CENG 783
Special topics in Deep Learning

Week 7 – Intro to CVPR & Human Vision
Sinan Kalkan
today

• A quick introduction to Computer Vision and Pattern Recognition
  • The problem(s) that we are trying to solve
  • The challenges
  • The approaches, especially the hand-designed feature-based ones

• Human vision system in a nutshell
  • Main steps of processing starting from the retina
  • Hierarchies of features in human vision system

• What can we learn for CVPR from human vision system?
What is Vision?

David Marr:
* “Understand what is where by looking”
The starting point

Grayscale Image

<table>
<thead>
<tr>
<th>y = 41</th>
<th>42</th>
<th>43</th>
<th>44</th>
<th>45</th>
<th>46</th>
<th>47</th>
<th>48</th>
<th>49</th>
<th>50</th>
<th>51</th>
<th>52</th>
<th>53</th>
<th>54</th>
<th>55</th>
</tr>
</thead>
<tbody>
<tr>
<td>210</td>
<td>209</td>
<td>204</td>
<td>202</td>
<td>197</td>
<td>247</td>
<td>143</td>
<td>71</td>
<td>64</td>
<td>80</td>
<td>84</td>
<td>54</td>
<td>54</td>
<td>57</td>
<td>58</td>
</tr>
<tr>
<td>206</td>
<td>196</td>
<td>203</td>
<td>197</td>
<td>195</td>
<td>210</td>
<td>207</td>
<td>56</td>
<td>63</td>
<td>58</td>
<td>53</td>
<td>53</td>
<td>61</td>
<td>62</td>
<td>51</td>
</tr>
<tr>
<td>201</td>
<td>207</td>
<td>192</td>
<td>201</td>
<td>198</td>
<td>213</td>
<td>156</td>
<td>69</td>
<td>65</td>
<td>57</td>
<td>55</td>
<td>52</td>
<td>53</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>216</td>
<td>206</td>
<td>211</td>
<td>193</td>
<td>202</td>
<td>207</td>
<td>208</td>
<td>57</td>
<td>69</td>
<td>60</td>
<td>55</td>
<td>77</td>
<td>49</td>
<td>62</td>
<td>51</td>
</tr>
<tr>
<td>221</td>
<td>206</td>
<td>211</td>
<td>194</td>
<td>196</td>
<td>197</td>
<td>220</td>
<td>56</td>
<td>63</td>
<td>60</td>
<td>55</td>
<td>46</td>
<td>97</td>
<td>58</td>
<td>106</td>
</tr>
<tr>
<td>209</td>
<td>214</td>
<td>224</td>
<td>199</td>
<td>194</td>
<td>193</td>
<td>204</td>
<td>173</td>
<td>64</td>
<td>60</td>
<td>59</td>
<td>51</td>
<td>62</td>
<td>56</td>
<td>48</td>
</tr>
<tr>
<td>214</td>
<td>212</td>
<td>213</td>
<td>208</td>
<td>191</td>
<td>190</td>
<td>214</td>
<td>60</td>
<td>62</td>
<td>66</td>
<td>76</td>
<td>51</td>
<td>49</td>
<td>55</td>
<td>48</td>
</tr>
<tr>
<td>213</td>
<td>215</td>
<td>215</td>
<td>207</td>
<td>208</td>
<td>180</td>
<td>172</td>
<td>188</td>
<td>69</td>
<td>72</td>
<td>55</td>
<td>49</td>
<td>56</td>
<td>52</td>
<td>56</td>
</tr>
<tr>
<td>209</td>
<td>205</td>
<td>214</td>
<td>205</td>
<td>204</td>
<td>196</td>
<td>187</td>
<td>196</td>
<td>86</td>
<td>62</td>
<td>66</td>
<td>87</td>
<td>57</td>
<td>60</td>
<td>48</td>
</tr>
<tr>
<td>208</td>
<td>209</td>
<td>205</td>
<td>203</td>
<td>202</td>
<td>186</td>
<td>174</td>
<td>185</td>
<td>149</td>
<td>71</td>
<td>63</td>
<td>55</td>
<td>55</td>
<td>45</td>
<td>56</td>
</tr>
<tr>
<td>207</td>
<td>210</td>
<td>211</td>
<td>199</td>
<td>217</td>
<td>194</td>
<td>183</td>
<td>177</td>
<td>209</td>
<td>90</td>
<td>62</td>
<td>64</td>
<td>52</td>
<td>93</td>
<td>52</td>
</tr>
<tr>
<td>208</td>
<td>205</td>
<td>209</td>
<td>209</td>
<td>197</td>
<td>194</td>
<td>183</td>
<td>187</td>
<td>187</td>
<td>239</td>
<td>58</td>
<td>68</td>
<td>61</td>
<td>51</td>
<td>56</td>
</tr>
<tr>
<td>204</td>
<td>206</td>
<td>203</td>
<td>209</td>
<td>195</td>
<td>203</td>
<td>188</td>
<td>185</td>
<td>183</td>
<td>221</td>
<td>75</td>
<td>61</td>
<td>58</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>200</td>
<td>203</td>
<td>199</td>
<td>236</td>
<td>188</td>
<td>197</td>
<td>183</td>
<td>190</td>
<td>183</td>
<td>196</td>
<td>122</td>
<td>63</td>
<td>58</td>
<td>64</td>
<td>66</td>
</tr>
<tr>
<td>205</td>
<td>210</td>
<td>202</td>
<td>203</td>
<td>199</td>
<td>197</td>
<td>196</td>
<td>181</td>
<td>173</td>
<td>186</td>
<td>105</td>
<td>62</td>
<td>57</td>
<td>64</td>
<td>63</td>
</tr>
</tbody>
</table>
Vision: An inverse problem
In fact, it is ill-posed.

