4th International Workshop on Shape Perception in Human and Computer Vision

Shape of Human Activities

Sinisa Todorovic

joint work with William Brendel
Activity Recognition

Activities with:

• Rich temporal structure

• Shared subactivities
Goal: Recognition and Segmentation

- Recognize activities
- Identify the start and end frames
- Explain recognition: space-time structure
- Segment people and objects

long jump  high jump
Prior Work – Video Representation

• Space-time points
  – Laptev & Schmid 08, Niebles & Fei-Fei 08,...

• Still human postures
  – Soatto 07, Ning & Huang 08,...

• Action Templates
  – Yao & Zhu 09,...

• Point tracks
  – Sukthankar & Hebert 10,...

• Motion segments
  – Gorelick & Irani 08, Pritch & Peleg 08,...
Prior Work – Video Representation

• Space-time points
  – Laptev & Schmid 08, Niebles & Fei-Fei 08,...

• Still human postures
  – Soatto 07, Ning & Huang 08,...

• Action Templates
  – Yao & Zhu 09,...

• Point tracks
  – Sukthankar & Hebert 10,...

• Motion segments
  – Gorelick & Irani 08, Pritch & Peleg 08,...
Prior Work – Activity Representation

• Classifiers, e.g., Bag-of-Words
  – Ke, Herbert ICCV’05
  – Hamid, Essa ICCV07
  – Laptev, Schmid CVPR’08
  – ...

• Graphical models, e.g., AND-OR
  – Ivanov, Bobick PAMI00
  – Xiang, Gong IJCV’06
  – Ryoo, Aggarwal ICCV’09
  – Gupta, Davis CVPR09
  – Liu, Zhu CVPR09
  – ...
Prior Work – Activity Representation

• Classifiers, e.g., Bag-of-Words
  – Ke, Herbert ICCV’05
  – Hamid, Essa ICCV07
  – Laptev, Schmid CVPR’08

• Graphical models, e.g., AND-OR
  – Ivanov, Bobick PAMI00
  – Xiang, Gong IJCV’06
  – Ryoo, Aggarwal ICCV’09
  – Gupta, Davis CVPR09
  – Liu, Zhu CVPR09

• Require many examples
• Narrow goal: classification

• Pre-fixed model structure
• Hard to learn
• Hard to infer
Hypothesis

• Point-based features provide poor cues

• More expressive models are needed
Hypothesis

• Point-based features provide poor cues
• More expressive models are needed

To bridge the semantic gap

• Use mid-level features: Activity shape
  – Less training examples
  – Allow simpler learning and inference
Spatiotemporal Segmentation

Irani & Peleg 94, Weiss 97, Shi & Malik 98, DeMenthon 02, Cohen 04, Greenspan et al. 02, Ahuja 05, Medioni 05, Todorovic 09, Essa 10,...
Objects occupy space-time tubes

Because they
  - are cohesive in space
  - have locally smooth trajectories in time
As the right scale is unknown...

The graph captures spatiotemporal structure
Activity Shape = Segmentation Graph

Attributes of nodes and edges:

— Intrinsic properties: $F$
  
  - Motion
  - Object shape
Activity Shape = Segmentation Graph

Attributes of nodes and edges:

– Intrinsic properties: \( F \)
  • Motion
  • Object shape

– Adjacency matrices: \( A \)
  • Allen temporal relations
  • Spatial relations
  • Compositional relations

\[ G = (V, E) = \{(A_1, F_1), \ldots, (A_L, F_L)\} \]
Activity Shape = Segmentation Graph

Attributes of nodes and edges:

– Intrinsic properties: $F$
  - Motion
  - Object shape

– Adjacency matrices: $A$
  - Allen temporal relations
  - Spatial relations
  - Compositional relations
Our Approach

In a new video:
• Recognize
• Segment
• Explain
Activity-Shape Model

Video = Graph instance sampled from the model
Activity-Shape Model

Video = Graph instance sampled from the model

Model = Probabilistic Graph Mixture

- **compositional**
  \[ \mathcal{W}_1 \circ (A_1, F_1) + \mathcal{W}_2 \circ (A_2, F_2) + \ldots + \mathcal{W}_L \circ (A_L, F_L) \]

- **spatial**

- **temporal**
Generative Process

video: \( G = \{(A_1, F_1), \ldots, (A_L, F_L)\} \)

adjacency matrix \( A_i = PA_i P^T + \eta_i \)

node descriptor \( F_i = PF_i + \xi_i \)

\( i = 1, 2, \ldots, L \)
Activity-Shape Model

adjacency matrix  

\[ A_i = P A_i P^T + \eta_i \]

permutation matrix  

node descriptor  

\[ F_i = P F_i + \xi_i \]

\( i = 1, 2, \ldots, L \)
Learning

GIVEN \( K \) training videos \( \{ G_k : k = 1, ..., K \} \)

\[
A_{ki} = P_k A_i P_k^T + \eta_i \quad \text{permutation matrices}
\]

\[
F_{ki} = P_k F_i + \xi_i
\]

\( i = 1, 2, ..., L \)
GIVEN $K$ training videos

adjacency matrix

$$A_{ki} = P_k A_i P_k^T + \eta_i$$

node descriptor

$$F_{ki} = P_k F_i + \xi_i$$

$i = 1, 2, \ldots, L$
Learning

GIVEN $K$ training videos

adjacency matrix

$$A_{ki} = P_k A_i P_k^T + \eta_i$$

permutation matrices

node descriptor

$$F_{ki} = P_k F_i + \xi_i$$

noise

$i = 1, 2, \ldots, L$
constraint on permutation matrices

\[ \forall k, \quad P_k P_k^T = I, \quad P_k \in \{0, 1\}^{m \times n} \]

Learning
Inference \[ = \text{Quadratic Integer Program} \]
correctly learned activity-characteristic tubes
Learning Results

correctly learned activity-characteristic tubes
activity "handshaking" detected and segmented characteristic tube
activity "kicking" detected and segmented characteristic tube
Classification on UTexas Dataset

<table>
<thead>
<tr>
<th></th>
<th>hand shaking</th>
<th>hugging</th>
<th>kicking</th>
<th>pointing</th>
<th>punching</th>
<th>pushing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our</td>
<td>81.7% 75%</td>
<td>89.6% 87.5%</td>
<td>68.6% 62.5%</td>
<td>66.4% 50%</td>
<td>84.5% 75%</td>
<td>82.7% 75%</td>
</tr>
</tbody>
</table>

human interaction activities
Conclusion

• Shape-based video representation enables:
  – Simpler activity models, learning, inference...
  – Richer interpretation: recognition + segmentation

• Difficulties
  – Correspondence between model and data features