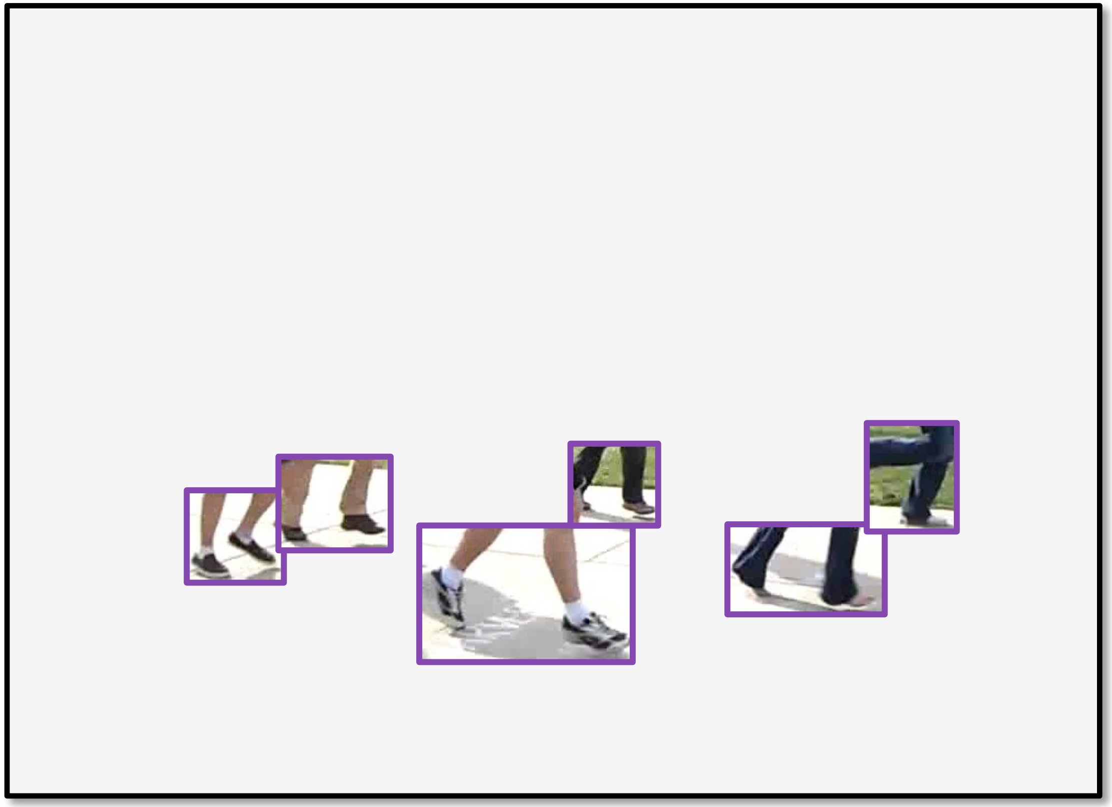
video/boxes

[Antic ECCV14]

Meaningful Parts of Group Behavior

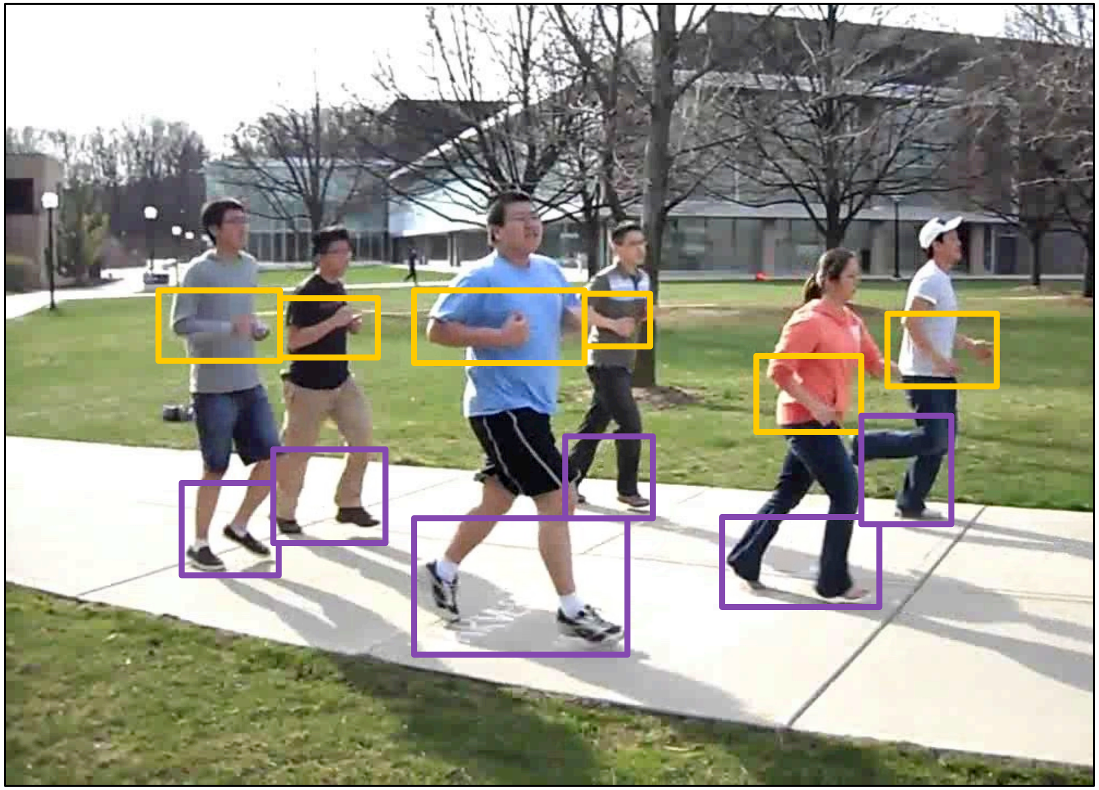
[Antic ECCV14]

Meaningful Parts of Group Behavior



[Antic ECCV14]

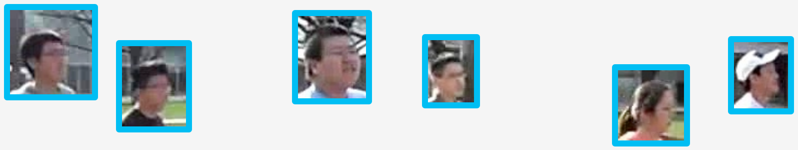
Meaningful Parts of Group Behavior



[Antic ECCV14]

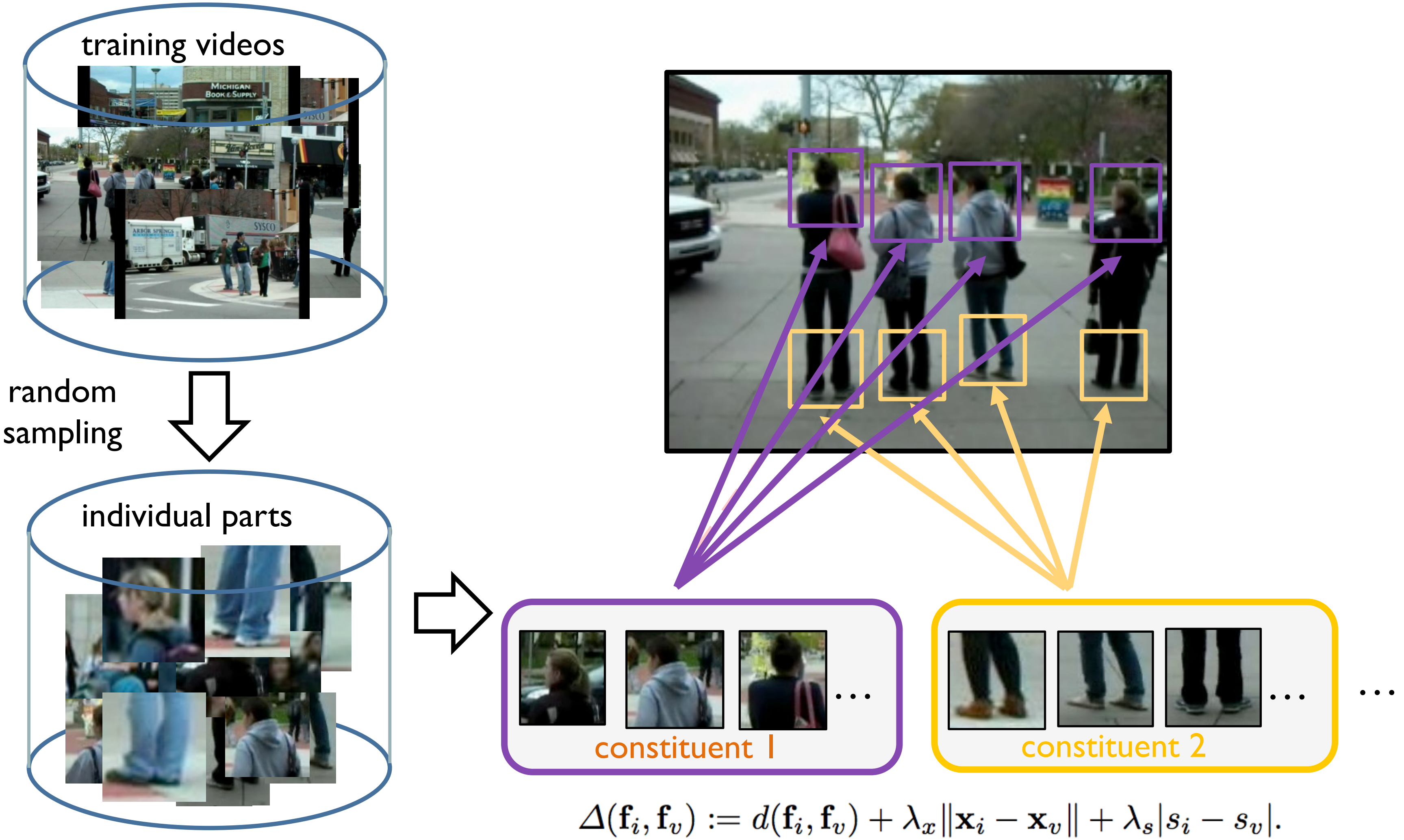
Less Meaningful Parts

Less meaningful parts



[Antic ECCV14]

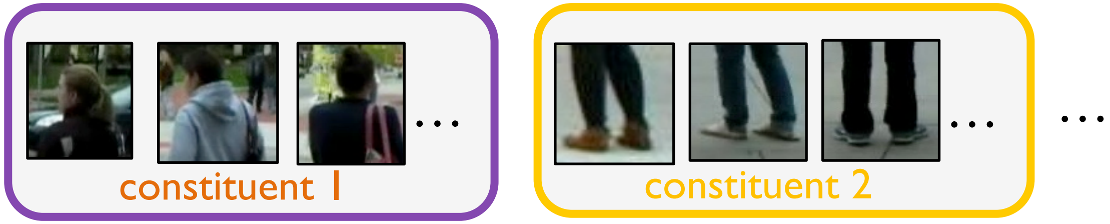
Learning Mid-level Constituents

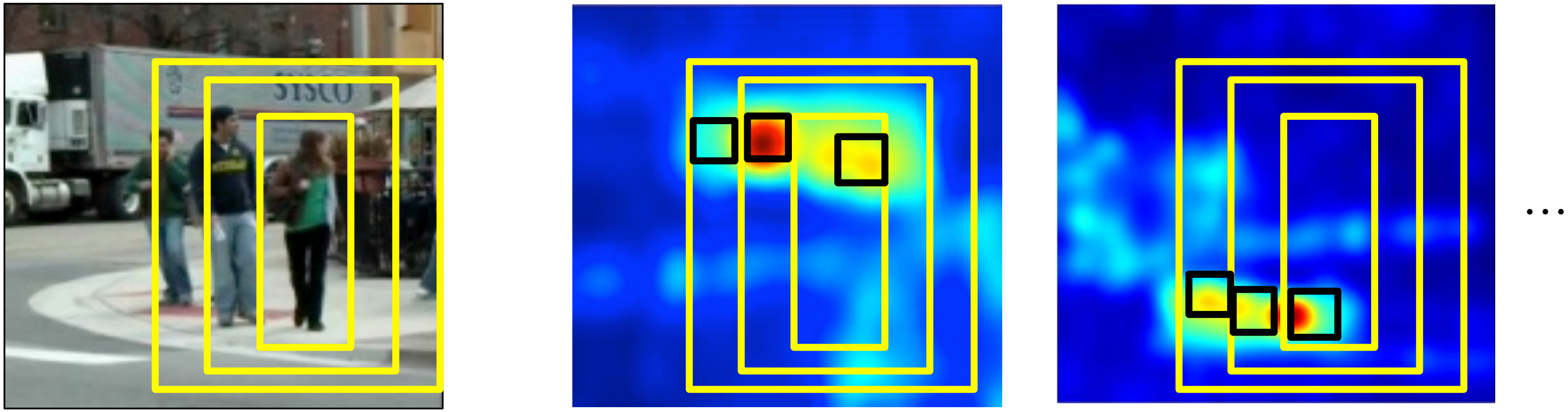


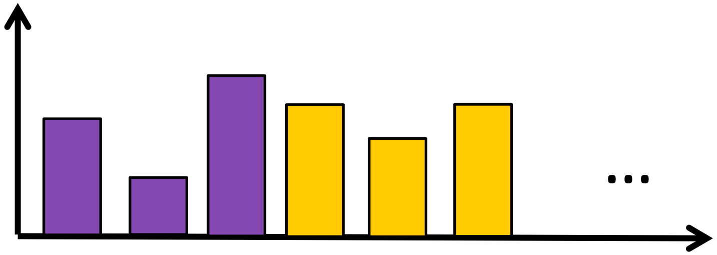
$$\Delta(\mathbf{f}_i, \mathbf{f}_v) := d(\mathbf{f}_i, \mathbf{f}_v) + \lambda_x \|\mathbf{x}_i - \mathbf{x}_v\| + \lambda_s |s_i - s_v|.$$

[Antic ECCV14]

Encoding Social Signal with Latent Constituents

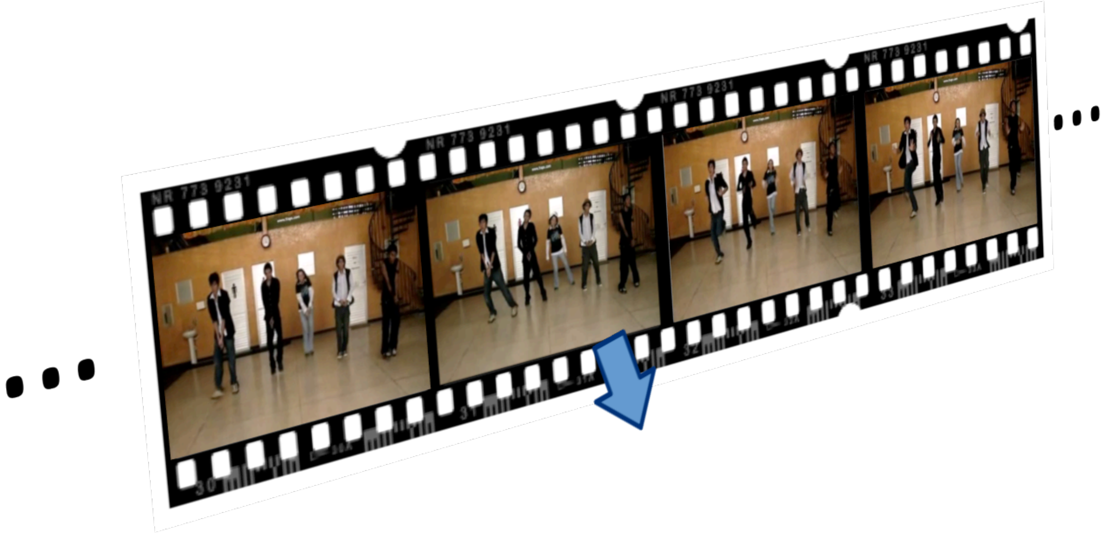






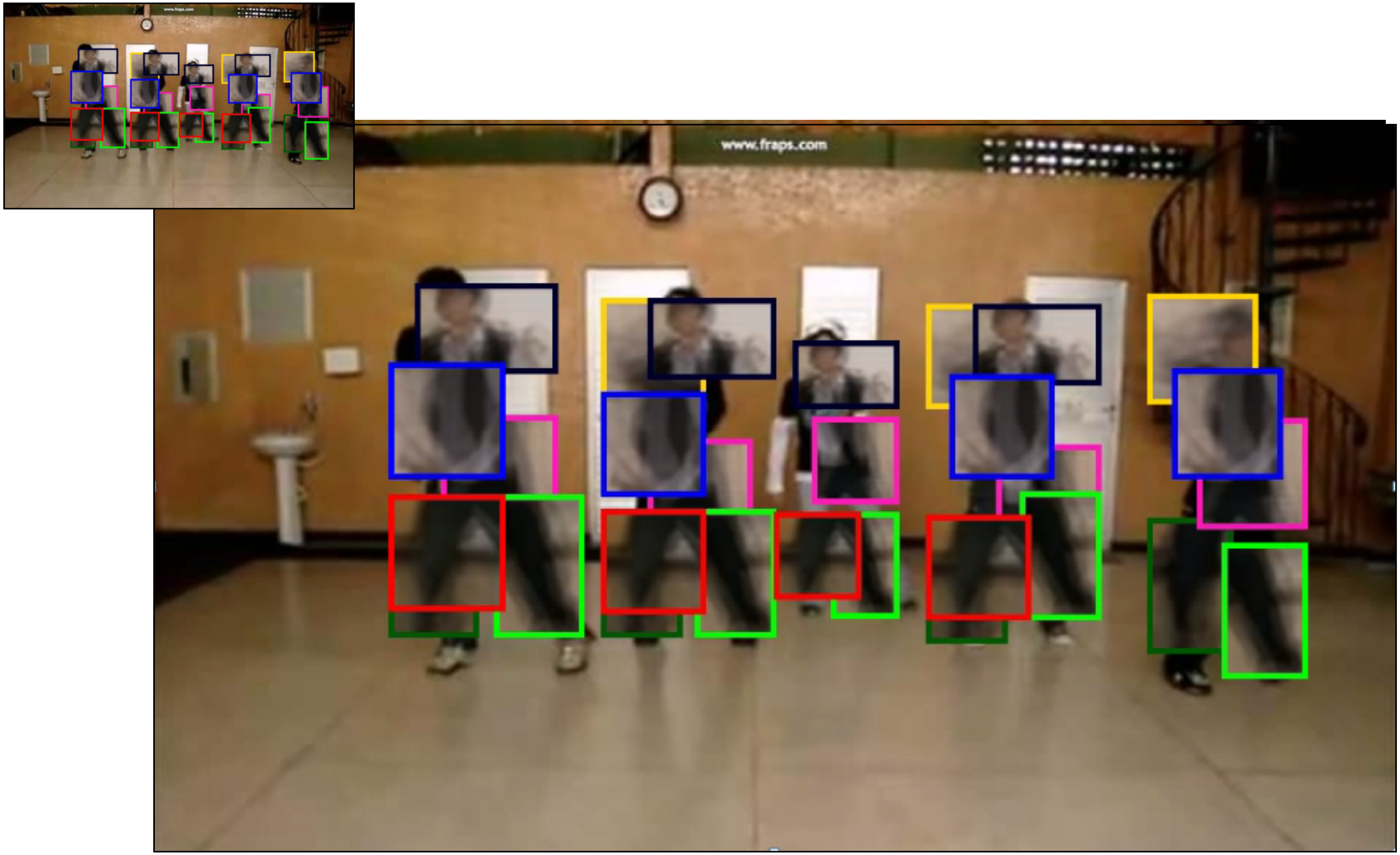
[Antic ECCV14]

