Action Recognition

Computer Vision

Jia-Bin Huang, Virginia Tech

Many slides from D. Hoiem
This section: advanced topics

• Convolutional neural networks in vision

• Action recognition

• Vision and Language

• 3D Scenes and Context
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)?
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?

To describe
What is the purpose of action recognition?

• To predict
What is the purpose of action recognition?

• To understand the intention and motivation

Why are they doing that?

to sell ice cream

to commute to work

to answer emergency call

to win race

Predicting Motivations of Actions by Leveraging Text, CVPR 2016
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

Space-Time Volumes

Blank et al. 2005
Representing Motion

Optical Flow with Split Channels

(a) original image

(b) optical flow $F_{x,y}$

(c) $F_x, F_y$

(d) $F^+_x, F^-_x, F^+_y, F^-_y$

(e) $Fb^+_x, Fb^-_x, Fb^+_y, Fb^-_y$

split into pos/neg channels

blurred pos/neg flow

Efros et al. 2003
Representing Motion
Tracked Points

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

Moving corner

Ball hits wall

Corner detectors in space-time

Balls collide

Balls collide (different scale)
Representing Motion
Space-Time Interest Points

Hand waves with high frequency

Hand waves with low frequency

Laptev 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**
- Training Images
- Image Features
- Classifier Training
- Trained Classifier
- Training Labels

**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
   - Image Gradients
   - Optical Flow

![Diagram of Spatio-temporal pyramids](image)
Features for Classifying Actions

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)
“Talk on phone”

“Get out of car”

Learning realistic human actions from movies, Laptev et al. 2008
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 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 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

Slide Credit: Yao/Fei-Fei
Human-Object Interaction

Holistic image based classification

Integrated reasoning
  • Human pose estimation
  • Object detection

Tennis racket
Human-Object Interaction

Holistic image based classification

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

Activity: Tennis Forehand

Slide Credit: 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

Difficult part appearance

Self-occlusion

Image region looks like a body part
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

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

Object detection is challenging

Small, low-resolution, partially occluded

Image region similar to detection target

- 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.

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

Mutual Context

Slide Credit: Yao/Fei-Fei
Mutual Context Model Representation

\[ A: \text{Tennis forehand, Croquet shot, Volleyball smash} \]

\[ O: \text{Tennis racket, Croquet mallet, Volleyball} \]

\[ H: \text{Intra-class variations} \]
- More than one \( H \) for each \( A \);
- Unobserved during training.

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

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

Slide Credit: Yao/Fei-Fei
Learning Results

Cricket defensive shot

Cricket bowling

Croquet shot

Slide Credit: Yao/Fei-Fei
Learning Results

Tennis
- forehand

Tennis
- serve

Volleyball
- smash

Slide Credit: Yao/Fei-Fei
Model Inference

The learned models
Model Inference

The learned models

Layout of the object and body parts.

Compositional Inference

\[ \Psi \left( A_1, H_1, O_1^*, \{ P_{1,n}^* \}_n \right) \]

[Chen et al, 2007]

Head detection

Torso detection

Tennis racket detection

Slide Credit: Yao/Fei-Fei
Model Inference

The learned models

Output

$$\Psi (A_1, H_1, O_1^*, \{P_{1,n}^*\}_n)$$

$$\Psi (A_K, H_K, O_K^*, \{P_{K,n}^*\}_n)$$

Slide Credit: Yao/Fei-Fei
Dataset and Experiment Setup

**Sport data set:** 6 classes
180 training (supervised with object and part locations) & 120 testing images

Tasks:
- Object detection;
- Pose estimation;
- Activity classification.

[Cricket defensive shot]
[Cricket bowling]
[Croquet shot]
[Tennis forehand]
[Tennis serve]
[Volleyball smash]

[Gupta et al, 2009]
**Dataset and Experiment Setup**

**Sport data set:** 6 classes
180 training (supervised with object and part locations) & 120 testing images

- **Cricket** defensive shot
- **Cricket** bowling
- **Croquet** shot
- **Tennis** forehand
- **Tennis** serve
- **Volleyball** smash

**Tasks:**
- **Object detection**;
- **Pose estimation**;
- **Activity classification**.

[Gupta et al, 2009]

Slide Credit: Yao/Fei-Fei
Object Detection Results

Sliding window  Pedestrian context  Our Method
[Andriluka et al, 2009]  [Dalal & Triggs, 2006]

Cricket bat

Cricket ball

Croquet mallet

Tennis racket

Volleyball

Slide Credit: Yao/Fei-Fei
Object Detection Results

![Object Detection Results](image)

**Small object**
- Sliding window
- Pedestrian context
- Our method

**Background clutter**
- Small object
- Background clutter

**Cricket ball**

**Volleyball**

Slide Credit: Yao/Fei-Fei
Dataset and Experiment Setup

**Sport data set**: 6 classes
180 training & 120 testing images

Tasks:
- Object detection;
- **Pose estimation**;
- Activity classification.

