Bag of Visual Words for Image Representation

Jianping Fan
Department of Computer Science
UNC-Charlotte

Course Website:
http://webpages.uncc.edu/jfan/itcs5152.html
Variability: Camera position
Illumination
Shape parameters
Within-class variations?

Recognition is all about modeling variability

Svetlana Lazebnik
Within-class variations
History of ideas in recognition

• 1960s – early 1990s: the geometric era
Variability: Θ Camera position Illumination Alignment

Shape: assumed known

Roberts (1965); Lowe (1987); Faugeras & Hebert (1986); Grimson & Lozano-Perez (1986); Huttenlocher & Ullman (1987)
Recall: Alignment

• Alignment: fitting a model to a transformation between pairs of features (*matches*) in two images

\[
\sum_i (T(x_i), x_i')
\]

Find transformation \( T \) that minimizes the residual between corresponding features in the two images.
Recognition as an alignment problem: Block world

Fig. 1. A system for recognizing 3-d polyhedral scenes. a) L.G. Roberts. b) A blocks world scene. c) Detected edges using a 2x2 gradient operator. d) A 3-d polyhedral description of the scene, formed automatically from the single image. e) The 3-d scene displayed with a viewpoint different from the original image to demonstrate its accuracy and completeness. (b) - (e) are taken from [64] with permission MIT Press.

J. Mundy, Object Recognition in the Geometric Era: a Retrospective, 2006

Representing and recognizing object categories is harder...

ACRONYM (Brooks and Binford, 1981)

Binford (1971), Nevatia & Binford (1972), Marr & Nishihara (1978)
Recognition by components

Biederman (1987)

Primitives (geons)

Objects


Svetlana Lazebnik
Generalized cylinders
Ponce et al. (1989)

Zisserman et al. (1995)

Forsyth (2000)

General shape primitives?

Svetlana Lazebnik
History of ideas in recognition

• 1960s – early 1990s: the geometric era
• 1990s: appearance-based models
Empirical models of image variability

**Appearance-based techniques**

Turk & Pentland (1991); Murase & Nayar (1995); etc.
Eigenfaces (Turk & Pentland, 1991)

<table>
<thead>
<tr>
<th>Experimental Condition</th>
<th>Correct/Unknown Recognition Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forced classification</td>
<td>96/0</td>
</tr>
<tr>
<td>Forced 100% accuracy</td>
<td>100/19</td>
</tr>
<tr>
<td>Forced 20% unknown rate</td>
<td>100/20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lighting</th>
<th>Orientation</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>85/0</td>
<td>100/39</td>
<td>100/60</td>
</tr>
<tr>
<td>94/20</td>
<td></td>
<td>74/20</td>
</tr>
</tbody>
</table>
Color Histograms

History of ideas in recognition

• 1960s – early 1990s: the geometric era
• 1990s: appearance-based models
• 1990s – present: sliding window approaches
Sliding window approaches
Sliding window approaches

- Turk and Pentland, 1991
- Belhumeur, Hespanha, & Kriegman, 1997
- Schneiderman & Kanade, 2004
- Viola and Jones, 2000
- Schneiderman & Kanade, 2004
- Argawal and Roth, 2002
- Poggio et al. 1993
History of ideas in recognition

• 1960s – early 1990s: the geometric era
• 1990s: appearance-based models
• Mid-1990s: sliding window approaches
• Late 1990s: local features
Local features for object instance recognition

Large-scale image search
Combining local features, indexing, and spatial constraints

Image credit: K. Grauman and B. Leibe
Large-scale image search
Combining local features, indexing, and spatial constraints

Philbin et al. ‘07
Large-scale image search
Combining local features, indexing, and spatial constraints

Google Goggles in Action
Click the icons below to see the different ways Google Goggles can be used.

