BIL 717
Image Processing

Erkut Erdem
Dept. of Computer Engineering
Hacettepe University

Visual Saliency
Where do you look on this image?
The squares depicts where 20 subjects looked
Adapted from T. Judd
Adapted from T. Judd
Adapted from T. Judd
Adapted from T. Judu
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
Terminology

- **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 [Julesz 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

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?
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>(search task)</td>
</tr>
<tr>
<td>Seo &amp; Milanfar (2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang &amp; Cottrell (2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUN model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goferman et al. (2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achanta (2010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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.

![Diagram of the Itti-Koch Model](image)
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

x: Image Location

G(x): saliency at x

Zhao Q, Koch C J Vis 2012;12:22
Feature Extraction

• Three basic visual features:

1. Intensity Contrast
   \[ I = \frac{r + g + b}{3} \]

2. Color
   \[ R = r - \frac{g + b}{2} \quad B = b - \frac{r + g}{2} \]
   \[ G = g - \frac{r + b}{2} \quad Y = \frac{r + g}{2} - \frac{|r - g|}{2} - b \]

3. Orientation
Feature Extraction

- Feature maps are extracted by considering center-surround differences between a “center” fine scale \( c \) and a “surround” coarser scale \( s \):

\[
c \in \{2, 3, 4\} \quad s = c + \delta, \quad \delta \in \{3, 4\}
\]

\[
\mathcal{I}(c, s) = |I(c) \ominus I(s)|
\]

\[
\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))|
\]

\[
\mathcal{BY}(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|
\]

\[
\mathcal{O}(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|
\]

\[
\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}
\]

In total, 42 different feature maps are estimated: 6 for intensity, 12 for color, and 24 for orientation.
Normalization operator

- Normalize values to fixed range [0..M]
- Compute the average of local maxima: m
- Multiply the feature map by \((M - m)^2\)
Saliency Estimation

- Compute individual saliency maps for each visual feature:

\[
\bar{I} = \bigoplus_{c=2}^{4} \oplus_{s=c+3}^{s} \mathcal{N}(I(c, s))
\]

\[
\bar{C}^- = \bigoplus_{c=2}^{4} \oplus_{s=c+3}^{s} \left[ \mathcal{N}(R \cdot G(c, s)) + \mathcal{N}(B \cdot Y(c, s)) \right]
\]

\[
\bar{O} = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} \mathcal{N}\left( \bigoplus_{c=2}^{4} \oplus_{s=c+3}^{s} \mathcal{N}(O(c, s, \theta)) \right)
\]

- Combine these individual saliency maps by averaging:

\[
S = \frac{1}{3}(\mathcal{N}(\bar{I}) + \mathcal{N}(\bar{C}) + \mathcal{N}(\bar{O}))
\]
Predicted Fixations

Input image

\[ \overrightarrow{\vec{c}} \]

\[ \vec{f} \]

\[ \delta \]

Output (FOA)

92 ms

145 ms

206 ms

260 ms

Slide credit: D. Hoeim

[Itti Koch Niebur 1998]
GBVS (Harel et al, 2007)

- A graph-based bottom-up model
- Several feature maps are extracted at multiple spatial scales and then represented as fully connected graphs.
  - Vertices: grid positions,
  - Edges: relationships between pairs of vertices

\[
d((i, j) \| (p, q)) \triangleq \left| \log \frac{M(i, j)}{M(p, q)} \right|, \quad w_1((i, j), (p, q)) \triangleq d((i, j) \| (p, q)) \cdot F(i - p, j - q)
\]

\[
F(a, b) \triangleq \exp \left( -\frac{a^2 + b^2}{2\sigma^2} \right).
\]

- The resulting graphs are used to define Markov chains in which their equilibrium distribution is treated as in the activation and saliency maps.
**AIM (Bruce and Tsotsos, 2006, 2009)**

- A bottom-up saliency model based on an information-theoretic perspective. (Shannon’s measure of Self-Information)
- A sparse set of basis functions is learned from the input image itself via ICA.
- Features representing the center region and its surrounding regions are described by means of these basis functions.
SUN (Zhang et al., 2009)

• Similar to AIM model, it employs high-level features derived via a visual dictionary, but these features are learned from a training set of natural images.

