Human action recognition: Recent progress, open questions and future challenges

Ivan Laptev
ivan.laptev@inria.fr
WILLOW, ENS/INRIA/CNRS, Paris
Computer vision grand challenge: Dynamic scene understanding

Objects:
cars, glasses, people, etc...

Actions:
drinking, running, door exit, car enter, etc...

Scene categories:
indoors, outdoors, street scene, etc...

Geometry:
Street, wall, field, stair, etc...

Constraints
Why analyzing people and human actions?
How many person pixels are in video?
How many person pixels are in video?

Movies: 35%

TV: 34%

YouTube: 40%
Applications

- Analyzing video archives
  - First appearance of N. Sarkozy on TV
  - Sociology research: Influence of character smoking in movies
  - Education: How do I make a pizza?

- Surveillance
  - Where is my cat?
  - Predicting crowd behavior
  - Counting people

- Graphics
  - Motion capture and animation
Example: Monitoring Russian presidential elections, March 4, 2012

- >80,000 election sites
- >84K hours (>95 years) of video in 1 day

• Interesting task: Automatic counting of voting actions
Example: Monitoring Russian presidential elections, March 4, 2012

- >80,000 election sites
- >84K hours (>95 years) of video in 1 day
- Interesting task: Detecting abnormal (depends on the country) events
What else can you find in 95 years of video (~75Tb)?
Problems

- Need to process very large amounts of video data
- Need to deal with large appearance variations, many classes

Drinking

Smoking
This talk:

Review of work on action recognition

Discussion: Do we ask the right questions?

Our more recent work
Motion perception (1973)

• “Moving Light Displays” (LED) inspired much of early work on human action recognition

A HOUGHTON MIFFLIN PRODUCTION

Copyright © 1971 by Houghton Mifflin Company

A Teaching Resource
At the Frontiers of Psychological Inquiry
Activities characterized by a pose
Activities characterized by a pose

Slide credit: A. Zisserman
Human pose estimation (2011)

Extension of LSVM model of Felzenszwalb et al.

Builds on Poslets idea of Bourdev et al.

Learns from lots of noisy annotations

Explores temporal continuity
Appearance-based methods: global shape

[A.F. Bobick and J.W. Davis, PAMI 2001]
Idea: summarize motion in video in a
Motion History Image (MHI):

Actions as spacetime shapes. 2007
Appearance methods: Shape

Pros:
+ Simple and fast
+ Works in controlled settings

Cons:
- Prone to errors of background subtraction
- Does not capture interior Structure and motion

Variations in light, shadows, clothing…
What is the background here?

Silhouette tells little about actions
Motion-based methods

Learning Parameterized Models of Image Motion
M.J. Black, Y. Yacoob, A.D. Jepson and D.J. Fleet, 1997

Recognizing action at a distance
Local feature methods

+ No segmentation needed
+ No object tracking needed
- Loss of global structure
Local feature methods: Why working?

- Finds similar events in pairs of video sequences
Bag-of-Features action recognition

Extraction of Local features

Occurrence histogram of visual words

Non-linear SVM with $\chi^2$ kernel

K-means clustering ($k=4000$)