A problem is well-posed if:

- A solution exists
- The solution is unique
- The solution depends continuously on the data,
Vision: An inverse problem (3)

Image

3D world

3D world

3D world

Torralba & Freeman
Challenges 1: viewpoint variation

Michelangelo 1475-1564

slide by Fei Fei, Fergus & Torralba
Challenges 1: view point variation

by Roger Shepard ("Turning the Tables")
Challenges 1: view point variation

Ames room

slide credit: A. Torralba
Challenges 1: view point variation
Challenges 2: illumination
Challenges 2: illumination

slide credit: A. Torralba
Challenges 2: illumination

slide credit: A. Torralba
Challenge 3: occlusion

Magritte, 1957

slide by Fei Fei, Fergus & Torralba
Challenge 4: scale

slide by FeiFei, Fergus & Torralba
Challenges 5: deformation
Challenges 6: background clutter

Klimt, 1913
Challenges 7: object intra-class variation

slide by Fei-Fei, Fergus & Torralba
Challenges 8: local ambiguity

slide by Fei-Fei, Fergus & Torralba
Challenges 8: Prior Knowledge

Now

• Image Gradient and Orientation
• Edges
• Corners
• Hessian matrix
Filtering

$g[m,n]$  

$\rightarrow$  

$\rightarrow$  

$f[m,n]$  

Slide: A. Torralba
Linear filtering

\[ f[m,n] = \sum_{k,l} h[m,n,k,l]g[k,l] \]

For a linear system, each output is a linear combination of all the input values:

In matrix form:

\[ F = H \cdot G \]
Linear filtering

\[ f[m,n] = I \otimes g = \sum_{k,l} h[m-k, n-l]g[k, l] \]
Linear filtering

\[ f[m,n] = I \otimes g = \sum_{k,l} h[m-k,n-l]g[k,l] \]

\[
\begin{array}{cccccccc}
111 & 115 & 113 & 111 & 112 & 111 & 111 & 111 \\
135 & 138 & 137 & 139 & 145 & 146 & 149 & 147 \\
163 & 168 & 188 & 196 & 206 & 202 & 206 & 207 \\
180 & 184 & 206 & 219 & 202 & 200 & 195 & 193 \\
189 & 193 & 214 & 216 & 104 & 79 & 83 & 77 \\
191 & 201 & 217 & 220 & 103 & 59 & 60 & 68 \\
195 & 205 & 216 & 222 & 113 & 68 & 69 & 83 \\
199 & 203 & 223 & 228 & 108 & 68 & 71 & 77 \\
\end{array}
\]

\[
\begin{array}{cccc}
-1 & 2 & -1 \\
-1 & 2 & -1 \\
-1 & 2 & -1 \\
\end{array}
\]

\[
\begin{array}{cccccccc}
? & -5 & 9 & -9 & 21 & -12 & 10 & ? \\
? & -29 & 18 & 24 & 4 & -7 & 5 & ? \\
? & -50 & 40 & 142 & -88 & -34 & 10 & ? \\
? & -41 & 41 & 264 & -175 & -71 & 0 & ? \\
? & -23 & 33 & 360 & -217 & -134 & -23 & ? \\
\end{array}
\]

\[
\begin{array}{cccccccc}
\end{array}
\]

Slide: A. Torralba
Impulse

\[ f[m,n] = I \otimes g = \sum_{k,l} h[m-k, n-l]g[k,l] \]
Shifts

\[ f[m,n] = I \otimes g = \sum_{k,l} h[m-k,n-l]g[k,l] \]
Rectangular filter

\[ g[m,n] \times h[m,n] = f[m,n] \]
Rectangular filter

\[ g[m,n] \ast h[m,n] = f[m,n] \]
Rectangular filter

\[ g[m,n] \ast h[m,n] = f[m,n] \]
Sharpening example

Original

Sharpened (differences are accentuated; constant areas are left untouched).
Sharpening

before

after

Slide: A. Torralba
Gaussian filter

\[ G(x, y; \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2 + y^2}{2\sigma^2}} \]

\[ \sigma = 1 \]

\[ \sigma = 2 \]

\[ \sigma = 4 \]
Global to Local Analysis

Dali

Slide: A. Torralba
\[
\begin{bmatrix}
-1 & 1
\end{bmatrix}
\]

\[g[m,n] \times [-1, 1] = h[m,n] \quad \Rightarrow f[m,n]\]
\( [-1 \ 1]^T \)
\[ h_x(x, y) = \frac{\partial h(x, y)}{\partial x} = \frac{-x}{2\pi\sigma^4} e^{-\frac{x^2+y^2}{2\sigma^2}} \]
\[ h_y(x, y) = \frac{\partial h(x, y)}{\partial x} = \frac{-y}{2\pi\sigma^4} e^{-\frac{x^2+y^2}{2\sigma^2}} \]

Magnitude: \[ h_x(x, y)^2 + h_y(x, y)^2 \]  
Edge strength

Angle: \[ \arctan\left( \frac{h_y(x, y)}{h_x(x, y)} \right) \]  
Edge orientation
Image Gradient & Orientation

\[ G = \sqrt{G_x^2 + G_y^2} \]

\[ \Theta = \arctan \left( \frac{G_y}{G_x} \right). \]
The scale of the smoothing filter affects derivative estimates, and also the semantics of the edges recovered.