Behavior Recognition with Constituents



[Antic ECCV14]

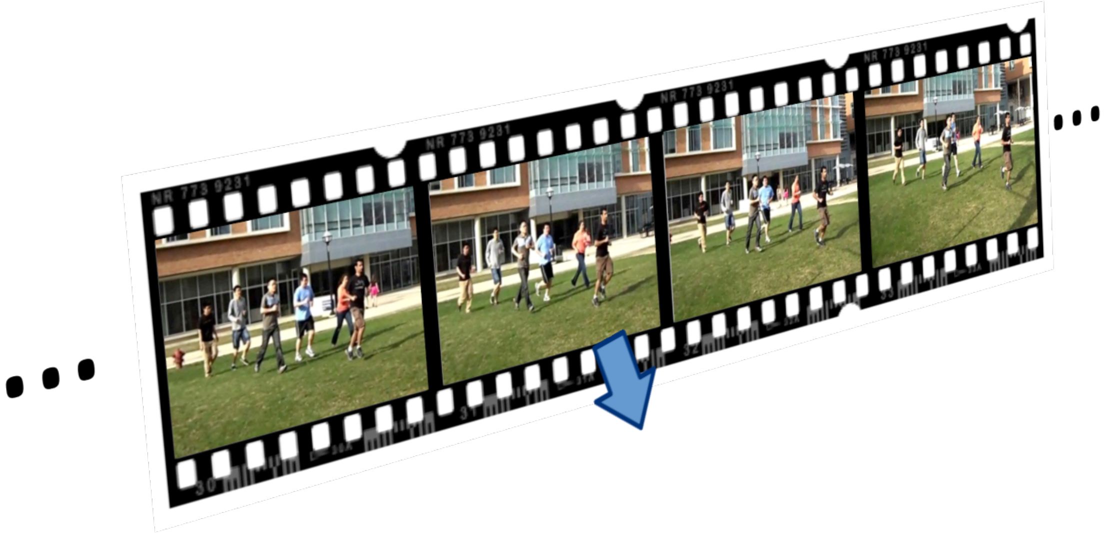
Behavior Recognition with Constituents



🟥 crossing 🟨 waiting 🟩 queueing 🟦 talking 🟦 dancing 🟪 jogging

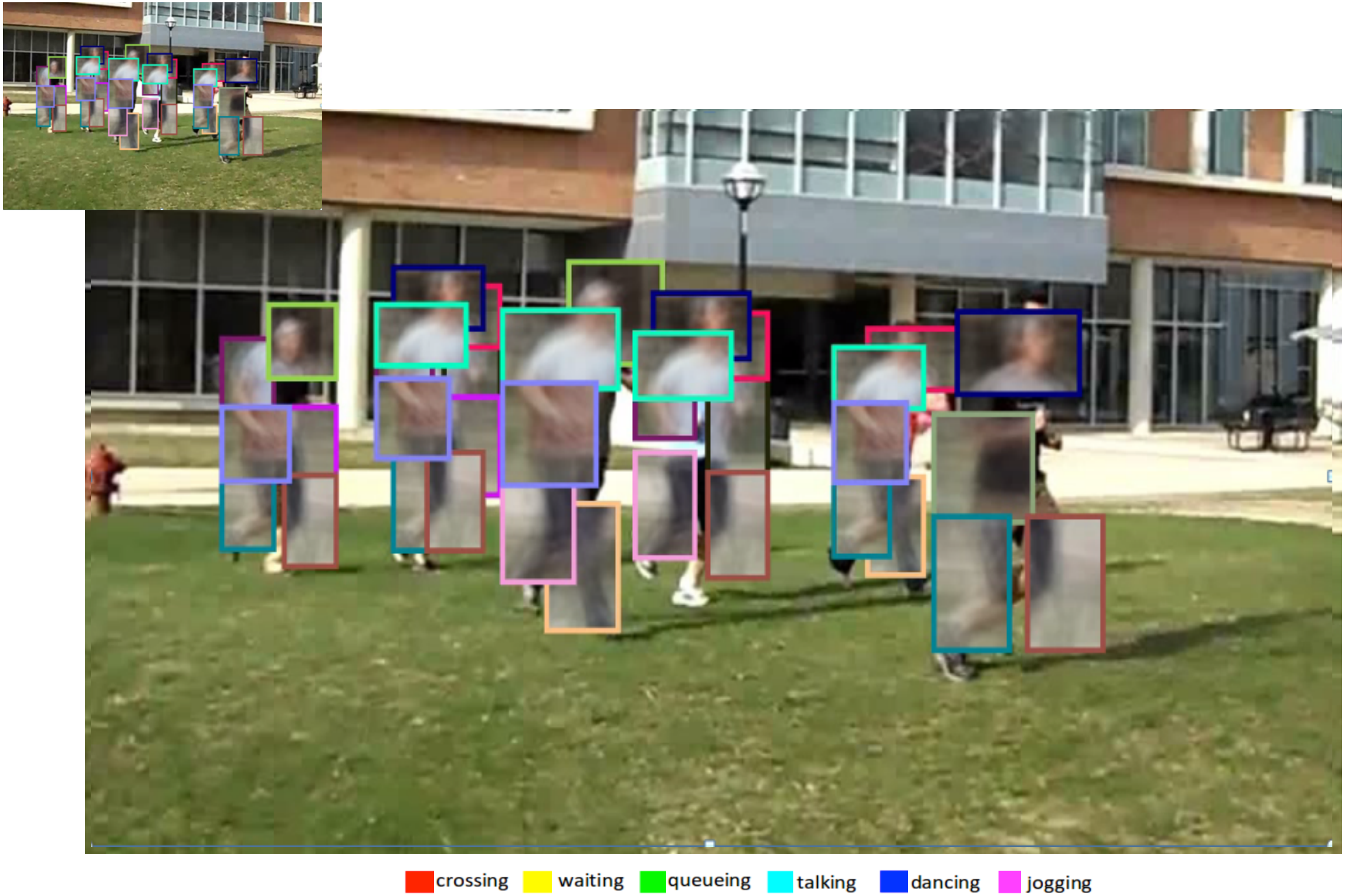
[Antic ECCV14]

Behavior Recognition with Constituents



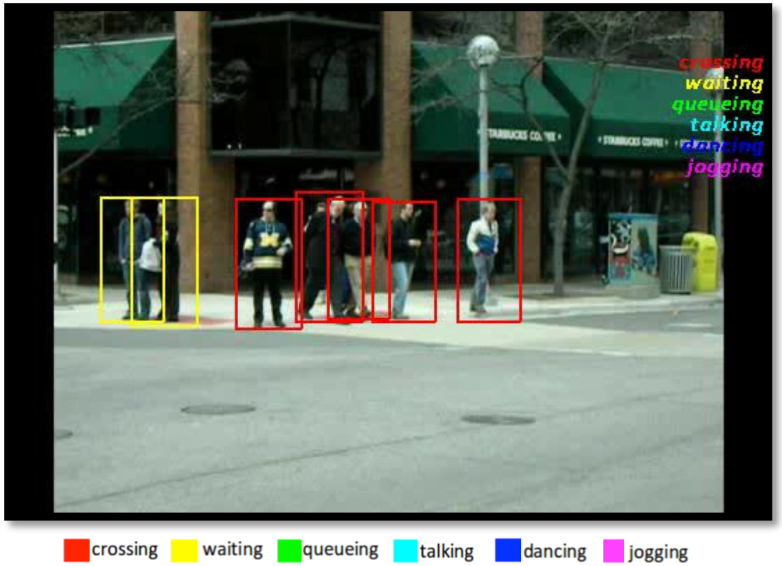
[Antic ECCV14]

Behavior Recognition with Constituents

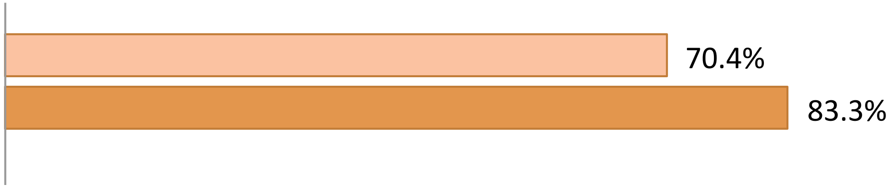


[Antic ECCV14]

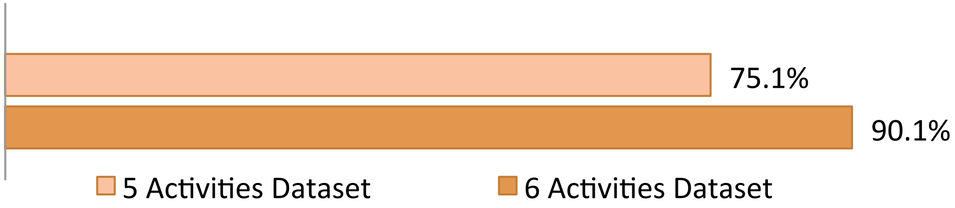
Quantitative Evaluation



Holistic approach (full b-boxes):



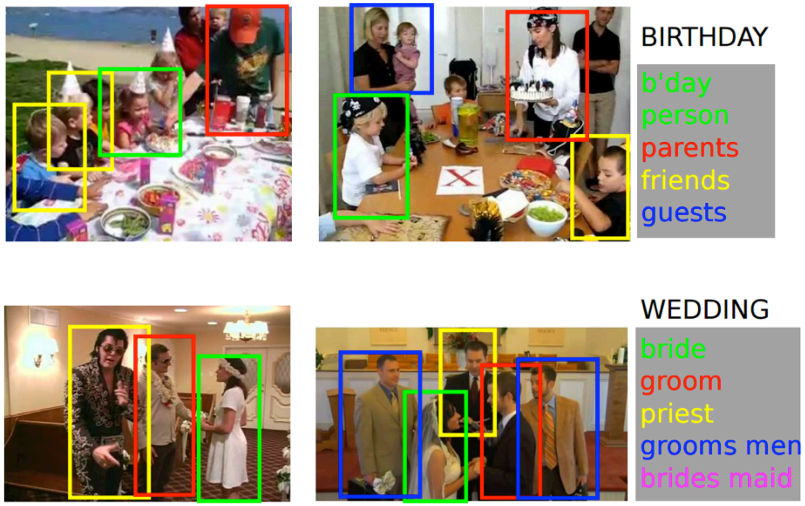
Latent constit's (functional grouping):

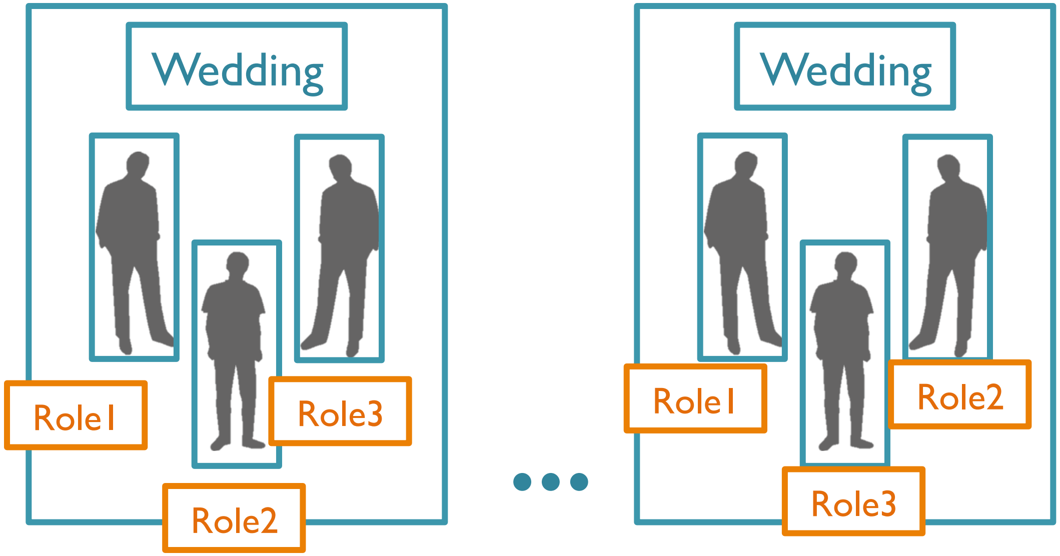


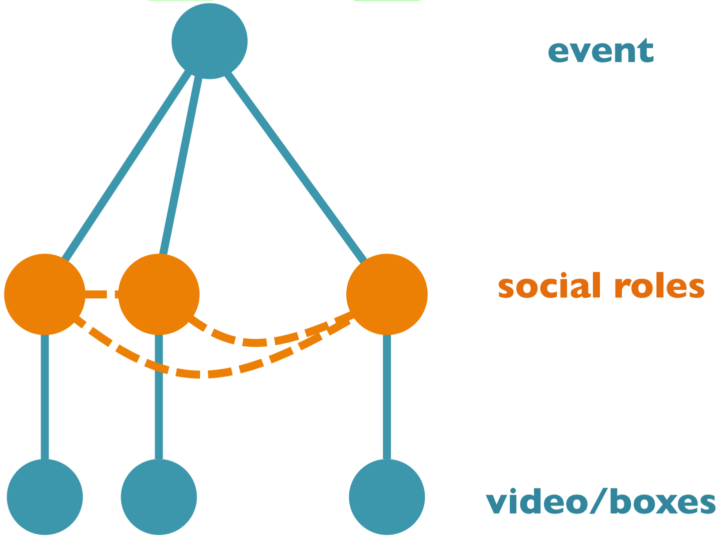
Social Role Discovery (Ramanathan CVPR13)

Input: videos with event labels, box tracks

Output: groups with activity label

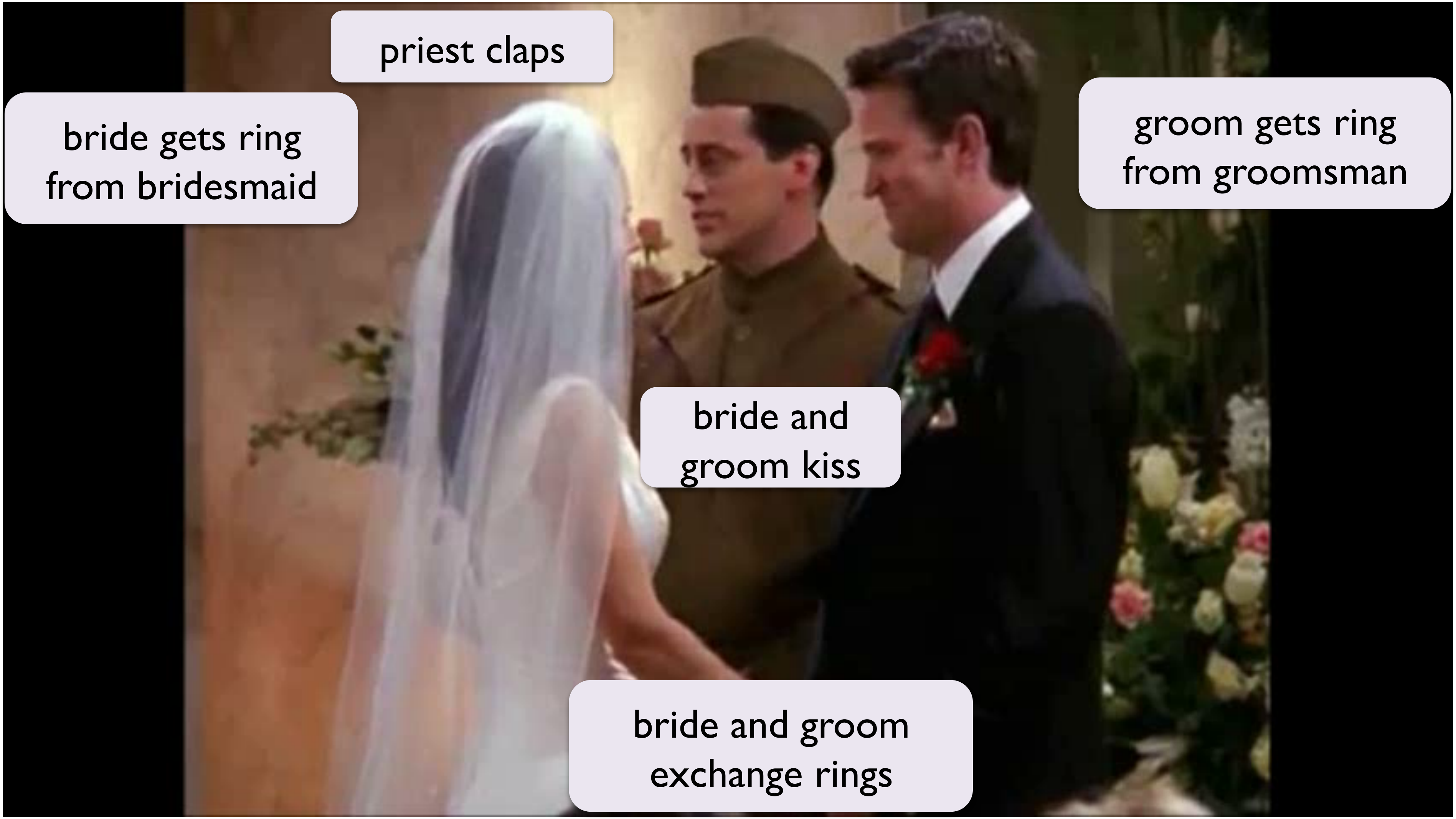






[Ramanathan CVPR13]

