Benchmarking Human Activity Recognition

Haowei Liu, Perceptual Computing, Intel
Rogerio Feris, IBM T.J. Watson Research Center
Ming-Ting Sun, University of Washington
6/21/2012
Introduction & Motivation

• Need datasets to benchmark different aspects of algorithms

• Need a common ground for researchers to evaluate/compare the performances of their approaches

• Need datasets that are good representatives of the problem being solved (solving the dataset vs. solving the problem)

• Focus and introduce the state of the art benchmarking video datasets for activity recognition
Outline

- Benchmark for Kinematic Activities
- Movie/Web Videos Benchmarks
- Benchmarks for Assisted Daily Life (ADL) Activities
- Video Surveillance Benchmarks
- Benchmarks for Group Activities
- Multi-Camera Benchmarks
- RGB-D Benchmarks
- Egocentric Benchmarks
Benchmarks for Kinematics Activities

KTH* and Weizmann+

- Low resolution (<200x200)
- Few background clutters
- Mostly frontal and side-on camera viewing angles
- High accuracy reported by many papers already

![Graphs showing performance metrics for KTH and Weizmann datasets](image)

Movie/Web Benchmarks

- Multiple camera viewing angles
- Camera motions
- Video qualities/ resolutions and clutters vary
- Multiple moving objects

**HMDB51 Dataset**

- Large set (51) of activity categories
- High intra-category variations
- Drastic appearance, scale, position changes of actors. Variations in camera motion and viewpoints

*HMDB51 sample frames*

Reported accuracy: 38.00% in [SadanandCVPR2012Action]
## Movie/Web Benchmarks

### Other datasets and reported results

<table>
<thead>
<tr>
<th>Classes</th>
<th>Clips</th>
<th>Resolution</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF Youtube(^1)</td>
<td>11</td>
<td>3185</td>
<td>240*320</td>
</tr>
<tr>
<td>UCF 50(^3)</td>
<td>50</td>
<td>&gt;5000</td>
<td>240*320</td>
</tr>
<tr>
<td>UCF Sports(^2)</td>
<td>9</td>
<td>182</td>
<td>480*720</td>
</tr>
<tr>
<td>Coffee/Cigarette(^4)</td>
<td>2</td>
<td>264</td>
<td>240*500</td>
</tr>
<tr>
<td>Hollywood1(^5)</td>
<td>8</td>
<td>400</td>
<td>240*500</td>
</tr>
<tr>
<td>Hollywood2(^6)</td>
<td>12</td>
<td>1707</td>
<td>240*500</td>
</tr>
<tr>
<td>Olympics(^7)</td>
<td>16</td>
<td>800</td>
<td>360*450</td>
</tr>
<tr>
<td>TRECVID MED(^{14})</td>
<td>15</td>
<td>32061</td>
<td>vary</td>
</tr>
</tbody>
</table>

\(^1\) [LiuCVPR09Recognizing], \(^2\) [RodriguezCVPR2008Action], \(^3\) [UCF50], \(^4\) [LaptevICCV07Retrieve], \(^5\) [LaptevCVPR2008Learning], \(^6\) [MarszalekCVPR2009Actions], \(^7\) [NieblesECCV2010Modeling], \(^8\) [GaidonCVPR2011Actom], \(^9\) [WangCVPR2011Action], \(^10\) [BrendelICCV2011Learning], \(^11\) [GilbertPAMI2010Action], \(^12\) [SadanandCVPR2012Action], \(^13\) [TangCVPR2012Learning], \(^14\) [TRECVID2011]
Benchmarks for Assisted Daily Living (ADL) Activities

ADL65 Dataset*

- Large set (65 categories) of high-resolution kitchen activities
- Fine-grained activities (low inter-class variability)
- Detailed annotations including time intervals and poses
- Provide classification & detection tasks
- Similar dataset: URADL+
  - high-res 10 kitchen activities.
  - reported accuracy: 96% in [WangCVPR2011ActionST]
- Scene and object info highly correlates with the activities

