The PASCAL Visual Object Classes Challenge 2011 (VOC2011)

Part 5 – Action Classification Taster Challenge

Mark Everingham
Luc Van Gool
Chris Williams
John Winn
Andrew Zisserman
Action Classification Taster Challenge

- Given the bounding box of a person, predict whether they are performing a given action.

  - Playing Instrument?
  - Reading?

- Encourage research on **still-image** activity recognition: more detailed understanding of image.
Ten Action Classes + “Other”

- Jumping
- Phone
- Playing Instrument
- Reading
- Riding Bike
- Riding Horse
- Running
- Taking Photo
- Using Computer
- Walking
Dataset Statistics

- ~5 times size of VOC2010 dataset

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actions</td>
<td>2,772</td>
<td>(608)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2,788</td>
</tr>
<tr>
<td>“Other”</td>
<td>575</td>
<td>(0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>575</td>
</tr>
</tbody>
</table>

VOC2010 counts shown in brackets

- Minimum ~200 people per action category
- Only subset of people are annotated (bounding box + actions)
- Actions are not mutually-exclusive
Methods

- **Train on VOC data: 7 Methods, 5 Groups**
  - Image classification within bounding box
    - SVM, bag of words/spatial pyramid
    - Multiple features: SIFT, PHOG, color SIFT, etc.
    - Random Forest, dense spatial pyramid features
  - Multi-scale context (bounding box neighbourhood, image)
  - Discriminative Saliency Map weighting BoW features
  - Unsupervised learning of relevant object detector

- **Train on own data: 3 Methods, 3 Groups**
  - Poselets + image context
  - Attribute classifiers + object detectors (trained on ImageNet)
  - Poselets + foreground/background descriptors (segmentation)
## AP by Class/Method

**VOC data**

<table>
<thead>
<tr>
<th>Method</th>
<th>Jumping</th>
<th>Phoning</th>
<th>Playing Instrument</th>
<th>Reading</th>
<th>Riding Bike</th>
<th>Riding Horse</th>
<th>Running</th>
<th>Taking Photo</th>
<th>Using Computer</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAENLEAR_DSAL</td>
<td>62.1</td>
<td>39.7</td>
<td>60.5</td>
<td>33.6</td>
<td>80.8</td>
<td>83.6</td>
<td>80.3</td>
<td>23.2</td>
<td>53.4</td>
<td>50.2</td>
</tr>
<tr>
<td>CAENLEAR_HOBJ_DSAL</td>
<td>71.6</td>
<td>50.7</td>
<td>77.5</td>
<td>37.8</td>
<td>86.5</td>
<td>89.5</td>
<td>83.8</td>
<td>25.1</td>
<td>58.9</td>
<td>59.2</td>
</tr>
<tr>
<td>MISSOURI_SSLMF</td>
<td>58.8</td>
<td>36.8</td>
<td>48.5</td>
<td>30.6</td>
<td>81.5</td>
<td>83.0</td>
<td>78.5</td>
<td>21.3</td>
<td>50.7</td>
<td>53.8</td>
</tr>
<tr>
<td>NUDTCONTEXT</td>
<td>65.9</td>
<td>41.5</td>
<td>57.4</td>
<td>34.7</td>
<td>88.8</td>
<td>90.2</td>
<td>87.9</td>
<td>25.7</td>
<td>54.5</td>
<td>59.5</td>
</tr>
<tr>
<td>NUDT_LL_SEMANTIC</td>
<td>66.3</td>
<td>41.3</td>
<td>53.9</td>
<td>35.2</td>
<td>88.8</td>
<td>90.0</td>
<td>87.6</td>
<td>25.5</td>
<td>53.7</td>
<td>58.2</td>
</tr>
<tr>
<td>STANFORD_RF_DENSEFTR_SVM</td>
<td>66.0</td>
<td>41.0</td>
<td>60.0</td>
<td>41.5</td>
<td>90.0</td>
<td>92.1</td>
<td>86.6</td>
<td>28.8</td>
<td>62.0</td>
<td>65.9</td>
</tr>
<tr>
<td>WVU_SVM-PHOWN</td>
<td>42.5</td>
<td>29.5</td>
<td>32.1</td>
<td>26.7</td>
<td>48.5</td>
<td>46.3</td>
<td>59.2</td>
<td>13.5</td>
<td>24.3</td>
<td>35.6</td>
</tr>
</tbody>
</table>

