Visual saliency

Erkut Erdem
Hacettepe University
Computer Vision Laboratory

Where do we look on these images?

The squares shows where 15 observers looked in eye tracking experiments

Slide credit: T. Judd

What is attention?

• Attention is an umbrella term which refers to the mechanisms by which relevant parts of sensory information are selected for further, more detailed processing, and the rest are discarded.

Every one knows what attention is. It is the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others.

William James, 1890

Why do perceptual systems need attention?

• Limited resources
  • Our visual system processes an enormous amount of data coming from the retina. $\sim 10^8$ bits/sec [Itti, 2000]

• Warning
  • noticing predators, sudden motion,

• Exploration
  • finding preys, locating objects, etc.

The amount of information coming down the optic nerve far exceeds what the brain is capable of fully processing and assimilating into conscious experience.
Attentional 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
  - very rapid, primitive, task-independent
- Top-down
  - slower, under cognitive control, task-dependent

Task-based visual attention

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

Theories of visual attention

- Feature-Integration Theory
  [Treisman & Gelade, 1980]
  - processing occurs in parallel and focused attention occurs in serial

- Guided Search Theory
  [Wolfe, 1989]
  - visual search relies on a combination of bottom-up and top-down activity
**Visual saliency**

- “Saliency at a given location is determined primarily by how **different** this location is from its **surround** in color, orientation, motion, depth, etc.” [Koch & Ullman, 1985]
- “Visual salience (or visual saliency) is the **distinct** subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention.” [Itti, 2007]

**Beyond biology: Applications in Computer Vision**

- Most computer vision algorithms have relied on brute-force (e.g. sliding window) strategies.
- Attentional mechanisms provide a relatively free and fast mechanism to select a few candidates while eliminating background clutter.
- To list a few of possible applications
  - scene classification [Siagian & Itti, 2007]
  - object recognition [Gao et al., 2009; Rutishauser et al., 2004]
  - object tracking [Butko et al., 2008]
  - robotics [Frintrop et al., 2006; Siagian & Itti, 2007]
  - content-based image resizing [Achanta & Susstrunk, 2009; Avidan & Shamir, 2007]

**Computational models of visual saliency**

- Can machines predict where humans look at a given image?

  - [Itti & Koch, 1998]
    - One of the first computational models of visual attention to predict where people look
    - A bottom-up model
    - An implementation of Koch & Ullman, 1985
    - It employs a multi-scale center-surround mechanism which imitates the workings of the retinal receptive field.

**Bottom-up models of visual saliency**

The common basic structure is:

1. Extract visual features,
2. Compute a saliency map for each feature channel
3. Compute a final saliency map by combining individual saliency maps

Image credit: [Zhao & Koch, 2012]
Feature integration step

- The most troublesome step
  - typically carried out by taking weighted average (linear summation).
  - But how different feature dimensions contribute to the overall saliency is still an open question! [Callaghan, 1989, 1990; Eckstein et al., 2000; Rosenholtz, 1999, 2001; Rosenholtz et al., 2004]

CovSal (Erdem and Erdem, 2013)

- The region covariance descriptor [Tuzel et al., 2006]
  - captures local image structures better than standard linear filters.
  - naturally provides nonlinear integration of different features by modeling their correlations.

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

$$\{f_i\}_{i=1}^{n} : d\text{-dimensional feature points inside } R$$

$$\begin{bmatrix} L(x,y) & a(x,y) & b(x,y) & \frac{\partial L(x,y)}{\partial x} & \frac{\partial L(x,y)}{\partial y} & x & y \end{bmatrix}^T$$

$$a_i = \begin{cases} \pm \sqrt{d} & \text{if } 1 \leq i \leq d \\ -\pm \sqrt{d} & \text{if } d+1 \leq i \leq 2d \end{cases}$$

$$C = LL^T$$ Cholesky decomposition

- Covariances alone can not explain changes in the means!

