Analysis:

Objects in Context
[Rabinovich, Vedaldi, Galleguillos, Wiewiora, Belongie]

&

Object Categorization using Co-Occurrence, Location and Appearance
[Galleguillos, Rabinovich, Belongie]

Utsav Prabhu
03/18/2009
Relations between objects
(Biederman et al., 1982)

1. Interposition
   Objects interrupt their background – fire hydrant in front of a building

2. Support
   Objects tend to rest on surfaces – car on a road

3. Probability
   Objects tend to be found in some scenes but not in others – cars with buildings, trees with grass,…

4. Position
   Given an object is probable in a scene, it often is found in some positions and not others – sky towards the top, grass towards the bottom

5. Familiar size
   Objects have a limited set of size relations with other objects – person larger than dog
Approaches using “probability” for object recognition

• Use low-level features across the image
  – Multiscale Conditional Random Fields for Image Labeling

• Use global scene features, such as gist
  – Using the Forest to See the Trees

• Focus of attention
  – Contextual Priming for Object Detection

• Generate a context feature for *each pixel*
  – A Critical View of Context
Semantic Context
Flowchart of approach used
Step 1: Segmentation

• Roadblocks:
  – Number of segments
  – Cues used to segment (pixel locations, color, texture,...)
  – Combination of the above cues

• Solution: Stability based segmentation
Stability based clustering
Take 1

1. Split the dataset into 2 disjoint subsets $A$ & $B$
2. Cluster $A$ into $k$ groups
3. Train a classifier $\varphi$ using the labels from the clustering algorithm
4. Cluster $B$ into $k$ groups
5. Also classify data in $B$ using the classifier $\varphi$
6. Compare the 2 results and determine a stability score
7. Repeat for a range of $k$
Stability based clustering
Take 2

1. Cluster the entire data into $k$ clusters
2. Perturb the data
   – Add noise
   – Perturb the positions of each data point
3. Cluster the data again using same $k$
4. Repeat steps 1-3 many times
5. Permute all the labelings except one (anchor)
6. Calculate a signature based on:
   \[ S(k) = \frac{1}{n - \frac{n}{k}} \left( \sum_{i=1}^{n} s_i - \frac{n}{k} \right) \]
   - Normalization coefficient
   - Indicates label agreement over all perturbations
   - Prevents bias for different values of $k$
7. Try all possible anchors & choose the one with highest stability

No classifier!
Reduced complexity!
Stability based segmentation

• Cues used: Color, Texture
• 9 different cue weightings used
• Noise added 20 times
• Segmentations for k=2 through k=9
Standard N-cut segmentation
Stability based segmentation

Image from MSRC database
Stability based segmentation - results

Images from MSRC database
Flowchart

Original Image → Pre-processing (segmentation) → Categorization → Post-processing (context) → O₁, ..., Oₘ

S₁, S₂, S₃, ..., Sₖ → {L₁,..,Lₙ}₁, {L₁,..,Lₙ}₂, {L₁,..,Lₙ}₃, ..., {L₁,..,Lₙ}ₖ
Bag of Features

1. Decompose the image into a collection of features
2. Map the features to a finite vocabulary of visual words
3. Compute a signature of these visual words
4. Feed the signatures into a classifier for labeling

features = SIFT, visual words = k-means, signature = histogram
Integrating BoF & stable segmentation

- Each segment (of the 54) is masked & zero-padded
- Compute the signature of each segment
  - Discard features which fall outside segment boundary
- Represent the image by ensemble of segment signatures

Reasons for doing this:
- Clustering features in segments incorporates coarse spatial information
- Masking makes features more shape-informative
- Improves SNR
Labeling segments

- Calculate signatures of ground truth segments of training images – $\Phi(I)$
- Calculate signatures of stable segments of test images - $\Phi(S)$
- Calculate L1 distance measure to each category:

$$d(S_q, c) = \min_i d(S_q, I_{ic}) = \min_i \| \phi(S_q) - \phi(I_{ic}) \|$$

- Construct a probability distribution over categories

$$p(c_i|S_q) = \left[1 - \frac{d(S_q, c_i)}{\sum_{j=1}^{n} d(S_q, c_j)}\right]$$
Categorization – Results