## Own data

<table>
<thead>
<tr>
<th>Method</th>
<th>Jumping</th>
<th>Phoning</th>
<th>Playing Instrument</th>
<th>Reading</th>
<th>Riding Bike</th>
<th>Riding Horse</th>
<th>Running</th>
<th>Taking Photo</th>
<th>Using Computer</th>
<th>Walking</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERKELEY_ACTION_POSELETS</td>
<td>59.5</td>
<td>31.3</td>
<td>45.6</td>
<td>27.8</td>
<td>84.4</td>
<td>88.3</td>
<td>77.6</td>
<td>31</td>
<td>47.4</td>
<td>57.6</td>
</tr>
<tr>
<td>STANFORD_COMBINE_ATTR_PART</td>
<td>66.7</td>
<td>41.1</td>
<td>60.8</td>
<td>42.2</td>
<td>90.5</td>
<td>92.2</td>
<td>86.2</td>
<td>28.8</td>
<td>63.5</td>
<td>64.2</td>
</tr>
<tr>
<td>STANFORD_MAPSVM_POSELET</td>
<td>27</td>
<td>29.3</td>
<td>28.3</td>
<td>23.8</td>
<td>71.9</td>
<td>82.4</td>
<td>67.3</td>
<td>20.1</td>
<td>26</td>
<td>46.4</td>
</tr>
</tbody>
</table>
Precision/Recall – Riding Horse

All results

- STANFORD_RF_DENSEFTR_SVM (92.1)
- NUDTCONTEXT (90.2)
- NUDT_LL_SEMANTIC (90.0)
- CAENLEAR_HOBI_DSAL (89.5)
- CAENLEAR_DSAL (83.6)
- MISSOURI_SSLMF (83.0)
- WVU_SVM-PHOW (46.3)
Precision/Recall – Taking Photo

The graph shows the precision and recall of various methods for identifying taking photo actions. The methods are:

- STANFORD_RF_DENSEFTR_SVM (28.8)
- NUDT_CONTEXT (25.7)
- NUDT_LL_SEMANTIC (25.5)
- CAENLEAR_HOBJ_DSAL (25.1)
- CAENLEAR_DSAL (23.2)
- MISSOURI_SSLMF (21.3)
- WVU_SVM-PHOW (13.5)
Max AP: 92.1% (riding horse) ... 28.8% (taking photo)
Median AP by Method

<table>
<thead>
<tr>
<th>Method</th>
<th>AP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAENLEAR_HOBJ_DSAL</td>
<td>65.4</td>
</tr>
<tr>
<td>STANFORD_RF_DENSEFTR_S</td>
<td>64.0</td>
</tr>
<tr>
<td>NUDT_CONTEXT</td>
<td>58.5</td>
</tr>
<tr>
<td>CAENLEAR_DSAL</td>
<td>57.0</td>
</tr>
<tr>
<td>NUDT_LL_SEMANTIC</td>
<td>56.1</td>
</tr>
<tr>
<td>MISSOURI_SLMF</td>
<td>52.3</td>
</tr>
<tr>
<td>WVU_SVM-PHOW</td>
<td>33.9</td>
</tr>
</tbody>
</table>
“True Positives” - Riding Horse

STANFORD_RF_DENSEFTR_SVM

NUDT_CONTEXT

CAENLEAR_HOBJ_DSAL
“False Negatives” - Riding Horse

STANFORD_RF_DENSEFTR_SVM

NUDT_CONTEXT

CAENLEAR_HOBJ_DSAL
“False Positives” - Riding Horse

STANFORD_RF_DENSEFTR_SVM

NUDT_CONTEXT

CAENLEAR_HOBJ_DSAL
“True Positives” - Jumping

CAENLEAR_HOBJ_DSAL

NUDT_LL_SEMANTIC

STANFORD_RF_DENSEFTR_SVM
“False Negatives” - Jumping

CAENLEAR_HOBJ_DSAL

NUDT_LL_SEMANTIC

STANFORD_RF_DENSEFTR_SVM
“False Positives” - Jumping

CAENLEAR_HOBJ_DSAL

NUDT_LL_SEMANTIC

STANFORD_RF_DENSEFTR_SVM
“True Positives” - Taking Photo

STANFORD_RF_DENSEFTR_SVM

NUDT_CONTEXT

CAENLEAR_HOBJ_DSAL
“False Negatives” - Taking Photo

STANFORD_RF_DENSEFTR_SVM

NUDT_CONTEXT

CAENLEAR_HOBJ_DSAL
“False Positives” - Taking Photo

STANFORD_RF_DENSEFTR_SVM

NUDTCONTEXT

CAENLEAR_HOBJ_DSAL
Prizes

- **Winner**
  - **STANFORD_RF_DENSEFTR_SVM**
    Bangpeng Yao, Aditya Khosla, Li Fei-Fei
    *Stanford University*

- **Runners-up**
  - **CAENLEAR_HOBJ_DSAL**
    Gaurav Sharma, Alessandro Prest, Frederic Jurie, Vittorio Ferrari, Cordelia Schmid
    *University of Caen, INRIA LEAR*
  - **NUDT_CONTEXT/NUDT_LL_SEMANTIC**
    Li Zhou, Dewen Hu, Zongtan Zhou
    *National University of Defense Technology, China*