Biological Perspective on Pattern Recognition

Presenter: Jianchao Yang
Course Instructor: Prof. Derek Hoiem
Computer vision follows human vision?

- We want computer to ‘see’ like human, how?

- Do we want the computer to ‘see’ more like a computer or more like a human?
Computer vs. Human

- Modern computers can perform many complex tasks much faster, more efficiently and precisely than human.
- In pattern recognition, a three-year-old can outperform the most sophisticated algorithms available today.
Topics today

- Object recognition with features inspired by visual cortex

- Face recognition by human: 20 results all computer vision researchers should know about.
Object Recognition with Features
Inspired by Visual Cortex

Thomas Serre et.al
Feature as a trade-off

- Template-based approach
  - Selective, but too rigid to capture the variations of object appearance

- Histogram-based approach
  - Robust to image transforms, but not selective enough

- “Learning the Discriminative Power-Invariance Trade-off”
Feature guided by biological systems

- Computer vision techniques inspired by human vision have never pasted V1

- Follows the standard model in primary visual cortex (V1)
Standard Model of Visual Cortex (V1)

- “Hierarchical models of object recognition in cortex”

- Simple version: four layers of alternating computation units of S units and C units
  - *Simple* S units: Gaussian-like tuning to increase selectivity.
  - *Complex* C units: Maximum pooling to increase invariance
Fact 1
Visual processing is hierarchical, aiming to build invariance to position and scale first and then to viewpoint and other transformations.

Fact 2
The receptive fields of the neurons as well as the complexity of their optimal stimuli increases.
Standard Model of Visual Cortex (cont.)

- **Fact 3**
  The initial processing of information is feed-forward.

- **Fact 4**
  Plasticity and learning probably occurs at all stages and certainly at the level of IT and PFC, the top layers of the hierarchy.
System overview

**Input Image**
- Gray-value

**S1**
- Apply battery of Gabor filters. Here we see filtration at 8 scales and 4 orientations (color indicates orientation). The full model uses 16 scales.

**C1**
- Local maximum over position and scale.

**S2**
- Filter (L2 RBF) with N previously seen patches \( P_i \) \( i = 1 \ldots N \). These patches are in C1 format. Each orientation in the patch is matched to the corresponding orientation in C1. The result is one image per C1 band per patch.

**C2**
- The C2 values are computed by taking a max over all S2 associated with a given patch. Thus, the C2 response has length N.
**$S_1$ Units**

- Multi-dimensional array of filters, arranged by scales and orientations
- Take the form of Gabor functions

\[
F(x, y) = \exp\left(-\frac{(x_o^2 + \gamma^2 y_o^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda} x_o\right), \quad \text{s.t.}
\]

\[
x_o = x \cos \theta + y \sin \theta \quad \text{and} \quad y_o = -x \sin \theta + y \cos \theta.
\]

- 16 Scales (7-by-7 to 37-by-37 pixels) and four orientations (0, 45, 90 and 135 degrees) used
C₁ Units

- Larger receptive fields (twice as large as S₁)
  - Add more tolerance to shift and size

- C₁ units spatially pool over S₁ inputs by max operation for each band and each orientation.
## $S_1$ and $C_1$ Parameters

