Part-based models and local feature layout for category recognition

Thursday, Nov 5
Kristen Grauman
UT-Austin

Announcements

- Pset 4 is out, due 11/24 (~2.5 weeks)
Coarse genres of approaches

• Alignment: hypothesize and test
  – Pose clustering with object instances
  – Indexing invariant features + verification

• Local features: as parts or words
  – Part-based models
  – Bags of words models

• Global appearance: “texture templates”
  – With or without a sliding window

Last time

• Alignment-based recognition: looking for object+pose that *fits* well with image.
  – Use good correspondences to designate hypotheses.
  – Can limit number of verifications performed by voting for most likely model parameters.
Alignment: fitting

- 3+ matches needed to solve for the parameters
- Use local invariant features for reliable matches:
  ✓ interest points relatively sparse
  ✓ very distinctive descriptors

Alignment: backprojection

- 3+ matches needed to solve for the parameters
- Once we have the model parameters, can “backproject”, meaning compute the hypothesized location of any other model points.
Alignment: verification

- 3+ matches needed to solve for the parameters
- Once we have the model parameters, can “backproject”, meaning compute the hypothesized location of any other model points.
- Verification: check for total agreement (e.g., do the image edges coincide with predicted model edges?)

A Real Example from SDSS

A single quad which we happened to try.

http://astrometry.net rowels@cs.toronto.edu
A Real Example from SDSS

The query image scaled, translated & rotated as specified by the quad.

http://astrometry.net roweis@cs.toronto.edu

A Real Example from SDSS

The proposed match, on which we run verification.

http://astrometry.net roweis@cs.toronto.edu
A Real Example from SDSS

http://astrometry.net  roweis@cs.toronto.edu

The verified answer, overlaid on the original catalogue.  The proposed match, on which we run verification.

Today

• Modeling categories with local features and spatial information.

• Transition from instances to categories.

Instance recognition  Category recognition
Key questions

- How to represent the local parts of categories?
- How to encode their spatial relationships?
Issue 1: Vocabularies and parts

- If we treat a visual word as a part, then each bin in descriptor space delineates variation allowed.
- Choice of quantization affects what we consider to be similar.

Pyramid match: main idea

Feature space partitions serve to “match” the local descriptors within successively wider regions.
Pyramid match: main idea

Pyramid match measures difficulty of a match at level \( i \), defined as the number of newly matched pairs at level \( i \).

For similarity, weights inversely proportional to bin size (or may be learned).

Normalize these kernel values to avoid favoring large sets.

\[
K_{\Delta}(X, Y) = \sum_{i=0}^{L} 2^{-i} I \left( H_X^{(i)}, H_Y^{(i)} \right) - I \left( H_X^{(i-1)}, H_Y^{(i-1)} \right)
\]

\( I(H_X, H_Y) = \sum_j \min(H_X(j), H_Y(j)) \)

\( = 3 \)

Histogram intersection counts number of possible matches at a given partitioning.

[Grauman & Darrell, ICCV 2005]
Issue 2: No spatial layout preserved

Slide by Bill Freeman, MIT
Spatial pyramid match

- Make a pyramid of bag-of-words histograms.
- Provides some loose (global) spatial layout information

Spatial pyramid match

- Can capture **scene** categories well---texture-like patterns but with some variability in the positions of all the local pieces.
Spatial pyramid match

• Can capture scene categories well—texture-like patterns but with some variability in the positions of all the local pieces.

Confusion table

<table>
<thead>
<tr>
<th></th>
<th>Office</th>
<th>Kitchen</th>
<th>Living Room</th>
<th>Bedroom</th>
<th>Store</th>
<th>Industrial</th>
<th>Tall Building</th>
<th>Inside City</th>
<th>森</th>
<th>Highway</th>
<th>Coast</th>
<th>Open Country</th>
<th>Mountain</th>
<th>Town</th>
<th>Suburb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Office</td>
<td>1</td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Kitchen</td>
<td>0.6</td>
<td>1</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Living Room</td>
<td>0.7</td>
<td>0.7</td>
<td>1</td>
<td>0.8</td>
<td>0.9</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Bedroom</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Store</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>1</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.9</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Industrial</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>1</td>
<td>0.8</td>
<td>0.9</td>
<td>0.7</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Tall Building</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>0.6</td>
<td>1</td>
<td>0.9</td>
<td>0.6</td>
<td>0.6</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Inside City</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.9</td>
<td>1</td>
<td>0.5</td>
<td>0.5</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>森</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>1</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Highway</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.6</td>
<td>0.5</td>
<td>0.7</td>
<td>1</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Coast</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.8</td>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Open Country</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
<td>1</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Mountain</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>0.7</td>
<td>0.6</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>1</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Town</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>1</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>Suburb</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.8</td>
<td>1</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Spatial pyramid match