- We additionally incorporate first-order statistics
  - Signapoints [Hong et al., 2009; Julier & Uhlmann, 1996]

$$s_i = \begin{cases} \pm \sqrt{d} & \text{if } 1 \leq i \leq d \\ -\pm \sqrt{d} & \text{if } d+1 \leq i \leq 2d \end{cases}$$

- Final representation: $$\Psi(C) = (\mu, s_1, \ldots s_d, s_{d+1}, \ldots, s_{2d})^T$$
CovSal (Erdem and Erdem, 2013)

• Visual dissimilarity between two patches $R_1$ and $R_2$ can be computed by using the following metrics:

For covariance descriptor:
$$p(C_1, C_2) = \sqrt{\sum_{i=1}^{n} b_i^2 \lambda_i(C_1, C_2)}$$

[Förstner & Moonen, 1999]

For sigma points descriptor:
$$||\Psi(C_i) - \Psi(C_j)||$$

CovSal (Erdem and Erdem, 2013)

• If the patch is highly dissimilar to the patches surrounding it $\rightarrow$ rare/salient
• Otherwise $\rightarrow$ common/non-salient

CovSal (Erdem and Erdem, 2013)

• The saliency of $R_i$ is defined as the weighted average of the dissimilarities between $R_i$ to the $m$ most similar regions around it.

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

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

Model 2
$$d(R_i, R_j) = \frac{||\Psi(C_i) - \Psi(C_j)||}{1 + ||x_i - x_j||}$$

weighting covariance distances by inverse spatial distance decreases the influence of visually similar nearby regions

CovSal (Erdem and Erdem, 2013)

• In an image, salient parts can and do appear over a wide range of scales.
• Saliency detection should be carried out simultaneously at multiple scales.

• Employ a fusion strategy to combine single-scale maps to come up with one final saliency map:

$$S(x) = G_o(x) * \prod_{k \in K} \tilde{S}^k(x)$$

Spatial coincidence assumption:
An image part is treated as salient if it is salient at all scales.
Center bias

- Experiments show that there is a tendency in humans to look towards the image center.

Saliency analysis at 5 different scales.

\[ S(x) = G(x) \prod_{k \in K} \hat{S}^k(x) \]

Scale-space extension to the image center and thus signifies the optimal location to initiate a visual search (Judd et al., 2009). It has been shown that adding the motor bias \( S(x) = G(x) \prod_{k \in K} \hat{S}^k(x) \) towards the image center, which is called the covariance descriptor of a region is \( \text{CovSal (Erdem and Erdem, 2013)} \).

The objects that can be treated as salient in an image are those that stand out in the saliency map computed from the actual fixation density map. For this reason, in predicting human eye fixations.

The human fixation maps are not reported here since they are not made public.

Some images from the MIT300 data set and their saliency maps computed by some subjects but rejected by the others. In these images, together with three other models Itti, DVA and SR) even if it is not specifically pointed out that the ROC based evaluation suffers from these sets and their saliency maps predicted by our model are presented in Figure 4.

Fixations for one observer

Fixations from 16 observers

Fixation map

Participation among the surrounding green peppers. Figure 4. (a) Input image. (b–d) Predicted saliency maps obtained at different scales (from the finest to the coarsest). (e) Final saliency map. The single-scale approach is easily extended to operate on multiple scales by employing a fusion strategy to combine these maps to come up with one final saliency map. The single-scale approach is easily extended to operate on multiple scales by employing a fusion strategy to combine these maps to come up with one final saliency map. The single-scale approach is easily extended to operate on multiple scales by employing a fusion strategy to combine these maps to come up with one final saliency map.
Most AUC scores, namely Itti, GBVS, DVA, CSD, and our approach (CovSal), computed saliency map from the actual fixation density map. For this reason, in false alarm rate. In addition, it does not consider the spatial deviation of the saliency, the authors of [Harel et al., 2007] pointed out that the ROC based evaluation su...
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.

