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
What does activity recognition involve?
Detection: are there people?
Action recognition

Structures

stand
run
fall
squat
Group activity recognition

help the fallen person
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
    - Intille & Bobick, 2001
    - Medioni et al, 2001
    - Loy et al, 2010
    - Lan et al, 2012
    - Amer et al, 2012

Number of People
Activity landscape

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

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

Talk

group-person interaction

person-person interaction
Model of Group Activities

$Y$: Talk, Queue

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

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

Activity

Action

Image evidence

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 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 $h_j$

Markov Random Field

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

- Clique weight potential

Clique

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 $h_j$
  - Learn structural connectivity among the actions.

Markov Random Field

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

Clique Clique weight potential

Obtained by structure learning
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 Clique weight potential
Model Learning

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

Goals:

Input:

\[ Y: \text{ talk} \]

\[ h: \]

stand-right

stand-left

stand-right

stand-left
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-left
  - stand-right
Model Learning

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

**Goals:**

Structural connectivity (hidden human-human interactions)

**Potential weights**
Model Learning

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

Goals:

- Structural connectivity
- Potential weights

Approach:

$$\max_{E=\{e\}} \sum_e w_e \psi_e$$
Model Learning

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

**Goals:**
- Structural connectivity
- Potential weights

**Notation**
- \( \psi_i \): Potential values of the \( i \)-th image.
- \( w_r \): Potential weights of the \( r \)-th activity.
- \( y(r) \): \( r \)-th activity class.
- \( \xi_i \): A slack variable for the \( i \)-th image.

**Approach:**
Max-margin learning
\[
\min_{\Psi, \xi} \frac{1}{2} \sum_{r} \|w_r\|_2^2 + \beta \sum_{i} \xi_i
\]

s.t. \( \forall i, r \) where \( y(r) \neq y(c_i) \),
\[
 w_{c_i} \cdot \psi_i - w_r \cdot \psi_i \geq 1 - \xi_i
\]
\( \forall i, \xi_i \geq 0 \)
Model Inference

The learned models

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

Activity, interactions, actions

coordinate ascent inference

Person detection
Visualization of the Results

crossing

waiting

queuing

walking

talking
Baselines

- SVM
- No connection
- Min-spanning tree
- $\varepsilon$-neighborhood graph
### Results – Collective Activity Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Overall</th>
<th>Mean per-class</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>70.9</td>
<td>68.6</td>
</tr>
<tr>
<td>no connection</td>
<td>75.9</td>
<td>73.7</td>
</tr>
<tr>
<td>min-spanning tree</td>
<td>73.6</td>
<td>70.0</td>
</tr>
<tr>
<td>(\varepsilon)-neighborhood graph, (\varepsilon=100)</td>
<td>74.3</td>
<td>72.9</td>
</tr>
<tr>
<td>(\varepsilon)-neighborhood graph, (\varepsilon=200)</td>
<td>70.4</td>
<td>66.2</td>
</tr>
<tr>
<td>(\varepsilon)-neighborhood graph, (\varepsilon=300)</td>
<td>62.2</td>
<td>62.5</td>
</tr>
<tr>
<td>complete graph</td>
<td>62.6</td>
<td>58.7</td>
</tr>
<tr>
<td>our approach</td>
<td>79.1</td>
<td>77.5</td>
</tr>
</tbody>
</table>
Nursing Home Data

• 22 short clips of fall + a 30-min non-fall clip, 5 actions, 2 group activities
<table>
<thead>
<tr>
<th>Method</th>
<th>Overall</th>
<th>Mean per-class</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>48.0</td>
<td>52.4</td>
</tr>
<tr>
<td>no connection</td>
<td>54.4</td>
<td>56.1</td>
</tr>
<tr>
<td>min-spanning tree</td>
<td>66.9</td>
<td>62.3</td>
</tr>
<tr>
<td>$\varepsilon$-neighborhood graph, $\varepsilon=100$</td>
<td>72.7</td>
<td>61.3</td>
</tr>
<tr>
<td>$\varepsilon$-neighborhood graph, $\varepsilon=200$</td>
<td>67.6</td>
<td>61.1</td>
</tr>
<tr>
<td>$\varepsilon$-neighborhood graph, $\varepsilon=300$</td>
<td>68.6</td>
<td>64.2</td>
</tr>
<tr>
<td>complete graph</td>
<td>70.6</td>
<td>62.2</td>
</tr>
<tr>
<td>our approach</td>
<td>71.5</td>
<td>67.4</td>
</tr>
</tbody>
</table>
• Tian Lan, Leonid Sigal, Greg Mori. Social Roles in Hierarchical Models for Human Activity Recognition. CVPR 2012
Semantic Descriptions 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
pass

social roles
attacker
first defenders
man-marking
defend-space
teammate

event
corner hit
free hit
attack play

first defenders

man-marking
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

**Y:**
- Corner hit
- Attack play

**r:**
- Attacker
- Man-marking

**h:**
- Pass
- jog

**x:** Concatenated HOG [Dalal & Triggs, 2005]
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: \( \text{loss} = \Delta(y, y_i) \)

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

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

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

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

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

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

\[ w_{y, r, h_i} \cdot \psi_i - w_{y, r, h} \cdot \psi_i \geq \text{loss} - \xi_i \]

\( \forall i, \xi_i \geq 0 \)
Model Inference - Query

The learned models

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

Score: \[ \Psi(Y^*, \{r^*_i\}_n, \{h^*_i\}_n, q) \]

Event, social roles, actions, queries

Maximize \[ \max_{y, r, h, q} \sum_e w_e \Psi_e \]

Person detection and tracking

Coordinate ascent inference
ESPN Broadcast Field Hockey Data

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

<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

- attack
- first defenders
- defenders-space
- defenders-person
- other

- Unary
- Full model
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:
    - $r_1$, $r_2$, ..., $r_N$
  - Social Role:
    - $h_1$, $h_2$, ..., $h_N$
  - Action:
    - $x_1$, $x_2$, ..., $x_N$

- **Full model**
  - Event:
    - $y$
  - Social Role:
    - $r_1$, $r_2$, ..., $r_N$
  - Action:
    - $h_1$, $h_2$, ..., $h_N$

- **Group activity**
  - Event:
    - $x_O$
  - Social Role:
    - $h_1$, $h_2$, ..., $h_N$
  - Action:
    - $x_1$, $x_2$, ..., $x_N$

### 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)
Conclusion

Structural Recognition of Human Activities
Acknowledgements

Tian Lan