• Can capture scene categories well—texture-like patterns but with some variability in the positions of all the local pieces.

<table>
<thead>
<tr>
<th>Level (vocabulary size: 200)</th>
<th>Single-level</th>
<th>Pyramid</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 (1 × 1)</td>
<td>72.2 ±0.6</td>
<td>79.0 ±0.5</td>
</tr>
<tr>
<td>1 (2 × 2)</td>
<td>77.9 ±0.6</td>
<td>81.1 ±0.3</td>
</tr>
<tr>
<td>2 (4 × 4)</td>
<td>79.4 ±0.3</td>
<td>81.1 ±0.3</td>
</tr>
<tr>
<td>3 (8 × 8)</td>
<td>77.2 ±0.4</td>
<td>80.7 ±0.3</td>
</tr>
</tbody>
</table>
• What will a grid binning of features over the whole image be sensitive to?

Semi-local configurations of words

• To avoid this translation sensitivity, can make the spatial configurations feature-centric.
• Continuing the analogy: common visual “phrases”

[Quack, Ferrari, Leibe, van Gool ICCV 2007]  [Sivic et al. ICCV 2005]
Parts+structure ("part-based") models

- Previous methods add some spatial layout to the bag-of-words model
- **Part-based** models represent a category by common parts (appearance) and their layout (shape/structure).
  - We’ll next look at two ways to encode the shape:
    - implicit and explicit

---

Shape representation in part-based models

"Star" shape model

- e.g. implicit shape model
- Parts mutually independent

N image features, P parts in the model

Slide credit: Rob Fergus
Implicit shape models

- Visual vocabulary is used to index votes for object position [a visual word = “part”]

B. Leibe, A. Leonardis, and B. Schiele, *Combined Object Categorization and Segmentation with an Implicit Shape Model*, ECCV Workshop on Statistical Learning in Computer Vision 2004
Implicit shape models: Training

1. Build vocabulary of patches around extracted interest points using clustering

How will this differ from a vocabulary built from, e.g., all frames in a movie?

Implicit shape models: Training

1. Build vocabulary of patches around extracted interest points using clustering
2. Map the patch around each interest point to closest word
Implicit shape models: Training

1. Build vocabulary of patches around extracted interest points using clustering
2. Map the patch around each interest point to closest word
3. For each word, store all positions it was found, relative to object center

Implicit shape models: Testing

1. Given new test image, extract patches, match to vocabulary words
2. Cast votes for possible positions of object center
3. Search for maxima in voting space
4. (Extract weighted segmentation mask based on stored masks for the codebook occurrences)

*What is the dimension of the Hough space?*
Implicit shape models: Testing

Example: Results on Cows

Original image

K. Grauman, B. Leibe
Example: Results on Cows

Interest points

K. Grauman, B. Leibe

Example: Results on Cows

Matched patches

K. Grauman, B. Leibe
Example: Results on Cows

Votes
K. Grauman, B. Leibe

Example: Results on Cows

1st hypothesis
K. Grauman, B. Leibe
Example: Results on Cows

2nd hypothesis

K. Grauman, B. Leibe

Example: Results on Cows

3rd hypothesis

K. Grauman, B. Leibe
Detection Results

• Qualitative Performance
  - Recognizes different kinds of objects
  - Robust to clutter, occlusion, noise, low contrast

Scoring an object detector

If prediction and ground truth are \textit{bounding boxes}, when do we have a correct detection?
Scoring an object detector

\[ a_o = \frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})} \]

We’ll say the detection is correct (a “true positive”) if the intersection of the bounding boxes, divided by their union, is > 50%.