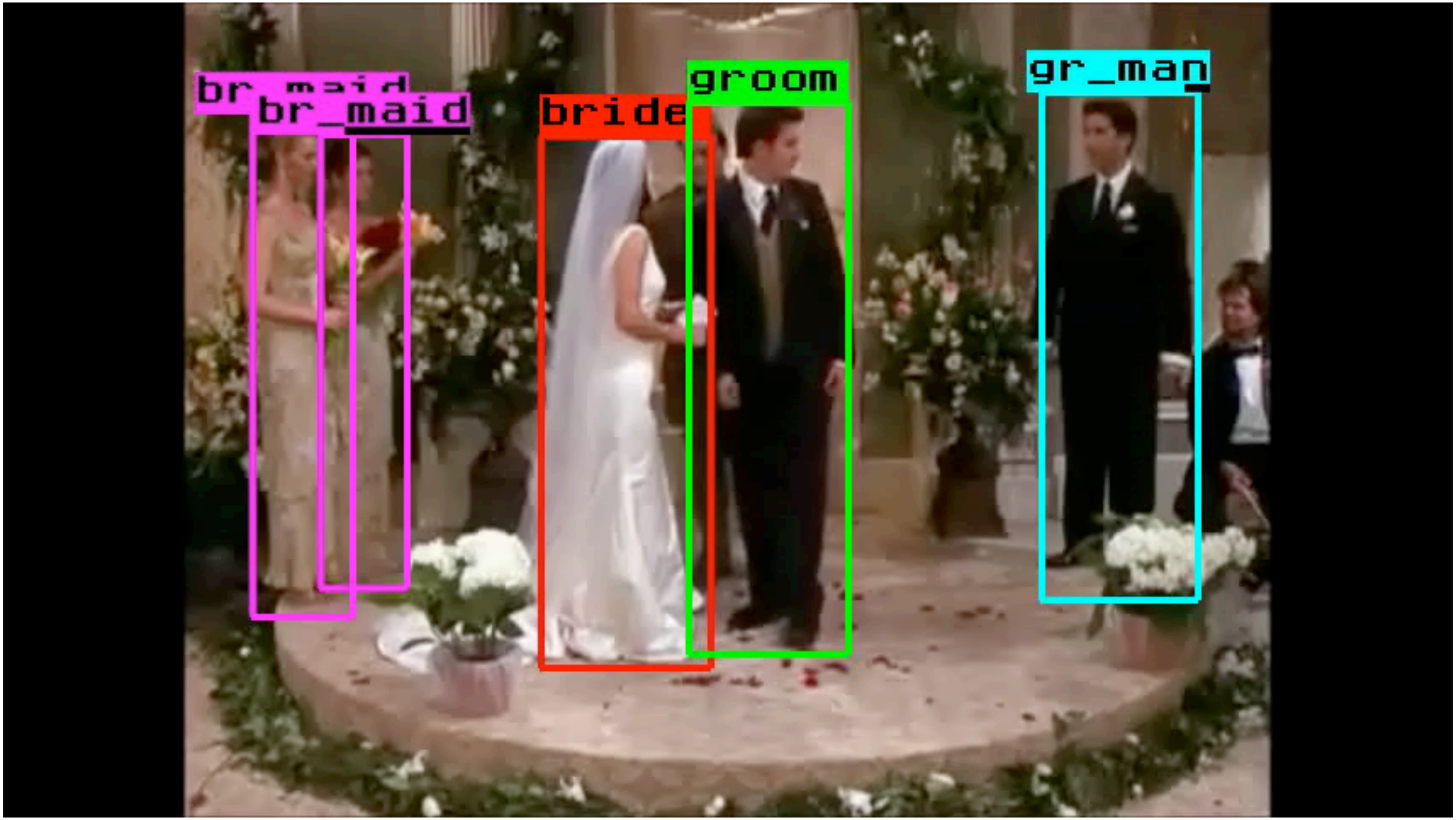
Humans in Social Setting


priest claps
bride gets ring from bridesmaid
groom gets ring from groomsman
bride and groom kiss
bride and groom exchange rings

Slide from Ramanathan et al, 2013

[Ramanathan CVPR13]

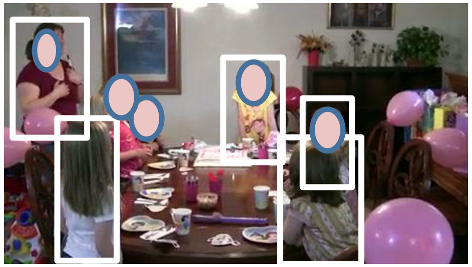
Goal: Identify social roles



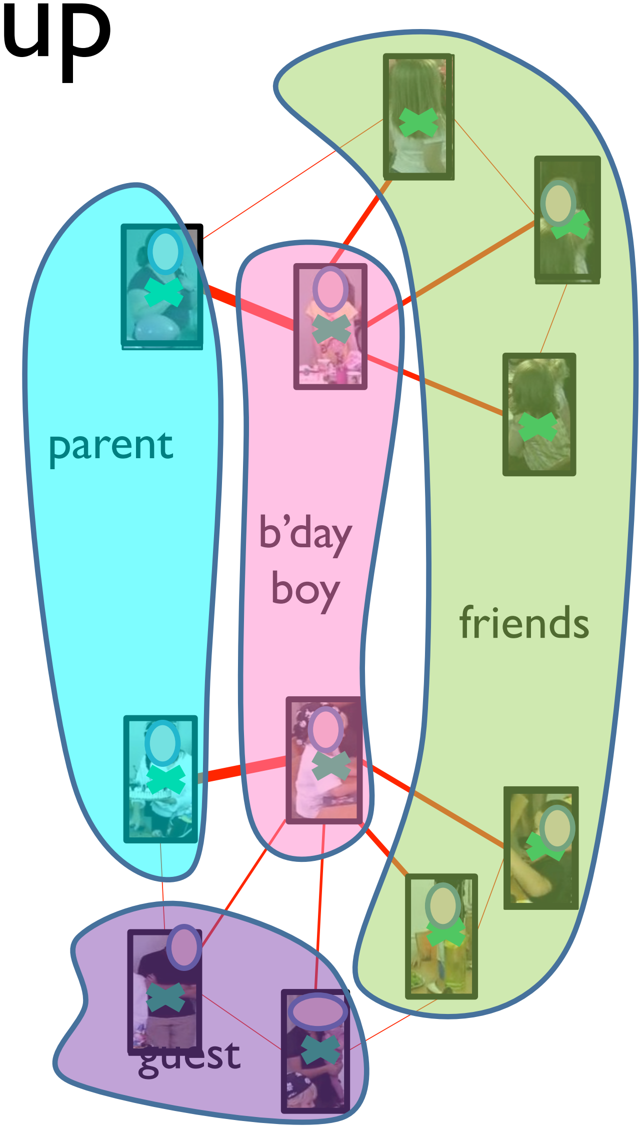
Slide from Ramanathan et al, 2013

[Ramanathan CVPR13]

Problem setup



⋮

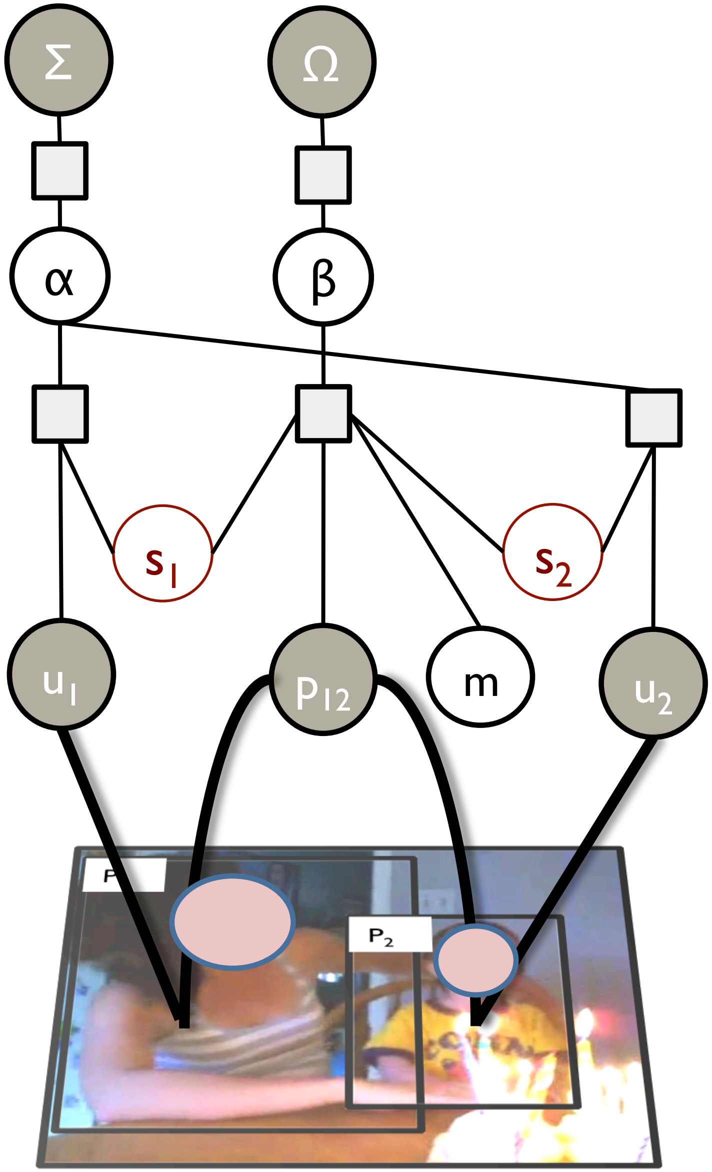


✖ person-specific role features

▬ inter-role interaction features

[Ramanathan CVPR13]

Social role model

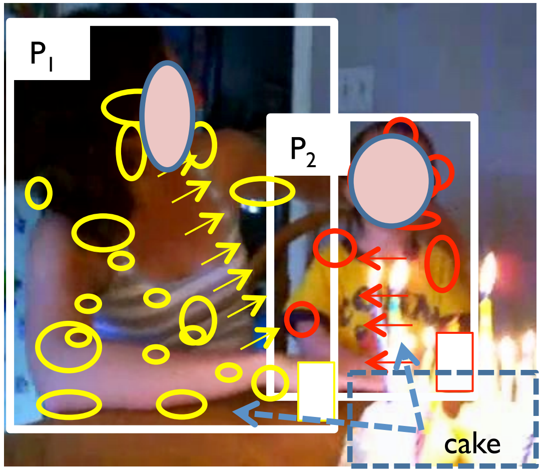


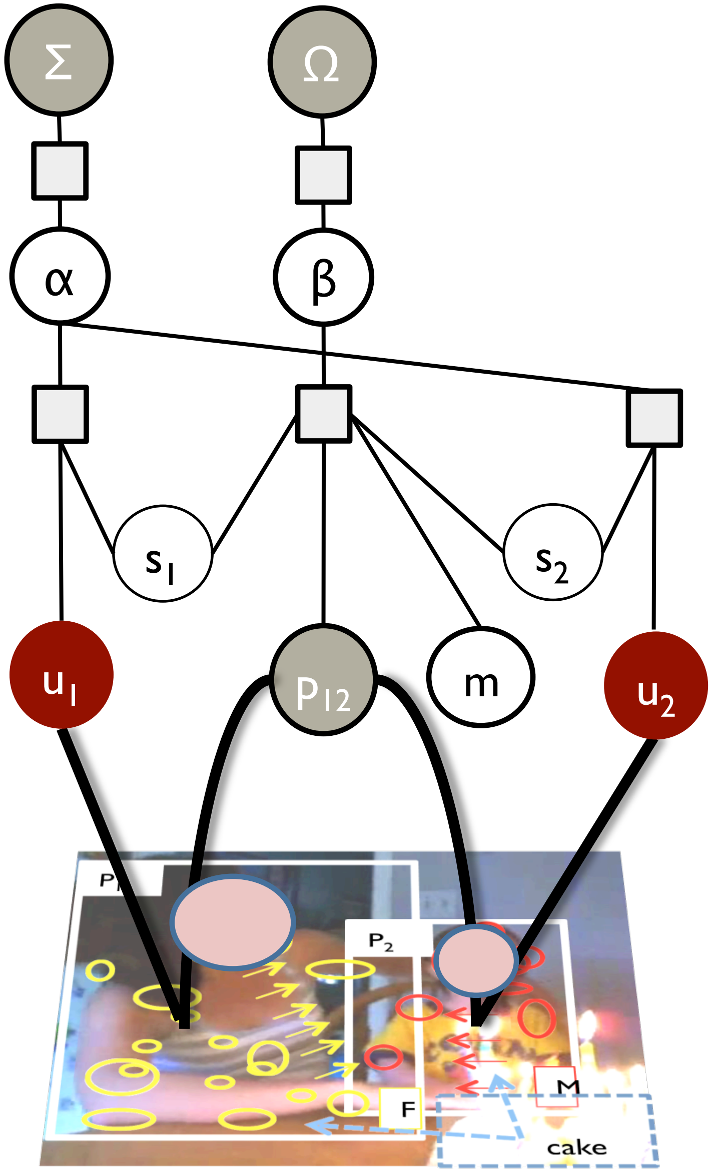
[Ramanathan CVPR13]

Social role model

Person-specific features

- Role specific **actions**
  - **HOG3D** from person tube
  - **Trajectory** of person
- **Color** and **Gender** features
- Object interaction features





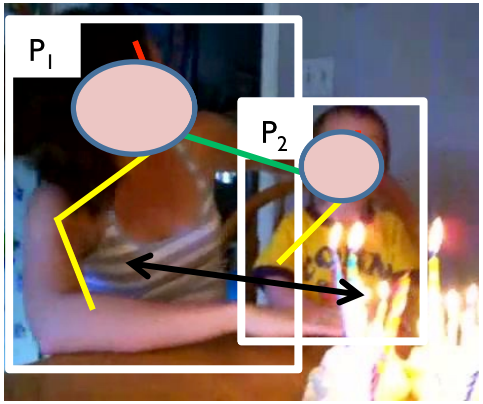
[Ramanathan CVPR13]

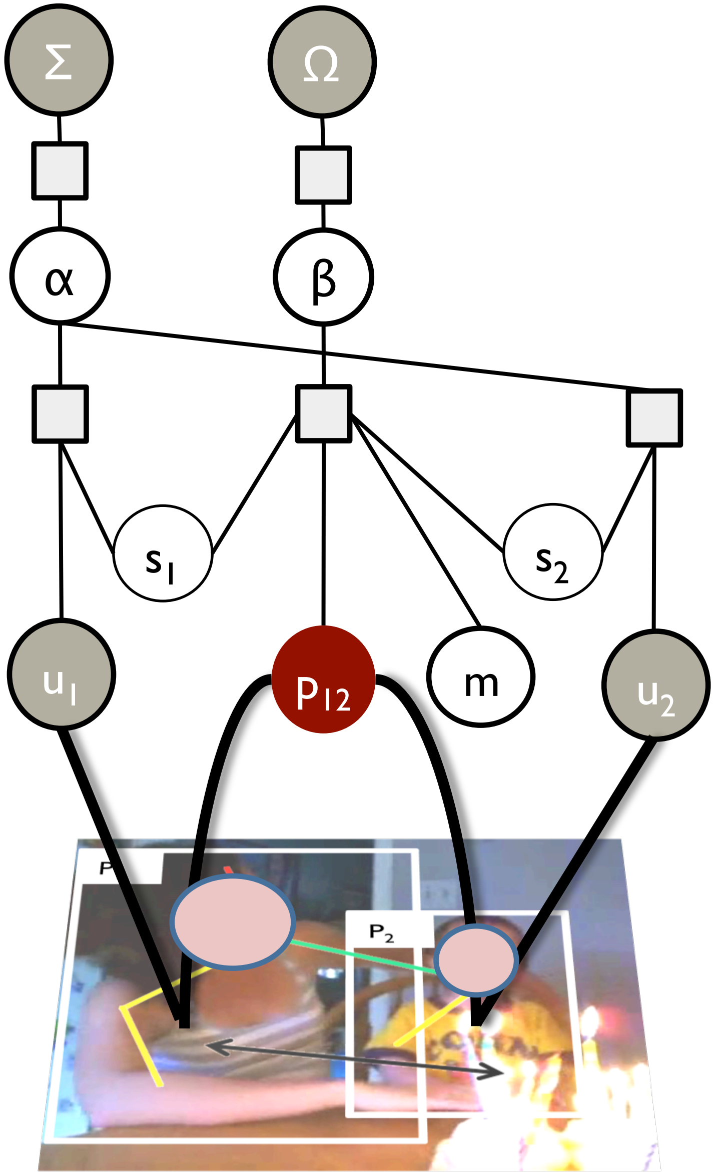
Social role model

Person-specific features

Inter-role features

- Spatio-temporal features
- Proxemic features





[Ramanathan CVPR13]

