Visual Object Recognition
Computational Models and Neurophysiological Mechanisms
Neuro 130/230. Harvard College/GSAS 78454

Web site:  http://tinyurl.com/visionclass
  → Class notes, Class slides, Readings Assignments

Location:  Biolabs 2062

Time:  Mondays 03:00 – 05:00

Lectures:
Faculty:  Gabriel Kreiman (and invited guests)
TA:  Will Xiao

Contact information:
Gabriel Kreiman
  gabriel.kreiman@tch.harvard.edu  617-919-2530
Will Xiao
  xiaow@fas.harvard.edu
Office Hours: Before class (Mondays 2pm), after class (Mondays 5pm). By appointment
Class 1 [09/09/2019]. Introduction to visual pattern recognition [Kreiman]
Class 2 [09/16/2019]. Natural image statistics and the retina [Kreiman]
Class 3 [09/23/2019]. Lesions and neurological examination of extrastriate cortex [Kreiman]
Class 4 [09/30/2019]. Psychophysics studies of visual object recognition [Kreiman]
Class 5 [10/07/2019]. Primary visual cortex [Kreiman]
October 14th: University Holiday
Class 6 [10/21/2019]. Adventures into terra incognita [Kasper Vinken]
Class 7 [10/28/2019]. High-level visual cognition [Kohitij Kar]
Class 12 [12/02/2019]. The operating system for vision [Kreiman]
Quick recap
Psychophysics: The study of the dependencies of psychological experiences upon the physical stimuli that generate them

Basic measures:

- **Reaction time** — The time taken by subjects to perform a task or make a judgment can give an indication (or at least an upper bound) of how long the necessary psychological (and hence neural) processing takes.

- **Performance** — Often inversely related to reaction time (speed-accuracy trade-off).

**Threshold** — Stimuli can be varied to determine the threshold for detection, discrimination, or some more complex psychological phenomenon.

- **Eye movements** — Can provide insights about tasks, goals, sampling, attention
• “Law” of Closure — The mind may experience elements it does not perceive through sensation, in order to complete a regular figure (that is, to increase regularity).
• Law of Similarity — The mind groups similar elements into collective entities or totalities. This similarity might depend on relationships of form, color, size, or brightness.
• Law of Proximity — Spatial or temporal proximity of elements may induce the mind to perceive a collective or totality.
• Law of Symmetry (Figure ground relationships) — Symmetrical images are perceived collectively, even in spite of distance.
• Law of Continuity — The mind continues visual, auditory, and kinetic patterns.
• Law of Common Fate — Elements with the same moving direction are perceived as a collective or unit.
Law of closure
Perceiving objects as whole even if they are not complete
Law of closure
Perceiving objects as whole even if they are not complete
Law of proximity
Grouping nearby elements
Law of similarity
Grouping similar elements

Similarity might depend on relationships of form, color, size, or brightness
Law of symmetry
Grouping elements that are symmetrical to each other

http://isle.hanover.edu/Ch05Object/Ch05SymmetryLaw.html
Law of continuity
Continuing visual, auditory, and kinetic patterns
Law of continuity
Continuing visual, auditory, and kinetic patterns
Law of common fate
Grouping elements that move together
What laws can you find here?
What features are important to recognize an object?
Recognition from minimal features
MIRCs
Minimal Recognizable Configurations

Ullman, PNAS 2016
Recognition of caricatures

Images: Hanoch Piven
Holistic representation of faces

Part-whole illusion

McKone et al, Frontiers in Psychology, 2013
Canonical views may help recognition

McKone et al, Frontiers in Psychology, 2013
Holistic representation of faces

Composite illusion

McKone et al, Frontiers in Psychology, 2013
Four key properties of visual recognition

• Selectivity
• Invariance
• Speed
• Large capacity
Tolerance to image transformations

- Scale
- Position
- Rotation (2D)
- Rotation (3D) – viewpoint
- Color
- Illumination
- Cues
- Clutter
- Occlusion
- Other non-rigid transformations (aging, expressions, etc)
Scale tolerance
One-shot learning for scale tolerance

Which one is it?
Tolerance to viewpoint and illumination changes
Visual recognition depends on experience
Recognition of images flashed for ~100 ms (demo)
Visual recognition can be extremely fast

Fig. 1. Choice saccade task. After a pseudo-random fixation period, a blank screen (gap period) for 200 ms preceded the simultaneous presentation of two natural scenes in the left and right hemifields (20 ms). The images were followed by two grey fixation crosses indicating the saccade landing positions.

