Data-driven Approaches in Social Dynamics
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

Collective Activity (C) “gathering”
Interaction (I) “approaching”
Atomic Activities (A) “facing-right” “walking”
Interaction (I) “approaching”
Tracklet(T)

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.

**Situations:**
1. Waiting
2. Queuing
3. 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 viewpoint and velocity.
  – Finite set of activity labels.
  – A bag of mid-level discriminative parts.
• Encode the context using a spatio-temporal descriptor.
Individual Posture Representation

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

Simple and well-studied.
Require the definition and annotation of the posture space.

Dalal and Triggs, CVPR 2005
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.
Encoding Context: Action Context

\[ C_i = \left[ \max_{j \in \mathcal{N}_1(i)} S_{1j}, \ldots, \max_{j \in \mathcal{N}_1(i)} S_{Kj}, \ldots, \max_{j \in \mathcal{N}_M(i)} S_{1j}, \ldots, \max_{j \in \mathcal{N}_M(i)} S_{Kj} \right] \]
Encoding Context: Spatio-Temporal Local (STL) Descriptor

Spatio-Temporal Local (STL) Descriptor

Frequency

Bin

e.g. SVM

[Choi VSWS09]
Encoding Context: Learning the Contextual Model

$P_1(c)$
Learning the Contextual Model

(a) Waiting
(b) Talking

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

Anchors are looking upward.
Collective Activity Dataset

- 44 videos with multiple people.
- Crossing, Waiting, Queuing, Walking, Talking.
- Leave-One-Video-Out.
Atomic Activity Feature v.s. Crowd Context

Classification Accuracy

- STIP (Baseline)

31.8
Qualitative Examples

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

[Choi CVPR11]
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.
Hierarchical Activity Model

Input: video with tracklets
Hierarchical Activity Model

[Choi ECCV12]
Hierarchical Activity Model

\[
\Psi(C, I, A, O, f) = \\
\Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) + \\
\Psi(C) + \Psi(I) + \Psi(A) - c^T f, \quad f \in S
\]
Atomic-Observation Potential

\[ \Psi(C, I, A, O, f) = \ \Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) + \Psi(C) + \Psi(I) + \Psi(A) - c^T f, \ f \in S \]

Atomic Activity Models

• Action: BoW with STIP
• Pose: HoG

[Dollar et al, 06; Niebles et al, 07]

[Dalal and Triggs, 05]
Interaction-Atomic Potential

\[ \Psi(C, I, A, O, f) = \Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) + \Psi(C) + \Psi(I) + \Psi(A) - c^T f, \quad f \in S \]

**I: Standing-in-a-line**

Model
Interaction-Atomic Potential

\[ \Psi(C, I, A, O, f) = \Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) + \Psi(C) + \Psi(I) + \Psi(A) - c^T f, \quad f \in S \]

I: Standing-in-a-line

[Choi ECCV12]
Collective-Interaction Potential

\[
\Psi(C, I, A, O, f) = \\
\Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) + \\
\Psi(C) + \Psi(I) + \Psi(A) - c^T f, \ f \in S
\]

[C: Queuing]
Collective-Interaction Potential

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

[C: Queuing]

[Choi ECCV12]
Collective-Observation Potential

\[ \Psi(C, I, A, O, f) = \Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) + \Psi(C) + \Psi(I) + \Psi(A) - c^T f, \ f \in S \]

Collective Activity
- STL of all targets

Choi et al., 09
Activity Transition Potential

\[
\Psi(C, I, A, O, f) = \\
\Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) + \\
\Psi(C) + \Psi(I) + \Psi(A) - c^T f, \ f \in S
\]

Smooth activity transition
Trajectory Estimation

\[
\Psi(C, I, A, O, f) = \\
\Psi(A, O) + \Psi(I, A, f) + \Psi(C, I) + \Psi(C, O) + \\
\Psi(C) + \Psi(I) + \Psi(A) - c^T f, \ f \in S
\]
Training the Graphical Model

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

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

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

Tsochantaridis et al, 2004
Example Classification Result

Interaction labels
AP: approaching
FE: facing-each-other
SR: standing-in-a-row
...
Example Classification Result

Atomic Activities
Action:
W - walking
S – standing

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

[Choi ECCV12]
Example Classification Result

Pair-Interactions
- AP: approaching
- FE: facing-each-other
- SS: standing-side-by-side
- SQ: standing-in-a-queue
Example Classification Result
Learning Latent Constituents for Group Activity Recognition (Antic ECCV14)

Input: video and box tracks
Output: group activity labels in time
Meaningful Parts of Group Behavior
Meaningful Parts of Group Behavior
Meaningful Parts of Group Behavior
Less Meaningful Parts