Feature quantization

Feature description

[Laptev, Marszałek, Schmid, Rozenfeld 2008]
Action classification results

<table>
<thead>
<tr>
<th>Channel</th>
<th>hoghof</th>
<th>flat</th>
<th>Chance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AnswerPhone</td>
<td>15.7</td>
<td>20.9</td>
<td>7.2</td>
</tr>
<tr>
<td>DriveCar</td>
<td>86.6</td>
<td>84.6</td>
<td>11.5</td>
</tr>
<tr>
<td>Eat</td>
<td>59.5</td>
<td>67.0</td>
<td>3.7</td>
</tr>
<tr>
<td>FightPerson</td>
<td>71.1</td>
<td>69.8</td>
<td>7.9</td>
</tr>
<tr>
<td>GetOutCar</td>
<td>29.3</td>
<td>45.7</td>
<td>6.4</td>
</tr>
<tr>
<td>HandShake</td>
<td>21.2</td>
<td>27.8</td>
<td>5.1</td>
</tr>
<tr>
<td>HugPerson</td>
<td>35.8</td>
<td>43.2</td>
<td>7.5</td>
</tr>
<tr>
<td>Kiss</td>
<td>51.5</td>
<td>52.5</td>
<td>11.7</td>
</tr>
<tr>
<td>Run</td>
<td>69.1</td>
<td>67.8</td>
<td>16.0</td>
</tr>
<tr>
<td>SitDown</td>
<td>58.2</td>
<td>57.6</td>
<td>12.2</td>
</tr>
<tr>
<td>SitUp</td>
<td>17.5</td>
<td>17.2</td>
<td>4.2</td>
</tr>
<tr>
<td>StandUp</td>
<td>51.7</td>
<td>54.3</td>
<td>16.5</td>
</tr>
</tbody>
</table>

Average precision (AP) for Hollywood-2 dataset
**Evaluation of local feature detectors and descriptors**

**Four types of detectors:**
- Harris3D [Laptev 2003]
- Cuboids [Dollar et al. 2005]
- Hessian [Willems et al. 2008]
- Regular dense sampling

**Four types of descriptors:**
- HoG/HoF [Laptev et al. 2008]
- Cuboids [Dollar et al. 2005]
- HoG3D [Kläser et al. 2008]
- Extended SURF [Willems’et al. 2008]

**Three human actions datasets:**
- KTH actions [Schuldt et al. 2004]
- UCF Sports [Rodriguez et al. 2008]
Space-time feature detectors

Harris3D

Cuboids

Hessian

Dense
## Results on KTH Actions

6 action classes, 4 scenarios, staged

### Detectors

<table>
<thead>
<tr>
<th>Detectors</th>
<th>Harris3D</th>
<th>Cuboids</th>
<th>Hessian</th>
<th>Dense</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG3D</td>
<td>89.0%</td>
<td>90.0%</td>
<td>84.6%</td>
<td>85.3%</td>
</tr>
<tr>
<td>HOG/HOF</td>
<td>91.8%</td>
<td>88.7%</td>
<td>88.7%</td>
<td>86.1%</td>
</tr>
<tr>
<td>HOG</td>
<td>80.9%</td>
<td>82.3%</td>
<td>77.7%</td>
<td>79.0%</td>
</tr>
<tr>
<td>HOF</td>
<td>92.1%</td>
<td>88.2%</td>
<td>88.6%</td>
<td>88.0%</td>
</tr>
<tr>
<td>Cuboids</td>
<td>-</td>
<td>89.1%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>E-SURF</td>
<td>-</td>
<td>-</td>
<td>81.4%</td>
<td>-</td>
</tr>
</tbody>
</table>

(Average accuracy scores)

- Best results for **sparse** Harris3D + HOF
- Dense features perform relatively poor compared to sparse features

[Wang, Ullah, Kläser, Laptev, Schmid, 2009]
## Results on UCF Sports

10 action classes, videos from TV broadcasts

### Detectors

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Harris3D</th>
<th>Cuboids</th>
<th>Hessian</th>
<th>Dense</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG3D</td>
<td>79.7%</td>
<td>82.9%</td>
<td>79.0%</td>
<td>85.6%</td>
</tr>
<tr>
<td>HOG/HOF</td>
<td>78.1%</td>
<td>77.7%</td>
<td>79.3%</td>
<td>81.6%</td>
</tr>
<tr>
<td>HOG</td>
<td>71.4%</td>
<td>72.7%</td>
<td>66.0%</td>
<td>77.4%</td>
</tr>
<tr>
<td>HOF</td>
<td>75.4%</td>
<td>76.7%</td>
<td>75.3%</td>
<td>82.6%</td>
</tr>
<tr>
<td>Cuboids</td>
<td>-</td>
<td>76.6%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>E-SURF</td>
<td>-</td>
<td>-</td>
<td>77.3%</td>
<td>-</td>
</tr>
</tbody>
</table>