Forsyth, 2002
Issues:

1) The gradient magnitude at different scales is different; which should we choose?
2) The gradient magnitude is large along thick trail; how do we identify the significant points?
3) How do we link the relevant points up into curves?
4) Noise.

The scale of the smoothing filter affects derivative estimates, and also the semantics of the edges recovered.

Forsyth, 2002
What is an edge?
What is an edge?

- Depth discontinuity
- Material change
- Texture boundary

Slide: A. Torralba
Edges

Slide: A. Torralba
Gaussian gradient boundaries

Human boundaries

Slide: A. Torralba
Corners, Junctions

- Non-accidental features (Witkin & Tenenbaum, 1983)
Corners as distinctive interest points

• Shifting a window in any direction should give a large change in intensity

“flat” region: no change in all directions
“edge”: no change along the edge direction
“corner”: significant change in all directions

Source: A. Efros
Harris Detector formulation

Change of intensity for the shift \([u, v]\):

\[
E(u, v) = \sum_{x, y} w(x, y) \left[ I(x + u, y + v) - I(x, y) \right]^2
\]

Window function

Shifted intensity

Intensity

Window function \(w(x, y) = \)

1 in window, 0 outside

or

Gaussian

Source: R. Szeliski
Taylor Series for 2D Functions

\[ f(x+u, y+v) = f(x, y) + uf_x(x, y) + vf_y(x, y) + \]

**First partial derivatives**

\[ \frac{1}{2!} \left[ u^2 f_{xx}(x, y) + uv f_{xy}(x, y) + v^2 f_{yy}(x, y) \right] + \]

**Second partial derivatives**

\[ \frac{1}{3!} \left[ u^3 f_{xxx}(x, y) + u^2 v f_{xxy}(x, y) + uv^2 f_{xyy}(x, y) + v^3 f_{yyy}(x, y) \right] + \]

**Third partial derivatives**

+ \ldots (Higher order terms)

**First order approx**

\[ f(x+u, y+v) \approx f(x, y) + uf_x(x, y) + vf_y(x, y) \]
Harris Corner Derivation

\[
\sum [I(x + u, y + v) - I(x, y)]^2
\]

\[
\approx \sum [I(x, y) + uI_x + vI_y - I(x, y)]^2 \quad \text{First order approx}
\]

\[
= \sum u^2I_x^2 + 2uvI_xI_y + v^2I_y^2
\]

\[
= \sum \begin{bmatrix} u \\ v \end{bmatrix} \begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \quad \text{Rewrite as matrix equation}
\]

\[
= \begin{bmatrix} u & v \end{bmatrix} \left( \sum \begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} \right) \begin{bmatrix} u \\ v \end{bmatrix}
\]
Harris Detector formulation

This measure of change can be approximated by:

\[ E(u, v) \approx \begin{bmatrix} u \\ v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix} \]

where \( M \) is a 2×2 matrix computed from image derivatives:

\[
M = \sum_{x, y} w(x, y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}
\]

Gradient with respect to \( x \), times gradient with respect to \( y \)

Sum over image region – area we are checking for corner

Source: R. Szeliski
Harris Detector formulation

where $M$ is a $2 \times 2$ matrix computed from image derivatives:

$$M = \sum_{x, y} w(x, y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}$$

Sum over image region – area we are checking for corner

Gradient with respect to $x$, times gradient with respect to $y$
Interpreting the eigenvalues of $\mathbf{M}$

Classification of image points using eigenvalues of $\mathbf{M}$:

- **“Corner”**
  - $\lambda_1$ and $\lambda_2$ are large,
  - $\lambda_1 \sim \lambda_2$;
  - $E$ increases in all directions

- **“Edge”**
  - $\lambda_2 \gg \lambda_1$

- **“Flat”** region
  - $\lambda_1$ and $\lambda_2$ are small;
  - $E$ is almost constant in all directions

Source: R. Szeliski
Corner response function

\[ R = \det(M) - \alpha \text{trace}(M)^2 = \lambda_1 \lambda_2 - \alpha(\lambda_1 + \lambda_2)^2 \]

\( \alpha \): constant (0.04 to 0.06)

Source: R. Szeliski
Harris Detector: Properties

- Rotation invariance

Ellipse rotates but its shape (i.e. eigenvalues) remains the same

Corner response $R$ is invariant to image rotation
Harris Detector: Properties

- Not invariant to image scale

All points will be classified as edges

Corner!
**Hessian Matrix**

- The eigenvalues of the 2D case can be used for corner & edge detection.

