Regionlets for Generic Object Detection

Xiaoyu Wang†, Ming Yang‡, Shenghuo Zhu† and Yuanqing Lin†

†NEC Labs America, Inc.
Cupertino, CA 95014, USA

‡Facebook, Inc.
Menlo Park, CA 94025, USA
Generic object detection

Train

Sheep

Potted plant
State of the art

Performance evolution on the PASCAL VOC 2007 object detection dataset (mean AP)

- 21.3% in 2008
- 26.4% in 2009
- 29.6% in 2010
- 33.8% in 2011
- 37.7% in 2013
- 33.7% in 2013
- 41.7% in 2013

2. A. Vedaldi, et. al. Multiple Kernels for Object Detection. ICCV 2009
State of the art

- The two representative object detection frameworks

1. Scanning window with Deformable Part-based Model (DPM)
2. Selective Search with Spatial Pyramid Matching (SS_SPM)
3. Regionlets (No deep CNN feature yet 😊)

<table>
<thead>
<tr>
<th></th>
<th>33.8%</th>
<th>33.7%</th>
<th>41.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Object Detection

12/15/2013 Regionlets for Generic Object Detection 5
Review: Feature extraction

- Feature design
  - HOG
  - SIFT, and many others...
- Feature extraction
  - Densely extracted over N x N pixel cells
Review: Deformation handling

- Deformable Part-based Model (DPM)
  - Specify the number of deformable parts
- Spatial Pyramid Matching
  - Specify the number of pyramids to build

- Do we have to pre-define model parameters to handle different degrees of deformation?
Review: Multi scales/viewpoints

- **DPM**
  - Resize an image to detect objects at a fixed scale
  - Multiple models, each deals with one viewpoint

- **Spatial Pyramid Matching**
  - No need to resize the image
  - One model, a codebook is used to encode features

- *Can we learn a model that can be easily adapted to arbitrary scales and viewpoints?*
Motivation

- Motivation: A flexible and general object-level representation with
  - Hassle free deformation handling
  - Arbitrary scales and aspect ratio handling

Regionlets!
Detection framework

2. B. Alexe, et. al. Measuring the objectness of image windows. PAMI 2012
Regionlet: Definition

- Region($R$): Feature extraction region
- Regionlet($r_1, r_2, r_3$): A sub-region in a feature extraction area whose position/resolution are relative and normalized to a detection window

Figure 1
Regionlet: Definition (cont.)

- Relative normalized position

![Diagram showing relative normalized position with images]

Figure 2
Regionlet: Feature extraction

Could be SIFT, HOG, LBP, Covariance features, whatever feature your like!
Regionlets: Training

- Constructing the regions/regionlets pool
  - Small region, fewer regionlets -> fine spatial layout
  - Large region, more regionlets -> robust to deformation

- Learning realBoost\(^1\) cascades
  - 16K region/regionlets candidates for each cascade
  - Learning of each cascade stops when the error rate is achieved (1% for positive, 37.5% for negative)
  - Last cascade stops after collecting 5000 weak classifiers
  - Result in 4-7 cascades
  - 2-3 hours to finish training one category on a 8-core machine

---

Regionlets: Testing

- No image resizing
- Any scale, any aspect ratio
- Adapt the model size to the same size as the object candidate bounding box

Regionlet based model  Applied to candidate boxes
Experiments

- **Datasets**
  - PASCAL VOC 2007, 2010
    - 20 object categories
  - ImageNet Large Scale Object Detection Dataset
    - 200 object categories

- **Investigated Features**
  - HOG
  - LBP
  - Covariance
  - Deep Convolutional Neural Network (DCNN) feature
    (only for the ImageNet challenge)
### Experiments: PASCAL VOC

Table 1. Performance on the PASCAL VOC 2007 dataset (Evaluated using Average Precision or mean Average Precision: mAP, no DCNN feature, no outside data)