<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>0.730</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>Harel &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>1.884</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>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>1.884</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>
</tbody>
</table>

Our approach with Covariances only: 0.767, 0.834, 1.184, 1.342, 3.142, 1.931, 0.408, 0.546
Covariances + means: 0.765, 0.834, 1.198, 1.396, 3.398, 1.896, 0.402, 0.548
Covariances + center: 0.840, 0.840, 1.753, 1.753, 1.901, 1.901, 0.561, 0.561
Covariances + means + center: 0.851, 0.851, 1.891, 1.898, 1.728, 1.728, 0.581, 0.581
Chance: 0.505, 0.803, -0.001, 0.969, -2.401, -2.401, 0.187, 0.479

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

<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>0.730</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>Harel &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>1.884</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>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>1.884</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>
</tbody>
</table>

Our approach with Covariances only: 0.767, 0.834, 1.184, 1.342, 3.142, 1.931, 0.408, 0.546
Covariances + means: 0.765, 0.834, 1.198, 1.396, 3.398, 1.896, 0.402, 0.548
Covariances + center: 0.840, 0.840, 1.753, 1.753, 1.901, 1.901, 0.561, 0.561
Covariances + means + center: 0.851, 0.851, 1.891, 1.898, 1.728, 1.728, 0.581, 0.581
Chance: 0.505, 0.803, -0.001, 0.969, -2.401, -2.401, 0.187, 0.479

MIT 1003 - quantitative results

MIT 1003 - qualitative results

Toronto - quantitative results

MIT 1003 - qualitative results
MIT 300 - quantitative results

<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>3.994</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>3.096</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>0.820</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>3.121</td>
<td>0.390</td>
<td>0.509</td>
</tr>
</tbody>
</table>

Our approach with Covariances + center: 0.800, Covariances + means + center: 0.806, Center: 1.000, Chance: 0.500

<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>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>3.096</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>0.820</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>3.121</td>
<td>0.390</td>
<td>0.509</td>
</tr>
</tbody>
</table>

Our approach with Covariances + center: 0.800, Covariances + means + center: 0.806, Center: 1.000, Chance: 0.500

<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>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>3.096</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>0.820</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>3.121</td>
<td>0.390</td>
<td>0.509</td>
</tr>
</tbody>
</table>

Our approach with Covariances + center: 0.800, Covariances + means + center: 0.806, Center: 1.000, Chance: 0.500

Salient object detection

• Task of identifying foreground objects that attract more attention.

• ImgSal data set [Li et al., 2012]
  • 235 natural color images
  • Six different categories:
    • large salient regions (50 images),
    • intermediate salient regions (80 images),
    • small salient regions (60 images),
    • cluttered backgrounds (15 images),
    • repeating distractors (15 images),
    • large and small salient regions (15 images)
### ReTargetMe data set

aka content aware image resizing

- automatically resizing an image to arbitrary aspect ratios while trying to preserve important content
- **ReTargetMe data set** [Li et al., 2012]
  - 80 images with 92 different resizing scenarios
  - categorized into nine groups:
    - lines/clear edges,
    - symmetry,
    - indoor
    - faces/people,
    - textual elements,
    - geometric structures,
    - recurring texture,
    - outdoor/nature
    - evident foreground objects,

### ImgSal - qualitative results

![Input image](input image)

![Saliency map](saliency map)

![Proto-objects](proto-objects)

![Ground truth labeling](ground truth labeling)

