SUN: A Model of Visual Salience Using Natural Statistics

Gary Cottrell
Lingyun Zhang    Matthew Tong
Tim Marks       Honghao Shan
Nick Butko      Javier Movellan
Collaborators

Lingyun Zhang
Matthew H. Tong
Tim Marks
Honghao Shan

QuickTime™ and a TIFF (LZW) decompressor are needed to see this picture.
Collaborators

Nicholas J. Butko

Javier R. Movellan
Visual Salience

- **Visual Salience** is some notion of what is *interesting* in the world - it captures our attention.

- Visual salience is important because it drives a decision we make a *couple of hundred thousand times a day* - where to look.
**Visual Salience**

- **Visual Salience** is some notion of what is *interesting* in the world - it captures our attention.
- But that’s kind of vague...
- The role of Cognitive Science is to make that explicit, by creating a *working model* of visual salience.
- A good way to do that these days is to use probability theory - because as everyone knows, the brain is Bayesian! ;-)

Data We Want to Explain

- Visual search:
  - Search asymmetry: A search for one object among a set of distractors is faster than vice versa.
  - Parallel vs. serial search (and the continuum in between): An item “pops out” of the display no matter how many distractors vs. reaction time increasing with the number of distractors (not emphasized in this talk…)

- Eye movements when viewing images and videos.
Audience participation!

Look for the unique item

Clap when you find it
What just happened?

- This phenomenon is called **the visual search asymmetry**:
  - Tilted bars are more easily found among vertical bars than vice-versa.
  - Backwards “s”’s are more easily found among normal “s”’s than vice-versa.
  - Long line segments are more easily found among short than vice-versa.
  - Unfamiliar race faces are more easily found among familiar than vice-versa.
Why is there an asymmetry?

- There are not too many *computational* explanations:
  - Prototypes do not pop out
  - We have a “race” feature
  - ???
- Our model of visual salience will naturally account for this.
Saliency Maps

- Koch and Ullman, 1985: the brain calculates an explicit saliency map of the visual world
- Their definition of saliency relied on center-surround principles
  - Points in the visual scene are salient if they differ from their neighbors
- In more recent years, there have been a multitude of definitions of saliency
Saliency Maps

- There are a number of candidates for the salience map: probably it is in LIP, the Lateral Intraparietal Sulcus, a region of the parietal lobe… but there may be representations of salience much earlier in the visual pathway - some even suggest in V1.
- But we won’t be talking about the brain today…
Probabilistic Saliency

- **Our basic assumption:**
  - The main goal of the visual system is to find potential targets that are important for survival, such as prey and predators.
  - The visual system should direct attention to locations in the visual field with a high probability of the target class or classes.
  - We will lump all of the potential targets together in one random variable, $T$.
  - For ease of exposition, we will leave out our location random variable, $L$ (See Matt Tong’s poster).
Probabilistic Saliency

- Notation: $x$ denotes a point in the visual field
  - $T_x$: binary variable signifying whether point $x$ belongs to a target class
  - $F_x$: the visual features at point $x$
- The task is to find the point $x$ that maximizes

$$p(T_x | F_x)$$

the probability of a target given the features at point $x$

- This quantity *is* the saliency of a point $x$
- *Note:* This is what most classifiers compute!
Probabilistic Saliency

- Taking the log and applying Bayes’ Rule results in:

\[
\log p(T_x|F_x) = \log \frac{p(F_x|T_x)p(T_x)}{p(F_x)} = \log p(F_x|T_x) + \log p(T_x) + \log \frac{1}{p(F_x)}
\]
Probabilistic Saliency

\[
\log p(F_x | T_x) + \log p(T_x) + \log \frac{1}{p(F_x)}
\]

- \( \log p(F_x | T_x) \)
  - Probabilistic description of the features of the target
  - Provides a form of **top-down** (endogenous, intrinsic) saliency
  - Some similarity to Iconic Search (Rao et al., 1995) and Guided Search (Wolfe, 1989)
\[ \log p(F_x | T_x) + \log p(T_x) + \log \frac{1}{p(F_x)} \]

- \( \log p(T_x) \)
  - Constant over locations for fixed target classes, so we can drop it.
  - Note: this is a stripped-down version of our model, useful for 20 minute presentations ;-
  - We usually include a location variable as well that encodes the prior probability of targets being in particular locations.
Probabilistic Saliency

\[ \log p(F_x | T_x) + \log p(T_x) + \log \frac{1}{p(F_x)} \]

