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     Will Xiao
gabriel.kreiman@tch.harvard.edu   xiaow@fas.harvard.edu
617-919-2530

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]

Starting from the very beginning

• Let there be light, and there was light.

• Objects reflect light

• Light photons impinge on the retina (Latin: small net)

• The retina conveys visual information to the brain

An oversimplified first-order description:

*The eye functions as a very sophisticated and spectacular digital camera*
An image as a collection of pixels

<table>
<thead>
<tr>
<th>57</th>
<th>53</th>
<th>58</th>
<th>63</th>
<th>44</th>
<th>41</th>
<th>66</th>
<th>93</th>
<th>68</th>
<th>25</th>
<th>67</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>52</td>
<td>117</td>
<td>130</td>
<td>121</td>
<td>124</td>
<td>119</td>
<td>130</td>
<td>94</td>
<td>34</td>
<td>58</td>
</tr>
<tr>
<td>65</td>
<td>106</td>
<td>67</td>
<td>71</td>
<td>84</td>
<td>152</td>
<td>164</td>
<td>142</td>
<td>150</td>
<td>145</td>
<td>143</td>
</tr>
<tr>
<td>111</td>
<td>64</td>
<td>47</td>
<td>55</td>
<td>98</td>
<td>104</td>
<td>117</td>
<td>124</td>
<td>130</td>
<td>147</td>
<td>147</td>
</tr>
<tr>
<td>79</td>
<td>44</td>
<td>40</td>
<td>67</td>
<td>89</td>
<td>80</td>
<td>78</td>
<td>91</td>
<td>107</td>
<td>97</td>
<td>87</td>
</tr>
<tr>
<td>68</td>
<td>44</td>
<td>51</td>
<td>60</td>
<td>66</td>
<td>61</td>
<td>61</td>
<td>69</td>
<td>66</td>
<td>52</td>
<td>48</td>
</tr>
<tr>
<td>47</td>
<td>79</td>
<td>99</td>
<td>57</td>
<td>47</td>
<td>44</td>
<td>47</td>
<td>54</td>
<td>46</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td>50</td>
<td>110</td>
<td>123</td>
<td>70</td>
<td>44</td>
<td>46</td>
<td>45</td>
<td>51</td>
<td>49</td>
<td>43</td>
<td>40</td>
</tr>
<tr>
<td>61</td>
<td>87</td>
<td>95</td>
<td>58</td>
<td>45</td>
<td>55</td>
<td>46</td>
<td>46</td>
<td>51</td>
<td>49</td>
<td>39</td>
</tr>
<tr>
<td>62</td>
<td>72</td>
<td>87</td>
<td>63</td>
<td>59</td>
<td>59</td>
<td>57</td>
<td>48</td>
<td>56</td>
<td>47</td>
<td>44</td>
</tr>
<tr>
<td>49</td>
<td>51</td>
<td>52</td>
<td>52</td>
<td>52</td>
<td>48</td>
<td>48</td>
<td>51</td>
<td>52</td>
<td>55</td>
<td>56</td>
</tr>
</tbody>
</table>
Natural images are special
We only encounter a small subset of the space of possible images

Consider an image of size 100 x 100 pixels
Assume a pixel can have 256 shades of gray
How many such images are possible?

Answer
For a size of 1x1 pixel, there are 256 possible images.
For a size of 1x2 pixels, there are $256^2$ possible images.
For a size of 100x100 pixels, there are $256^{10000}$ possibilities.*

Yet, we only encounter a small fraction of these possibilities in natural images

*Some of those are “related” by translation, rotation or inversion, etc
Spatial regularities in natural scenes

The properties of nearby points are correlated

Figure 3  (a) Joint distributions of image pixel intensities separated by three different distances. (b) Autocorrelation function.
Images contain different spatial frequencies
Some are more prevalent than others

Low frequencies  Middle frequencies  High frequencies

Natural image statistics

Power spectrum $\sim 1/f^2$

$log(f(w)) = \alpha \log(w) + c$

Note: Scale invariance

$w' \rightarrow aw$

$log(f(w')) = \beta \log(w) + d$

There are multiple examples of power law distributions in physics, biology and social sciences

Figure 4  Power spectrum of a natural image (solid line) averaged over all orientations, compared with $1/f^2$ (dashed line).