Images from MSRC database
Incorporating semantic context

• What we have:
  – Image I with segments \( \{S_1, S_2, \ldots S_k\} \)
  – Marginal probabilities \( p(c_i | S_j) \)

• What we want:
  – Segment labels \( \{c_1, c_2, \ldots c_k\} \) for segments \( \{S_1, S_2, \ldots S_k\} \) which are *in semantic contextual agreement* with each other
CRF framework

\[ p(c_1 \ldots c_k | S_1 \ldots S_k) = \frac{B(c_1 \ldots c_k) \prod_{i=1}^{k} A(i)}{Z(\phi, S_1 \ldots S_k)}, \text{ with} \]

\[ A(i) = p(c_i | S_i) \text{ and } B(c_1 \ldots c_k) = \exp \left( \sum_{i,j=1}^{k} \phi(c_i, c_j) \right) \]

• Separate marginal terms from pair-wise interaction potentials \( \Phi(c_i, c_j) \)

• Where do we get \( \Phi(c_i, c_j) \) from?
  – Co-occurrence matrix from training dataset
  – Google Sets
Co-occurrence matrix

Diagonal entries = frequency of object in training set

Off-diagonal entries = label co-occurrence counts

\( \Phi(c_i, c_j) \) is learned from this data using MLE, gradient descent, importance sampling, monte carlo integration, ...

Can we use values from the co-occurrence matrix directly?
Images from MSRC database
Figure 6. Examples of MSRC (first 3) and PASCAL (last 3) test images, where contextual constraints have improved the categorization accuracy. Results are shown in two different ways, one for each dataset. In MSRC, full segmentations of highest average categorization accuracy are shown; in PASCAL individual segments of highest categorization accuracy are shown. (a) Original Segmented Image. (b) Categorization without contextual constraints. (c) Categorization with co-occurrence contextual constraints derived from the training data. (d) Ground Truth.
Google sets

- Automatically create sets of *possibly related* items from a few examples
- Based on search statistics, trends, web page content, dictionary / thesaurus, wikipedia, ...

Can Google Sets provide a true semantic context based grouping criterion?
Google sets – sanity check

- Query #1: “dog”
  - Results: “dog” “cat” “trackbacks 0” “ティムティム” “canine” “canid” “bird” “pets” “dogs” “horse” “edit” “comments 0” “puppy”
  - Categories found in the results: “cat”, “bird”

Q: How often do dogs and {cats, birds} appear in the same image?
Lets look at the largest annotated database we have: LabelMe.

- Number of images containing dogs = 223
- Number of images containing dogs and cats = 0
- Number of images containing dogs and birds = 0
Google sets – sanity check

• Query #2: “cow”
  – Results: “cow” “pig” “horse” “dog” “cat” “bear” “sheep” “duck” “rabbit” “chicken” “goat” “cash” “animal” “calf”
  – Categories found in the results: “dog” “cat” “sheep” “bird”

  – Number of images containing cows = 33
  – Number of images containing cows and dogs = 0
  – Number of images containing cows and cats = 0
  – Number of images containing cows and sheep = 0
  – Number of images containing cows and birds = 0
Google sets – sanity check

• Query #3: “car”
  – Results: “car” “truck” “auto” “train” “parking” “cars” “boat” “suv” “bus” “motorcycle” “hotel”
  – Categories found in the results: “boat” “bike”
  – Number of images containing cars = 6600
  – Number of images containing cars and boats = 0
  – Number of images containing cars and bikes = 1
Conclusion: Google sets is not really a good source for a semantic context based grouping criterion
Experimental Results

- MSRC & PASCAL datasets

<table>
<thead>
<tr>
<th></th>
<th>No Context</th>
<th>Google Sets</th>
<th>Using Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSRC</td>
<td>45.0%</td>
<td>58.1%</td>
<td>68.4%</td>
</tr>
<tr>
<td>PASCAL</td>
<td>61.8%</td>
<td>63.4%</td>
<td>74.2%</td>
</tr>
</tbody>
</table>