$$H(f) = \begin{bmatrix}
\frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\
\frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2}
\end{bmatrix}.$$  

Harris:

$$M = \sum_{x,y} w(x, y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}.$$
Outline of approaches in Computer Vision & Pattern Recognition

Preprocessing
- Noise removal
- Contrast/color enhancement
- Background subtraction

Feature Extraction
- Keypoint detection
- Keypoint extraction

Learning / Classification

Models
FEATURES WIDELY USED IN COMPUTER VISION
Feature Extraction: Overview

- Two phases:
  - Keypoint (interest point) detection
  - Keypoint description

- Desired properties:
  - Repeatable
  - Scale invariant
  - Rotation invariant
  - Translation invariant
Feature Extraction:
Keypoint (interest point) detection

• What can be keypoints?
  – Edges
  – Corners
  – Segments / blobs
  – ...

(a) Color Labels (ACA)  (b) Texture Classes
(c) Crude Segmentation (d) Final Segmentation
Feature Extraction:
Keypoint (interest point) detection

- **Difference of Gaussians**
  - Gaussian is a smoothing function:
    \[
    G(x, y) = \frac{1}{2\pi \sigma^2} e^{-\frac{x^2 + y^2}{2\sigma^2}}
    \]
  - When a difference of Gaussians of different widths is applied, we get edge detection.
    \[
    \Gamma_{\sigma_1, \sigma_2}(x) = I \ast \frac{1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_1^2}} - I \ast \frac{1}{\sigma_2 \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_2^2}}.
    \]
Feature Extraction: Scale-space representation

- A single scale (width of the kernel) is not sufficient.
  - **Solution**: apply DoG or LoG at different scales, and find the maximum among scales as well as space.
Feature Extraction:
Keypoint description

• Keypoint detection gives us “interesting point”s and a rough estimate of the scales (widths).
• We need to represent the information at these points:
  – Color
  – Texture
  – Gradient response
  – Image orientation
  – …

Feature Extraction:
Keypoint descriptors

• Distribution-based
  – Scale-Invariant Feature Transform, Speeded-up Robust Features, ...

• Frequency or filter-based
  – Fourier Descriptors, Gabor Filters, ...

• Differential descriptors
  – Steerable filters, Histogram of Oriented Gradients, ...

• Moment-based
  – Zernike moments, ...

• Others
  – Local Binary Patterns
  – Global Binary Patterns

Feature Extraction:

Scale-Invariant Feature Transform

1. Construct Scale Space
2. Take Difference of Gaussians
3. Locate DoG Extrema
4. Sub Pixel Locate Potential Feature Points
5. Filter Edge and Low Contrast Responses
6. Assign Keypoints Orientations
7. Build Keypoint Descriptors
8. Go Play with Your Features!!

Jason Clemons
Feature Extraction:
Scale-Invariant Feature Transform (SIFT)

1. Scale gives the width of the interest point.
2. Compute peak orientation to determine the orientation of the interest region
   – Weighted by the gradient and Gaussian centered at the interest point
3. Divide the region into 4x4 grids, and for each grid, compute local gradient orientation histogram (8-bin each)
4. This gives us a feature vector of size $4 \times 4 \times 8 = 128$.

Lowe, David G. (2004). Distinctive image features from scale-invariant key points. *IJCV.*
Feature Extraction:

Speeded-up Robust Features (SURF)

1. Uses Hessian as the detector.
2. Scale-space is used similar to SIFT.
3. Orientation is estimated from a circular region around the interest point.
   - Gradient and orientation are extracted using a fast version of Haar wavelets.
4. The region is divided into a 4x4 grid. In each grid, $\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|$ are computed.
5. This gives us a feature vector of size 4x4x4=64.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding, 2008
Feature Extraction:

Histogram of Oriented Gradients (HOG)

- **Dense feature extraction**
  - The image is divided into windows.
- **Each window is divided into blocks of cells.**
  - In each cell, a histogram of orientations weighted by gradients is calculated.
- **The concatenation of histograms is the feature vector.**

Feature Extraction:
Histogram of Oriented Gradients (HOG)

Feature Extraction: Shapeme Histograms

- By Mori et al. (2005)

Shape Context histograms + k-means clustering + histogram

Shape Context (Belongi et al., 2002)
Feature Extraction: Global Binary Patterns


\[ (c_m^h, c_m^v) \]

(a) Vertical multiplication
\[ (2^{-j_1} 2^{-j_2} 2^0 2^{-1} 2^{-2} 2^{-3}) \]

(b) Horizontal multiplication

(c) Horizontal sum

\[ GBP_{h,v} = GBP_h \oplus GBP_v \]
Feature Extraction: Bag of Words

- Instead of using features directly, first group them into categories ("words") and then use those words.
  - More efficient
  - Better performance

Legend:
- Training local feature
- Cluster 1
- Cluster 2
- Cluster 3
- Test local feature

Local visual features of training images → Clustering → Clustered training features → Testing → Local visual features of test images → Word assignment → Bag of words

Normalization:
\[ \frac{12^2 + 4^2 + 3^2}{13} = 13 \]