- \(-\log p(F_x)\)
  - This is called the **self-information** of this variable
  - It says that rare **feature values** attract attention
  - Independent of task
  - Provides notion of **bottom-up** (exogenous, extrinsic) saliency
Probabilistic Saliency

Now we have two terms:

- **Top-down** saliency
- **Bottom-up** saliency

Taken together, this is the *pointwise mutual information* between the features and the target.

\[
\log P(F_x | T_x) - \log P(F_x)
\]
Math in Action: Saliency Using “Natural Statistics”

- For most of what I will be telling you about next, we use only the $-\log p(F)$ term, or bottom up salience.
- Remember, this means rare feature values attract attention.
Math in Action: Saliency Using “Natural Statistics”

- Remember, this means rare feature values attract attention.
- This means two things:
  - We need some features (that have values)! What should we use?
  - We need to know when the values are unusual: So we need *experience*. 
Math in Action: Saliency Using “Natural Statistics”

- Experience, in this case, means collecting statistics of how the features respond to natural images.

- We will use two kinds of features:
  - Difference of Gaussians (DOGs)
  - Independent Components Analysis (ICA) derived features
Feature Space 1: Differences of Gaussians

These respond to differences in brightness between the center and the surround.

We apply them to three different color channels separately (intensity, Red-Green and Blue-Yellow) at four scales: 12 features total.
Feature Space 1: Differences of Gaussians

- Now, we run these over Lingyun’s vacation photos, and record how frequently they respond.
Feature Space 2: Independent Components
Learning the Distribution

We fit a generalized Gaussian distribution to the histogram of each feature.

\[ p(F_i; \sigma_i, \theta_i) = \frac{\theta_i}{2\sigma_i \cdot \Gamma\left(\frac{1}{\theta_i}\right)} \exp\left(-\frac{F_i}{\sigma_i} \left| \frac{\theta_i}{\sigma_i} \right|^\theta \right) \]

where \( F_i \) is the \( i^{th} \) filter response,
\( \theta_i \) is the shape parameter and \( \sigma_i \) is the scale parameter.

\[ \log p(F_i) = \text{const.} - \left| \frac{F_i}{\sigma_i} \right|^\theta \]
• This is $P(F)$ for four different features.
• Note these features are *sparse* - I.e., their most frequent response is near 0.
• When there is a big response (positive or negative), it is interesting!
Bottom-up Saliency

- We have to estimate the joint probability from the features.
- If all filter responses are independent:
  \[- \log p(F) = - \sum_i \log p(F_i)\]
- They’re not independent, but we proceed as if they are. (ICA features are “pretty independent”)
- Note: No weighting of features is necessary!
### Qualitative Results: BU Saliency

<table>
<thead>
<tr>
<th>Original Image</th>
<th>Human fixations</th>
<th>DOG Salience</th>
<th>ICA Salience</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="fixations1.png" alt="Fixations 1" /></td>
<td><img src="dog_saliency1.png" alt="DOG Salience 1" /></td>
<td><img src="ica_saliency1.png" alt="ICA Salience 1" /></td>
</tr>
<tr>
<td><img src="image2.png" alt="Image 2" /></td>
<td><img src="fixations2.png" alt="Fixations 2" /></td>
<td><img src="dog_saliency2.png" alt="DOG Salience 2" /></td>
<td><img src="ica_saliency2.png" alt="ICA Salience 2" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Image 3" /></td>
<td><img src="fixations3.png" alt="Fixations 3" /></td>
<td><img src="dog_saliency3.png" alt="DOG Salience 3" /></td>
<td><img src="ica_saliency3.png" alt="ICA Salience 3" /></td>
</tr>
</tbody>
</table>
### Qualitative Results: BU Saliency

<table>
<thead>
<tr>
<th>Original Image</th>
<th>Human fixations</th>
<th>DOG Salience</th>
<th>ICA Salience</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Original Image" /></td>
<td><img src="image2.png" alt="Human fixations" /></td>
<td><img src="image3.png" alt="DOG Salience" /></td>
<td><img src="image4.png" alt="ICA Salience" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Original Image" /></td>
<td><img src="image6.png" alt="Human fixations" /></td>
<td><img src="image7.png" alt="DOG Salience" /></td>
<td><img src="image8.png" alt="ICA Salience" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Original Image" /></td>
<td><img src="image10.png" alt="Human fixations" /></td>
<td><img src="image11.png" alt="DOG Salience" /></td>
<td><img src="image12.png" alt="ICA Salience" /></td>
</tr>
</tbody>
</table>
Qualitative Results: BU Saliency
Quantitative Results: BU Saliency