Table 1
Summary of behavioural results. Participant numbers correspond to those in Fig. 4

<table>
<thead>
<tr>
<th>Subject</th>
<th>N</th>
<th>Accuracy (%)</th>
<th>Median RT (ms)</th>
<th>Min RT (ms)</th>
<th>Mean start (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>682</td>
<td>96.3</td>
<td>227</td>
<td>130</td>
<td>143</td>
</tr>
<tr>
<td>2</td>
<td>774</td>
<td>93.3</td>
<td>200</td>
<td>130</td>
<td>136</td>
</tr>
<tr>
<td>3</td>
<td>726</td>
<td>81.8</td>
<td>201</td>
<td>130</td>
<td>129</td>
</tr>
<tr>
<td>4</td>
<td>563</td>
<td>80.1</td>
<td>191</td>
<td>120</td>
<td>126</td>
</tr>
<tr>
<td>5</td>
<td>672</td>
<td>86.6</td>
<td>159</td>
<td>130</td>
<td>133</td>
</tr>
<tr>
<td>6</td>
<td>675</td>
<td>86.1</td>
<td>224</td>
<td>150</td>
<td>143</td>
</tr>
<tr>
<td>7</td>
<td>574</td>
<td>90.2</td>
<td>204</td>
<td>140</td>
<td>129</td>
</tr>
<tr>
<td>8</td>
<td>653</td>
<td>94.0</td>
<td>213</td>
<td>150</td>
<td>147</td>
</tr>
<tr>
<td>9</td>
<td>694</td>
<td>96.7</td>
<td>251</td>
<td>180</td>
<td>200</td>
</tr>
<tr>
<td>10</td>
<td>534</td>
<td>89.7</td>
<td>236</td>
<td>180</td>
<td>124</td>
</tr>
<tr>
<td>11</td>
<td>739</td>
<td>90.0</td>
<td>253</td>
<td>190</td>
<td>205</td>
</tr>
<tr>
<td>12</td>
<td>652</td>
<td>96.6</td>
<td>276</td>
<td>200</td>
<td>235</td>
</tr>
<tr>
<td>13</td>
<td>703</td>
<td>95.0</td>
<td>238</td>
<td>160</td>
<td>173</td>
</tr>
<tr>
<td>14</td>
<td>769</td>
<td>98.7</td>
<td>301</td>
<td>230</td>
<td>251</td>
</tr>
<tr>
<td>15</td>
<td>529</td>
<td>77.1</td>
<td>233</td>
<td>160</td>
<td>235</td>
</tr>
<tr>
<td>All</td>
<td>8998</td>
<td>90.1</td>
<td>228</td>
<td>120</td>
<td>140</td>
</tr>
</tbody>
</table>

The second column of this table indicates the total number of trials per participant (see Section 2 for details). Columns 3–5 give the mean accuracy, median and minimum reaction time values for each participant shown in Figs. 3B and C. The last column indicates the onset latency of the mean eye trace for each participant (see Fig. 5).

Is information integrated over time?

Original image

Atom 1

Atom 2

Atom 3

Which category?
Animal
Person
Plant
Vehicle

500 ms fixation
with eye tracking

500 ms of
170 Hz noise

Time

SOA

SOA

500 ms noise,
then response screen

Rapid decay in recognition of asynchronously presented object parts

Brief asynchrony disrupts object recognition

The visual system has a very large capacity
Object recognition from partial information
Presence of the occluder can help
Pattern completion: Objects can be recognized from partial information
Object completion task

Unmasked

Whole

Animal
Chair
Vehicle
Fruit
Face

response

Partial

Animal
Chair
Vehicle
Fruit
Face

response

500 ms

33-150 ms

500 ms

500 ms
Strong robustness to limited visibility

![Graph showing performance vs. percent visible]

Tang et al, PNAS 2018
Backward masking allows investigation of computational processing times.
Object completion task (masking)
Backward masking disrupts pattern completion
Beyond pixels – Context matters
Context example
Context example
Eye movements are critical for scene understanding
Example visual search experiment

Zhang et al, Nature Communications 2018
Four key properties of visual search

1. **Selectivity**
   - Distinguishing target from distractors

2. **Invariance**
   - Finding target irrespective of changes in appearance

3. **Efficiency**
   - Rapid search, avoiding exhaustive exploration

4. **Generalization**
   - No training required

---

**Waldo, Wally, Charlie, Walter**

---

**3. Efficiency**
- [Rapid sampling, avoiding exhaustive exploration]

1. Extract initial sensory map → Call **VisualSampling**
2. Propose image gist → Call **RapidPeripheralAssessment**
3. Propose foveal objects → Call **FovealRecognition**
4. Inference from 1+2+3 → Call **PatternCompletion**
5. Temporary information storage → Call **VisualBuffer**
6. Task-dependent sampling → Call **TargetAttentionProposal**
7. Active sampling → Call **EyeMovementImplementation**
8. Detect people → Call **PeopleDetection**
9. Determine spatial relationships → Call **SpatialRelationships**
10. Repeat steps 3+4+5
11. Repeat steps 6-7
12. Repeat 8-9
13. Got answer? → Call **TaskTerminationDecision**
14. If satisfactory, answer the question → Call **TaskReport**

---

**Wald**

---
Visual illusions: The visual system does not always get it right
Doubles?

http://www.francoisbrunelle.com/

Francois Brunelle
Further reading


Original articles cited in class (see lecture notes for complete list)

- McKone et al, Frontiers in Psychology, 2013