Today

• Human pose and actions: Introduction
• Estimating human pose
• Recognizing human actions
  – Using specialized features
  – Using pose
  – Using objects
  – From ego-centric video
Next time (Last class)

• Review for the final exam + OMETs
• By Wednesday night, post on Piazza questions or anything you want me to review, for participation credit

• Extra office hours on Friday, 2-3pm
Final Exam

• Monday, Dec. 14, 12pm
• Same room (5502 Sennott Square)
• Similar to midterm exam (mostly short questions and a few problems), but longer (100 points)
• Will only cover topics discussed after midterm (but some of these use topics from first half)
Homework 4

Mean = 77.16, median = 99, max = 123
Homework 5

• Due Thursday
• See Piazza for correction about how to get probabilities for Part III
Participation

- Tentative grades entered on CourseWeb
- Median is 80%
What is an action/activity?

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)?
• Could be more or less complex

Adapted from Derek Hoiem
Terminology: 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>
How can we identify actions?

Motion

Pose

Held

Objects

Nearby

Objects

Derek Hoiem
Today

• Human pose and actions: Introduction

• Estimating human pose

• Recognizing human actions
  – Using specialized features
  – Using pose
  – Using objects
  – From ego-centric video
Real-Time Human Pose Recognition in Parts from Single Depth Images

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, Andrew Blake

Best paper award at CVPR 2011

Adapted from Jamie Shotton
The mission

- Recognize large variety of human poses, all shapes & sizes
- Limited compute budget
  - super-real time on Xbox 360 to allow games to run concurrently

Adapted from Jamie Shotton
The approach: body part recognition

- right hand
- neck
- right elbow
- left shoulder

Jamie Shotton
Body part recognition

- No temporal information
  - frame-by-frame

- Local pose estimate of parts
  - each pixel & each body joint treated independently
  - reduced training data and computation time

- Very fast
  - simple depth image features
  - parallel decision forest classifier
The Kinect pose estimation pipeline

capture depth image & remove bg

infer body parts per pixel

cluster pixels to hypothesize body joint positions

fit model & track skeleton

Jamie Shotton
Classifying pixels

- Compute $P(c_i | w_i)$
  - pixels $i = (x, y)$
  - body part $c_i$
  - image window $w_i$

- Discriminative approach
  - learn classifier $P(c_i | w_i)$ from training data
Synthetic training data

Record mocap
500k frames
distilled to 100k poses

Retarget to several models

Render (depth, body parts) pairs

Train invariance to:

Jamie Shotton
Fast depth image features

- Depth comparisons
  - Very fast to compute

\[
f_\Theta(I, x) = d_I(x) - d_I(x + \Delta)
\]

Adapted from Jamie Shotton
To classify pixel $x$, start here

$\Theta(I, x; \Delta_1) > t_1$

$\Theta(I, x; \Delta_2) > t_2$

Adapted from Jamie Shotton
Depth of trees

input depth | ground truth parts | inferred parts (soft)

depth 18
Depth of trees

Average per-class accuracy vs. Depth of trees for synthetic and real test data.

- **Synthetic test data**
  - 900k training images
  - 15k training images

- **Real test data**
  - 900k training images
  - 15k training images

Jamie Shotton
Decision forest classifier

- Trained on different random subset of images
  - “bagging” helps avoid over-fitting
- Average tree posteriors $P(c|I, x) = \frac{1}{T} \sum_{t=1}^{T} P_t(c|I, x)$

[Amit & Geman 97]
[Breiman 01]
[Geurts et al. 06]
Number of trees

Average per-class accuracy

<table>
<thead>
<tr>
<th>Number of trees</th>
<th>1 tree</th>
<th>3 trees</th>
<th>6 trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>40%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ground truth

inferred body parts (most likely)
input depth

inferred body parts

front view

side view

top view

inferred joint positions (modes found using mean shift)

no tracking or smoothing

Jamie Shotton
input depth

inferred body parts

front view

side view

top view

inferred joint positions (modes found using mean shift)

no tracking or smoothing
Today

• Human pose and actions: Introduction
• Estimating human pose
• Recognizing human actions
  – Using specialized features
  – Using pose
  – Using objects
  – From ego-centric video
Representing Actions