### ImgSal - quantitative results

<table>
<thead>
<tr>
<th></th>
<th>Large salient regions</th>
<th>Intermediate salient regions</th>
<th>Small salient regions</th>
<th>Cluttered backgrounds</th>
<th>Repeating distractors</th>
<th>Large and small salient regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>DSC</td>
<td>AUC</td>
<td>DSC</td>
<td>AUC</td>
<td>DSC</td>
<td>AUC</td>
</tr>
<tr>
<td>Li et al. (1998)</td>
<td>0.897</td>
<td>0.610</td>
<td>0.897</td>
<td>0.473</td>
<td>0.937</td>
<td>0.401</td>
</tr>
<tr>
<td>Harel et al. (2007)</td>
<td>0.945</td>
<td>0.694</td>
<td>0.925</td>
<td>0.529</td>
<td>0.951</td>
<td>0.463</td>
</tr>
<tr>
<td>Torralba et al. (2006)</td>
<td>0.790</td>
<td>0.469</td>
<td>0.825</td>
<td>0.377</td>
<td>0.929</td>
<td>0.372</td>
</tr>
<tr>
<td>Hou &amp; Zhang (2007)</td>
<td>0.833</td>
<td>0.534</td>
<td>0.861</td>
<td>0.448</td>
<td>0.939</td>
<td>0.411</td>
</tr>
<tr>
<td>Zhang et al. (2008)</td>
<td>0.760</td>
<td>0.461</td>
<td>0.813</td>
<td>0.391</td>
<td>0.895</td>
<td>0.386</td>
</tr>
<tr>
<td>Bruce &amp; Tsotsos (2009)</td>
<td>0.798</td>
<td>0.480</td>
<td>0.825</td>
<td>0.383</td>
<td>0.914</td>
<td>0.357</td>
</tr>
<tr>
<td>Seo &amp; Milanfar (2009)</td>
<td>0.842</td>
<td>0.563</td>
<td>0.896</td>
<td>0.474</td>
<td>0.948</td>
<td>0.430</td>
</tr>
<tr>
<td>Goferman et al. (2010)</td>
<td>0.905</td>
<td>0.636</td>
<td>0.950</td>
<td>0.610</td>
<td>0.970</td>
<td>0.553</td>
</tr>
<tr>
<td>Our approach with Covariances only</td>
<td>0.920</td>
<td>0.666</td>
<td>0.928</td>
<td>0.548</td>
<td>0.967</td>
<td>0.470</td>
</tr>
<tr>
<td>Covariances + means</td>
<td>0.866</td>
<td>0.614</td>
<td>0.924</td>
<td>0.584</td>
<td>0.972</td>
<td>0.586</td>
</tr>
<tr>
<td>Covariances + center</td>
<td>0.919</td>
<td>0.681</td>
<td>0.909</td>
<td>0.517</td>
<td>0.919</td>
<td>0.329</td>
</tr>
<tr>
<td>Covariances + means + center</td>
<td>0.865</td>
<td>0.673</td>
<td>0.912</td>
<td>0.580</td>
<td>0.954</td>
<td>0.508</td>
</tr>
</tbody>
</table>

---

**Journal of Vision**
Fig. 7. The authors built a class independent average object map from predictions where those objects are more likely to be found. In maps with relatively small variations compatible with bounding boxes of all classes (of the PASCAL VOC 2008 dataset) were considered as relevant regions. A dataset of 93 images. Their map is crisper than the ones in the authors built a class independent average object map from predictions where those objects are more likely to be found. In maps obtained using the training and Feature selection de Campos et al., CVIU 20

3.2.3. Hybrid saliency maps

All things being equal, eye movements also tend to be directed (the same pixels). However, clear relationships also exist between (ac o m p l e x s c e n e , p e o p l e o v e r w h e l m i n g l yc h o o s et od i r e c tt h e i r decision being (Appropriateness

1.2. INFORMATION FROM GAZE

Itti et al. (1998) 7.82 8.85 6.48 8.45 6.80 8.50 6.13 8.41 8.08 9.29 8.06 6.00 4.67 6.62 7.41 8.10 9.70
Zhong et al. (2008) 9.95 7.87 10.38 8.07 9.10 7.60 8.59 9.87 9.26 9.75 9.67 8.61 11.00 9.95 11.00 10.00 10.00
Seo & Mianfar (2008) 2.84 8.25 9.67 9.00 7.60 7.90 9.75 9.75 8.61 7.95 8.25 8.71 9.65 7.00 11.17 7.59 9.41 7.00 8.20
Golkorman et al. (2012) 0.78 6.80 9.07 5.60 9.30 6.70 8.63 5.75 8.87 7.00 7.12 7.88 4.67 7.33 7.86 5.97 7.60 8.10
Our approach with Covariances only 5.85 5.56 5.81 5.69 6.50 4.40 6.63 6.50 5.29 5.05 5.24 5.82 7.00 6.50 4.59 5.68 5.60 4.60
Covariances + means 9.04 5.00 6.01 6.00 9.90 5.10 7.89 5.39 6.68 6.63 8.24 8.42 8.67 7.00 8.59 8.91 10.50 4.30
Covariances – center 2.84 5.67 5.80 5.52 2.88 4.70 3.26 5.64 2.88 5.75 5.24 5.82 7.00 6.50 4.59 5.68 5.60 4.60
Covariances – means 4.75 5.75 5.75 5.75 4.70 5.80 4.87 4.03 4.76 5.97 4.18 5.53 5.17 4.33 4.33 4.33 4.33 4.33 4.33
Recurrent – center 7.53 3.35 7.93 4.97 3.35 4.80 7.60 4.80 7.45 6.32 7.94 5.00 9.17 2.00 4.37 4.73 8.70 3.60

