Natural Scene Categorization: from Humans to Computers

Li Fei-Fei

Beckman Institute, ECE, Psychology

http://visionlab.ece.uiuc.edu
#1: natural scene categorization entails little attention (Rufin VanRullen, Christof Koch, Pietro Perona)

#2: what can we perceive within a glance of a scene – a working definition for ‘gist’ (Asha Iyer, Christof Koch, Pietro Perona)

#3: local patches, and some intermediate level information – a hierarchical Bayesian algorithm for natural scene categorization (Pietro Perona)

An outdoor scene, I think. reminded me of a city... like walking in a park in New York or something. there seemed to be trees and a road and then this large skyscraper in the background.
• #1: natural scene categorization entails little attention

Reference: Li et al. 2002; Fei-Fei et al. 2005
Our question

1. How critical is attention in natural scene recognition?

2. How does this compare to other recognition tasks?
animals
vehicle

synthetic stimuli

Li et al. 2002
Li et al. 2002
Without color…

![Graph showing normalized central perf (%) vs. normalized period perf (%)](image)

Additional images of black and white photos are shown on the right side of the page.
Effect of “meaningful” category

randomly rotated
Target vs. Distractor
( masked by )

fixed rotation
Target vs. Distractor
( masked by )

upright position
Target vs. Distractor
( masked by )
• #2: what can we perceive within a glance of a scene – a working definition for ‘gist’

Reference: Fei-Fei et al. submitted
What is “gist”?  

- sensory data, e.g. “color”, “size”, etc.

Wolfe, 1998; Goffaux et al. 2005
What is “gist”?

- sensory data, e.g. “color”, “size”, etc.
- “inventory of some of the objects (and textures)”

Wolfe, 1998
What is “gist”?  

- sensory data, e.g. “color”, “size”, etc.  
- “inventory of some of the objects”  
- “some relationships between objects”
What is “gist”? 

- sensory data, e.g. “color”, “size”, etc.
- “inventory of some of the objects”
- “some relationships between objects”
- “layout”

Biederman et al. 1987, Wolfe, 1998
What is “gist”? 

- sensory data, e.g. “color”, “size”, etc.
- “inventory of some of the objects”
- “some relationships between objects”
- “layout”
- “stiffness”

Wolfe, 1998
What is “gist”? 

- sensory data, e.g. “color”, “size”, etc.
- “inventory of some of the objects”
- “some relationships between objects”
- “layout”
- “stufness”
- scene category
What do people see in a glance?
Stage I: Collect Image Description

--- Illustration of 1 Trial

Subject types freely what he/she saw in the image

Please type your description here:

An outdoor scene, I think. reminded me a city... like walking in a park in New York or something. there seemed to be trees and a road and then this large skyscraper in the background.

Image onset: t = 0 msec

Mask onset: t = SOA

1 of 7 possible SOA’s (msec):
27, 40, 53, 67, 80, 120, 500

Image onset: t = 0 msec

Mask onset: t = SOA

1 of 7 possible SOA’s (msec):
27, 40, 53, 67, 80, 120, 500
PT = 27ms
Couldn't see much; it was mostly dark w/ some square things, maybe furniture. (Subject: AM)

PT = 40ms
This looked like an indoor shot. Saw what looked like a large framed object (a painting?) on a white background (i.e., the wall). (Subject: RW)

PT = 500ms
This is indoors. It's must be a rich person's house. There are many paintings on the wall. The largest painting might have a fireplace beneath it. I think the largest painting was that of a man standing erect. The room is richly decorated and it looks like one of the rooms in Mr. Darcy's house in the A&E movie Pride and Prejudice. Or maybe it more closely resembles one of the rooms where the one of the rooms in Huntington's house (at the Huntington).

PT = 67ms
I saw the interior of a room in a house. There was a picture to the right, that was black, and possibly a table in the center. It seemed like a formal dining room. (Subject: JB)

Fei-Fei et al. submitted
I could make out some kind of circular shapes near the bottom of the picture. These reminded me of those round life preservers that are on ships. There was also a man standing on top of some wooden structure.
Scene level

Evaluation Score

- outdoor
- indoor
- mm.
- outdoor
- hh.
- rooms
- nat.
- outdoor
- office/class
- dine/rest
- urban/city
- field/park
- store
- sky
- water

PT = 107
PT = 500
Object level

![Graph showing evaluation scores for different categories: animate obj, people, body/figure, ethnicity, appearance, large mammal, animal, mammal, gender/age, bird, cat/dog, hill/mnt., plant, rocks, vehicle, building, desk/table, inanimate obj, obj, structure, furniture, car. The graph compares evaluation scores at PT = 107 and PT = 500, with error bars indicating variability. Significant differences are marked with an asterisk (*)].
(Social) Events

![Bar chart showing evaluation scores for different events.]