Tracked Points

Adapted from Derek Hoiem

Matikainen et al. 2009
Representing Actions

Space-Time Interest Points

- Corner detectors in space+time

Adapted from Derek Hoiem
Representing Actions

*Spatio-temporal interest points*

*Spatial interest points*
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, hof h3x1 t3)</td>
</tr>
<tr>
<td>Action GetOutCar</td>
<td>21.9%</td>
<td>14.9%</td>
<td>22.5% (hof o2x2 t1)</td>
<td>41.5% (hof o2x2 t1, hof h3x1 t1)</td>
</tr>
<tr>
<td>Action HandShake</td>
<td>18.6%</td>
<td>12.1%</td>
<td>23.7% (hog h3x1 t1)</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 t2, 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 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>

Derek Hoiem, figures from Ivan Laptev
Today

• Human pose and actions: Introduction
• Estimating human pose
• Recognizing human actions
  – Using specialized features
  – Using pose
  – Using objects
  – From ego-centric video
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

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

Yao/Fei-Fei
Human pose estimation & Object detection

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
Human pose estimation & Object detection

Object detection is challenging

- Viola & Jones, 2001
- Lampert et al, 2008
- Divvala et al, 2009
- Vedaldi et al, 2009
Human pose estimation & Object detection

Facilitate

Given the pose is estimated.
Human pose estimation & Object detection

Mutual Context

Yao/Fei-Fei
Activity Classification Results

- **Cricket shot**
  - Gupta et al., 2009: 78.9%
  - Bag-of-words SIFT+SVM: 52.5%

- **Tennis forehand**
  - Our model: 83.3%
  - Gupta et al., 2009: 78.9%
  - Bag-of-words SIFT+SVM: 52.5%
Today

• Human pose and actions: Introduction
• Estimating human pose
• Recognizing human actions
  – Using specialized features
  – Using pose
  – Using objects
  – From ego-centric video
Detecting Activities of Daily Living in First-person Camera Views

Hamed Pirsiavash, Deva Ramanan

CVPR 2012
Motivation

A sample video of Activities of Daily Living
Applications
Tele-rehabilitation

Long-term at-home monitoring

Applications
Life-logging

So far, mostly “write-only” memory!

This is the right time for computer vision community to get involved.

Wearable ADL detection

It is easy to collect natural data
Challenges
What features to use?

Low level features  High level features
(Weak semantics)  (Strong semantics)

Space-time interest points  Human pose
Laptev, IJCV’05

Difficulties of pose:
• Detectors are not accurate enough
• Not useful in first person camera views
Challenges
What features to use?

Low level features
(Weak semantics)

High level features
(Strong semantics)

Space-time interest points
Laptev, IJCV’05

Human pose

Object-centric features

Difficulties of pose:
• Detectors are not accurate enough
• Not useful in first person camera views
Challenges
Long-scale temporal structure

“Classic” data: boxing

Wearable data: making tea

- Start boiling water
- Do other things (while waiting)
- Pour in cup
- Drink tea

Adapted from Hamed Pirsiavash
Appearance feature: bag of objects

Video clip

Bag of detected objects

SVM classifier
Temporal pyramid
Coarse to fine correspondence matching with a multi-layer pyramid

Inspired by “Spatial Pyramid” CVPR’06 and “Pyramid Match Kernels” ICCV’05
Accuracy on 18 action categories

- Our model: 40.6%
- STIP baseline: 22.8%
Summary: Human actions

• Action recognition still an open problem
  – How to represent actions?
• Types of data: atomic and more complex actions, ego-centric video
• Common representations
  – Space-time interest points
  – Pose
  – Objects (and temporal pyramids of objects)
• Pose
  – Can be approached as a classification problem using depth data