<table>
<thead>
<tr>
<th></th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPM</td>
<td>33.2</td>
<td>60.3</td>
<td>10.2</td>
<td>16.1</td>
<td>27.3</td>
<td>54.3</td>
<td>58.2</td>
<td>23.0</td>
<td>20.0</td>
<td>24.1</td>
<td>26.7</td>
<td>12.7</td>
<td>58.1</td>
<td>48.2</td>
<td>43.2</td>
<td>12.0</td>
<td>21.1</td>
<td>36.1</td>
<td>46.0</td>
<td>43.5</td>
<td>33.7</td>
</tr>
<tr>
<td>SS_SPM</td>
<td>43.5</td>
<td>46.5</td>
<td>10.4</td>
<td>12.0</td>
<td>9.3</td>
<td>49.4</td>
<td>53.7</td>
<td>39.4</td>
<td>12.5</td>
<td>36.9</td>
<td>42.2</td>
<td>26.4</td>
<td>47.0</td>
<td>52.4</td>
<td>23.5</td>
<td>12.1</td>
<td>29.9</td>
<td>36.3</td>
<td>42.2</td>
<td>48.8</td>
<td>33.8</td>
</tr>
<tr>
<td>Objectness</td>
<td>28.6</td>
<td>54.5</td>
<td>1.1</td>
<td>14.2</td>
<td>26.7</td>
<td>42.0</td>
<td>50.2</td>
<td>18.2</td>
<td>16.5</td>
<td>17.5</td>
<td>26.2</td>
<td>7.7</td>
<td>46.8</td>
<td>39.6</td>
<td>36.2</td>
<td>11.6</td>
<td>14.1</td>
<td>23.1</td>
<td>34.8</td>
<td>39.2</td>
<td>27.4</td>
</tr>
<tr>
<td>Regionlets-S</td>
<td>50.8</td>
<td>44.6</td>
<td>17.0</td>
<td>23.5</td>
<td>16.7</td>
<td>48.9</td>
<td>67.6</td>
<td>39.1</td>
<td>16.5</td>
<td>32.4</td>
<td>44.0</td>
<td>18.9</td>
<td>52.1</td>
<td>46.6</td>
<td>36.6</td>
<td>13.8</td>
<td>33.8</td>
<td>27.6</td>
<td>55.5</td>
<td>50.4</td>
<td>36.8</td>
</tr>
<tr>
<td>Regionlets-M</td>
<td>54.2</td>
<td>52.0</td>
<td>20.3</td>
<td>24.0</td>
<td>20.1</td>
<td>55.5</td>
<td>68.7</td>
<td>42.6</td>
<td>19.2</td>
<td>44.2</td>
<td>49.1</td>
<td>26.6</td>
<td>57.0</td>
<td>54.5</td>
<td>43.4</td>
<td>16.4</td>
<td>36.6</td>
<td>37.7</td>
<td>59.4</td>
<td>52.3</td>
<td>41.7</td>
</tr>
</tbody>
</table>

Table 2: Performance comparison with state of the art

<table>
<thead>
<tr>
<th></th>
<th>VOC 2007</th>
<th>VOC 2010</th>
<th>Results year</th>
</tr>
</thead>
<tbody>
<tr>
<td>DPM(WC) [12]</td>
<td>35.4</td>
<td>33.4</td>
<td>2008</td>
</tr>
<tr>
<td>UCI.2009 [7]</td>
<td>27.1</td>
<td>N/A</td>
<td>2009</td>
</tr>
<tr>
<td>NLPR(WC) [10]</td>
<td>N/A</td>
<td>36.8</td>
<td>2010</td>
</tr>
<tr>
<td>MITUCLA(WC) [10]</td>
<td>N/A</td>
<td>36.0</td>
<td>2010</td>
</tr>
<tr>
<td>UVA [10]</td>
<td>N/A</td>
<td>32.9</td>
<td>2010</td>
</tr>
<tr>
<td>MIT.2010 [31]</td>
<td>29.6</td>
<td>N/A</td>
<td>2010</td>
</tr>
<tr>
<td>Song et al. (WC) [23]</td>
<td>37.7</td>
<td>36.8</td>
<td>2011</td>
</tr>
<tr>
<td>Li et al. (WC) [18]</td>
<td>35.2</td>
<td>N/A</td>
<td>2011</td>
</tr>
<tr>
<td>SS_SPM [25]</td>
<td>33.8</td>
<td>34.1</td>
<td>2011</td>
</tr>
<tr>
<td>Cinbis et al. (WC) [5]</td>
<td>35.0</td>
<td>N/A</td>
<td>2012</td>
</tr>
<tr>
<td>Ours (Regionlets)</td>
<td>41.7</td>
<td>39.7</td>
<td>2013</td>
</tr>
</tbody>
</table>
## Experiments: ImageNet

### ImageNet Challenge

<table>
<thead>
<tr>
<th>Methods</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>UvA-EuVision</td>
<td>22.6% (with DCNN feature)</td>
</tr>
<tr>
<td>Regionlets with deep features&lt;sup&gt;(1)&lt;/sup&gt;</td>
<td>20.9% (with DCNN feature)</td>
</tr>
<tr>
<td>Regionlets without deep features&lt;sup&gt;(2)&lt;/sup&gt;</td>
<td>19.6% (no DCNN feature)</td>
</tr>
<tr>
<td>OverFeat-NYU</td>
<td>19.4% (DCNN)</td>
</tr>
<tr>
<td>Toronto A</td>
<td>11.2% (N/A)</td>
</tr>
<tr>
<td>SYSU_Vision</td>
<td>10.5% (N/A)</td>
</tr>
</tbody>
</table>

<sup>(1)</sup> The result of using only a single method and single set of parameters, no context. No combining!

<sup>(2)</sup> The result of using traditional features only – no DCNN features were used.

Check our presentation at the ILSVRC2013 workshop for more details!
Running speed

- 0.2 second per image using a single core if candidate bounding boxes are given, real time (>30 frames per second) using 8 cores.

- 2 seconds per image to generate candidate bounding boxes.

- 2-3 hours to finish training one category on a 8-core machine.
Conclusions

- A new object representation for object detection
  - Non-local max-pooling of regionlets
  - Relative normalized locations of regionlets
  - Flexibility to incorporate various types of features

- A principled data-driven detection framework, effective in handling deformation, multiple scales, multiple viewpoints

- Superior performance with a fast running speed