Regionlets for Generic Object Detection

A test on ImageNet

Xiaoyu Wang †
Miao Sun ‡
Yuanqing Lin †
Tony X. Han ‡
Shenghuo Zhu †

† NEC Laboratories America
Relentless passion for innovation

‡ University of Missouri
Introduction

- Generic object detection is challenging
  - Rich deformation
  - Arbitrary scales
  - Arbitrary viewpoints

- Limitations of current state of the art
  - Hand-crafted parameters to handle different degrees of deformation
  - Sub-optimal multiple scales/viewpoints handling
Motivation

- A flexible and general object-level representation
  - Data-driven deformation handling
  - Multiple scales/viewpoints handling using a single and flexible model (Detecting an object at its original scale and aspect ratio)
  - Fast and easy to be extended with different features
Detection Framework

1. B. Alexe, et. al. What is an object? CVPR 2010
3. X. Wang, et. al. Regionlets for Generic Object Detection. ICCV 2013
What is regionlet?

- 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

Figure 2

Traditional:
\[(l, t, r, b)\]
\[(50, 50, 180, 180)\]

Normalized:
\[
\left(\frac{l}{w'}, \frac{t}{h'}, \frac{r}{w'}, \frac{b}{h'}\right)
\]
\[(.25, .25, .90, .90)\]
Regionlet: Feature extraction

Figure 3

Non-local pooling

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

- Constructing the regions/regionlets pool
  - Uniformly sample the position/configuration space of regions/regionlets

- 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

Deformation Handling

- Two-layers deformation handling
  - Data-driven feature extraction region
    - Larger region -> more robust to deformation
    - Small region -> finer spatial layout
  - Data-driven non-local max-pooling over regionlets
    - Permutation invariance among regionlets
    - Exclusive feature representation among regionlets
Arbitrary scale/viewpoints handling

- Coordinates of regionlets are normalized in a model
- Absolute regionlets coordinates are computed on the fly based on
  - The normalized coordinates
  - Resolution of the detection window

Figure 4
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
Regionlets on PASCAL

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>Regionlets</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 [12]</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 [25]</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 [3]</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>Regionlets</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</td>
<td>27.1</td>
<td>N/A</td>
<td>2009</td>
</tr>
<tr>
<td>INRIA_2009</td>
<td>28.9</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><strong>41.7</strong></td>
<td><strong>39.7</strong></td>
<td>2013</td>
</tr>
</tbody>
</table>

12/17/2013
Regionlets on PASCAL

- Regionlets with Deep CNN feature (outside data)

Table 3. Performance with Deep CNN feature

<table>
<thead>
<tr>
<th>Deep CNN convolutional layer feature (outside data)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN(ImageNet) + layer5 + SVM$^1$</td>
<td>40.1%</td>
</tr>
<tr>
<td>CNN(ImageNet) + layer5 + Hand-crafted feature + Regionlets</td>
<td>49.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Deep CNN fine-tuned full connected layer feature (outside data)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN(fine-tuned on PASCAL) + FC$_7$ + SVM$^1$</td>
<td>48.0%</td>
</tr>
</tbody>
</table>

Will Regionlets model perform at 49.3% + 7.9% = 57.2% using fine-tuned full connected layer feature?

## Regionlets on 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(1)</td>
<td>20.9% (with DCNN feature)</td>
</tr>
<tr>
<td><strong>Regionlets without deep features</strong></td>
<td><strong>19.6% (no DCNN feature)</strong></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>

(1) It’s a preliminary result, we have a better performance now!
Regionlets on ImageNet

- Performance on the validation dataset

![Graph showing performance on the validation dataset with a mean AP of 0.163. The x-axis represents sorting categories using AP, and the y-axis shows AP value. The graph shows a descending trend from approximately 0.7 to 0.1 as categories are sorted.](image-url)
Regionlets on ImageNet

- Top 3 easiest categories: butterfly
Regionlets on ImageNet

- Top 3 easiest categories: Basketball
Regionlets on ImageNet

- Top 3 easiest categories: Dog
Regionlets on ImageNet

- Top 3 hardest categories: backpack
Regionlets on ImageNet

- Top 3 hardest categories: Spatula
Regionlets on ImageNet

- Top 3 hardest categories: Ladle
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 (.2 seconds per image)