Where do you look on this image?
The squares depicts where 20 subjects looked.
Saliency and Visual Attention
Adapted from T. Judd
Adapted from T. Judd
Adapted from T. Judd
Adapted from T. Judd
Adapted from T. Judd
Adapted from T. Judd
Adapted from T. Judd
Adapted from T. Judd
Adapted from T. Judd
need to prioritize the visual information and decide what is most important

Adapted from T. Judd
Scene Analysis and Eye Movements

• Our visual system processes an enormous amount of data coming from the retina. $\sim 10^9$ bits/sec (Itti, PhD Thesis, 2000)

• We have developed some selection mechanisms to find the most relevant parts of a scene.
  – Warning (noticing predators, sudden motion, etc.)
  – Exploration (finding preys, locating objects, etc.)

Adapted from T. Judd, M. S. Lewicki
Attention and Scene Analysis

• **Attention** is a complex set of interrelated processes.
  – selection of information (bottom-up)
  – integration of that information with existing knowledge (top-down)

• Why do perceptual systems **need attention**?
  – limited resources
  – Even though we have $10^{12}$ neurons, the brain is still not sufficient to process all the information coming out of the retina
  – simplifies the problem computationally by selecting information
  – perceptual constancy by separating the “foreground” from the “background”

Adapted from M. S. Lewicki
• **Gaze** refers to coordinated eye-head movements during shifts in visual attention.

• **Fixation** is the maintaining of the visual gaze on a single location.

• **Saliency** describes the distinctive nature of regions/objects about how they stand out in relation to their surroundings (grab our attention).
What is attention?

• "Everyone knows what attention is."
  
  William James

• Theories of attention
  – Feature-Integration Theory
    [Treisman and Gelade 1980]
  – Textons [Julezs 1981]
  – Guided Search Theory [Wolfe 1989]
  – Similarity [Duncan and Humphreys 1987]
Theories of Visual Attention

- Feature-Integration Theory [Treisman and Gelade 1980]

- Textons [Julezs 1981]

- Similarity [Duncan and Humphreys 1987]

- Guided Search Theory [Wolfe 1989]

[Healey and Enns 2011]
Attention Mechanisms

• **Attention** is a complex set of interrelated processes
  - selection of information (bottom-up)
  - integration of that information with existing knowledge (top-down)

• Bottom-up attention
  - very rapid, task-independent

• Top-down attention
  - slower, task dependent
Top-down Attention

Yarbus (1967) was the first to show that task influences fixation locations.

“They did not expect him” by Repin

Adapted from T. Judd
Saccades and Fixations

What are the material circumstances of the family?

What were they doing before arrival?

Remember object and person positions

Free Examine

What are their ages?

Remember the clothes

How long has the unexpected visitor been away?

Slide credit: D. Hoeim

[Yarbus 1967]
Computational Models of Visual Attention

• Can machines predict where the humans look at a given image?
Itti-Koch Model (1998)

- *First computational model of visual attention to predict where people look*
- An implementation of Koch & Ullman, 1985
- employs a multi-scale center-surround mechanism which imitates the workings of the retinal receptive field
Bottom-up Saliency Models

(1) Extract visual features

(2) Compute a saliency map for each feature channel

(3) Compute a final saliency map by combining individual saliency maps

Zhao Q, Koch CJ Vis 2012;12:22
Predicted Fixations

Slide credit: D. Hoeim

[Itti Koch Niebur 1998]
Many models of saliency have been introduced

<table>
<thead>
<tr>
<th>Biologically Inspired</th>
<th>Mathematically Inspired</th>
<th>Add top-down features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerf et al. (2007)</td>
<td>Graphical Model</td>
<td></td>
</tr>
<tr>
<td>Rosenholtz (1999)</td>
<td>Esaliency</td>
<td></td>
</tr>
<tr>
<td>Le Meur et al. (2006)</td>
<td>Information theoretic approach</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“Surprise” model</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from T. Judd
## Which one is the best?

<table>
<thead>
<tr>
<th>Biologically Inspired</th>
<th>Mathematically Inspired</th>
<th>Add top-down features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerf et al. (2007)</td>
<td>Graphical Model</td>
<td></td>
</tr>
<tr>
<td>Rosenholtz (1999)</td>
<td>Esaliency</td>
<td></td>
</tr>
<tr>
<td>Le Meur et al. (2006)</td>
<td>Information theoretic approach</td>
<td></td>
</tr>
<tr>
<td>SUN model</td>
<td>Itti and Baldi (2006)</td>
<td></td>
</tr>
<tr>
<td>Goferman et al. (2009)</td>
<td>“Surprise” model</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Elazary and Itti (2010)</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from T. Judd
Which one is the best?

<table>
<thead>
<tr>
<th>Biologically Inspired</th>
<th>Mathematically Inspired</th>
<th>Add top-down features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerf et al. (2007)</td>
<td>Graphical Model</td>
<td>(search task)</td>
</tr>
<tr>
<td>SUN model</td>
<td>“Surprise” model</td>
<td></td>
</tr>
</tbody>
</table>

Benchmark needed!

Adapted from T. Judd
How benchmark data sets are formed?

Through eye tracking experiments

Adapted from T. Judd
Fixations for one observer

Adapted from T. Judd
Fixations from 15 observers

Adapted from T. Judd
Fixation map created from gaussian convolution over fixations

Adapted from T. Judd
Human Eye Fixations Data Sets

• Bruce data set [Bruce and Tsotsos 2006]
  – 120 color images of indoor and outdoor scenes
  – Eye movement data from 20 subjects

• MIT 1003 data set [Judd et al. 2009]
  – 1003 natural color images (779 landscape + 228 portrait)
  – Eye movement data from 15 subjects

• MIT 300 data set [Judd et al. 2012]
  – 300 natural images (223 landscape + 77 portrait)
  – Eye movement data from 39 subjects
Sample Images

Bruce data set

MIT 1003 data set

MIT 300 data set
Why are fixations center biased?