ReTargetMe - qualitative results

Beyond saliency - feature selection

Aesthetic class prediction, Wong and Low, ICIP 2009

Scene recognition, Fornoni and Caputo, BMVC 2012

Image classification, de Campos et al., CVIU 2012

Object detection, Yun et al., CVPR 2013

ReTargetMe - quantitative results
Beyond saliency - as a feature

Learning saliency,
Judd et al., ICCV 2009, Borji, CVPR 2012

Generic objectness,
Alexe et al., CVPR 2010

Memorability prediction,
Khosla et al., NIPS 2012
Mancas and le Meur, ICIP 2013

Problems with saliency models?

- Important information may not be visually salient (e.g., stop sign in a cluttered scene)
- Salient information may not be important
- Cannot account for many fixations when there is a task

Top-down saliency estimation

- A superpixel-based top-down saliency model via joint discriminative dictionary and CRF learning
- Task: Task-driven such as detecting an object instance from a certain category
**Top-down saliency estimation**

**Training:**
1. Segment the images into superpixels and represent them with the sigma points descriptor.
2. Extract the objectness maps.
3. Jointly learn the dictionary and the CRF parameters for each object category.

**Testing:**
1. Segment the images into superpixels and represent them with the sigma points descriptor.
2. Compute the sparse codes of superpixels with dictionaries learned from data.
3. Estimate the objectness map.
4. Use the CRF model to infer the saliency scores.

---

**CRF and dictionary learning**

- Construct a CRF model with nodes representing the superpixels and edges describing the connections among them.

\[
\log P(Y|X, D, \theta) = \sum_{i \in V} \psi_i(y_i, x_i; D, \theta) + \sum_{i \in V} \gamma_i(y_i, x_i; \theta) + \sum_{(i,j) \in E} \phi_{i,j}(y_i, y_j, x_i, x_j; \theta) - \log Z(\theta, D)
\]

- **Dictionary potential:** Use a sparse codes-based linear classifier as a unary potential.

\[
\psi_i(y_i, x_i; D, \theta) = -y_i \cdot w^T \cdot \alpha_i
\]

\[
\alpha_i(x_i, D) = \arg \min_{\alpha} \frac{1}{2} \|x_i - D\alpha\|^2 + \lambda \|\alpha\|_1
\]
log \( P(Y|X, D, \theta) = \sum_{i \in V} \psi_i(y_i, x_i; D, \theta) + \sum_{i \in V} \gamma_i(y_i, x_i; \theta) \)

\[ \text{dictionary potential} \]

\[ \text{objectness potential} \]

\[ + \sum_{(i,j) \in E} \phi_{i,j}(y_i, y_j, x_i, x_j; \theta) - \log Z(\theta, D) \]

\[ \text{edge potential} \]

- **Objectness potential**: a class-independent unary potential
  \[ \gamma_i(y_i, x_i; \theta) = -\beta y_i (2P(obj|x_i) - 1) \]

**CRF and dictionary learning**

- **Learning**: Simultaneously learn the CRF parameters \( \theta \) and the dictionary \( D \) by optimizing:
  \[ (D^*, \theta^*) = \arg \max_{D, \theta} \prod_{m=1}^{M} P(Y^{(m)}|X^{(m)}, D, \theta) \]