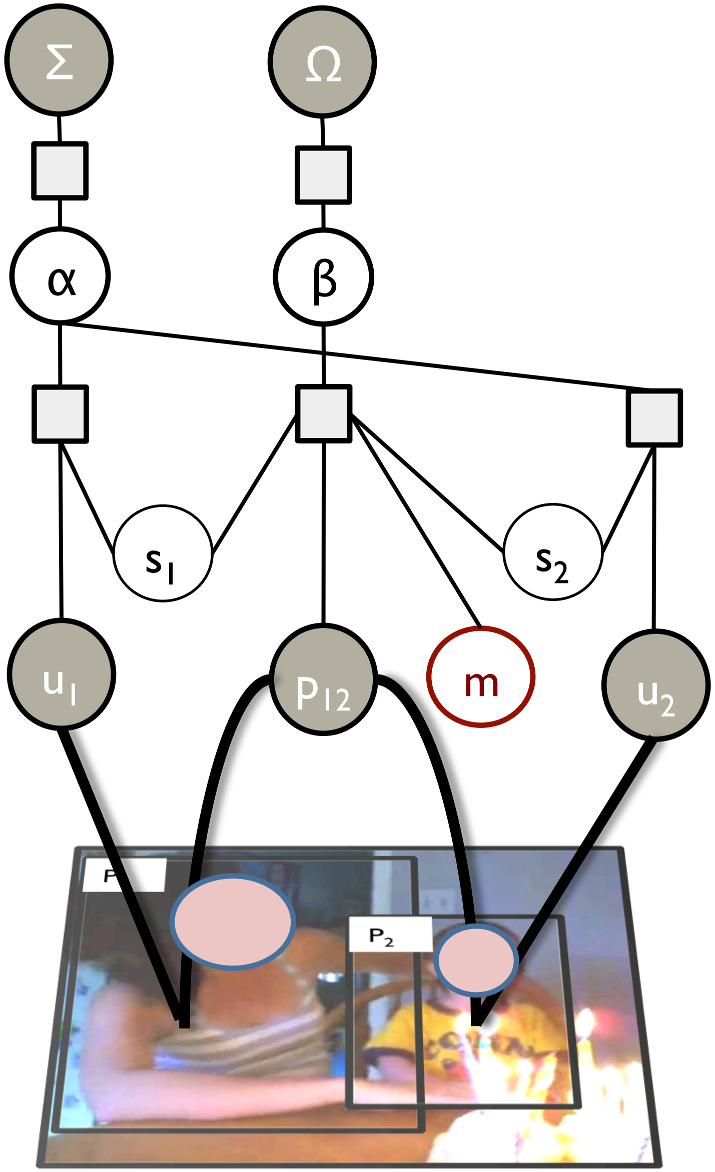
Social role model

Person-specific features

Inter-role features

Reference role

only interactions with reference considered for tractable inference



[Ramanathan CVPR13]

Social role model

Person-specific features

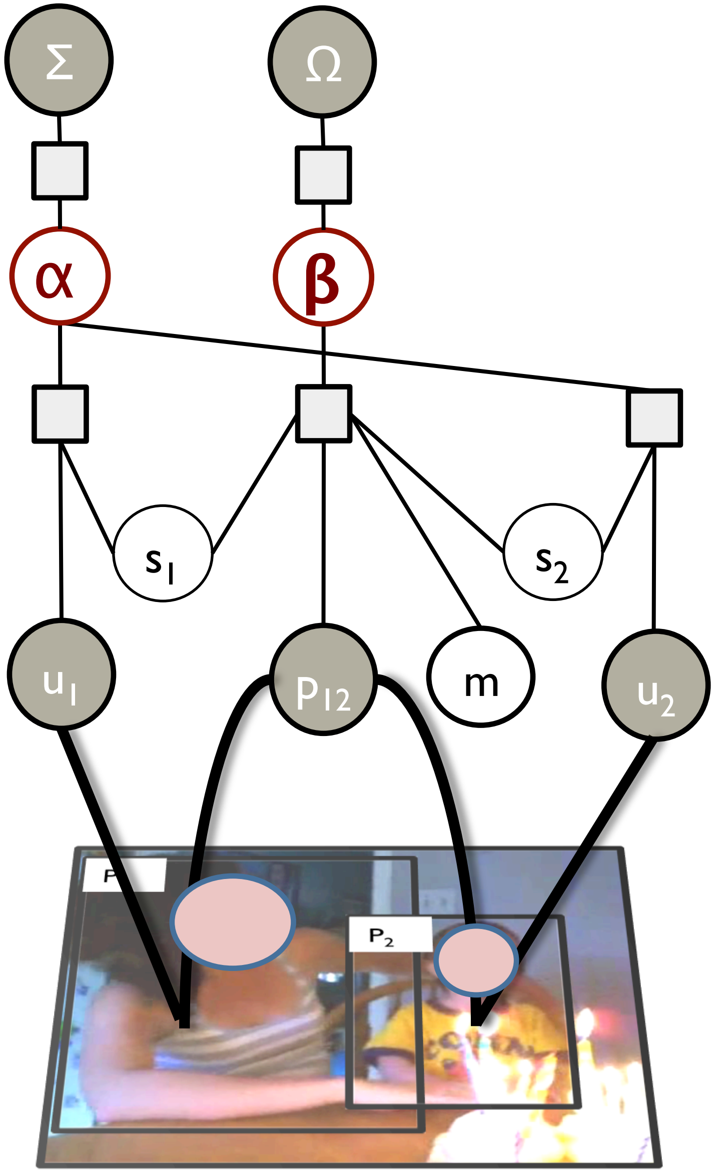
Inter-role features

Reference role

Model parameters

α is the person-specific feature weight

β is the inter-role feature weight



[Ramanathan CVPR13]

Social role model

Person-specific features

Inter-role features

Reference role

Model parameters

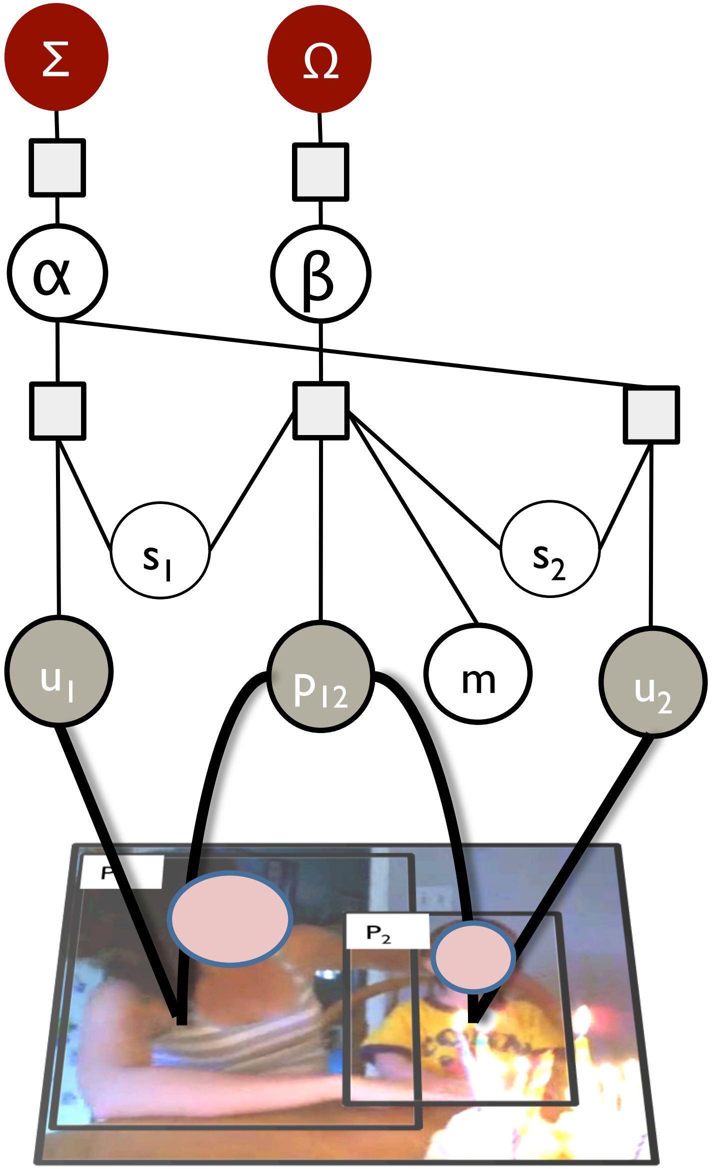
α is the person-specific feature weight

β is the inter-role feature weight

Σ

Ω

Gaussian priors for regularization

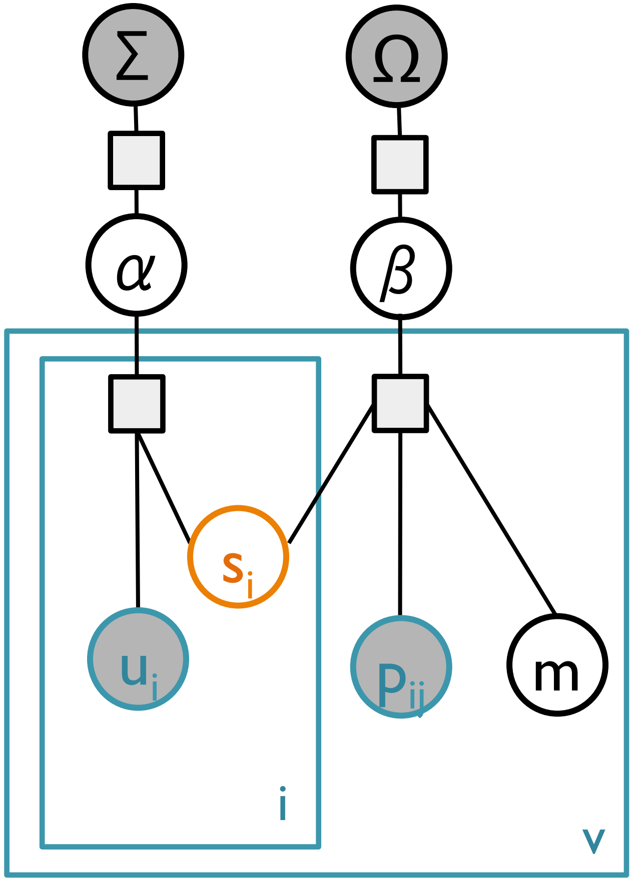


[Ramanathan CVPR13]

Social role model

Variational Inference

Jointly *Learn model parameters.*
Assign social roles.



[Ramanathan CVPR13]

Dataset

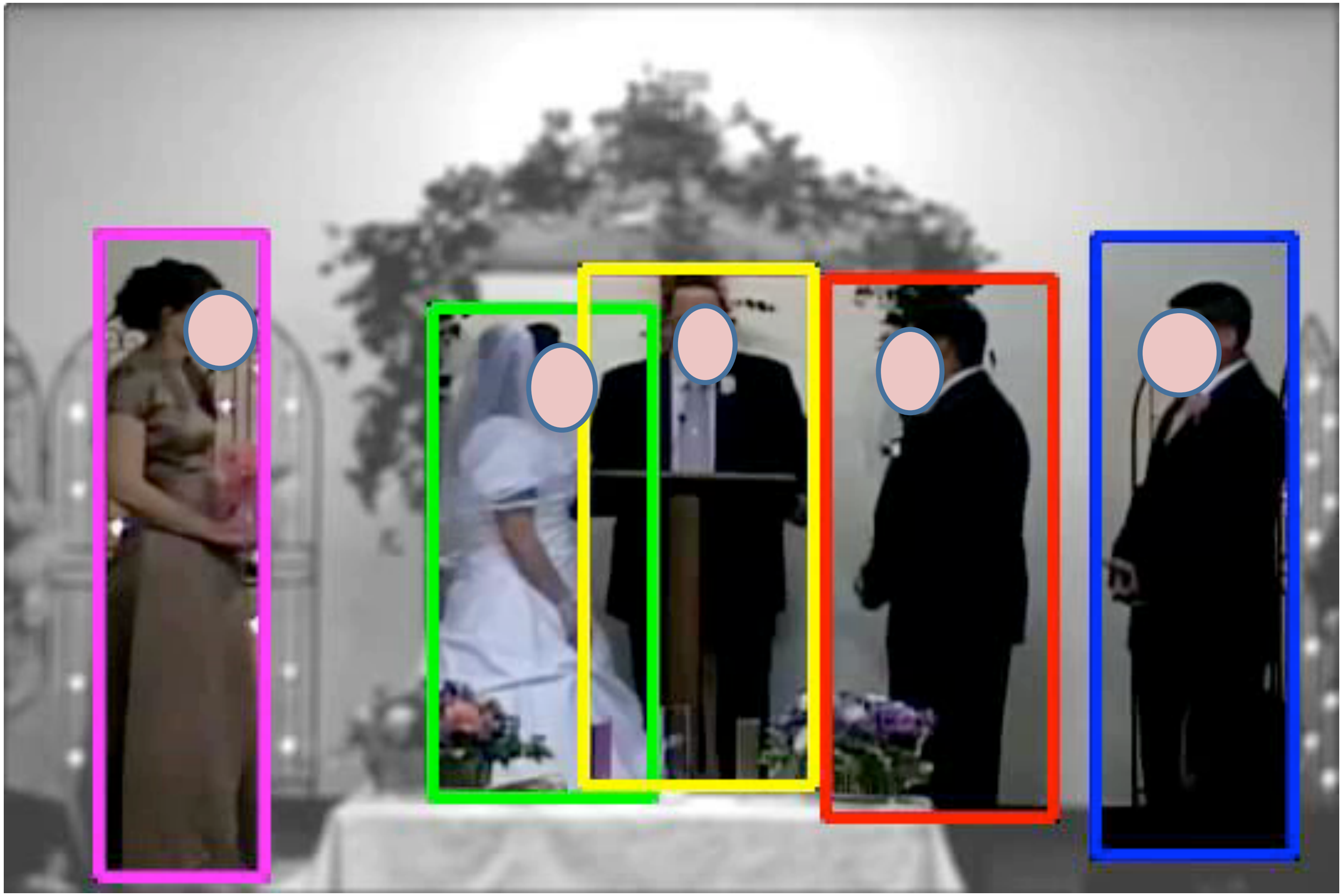
- Youtube Social Role Dataset
  - Available at http://vision.stanford.edu/vigneshr_release_data/youtube_CVPR13_social.tar.gz
  - Only the event type is provided.
  - Social roles are discovered in unsupervised fashion.

[Ramanathan CVPR13]

Results: role clusters

[Ramanathan CVPR13]