Reported average precision for
Classification: 59.2%
Detection: 45.0%
in [RohrbachCVPR2012Database]

* [RohrbachCVPR2012Database]
+ [MessingICCV2009Activity]
Benchmarks for Group Activities

**UT-Interaction***

- 6 classes of 2-person interaction activities
- Detailed annotation with time intervals/bounding boxes
- Camera jittering
- Pedestrians in the background
- Concurrent activities
- Similar datasets: Collective Dataset\(^1\), BEHAVE\(^2\)

*\[RyooICPR2010Overview\]

\(^1\)[ChoiICCV2009Collective]

\(^2\)[BlunsdenBMVA2010Behave]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AmerICCV2011Chain</td>
<td>75.75%</td>
</tr>
<tr>
<td>BrendelICCV2011Learning</td>
<td>78%</td>
</tr>
<tr>
<td>RyooICCV2011Early</td>
<td>85%</td>
</tr>
<tr>
<td>GaurICCV2011SFG</td>
<td>72%</td>
</tr>
</tbody>
</table>
Virat Ground Video Dataset*

- Realistic scenarios (non-actors)
- Multi spatial-temporal resolutions
- Diverse scenes (16 scenes) and event types (23)
- Multiple objects and concurrent activities
- Different camera perspectives
- Detailed annotations including time intervals, bounding boxes, and tracks
- People/facility, people/vehicle interaction

* [OhCVPR2011Virat]
Benchmarks for Long Term Surveillance

Virat Aerial Video Dataset

- Camera motion
- Low resolution of human figures
- Similarity across actions from high altitude
- Time-varying viewpoints and scales
- Shadows and interrupted tracking

sample frames for different scenes and viewpoints
Benchmarks for Long Term Surveillance

Evaluations

Average accuracy on aerial dataset: 38% in [ChenCVPR11Modeling]

Average hit rate on ground dataset: 33% in [OhCVPR2011Virat]

[OhCVPR2011Virat]

Average hit rate on ground dataset: 33% in [OhCVPR2011Virat]

Average accuracy on aerial dataset: 38% in [ChenCVPR11Modeling]

[ChenCVPR11Modeling]
Benchmarks for Long Term Surveillance

TRECVID 2011 Surveillance Event Detection (SED) Dataset

- ~100 hours indoor airport surveillance videos
- 7 events including 2 single person events, 2 person-object interaction events and 3 multi-person events
- Different background clutters and traffic due to different camera placements
- Best Normalized Detection Cost Rate (NDCR): 0.8~2

\[ NDCR = P_{miss} + \frac{Rate_{FA}}{Rate_{event}} \]

Perfect NDCR: 0

camera placements: controlled access door, waiting area, debarkation area, elevator door, transit area
Multi-Camera Benchmarks

IXMAS Dataset*

- Provide 5-view videos of 13 kinematic activities
- Provide silhouette, reconstructed volumes, and calibration information
- 3D information is available
- Suitable to evaluate view dependent models

*WeinlandICCV07Action

![Sample frames](image_url)
Multi-Camera Benchmarks

IXMAS Dataset - Evaluations

- **Accuracy by using multi-cameras***

<table>
<thead>
<tr>
<th>Combinations</th>
<th>1,2,3,4,5</th>
<th>1,2,3,4</th>
<th>1,2,3,5</th>
<th>1,2,3</th>
<th>1,3,5</th>
<th>1,3</th>
<th>2,4</th>
<th>3,5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>88.20%</td>
<td>88.20%</td>
<td>89.40%</td>
<td>87.70%</td>
<td>88.40%</td>
<td>86.60%</td>
<td>82.40%</td>
<td>83.30%</td>
</tr>
</tbody>
</table>

Recognition using multi-view information (5~15% improvements over single view)

- **View transfer evaluations***

Evaluating cross-view model effectiveness

<table>
<thead>
<tr>
<th>%</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* [WuCVPR2011Action]

