Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories

To appear in CVPR 2006

Svetlana Lazebnik (slazebni@uiuc.edu)
Beckman Institute, University of Illinois at Urbana-Champaign

Cordelia Schmid (cordelia.schmid@inrialpes.fr)
INRIA Rhône-Alpes, France

Jean Ponce (ponce@di.ens.fr)
Ecole Normale Supérieure, France

http://www-cvr.ai.uiuc.edu/ponce_grp
Overview

• A “pre-attentive” approach: recognize the scene as a whole without examining its constituent objects  

• Inspiration: *locally orderless images*  
  Koenderink & Van Doorn (1999)

• Previous work: “subdivide-and-disorder” strategy

Szummer & Picard (1997)  
Gist: Torralba et al. (2003)
Spatial pyramid representation

- Extension of a bag of features
- Locally orderless representation at several levels of resolution
- Based on *pyramid match kernels* Grauman & Darrell (2005)
  - *Grauman & Darrell*: build pyramid in feature space, discard spatial information
  - *Our approach*: build pyramid in image space, quantize feature space
Pyramid matching

Indyk & Thaper (2003), Grauman & Darrell (2005)

Find maximum-weight matching (weight is inversely proportional to distance)

Original images

Feature histograms:
Level 3

Level 2

Level 1

Level 0

Total weight (value of pyramid match kernel): \( I_3 + \frac{1}{2}(I_2 - I_3) + \frac{1}{4}(I_1 - I_2) + \frac{1}{8}(I_0 - I_1) \)
Feature extraction

Weak features

Edge points at 2 scales and 8 orientations (vocabulary size 16)

Strong features

SIFT descriptors of 16x16 patches sampled on a regular grid, quantized to form visual vocabulary (size 200, 400)
Scene category dataset
Fei-Fei & Perona (2005), Oliva & Torralba (2001)
http://www-cvr.ai.uiuc.edu/ponce_grp/data

Multi-class classification results (100 training images per class)

<table>
<thead>
<tr>
<th>Level</th>
<th>Weak features (vocabulary size: 16)</th>
<th>Strong features (vocabulary size: 200)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single-level</td>
<td>Pyramid</td>
</tr>
<tr>
<td>0 (1 × 1)</td>
<td>45.3 ±0.5</td>
<td></td>
</tr>
<tr>
<td>1 (2 × 2)</td>
<td>53.6 ±0.3</td>
<td>56.2 ±0.6</td>
</tr>
<tr>
<td>2 (4 × 4)</td>
<td>61.7 ±0.6</td>
<td>64.7 ±0.7</td>
</tr>
<tr>
<td>3 (8 × 8)</td>
<td>63.3 ±0.8</td>
<td>66.8 ±0.6</td>
</tr>
</tbody>
</table>

Fei-Fei & Perona: 65.2%
Scene category retrieval

Query

Retrieved images

- kitchen
- living room
- living room
- office
- living room
- living room
- living room
- living room
- inside city
- store
- mountain
- forest
- tall bldg
- inside city
- inside city
- inside city
- inside city
- tall bldg
- inside city
- mountain
- mountain
- mountain
Scene category confusions

Difficult indoor images

- Kitchen
- Living room
- Bedroom
### Caltech101 dataset

Fei-Fei et al. (2004)


---

#### Multi-class classification results (30 training images per class)

<table>
<thead>
<tr>
<th>Level</th>
<th>Weak features (16)</th>
<th>Strong features (200)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Single-level</td>
<td>Pyramid</td>
</tr>
<tr>
<td>0</td>
<td>15.5 ±0.9</td>
<td>41.2 ±1.2</td>
</tr>
<tr>
<td>1</td>
<td>31.4 ±1.2</td>
<td>55.9 ±0.9</td>
</tr>
<tr>
<td>2</td>
<td>47.2 ±1.1</td>
<td>63.6 ±0.9</td>
</tr>
<tr>
<td>3</td>
<td>52.2 ±0.8</td>
<td>60.3 ±0.9</td>
</tr>
</tbody>
</table>

---
Caltech101 comparison

Zhang, Berg, Maire & Malik, 2006
Caltech101 challenges

Top five confusions

<table>
<thead>
<tr>
<th>class 1 / class 2</th>
<th>class 1 mis-classified as class 2</th>
<th>class 2 mis-classified as class 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ketch / schooner</td>
<td>21.6</td>
<td>14.8</td>
</tr>
<tr>
<td>lotus / water lily</td>
<td>15.3</td>
<td>20.0</td>
</tr>
<tr>
<td>crocodile / crocodile head</td>
<td>10.5</td>
<td>10.0</td>
</tr>
<tr>
<td>crayfish / lobster</td>
<td>11.3</td>
<td>9.1</td>
</tr>
<tr>
<td>flamingo / ibis</td>
<td>9.5</td>
<td>10.4</td>
</tr>
</tbody>
</table>

Easiest and hardest classes

- **minaret (97.6%)**
- **windsor chair (94.6%)**
- **joshua tree (87.9%)**
- **okapi (87.8%)**
- **cougar body (27.6%)**
- **beaver (27.5%)**
- **crocodile (25.0%)**
- **ant (25.0%)**

**Sources of difficulty:** lack of texture, camouflage, “thin” objects, highly deformable shape
Graz dataset

Opelt et al. (2004)

http://www.emt.tugraz.at/~pinz/data/

Detection results (100 pos./100 neg. training images)

<table>
<thead>
<tr>
<th>Class</th>
<th>Spatial pyramid ($M = 200$)</th>
<th>Opelt et al. (2004)</th>
<th>Zhang et al. (2005)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$L = 0$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bikes</td>
<td>$82.4 \pm 2.0$</td>
<td>86.5</td>
<td>92.0</td>
</tr>
<tr>
<td>People</td>
<td>$79.5 \pm 2.3$</td>
<td>80.8</td>
<td>88.0</td>
</tr>
<tr>
<td></td>
<td>$L = 2$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Global spatial regularities (natural scene statistics) help even in databases with high geometric variability!