Frontiers of 
*Human Activity Analysis*

J. K. Aggarwal
Michael S. Ryoo
Kris M. Kitani
Overview
Machine point of view

- Activities as videos
  - Activity = a particular set of videos
Activity classification

- Simple task of identifying videos
  - Categorize given videos into their types.
    - Known, limited number of classes
    - Assumes that each video contains a single activity

Push
Punch
Kick
Shake
Hug
Activity classification

- **Activity categorization**
  - Input = a video segment containing 1 activity
Activity detection

- **Search** for the particular time interval
  - <starting time, ending time>
  - Video segment containing the activity

**Input:**
continuous video stream
Activity detection by classification

- Binary classifier
- Sliding window technique
  - Classify all possible time intervals
Recognition process

- Represent videos in terms of features
  - Captures properties of activity videos

- Recognize activities by comparing video representations
  - Decision boundary
Taxonomy

- Approach based taxonomy
  - Recognition approaches can be categorized.

Human activity recognition

Single-layered approaches
- Space-time approaches
  - Trajectories
  - Volumes
  - Local features
  - Exemplar-based
  - State-based

Hierarchical approaches
- Sequential approaches
- Statistical
- Syntactic
- Description-based

Aggarwal and Ryoo, ACM CSUR 2011
Single layered vs. hierarchical

- Single layered approaches

- Hierarchical approaches

Feature extraction

Stretching \(<1, 20>\)
Withdrawning \(<21, 30>\)
These approaches recognize actions directly from a sequence of images.

- **Single-layered approaches**
  - **Space-time approaches**
    - **Trajectories**
      - Template matching
        - [Campbell and Bobick '95]
        - [Rao and Shah '01]
      - Neighbor-based (including SVM)
        - [Sato and Aggarwal '04]
    - **Space-time volume**
      - [Bobick and J. Davis '01]
      - [Shechtman and Irani '05]
      - [Rodriguez et al. '08]
    - **Space-time features**
      - [Zelnik-Manor '01]
      - [Laptev and Lindeberg '03]
      - [Dollar et al. '05]
      - [Efros et al. '03]
      - [Yilmaz and Shah '05]
      - [Ke et al. '07]
      - [Shul'et al. '04]
      - [Blank et al. '05]
      - [Scovanner et al. '07]
      - [Laptev et al. '08]
    - **Data-based**
      - [Darrell and Pentland '93]
      - [Gavrila and L. Davis '95]
      - [Yacoob and Black '98]
      - [Ali and Aggarwal '01]
      - [Veeraraghavan et al. '06]
      - [Lubliner et al. '06]
      - [Jiang et al. '06]
      - [Vaswani et al. '03]^6
    - **State model-based**
      - [Yamato et al. '92]
      - [Starner and Pentland '95]
      - [Bregler '97]
      - [Bobick and Wilson '97]
      - [Oliver et al. '00]
      - [Park and Aggarwal, '04]
      - [Natarajan and Nevatia '07]
      - [Moore et al. '99]^9
      - [Gupta and Davis '07]^9
      - [Filipovych and Ribeiro '08]^9
  - **Sequential approaches**
    - Statistical matching
      - [Sheikh et al. '05]
      - [Ekan and Shah '05]^9
      - [Chomat and Crowley '99]
      - [Nebles et al. '06, '08]
      - [Wong et al. '07]
      - [Le et al. '04]^9
      - [Chomat and Crowley '99]
      - [Nebles et al. '06, '08]
      - [Wong et al. '07]
      - [Le et al. '04]^9
Single layered approaches

- Action representation
  - Video volumes themselves
  - Features directly extracted from videos

- Action classification
  - Machine learning techniques
    - Support vector machines
    - Hidden Markov models
### Taxonomy – Hierarchical

<table>
<thead>
<tr>
<th></th>
<th>Statistical approaches</th>
<th>Syntactic approaches</th>
<th>Description-based approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human actions</td>
<td>[Nguyen et al. '05]</td>
<td></td>
<td>[Pinhanez and Bobick '98]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Gupta et al. '09]</td>
<td>[Intille and Bobick '99]</td>
</tr>
<tr>
<td>Human-Human interactions</td>
<td>[Oliver et al. '02]</td>
<td>[Ivanov and Bobick '00]</td>
<td>[Vu et al. '03]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Joo and Chellapha '06]</td>
<td>[Ghanem et al. '04]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[Ryoo and Aggarwal '06, '09a]</td>
</tr>
<tr>
<td>Human-Object interactions</td>
<td>[Shi et al. '04]O</td>
<td>[Moore and Essa '02]O</td>
<td>[Siskind '01]O</td>
</tr>
<tr>
<td></td>
<td>[Yu and Aggarwal '06]O</td>
<td>[Minnen et al. '03]O</td>
<td>[Nevatia et al. '03, '04]O</td>
</tr>
<tr>
<td></td>
<td>[Damen and Hogg '09]O</td>
<td>[Kitani et al. '07]O</td>
<td>[Ryoo and Aggarwal '07]O</td>
</tr>
<tr>
<td>Group activities</td>
<td>[Cupillard et al. '02]G</td>
<td></td>
<td>[Ryoo and Aggarwal '08, '10]G</td>
</tr>
<tr>
<td></td>
<td>[Gong and Xiang '03]G</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Zhang et al.'06]G</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[Dai et al.'08]G</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical approaches

- Layered approaches
  - Activities in terms of sub-events.
  - Human interactions
    - Multiple agents

- Suitable for activity-level recognition
Hierarchical approaches

- Activities as semantic structures
  - Activity = a concatenation of its sub-events
  - Human-oriented: high-level
  - Hierarchically organized representations

**Handshake** = “two persons do **shake-action** (stretches, stays stretched, withdraw) simultaneously, while touching”.

```
this == Push_interactions(p1,p2)
 i = Stretch(p1’s arm) j = Stay_Stretched(p1’s arm)
 k = Touching(p1, p2) l = Depart(p2, p1)
```

- Fighting → Punching : 0.3
  | Punching Fighting : 0.7
- Punching → stretch withdraw : 0.8
  | stretch stay_withdrawn : 0.1
  | stay_stretched withdraw : 0.1