Simoncelli and Olshausen 2001
There are also strong correlations in time

The visual input is largely static, except for:

- External object movements
- Head movements
- Eye movements

The visual image is largely static over hundreds of milliseconds.

Silent Reading 225-250 ms fixation, 2 degrees saccade size (8-9 letters)
Scene Perception 260-330 ms fixation, 4 degrees saccade

“Slowness” has been proposed as a constraint for learning about objects (Földiak 1991, Stringer et al 2006, Wiskott et al 2002, Li et al 2008)
Saccadic eye movements
Example pattern of fixations while examining an image
Microsaccades are important for perception.

The eye constantly makes very small “fixational” movements, a.k.a. microsaccades.

When visual stimulation is held fixed relative to the retina, visual perception rapidly fades (Ratliff & Riggs, 1950).

These microsaccades may be critical for counteracting perceptual fading (McCamy, 2012).
The image is focused onto the retina.

Images are inverted by the eye.

- Object
- Eye lens
- Focal plane
Why don’t we see everything upside down?

Perception can adapt to reversing the image.
Some preliminary experiments on vision without inversion of the retinal image. Stratton, 1896
The blind spot

In octopi, light reaches directly to the photoreceptors before reaching the ganglion cells.
“Seeing” the blind spot
Evolution of the retina

4. Blind spot is only present in vertebrates

1. In octopi, light reaches directly the photoreceptors before reaching the ganglion cells
Sizes are measured in degrees of visual angle

\[ \alpha / 2 = \arctan \left( \frac{h}{2d} \right) \]

Size of the moon ~ 0.5 degrees
Thumb at arms length ~ 2 degrees
The retina: An amazingly beautiful circuitry composed of many different cell types

- ~0.5 mm thick
- 5 x 5 cm retinal area
- Three cellular layers
- Rods (low-illumination conditions, \(~10^8\))
- Cones (color sensitivity, \(~10^6\))
- Blind spot
- Fovea (rod free, ~0.5 mm, ~1.7 deg)
- Midget ganglion cells (small dendritic arbors)
- Parasol ganglion cells (large dendritic arbors)

Dowling (2007), Scholarpedia, 2:3487
The retina: Some cells fire action potentials whereas other cells show graded responses

- Photoreceptors transduce incoming light input into electrical signals
- Rod to bipolar convergence increases rod-pathway sensitivity
- Cones, rods, horizontal and bipolar cells are non-spiking neurons
- Many different types of amacrine cells (some of which fire action potentials)
- Retinal ganglion cells fire action potentials and carry the output signals

Rods see largely in grayscale
Non-uniform sampling of the visual field

There is much more detail at the fovea
To be, or not to be, that is the question:  
Whether 'tis nobler in the mind to suffer 
The slings and arrows of outrageous fortune,  
Or to take Arms against a Sea of troubles,  
And by opposing end them: to die, to sleep 
No more; and by a sleep, to say we end 
The heart-ache, and the thousand natural shocks 
That Flesh is heir to? 'Tis a consummation 
Devoutly to be wished. To die, to sleep,  
perchance to **Dream**; aye, there's the rub, 
For in that sleep of death, what dreams may come,  
When we have shuffled off this mortal coil,  
Must give us pause. There's the respect 
That makes Calamity of so long life:  
For who would bear the Whips and Scorns of time, 
The Oppressor's wrong, the *proud* man's Contumely,  
The pangs of *dispised* Love, the Law's delay,  
The insolence of Office, and the spurns
There is much more detail at the fovea.
Example psychophysics task
Indistinguishable images due to reduced acuity in the periphery

Picture 1 is flashed
Picture 2 is flashed
Picture 3 is flashed

Was there an L or a T in the center?

Was picture 3 more similar to picture 1 or picture 2?