• Two sets of features based on:
  1. Difference of Gaussians (DoG) filters and 2. ICA

• A global definition of saliency in which the salient parts are estimated by considering the global rarity of the local visual features in the entire image.
CSD (Goferman et al., 2010)

- A context-aware saliency model based on four principles:
  1. Local color and contrast information
  2. Frequently occurring global features (maintain unique features)
  3. Visual organization rules (Gestalt)
  4. High-level features (Faces, Objects, People, etc.)
CSD (Goferman et al., 2010)

- Unique appearance → salient
- Position is important!

\[ d(p_i, p_j) = \frac{d_{\text{color}}(p_i, p_j)}{1 + c \cdot d_{\text{pos}}(p_i, p_j)} \]

Adapted from S. Goferman
Single and multi-scale saliency

• A multi-scale analysis for saliency estimation:

\[ S_i^r = 1 - \exp \left\{ - \frac{1}{K} \sum_{k=1}^{K} d(p_i^r, q_k^r) \right\} \]

Single scale saliency formula

\[ \bar{S}_i = \frac{1}{M} \sum_{r \in R} S_i^r \]

multi-scale saliency formula

Adapted from S. Goferman
CSD (Goferman et al., 2010) - Summary

• Single-scale saliency
  \[ S_i^r = 1 - \exp\left\{ - \frac{1}{K} \sum_{k=1}^{K} d(p_i^r, q_k^r) \right\} \]

• Multiple scales
  \[ \bar{S}_i = \frac{1}{M} \sum_{r \in R} S_i^r \]

• Final saliency
  \[ \hat{S}_i = \bar{S}_i (1 - d_{foci}(i)) \]

Adapted from S. Goferman
• A Bayesian contextual guidance model for visual search tasks which combines low-level salience and scene context:

\[
p(O = 1, X | L, G) = \frac{1}{p(L | G)} p(L | O = 1, X, G) p(X | O = 1, G) p(O = 1 | G)
\]

- Bottom-up factors
- Top-down target knowledge
- Context-based priors
- Target presence
Learning Saliency (Judd et al, 2009)

• Formulate saliency estimation as a supervised learning problem.

• Learn a classification function $f : \mathbb{R}^D \rightarrow \{+1, -1\}$ with $D$ denoting the dimension of the features, which returns the label (+1) for salient (fixated) points, and (−1) for the non-salient points.

• Based on training SVMs to determine the saliency of a pixel.

• Simple low-level features like intensity, orientation and color along with some mid- and high-level features such as responses of horizon line detectors, face and pedestrian detectors.

• Learning of specific feature weights and normalization schemes in the integration step.
Feature Integration

• It still remains an open question how different feature dimensions contribute to the overall saliency.

• Common strategies are:
  – summation, max, etc.

• Some studies try to overcome this issue by finding optimal values for feature weights in a supervised manner as in Judd et al., 2009.
CovSal (Erdem and Erdem, 2012)

- Covariance matrices of simple image features (aka region covariance; Tuzel, Porikli, & Meer, 2006) are used as meta-features for saliency estimation.
- Region covariances capture local image structures better than standard linear filters.
- More importantly, they naturally provide nonlinear integration of different features by modeling their correlations.
CovSal (Erdem and Erdem, 2012)

(a) Input image

(b) Covariance matrices for the regions A-E

- Very simple visual features:

\[
F(x, y) = \begin{bmatrix}
  x & y & L(x, y) & a(x, y) & b(x, y) \\
  \frac{\partial I(x, y)}{\partial x} & \frac{\partial I(x, y)}{\partial y}
\end{bmatrix}^T
\]

(c) Our saliency

(d) Itti’s saliency
Region Covariances

• Let I denote an image, and F be the feature image extracted from I:

\[ F(x,y) = \Phi(I,x,y) \]

with \( \Phi \) denoting the d-dimensional function of features.

• Then, a region R inside F can be represented with a \( d \times d \) covariance matrix \( C_R \) of the feature points:

\[ C_R = \frac{1}{n-1} \sum_{i=1}^{n} (f_i - \mu)(f_i - \mu)^T \]

• Distance between two covariances \( C_1 \) and \( C_2 \) can be estimated using special metrics:

\[ \rho(C_1,C_2) = \sqrt{\sum_{i=1}^{n} \ln^2 \lambda_i(C_1,C_2)} \]

Covariance matrices do not lie on Euclidean space.
Model 1: Saliency using covariance features

- The saliency of a region $R_i$ is defined as the weighted average of the dissimilarities between $R_i$ to the $m$-most similar regions around it:

$$S(R_i) = \frac{1}{m} \sum_{j=1}^{m} d(R_i, R_j)$$

- Dissimilarity measure $d(R_i, R_j)$ is defined as:

$$d(R_i, R_j) = \frac{\rho(C_i, C_j)}{1 + ||x_i - x_j||}$$

Weighting the covariance distance by inverse spatial distance makes the nearby regions have more influence on the saliency computation.
Model 2: Saliency using covariance and mean features

- First-order statistics can be also valuable for saliency.

