Activity Recognition

Computer Vision
CS 143, Brown

James Hays
What is an action?

Action: a transition from one state to another

• Who is the actor?
• How is the state of the actor changing?
• What (if anything) is being acted on?
• How is that thing changing?
• What is the purpose of the action (if any)?
Human activity in video

No universal terminology, but approximately:

• **“Actions”**: atomic motion patterns -- often gesture-like, single clear-cut trajectory, single nameable behavior (e.g., sit, wave arms)

• **“Activity”**: series or composition of actions (e.g., interactions between people)

• **“Event”**: combination of activities or actions (e.g., a football game, a traffic accident)

Adapted from Venu Govindaraju
How do we represent actions?

Categories
Walking, hammering, dancing, skiing, sitting down, standing up, jumping

Poses

Nouns and Predicates
<man, swings, hammer>
<man, hits, nail, w/ hammer>
What is the purpose of action recognition?
Surveillance

http://users.isr.ist.utl.pt/~etienne/mypubs/Avinetal06PETS.pdf
Interfaces

We will soon launch our beta product. Stay tuned and be the first to control YouTube, Hulu, Vevo or Netflix through a flick of fingers.

2011
Interfaces

How can we identify actions?

- Motion
- Pose
- Held Objects
- Nearby Objects
Representing Motion

Optical Flow with Motion History

sit-down

sit-down MHI
Representing Motion

Optical Flow with Split Channels
Representing Motion

Tracked Points

Matikainen et al. 2009
Representing Motion
Space-Time Interest Points

Corner detectors in space-time
Representing Motion
Space-Time Interest Points

Hand waves with high frequency

Hand waves with low frequency
Representing Motion

Space-Time Volumes

Blank et al. 2005
Examples of Action Recognition Systems

• Feature-based classification

• Recognition using pose and objects
Action recognition as classification

Retrieving actions in movies, Laptev and Perez, 2007
Remember image categorization...

Training Images

Training

Image Features

Classifier Training

Training Labels

Trained Classifier
Remember image categorization...

Training

- Training Images
- Image Features
- Classifier Training
- Trained Classifier
- Training Labels
- Outdoor Prediction

Testing

- Test Image
- Image Features
- Trained Classifier
- Prediction Outdoor
Remember spatial pyramids....

Compute histogram in each spatial bin
Features for Classifying Actions

1. Spatio-temporal pyramids (14x14x8 bins)
   - Image Gradients
   - Optical Flow
Features for Classifying Actions

2. Spatio-temporal interest points

Corner detectors in space-time

Descriptors based on Gaussian derivative filters over x, y, time
Classification

- Boosted stubs for pyramids of optical flow, gradient
- Nearest neighbor for STIP
Searching the video for an action

1. Detect keyframes using a trained HOG detector in each frame
2. Classify detected keyframes as positive (e.g., “drinking”) or negative (“other”)
Accuracy in searching video

With keyframe detection

Without keyframe detection

PR drinking

- OF5Hist-KFtrained (ap: 0.434)
- OFGrad9Hist-KFtrained (ap: 0.343)
- OFGrad9Hist (ap: 0.179)
- OF5Hist (ap: 0.048)

precision vs. recall graph
Learning realistic human actions from movies, Laptev et al. 2008

“Talk on phone”

“Get out of car”
Approach

• Space-time interest point detectors
• Descriptors
  – HOG, HOF
• Pyramid histograms (3x3x2)
• SVMs with Chi-Squared Kernel
Results