<table>
<thead>
<tr>
<th>Model</th>
<th>KL(SE)</th>
<th>ROC(SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Itti et al. (1998)</td>
<td>0.1130(0.0011)</td>
<td>0.6146(0.0008)</td>
</tr>
<tr>
<td>Bruce &amp; Tsotsos (2006)</td>
<td>0.2029(0.0017)</td>
<td>0.6727(0.0008)</td>
</tr>
<tr>
<td>Gao &amp; Vasconcelos (2007)</td>
<td>0.1535(0.0016)</td>
<td>0.6395(0.0007)</td>
</tr>
<tr>
<td>SUN (DoG)</td>
<td>0.1723(0.0012)</td>
<td>0.6570(0.0007)</td>
</tr>
<tr>
<td>SUN (ICA)</td>
<td><strong>0.2097(0.0016)</strong></td>
<td>0.6682(0.0008)</td>
</tr>
</tbody>
</table>

- These results use the KL distance and ROC between the human fixations and human fixations for the *same point in other images*, which underestimates performance but allows fair comparisons.

- *Not* doing this makes a Gaussian blob centered on the image outperform every known method except a location prior, because of the center bias of eye fixations (Parkhurst & Neibur 2003; Tatler et al., 2005; Zhang et al., 2008)
Related Work

- Torralba et al. (2003) derives a similar probabilistic account of saliency, but:
  - Starts with a decision of whether the target is or is not present
  - Uses current image’s statistics
  - Emphasizes effects of global features and scene gist

- Bruce and Tsotsos (2006) also use self-information as bottom-up saliency
  - Uses current image’s statistics
Related Work

- The use of the current image’s statistics means:
  - These models follow a very different principle: finds rare feature values \textit{in the current image} instead of \textit{unusual feature values in general: novelty}.
- As we’ll see, novelty helps explain several search asymmetries
- Models using the current image’s statistics are unlikely to be neurally computable in the necessary timeframe, as the system must collect statistics from entire image to calculate local saliency at each point
Our definition of bottom-up saliency leads to a clean explanation of several search asymmetries (Zhang, Tong, and Cottrell, 2007)

- All else being equal, targets with uncommon feature values are easier to find

- Examples:
  - Treisman and Gormican, 1988 - A tilted bar is more easily found among vertical bars than vice versa
  - Levin, 2000 - For Caucasian subjects, finding an African-American face in Caucasian faces is faster due to its relative rarity in our experience (basketball fans who have to identify the players do not show this effect).
Search Asymmetry Results
Top-down salience in Visual Search

- Suppose we actually have a target in mind - e.g., find pictures, or mugs, or people in scenes.
- As I mentioned previously, the original (stripped down) salience model can be implemented as a classifier applied to each point in the image.
- When we include location, we get (after a large number of completely unwarranted assumptions):

\[
\log \text{salience}_x = -\log p(F = f_x) + \log p(F = f_x | T_x = 1) + \log p(T_x = 1 | L = l)
\]

Self-information: Bottom-up saliency
Log likelihood: Top-down knowledge of appearance
Location prior: Top-down knowledge of target's location
Qualitative Results (mug search)

- Where we disagree the most with Torralba et al. (2006)
  - GIST

- SUN
Qualitative Results (picture search)

- Where we disagree the most with Torralba et al. (2006)
  - GIST
  - SUN
Where we agree the most with Torralba et al. (2006)

GIST

SUN
This is an example where SUN and humans make the same mistake due to the similar appearance of TV’s and pictures (the black square in the upper left is a TV!).
Quantitative Results

- Area Under the ROC Curve (AUC) gives basically identical results.
Saliency of Dynamic Scenes

- Created spatiotemporal filters
  - Temporal filters: Difference of exponentials (DoE)
    - Highly active if change
    - If features stay constant, goes to zero response
    - Resembles responses of some neurons (cells in LGN)
  - Easy to compute
  - Convolve with spatial filters to create spatiotemporal filters
Saliency of Dynamic Scenes