Normalized bag of words (length one)
EARLY VISION
Sensory Coding

• Efficient Coding Hypothesis
  – “The goal of early vision (or, early visual processing) is to provide an efficient representation of the incoming visual signal”

• For a review & critics:
  – (Simoncelli & Olshausen, 2001; Simoncelli, 2003)

Independent Component Analysis:

\[
\begin{align*}
&= s_1 \cdot \begin{array}{c}
\end{array} + s_2 \cdot \begin{array}{c}
\end{array} + \cdots + s_k \cdot \begin{array}{c}
\end{array}
\end{align*}
\]
Gabor Filter

- Gaussian multiplied by a sinusoidal.

\[ g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp \left( -\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2} \right) \exp \left( i \left( \frac{2\pi x'}{\lambda} + \psi \right) \right) \]

Real

\[ g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp \left( -\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2} \right) \cos \left( \frac{2\pi x'}{\lambda} + \psi \right) \]

Imaginary

\[ g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp \left( -\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2} \right) \sin \left( \frac{2\pi x'}{\lambda} + \psi \right) \]

where

\[ x' = x \cos \theta + y \sin \theta \]

and

\[ y' = -x \sin \theta + y \cos \theta \]
Gabor Filters vs. Cortical Receptive Fields

Fig. 1. Perspective plot of the two-dimensional profile of sensitivity of the model of the receptive field employed in this analysis. (a) The whole shape. (b) A cross section in the axis parallel to the preferred orientation of the receptive field, which highlights the Gaussian distribution used to describe the height ($y_0$) dimension. (c) A cross section in the dimension orthogonal to the preferred orientation, which highlights the d-DOG-S function used to model the structure of the excitatory and inhibitory zones in the receptive field.

Copyright © 1988 by the Optical Society of America and reprinted by permission of the copyright owner.

Two-dimensional spatial structure of receptive fields in monkey striate cortex

A. J. Parker and M. J. Hawken
Gabor Filters vs. Cortical Receptive Fields

An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex

JUDSON P. JONES AND LARRY A. PALMER

FIG. 2. The two-dimensional (2D) Gabor filter fits simple cell 2D spatial response profiles. Each part of this figure illustrates a 2D spatial response profile, the corresponding least-squared error best-fitting 2D Gabor filter, and the residual error, that function of space which remains after the 2D Gabor filter has been subtracted from the data.
Retinal Receptive Fields

Receptive field structure in ganglion cells: On-center Off-surround

Stimulus condition

Electrical response

© Stephen E. Palmer, 2002
Receptive field structure in ganglion cells: On-center Off-surrond
Receptive field structure in **ganglion cells**: On-center Off-surround

Stimulus condition  Electrical response

Response

Time

© Stephen E. Palmer, 2002
Receptive field structure in **ganglion cells**: On-center Off-surrond

Stimulus condition  
Electrical response

Response  
Time
Retinal Receptive Fields

Receptive field structure in **ganglion cells**: On-center Off-surround

Stimulus condition  Electrical response

Response

Time

© Stephen E. Palmer, 2002
Receptive field structure in ganglion cells: On-center Off-surround

Stimulus condition

Electrical response

Response

Time

© Stephen E. Palmer, 2002
Retinal Receptive Fields

RF of On-center Off-surround cells

Neural Response

Receptive Field

Response Profile

Center

Surround

On

Off

Firing Rate

Horizontal Position

© Stephen E. Palmer, 2002
RF of Off-center On-surround cells

Retinal Receptive Fields

Neural Response

Surround

Center

On

Off

Receptive Field

+ -

Response Profile

Firing Rate

on-surround

off-center

Horizontal Position

© Stephen E. Palmer, 2002
http://fourier.eng.hmc.edu/e180/lectures/v1/node7.html
Cortical Receptive Fields

Simple Cells: “Line Detectors”

A. Light Line Detector
   - - - - + + + + - - - -
   - - - - + + + + - - - -
   - - - - + + + + - - - -
   - - - - + + + + - - - -
   - - - - + + + + - - - -
   - - - - + + + + - - - -
   - - - - + + + + - - - -
   - - - - + + + + - - - -

Firing Rate
Horizontal Position

B. Dark Line Detector
   + + + + - - - - + + + +
   + + + + - - - - + + + +
   + + + + - - - - + + + +
   + + + + - - - - + + + +
   + + + + - - - - + + + +
   + + + + - - - - + + + +
   + + + + - - - - + + + +
   + + + + - - - - + + + +

Firing Rate
Horizontal Position

© Stephen E. Palmer, 2002
Cortical Receptive Fields

Simple Cells: “Edge Detectors”

C. Dark-to-light Edge Detector

D. Light-to-dark Edge Detector

Firing Rate

Horizontal Position

© Stephen E. Palmer, 2002
The early ‘low-level filtering’ achieves efficient encoding by finding intensity changes.
  – With Gabor-like filters.
This is one of the important evidence for the hypothesis that our visual system is tuned to the statistical regularities in the environment.
HUMAN VISION IN A NUTSHELL

(Slides are mostly from Norbert Krueger)
Deep Hierarchies in the Primate Visual Cortex: What Can We Learn For Computer Vision?

