Natural Scene Recognition: From Humans to Computers

Li Fei-Fei

1. Computer Science Department
2. Psychology Department
A picture is worth a thousand words.
--- Confucius
or *Printers’ Ink* Ad (1921)
blue
rugged
white and red
bright
textured structure
green
porous
grey
elongated shapes
monument
buildings
cloudy sky
trees
mountains
people
day time
walking
street
vendors
• To understand human visual intelligence by via psychophysical and physiological experiments

• To build intelligent visual algorithms for machines and robots
Potter, Biederman, etc. 1970s
Delorme, et al. 1998
A feed-forward mechanism?

Feature Integration Theory

Treisman et al. 1980
Visual Search:
find the green-vertical bar
Conjunction of features

Reaction Time

Single feature

# of distractors
less attentional load more
Peripheral categorization perf. (%) vs. Central discrimination perf. (%)

- Interference at approximately 75% on both axes.
Peripheral categorization perf. (%)

Central discrimination perf. (%)

~75

interference
Peripheral categorization perf. (%) vs. Central discrimination perf. (%)

- Approx. 75% performance in both categories

No interference
individual results

peripheral task performance (%)

central task performance (%)

Fei-Fei et al. *PNAS*, 2002
Compare to seemingly simpler tasks
Compare to seemingly simpler tasks
Compare to seemingly simpler tasks
Compare to seemingly simpler tasks
Are animals special?
Without color…

Fei-Fei et al. *Vis. Cog.*, 2004
Categorization without attention: Single Image vs. Double Images

Fei-Fei et al. Vis. Cog., 2004
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 )
F.I.T. predicted...

less attentional load more
Our data shows...

Li et al. 2002; Fei-Fei et al. 2005
Rapid Perception of Natural Scenes

- Where/how does this happen?
500 ms 32-45 ms 500 ms < 2000 ms

6 AFC, N = 4, error bars: s.e.m.
Behavioral Performance

<table>
<thead>
<tr>
<th>Viewed image category (ground truth)</th>
<th>beaches</th>
<th>buildings</th>
<th>forests</th>
<th>highways</th>
<th>industry</th>
<th>mountains</th>
</tr>
</thead>
<tbody>
<tr>
<td>beaches</td>
<td>0.53</td>
<td>0.03</td>
<td>0.05</td>
<td>0.12</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>buildings</td>
<td>0.02</td>
<td>0.88</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>forests</td>
<td>0.03</td>
<td>0.05</td>
<td>0.75</td>
<td>0.01</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>highways</td>
<td>0.04</td>
<td>0.05</td>
<td>0.02</td>
<td>0.74</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>industry</td>
<td>0.03</td>
<td>0.18</td>
<td>0.05</td>
<td>0.04</td>
<td>0.60</td>
<td>0.03</td>
</tr>
<tr>
<td>mountains</td>
<td>0.07</td>
<td>0.05</td>
<td>0.07</td>
<td>0.04</td>
<td>0.03</td>
<td>0.66</td>
</tr>
</tbody>
</table>

chance: 0.167
### PPA: Parahippocampal Place Area

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Intact</th>
<th>Scrambled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faces</td>
<td><img src="image1.jpg" alt="Faces" /></td>
<td><img src="image2.jpg" alt="Scrambled Faces" /></td>
</tr>
<tr>
<td>Objects</td>
<td><img src="image3.jpg" alt="Objects" /></td>
<td><img src="image4.jpg" alt="Scrambled Objects" /></td>
</tr>
<tr>
<td>Houses</td>
<td><img src="image5.jpg" alt="Houses" /></td>
<td><img src="image6.jpg" alt="Scrambled Houses" /></td>
</tr>
<tr>
<td>Scenes</td>
<td><img src="image7.jpg" alt="Scenes" /></td>
<td><img src="image8.jpg" alt="Scrambled Scenes" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Results</th>
<th>IS Scr Int</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intact</td>
<td>0.12</td>
</tr>
<tr>
<td>Scramble</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>-0.08</td>
</tr>
</tbody>
</table>

*Epstein & Kanwisher, 1998*
Pattern Recognition

Training Images:
- Beaches
- Brain images

Testing Image:
- Beach image
- Brain image

Select voxels

Statistical Pattern Recognition Algorithm
- Train
- Test

Guess "beach"
Pattern Recognition

Support Vector Machine (SVM)

Gaussian Naïve Bayes (GNB)

Neural Networks

Statistical Pattern Recognition Algorithm
Experimental Setup (fMRI)

- Passive viewing
- 6 blocks per run (all 6 categories)
- 12 runs for each subject
- Alternating runs feature upright or inverted images
Voxel Selection

Univariate Multiple Regression

Univariate Multiple Regression

F-statistic

beaches
forests
highways
buildings
industry
mountains

beaches
forests
highways
buildings
industry
mountains
Decoding Performance

N = 4 error bars: s.e.m.