[Gupta et al, 2009]
## Human Pose Estimation Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Torso</th>
<th>Upper Leg</th>
<th>Lower Leg</th>
<th>Upper Arm</th>
<th>Lower Arm</th>
<th>Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramanan, 2006</td>
<td>.52</td>
<td>.22</td>
<td>.22</td>
<td>.21</td>
<td>.28</td>
<td>.17</td>
</tr>
<tr>
<td>Andriluka et al, 2009</td>
<td>.50</td>
<td>.31</td>
<td>.30</td>
<td>.31</td>
<td>.27</td>
<td>.18</td>
</tr>
<tr>
<td>Our full model</td>
<td>.66</td>
<td>.43</td>
<td>.39</td>
<td>.44</td>
<td>.34</td>
<td>.44</td>
</tr>
</tbody>
</table>

Slide Credit: Yao/Fei-Fei
### Human Pose Estimation Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Torso</th>
<th>Upper Leg</th>
<th>Lower Leg</th>
<th>Upper Arm</th>
<th>Lower Arm</th>
<th>Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramanan, 2006</td>
<td>.52</td>
<td>.22</td>
<td>.22</td>
<td>.21</td>
<td>.28</td>
<td>.42</td>
</tr>
<tr>
<td>Andriluka et al, 2009</td>
<td>.50</td>
<td>.31</td>
<td>.30</td>
<td>.31</td>
<td>.27</td>
<td>.45</td>
</tr>
<tr>
<td>Our full model</td>
<td>.66</td>
<td>.43</td>
<td>.39</td>
<td>.44</td>
<td>.34</td>
<td>.58</td>
</tr>
</tbody>
</table>

Tennis serve model  
Our estimation result  
Andriluka et al, 2009  
Volleyball smash model  
Our estimation result  
Andriluka et al, 2009
### Human Pose Estimation Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Torso</th>
<th>Upper Leg</th>
<th>Lower Leg</th>
<th>Upper Arm</th>
<th>Lower Arm</th>
<th>Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramanan, 2006</td>
<td>.52</td>
<td>.22</td>
<td>.22</td>
<td>.21</td>
<td>.28</td>
<td>.17</td>
</tr>
<tr>
<td>Andriluka et al, 2009</td>
<td>.50</td>
<td>.31</td>
<td>.30</td>
<td>.31</td>
<td>.27</td>
<td>.18</td>
</tr>
<tr>
<td>Our full model</td>
<td>.66</td>
<td>.43</td>
<td>.39</td>
<td>.44</td>
<td>.34</td>
<td>.44</td>
</tr>
<tr>
<td>One pose per class</td>
<td>.63</td>
<td>.40</td>
<td>.36</td>
<td>.41</td>
<td>.31</td>
<td>.38</td>
</tr>
</tbody>
</table>

Slide Credit: Yao/Fei-Fei
Dataset and Experiment Setup

**Sport data set:** 6 classes
180 training & 120 testing images

**Tasks:**
- Object detection;
- Pose estimation;
- **Activity classification.**

[Gupta et al, 2009]
Activity Classification Results

Our model | Gupta et al, 2009 | Bag-of-words SIFT+SVM
---|---|---
83.3% | 78.9% | 52.5%

Classification accuracy

Cricket shot

Tennis forehand

Slide Credit: Yao/Fei-Fei
Motion features – Dense Trajectory

Dense sampling in each spatial scale

Tracking in each spatial scale separately

Trajectory description

N

HOG

Σ

HOF

Σ

MBH

Σ

Action Recognition by Dense Trajectories, CVPR 2011
Action Recognition with Improved Trajectories, ICCV 2013
Video classification with CNNs

Single Frame  Late Fusion  Early Fusion  Slow Fusion

Large-scale Video Classification with Convolutional Neural Networks, CVPR 2014
Video classification with CNNs

Sports Video Classification
Two-stream CNN

Spatial stream ConvNet
- conv1: 7x7x96, stride 2, norm. pool 2x2
- conv2: 5x5x256, stride 2, norm. pool 2x2
- conv3: 3x3x512, stride 1
- conv4: 3x3x512, stride 1
- conv5: 3x3x512, stride 1, pool 2x2
- full6: 4096 dropout
- full7: 2048 dropout
- softmax

Temporal stream ConvNet
- conv1: 7x7x96, stride 2, norm. pool 2x2
- conv2: 5x5x256, stride 2, norm. pool 2x2
- conv3: 3x3x512, stride 1
- conv4: 3x3x512, stride 1
- conv5: 3x3x512, stride 1, pool 2x2
- full6: 4096 dropout
- full7: 2048 dropout
- softmax

Input video
Single frame
Multi-frame optical flow
Class score fusion

Two-Stream Convolutional Networks for Action Recognition in Videos, NIPS 2014
3D Convolutional Networks

(a) 2D convolution

(b) 2D convolution on multiple frames

(c) 3D convolution

Conv1a 64 Pool1
Conv2a 128 Pool2
Conv3a 256 Pool3
Conv3b 256
Conv4a 512 Pool4
Conv4b 512
Conv5a 512 Pool5
Conv5b 512
fc6 4096
fc7 4096

Imagenet

C3D

Learning Spatiotemporal Features with 3D Convolutional Networks, ICCV 2015
Action recognition -> Semantic role Labeling
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 (could be framed in terms of effect or intent)
  • 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