Norbert Krüger, Peter Janssen, Sinan Kalkan, Markus Lappe, Aleš Leonardis, Justus Piater, Antonio J. Rodríguez-Sánchez, Laurenz Wiskott

Abstract—Computational modeling of the primate visual system yields insights of potential relevance to some of the challenges that computer vision is facing, such as object recognition and categorization, motion detection and activity recognition or vision-based navigation and manipulation. This article reviews some functional principles and structures that are generally thought to underlie the primate visual cortex, and attempts to extract biological principles that could further advance computer vision research. Organized for a computer vision audience, we present functional principles of the processing hierarchies present in the primate visual system considering recent discoveries in neurophysiology. The hierarchal processing in the primate visual system is characterized by a sequence of different levels of processing (in the order of ten) that constitute a deep hierarchy in contrast to the flat vision architectures predominantly used in today’s mainstream computer vision. We hope that the functional description of the deep hierarchies realized in the primate visual system provides valuable insights for the design of computer vision algorithms, fostering increasingly productive interaction between biological and computer vision research.

Index Terms—Computer Vision, Deep Hierarchies, Biological Modeling

1 INTRODUCTION

The history of computer vision now spans more than half a century. However, general, robust, complete satisfactory solutions to the major problems such as large-scale object, scene and activity recognition and categorization, as well as vision-based manipulation are still beyond reach of current machine vision systems. Biological visual systems, in particular those of primates, seem to accomplish these tasks almost effortlessly and have been, therefore, often used as an inspiration for computer vision researchers.

Interactions between the disciplines of “biological vision” and “computer vision” have varied in intensity throughout...

David Hubel and Torsten Wiesel

The Nobel Prize in Medicine 1981
Some remarks on the interaction of human vision research and computer vision

- David Marr 1982: *Vision: A computational investigation into the human representation and processing of visual information*

- 3 Stages
  - Primal Sketch: Multi-scale Edge Detection
  - 2.5D Sketch: Viewer centered Scene Representation
  - 3D Sketch: Object Centered Representation
Marr’s Paradigm

- Marr: “Visual processing is modular”

Marr’s Paradigm


From S. Lehar:
Marr’s Paradigm

- Marr’s 3D Model description:

Why did that ‘fail’? Two reasons

• The project was too ambitious at Marr’s time
  – Lack of knowledge on low-level modalities
    • Optic flow
    • Edge detection
    • Stereo
    • Structure-from-Motion

• Lack of computational resources
  – Slow clock frequency
  – No GPUs
‘Computer Vision’ and ‘Biological Vision’

• In the 80th and 90th there was a strong link

• This link has been kind of diluted from ‘both sides’
  – Computer Vision became a sub-discipline of Machine Learning
  – Many neurophysiologists have given up on understanding the brain on a functional level

• ‘Biologically inspired’ got a somehow bad reputation
  – Not efficient
  – Everything could somehow be biologically inspired
Maybe a restart is worthwhile

- Much better understanding of early vision
- Significantly larger computational resources
- Still many unsolved problems in CV
- Aim of the paper
  - Distill essential knowledge on the human visual system for Engineers
Overview

• The primate’s vision system: A deep Hierarchy
• Reflections
Basic facts

- 55% of the neo-cortex of the primate brain is concerned with vision
- Division in
  - Occipital Cortex
  - Dorsal Pathway
  - Ventral Pathway
Vision for recognition & action

- **Ventral pathway:**
  - Provides “what”
  - Object recognition

- **Dorsal pathway:**
  - Provides “where”
  - Vision for action
Dr. Alesha Sivartha in the late 1800s
(published in his metaphysical book *The Book of Life: The Spiritual and Physical Constitution of Man*)