- Bayesian Saliency (Itti and Baldi, 2006):
  - Saliency is Bayesian “surprise” (different from self-information)
  - Maintain distribution over a set of models attempting to explain the data, P(M)
  - As new data comes in, calculate saliency of a point as the degree to which it makes you alter your models
    - Total surprise: $S(D, M) = KL(P(M|D); P(M))$
  - Better predictor than standard spatial salience
  - Much more complicated (~500,000 different distributions being modeled) than SUN dynamic saliency (days to run vs. hours or real-time)
Saliency of Dynamic Scenes

- In the process of evaluating and comparing, we discovered how much the center-bias of human fixations was affecting results.
- Most human fixations are towards the center of the screen (Reinagel, 1999)

Accumulated human fixations from three experiments:

- Bruce and Tsotsos (2005)
- Einhauser and Konig (2006)
- Itti and Baldi (2006)
Saliency of Dynamic Scenes

- Results varied widely depending on how edges were handled
  - How is the invalid portion of the convolution handled?

Accumulated saliency of three models:

- Itti and Koch (1998)
- Bruce and Tsotsos (2005)
- Gao and Vasconcelos (2007)
Saliency of Dynamic Scenes

<table>
<thead>
<tr>
<th>Method</th>
<th>KL</th>
<th>ROC area</th>
<th>% above median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.0</td>
<td>0.5</td>
<td>50%</td>
</tr>
<tr>
<td>Bayesian Surprise</td>
<td>0.1332</td>
<td>0.6472</td>
<td>70.91%</td>
</tr>
<tr>
<td>Dynamic Saliency</td>
<td>0.1001</td>
<td>0.6262</td>
<td>70.91%</td>
</tr>
<tr>
<td>Dynamic Saliency (w/border)</td>
<td>0.1815</td>
<td>0.6596</td>
<td>75.37%</td>
</tr>
<tr>
<td>Centered Gaussian</td>
<td>0.4415</td>
<td>0.7641</td>
<td>86.89%</td>
</tr>
</tbody>
</table>

Initial results
Measures of Dynamic Saliency

- Typically, the algorithm is compared to the human fixations within a frame
  - I.e., how salient is the human-fixated point according to the model versus all other points in the frame
  - This measure is subject to the center bias - if the borders are down-weighted, the score goes up
Measures of Dynamic Saliency

- An alternative is to compare the salience of the human-fixated point to the same point across frames
  - Underestimates performance, since often locations are genuinely more salient at all time points (ex. an anchor’s face during a news broadcast)
  - Gives any static measure (e.g., centered-Gaussian) a baseline score of 0.
  - This is equivalent to sampling from the distribution of human fixations, rather than uniformly
  - On this set of measures, we perform comparably with (Itti and Baldi, 2006)
Saliency of Dynamic Scenes

Results using non-center-biased metrics on the human fixation data on videos from Itti(2005) - 4 subjects/movie, 50 movies, ~25 minutes of video.

<table>
<thead>
<tr>
<th>Method</th>
<th>KL</th>
<th>ROC area</th>
<th>% above median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0</td>
<td>0.5</td>
<td>50%</td>
</tr>
<tr>
<td>Bayesian Surprise</td>
<td>0.0344</td>
<td>0.5808</td>
<td>61.66%</td>
</tr>
<tr>
<td>Dynamic Saliency</td>
<td>0.0409</td>
<td>0.5818</td>
<td>62.39%</td>
</tr>
</tbody>
</table>
Demo...
Summary and Conclusions

- It is a good idea to start from first principles.
- Often the simplest model is best.
- Our model of salience rocks.
  - It does bottom up
  - It does top down
  - It does video (fast!)
  - It naturally accounts for search asymmetries
Summary and Conclusions

But, as is usually the case with grad students, Lingyun didn’t do everything I asked...

We are beginning to explore models based on utility: Some targets are more useful than others, depending on the state of the animal.

We are also looking at using our hierarchical ICA model, to get higher-level features.
Summary and Conclusions

- And a foveated retina,
- And updating the salience based on where the model looks (as is actually seen in LIP).
Thanks!!

- To you for listening
- To Thomas for inviting me
- To my students for being smart
- To my funders:
  - NSF (funded the Temporal Dynamics of Learning Center), NIMH (funded me), and the McDonnell Foundation’s grant to the Perceptual Expertise Network