Whole brain (1000 voxels)

- chance

hit rate

beaches buildings forests highways industry mountains
Decoding Performance
Decoding Performance

N = 4 error bars: s.e.m.
## Decoding Performance

The table below shows the classifier prediction accuracy for each category, compared to the chance level of 0.167.

<table>
<thead>
<tr>
<th>Viewed image category (ground truth)</th>
<th>Classifier prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>beaches</td>
<td>0.37</td>
</tr>
<tr>
<td>buildings</td>
<td>0.10</td>
</tr>
<tr>
<td>forests</td>
<td>0.08</td>
</tr>
<tr>
<td>highways</td>
<td>0.27</td>
</tr>
<tr>
<td>industry</td>
<td>0.11</td>
</tr>
<tr>
<td>mountains</td>
<td>0.13</td>
</tr>
</tbody>
</table>

The chance level is marked as 0.167, indicating the expected accuracy by chance.
Decoding Performance

N = 4 error bars: s.e.m.
Retinotopic Areas

N = 4 error bars: s.e.m.
Retinotopic Areas Excluded

N = 4 error bars: s.e.m.
Place Network (PPA + RSC)

N = 4 error bars: s.e.m.
Training: upright only; Testing: upright & inverted blocks intermixed
Scene inversion effect

![Graph showing hit rate for upright and inverted images. The upright condition has a higher hit rate, denoted by an asterisk (*) indicating a significant difference. The dashed line represents chance level.]
Scene inversion effect

hit rate

- Upright
- Inverted
- chance

*
Scene inversion effect

- **Upright**
- **Inverted**

- chance

**hit rate**

- Behavior
- Whole brain
- Retinotopic
- Place network

- * indicates significance
- ** indicates high significance
<table>
<thead>
<tr>
<th>Fei-Fei &amp; Perona, CVPR 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>livingroom</td>
</tr>
<tr>
<td>kitchen</td>
</tr>
<tr>
<td>office</td>
</tr>
<tr>
<td>ins. city</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
learning
category models
(and/or) classifiers
feature detection & representation
image representation

codewords dictionary
categorization
recognition
category decision
1. feature detection & representation

2. codewords dictionary

3. image representation
1. Feature detection and representation

extract interest points

- DoG
- Saliency detector (Kadir and Brady)
- grid
1. Feature detection and representation

- SIFT (Lowe ’99)
- Gray scale values

Represent interest points
2. Codewords dictionary formation
3. Image representation

The diagram shows a bar chart with bars representing the frequency of codewords. The bars are color-coded and ordered from left to right based on their frequency. The codewords are depicted as images below the bars. The image on the right is a black and white picture of a coastal scene.
3. Image representation

frequency

codewords
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, just as a movie screen is a copy of the visual image. Hubel and Wiesel’s discoveries have revealed how we know the visual perception to be more complex. By following the visual impulses along their path to the various cell layers of the optical cortex, Hubel and Wiesel have been able to demonstrate that the message about the image falling on the retina undergoes step-wise 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% jump in exports to $750bn, compared with a 18% rise in imports to $660bn. This would also annoy the US, which has long argued that China’s exports are unfairly helped by a deliberately undervalued yuan. Beijing agrees that the surplus is too high, but says the yuan is only one factor. Bank of China governor Zhou Xiaochuan said the country also needed to do more to boost domestic demand so that more goods stayed 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 trade 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.
Learning

- Feature detection & representation
- Image representation
- Codewords dictionary

Category models (and/or) classifiers
A Generative Model

LDA: Blei, Ng, & Jordan. 2003
A Generative Model

scene category

discrete variable: \( c \sim p(c | \eta) \)

\[
p(c|\eta) = \begin{cases} 
\text{forest} & 1 \\
\text{coast} & 2 \\
\text{kitchen} & 1 \\
\text{mountain} & 1 
\end{cases}
\]
A Generative Model

mixing parameter for the latent topics

\[ \pi \sim p(\pi | c, \theta) \]
\[ \sim \text{Dir}(\pi | c, \theta) \]

where \( \sum_{k=1}^{K} \pi_k = 1 \)
A Generative Model

\[ z \sim p(z | \pi) \sim \text{Mult}(z | \pi) \]

discrete variable:

topics proportion of themes

topic #13

topic #15
A Generative Model

\[ \ell, \theta \]

\[ c \rightarrow \pi \]

\[ z \]

\[ x \]

\[ \eta \]

\[ \beta \]

\[ K \]

\[ I \]

\[ x \sim p(x | z, \beta) \]

\[ \sim \text{Mult} \ (x | z, \beta) \]

expected value of $\beta$ given ‘$z=13$’

discrete variable:

patch label

codeword #265
A Generative Model

expected value of $\beta$ given ‘$z=15$’
Find the ‘best’ $\theta$ and $\beta$

**joint probability**

$$p(x, z, \pi|\theta, \beta, c) = p(\pi|c, \theta) \prod_n p(z_n|\pi)p(x_n|z_n, \beta)$$

$$p(x|\theta, \beta, c) = \int p(\pi|c, \theta) \left( \prod_n \sum_{z_n} p(z_n|\pi)p(x_n|z_n, \beta) \right) d\pi$$