<table>
<thead>
<tr>
<th>Landmark</th>
<th>Book</th>
<th>Contact Info</th>
<th>Artwork</th>
<th>Places</th>
<th>Wine</th>
<th>Logo</th>
</tr>
</thead>
</table>

Available on phones that run Android 1.5+ (i.e. Donut or Eclair)
History of ideas in recognition

• 1960s – early 1990s: the geometric era
• 1990s: appearance-based models
• Mid-1990s: sliding window approaches
• Late 1990s: local features
• Early 2000s: parts-and-shape models
Parts-and-shape models

- Model:
  - Object as a set of parts
  - Relative locations between parts
  - Appearance of part

Figure from [Fischler & Elschlager 73]
Constellation models

Pictorial structure model

Fischler and Elschlager(73), Felzenszwalb and Huttenlocher(00)

Pr(P_\text{tor}, P_\text{arm}, \ldots | \text{Im}) \propto \prod_{i,j} \Pr(P_i | P_j) \prod_i \Pr(\text{Im}(P_i))

generation of part geometry

part appearance
Discriminatively trained part-based models

History of ideas in recognition

• 1960s – early 1990s: the geometric era
• 1990s: appearance-based models
• Mid-1990s: sliding window approaches
• Late 1990s: local features
• Early 2000s: parts-and-shape models
• Mid-2000s: bags of features
Bag-of-features models
Bag-of-features models

Object ➔ Bag of ‘words’
Objects as texture

• All of these are treated as being the same

• No distinction between foreground and background: scene recognition?
Origin 1: Texture recognition

- Texture is characterized by the repetition of basic elements or *textons*

- For stochastic textures, it is the identity of the textons, not their spatial arrangement, that matters

Origin 1: Texture recognition

- Texture recognition
- Universal texton dictionary
- Histogram

Origin 2: Bag-of-words models

Origin 2: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary  
  
  Salton & McGill (1983)
Origin 2: Bag-of-words models

- Orderless document representation: frequencies of words from a dictionary
  Salton & McGill (1983)
Origin 2: Bag-of-words models

Bag-of-features steps

1. Extract features
2. Learn “visual vocabulary”
3. Quantize features using visual vocabulary
4. Represent images by frequencies of “visual words”
1. Feature extraction

- Regular grid or interest regions
1. Feature extraction

- Detect patches
- Normalize patch
- Compute descriptor

Slide credit: Josef Sivic
1. Feature extraction

Slide credit: Josef Sivic
2. Learning the visual vocabulary

Slide credit: Josef Sivic
2. Learning the visual vocabulary

[Diagram showing clustering process]

Slide credit: Josef Sivic
2. Learning the visual vocabulary

Visual vocabulary

Clustering

Slide credit: Josef Sivic
Clustering and vector quantization

- Clustering is a common method for learning a visual vocabulary or codebook
  - Unsupervised learning process
  - Each cluster center produced by k-means becomes a codevector
  - Codebook can be learned on separate training set
  - Provided the training set is sufficiently representative, the codebook will be “universal”

- The codebook is used for quantizing features
  - A *vector quantizer* takes a feature vector and maps it to the index of the nearest codevector in a codebook
  - Codebook = visual vocabulary
  - Codevector = visual word
Example codebook

Source: B. Leibe
Visual vocabularies: Issues

• How to choose vocabulary size?
  • Too small: visual words not representative of all patches
  • Too large: quantization artifacts, overfitting

• Computational efficiency
  • Vocabulary trees
    (Nister & Stewenius, 2006)
But what about layout?

All of these images have the same color histogram
Spatial pyramid

Compute histogram in each spatial bin
Spatial pyramid representation

Extension of a bag of features

Locally orderless representation at several levels of resolution

Lazebnik, Schmid & Ponce (CVPR 2006)
Spatial pyramid representation

Extension of a bag of features
Locally orderless representation at several levels of resolution

Lazebnik, Schmid & Ponce (CVPR 2006)
Spatial pyramid representation

Extension of a bag of features

Locally orderless representation at several levels of resolution

Lazebnik, Schmid & Ponce (CVPR 2006)
Scene category dataset

Multi-class classification results

<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>
### Caltech101 dataset


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>32.8 ±1.3</td>
</tr>
<tr>
<td>2</td>
<td>47.2 ±1.1</td>
<td>49.3 ±1.4</td>
</tr>
<tr>
<td>3</td>
<td>52.2 ±0.8</td>
<td>54.0 ±1.1</td>
</tr>
</tbody>
</table>
Bags of features for action recognition

Space-time interest points

History of ideas in recognition

• 1960s – early 1990s: the geometric era
• 1990s: appearance-based models
• Mid-1990s: sliding window approaches
• Late 1990s: local features
• Early 2000s: parts-and-shape models
• Mid-2000s: bags of features
• Present trends: combination of local and global methods, context, deep learning