Sample 1 or 2?
The retina has a huge dynamic range especially at high background brightnesses. Likewise it is uncertain that all summation has ceased for the stimulus of 19 degrees$^2$ (4.90 diameter) and 1 sec duration, used for measuring threshold intensity. For this reason thresholds were determined at intermediate durations and areas at the various background intensities on one subject. It was hoped that these results would also show whether the changes in total summation involved changes in both temporal and spatial summation, and whether the changes took place in the upper limits of complete summation (Bunsen-Roscoe and Ricco ranges), or whether the amounts, or ranges, of partial summation were involved.

Vision works well in moonless nights and in bright sunlight (~$10^9$ more light intensity)

There are several mechanisms that support light adaptation:
- Changing pupil size
- Reliance on rods vs. cones
- Photopigment bleaching
- Feedback from horizontal cells to photoreceptors

These mechanisms are relatively slow

Barlow, 1958
The receptive field

Neurons throughout the visual system are very picky about the stimulus location.

This cartoon neuron responds only when a flash of light appears in the periphery, in the lower left quadrant.

Blumberg and Kreiman, 2010
Physiology of retinal ganglion cells

The receptive field of most RGCs have a center-surround structure.

Kuffler, S. (1953) J. Neurophys. 16: 37-68

About 1.2 million projections from each retina towards the brain

Minority of RGCs have more complex response properties:

- Phasic cells respond briefly to stimulus onset, offset, or both
- Some phasic cells respond selectively to edge orientation
- Suppressed-by-contrast cells fire except when an edge is present in receptive field
- Bistratified RGCs lack surrounds and are color-sensitive
- Color-opponent cells have centers and surrounds with opposing color preferences
- Intrinsically photosensitive RGCs contain photoreceptors and project to regions controlling pupil size, circadian rhythm, etc.
- Direction-sensitive cells respond to direction of motion of light or dark spots

These cells likely account for approximately 10% of RGCs
Unclear to what extent, these cells contribute to visual object recognition

Stone and Fukuda, *Journal of Neurophysiology* 1974
Cleland and Levick, *Journal of Neurophysiology* 1974
Berson et al., *Science* 2002
## Eye versus digital cameras

<table>
<thead>
<tr>
<th></th>
<th>One eye</th>
<th>Camera (phone)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angle of view</td>
<td>~ 120 degrees</td>
<td>~ 70 degrees</td>
</tr>
<tr>
<td>Sampling</td>
<td>Eccentricity dependent</td>
<td>Uniform</td>
</tr>
<tr>
<td>Resolution</td>
<td>~ $10^7$ “MP” in the fovea</td>
<td>~ 20 MP</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>High sensitivity in low light conditions</td>
<td>Struggle in low light and requires very long integration time</td>
</tr>
<tr>
<td>Energy efficiency</td>
<td>Small fraction of 3 meals a day, say 14 hours of continuous use. (2500 calories / day ~ 0.12 Watts)</td>
<td>Continuous video recording: ~ 3 hours on one battery life (~5 Watts, hence &gt; 12 watts a day)</td>
</tr>
<tr>
<td>Post-processing</td>
<td>The human brain!</td>
<td>Minimal computer vision*</td>
</tr>
</tbody>
</table>
# Eyes in non-human animals

<table>
<thead>
<tr>
<th></th>
<th>Humans</th>
<th>Other animals</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Wavelength</strong></td>
<td>“Visible” spectrum</td>
<td>Also UV, IR</td>
<td>mice, dogs, birds: UV snakes: IR</td>
</tr>
<tr>
<td><strong>Cones</strong></td>
<td>3</td>
<td>1-30</td>
<td>bat, racoon: 1 cat, dogs: 2 some dragon flies: 30</td>
</tr>
<tr>
<td><strong>Polarization</strong></td>
<td>No</td>
<td>Some animals</td>
<td>cuttlefish</td>
</tr>
<tr>
<td><strong>Number of eyes</strong></td>
<td>2</td>
<td>Up to 10</td>
<td>spiders: 8-12 horseshoe crabs: 10</td>
</tr>
<tr>
<td><strong>Angle</strong></td>
<td>Binocular: ~120° Up to 180°</td>
<td>Up to 300°</td>
<td>rabbits: no binocular zone in the center cows, horses: 300°</td>
</tr>
<tr>
<td><strong>Eyes</strong></td>
<td>Correlated</td>
<td>Sometimes independent</td>
<td>Chameleon: can move eyes independently</td>
</tr>
<tr>
<td><strong>Resolution</strong></td>
<td>~100M receptors</td>
<td>Wide variation</td>
<td>Star fish: 200 receptors Birds: higher acuity than humans</td>
</tr>
</tbody>
</table>
To cortex, through the thalamus: The Lateral Geniculate Nucleus (LGN)