Results: role clusters



Results: role clusters

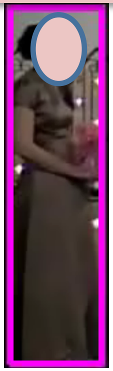
Bride

reference

Groom

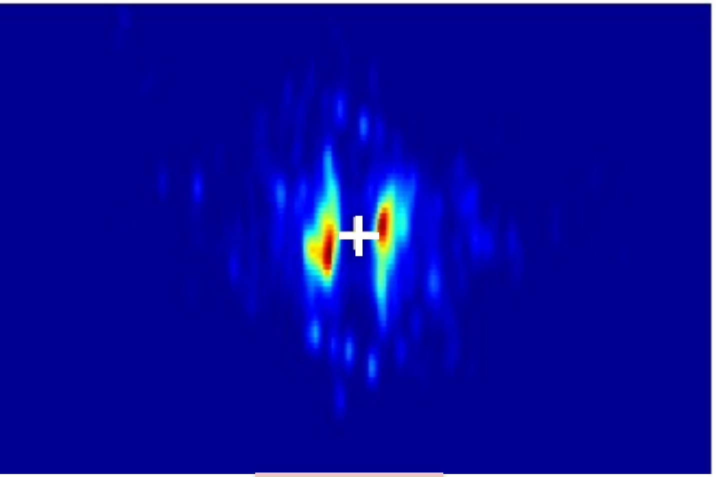
Priest

Brides maid



Grooms men

Spatial Relations



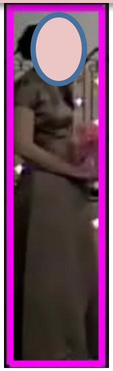
Bride

reference

Groom

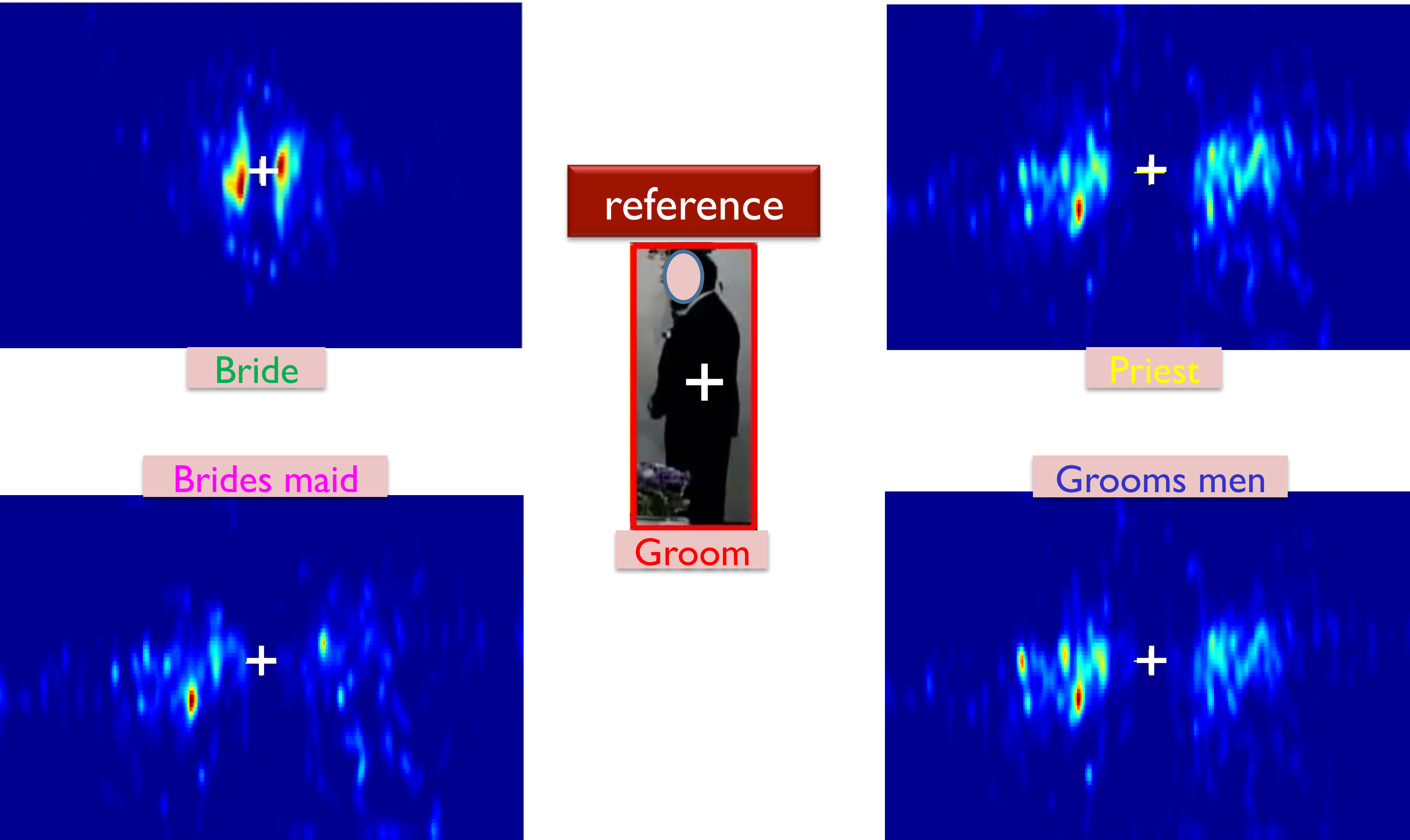
Priest

Brides maid



Grooms men

Spatial Relations



reference

Bride

Priest

Brides maid

Groom

Grooms men

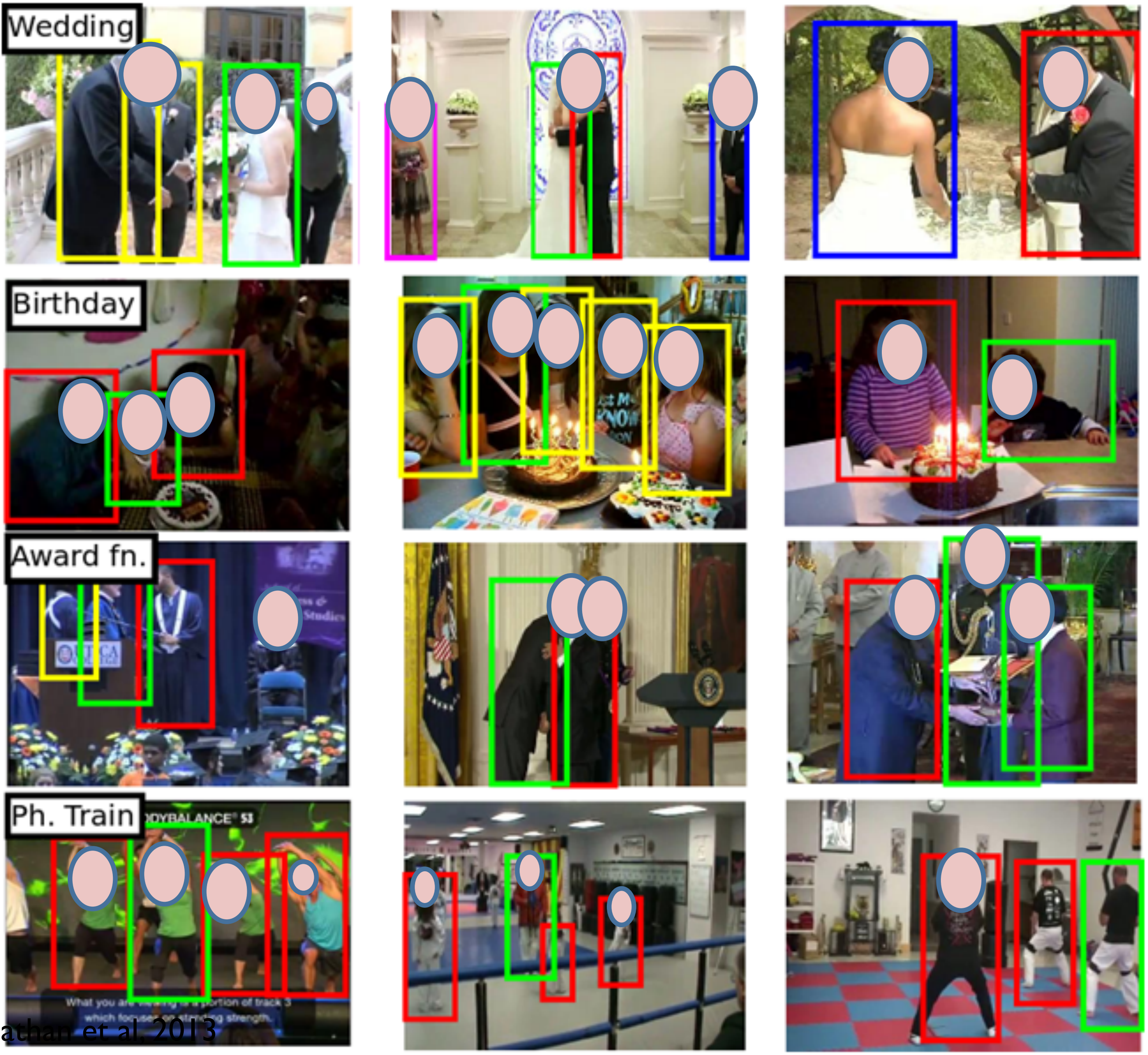
Results – Role clusters

bride
groom
priest
grooms men
brides maid

b'day person
parent
friends
guest

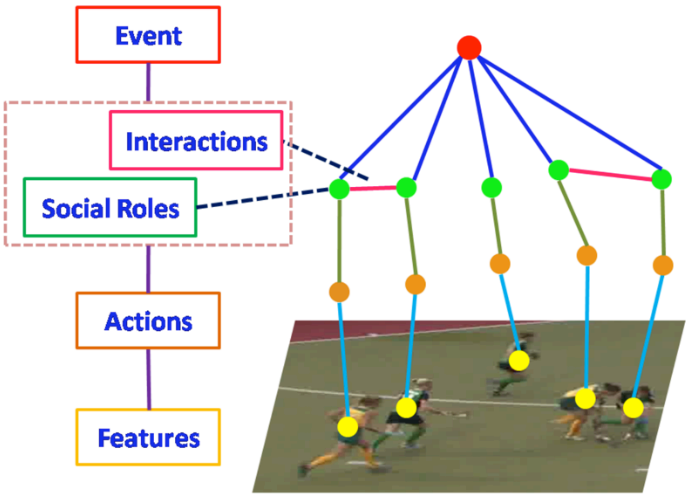
presenter
recipient
host
distributor

instructor
presenter



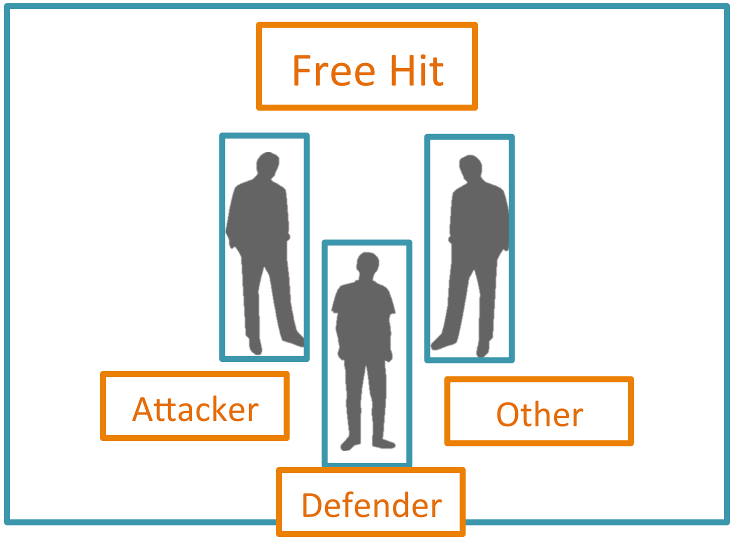
Slide from Ramanathan et al 2013

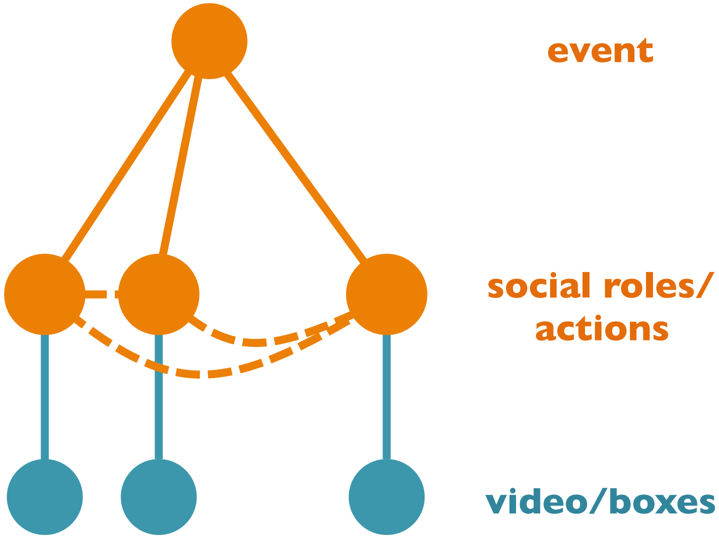
Social Role Discovery (Lan CVPR12)



Input: a video with box tracks

Output: social role and activity labels





[Lan CVPR12]

Semantic Description of Videos

actions

walk
run
jog
bend
shoot
dribble
…

social roles

attacker
first defender
man-marking
defend-space
Teammate
…

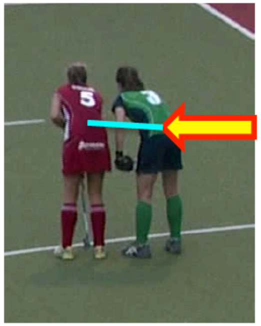
event

corner-hit
free-hit
attack play
…

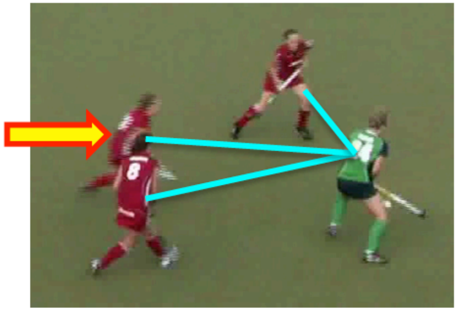
Social roles

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

man-marking

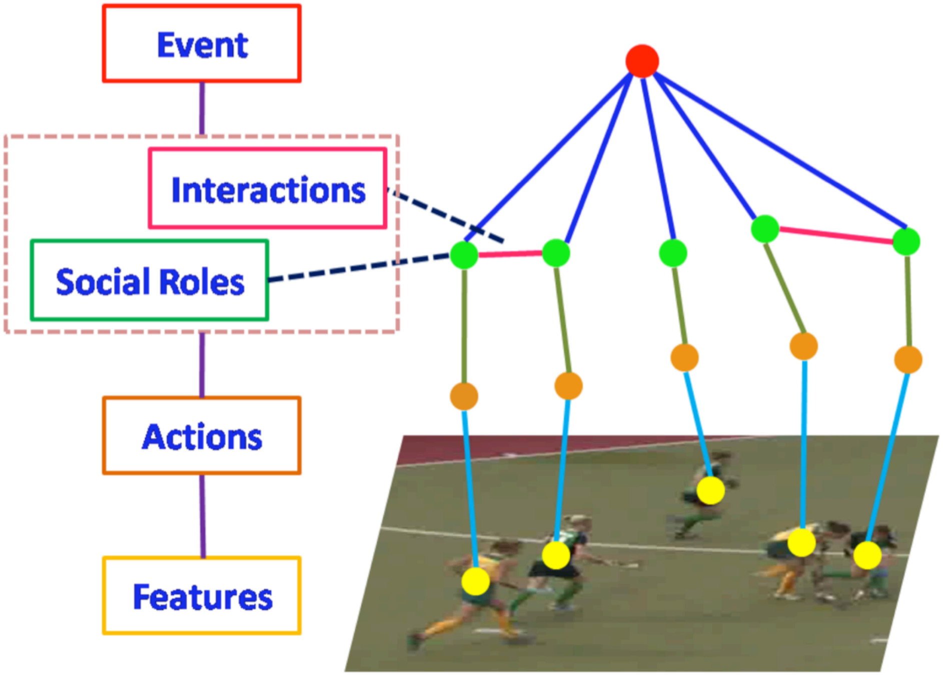


first defenders



[Lan CVPR12]

