Salience Using Natural Statistics: Comparing Local and Learned Statistics
Matthew H. Tong, Lingyun Zhang, and Garrison W. Cottrell
Department of Computer Science & Engineering, UCSD

**Motivation**

- **Visual Salience** is a measure of what is interesting in the world and therefore captures our attention.
- Visual salience is important because it helps drive a decision we make *five hundred thousand times a day* - where to look.
- We define salience within a probabilistic framework, based on a simple goal of the visual system.
- Our model is very simple, efficient, and state of the art.
- SUN differs from other models in its reliance on *learning and experience*.

**Basic assumptions**

- A key goal of the visual system is to find objects that are potentially important for survival or the task at hand.
- The visual system should direct attention to locations in the visual field that have a high probability of containing such objects.

**Definition of saliency**

Notation:
- \( x \): a point in the visual field
- \( C \): occurrence (presence) of a target at a point \( x \)
- \( F \): visual features at \( x \)
- \( L \): location (pixel or retinal coordinates) of a point \( x \) in the visual field.
- \( s_x \): saliency of point \( x \)

We define the salience of a point as the probability of a target at that point:

\[
x_x = p(C = 1 | F = f_x, L = L_x) = p(F = f_x, L = L_x | C = 1)p(C = 1)
\]

We use a simplification that features and location are independent and conditionally independent given \( C = 1 \):

\[
p(F = f_x, L = L_x | C = 1) = p(F = f_x)p(L = L_x)
\]

With these assumptions, our definition of saliency can be rewritten as:

\[
x_x = \frac{1}{\log p(L = L_x)} p(F = f_x | C = 1)p(C = 1)
\]

Since we care only about the ordering of points by salience, we can apply any monotonic function to saliency; here we take the logarithm.

**Features**

- We compared two sets of features:
  - 12 differences of Gaussians (4 scales, 3 color channels)
  - 362 11x11 ICA features, trained on natural images
- We assume the features are independent for ease of computation and model the distributions with generalized Gaussians.

**Free viewing: Salience is novelty**

- Bottom-up salience is the self-information of the features observed at a point. In the absence of a known target, this term dominates.
- Rare or novel features attract attention
- All probability distributions are learned through experience, and are not dependent on the statistics of the current scene.

**Comparison with AIM**

- AIM is an Attention Information Maximisation (Bruce and Tsotsos, 2006; 2009) that is also based on self-information but uses the statistics of the current image.
- In our 2008 work, SUN and AIM performed comparatively, but used different features.
- Bruce and Tsotsos also suggest AIM can handle the asymmetries we claim pose difficulty for them.
- We can compare predictions using the same feature sets
  - Shown here are two ICA filter sets: a 21x21 pixel set and a 31x31 pixel set, both with their dimensionality reduced with PCA.

**Tilted Bar Asymmetry**

- Our claims about tilted bars holds across filter sets.
- It’s even less clear how AIM could get upside down S’s.

**Qualitative results**

<table>
<thead>
<tr>
<th>Image</th>
<th>SUN(DoG)</th>
<th>SUN(ICA)</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
</tbody>
</table>

**Quantitative results**

<table>
<thead>
<tr>
<th>Model</th>
<th>KL(SE)</th>
<th>ROC(SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUN(DoG)</td>
<td>0.697(0.004)</td>
<td>0.697(0.004)</td>
</tr>
<tr>
<td>SUN(ICA)</td>
<td>0.697(0.004)</td>
<td>0.697(0.004)</td>
</tr>
</tbody>
</table>

**Quantitative results**

- These results use the KL distance and ROC between the human fixations and human fixation for the same point in other images, which underestimates performance but compensates for the central bias and allows fair comparisons.
- Not accounting for the central bias places too much importance to how borders are handled and leads to a central Gaussian ‘outperforming’ most models.

**Search asymmetries**

- Many search asymmetries cannot be explained by systems that get their statistics from the current image!

**Strengths and weakness**

- There are visual search and free viewing results where local image statistics best account for the data.
- Other times, prior experience seems to play a significant role.
  - Example: For this data set, AIM found natural textures like trees and grass salient. SUN was trained on natural images and knows tree limbs and grass are common.

**Further analysis**

- If visual statistics take time to gather, we could see better performance with SUN on early fixations
  - This does not appear to be the case
  - Both achieve best predictive accuracy on the second fixation
- Since both models have examples where they make superior predictions, is it possible to combine the two models?
  - When predicting fixations while free viewing, SUN’s predictions are superior a quarter of the time.
- For our first attempt at combining models (by averging or taking the max of the two maps) the predictions do not improve.

**Future directions**

- Continue examining asymmetries and pop-out.
- Combine local and learned statistics
- Update statistics online
- Use features that capture more local contrast
- Difference of Gaussians over feature responses
- RCA

**Conclusions**

- Many models of salience are based around the idea that we look at what is surprising.
- One definition of surprise is self-information which can be either based on other regions of the current scene or prior experience.
- Both local and learned statistics can account for some of the data.
- Our model needs to be improved by taking into account context and local statistics.