- **Sigma Points Representation:**
  Every SPD matrix has a unique factorization.

- Let $C$ be a $d \times d$ covariance matrix, the corresponding set of Sigma Points can be computed as:
  \[
  s_i = \begin{cases} 
  \alpha \sqrt{d} L_i & \text{if } 1 \leq i \leq d \\
  -\alpha \sqrt{d} L_i & \text{if } d + 1 \leq i \leq 2d 
  \end{cases}
  \]
  where $L_i$ is the $i^{th}$ column of the lower triangular matrix $L$ obtained with the Cholesky decomposition $C = LL^T$.

- First-order statistics can be easily incorporated to obtain an enriched feature vector: 
  \[
  \Psi(C) = (\mu, s_1, \ldots s_d, s_{d+1}, \ldots, s_{2d})^T
  \]
Model 2: Saliency using covariance and mean features

• The saliency of a region $R_i$ is given by:

$$S(R_i) = \frac{1}{m} \sum_{j=1}^{m} d'(R_i, R_j)$$

• Dissimilarity measure $d(R_i, R_j)$ is defined as:

$$d'(R_i, R_j) = \frac{||\Psi(C_i) - \Psi(C_j)||}{1 + ||x_i - x_j||}$$
Incorporating center bias

- Experiments on human eye fixations demonstrate that there is a tendency in humans to look towards the image center, which is called the center bias.

- This bias is mainly explained by several factors:
  - Photographer bias,
  - Viewing strategy,
  - Motor bias

- We included a center bias into our second model by defining the saliency of region $R_i$ as follows:

$$S'(R_i) = \left(1 - \frac{||x_i - x_c||}{Z}\right) \cdot S(R_i)$$
Scale-space extension

- The objects that can be treated as salient in an image can and do appear over a wide range of scales.
- Saliency detection should be carried out simultaneously at all possible scales.
- Employ a fusion strategy to combine single-scale maps to come up with one final saliency map:

\[
S(x) = G_\sigma(x) \prod_{k \in K} \hat{S}^k(x)
\]
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>(search task)</td>
</tr>
<tr>
<td>SUN model</td>
<td>Itti and Baldi (2006)</td>
<td></td>
</tr>
<tr>
<td>Navalpakkam and Itti (2005)</td>
<td>“Surprise” model</td>
<td></td>
</tr>
<tr>
<td>Goferman et al. (2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achanta (2010)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adapted from T. Judd
Which one is the best?

Biologically Inspired
Itti and Koch (1998)
Cerf et al. (2007)
Hou and Zhang (2007)
Rosenholtz (1999)
Itti and Baldi (2006)
Le Meur et al. (2006)
Seo & Milanfar (2009)
Zhang & Cottrell (2008)
SUN model
Goferman et al. (2009)
Achanta (2010)

Mathematically Inspired
Herral et al. (2007)
Graphical Model
Avraham and Lindenbaum (2009)
Esaliency
Bruce and Tsotsos (2009)
Information theoretic approach
Kienzle et al., (2007)
Gao and Vasconulos (2005)
Itti and Baldi (2006)
“Surprise” model
Navalpakkam and Itti (2005)
Elazary and Itti (2010)

Add top-down features
Ehinger et al., (2009)
(search task)
Oliva et al. (2003)
Torralba et al. (2006)
Zhang et al. (2008)
Kanan et al. (2009)

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>Seo &amp; Milanfar (2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zhang &amp; Cottrell (2008)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUN model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goferman et al. (2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Achanta (2010)</td>
<td></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

Fixation map

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

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

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

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

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

Area under ROC curve (AUC) as a performance measure

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

Thresholded Saliency Map

Receiver Operating Characteristic curve

Area under ROC curve (AUC) as a performance measure

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

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

ROC curve always starts at 0 ends at 1

Adapted from T. Judd
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

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

% of human fixations (True positives)

Percent Salient (False positives)

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

Thresholded Center 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 Center 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