Hierarchical Model

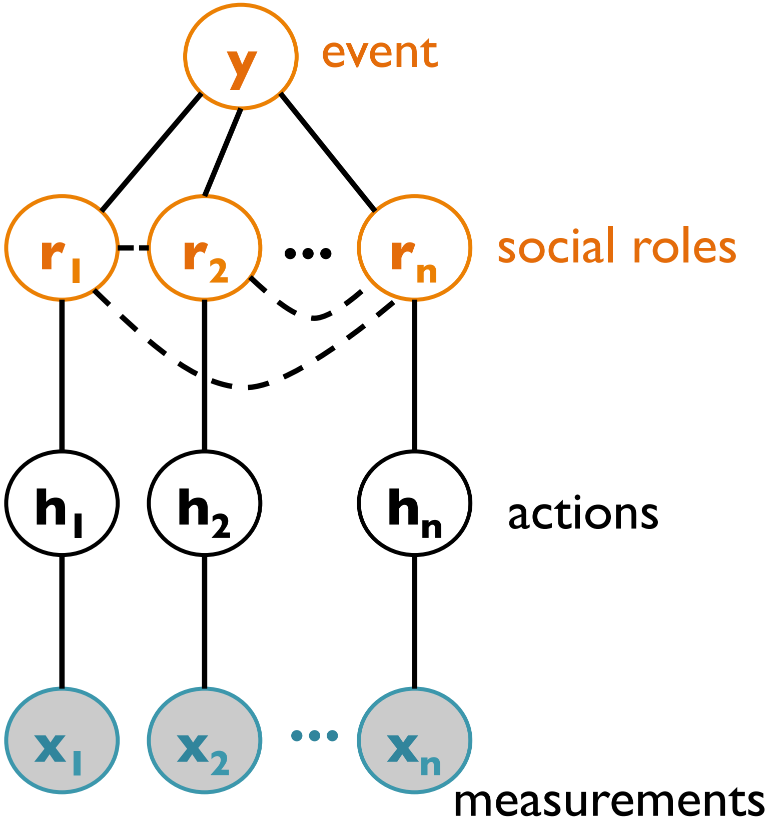

Event

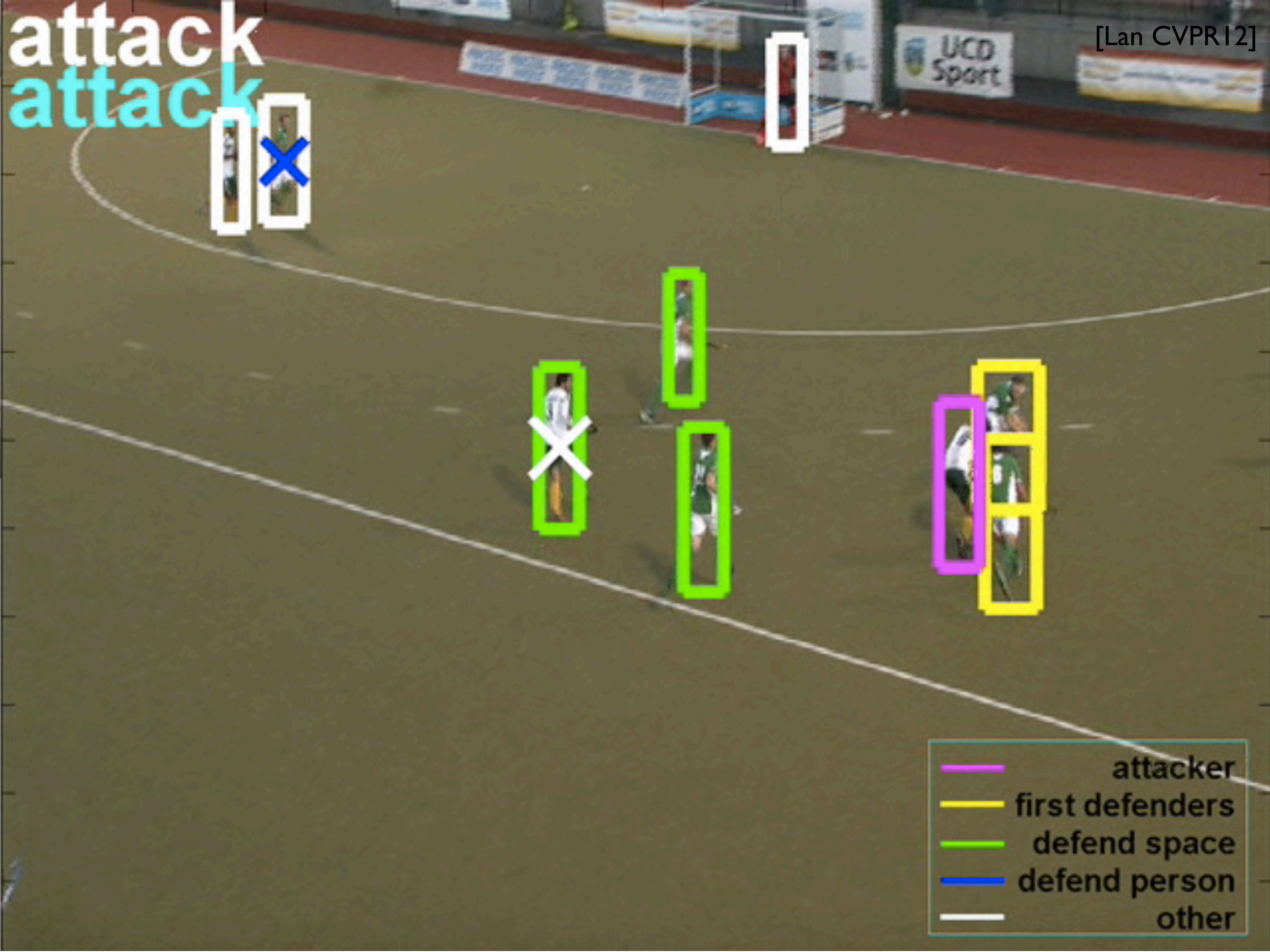
Interactions

Social Roles

Actions

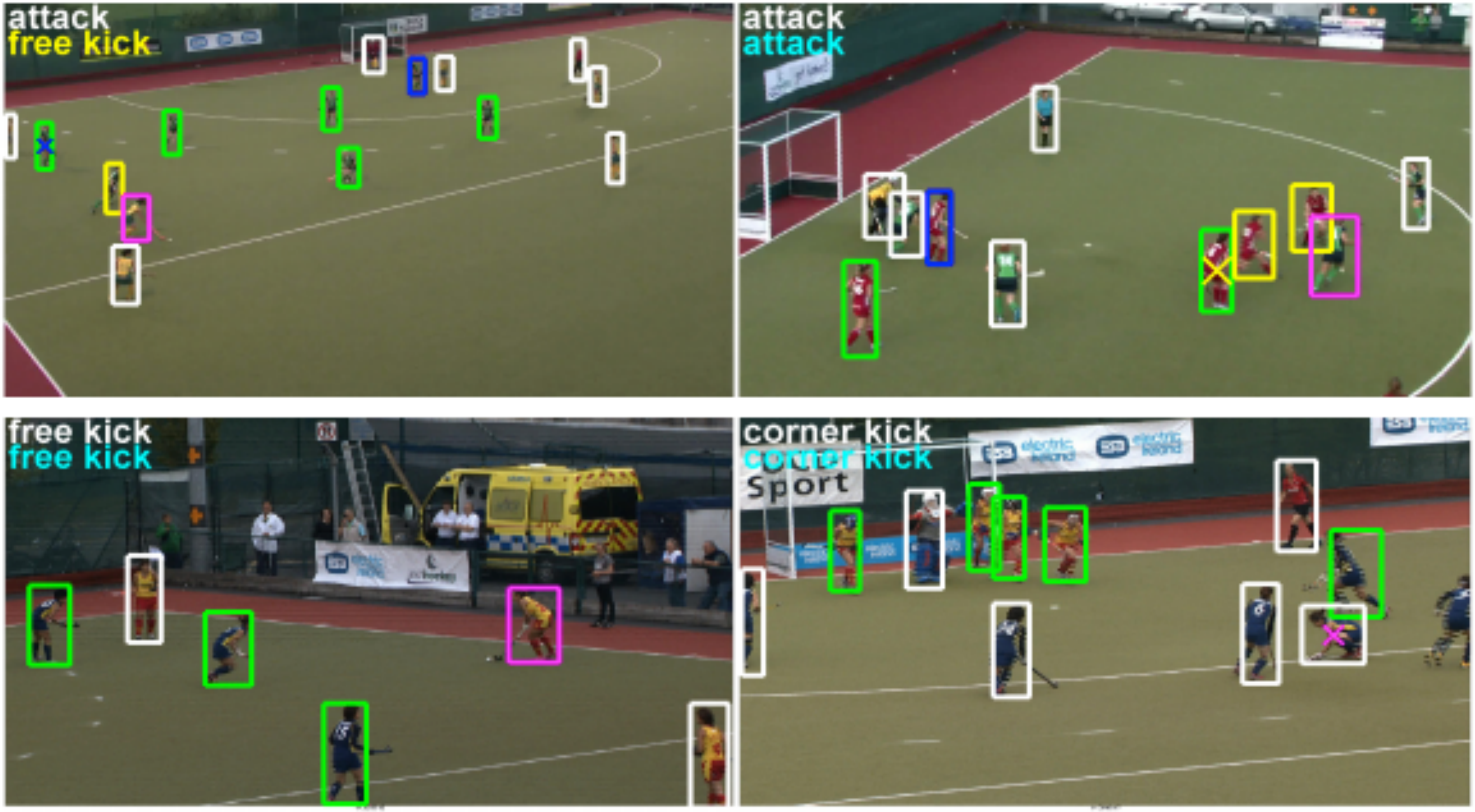
Features




attack
attack
attacker
first defenders
defend space
defend person
other

[Lan CVPR12]

[Lan CVPR12]

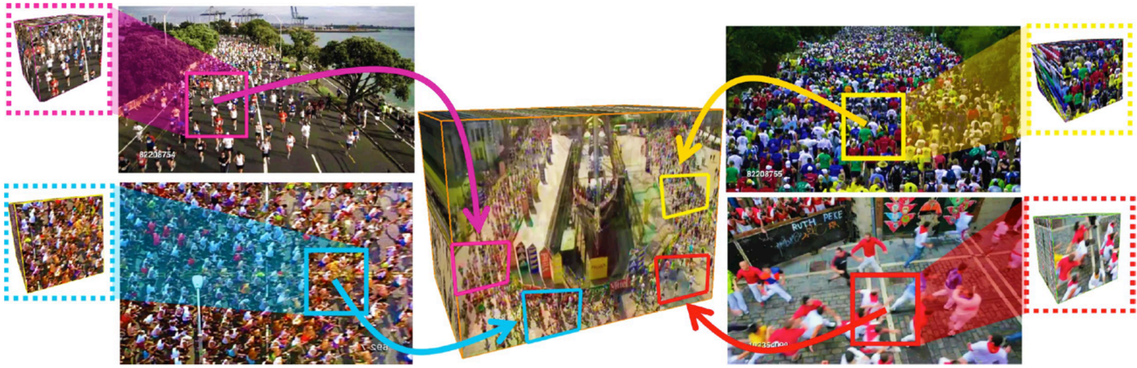


attacker, first defender, defend space, defend person, other

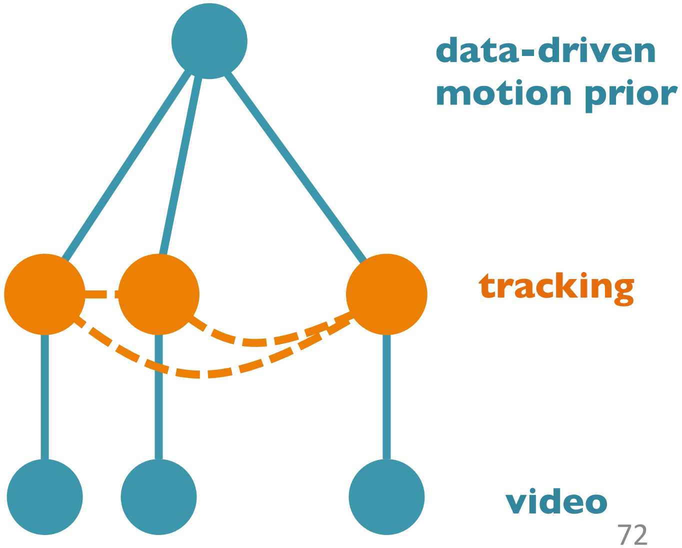
Data-driven Crowd Analysis (Rodriguez ICCV11)

Input: a crowd video

Output: individual tracking

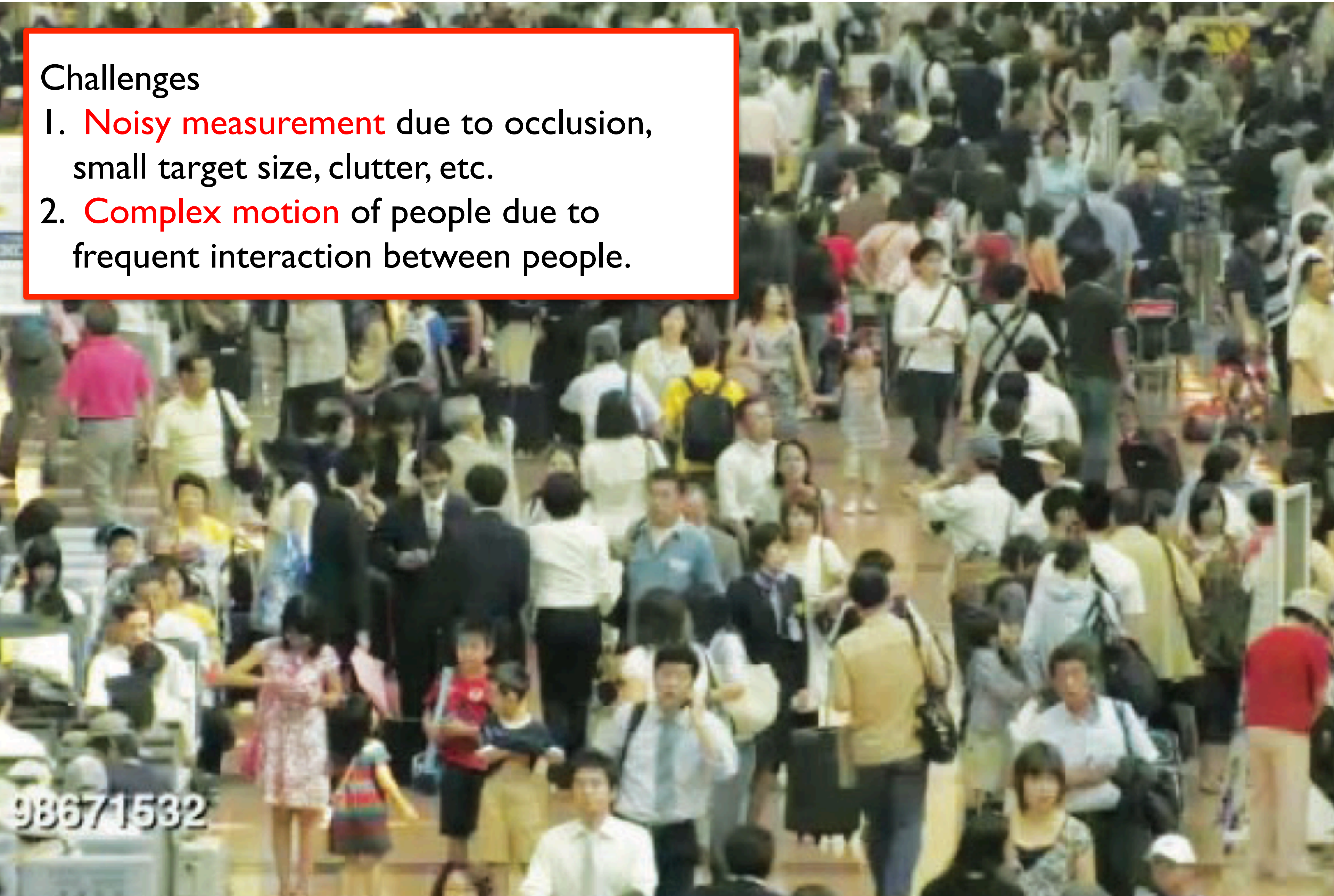


Crowd video



[Rodriguez ICCV11]

[Rodriguez ICCV11]



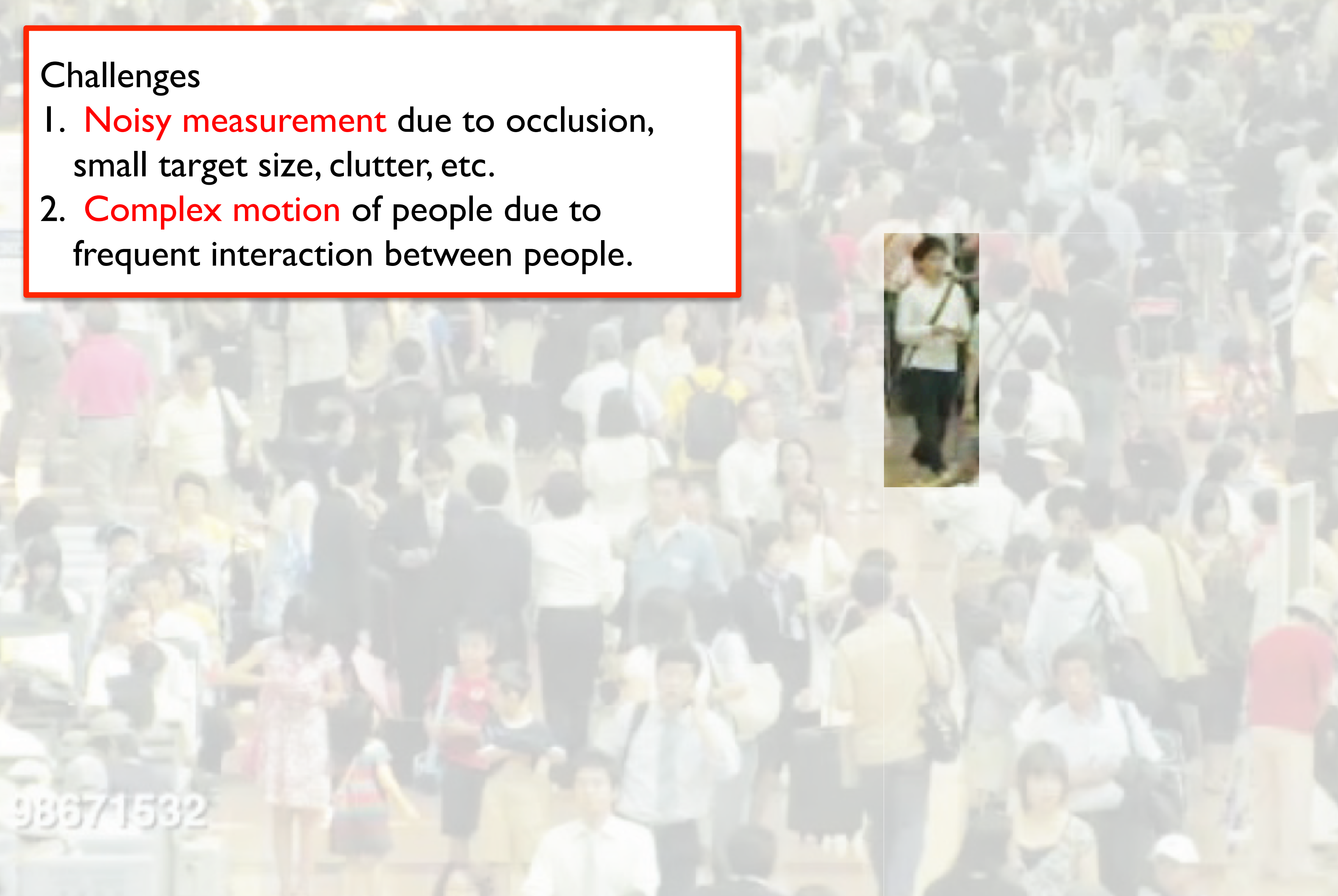
Challenges

1. Noisy measurement due to occlusion, small target size, clutter, etc.
2. Complex motion of people due to frequent interaction between people.

[Rodriguez ICCV11]

Challenges

1. Noisy measurement due to occlusion, small target size, clutter, etc.
2. Complex motion of people due to frequent interaction between people.



[Rodriguez ICCV11]

Challenges

1. Noisy measurement due to occlusion, small target size, clutter, etc.
2. Complex motion of people due to frequent interaction between people.

Solution

1. See larger are to encode collective signal.

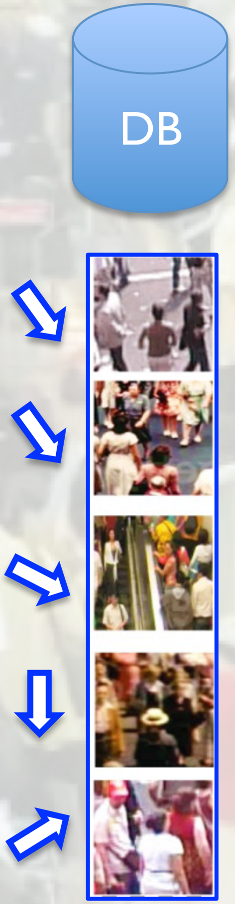
[Rodriguez ICCV11]

Challenges

1. Noisy measurement due to occlusion, small target size, clutter, etc.
2. Complex motion of people due to frequent interaction between people.