(Average precision scores)

- Best results for **dense** + HOG3D

[Wang, Ullah, Kläser, Laptev, Schmid, 2009]
Results on Hollywood-2

12 action classes collected from 69 movies

<table>
<thead>
<tr>
<th>Detectors</th>
<th>Harris3D</th>
<th>Cuboids</th>
<th>Hessian</th>
<th>Dense</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG3D</td>
<td>43.7%</td>
<td>45.7%</td>
<td>41.3%</td>
<td>45.3%</td>
</tr>
<tr>
<td>HOG/HOF</td>
<td>45.2%</td>
<td>46.2%</td>
<td>46.0%</td>
<td>47.4%</td>
</tr>
<tr>
<td>HOG</td>
<td>32.8%</td>
<td>39.4%</td>
<td>36.2%</td>
<td>39.4%</td>
</tr>
<tr>
<td>HOF</td>
<td>43.3%</td>
<td>42.9%</td>
<td>43.0%</td>
<td>45.5%</td>
</tr>
<tr>
<td>Cuboids</td>
<td>-</td>
<td>45.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>E-SURF</td>
<td>-</td>
<td>-</td>
<td>38.2%</td>
<td>-</td>
</tr>
</tbody>
</table>

(Average precision scores)

- Best results for dense + HOG/HOF

[Wang, Ullah, Kläser, Laptev, Schmid, 2009]
More recent local methods I

- Y. and L. Wolf, "Local Trinary Patterns for Human Action Recognition ", ICCV 2009 + ECCV 2012 extension

- P. Matikainen, R. Sukthankar and M. Hebert "Trajectons: Action Recognition Through the Motion Analysis of Tracked Features" ICCV VOEC Workshop 2009,

More recent local methods II

- Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification, J.C. Niebles, C.-W. Chen and L. Fei-Fei, ECCV 2010

- Recognizing Human Actions by Attributes J. Liu, B. Kuipers, S. Savarese, CVPR 2011
## Dense trajectory descriptors

*Wang et al. CVPR’11*

<table>
<thead>
<tr>
<th></th>
<th>KTH</th>
<th>YouTube</th>
<th>Hollywood2</th>
<th>UCF sports</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laptev et al. [5]</td>
<td>91.8%</td>
<td>Liu et al. [45]</td>
<td>71.2%</td>
<td>Wang et al. [17]</td>
</tr>
<tr>
<td>Kovashka et al. [53]</td>
<td>94.53%</td>
<td>Ikizler-Cinbis et al. [35]</td>
<td>75.21%</td>
<td>Taylor et al. [58]</td>
</tr>
<tr>
<td>Yuan et al. [60]</td>
<td>93.7%</td>
<td>Brendel et al. [51]</td>
<td>77.8%</td>
<td>Ullah et al. [43]</td>
</tr>
<tr>
<td>Le et al. [52]</td>
<td>93.9%</td>
<td>Le et al. [52]</td>
<td>75.8%</td>
<td>Gilbert et al. [61]</td>
</tr>
<tr>
<td>Gilbert et al. [61]</td>
<td>94.5%</td>
<td>Bhattacharya et al. [62]</td>
<td>76.5%</td>
<td>Le et al. [52]</td>
</tr>
<tr>
<td>MBH</td>
<td>95.0%</td>
<td>MBH</td>
<td>80.6%</td>
<td>MBH</td>
</tr>
<tr>
<td>Combined</td>
<td>94.2%</td>
<td>Combined</td>
<td>84.1%</td>
<td>Combined</td>
</tr>
<tr>
<td>MBH+STP</td>
<td>95.3%</td>
<td>MBH+STP</td>
<td>83.0%</td>
<td>MBH+STP</td>
</tr>
<tr>
<td>Combined+STP</td>
<td>94.4%</td>
<td>Combined+STP</td>
<td>85.4%</td>
<td>Combined+STP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IXMAS</th>
<th>UIUC</th>
<th>Olympic Sports</th>
<th>UCF50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tran et al. [50]</td>
<td>80.22%</td>
<td>Tran et al. [50]</td>
<td>98.7%</td>
<td>Brendel et al. [56]</td>
</tr>
<tr>
<td>Junejo et al. [63]</td>
<td>79.6%</td>
<td>Niebles et al. [49]</td>
<td>72.1%</td>
<td></td>
</tr>
<tr>
<td>Wu et al. [54]</td>
<td>88.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MBH</td>
<td>91.8%</td>
<td>MBH</td>
<td>97.1%</td>
<td>MBH</td>
</tr>
<tr>
<td>Combined</td>
<td>93.5%</td>
<td>Combined</td>
<td>98.4%</td>
<td>Combined</td>
</tr>
<tr>
<td>MBH+STP</td>
<td>91.9%</td>
<td>MBH+STP</td>
<td>98.1%</td>
<td>MBH+STP</td>
</tr>
<tr>
<td>Combined+STP</td>
<td>93.6%</td>
<td>Combined+STP</td>
<td>98.3%</td>
<td>Combined+STP</td>
</tr>
</tbody>
</table>
Action recognition datasets