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

Best performance

Chance performance

Adapted from T. Judd
Density map from human fixation data
[Harel et al. 2007]
[Goferman et al. 2010]
Our Model 1
Our Model 2
Comparison to the state-of-the-art models
Comparison to the state-of-the-art models

Human density map
Itti et al. (1998)
Harel et al. (2007)

Torralba et al. (2006)
Hou and Zhang (2007)
Zhang et al. (2008)

Bruce and Tsotsos (2009)
Seo and Milanfar (2009)
Goferman et al. (2010)
Comparison to the state-of-the-art models

Our approach with

- covariances only
- covariances + center
- covariances + means
- covariances + means + center
Comparison to the state-of-the-art models
Comparison to the state-of-the-art models

Human density map  Itti et al. (1998)  Harel et al. (2007)
Comparison to the state-of-the-art models

Our approach with

covariances only

covariances + center

covariances + means

covariances + means + center
Comparison to the state-of-the-art models
Comparison to the state-of-the-art models

Human density map

Itti et al. (1998)

Harel et al. (2007)

Torralba et al. (2006)

Hou and Zhang (2007)

Zhang et al. (2008)

Bruce and Tsotsos (2009)

Seo and Milanfar (2009)

Goferman et al. (2010)
Comparison to the state-of-the-art models

Our approach with

covariances only

covariances + center

covariances + means

covariances + means + center
Quantitative Comparisons

The DSC is a measure of set agreement defined by

\[
\text{DSC} = \frac{2 \times TP}{(TP + FP) + (TP + FN)}
\]

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 that 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 were obtained by setting the threshold as the average intensity of the saliency map plus one standard deviation. Our saliency model detected the salient objects accurately under these difficult scenarios.

We provide quantitative analysis of our model and the state-of-the-art saliency models on the ImgSal data set in Table 4.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC Without CB</th>
<th>AUC With CB</th>
<th>NSS Without CB</th>
<th>NSS With CB</th>
<th>EMD Without CB</th>
<th>EMD With CB</th>
<th>Similarity Without CB</th>
<th>Similarity With CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itti et al. (1998)</td>
<td>0.771</td>
<td>0.825</td>
<td>1.137</td>
<td>1.264</td>
<td>2.906</td>
<td>2.002</td>
<td>0.397</td>
<td>0.521</td>
</tr>
<tr>
<td>Harel et al. (2007)</td>
<td>0.829</td>
<td>0.835</td>
<td>1.533</td>
<td>1.533</td>
<td>2.014</td>
<td>1.886</td>
<td>0.519</td>
<td>0.556</td>
</tr>
<tr>
<td>Torralba et al. (2006)</td>
<td>0.710</td>
<td>0.832</td>
<td>0.805</td>
<td>1.185</td>
<td>3.467</td>
<td>1.868</td>
<td>0.330</td>
<td>0.528</td>
</tr>
<tr>
<td>Hou &amp; Zhang (2007)</td>
<td>0.736</td>
<td>0.835</td>
<td>0.964</td>
<td>1.271</td>
<td>3.791</td>
<td>1.959</td>
<td>0.360</td>
<td>0.550</td>
</tr>
<tr>
<td>Zhang et al. (2008)</td>
<td>0.718</td>
<td>0.832</td>
<td>0.884</td>
<td>1.194</td>
<td>3.954</td>
<td>1.968</td>
<td>0.347</td>
<td>0.541</td>
</tr>
<tr>
<td>Bruce &amp; Tsotsos (2009)</td>
<td>0.728</td>
<td>0.835</td>
<td>0.896</td>
<td>1.165</td>
<td>3.127</td>
<td>1.809</td>
<td>0.351</td>
<td>0.535</td>
</tr>
<tr>
<td>Seo &amp; Milanfar (2009)</td>
<td>0.766</td>
<td>0.845</td>
<td>1.100</td>
<td>1.320</td>
<td>3.222</td>
<td>1.759</td>
<td>0.415</td>
<td>0.579</td>
</tr>
<tr>
<td>Goferman et al. (2010)</td>
<td>0.784</td>
<td>0.841</td>
<td>1.272</td>
<td>1.370</td>
<td>3.520</td>
<td>1.819</td>
<td>0.431</td>
<td>0.574</td>
</tr>
<tr>
<td>Our approach with</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covariances only</td>
<td>0.767</td>
<td>0.834</td>
<td>1.184</td>
<td>1.342</td>
<td>3.142</td>
<td>1.931</td>
<td>0.408</td>
<td>0.546</td>
</tr>
<tr>
<td>Covariances + means</td>
<td>0.765</td>
<td>0.834</td>
<td>1.198</td>
<td>1.396</td>
<td>3.398</td>
<td>1.896</td>
<td>0.402</td>
<td>0.548</td>
</tr>
<tr>
<td>Covariances + center</td>
<td>0.840</td>
<td>0.840</td>
<td>1.753</td>
<td>1.753</td>
<td>1.901</td>
<td>1.901</td>
<td>0.561</td>
<td>0.561</td>
</tr>
<tr>
<td>Covariances + means + center</td>
<td><strong>0.851</strong></td>
<td><strong>0.851</strong></td>
<td><strong>1.891</strong></td>
<td><strong>1.898</strong></td>
<td><strong>1.728</strong></td>
<td><strong>1.728</strong></td>
<td><strong>0.581</strong></td>
<td><strong>0.581</strong></td>
</tr>
<tr>
<td>Center</td>
<td>–</td>
<td>0.803</td>
<td>–</td>
<td>0.969</td>
<td>–</td>
<td>2.401</td>
<td>–</td>
<td>0.478</td>
</tr>
<tr>
<td>Chance</td>
<td>0.505</td>
<td>0.803</td>
<td>–0.001</td>
<td>0.969</td>
<td>5.159</td>
<td>2.339</td>
<td>0.187</td>
<td>0.479</td>
</tr>
</tbody>
</table>