+ [LiCVPR2012Discriminative]
RGB-D Benchmarks
MSRAAction3D and DailyActivity3D*

- Both recorded with commercially available depth sensor
- MSRAAction3D consists of 20 kinematic activities
- DailyActivity3D consists of 16 living room activities involving different objects e.g. writing on a paper, answering phone

Accuracy: 88.2% *

MSRAAction3D sample frames

Accuracy: 85.7%*

*[WangCVPR2012Mining]

DailyActivity3D sample frames
Egocentric Benchmarks

Egocentric ADL Datasets\(^1\)

- Complex object interactions (42 objects)
- Large set of actions (18 actions), and sites (20 homes)
- Longer activities
- Large variations of object appearance
- Scene/Clutter variations
- Similar datasets: GTEA\(^2\) (GeorgiaTech Egocentric) and Intel Egocentric Dataset\(^3\)
- Strong priors for hand locations

Reported classification accuracy: 40.6\% in [PirsiavashCVPR2012 Detecting]

\(^1\)[PirsiavashCVPR2012Detecting]  
\(^2\)[FathiICCV2011Understanding]  
\(^3\)[RenEgo2009Ego]
Egocentric Benchmarks

Other Egocentric Datasets

- UEC Sports\(^1\)
  - Large set of outdoor activities
  - Large motion and blur

- GeorgiaTech\(^2\)
  - Egocentric social activity recognition

- Other datasets: Egocentric novelty detection\(^3\), and egocentric video summary\(^4\)

