Emerging Topics in Human Activity Recognition

Michael S. Ryoo  NASA-JPL
Ivan Laptev  INRIA
Greg Mori  Simon Fraser University
Sangmin Oh  Kitware

CVPR tutorial on 2014/06/23
Introduction
Introduction

Computer Vision – Video Understanding

- Labeling of events by humans in a given video

Person 1 – teases P2, runs away
Person 3 – kicks P1
Person 4 – stops fighting

- Development of automated perception algorithms
Why video analysis?

Data:

- ~2.5 Billion new images / month
- ~5K image uploads every min.
- ~30M surveillance cameras in US
  => ~700K video hours/day
- TV-channels recorded since 60’s
- >34K hours of video upload every day
- And even more with future wearable devices
Why video analysis?

Applications:

- First appearance of N. Sarkozy on TV
- Sociology research: Influence of character smoking in movies
- Education: How do I make a pizza?
- Where is my cat?
- Predicting crowd behavior
- Counting people
- Motion capture and animation
Why video analysis?

Applications:

Unconstrained video search
Why human activities?

How many person-pixels are in the video?
Why human activities?

How many person-pixels are in the video?

- Movies: 35%
- TV: 34%
- YouTube: 40%
How many person pixels in our daily life?

Wearable camera data: Microsoft SenseCam dataset
How many person pixels in our daily life?

Wearable camera data: Microsoft SenseCam dataset

~4%
Activity recognition

*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

Push/not-push Classifier

Sliding window technique

• Classify all possible time intervals
Activity Recognition with postures

Hierarchical activity recognition

Interaction:
Person “pushed” by Person in time interval <4, 20>

Action recognition

<1,20> : Facing right
<1,20> : Arm staying
<1,20> : Leg staying
<1,20> : Facing left
<4,20> : Arm stretching
<1,20> : Leg staying

Features from each frame

1 : Arm withdrawn
8 : Arm withdrawn
13 : Arm withdrawn
1 : Arm withdrawn
8 : Arm somewhat stretched
13 : Arm fully stretched

Body-part tracking

Input sequences

- Interaction
- Action
  - Sequence of features
- Features
  - Numerical status of a body part
- Tracking
  - Estimates locations of human body parts
Activity Recognition with video features

Activity recognition

Activity: “Shaking” occurred

Spatio-temporal features

Videos as 3-D X YT volumes

Decision boundary for histogram of video patches (i.e., space-time features)

Input sequences

- Activity recognition
  - Decision boundary

- Activity representation
  - A set of local spatio-temporal features

- Features
  - Information in local video patches with salient movements
History

1973

Johansson’s experiments [Johansson 1973]

1992

Tennis action recognition [Yamato et al. 1992]

1995

American sign language (ASL) recognition [Starner and Pentland 1995]
History

3-D posture-based recognition [Gavrila and L. Davis 1995]

Motion history images [J. Davis, Bobick 1997]

Human-human interactions [Oliver, Rosario, Pentland 2000]
History

Time

Far-field action recognition
[Efros, Berg, Mori, Malik 2003]

KTH
Local spatio-temporal features
[Laptev 2005]

Hollywood
Movies
[Laptev 2008]

UT-Interaction
Spatio-temporal relations
[Ryoo and Aggarwal 2009]
Dimension 1: type of videos

Different types of videos and their dataset

**Surveillance videos**
- Static cameras
- Side or top view
- Simple background

**Movies and user videos**
- Moving cameras
- Side view
- Dynamic

**Sports videos**
- Video segments
- Side or top view
- Objects/people

**First-person videos**
- Moving cameras
- 1st-person view
- Very dynamic
Dimension 2: levels of human activities

There are various types of activities

- The ultimate goal is to make computers recognize all of them reliably.

Levels of human activities:
- gestures
- actions
- human-object interactions
- interactions
- group activities
Dimension 3: structure in activities

Different levels of structure complexity (temporal/spatial)

Complexity of structure in human activities
## Existing works

Based on ‘video type’ and ‘activity level’ dimensions

<table>
<thead>
<tr>
<th>Video Type</th>
<th>Actions</th>
<th>Human-object interactions</th>
<th>Human-human interactions</th>
<th>Group activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surveillance videos</td>
<td>[Laptev 05]</td>
<td>[Oh et al. 11]</td>
<td>[Ryoo and Aggarwal 09]</td>
<td>[Ryoo and Aggarwal 08,11] [Lan, Wang, Yang, Mori 10]</td>
</tr>
<tr>
<td>Movies and user videos</td>
<td>[Laptev 07]</td>
<td>[Marszałek, Laptev, Schmid 09]</td>
<td>[Rodriguez, Ahmed, Shah 08]</td>
<td>[Lan, Sigil, Mori 12]</td>
</tr>
<tr>
<td>Sports videos</td>
<td>[Efros, Berg, Mori, Malik 2003]</td>
<td>[Rodriguez, Ahmed, Shah 08]</td>
<td>[Niebles, Chen, Fei-Fei 10]</td>
<td>[Lan, Sigil, Mori 12]</td>
</tr>
<tr>
<td>First-person videos</td>
<td>[Kitani et al. 11]</td>
<td>[Fathi et al. 2011]</td>
<td>[Ryoo and Matthies 13]</td>
<td>[Lan, Sigil, Mori 12]</td>
</tr>
</tbody>
</table>
# Existing works

Based on ‘video type’ and ‘activity level’ dimensions

<table>
<thead>
<tr>
<th></th>
<th>Actions</th>
<th>Human-object interactions</th>
<th>Human-human interactions</th>
<th>Group activities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surveillance videos</strong></td>
<td>[Laptev 05]</td>
<td>[Oh et al. 11]</td>
<td>[Ryoo and Aggarwal 09]</td>
<td>[Ryoo and Aggarwal 08,11]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>[Vahdat, Gao, Ranjbar, Mori 11]</td>
<td>[Lan, Wang, Yang, Mori 10]</td>
</tr>
<tr>
<td><strong>Movies and user videos</strong></td>
<td>[Laptev 07]</td>
<td>[Marszałek, Laptev, Schmid 09]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Kim, Oh, Vahdat, Cannons, Perera, Mori 13]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sports videos</strong></td>
<td>[Efros, Berg, Mori, Malik 2003]</td>
<td>[Rodriguez, Ahmed, Shah 08]</td>
<td>[Niebles, Chen, Fei-Fei 10]</td>
<td>[Lan, Sigil, Mori 12]</td>
</tr>
<tr>
<td><strong>First-person videos</strong></td>
<td>[Kitani et al. 11]</td>
<td>[Fathi et al. 2011]</td>
<td>[Pirsiavash and Ramanan 2012]</td>
<td>[Ryoo and Matthies 13]</td>
</tr>
</tbody>
</table>
Why difficult?

- **Large variations in appearance:** occlusions, non-rigid motion, viewpoint changes, clothing...

- **Manual collection of training samples is prohibitive:** many action classes, rare occurrence

- **Action vocabulary is not well-defined**
Challenges - variations