<table>
<thead>
<tr>
<th>Scale band $S$</th>
<th>Spatial pooling grid ($N_S \times N_S$)</th>
<th>Overlap $\Delta_S$</th>
<th>$C_1$ layer filter size $s$</th>
<th>$S_1$ layer Gabor $\sigma$</th>
<th>$S_1$ layer Gabor $\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band 1</td>
<td>$8 \times 8$</td>
<td>4</td>
<td>$7 \times 7$</td>
<td>2.8</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$9 \times 9$</td>
<td>3.6</td>
<td>4.6</td>
</tr>
<tr>
<td>Band 2</td>
<td>$10 \times 10$</td>
<td>5</td>
<td>$11 \times 11$</td>
<td>4.5</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$13 \times 13$</td>
<td>5.4</td>
<td>6.8</td>
</tr>
<tr>
<td>Band 3</td>
<td>$12 \times 12$</td>
<td>6</td>
<td>$15 \times 15$</td>
<td>6.3</td>
<td>7.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$17 \times 17$</td>
<td>7.3</td>
<td>9.1</td>
</tr>
<tr>
<td>Band 4</td>
<td>$14 \times 14$</td>
<td>7</td>
<td>$19 \times 19$</td>
<td>8.2</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$21 \times 21$</td>
<td>9.2</td>
<td>11.5</td>
</tr>
<tr>
<td>Band 5</td>
<td>$16 \times 16$</td>
<td>8</td>
<td>$23 \times 23$</td>
<td>10.2</td>
<td>12.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$25 \times 25$</td>
<td>11.3</td>
<td>14.1</td>
</tr>
<tr>
<td>Band 6</td>
<td>$18 \times 18$</td>
<td>9</td>
<td>$27 \times 27$</td>
<td>12.3</td>
<td>15.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$29 \times 29$</td>
<td>13.4</td>
<td>16.8</td>
</tr>
<tr>
<td>Band 7</td>
<td>$20 \times 20$</td>
<td>10</td>
<td>$31 \times 31$</td>
<td>14.6</td>
<td>18.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$33 \times 33$</td>
<td>15.8</td>
<td>19.7</td>
</tr>
<tr>
<td>Band 8</td>
<td>$22 \times 22$</td>
<td>11</td>
<td>$35 \times 35$</td>
<td>17.0</td>
<td>21.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$37 \times 37$</td>
<td>18.2</td>
<td>22.8</td>
</tr>
</tbody>
</table>
$S_2$ Units

- Behaves as radial basis function units
- The response $r$ of the corresponding $S_2$ unit for an input patch $X$ from $C_1$ is given by:

$$r = \exp\left(-\beta \|X - P_i\|^2\right)$$

- $P_i$ is the learned prototype. $S_2$ maps are computed across all positions for each of the eight scale bands.
C₂ Units

- Max pooling over all scales and positions for each S₂ type over the entire S₂ lattice.

- C₂ is totally invariant to shift and scale changes.

- Final feature is the combination of C₁ and C₂ (local invariance and global invariance)
System Revisit

**Input Image**
Gray-value

**S1**
Apply battery of Gabor filters. Here we see filtration at 8 scales and 4 orientations (color indicates orientation). The full model uses 16 scales.

**C1**
Local maximum over position and scale.

**S2**
Filter (L2 RBF) with N previously seen patches \( \{ P_i \}_{i=1}^N \). These patches are in C1 format. Each orientation in the patch is matched to the corresponding orientation in C1. The result is one image per C1 band per patch.

**C2**
The C2 values are computed by taking a max over all S2 associated with a given patch. Thus, the C2 response has length N.
System Revisit: example

Four S1 scale bands are shown here. The levels in these images have been scaled for illustrative purposes. Color indicates orientation, as above.

Only one S2 scale is shown for each patch.
Evaluation

- Detection on five classes of Caltech 101
- Compared with key-point SIFT features
Evaluation (cont.)

- Street scene database
My comments about the paper

- Use neuroscience as a guide to build a complex system for object recognition.

- The feature is somewhat robust to shift and scale changes. C1 is still too rigid on spatial variations.

- $C_2$ is somewhat like the histogram feature.
Questions to ask

- Should we use neuroscience as a guide for computer vision?

- If not, what should be the principle for constructing a complicated system that work for real applications?
Face Recognition by Humans

20 Results all Computer Vision Researcher Should Know About
Result 1

- Humans can recognize faces in extremely low resolution images.
Result 3

- High-frequency information by itself does not lead to good face recognition performance
Eyebrows are among the most important for recognition.
Result 6

- Both internal and external facial cues are important and they exhibit non-linear interactions
Result 7

- The important configural relations appear to be independent across the width and height dimensions
Result 8

- Vertical inversion dramatically reduces recognition performance
Result 12

- Contrast polarity inversion dramatically impairs recognition performance, possibly due to compromised ability to use pigmentation cues
Result 15

- Motion of faces appears to facilitate subsequent recognition
Result 19

- Latency of responses to faces in IT cortex is about 120 ms, suggesting a largely feed-forward computation
Human memory for briefly seen faces is rather poor
Discussion