From: van Essen 1992
Half of the brain in 15 minutes
## Basic Facts

<table>
<thead>
<tr>
<th>Area</th>
<th>Size (mm²)</th>
<th>RFS</th>
<th>Latency (ms)</th>
<th>co/bi lat.</th>
<th>rt/st/cl/co</th>
<th>CI/SI/PI/OI</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-cortical processing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retina</td>
<td>1018</td>
<td>0.01</td>
<td>20-40</td>
<td>bl</td>
<td>+/-/-/+</td>
<td>-/-/-</td>
<td>sensory input, contrast computation, relay, gating</td>
</tr>
<tr>
<td>LGN</td>
<td>0.1</td>
<td></td>
<td>30-40</td>
<td>co</td>
<td>+/-/-/+</td>
<td>-/-/-</td>
<td></td>
</tr>
<tr>
<td>V1</td>
<td>1120</td>
<td>3</td>
<td>30-40</td>
<td>co</td>
<td>+/-/-/+</td>
<td>-/-/-</td>
<td>generic feature processing</td>
</tr>
<tr>
<td>V2</td>
<td>1190</td>
<td>4</td>
<td>40</td>
<td>co</td>
<td>+/-/-/+</td>
<td>-/-/-</td>
<td>generic feature processing</td>
</tr>
<tr>
<td>V3/V3A/VP</td>
<td>325</td>
<td>6</td>
<td>50</td>
<td>co</td>
<td>+/-/-/+</td>
<td>-/-/-</td>
<td>generic feature processing</td>
</tr>
<tr>
<td>V4/VOT/V4t</td>
<td>650</td>
<td>8</td>
<td>70</td>
<td>co</td>
<td>+/-/-/+</td>
<td>+/-/-/+</td>
<td>generic feature processing / color motion</td>
</tr>
<tr>
<td>MT</td>
<td>55</td>
<td>7</td>
<td>50</td>
<td>co</td>
<td>+/-/-/+</td>
<td>+/-/-/+</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>3340</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ventral Pathway / What (Object Recognition and Categorization)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TEO</td>
<td>590</td>
<td>3-5</td>
<td>70</td>
<td>co</td>
<td>(+/-/-/+</td>
<td>-/-/-/+</td>
<td>object recognition and categorization</td>
</tr>
<tr>
<td>TE</td>
<td>180</td>
<td>10-20</td>
<td>80-90</td>
<td>bl</td>
<td>+/-/-/+</td>
<td>+/-/-/+</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>770</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dorsal Pathway / Where and How (Coding of Action Relevant Information)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MST</td>
<td>60</td>
<td>&gt;30</td>
<td>60-70</td>
<td>bl</td>
<td>+/-/-/-</td>
<td>I</td>
<td>optic flow, self-motion, pursuit</td>
</tr>
<tr>
<td>VIP</td>
<td>40</td>
<td>10-30</td>
<td>50-60</td>
<td>bl</td>
<td>+/-/-/-</td>
<td>+/-/-/?</td>
<td>optic flow, touch, near extra personal space</td>
</tr>
<tr>
<td>7a</td>
<td>115</td>
<td>&gt;30</td>
<td>90</td>
<td>bl</td>
<td>(+/-/-/-</td>
<td>+/-/-/?</td>
<td>Optic flow, heading</td>
</tr>
<tr>
<td>LIP</td>
<td>55</td>
<td>12-20</td>
<td>50</td>
<td>cl</td>
<td>+/-/-/-</td>
<td>+/-/-/-</td>
<td>salience, saccadic eye movements</td>
</tr>
<tr>
<td>AIP</td>
<td>35</td>
<td>5-7</td>
<td>60</td>
<td>bl</td>
<td>+/-/-/?</td>
<td>+/-/-/?</td>
<td>grasping</td>
</tr>
<tr>
<td>MIP</td>
<td>55</td>
<td>10-20</td>
<td>100</td>
<td>co</td>
<td>+/-/-/?</td>
<td></td>
<td>reaching</td>
</tr>
<tr>
<td>Sum</td>
<td>585</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 1**

Basic facts on the different areas of the macaque visual cortex based on different sources [44], [28], [95], [141], [161]. *First column:* Name of Area. *Second column:* Size of area in mm². *?’* indicates that this information is not available. *Third column:* Average receptive field size in degrees at 5 degree of eccentricity. *Fourth column:* Latency in milliseconds. *Fifth Column:* Contra versus bilateral receptive fields. *Sixth Column:* Principles of organization: Retinotopic (rt), spatiotopic (st), clustered (cl) columnar (co). *Seventh Column:* Invariances in representation of shape: Cue-Invariance (Ci), Size Invariance (Si), Position Invariance (Pi), Occlusion Invariance (Oi). *’I’* indicates that this entry is irrelevant for the information coded in these areas. *Eighth Column:* Function associated to a particular area.
Pre-cortical Areas
Precortical Areas

Retina

LGN

- No Feature Transformation
- Preparing for Stereo
Occipital Cortex

- **V1**: Primary visual cortex
- **V2**: Second visual area
- **V3/V3A**: Third visual area
- **V4**: Fourth visual area
- **TE (AIT)**
- **TEO (PIT)**
- **MST**: Middle suprasylvian gyrus
- **VIP**: Ventral intraparietal area
- **LIP**: Lateral intraparietal area
- **AIP**: Anterior intraparietal area
- **PMD, PMV**: Premotor cortex
- **FEF, Frontal Cortex, Occulomotor SC, Brain Stem**: Prefrontal cortex
- **Arm control, PMD**: Arm control
- **Hypocampus Memory**: Memory
- **Ventral pathway**: Ventral stream
- **Dorsal pathway**: Dorsal stream

**Dorsal Stream**
- **Dorsal lateral geniculate nucleus**
- **Thalamus**
- **Striate cortex (primary visual cortex)**
- **Extrastriate cortex**
- **Ventral Stream**
- **Inferior temporal cortex: Second level of visual association cortex**

**Object recognition pathway**

**Space/Action pathway**

**Retina**

**Second level of visual association cortex in parietal lobe**
Occipital Cortex

• More than 70% of the visual cortex
  – Occipital Cortex 3340mm\(^2\)
  – Ventral Pathway 770mm\(^2\)
  – Dorsal Pathway 585mm\(^2\)

• Processing
  – Task unspecific generic scene representation
Occipital Cortex: V1 and V2
Concept of Hue as Object Property
Linguistic Concept of ‘red’ or ‘blue’

2D Motion  3D Motion
Ventral Pathway
Ventral Pathway

• Ca. 17% of the visual cortex
  – Occipital Cortex 3340mm²
  – Ventral Pathway 770mm²
  – Dorsal Pathway 585mm²

• Processing
  – Object Recognition and Categorization
  – Many suggestions for how to divide into areas
Ventral Pathway: TEO and TE

Orientation and shape selectivity.