- **Social Interactions**
  - PT = 107: 0.5
  - PT = 500: 0.6

- **Sports and Games**
  - PT = 107: 0.5
  - PT = 500: 0.6

- **Stage Performance**
  - PT = 107: 0.5
  - PT = 500: 0.6
• #3: local patches, and some intermediate level information – a hierarchical Bayesian algorithm for natural scene categorization

Reference: Fei-Fei et al. CVPR 2005
• **global cues: colors, textures, etc.**

Jain, Zhang et al. (1998)

Figure 7: 2D Feature space plots showing (a) edge direction coherence vector and (b) color coherence vectors; + represents the landscape patterns and © represents the city patterns; only a subset of 2,710 patterns have been plotted here for clarity of display.

Szummer et al. (1998)
• **global cues: frequency**

- openness
- ruggedness
- roughness

Oliva, Torralba (1999, 2001)
local patch based idea

Concept
Occurrences

<table>
<thead>
<tr>
<th>Object</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>sky</td>
<td>14.0%</td>
</tr>
<tr>
<td>water</td>
<td>32.5%</td>
</tr>
<tr>
<td>grass</td>
<td>0.0%</td>
</tr>
<tr>
<td>trunks</td>
<td>0.0%</td>
</tr>
<tr>
<td>foliage</td>
<td>6.5%</td>
</tr>
<tr>
<td>fields</td>
<td>0.0%</td>
</tr>
<tr>
<td>rocks</td>
<td>31.0%</td>
</tr>
<tr>
<td>flowers</td>
<td>0.0%</td>
</tr>
<tr>
<td>sand</td>
<td>16.0%</td>
</tr>
</tbody>
</table>

Vogel & Schiele (2004)
Our intuitions

• local patch based

• intermediate level themes within scenes

Fei-Fei & Perona (CVPR 2005)
Our intuitions

• local patch based

• intermediate level themes

• weakly supervised
  – no human annotation of local patches and intermediate level themes

Fei-Fei & Perona (CVPR 2005)
Image → Bag of ‘words’
Of all the sensory impressions proceeding to the brain, the visual experiences are the dominant ones. Our perception of the world around us is based essentially on the messages that reach the brain from our eyes. For a long time it was thought that the retinal image was transmitted point by point to visual centers in the brain as a movie screen. Through the discoveries of Hubel and Wiesel we now know that the perception of visual images is a more complex process involving a stepwise analysis in a system of nerve cells stored in columns. In this system each cell has its specific function and is responsible for a specific detail in the pattern of the retinal image.

China is forecasting a trade surplus of $90bn (£51bn) to $100bn this year, a threefold increase on 2004’s $32bn. The Commerce Ministry said the surplus would be created by a predicted 30% increase in exports to $750bn, compared with a rise of 18% in imports to $660bn. This is likely to annoy the US, which has long argued that China’s exports are unfairly helped by a deliberately undervalued yuan. Beijing agrees the surplus is too high, but says the yuan is only one factor. Bank of China governor Zhou Xiaochuan said the country also needs to boost domestic demand so more goods stay within the country. China increased the value of the yuan against the dollar by 2.1% in July and permitted it to trade within a narrow band, but the US wants the yuan to be allowed to rise freely. However, Beijing has made it clear that it will take its time and tread carefully before allowing the yuan to rise further in value.
"beach"
<table>
<thead>
<tr>
<th></th>
<th>highway</th>
<th>insidecity</th>
<th>tallbuildings</th>
<th>street</th>
<th>suburb</th>
<th>forest</th>
<th>coast</th>
<th>mountain</th>
<th>opencountry</th>
<th>bedroom</th>
<th>kitchen</th>
<th>livingroom</th>
<th>office</th>
</tr>
</thead>
<tbody>
<tr>
<td>highway</td>
<td>74</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>insidecity</td>
<td>58</td>
<td>10</td>
<td>6</td>
<td>8</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>tallbuildings</td>
<td>4</td>
<td>76</td>
<td>10</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>street</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>78</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>suburb</td>
<td>94</td>
<td>2</td>
<td>12</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>forest</td>
<td></td>
<td></td>
<td>88</td>
<td>12</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>coast</td>
<td>2</td>
<td>2</td>
<td>78</td>
<td>20</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>mountain</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>opencountry</td>
<td>8</td>
<td>10</td>
<td>16</td>
<td>10</td>
<td>10</td>
<td>48</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>bedroom</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>kitchen</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>livingroom</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>office</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>64</td>
</tr>
</tbody>
</table>
Our model

Traditional texton model
model distance based on theme distribution
segmentation by themes

codeword images

theme images
Summary

• natural scene categorization entails little attention

• ‘gist’ of a scene includes much information on objects, scenes and beyond

• propose a hierarchical Bayesian algorithm for natural scene categorization using local patches
references