- KTH Actions, 6 classes, 2391 video samples [Schuldt et al. 2004]
- Weizman, 10 classes, 92 video samples, [Blank et al. 2005]

- UCF YouTube, 11 classes, 1168 samples, [Liu et al. 2009]

- Hollywood-2, 12 classes, 1707 samples, [Marszałek et al. 2009]

- UCF Sports, 10 classes, 150 samples, [Rodriguez et al. 2008]

- Olympic Sports, 16 classes, 783 samples, [Niebles et al. 2010]

- HMDB, 51 classes, ~7000 samples, [Kuehne et al. 2011]

- PASCAL VOC 2011 Action Classification Challenge, 10 classes, 3375 image samples
Where to go next?
Is action classification the right problem?

- Is action vocabulary well-defined?

Examples of “Open” action:

- What granularity of action vocabulary shall we consider?
Do we want to learn *person-throws-cat-into-trash-bin* classifier?

Source: http://www.youtube.com/watch?v=eYdUZdan5i8
How action recognition is related to other visual recognition tasks?
We can recognize cars and roads, What’s next?
What is missing in current methods?
What is missing in current methods?

Object detection/classification won’t help us to safely cross the street
A plain has crashed, the cabin is broken, somebody is likely to be injured or dead.
Limitations of Current Methods

What is unusual in this scene?  Is this scene dangerous?  What is intention of this person?

What is unusual in this scene?
Next challenge

Shift the focus of computer vision

Object, scene and action recognition

Recognition of objects’ function and people’s intentions

Is this a picture of a dog?
Is the person running in this video?

What people do with objects?
How they do it?
For what purpose?

Enable new applications
Motivation

• Exploit the link between human pose, action and object function.

• Use human actors as active sensors to reason about the surrounding scene.
Scene semantics from long-term observation of people

ECCV 2012

V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, A. Efros
Goal

Recognize objects by the way people interact with them.

Time-lapse “Party & Cleaning” videos

Lots of person-object interactions, many scenes on YouTube

Semantic object segmentation

- Red: Sofa
- Teal: Shelf
- Gray: Floor
- Yellow: Table
- Green: Tree
- Black: Wall
New “Party & Cleaning” dataset
Goal

Recognize objects by the way people interact with them.