- exact inference is intractable
- use Variational Inference
A Generative Model

Variational Inference

\[
\gamma_{ck} = \theta_{ck}^0 + \sum_{n}^{N} \langle \delta(z_n^k = 1) \rangle \\
\langle \log \pi_{ck} \rangle = \Psi(\gamma_{ck}) - \Psi\left(\sum_{k} \gamma_{ck}\right)
\]

\[
\langle \delta(z_n^k = 1) \rangle = \exp\left\{ \langle \log \pi_{ck} \rangle + \sum_{t}^{T} \langle \log \beta_{kt} \rangle \delta(x_n^t = 1) \right\}
\]

\[
\xi_{kt} = \zeta^0 + \sum_{i}^{I} \sum_{n}^{N} \langle \delta(z_{i,n}^k = 1) \rangle \delta(x_{i,n}^t = 1) \\
\langle \log \beta_{kt} \rangle = \Psi(\xi_{kt}) - \Psi\left(\sum_{t} \xi_{kt}\right)
\]

Maximum Likelihood estimation (Minka 2000)
category models (and/or) classifiers

Recognition

codewords dictionary

category decision
\[ c = \arg \max_c p(x \mid c, \theta, \beta) \]
<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>14</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>4</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>4</td>
<td>2</td>
<td>4</td>
<td>2</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>4</td>
<td>2</td>
<td>2</td>
<td>4</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>78</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>suburb</td>
<td>2</td>
<td>88</td>
<td>12</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>forest</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>70</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>coast</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>78</td>
<td>78</td>
<td>20</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>6</td>
<td>8</td>
<td>70</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>opencountry</td>
<td>8</td>
<td>8</td>
<td>10</td>
<td>16</td>
<td>10</td>
<td>10</td>
<td>48</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</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>2</td>
<td>4</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>4</td>
<td>4</td>
<td>4</td>
<td>18</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>livingroom</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>18</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>office</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>12</td>
<td>12</td>
<td>8</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>64</td>
</tr>
</tbody>
</table>
model distance based on topic distribution
Change blindness

Rensink, O’regan, Simon, etc.
Change blindness

Rensink, O’regan, Simons, etc.
what **DO** we 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: $t = 0 \text{ msec}$

Mask onset: $t = PT$

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

Image onset: $t = 0 \text{ msec}$
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 = 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. JoV 2007
Response attributes??

Fei-Fei et al. JoV, 2007
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.
What’s in a glance?

Average fixation time (one glance) = 120-200ms
What’s in a glance?

PT = 107ms

This is outdoors. A black, furry dog is running/walking towards the right of the picture. His tail is in the air and his mouth is open. Either he had a ball in his mouth or he was chasing after a ball. (Subject EC)

PT = 500ms

I saw a black dog carrying a gray frisbee in the center of the photograph. The dog was walking near the ocean, with waves lapping up on the shore. It seemed to be a gray day out. (Subject JB)

A room full of musical instruments. A piano in the foreground, a harp behind that, a guitar hanging on the wall (to the right). It looked like there was also a window behind the harp, and perhaps a bookcase on the left. (Subject RW)

Fei-Fei et al. JoV, 2007
Scene level

Fei-Fei et al. JoV, 2007
Object level

Fei-Fei et al. JoV, 2007
(Social) Events

Fei-Fei et al. JoV, 2007
What, where and who? Classifying events by scene and object recognition

event: Rowing

scene: Lake

Head of the Charles, October, 2003

Athlete

Rowing boat

Tree

Water

L.-J. Li & L. Fei-Fei ICCV 2007
Average Perf. = 73.4%

L.-J. Li & L. Fei-Fei ICCV 2007
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

Rufin VanRullen, Pietro Perona, Christof Koch

#2: decoding the neural representation of natural scene categories

Eamon Caddigan, Dirk Walther, Diane Beck

#3: what can we perceive within a glance of a scene?

Asha Iyer, Pietro Perona, Christof Koch

#4: Bayesian graphical models for natural scene categorization and event recognition

Pietro Perona, Li-Jia Li
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.

Thank you! Thank you!

#1: natural scene categorization entails little attention (Rufin VanRullen, Pietro Perona, Christof Koch)

#2: decoding the neural representation of natural scene categories (Eamon Caddigan, Dirk Walther, Diane Beck)

#3: what can we perceive within a glance of a scene? (Asha Iyer, Pietro Perona, Christof Koch)

#4: Bayesian graphical models for natural scene categorization and event recognition (Pietro Perona, Li-Jia Li)