<table>
<thead>
<tr>
<th>Task</th>
<th>HoG BoF</th>
<th>HoF BoF</th>
<th>Best channel</th>
<th>Best combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>KTH multi-class</td>
<td>81.6%</td>
<td>89.7%</td>
<td>91.1% (hof h3x1 t3)</td>
<td>91.8% (hof 1 t2, hog 1 t3)</td>
</tr>
<tr>
<td>Action AnswerPhone</td>
<td>13.4%</td>
<td>24.6%</td>
<td>26.7% (hof h3x1 t3)</td>
<td>32.1% (hof o2x2 t1, hog h3x1 t3)</td>
</tr>
<tr>
<td>Action GetOutCar</td>
<td>21.9%</td>
<td>14.9%</td>
<td>22.5% (hof o2x2 1)</td>
<td>41.5% (hof o2x2 t1, hog h3x1 t1)</td>
</tr>
<tr>
<td>Action HandShake</td>
<td>18.6%</td>
<td>12.1%</td>
<td>23.7% (hog h3x1 1)</td>
<td>32.3% (hog h3x1 t1, hog o2x2 t3)</td>
</tr>
<tr>
<td>Action HugPerson</td>
<td>29.1%</td>
<td>17.4%</td>
<td>34.9% (hog h3x1 t2)</td>
<td>40.6% (hog 1 t2, hog o2x2 t1, hog h3x1 t2)</td>
</tr>
<tr>
<td>Action Kiss</td>
<td>52.0%</td>
<td>36.5%</td>
<td>52.0% (hog 1 1)</td>
<td>53.3% (hog 1 t1, hog 1 t1, hog o2x2 t1)</td>
</tr>
<tr>
<td>Action SitDown</td>
<td>29.1%</td>
<td>20.7%</td>
<td>37.8% (hog 1 t2)</td>
<td>38.6% (hog 1 t2, hog 1 t3)</td>
</tr>
<tr>
<td>Action SitUp</td>
<td>6.5%</td>
<td>5.7%</td>
<td>15.2% (hog h3x1 t2)</td>
<td>18.2% (hog o2x2 t1, hog o2x2 t2, hog h3x1 t2)</td>
</tr>
<tr>
<td>Action StandUp</td>
<td>45.4%</td>
<td>40.0%</td>
<td>45.4% (hog 1 1)</td>
<td>50.5% (hog 1 t1, hog 1 t2)</td>
</tr>
</tbody>
</table>
Action Recognition using Pose and Objects

Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities, B. Yao and Li Fei-Fei, 2010
Human-Object Interaction

Holistic image based classification

Integrated reasoning
- Human pose estimation
Human-Object Interaction

Holistic image based classification

Integrated reasoning
  • Human pose estimation
  • Object detection
Human-Object Interaction

Holistic image based classification

Integrated reasoning
- Human pose estimation
- Object detection
- Action categorization

HOI activity: Tennis Forehand
Human pose estimation & Object detection

Human pose estimation is challenging.

- Difficult part appearance
- Self-occlusion
- Image region looks like a body part

- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009
Human pose estimation is challenging.

- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009
Human pose estimation & Object detection

Facilitate

Given the object is detected.
Human pose estimation & Object detection

- Small, low-resolution, partially occluded
- Image region similar to detection target

Object detection is challenging

- Viola & Jones, 2001
- Lampert et al, 2008
- Divvala et al, 2009
- Vedaldi et al, 2009

Slide Credit: Yao/Fei-Fei
Human pose estimation & Object detection

Object detection is challenging

- Viola & Jones, 2001
- Lampert et al, 2008
- Divvala et al, 2009
- Vedaldi et al, 2009

Slide Credit: Yao/Fei-Fei
Human pose estimation & Object detection

Facilitate

Given the pose is estimated.

Slide Credit: Yao/Fei-Fei
Human pose estimation & Object detection

Mutual Context
Mutual Context Model Representation

A:
- Tennis forehand
- Croquet shot
- Volleyball smash

O:
- Tennis racket
- Croquet mallet
- Volleyball

H:
- Intra-class variations
  - More than one H for each A;
  - Unobserved during training.

P: \( l_p \): location; \( \theta_p \): orientation; \( s_p \): scale.

\( f \): Shape context. [Belongie et al, 2002]

Slide Credit: Yao/Fei-Fei
Activity Classification Results

- Cricket shot
- Tennis forehand

<table>
<thead>
<tr>
<th>Model</th>
<th>Classification accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model</td>
<td>83.3%</td>
</tr>
<tr>
<td>Gupta et al, 2009</td>
<td>78.9%</td>
</tr>
<tr>
<td>Bag-of-words SIFT+SVM</td>
<td>52.5%</td>
</tr>
</tbody>
</table>

Slide Credit: Yao/Fei-Fei
Take-home messages

• Action recognition is an open problem.
  – How to define actions?
  – How to infer them?
  – What are good visual cues?
  – How do we incorporate higher level reasoning?
Take-home messages

• Some work done, but it is just the beginning of exploring the problem. So far...
  – Actions are mainly categorical
  – Most approaches are classification using simple features (spatial-temporal histograms of gradients or flow, s-t interest points, SIFT in images)
  – Just a couple works on how to incorporate pose and objects
  – Not much idea of how to reason about long-term activities or to describe video sequences