Less meaningful parts
Learning Mid-level Constituents

\[ \Delta(f_i, f_v) := d(f_i, f_v) + \lambda_x \| x_i - x_v \| + \lambda_s | s_i - s_v |. \]
Encoding Social Signal with Latent Constituents

constituent 1

constituent 2

[Antic ECCV14]
Behavior Recognition with Constituents
Behavior Recognition with Constituents
Behavior Recognition with Constituents
Behavior Recognition with Constituents
Quantitative Evaluation

Holistic approach (full b-boxes):

Latent constit’s (functional grouping):

5 Activities Dataset
6 Activities Dataset

70.4% 83.3%
75.1% 90.1%
Input: videos with event labels, box tracks
Output: groups with activity label
Humans in Social Setting

- Bride gets ring from bridesmaid
- Priest claps
- Bride and groom exchange rings
- Groom gets ring from groomsman
- Bride and groom kiss

Slide from Ramanathan et al, 2013 [Ramanathan CVPR13]
Goal: Identify social roles

Slide from Ramanathan et al, 2013
Problem setup

- person-specific role features
- inter-role interaction features

[Ramanathan CVPR13]
Social role model
Social role model

Person-specific features

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

[Image: Diagram of social role model with nodes labeled \( \Sigma \), \( \Omega \), \( \alpha \), \( \beta \), \( s_1 \), \( s_2 \), \( u_1 \), \( u_2 \), \( P_1 \), \( P_2 \), and \( P_{12} \).]
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

[Reference: Ramanathan CVPR13]
Social role model

Person-specific features
Inter-role features
Reference role

Model parameters

\( \alpha \) is the person-specific feature weight

\( \beta \) is the inter-role feature weight
Social role model

Person-specific features
Inter-role features
Reference role

**Model parameters**

- $\alpha$ is the person-specific feature weight
- $\beta$ is the inter-role feature weight
- $\Sigma$ Gaussian priors for regularization

---

[Ramanathan CVPR13]
Social role model

Variational Inference

Jointly

Learn model parameters.
Assign social roles.
Dataset

- Youtube Social Role Dataset
  - Only the event type is provided.
  - Social roles are discovered in unsupervised fashion.

[Ramanathan CVPR13]
Results: role clusters
Results: role clusters
Results: role clusters

- Bride
- Bridesmaid
- Priest
- Grooms men
- Groom (Reference)
Spatial Relations

Bride

Bridesmaid

Priest

Groom

Grooms men
Spatial Relations

Bride

Brides maid

Priest

Grooms men

Groom

reference
Results – Role clusters

Slide from Ramanathan et al, 2013
Social Role Discovery (Lan CVPR12)

Input: a video with box tracks
Output: social role and activity labels

Free Hit
Attacker
Defender

Event
Interactions
Social Roles
Actions
Features

Event
social roles/actions
video/boxes
Semantic Description of Videos

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

actions
walk  
run  
jog  
bend  
shoot  
dribble  

Social roles
attacker 
first defender 
man-marking 
defend-space 
Teammate  

event
corner-hit  
free-hit  
attack play  

First defenders

Semantic Description of Videos

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

Social Roles

Actions

Features

Event

Interactions

$$y$$
event

$$r_1$$ social roles

$$r_2$$

$$\ldots$$

$$r_n$$

$$h_1$$

$$h_2$$ actions

$$\ldots$$

$$h_n$$

$$x_1$$ measurements

$$x_2$$

$$\ldots$$

$$x_n$$

[Lan CVPR12]
attacker, first defender, defend space, defend person, other
Data-driven Crowd Analysis (Rodriguez ICCV11)

Input: a crowd video
Output: individual tracking
Collective signal => help understand the videos better.

[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.
Challenges

1. **Noisy measurement** due to occlusion, small target size, clutter, etc.
2. **Complex motion** of people due to frequent interaction between people.
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 area to encode **collective signal**.
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 area to encode **collective signal**.
2. Transfer **the motion prior of data** with similar collective signal.
Mixture of Other Videos
Framework

Unsupervised Motion Prior Learning with CTM.

Region Representation (HOG3D)

Individual Optical Flow Tracking

Scene/Region Matching

Kalman Filter

tracking
Tracking in Crowd Videos

Ground Truth, Tracking Results
Tracking in Crowd Videos

Ground Truth, Tracking Results
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
Real Video from Kaaba during Hajj
Simulation Results
Simulation Result - Zoomed
Density of People

Density (people/m²)

Time: 0 s
Speed of People