Scoring an object detector

If the detector can produce a confidence score on the detections, then we can plot the rate of true vs. false positives as a threshold on the confidence is varied.

\[ TPR = \text{fraction of positive examples that are correctly labeled.} \]

\[ FPR = \text{fraction of negative examples that are misclassified as positive.} \]
Shape representation in part-based models

“Star” shape model

Fully connected constellation model

- e.g. implicit shape model
- Parts mutually independent
- e.g. Constellation Model
- Parts fully connected

N image features, P parts in the model

Part-based models: constellation of fully connected parts

Slide credit: Rob Fergus

Slide by Bill Freeman, MIT
Probabilistic constellation model

\[ P(\text{image} \mid \text{object}) = P(\text{appearance, shape} \mid \text{object}) \]

Part descriptors

Part locations

Candidate parts

Source: Lana Lazebnik

Probabilistic constellation model

\[ P(\text{image} \mid \text{object}) = P(\text{appearance, shape} \mid \text{object}) \]

Part 1

Part 3

Part 2

Source: Lana Lazebnik
Probabilistic constellation model

\[ P(\text{image} \mid \text{object}) = P(\text{appearance}, \text{shape} \mid \text{object}) \]
\[ = \max_h P(\text{appearance} \mid h, \text{object}) p(\text{shape} \mid h, \text{object}) p(h \mid \text{object}) \]

h: assignment of features to parts

Source: Lana Lazebnik
Probabilistic constellation model

\[ P(\text{image} \mid \text{object}) = P(\text{appearance}, \text{shape} \mid \text{object}) = \max_h P(\text{appearance} \mid h, \text{object}) p(\text{shape} \mid h, \text{object}) p(h \mid \text{object}) \]

Distribution over joint part positions

2D image space

Example results from constellation model:

data from four categories

Slide from Li Fei-Fei http://www.vision.caltech.edu/feifeili/Resume.htm
Face model Appearance: 10 patches closest to mean for each part

Recognition results

Test images: size of circles indicates score of hypothesis

Fergus et al. CVPR 2003
Appearance: 10 patches closest to mean for each part

Motorbike model

Recognition results

Spotted cat model

Recognition results

Fergus et al. CVPR 2003
Comparison

<table>
<thead>
<tr>
<th>class</th>
<th>bag of features</th>
<th>bag of features</th>
<th>Part-based model</th>
</tr>
</thead>
<tbody>
<tr>
<td>airplanes</td>
<td>98.8</td>
<td>97.1</td>
<td>90.2</td>
</tr>
<tr>
<td>cars (rear)</td>
<td>98.3</td>
<td>98.6</td>
<td>90.3</td>
</tr>
<tr>
<td>cars (side)</td>
<td>95.0</td>
<td>87.3</td>
<td>88.5</td>
</tr>
<tr>
<td>faces</td>
<td>100</td>
<td>99.3</td>
<td>96.4</td>
</tr>
<tr>
<td>motorbikes</td>
<td>98.5</td>
<td>98.0</td>
<td>92.5</td>
</tr>
<tr>
<td>spotted cats</td>
<td>97.0</td>
<td>—</td>
<td>90.0</td>
</tr>
</tbody>
</table>

Source: Lana Lazebnik

Shape representation in part-based models

"Star" shape model

- e.g. implicit shape model
- Parts mutually independent
- Recognition complexity: $O(NP)$
- Method: Gen. Hough Transform

Fully connected constellation model

- e.g. Constellation Model
- Parts fully connected
- Recognition complexity: $O(N^P)$
- Method: Exhaustive search

N image features, P parts in the model

Slide credit: Rob Fergus
Weaker supervision to learn object parts

Summary

- Modeling categories with local features and spatial information:
  - Histograms, configurations of visual words to capture global or local layout in the bag-of-words framework
    - Pyramid match, semi-local features
  - Part-based models to encode category’s part appearance together with 2d layout,
  - Allow detection within cluttered image
    - “implicit shape model”: shape based on layout of all parts relative to a reference part; Generalized Hough for detection
    - “constellation model”: explicitly model mutual spatial layout between all pairs of parts; exhaustive search for best fit of features to parts