• Layers 2, 3 and 5 receive ipsilateral eye’s input
• Layers 1, 4 and 6 receive contralateral eye’s input
• Layers 1-2: magnocellular layers, input from parasol retinal ganglion cells (∼ motion)
• Layers 3-6: parvocellular layers, input from midget ganglion cells (∼ color)
• On and Off center cells are present in all layers
• LGN does not project back to the retina

NOTE: Most of the input to the LGN comes from visual cortex and not from the retina! (e.g. Douglas and Martin 2004)
Difference of Gaussians

The center-surround structure can be described by a difference of gaussians (Mexican-hat)

\[ D(x,y) = \pm \left( \frac{1}{2\pi\sigma_{\text{cen}}^2} \exp\left[ -\frac{x^2 + y^2}{2\sigma_{\text{cen}}^2} \right] - \frac{B}{2\pi\sigma_{\text{sur}}^2} \exp\left[ -\frac{x^2 + y^2}{2\sigma_{\text{sur}}^2} \right] \right) \]

Center response (\(\sigma_{\text{cen}}\))

Surround response (\(\sigma_{\text{sur}}\))
Neurons respond with transient bursts of activity

Dynamic receptive fields in the retina/LGN

\[ D(x,y,t) = \pm \left( \frac{D_{\text{cen}}(t)}{2\pi\sigma_{\text{cen}}^2} \exp\left[ -\frac{x^2 + y^2}{2\sigma_{\text{cen}}^2} \right] - \frac{BD_{\text{sur}}(t)}{2\pi\sigma_{\text{sur}}^2} \exp\left[ -\frac{x^2 + y^2}{2\sigma_{\text{sur}}^2} \right] \right) \]

\[ D_{\text{cen}}(t) = \alpha_{\text{cen}}^2 t \exp\left[ -\alpha_{\text{cen}} t \right] - \beta_{\text{cen}}^2 t \exp\left[ -\beta_{\text{cen}} t \right] \]

\[ D_{\text{sur}}(t) = \alpha_{\text{sur}}^2 t \exp\left[ -\alpha_{\text{sur}} t \right] - \beta_{\text{sur}}^2 t \exp\left[ -\beta_{\text{sur}} t \right] \]

Difference of Gaussians in space and time
The center-surround structure can also be seen in receptive field dynamics

Information from the retina goes to many places
Subcortical visual pathways

Retinal projections

<table>
<thead>
<tr>
<th><strong>Lateral geniculate nucleus (LGN)</strong></th>
<th>Main visual pathway in birds, reptiles, fish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Superior Colliculi</td>
<td>Implicated in saccade generation in mammals</td>
</tr>
<tr>
<td>Suprachiasmatic Nucleus</td>
<td>Hypothalamus: involved in circadian rhythms</td>
</tr>
<tr>
<td>Pretectum</td>
<td></td>
</tr>
<tr>
<td>Pregeniculate</td>
<td></td>
</tr>
<tr>
<td>Accessory optic system</td>
<td>Primates can recognize objects after lesions to the Superior Colliculus or Suprachiasmatic Nucleus, but not after lesions to V1 (Gross 1994).</td>
</tr>
</tbody>
</table>
Further reading

• Class notes: http://tinyurl.com/vision-class

Some of the original articles cited in class (see lecture notes for full list)

• Simoncelli and Olshausen. Annual Review of Neuroscience 2001
• Dowling J. Scholarpedia 2007.
• Felleman and Van Essen. Cerebral Cortex 1991.
• Kuffler. Journal of Neurophysiology 1953.