Table 1. Average Categorization Accuracy.
Figure 5. Confusion matrices of average categorization accuracy for MSRC and PASCAL datasets. First row: MSRC dataset; second row: PASCAL dataset. (a) Categorization with no contextual constraints. (b) Categorization with Google Sets context constraints. (c) Categorization with Ground Truth context constraints.
Discussion

• Does co-occurrence truly represent the semantic context of an object?
• Does masking and zero-padding each segment incorporate any kind of shape-information about the segment?
• Should context have the last say in a feed-forward model?
Incorporating spatial context
Spatial context descriptor

• Pair-wise feature
• 3-dimensional descriptor:

\[ F_{ij} = (\mu_{ij}, O_{ij}, O_{ji})^T \quad \forall i, j \in \Omega, i \neq j \]

\[ o_{ij} = \frac{\beta_i \cap \beta_j}{\beta_i} \]
\[ \mu_{ij} = \mu_{yi} - \mu_{yj} \]

– \( \mu_{ij} \) is the difference in y components of centroids of the 2 objects
– \( \beta_i \) is the bounding box / pixel mask of object i
Spatial context feature - example
Spatial context feature

- Vector quantize this descriptor into four groups: *above, below, inside, around*
Locations and co-occurrences
Updated CRF model

\[
p(c_1 \ldots c_k | S_1 \ldots S_k) = \frac{B(c_1 \ldots c_k) \prod_{i=1}^{k} p(c_i | S_i)}{Z(\phi_0, \ldots \phi_r, S_1 \ldots S_k)}
\]

with \( B(c_1 \ldots c_k) = \exp \left( \sum_{i,j=1}^{k} \sum_{r=0}^{q} \alpha_r \phi_r (c_i, c_j) \right) \)
Experimental Results

<table>
<thead>
<tr>
<th>Categories</th>
<th>Semantic Context [18]</th>
<th>CoLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>building</td>
<td>0.85</td>
<td><strong>0.91</strong></td>
</tr>
<tr>
<td>grass</td>
<td>0.94</td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td>tree</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>cow</td>
<td>0.36</td>
<td>0.41</td>
</tr>
<tr>
<td>sheep</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>sky</td>
<td>0.89</td>
<td><strong>0.97</strong></td>
</tr>
<tr>
<td>aeroplane</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>water</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>face</td>
<td>0.80</td>
<td><strong>0.81</strong></td>
</tr>
<tr>
<td>car</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td>bike</td>
<td>0.59</td>
<td><strong>0.60</strong></td>
</tr>
<tr>
<td>flower</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>sign</td>
<td>0.54</td>
<td>0.54</td>
</tr>
<tr>
<td>bird</td>
<td>0.54</td>
<td>0.52</td>
</tr>
<tr>
<td>book</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>chair</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>road</td>
<td>0.94</td>
<td><strong>0.96</strong></td>
</tr>
<tr>
<td>cat</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td>dog</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td>body</td>
<td>0.75</td>
<td><strong>0.77</strong></td>
</tr>
<tr>
<td>boat</td>
<td>0.76</td>
<td><strong>0.81</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Categories</th>
<th>Semantic Context [18]</th>
<th>CoLA</th>
</tr>
</thead>
<tbody>
<tr>
<td>aeroplane</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>bicycle</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>bird</td>
<td>0.18</td>
<td><strong>0.14</strong></td>
</tr>
<tr>
<td>boat</td>
<td>0.28</td>
<td><strong>0.42</strong></td>
</tr>
<tr>
<td>bottle</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>bus</td>
<td>0.46</td>
<td><strong>0.50</strong></td>
</tr>
<tr>
<td>car</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td>cat</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>chair</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>cow</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>dining table</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>dog</td>
<td>0.32</td>
<td><strong>0.29</strong></td>
</tr>
<tr>
<td>horse</td>
<td>0.12</td>
<td><strong>0.15</strong></td>
</tr>
<tr>
<td>motorbike</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>person</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>potted plant</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>sheep</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>sofa</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>train</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>tv monitor</td>
<td>0.62</td>
<td>0.62</td>
</tr>
</tbody>
</table>

Table 1. Comparison of recognition accuracy between the models for MSRC and PASCAL categories. Results in bold indicate an increase in performance by our model. A decrease in performance is shown in italics.
Thank You