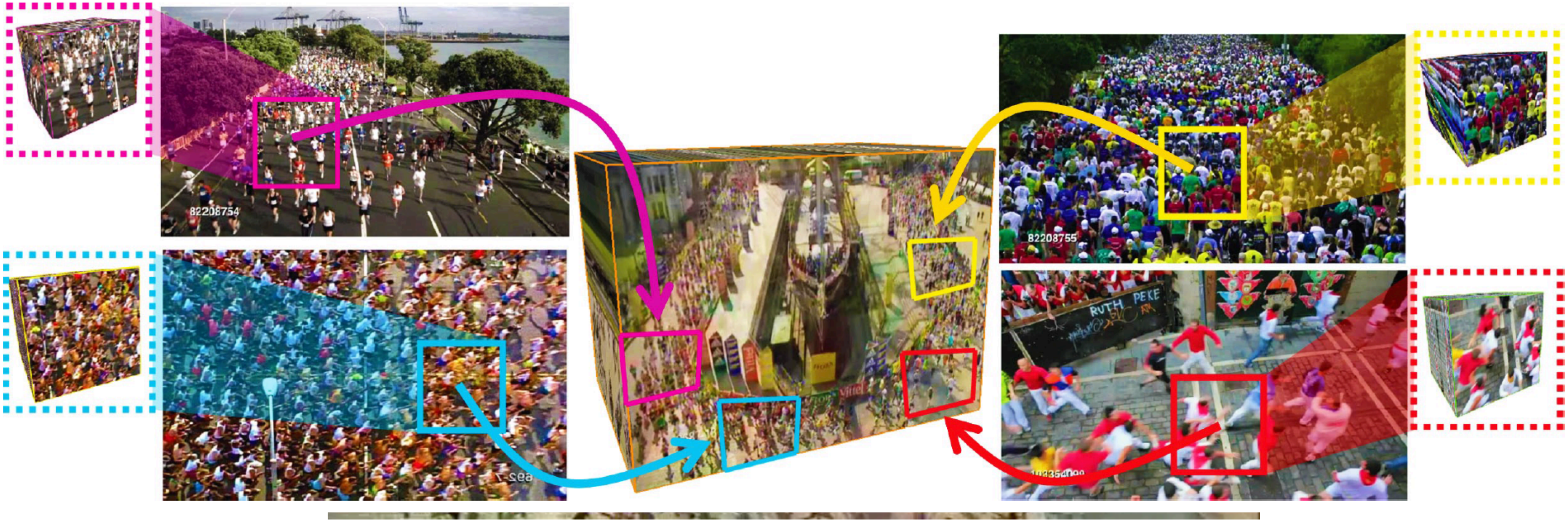
Solution

1. See larger are to encode collective signal.
2. Transfer the motion prior of data with similar collective signal.



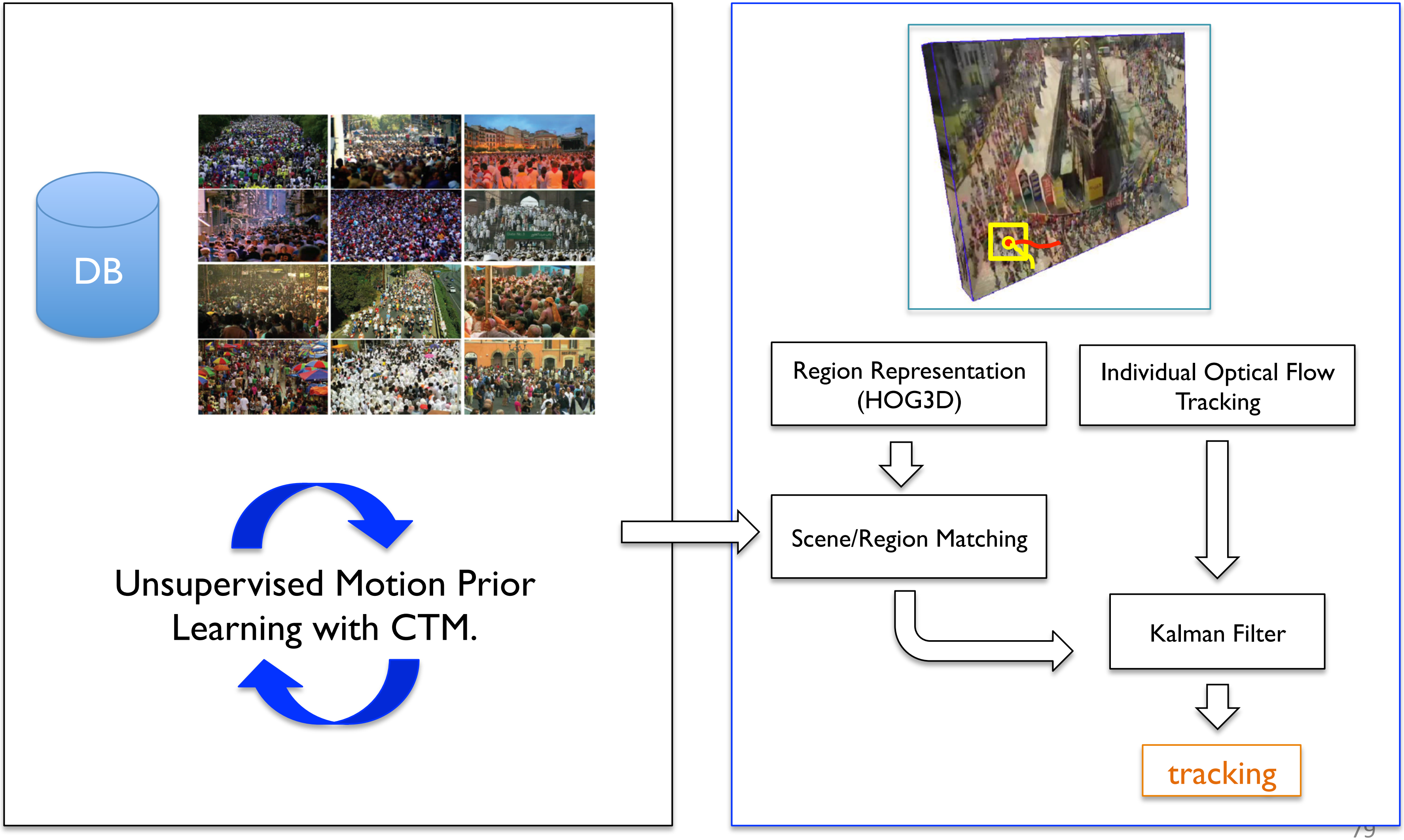
[Rodriguez ICCV11]

Mixture of Other Videos



[Rodriguez ICCV11]

Framework



DB

Unsupervised Motion Prior Learning with CTM.

Region Representation (HOG3D)

Individual Optical Flow Tracking

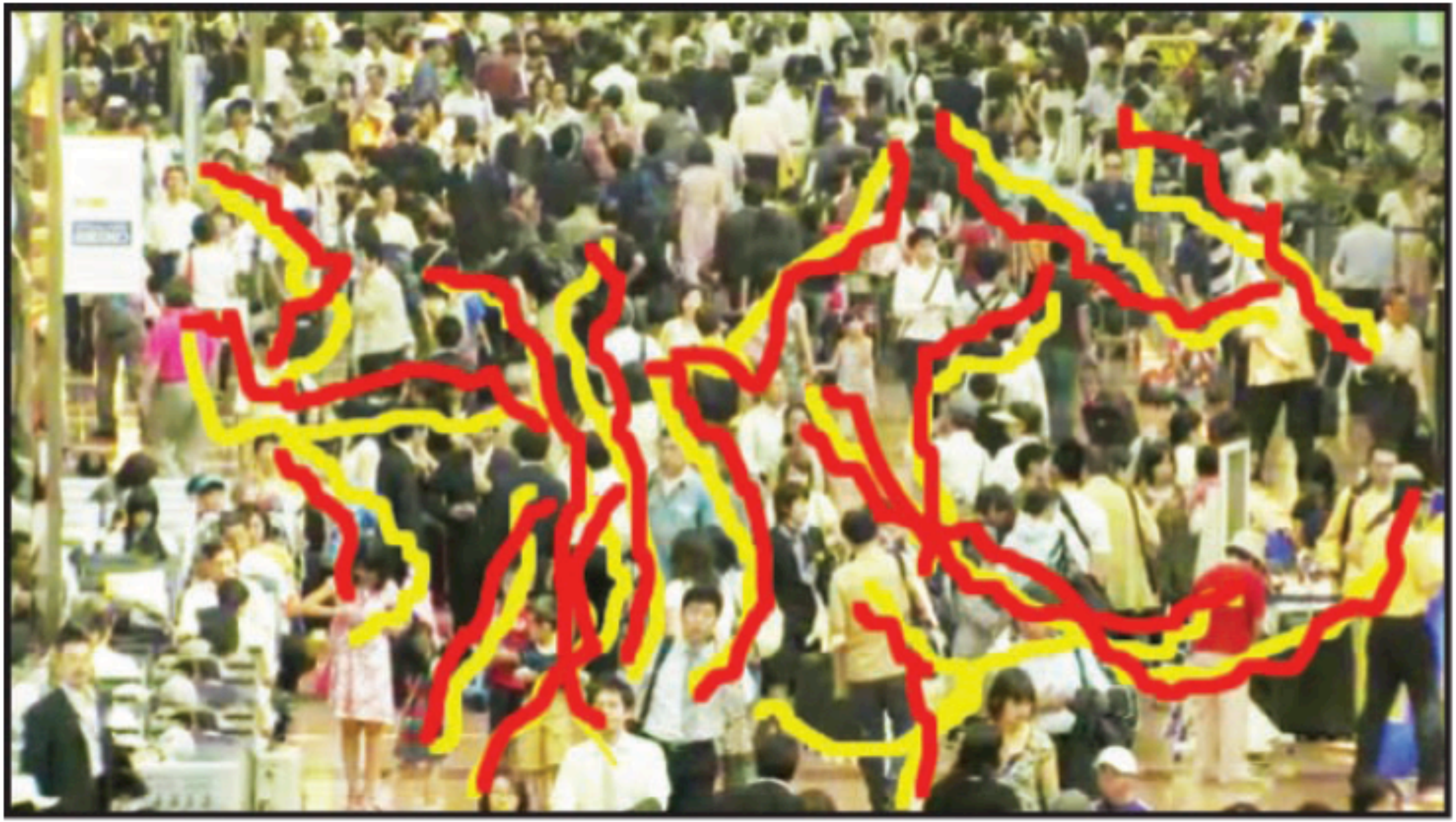
Scene/Region Matching

Kalman Filter

tracking

[Rodriguez ICCV11]

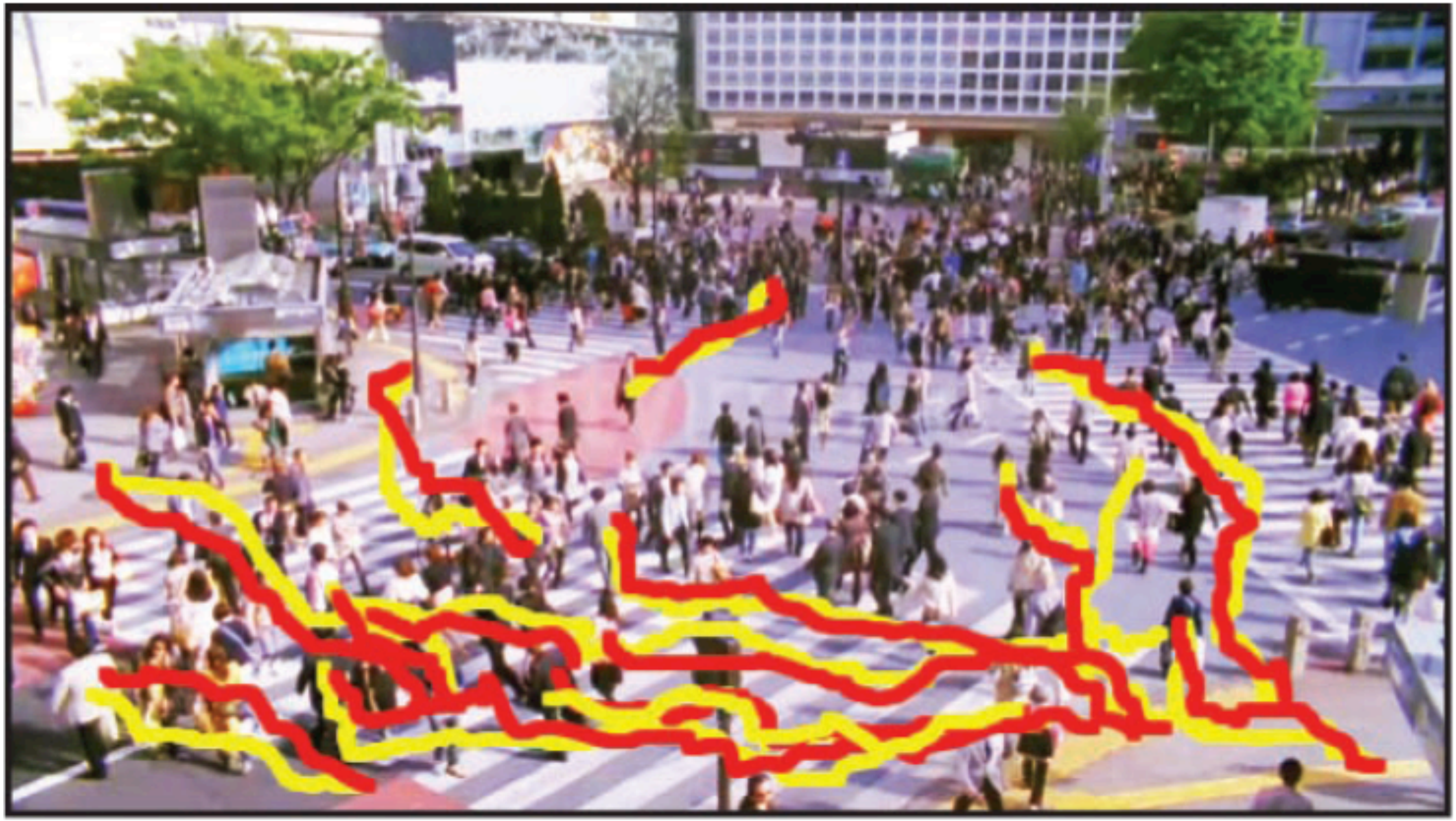
Tracking in Crowd Videos



Ground Truth, Tracking Results

[Rodriguez ICCV11]

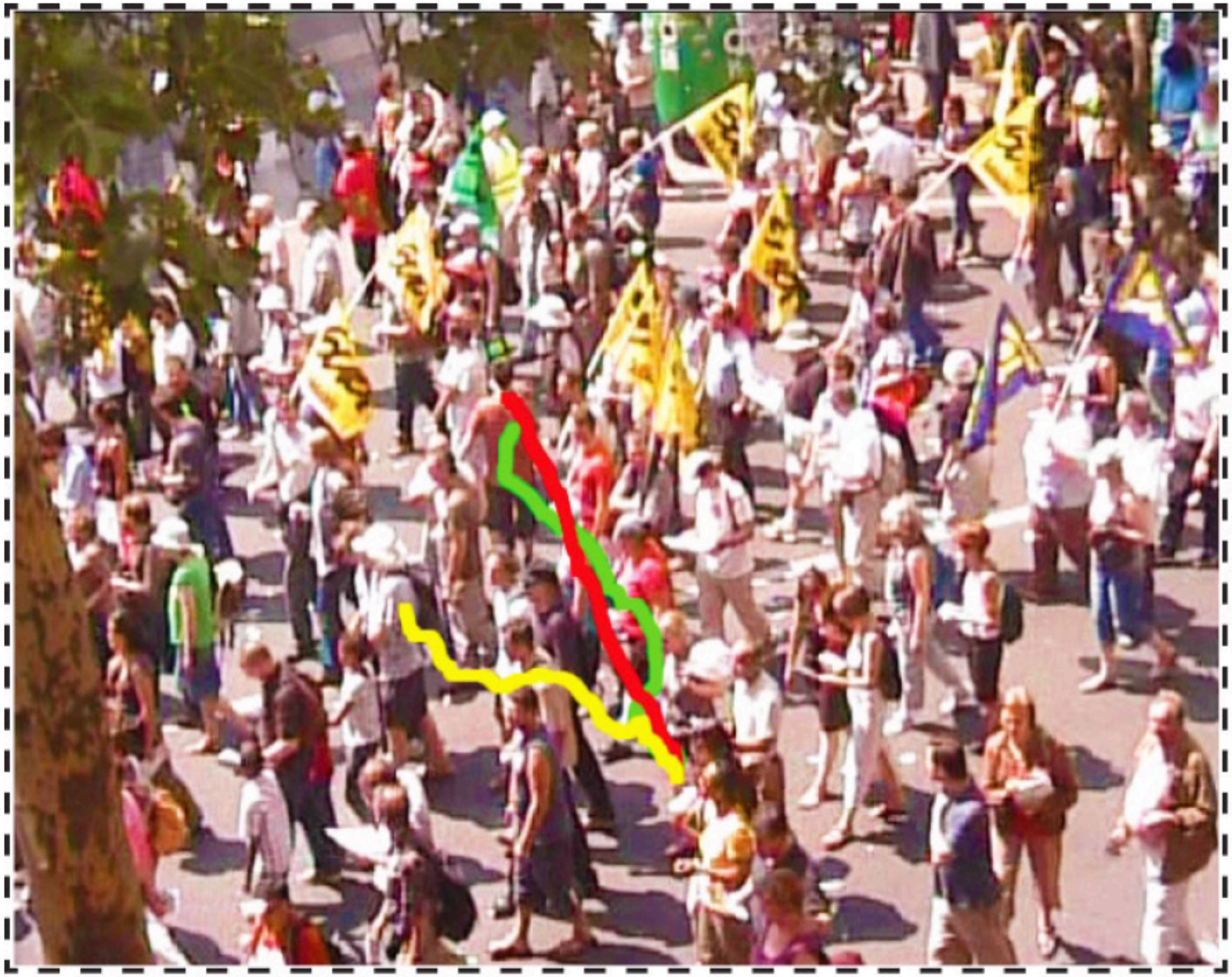
Tracking in Crowd Videos



Ground Truth, Tracking Results

[Rodriguez ICCV11]

Tracking in Rare Event



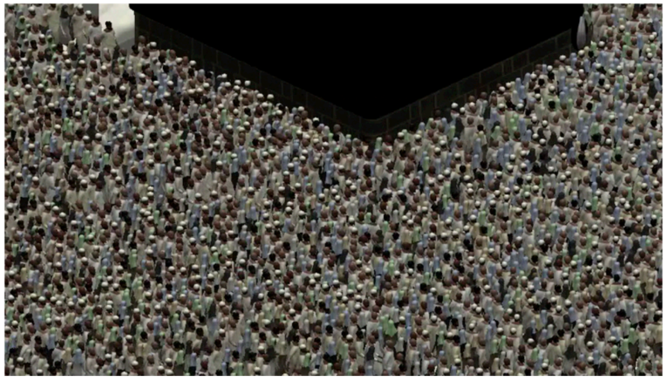
Ground-truth, Base-line, Data-driven racking Results

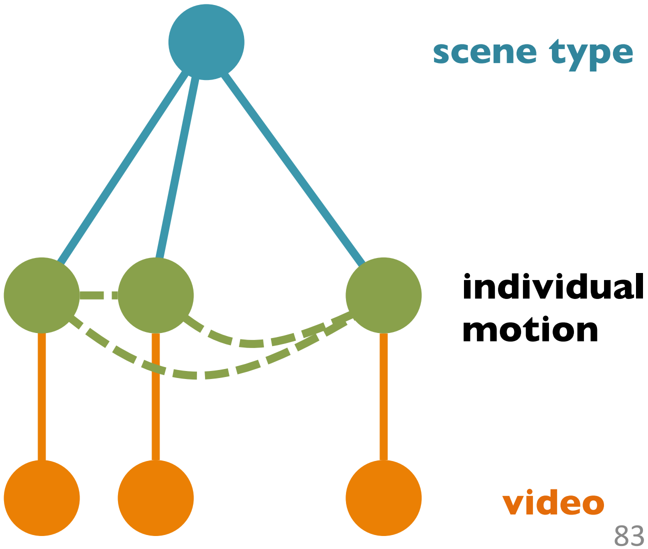
Application: Crowd Simulation [Curtis LC11]