Selectivity to complex features / shapes

Tanaka
Dorsal Pathway

- Hypocampus
- Prefrontal cortex (non motor)
- FEF, frontal SC, Brain Stem
- PMD, PMV, Prefrontal cortex (Hand control, rule based behaviors)
- Vestibular information about arm, eye and head position

- Occipital cortex
- Dorsal lateral geniculate nucleus
- Thalamus
- Extrastriate cortex
- Inferior temporal cortex: Second level of visual association cortex
Dorsal Pathway

• Ca 15% of the visual cortex
  – Occipital Cortex 3340mm²
  – Ventral Pathway 770mm²
  – Dorsal Pathway 585mm²

• Processing
  – Much less known than Ventral Pathway
  – Many more distinguished areas
  – Coding visual information related to action and position in space
Dorsal Pathway

CIP
Cue invariant 3D shape

MST
Ego-motion

AIP
Hand shape and affordances

MIP
Reaching

VIP
Ego-space

LIP
Saccadic related retinotopic repr.
Overview

• Some annoying prior remarks
• The primate’s vision system: A deep Hierarchy
• What can we learn?
What can we learn?

• The primate’s vision
  – uses a **huge amount of resources** in the brain organized in a **deep hierarchy**
  – devotes its main vision resources to a **two step computation of a rich set of descriptors in a generic representation**
Deep Hierarchy

- Richness of representation
- Deep Hierarchy versus flat Architectures
- Separation of information
- Feedback
- Learning versus hard-wiring
Flat versus deep Hierarchies

Deep Hierarchy

<table>
<thead>
<tr>
<th>Task V1</th>
<th>Task V2</th>
<th>Task V3</th>
<th>Task Vn</th>
<th>Task D1</th>
<th>Task D2</th>
<th>Task D3</th>
<th>Task Dn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 5 (ventral)</td>
<td>Level 5 (dorsal)</td>
<td>Level 4</td>
<td>Level 3</td>
<td>Level 2</td>
<td>Level 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Flat Hierarchy

<table>
<thead>
<tr>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
<th>Task 5</th>
<th>Task 6</th>
<th>Task 7</th>
<th>Task 8</th>
<th>Task n</th>
</tr>
</thead>
</table>

Some kind of Features
Example of a flat hierarchy

The frog does not seem to see or, at any rate, is not concerned with the detail of stationary parts of the world around him. He will starve to death surrounded by food if it is not moving. His choice of food is determined only by size and movement. He will leap to capture any object the size of an insect or worm, providing it moves like one. He can be fooled easily not only by a bit of dangled meat but by any moving small object. His sex life is conducted by sound and touch. His choice of paths in escaping enemies does not seem to be governed by anything more devious than leaping to where it is darker. Since he is equally at home in water and on land, why should it matter where he lights after jumping or what particular direction he takes? He does remember a moving thing providing it stays within his field of vision and he is not distracted.

Increasing Level of Abstraction
Increasing Level of Abstraction
Flat versus deep hierarchies

- Flat Hierarchies are inefficient
  - No sharing of computational resources
  - Transfer of experience across tasks is facilitated within the same representations
What can we learn?

• The primate’s vision
  – uses a huge amount of resources in the brain organized in a deep hierarchy
  – devotes its main vision resources to a two step computation of a rich set of descriptors in a generic representation
V1 and V2: A Zoo of Features

Local Processing      Semi-Global Processing
Perspective

- A stable, rich and disambiguated generic scene representation
What does human vision suggest?

• There is evidence that a good part of the local feature processing in V1 works from scratch
  – Hubel and Wiesel, Bonhoefer et al.

• However, lateral connections are probably heavily influenced by statistics of experience

• Gestalt laws and pictorial depth cues are used only after 6 months of development

• Possible Approach
  – Understand how extractable 3D features in structured regions relate to not yet computed 3D structures at homogeneous areas
  – If there are strong conditional probabilities that would be a good sign
  – Integrate semantic scene understanding and depth processing
Gestalt laws and Pictorial Depth Cues

- Proximity
- Similarity
- Closure
- Area
- Symmetry
- Continuity
- Shading
- Occlusion
- Texture Gradient
- Familiar Size
- Linear Perspective
Conclusion

• The primate’s vision system: A deep Hierarchy

• Can we learn something?

• What can we learn?
  – Deep Hierarchies
  – Large resources to generic scene representation

• My personal believe:
  – We need more structure than Deep ANNs provide

• But
  – I might actually be wrong
FINAL REMARKS
Many similarities between deep learning and our brain

Sensory Coding

- **Efficient Coding Hypothesis**
  - “The goal of early vision (or, early visual processing) is to provide an efficient representation of the incoming visual signal”

- **For a review & critics:**
  - (Simoncelli & Olshausorden, 2001; Simoncelli, 2003)

Independent Component Analysis:

\[ s = s_1 \cdot s_2 + \cdots + s_k \]  

(Kalkan et al., 2008)

(Olshausen & Field, 1996)

(Hyvarinen, 2010)
Compare the basis functions of patches with the following.