- photographer bias
- viewing strategy

Adapted from T. Judd
Evaluation Metrics

- Area under ROC curve (AUC)
- Earth Mover’s Distance (EMD)
- Normalized Scanpath Saliency (NSS)
- Similarity Score
- ...
Area under ROC curve (AUC) as a performance measure

Saliency Map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure
Area under ROC curve (AUC) as a performance measure

Thresholded Saliency Map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)
Area under ROC curve (AUC) as a performance measure

Thresholded Saliency Map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Percentage of fixations are calculated that lie within the salient portion of the map

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Thresholded Saliency Map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Thresholded Saliency Map

Receiver Operating Characteristic curve

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Thresholded Saliency Map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Thresholded Saliency Map

Receiver Operating Characteristic curve

% of human fixations (True positives) vs Percent Salient (False positives)

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Thresholded Saliency Map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

ROC curve always starts at 0 ends at 1
Area under ROC curve (AUC) as a performance measure

Thresholded Center Map

Receiver Operating Characteristic curve

% of human fixations (True positives) vs. Percent Salient (False positives)

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Thresholded Center Map

Receiver Operating Characteristic curve

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure
Area under ROC curve (AUC) as a performance measure

Thresholded Center Map

Receiver Operating Characteristic curve

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure
Area under ROC curve (AUC) as a performance measure

Random Noise map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Random Noise map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)

Chance performance

Adapted from T. Judd
Area under ROC curve (AUC) as a performance measure

Perfect saliency map

Receiver Operating Characteristic curve

% of human fixations (True positives)

Percent Salient (False positives)

Best performance

Chance performance

Adapted from T. Judd
Density map from human fixation data
[Itti et al. 1998]
[Harel et al. 2007]
[Goferman et al. 2010]
Our Model
Density map from human fixation data
Density map from human fixation data
[Harel et al. 2007]
Our Model
Some other comparative results

<table>
<thead>
<tr>
<th>Human density map</th>
<th>Itti et al. (1998)</th>
<th>Harel et al. (2007)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our approach</strong></td>
<td>Covariance</td>
<td>Covariance + CB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Human density map</th>
<th>Itti et al. (1998)</th>
<th>Harel et al. (2007)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our approach</strong></td>
<td>Covariance</td>
<td>Covariance + CB</td>
</tr>
</tbody>
</table>

**Evaluation scores**

In (Li et al., 2012), the AUC score and the maximal value of the Dice Similarity Coefficient (DSC) curve are used in quantitative evaluation. The predicted saliency maps are thresholded and the thresholded binary maps are compared against the binary ground truth images provided in the data set using these scores.

The DSC is a measure of set agreement defined by

$$DSC = \frac{2TP}{2TP + FP + FN}$$

where $TP$ is the true positive, $FP$ is the false positive, and $FN$ is the false negative counts. A DSC value of 1 indicates a perfect agreement whereas a DSC value of 0 means no overlap, so a good salient object model should give a DSC value close to 1.

**Performance**

Detecting salient objects on the ImgSal data set poses some great challenges such as variation in scale, cluttered backgrounds, repeating distractors, etc. The images contain one or more objects which are distinguishable from the background by their visual characteristics but with different difficulty levels. In Figure 7, we present some qualitative examples. The illustrated object maps are...
Understanding attention enables applications in computer graphics & vision, design

- image cropping / thumbnailing
- image and video compression
- non photorealistic rendering
- scene understanding
- advertising and package design
- web usability
- localization / recognition
- object detection
- navigational assistance
- robot active vision
- surveillance systems
- assistive technology for blind or low-vision people
Detecting Salient Objects

• The task is to detect foreground objects that attracts more attention in a given image
Detecting Salient Objects

• The task is to detect foreground objects that attracts more attention in a given image
Detecting Salient Objects

• ImgSal data set [Li, Levine, An, Xu & He, 2012]
• 235 natural images of size 480 x 640 pixels
• 19 subjects are asked to label the most salient objects in the images presented to them
• 6 different categories
  – large salient regions,
  – intermediate salient regions,
  – small salient regions,
  – cluttered backgrounds,
  – repeating distractors,
  – large and small salient regions)
Action Recognition and Gaze

- Predicted human fixations can improve action recognition performances (Mathe and Sminchisescu, ECCV 2012)
Fig. 1. Heat maps generated from the fixations of 16 human subjects while viewing 5 videos selected from the Hollywood-2 and UCF Sports datasets. Notice that fixated locations are generally tightly clustered. This is suggestive of a significant degree of consistency among human subjects in terms of the spatial distribution of their visual attention. See fig. 2b,c for quantitative studies.

The divide is well epitomized by the lack of matching large scale datasets that would provide recordings of the workings of the human visual system, in the context of a visual recognition task, at different levels of interpretations including neural systems or eye movements. The human eye movement level, defined by image fixations and saccades, is potentially the less controversial to measure and analyze. It is sufficiently 'high-level' or 'behavioral' for the computer vision community to rule-out, to some degree at least, open-ended debates on where and what should one record, as could be the case, for instance with neural systems in different brain areas[5]. Besides, our goals in this context are pragmatic: fixations provide a sufficiently high-level signal that can be precisely registered with the
Summary

• Can machines predict where the humans look at a given image?

• State-of-the-art computational models of visual attention reach a certain point of success.

• Selection mechanisms provided by the computational saliency models could improve several machine perception applications.
Thanks for your attention!