Table 1. Performance comparisons of the saliency models on the Toronto data set. Chance and Center are the baselines, which respectively stand for the random and the centered Gaussian models. CB denotes center bias. The best performing model is shown in bold type.
Quantitative Comparisons

The DSC is a measure of set agreement defined by

$$\text{DSC} = \frac{2 \times TP}{(TP + FP) + (TP + FN)}$$

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 that 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 were obtained by setting the threshold as the average intensity of the saliency map plus one standard deviation. Our saliency model detected the salient objects accurately under these difficult scenarios.

We provide quantitative analysis of our model and the state-of-the-art saliency models on the ImgSal data set in Table 4. The proposed models outperformed the other saliency models in three out of six categories, and it was the second best or third best model in other categories.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC Without CB</th>
<th>AUC With CB</th>
<th>NSS Without CB</th>
<th>NSS With CB</th>
<th>Similarity Without CB</th>
<th>Similarity With CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itti et al. (1998)</td>
<td>0.741</td>
<td>0.827</td>
<td>0.921</td>
<td>1.170</td>
<td>0.273</td>
<td>0.402</td>
</tr>
<tr>
<td>Harel et al. (2007)</td>
<td>0.791</td>
<td>0.829</td>
<td>1.150</td>
<td>1.182</td>
<td>0.319</td>
<td>0.415</td>
</tr>
<tr>
<td>Torralba et al. (2006)</td>
<td>0.700</td>
<td>0.832</td>
<td>0.771</td>
<td>1.156</td>
<td>0.244</td>
<td>0.412</td>
</tr>
<tr>
<td>Hou &amp; Zhang (2007)</td>
<td>0.713</td>
<td>0.833</td>
<td>0.855</td>
<td>1.200</td>
<td>0.264</td>
<td>0.421</td>
</tr>
<tr>
<td>Zhang et al. (2008)</td>
<td>0.703</td>
<td>0.834</td>
<td>0.829</td>
<td>1.177</td>
<td>0.261</td>
<td>0.418</td>
</tr>
<tr>
<td>Bruce &amp; Tsotsos (2009)</td>
<td>0.709</td>
<td>0.835</td>
<td>0.813</td>
<td>1.148</td>
<td>0.254</td>
<td>0.415</td>
</tr>
<tr>
<td>Seo &amp; Milanfar (2009)</td>
<td>0.712</td>
<td>0.836</td>
<td>0.826</td>
<td>1.171</td>
<td>0.263</td>
<td>0.424</td>
</tr>
<tr>
<td>Goferman et al. (2010)</td>
<td>0.758</td>
<td>0.840</td>
<td>1.053</td>
<td>1.241</td>
<td>0.297</td>
<td>0.431</td>
</tr>
<tr>
<td>Our approach with</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covariances only</td>
<td>0.715</td>
<td>0.826</td>
<td>0.862</td>
<td>1.169</td>
<td>0.261</td>
<td>0.410</td>
</tr>
<tr>
<td>Covariances + means</td>
<td>0.740</td>
<td>0.832</td>
<td>0.940</td>
<td>1.240</td>
<td>0.287</td>
<td>0.417</td>
</tr>
<tr>
<td>Covariances + center</td>
<td>0.833</td>
<td>0.833</td>
<td>1.468</td>
<td>1.486</td>
<td>0.417</td>
<td>0.418</td>
</tr>
<tr>
<td>Covariances + means + center</td>
<td><strong>0.843</strong></td>
<td><strong>0.843</strong></td>
<td><strong>1.488</strong></td>
<td><strong>1.543</strong></td>
<td><strong>0.428</strong></td>
<td><strong>0.432</strong></td>
</tr>
<tr>
<td>Center</td>
<td>–</td>
<td>0.810</td>
<td>–</td>
<td>1.004</td>
<td>–</td>
<td>0.379</td>
</tr>
<tr>
<td>Chance</td>
<td>0.500</td>
<td>0.810</td>
<td>−0.000</td>
<td>1.004</td>
<td>0.131</td>
<td>0.383</td>
</tr>
</tbody>
</table>

