Social Roles in Hierarchical Models for Human Activity Recognition

Greg Mori
School of Computing Science
Simon Fraser University

CVPR POCV Workshop
June 23, 2014

Joint work with: Tian Lan and Leonid Sigal
What does activity recognition involve?
Detection: are there people?
Action recognition

Structures

stand
run
fall
squat
Group activity recognition

help the fallen person
Intention/social role

- watch
- get help
- comfort
Role of Context in Actions

Is this a fallen person?
Role of Context in Actions

Who has the puck?
### Semantic Descriptions of Videos

#### Actions
- walk
- run
- jog
- bend
- shoot
- dribble
- pass

#### Social Roles
- attacker
- first defenders
- 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.
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?

Social Roles
- attacker
- man-marking
- teammate
System Overview
Activity Hierarchy Model Representation

$Y$: Corner hit, Attack play

$r$: Attacker, Man-marking

$h$: Pass, jog

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

Video evidence

Event

Social Role

Action

$\bigcirc\ h_1\ r_1\ \ldots\ r_N\ h_N\ \ldots\ h_2\ r_2\ \ldots\ r_1\ \bigcirc$

Lan, Sigal, and Mori CVPR 2012
Model Learning

\[ \Psi = \sum_{e \in E} w_e \psi_e \]

\[
\Psi = \sum w_e \psi_e
\]

\[
\psi_e \in E
\]

\[
\sum_{i} \xi_i
\]

\[
\sum \xi_i
\]

\[
\psi = \psi_e \in E
\]

\[
\psi \in E
\]

\[
\Psi = \sum \psi_e \in E
\]

Query for event: \( \text{loss} = \Delta(y, y_i) \)

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

Query for social roles: \( \text{loss} = \Delta(r, r_i) \)

Query for actions: \( \text{loss} = \Delta(h, h_i) \)

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

\[
\min_{\|w\|_2^2, \xi} \frac{1}{2} \sum_{i} \xi_i
\]

\[
\min \frac{1}{2} \|w\|_2^2 + \beta \sum \xi_i
\]

s.t. \( \forall i, y, r, h \)

\[
w_{y_i} h_i \cdot \psi_i - w_{yrh} \cdot \psi_i = \text{loss} - \xi_i
\]

\[\forall i, \xi_i \geq 0\]
Model Inference
- Query

Event, social roles, actions, queries

Score: \( \Psi(Y^*, \{r_{1,n}^*\}, \{h_{1,n}^*\}, q) \)

max \( \sum_{y,r,h \neq q} w_e \psi_e \)

The learned models

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

Person detection and tracking

Coordinate ascent inference

- Query
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)

### Diagram

- **Unary**
  - Event
  - Social Role
  - Action

- **Full model**
  - Event
  - Social Role
  - Action

- **Group activity**
  - Event

### Table

<table>
<thead>
<tr>
<th>Method</th>
<th>Action</th>
<th>Role</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unary</td>
<td>40.9</td>
<td>35.0</td>
<td>73.2</td>
</tr>
<tr>
<td>Full model</td>
<td>42.0</td>
<td>50.1</td>
<td>80.5</td>
</tr>
<tr>
<td>Action model (HOG+SVM)</td>
<td>38.7</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Group activity [Lan et al. PAMI 12]</td>
<td>N/A</td>
<td>N/A</td>
<td>78.5</td>
</tr>
</tbody>
</table>
Results – Query for Social Roles (Nursing Home)
ESPN Broadcast Field Hockey Data

- 58 videos, 11 actions, 5 social roles, 3 scene-level events
# Results – Scene Labeling

## Network Structures

- **Unary**
  - Event
  - Social Role
  - Action

- **Full model**

<table>
<thead>
<tr>
<th>Method</th>
<th>Action</th>
<th>Role</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>unary</td>
<td>21.5</td>
<td>21.7</td>
<td>56.9</td>
</tr>
<tr>
<td>Full model</td>
<td>28.8</td>
<td>44.0</td>
<td>62.8</td>
</tr>
<tr>
<td>action model (HOG+SVM)</td>
<td>26.1</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Results – Query for Social Roles
Conclusion

Structural Recognition of Human Activities
Acknowledgements

Tian Lan
BACKUP SLIDES
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

**Image evidence**
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 $h_j$

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 Clique weight potential

Activity

Obtained by structure learning

Image evidence

• 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.
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

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

Clique Clique weight potential

Image evidence

Activity

Action

$\psi_e(Y, x_0)$

$\psi_e(h_i, x_i)$
Model Learning

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

**Goals:**

**Input:**
- $Y$: talk
- $h$: stand-right, stand-left

**Diagram:**
- Nodes represent activities and actions.
- $x_0, x_1, \ldots, x_N$ are input variables.
- $h_1, h_2, \ldots, h_N$ are hidden variables.
- $y$ is the output variable.
Model Learning

\[ \Psi = \sum_{e \in E} w_e \psi_e \]

**Goals:**

**Structural connectivity (hidden human-human interactions)**

Potential weights

**Input:**

- **Y:** talk
- **h:**
  - stand-left
  - stand-right
  - stand-right
  - stand-left
Model Learning

\[ \Psi = \sum_{e \in E} w_e \psi_e \]

\( \Psi \): talk

**Input:**

**Y:** talk

**Input:**

**h:**

stand-left

stand-left

stand-right

stand-right

**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
  \[ \max_{E = \{e\}} \sum_{e} w_e \psi_e \]
The learned models

Model Inference

\[ \Psi \left( Y^*, e^*, \{ h_{1,n}^* \}_n \right) \]

Activity, interactions, actions

Coordinate ascent inference

Person detection

Talk

stand-left

stand-right

s-r

s-r

\( I \)
Activity Hierarchy Model Representation

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