---

**Quantitative analysis**

**EER results on the Graz-02 dataset**

<table>
<thead>
<tr>
<th></th>
<th>Bike</th>
<th>Car</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>Margolin et al. (2013)</td>
<td>25.6</td>
<td>16.9</td>
<td>17.4</td>
</tr>
<tr>
<td>Perazzi et al. (2012)</td>
<td>11.4</td>
<td>13.8</td>
<td>14.3</td>
</tr>
<tr>
<td>Yang and Zhang (2013)</td>
<td>14.8</td>
<td>13.7</td>
<td>14.9</td>
</tr>
<tr>
<td>Objectness (Alexe et al., 2010)</td>
<td>53.5</td>
<td>48.3</td>
<td>43.5</td>
</tr>
<tr>
<td>Aldavert et al. (2010)</td>
<td>71.9</td>
<td>64.9</td>
<td>58.6</td>
</tr>
<tr>
<td>Khan and Tappen (2013)</td>
<td>72.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Marszalek and Schmid (2012)</td>
<td>61.8</td>
<td>53.8</td>
<td>44.1</td>
</tr>
<tr>
<td>Yang and Yang (2012)</td>
<td>62.4</td>
<td>60.0</td>
<td>62.0</td>
</tr>
<tr>
<td>Our approach (setting 1)</td>
<td>71.9</td>
<td>61.9</td>
<td>65.5</td>
</tr>
<tr>
<td>Our approach (setting 2)</td>
<td>71.7</td>
<td>62.0</td>
<td>64.9</td>
</tr>
<tr>
<td>Our approach (setting 3)</td>
<td>73.9</td>
<td>68.4</td>
<td>68.2</td>
</tr>
</tbody>
</table>
Qualitative analysis

Saliency maps on the Graz-02 dataset

Main insights from natural tasks

- Vision is active not passive.
  - Specific information is usually acquired at the fixation point.
  - Information is acquired “just-in-time”.
- Fixations patterns reflect learning at several levels:
  - what objects are relevant
  - where information is located
  - order of sub-tasks/properties of world.
- Fixations tightly linked to actions.

Developments in eye tracking

- Head free:
  - Head mounted IR video-based systems
  - Remote systems with head tracking!
  - Scene camera

Ego Centric Vision a.k.a First person vision

(Focus on Bloom et al., ICCVW 2011)

- Information is acquired “just-in-time”. 
- Fixations tightly linked to actions.

Ego Centric Vision a.k.a First person vision

(Focus on Bloom et al., ICCVW 2011)

- Information is acquired “just-in-time”. 
- Fixations tightly linked to actions.

Ego Centric Vision a.k.a First person vision

(Focus on Bloom et al., ICCVW 2011)

- Information is acquired “just-in-time”. 
- Fixations tightly linked to actions.
dataset used for action recognition. Presented in [10], it is one of the largest and most challenging

One objective of this work is to introduce eye movement recordings for the PASCAL VOC image

3 Action from a Single Image – New Human Eye Movement Dataset

aware of any eye movement models that are learned from eye movement data.

inverse optimal control setting, which allows, in principle, an arbitrary time horizon.

reward, albeit our reward function is learned instead of being pre-specified, and we work in an

some resemblance with these later methods, in that we also aim at maximizing the future expected

visual saliency measures can be used to obtain scanpaths[11] in conjunction with non-maximum

and video[15, 19]. The prediction of eye movements has been less studied. In contrast, predefined

overview). Recently, the trend has been to learn saliency models from fixation data in images[13, 22]

recognition from video appears in our prior work[19].

by the task[5]. A quantitative analysis of task influence on visual search in the context of action

cal properties like the saccade amplitude and the fixation duration have been shown to be influenced

for picture viewing, but these groundbreaking studies have been fundamentally qualitative. Statisti-

learning techniques for saliency modeling and eye movement prediction.

of magnitude larger than the existing image databases. This makes it adequate to using machine

have been collected under free-viewing, and the few task controlled ones[14, 7] have been designed

2 Related Work

• significantly improved estimates. Section

and advanced computer vision descriptors in order to learn task sensitive reward functions based on

trainable, eye movement prediction model. The method combines inverse reinforcement learning

visual attention patterns, both spatial and sequential. Our findings are presented in

Figure 1: Saliency maps obtained from the gaze patterns of 12 viewers under action recognition (left

Mathe and Sminchisescu, NIPS 2013

Shapovalova et al., NIPS 2013

Fathi and Rehg, CVPR 2013

Li et al., ICCV 2013

Figure 1:

§

Fathi and Rehg, CVPR 2013

Li et al., ICCV 2013

63

62