Table 2. Performance comparisons of the saliency models on the MIT1003 data set. The best performing model is shown in bold type.
Quantitative Comparisons

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC Without CB</th>
<th>AUC With CB</th>
<th>EMD Without CB</th>
<th>EMD With CB</th>
<th>Similarity Without CB</th>
<th>Similarity With CB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itti et al. (1998)</td>
<td>0.750</td>
<td>0.806</td>
<td>4.560</td>
<td>3.394</td>
<td>0.405</td>
<td>0.493</td>
</tr>
<tr>
<td>Harel et al. (2007)</td>
<td>0.801</td>
<td>0.813</td>
<td>3.574</td>
<td>3.315</td>
<td>0.472</td>
<td>0.501</td>
</tr>
<tr>
<td>Torralba et al. (2006)</td>
<td>0.684</td>
<td>0.806</td>
<td>4.715</td>
<td><strong>3.036</strong></td>
<td>0.343</td>
<td>0.488</td>
</tr>
<tr>
<td>Hou &amp; Zhang (2007)</td>
<td>0.682</td>
<td>0.804</td>
<td>5.368</td>
<td>3.200</td>
<td>0.319</td>
<td>0.487</td>
</tr>
<tr>
<td>Zhang et al. (2008)</td>
<td>0.672</td>
<td>0.799</td>
<td>5.088</td>
<td>3.296</td>
<td>0.340</td>
<td>0.473</td>
</tr>
<tr>
<td>Bruce &amp; Tsotsos (2009)</td>
<td>0.751</td>
<td><strong>0.820</strong></td>
<td>4.236</td>
<td>3.085</td>
<td>0.390</td>
<td>0.507</td>
</tr>
<tr>
<td>Goferman et al. (2010)</td>
<td>0.742</td>
<td>0.815</td>
<td>4.900</td>
<td>3.219</td>
<td>0.390</td>
<td><strong>0.509</strong></td>
</tr>
<tr>
<td>Our approach with</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Covariances + center</td>
<td>0.800</td>
<td>0.800</td>
<td>3.422</td>
<td>3.422</td>
<td>0.487</td>
<td>0.487</td>
</tr>
<tr>
<td>Covariances + means + center</td>
<td><strong>0.806</strong></td>
<td><strong>0.811</strong></td>
<td><strong>3.109</strong></td>
<td><strong>3.109</strong></td>
<td><strong>0.502</strong></td>
<td>0.503</td>
</tr>
<tr>
<td>Center</td>
<td>–</td>
<td>0.783</td>
<td>–</td>
<td>3.719</td>
<td>–</td>
<td>0.451</td>
</tr>
<tr>
<td>Chance</td>
<td>0.503</td>
<td>0.783</td>
<td>6.352</td>
<td>3.506</td>
<td>0.327</td>
<td>0.482</td>
</tr>
<tr>
<td>Judd et al. (2009)</td>
<td>0.811</td>
<td>0.813</td>
<td>3.130</td>
<td>3.130</td>
<td>0.506</td>
<td>0.511</td>
</tr>
</tbody>
</table>

Table 3. Performance comparisons of the saliency models on the MIT300 data set. The best performing model is shown in bold type.
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)
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!