Lots of person-object interactions, many scenes on YouTube

Time-lapse “Party & Cleaning” videos

Semantic object segmentation

- Sofa
- Shelf
- Floor
- Table
- Tree
- Wall
Pose vocabulary
Pose histogram
Some qualitative results
Quantitative results

<table>
<thead>
<tr>
<th></th>
<th>DPM</th>
<th>Hedau</th>
<th>(A+L)</th>
<th>(P)</th>
<th>(A+P)</th>
<th>(A+L+P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wall</td>
<td>—</td>
<td>75±3.9</td>
<td>76±1.6</td>
<td>76±1.7</td>
<td>82±1.2</td>
<td>81±1.3</td>
</tr>
<tr>
<td>Ceiling</td>
<td>—</td>
<td>47±20</td>
<td>53±8.0</td>
<td>52±7.4</td>
<td>69±6.7</td>
<td>69±6.6</td>
</tr>
<tr>
<td>Floor</td>
<td>—</td>
<td>59±3.1</td>
<td>64±5.5</td>
<td>65±3.6</td>
<td>76±3.2</td>
<td>76±2.9</td>
</tr>
<tr>
<td>Bed</td>
<td>31±20</td>
<td>12±7.2</td>
<td>14±5.0</td>
<td>21±5.8</td>
<td>27±13</td>
<td>26±13</td>
</tr>
<tr>
<td>Sofa/Armchair</td>
<td>26±9.4</td>
<td>26±10</td>
<td>34±3.3</td>
<td>32±6.5</td>
<td>44±5.4</td>
<td>43±5.8</td>
</tr>
<tr>
<td>Coffee Table</td>
<td>11±5.4</td>
<td>11±5.2</td>
<td>11±4.4</td>
<td>12±4.3</td>
<td>17±10</td>
<td>17±9.6</td>
</tr>
<tr>
<td>Chair</td>
<td>9.5±3.9</td>
<td>6.3±2.8</td>
<td>8.3±2.7</td>
<td>5.8±1.4</td>
<td>11±5.4</td>
<td>12±5.9</td>
</tr>
<tr>
<td>Table</td>
<td>15±6.4</td>
<td>18±3.8</td>
<td>17±3.9</td>
<td>16±7.1</td>
<td>22±6.2</td>
<td>22±6.4</td>
</tr>
<tr>
<td>Wardrobe/Cupboard</td>
<td>27±10</td>
<td>27±8.2</td>
<td>28±6.4</td>
<td>22±1.1</td>
<td>36±7.4</td>
<td>36±7.2</td>
</tr>
<tr>
<td>Christmas tree</td>
<td>50±3.3</td>
<td>55±12</td>
<td>72±1.8</td>
<td>20±6.0</td>
<td>76±6.2</td>
<td>77±5.5</td>
</tr>
<tr>
<td>Other Object</td>
<td>12±6.4</td>
<td>11±1.2</td>
<td>7.9±1.9</td>
<td>13±4.2</td>
<td>16±8.3</td>
<td>16±8.2</td>
</tr>
<tr>
<td>Average</td>
<td>23±1.8</td>
<td>31±2.0</td>
<td>35±2.4</td>
<td>30±1.7</td>
<td>43±4.4</td>
<td>43±4.3</td>
</tr>
</tbody>
</table>

A: Appearance (SIFT) histograms;
L: Location;
P: Pose histograms

Hedau: Hedau et al., Recovering the spatial layout of cluttered rooms. In: ICCV. (2009)
DPM: Felzenszwalb et al., Object detection with discriminatively trained part based models. PAMI (2010)
Using our model as pose prior

Given a bounding box and the ground truth segmentation, we fit the pose clusters in the box and score them by summing the joint’s weight of the underlying objects.
Using our model as pose prior
Conclusions

- BOF methods give state-of-the-art results for action recognition in realistic data. Better models are needed.

- Action classification (and temporal action localization) are often ill-defined problems.

- Targeting more realistic problems with functional models of objects and scenes can be the next challenge.