\(^1\)[KitaniCVPR2011Fast], \(^2\)[FathiCVPR2012Social], \(^3\)[AghazadehCVPR2011Novelty], \(^4\)[LeeCVPR2012Discovering]
<table>
<thead>
<tr>
<th>Class</th>
<th>Classes</th>
<th>Resolution</th>
<th>Sequences</th>
<th>Frames</th>
<th>URL</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>KTH</td>
<td>6</td>
<td>160*120</td>
<td>600</td>
<td>~500k</td>
<td><a href="http://www.nada.kth.se/cvap/actions/">http://www.nada.kth.se/cvap/actions/</a></td>
<td>N.A</td>
</tr>
<tr>
<td>Weizmann</td>
<td>10</td>
<td>180*144</td>
<td>90</td>
<td>~8k</td>
<td><a href="http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html">http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html</a></td>
<td>extracted foreground mask</td>
</tr>
<tr>
<td>HMDB51</td>
<td>51</td>
<td>~240*480</td>
<td>7000</td>
<td>~800K</td>
<td><a href="http://serre-lab.clps.brown.edu/resources/HMDB/">http://serre-lab.clps.brown.edu/resources/HMDB/</a></td>
<td>meta info regarding video quality, angle, and camera motion</td>
</tr>
<tr>
<td>UCF Youtube</td>
<td>11</td>
<td>240*320</td>
<td>1168</td>
<td>~80K</td>
<td><a href="http://www.cs.ucf.edu/vision/public_html/data.html#UCF%20YouTube%20Action%20Dataset">http://www.cs.ucf.edu/vision/public_html/data.html#UCF%20YouTube%20Action%20Dataset</a></td>
<td>N.A</td>
</tr>
<tr>
<td>UCF 50</td>
<td>50</td>
<td>240*320</td>
<td>&gt;500</td>
<td>&gt;150K</td>
<td><a href="http://www.cs.ucf.edu/vision/public_html/data.html#UCF50">http://www.cs.ucf.edu/vision/public_html/data.html#UCF50</a></td>
<td>N.A</td>
</tr>
<tr>
<td>Coffee/Cigarette</td>
<td>2</td>
<td>240*500</td>
<td>1</td>
<td>36K</td>
<td><a href="http://www.di.ens.fr/~laptev/download.html">http://www.di.ens.fr/~laptev/download.html</a></td>
<td>Space-time cuboid, key frame and head position</td>
</tr>
<tr>
<td>Hollywood1</td>
<td>8</td>
<td>240*500</td>
<td>~600</td>
<td>~400K</td>
<td><a href="http://www.di.ens.fr/~laptev/download.html">http://www.di.ens.fr/~laptev/download.html</a></td>
<td>Time interval</td>
</tr>
<tr>
<td>Hollywood2</td>
<td>12</td>
<td>240*500</td>
<td>~600</td>
<td>~600K</td>
<td><a href="http://www.di.ens.fr/~laptev/download.html">http://www.di.ens.fr/~laptev/download.html</a></td>
<td>Time interval</td>
</tr>
<tr>
<td>Dataset</td>
<td>Classes</td>
<td>Resolution</td>
<td>Sequences</td>
<td>Frames</td>
<td>URL</td>
<td>Annotation</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>------------</td>
<td>-----------</td>
<td>--------</td>
<td>------------------------------------------</td>
<td>-------------------------------------------</td>
</tr>
<tr>
<td>ADL65</td>
<td>65</td>
<td>1624x1224</td>
<td>44</td>
<td>~881K</td>
<td>N.A</td>
<td>time-interval and body pose</td>
</tr>
<tr>
<td>URADL</td>
<td>10</td>
<td>1280x720</td>
<td>50</td>
<td>~75K</td>
<td><a href="http://www.cs.rochester.edu/~rmessing/uradl/">http://www.cs.rochester.edu/~rmessing/uradl/</a></td>
<td>N.A</td>
</tr>
<tr>
<td>Ut-interaction</td>
<td>6</td>
<td>480*720</td>
<td>20</td>
<td>~36K</td>
<td><a href="http://cvrc.ece.utexas.edu/SDHA2010/Human">http://cvrc.ece.utexas.edu/SDHA2010/Human</a> Interaction.html</td>
<td>time-interval and bounding boxes of subjects</td>
</tr>
<tr>
<td>Collective</td>
<td>5</td>
<td>480*720</td>
<td>44</td>
<td>~13K</td>
<td><a href="http://www.eecs.umich.edu/vision/activity-dataset.html">http://www.eecs.umich.edu/vision/activity-dataset.html</a></td>
<td>locations of the subjects, bounding box, and pose info</td>
</tr>
<tr>
<td>BEHAVE</td>
<td>10</td>
<td>480*640</td>
<td>4</td>
<td>~300K</td>
<td><a href="http://groups.inf.ed.ac.uk/vision/BEHAVEVEDATA/INTERACTIONS/">http://groups.inf.ed.ac.uk/vision/BEHAVEVEDATA/INTERACTIONS/</a></td>
<td>Interval, group id, bounding box</td>
</tr>
<tr>
<td>IXMAS</td>
<td>13</td>
<td>390x291</td>
<td>36</td>
<td>~40K*</td>
<td><a href="http://4drepository.inrialpes.fr/">http://4drepository.inrialpes.fr/</a></td>
<td>silhouettes, reconstructed volume</td>
</tr>
<tr>
<td>VIRAT</td>
<td>23</td>
<td>1920 x 1080</td>
<td>N.A</td>
<td>~1620K</td>
<td><a href="http://www.viratdata.org/">http://www.viratdata.org/</a></td>
<td>object tracks, subject bounding boxes, event interval</td>
</tr>
</tbody>
</table>
## Summary Table

<table>
<thead>
<tr>
<th></th>
<th>classes</th>
<th>res.</th>
<th>seqs</th>
<th>frames</th>
<th>url</th>
<th>annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>EgoADL</td>
<td>18</td>
<td>N.A</td>
<td>N.A</td>
<td>1M</td>
<td>N.A</td>
<td>object bounding box, tracks and labels, hand positions</td>
</tr>
</tbody>
</table>
Conclusions and Challenges

- We covered most of the major benchmarking datasets – good starting points for people new to the field
- Creating a benchmark that captures the level of complexity of real world problems is still hard
- Few reported cross dataset performance or model generality evaluation
- The trend is to create large scale benchmarking datasets with detailed annotations - need better tools (e.g. LabelMe Video) or technology (e.g. Machine-aided Mechanical Turk) for good quality and large annotation tasks
Bibliography


Bibliography (continued)


Bibliography (continued)


Bibliography (continued)