Input: scene type

Output: video of a crowd

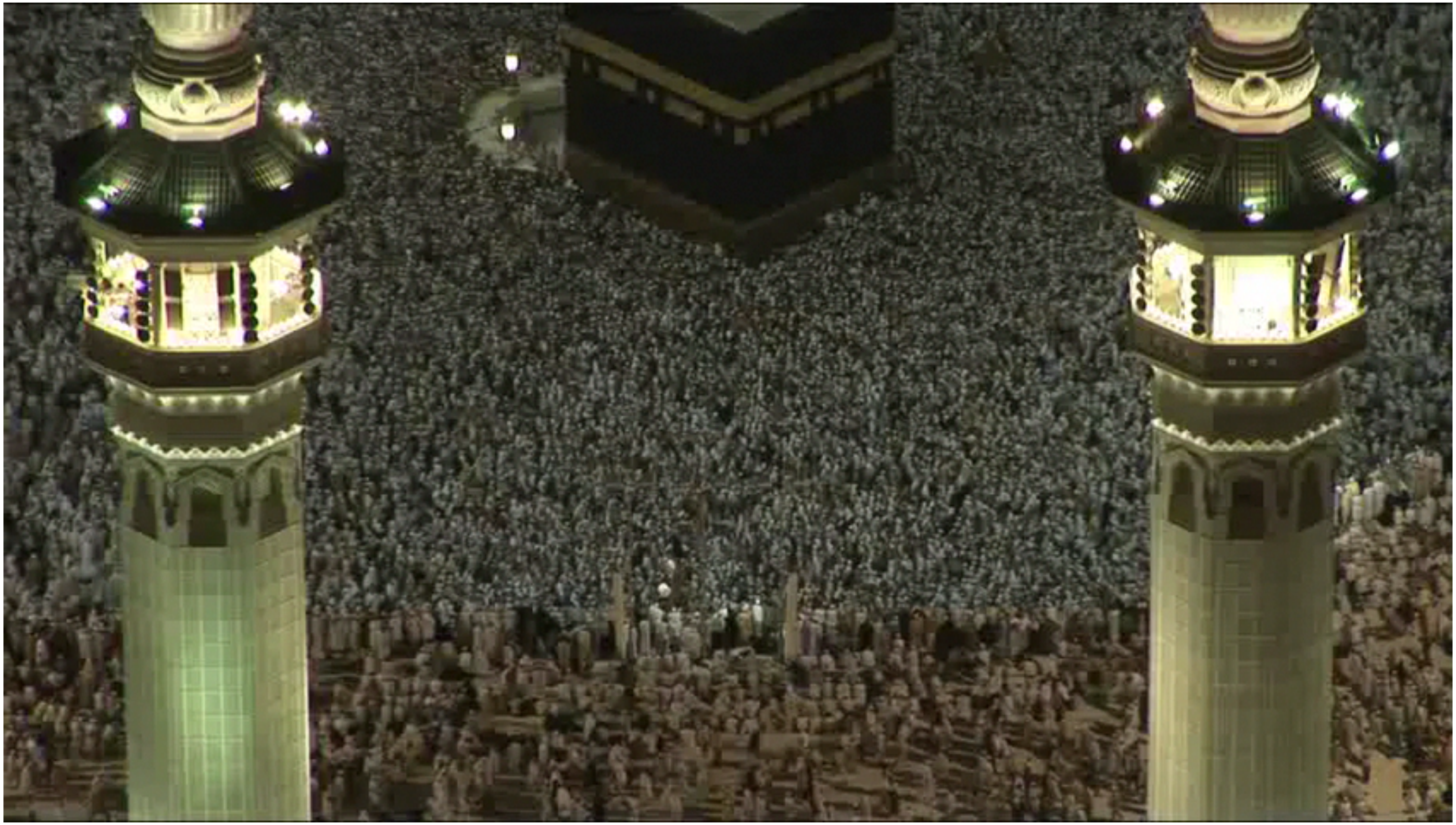
Model: Social force + FSM





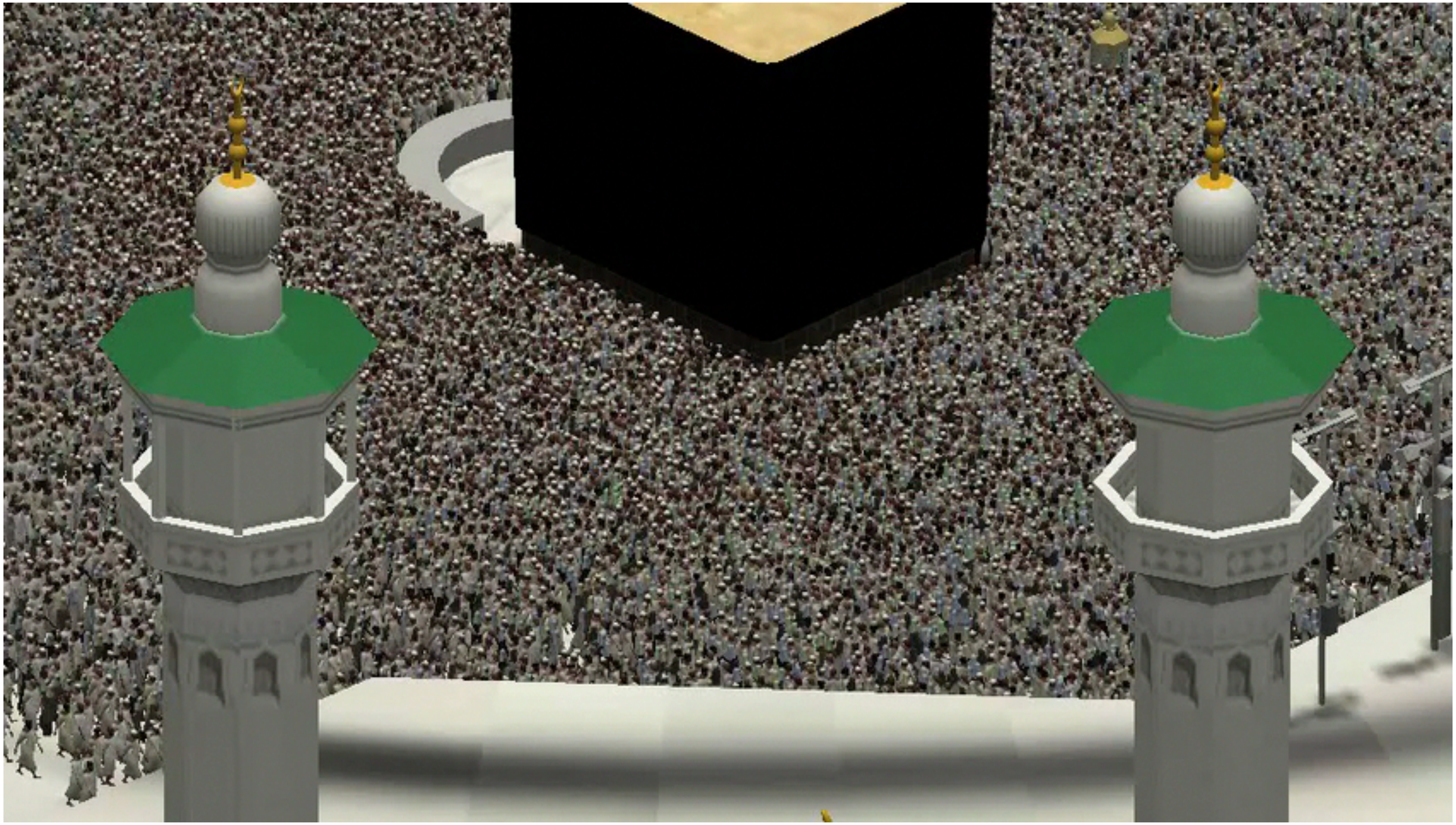
[Curtis LC11]

Real Video from Kaaba during Hajj



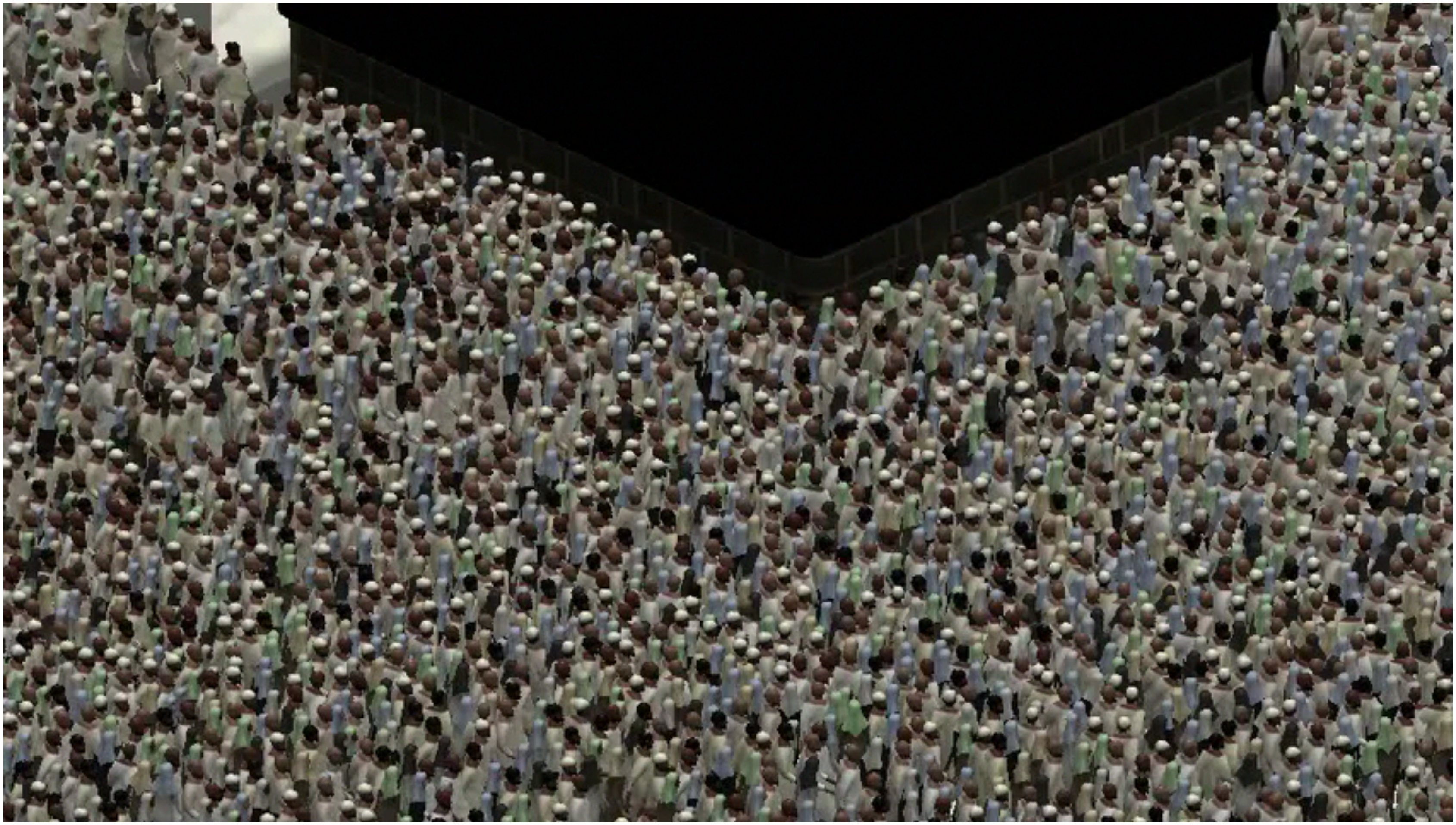
[Curtis LC11]

Simulation Results



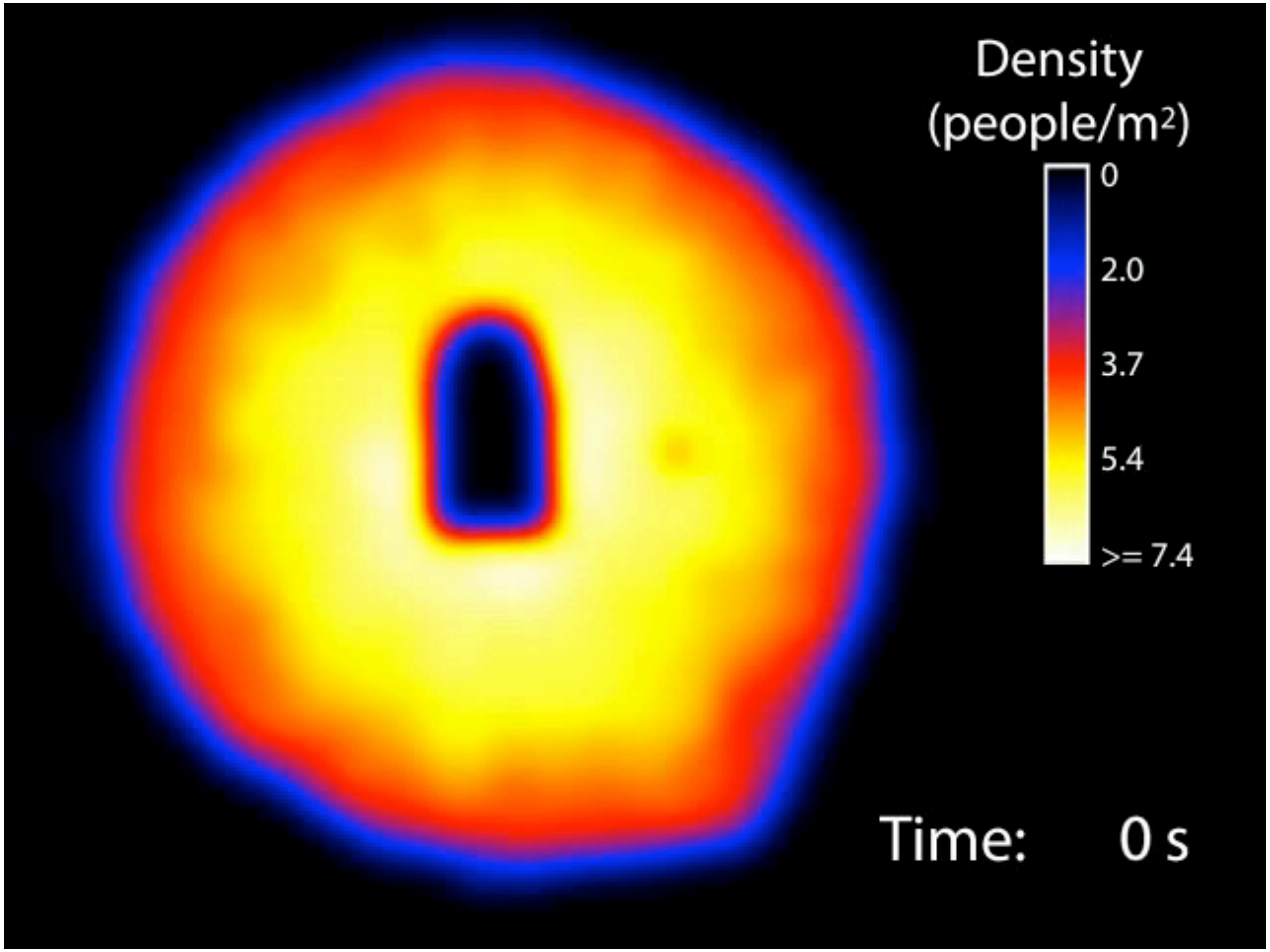
[Curtis LC11]

Simulation Result - Zoomed



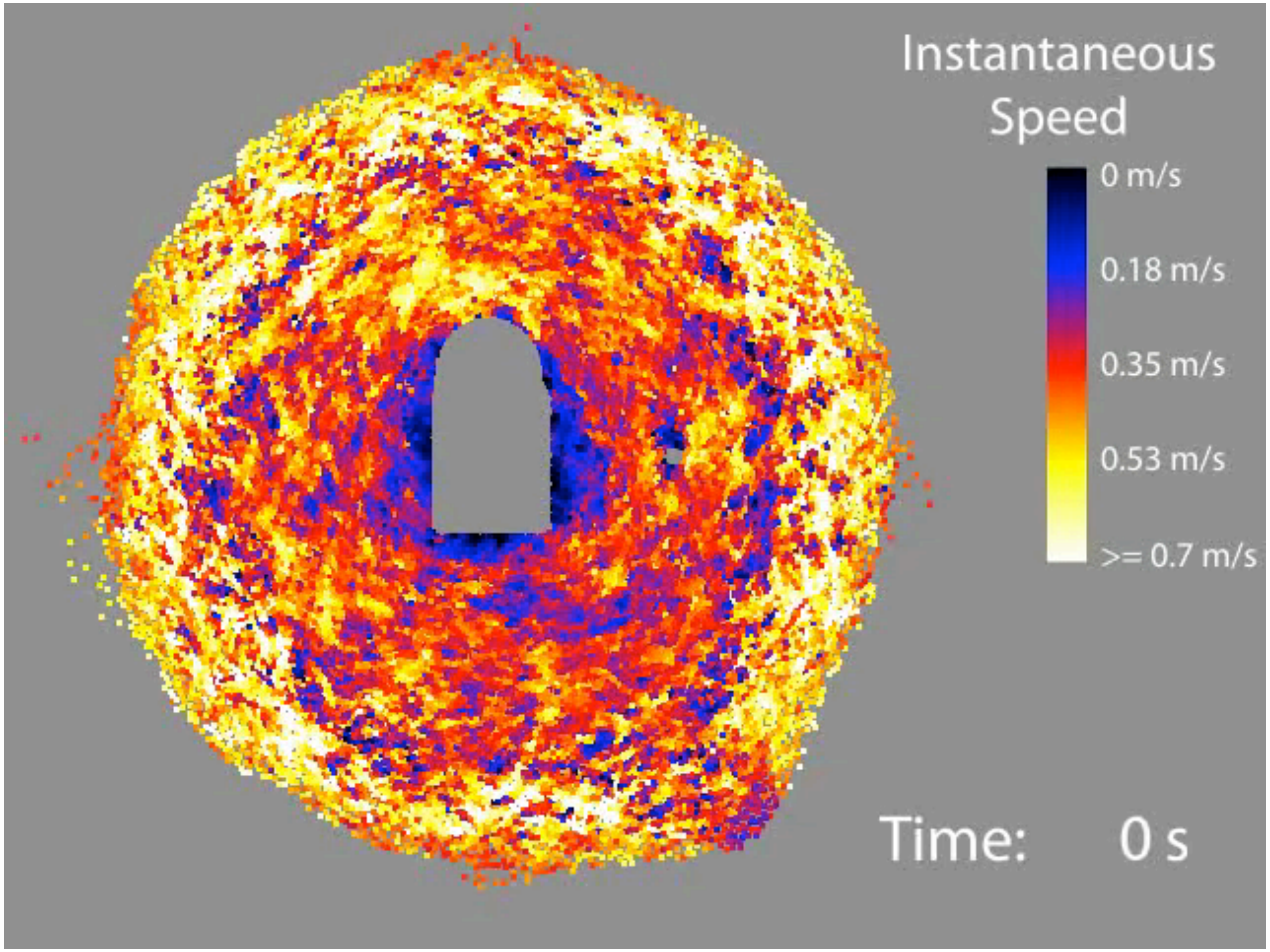
[Curtis LC11]

Density of People



[Curtis LC11]

Speed of People


Instantaneous Speed
0 m/s
0.18 m/s
0.35 m/s
0.53 m/s
>= 0.7 m/s
Time: 0 s