Visual Object Recognition

Neurobiology 230 – Harvard / GSAS 78454

Gabriel Kreiman

Email: gabriel.kreiman@tch.harvard.edu
Phone: 617-919-2530
Web site: http://tinyurl.com/vision-class
Dates: Mondays
Time: 3:30 – 5:30 PM
Location: Biolabs 1058
Starting from the very beginning

• Objects reflect light

• Light photons impinge on the retina

• The retina conveys visual information to the brain

An oversimplified (and rather erroneous) first-order description:

*The retina functions as a very sophisticated and spectacular digital camera*
Natural images are special
We only encounter a small subset of the space of possible images

• Consider a grayscale image with 256 possible tones
• Consider an image of size 100 x 100 pixels
• 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 or rotation or inversion, etc
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
Spatial aspects of 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. Simoncelli and Olshausen 2001
Natural image statistics
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 (Foldiak 1991, Stringer et al 2006, Wiskott et al 2002, Li et al 2008)
The image is focused onto the retina
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>
The retina
A 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 (high-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
Rods see largely in grayscale
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
- Retinal ganglion cells fire action potentials and carry the output signals

There is much more detail at the fovea.
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 RGC has a center-surround structure

Kuffler, S. (1953)
J. Neurophys. 16: 37-68
Dynamic receptive fields in the retina/LGN

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

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

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

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

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

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

Surround response \((\sigma_{sur})\)

Difference of Gaussians
The center-surround structure can be described by a difference of gaussians (mexican-hat)
The lateral geniculate nucleus (LGN) is the main visual part of the thalamus:
- 6 layers
- Layers 2, 3 and 5 receive ipsilateral input
- Layers 1, 4 and 6 receive contralateral input
- Layers 1-2: magnocellular cells that receive input from M ganglion cells
- Layers 3-6: parvocellular cells that receive input from P ganglion cells
- Right and left visual hemifields are separate in the LGN
- Right and left eyes are separate in the LGN
- The visual field is represented multiple times in the LGN
- 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)

Subcortical visual pathways

Retinal projections

- Lateral geniculate nucleus (LGN) – Thalamus
- Superior Colliculi – Main visual pathway in birds, reptiles, fish
- Suprachiasmatic Nucleus – Hypothalamus: involved in circadian rhythms
- Pretectum
- Pregeniculate
- Accessory optic system

Primates can recognize objects after lesions to the Superior Colliculus but not after lesions to V1 (Gross 1994 for historical overview).
Visual system circuitry
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
Felleman and Van Essen. Cerebral Cortex 1991.