Task-driven Saliency Using Natural Statistics (SUN)
Matthew H. Tong, Christopher Kanan, Lingyun Zhang, and Garrison W. Cottrell
Department of Computer Science & Engineering, UCSD

Motivation
- Mathematical models of saliency are important for understanding human vision, as well as for computer vision applications.
- Many models of human saccades and visual saliency address the implementation, but fail to address the goal of the computation.
- We define saliency within a Bayesian framework, based on the assumption of a simple goal for the visual system.
- Within our approach multiple forms of top-down knowledge are naturally combined with bottom-up saliency.

Definition of saliency
Notation:
- z: a point in the visual field.
- f: the feature values of a point as they are perceptually represented by the visual system.
- L: the location of a point in the visual field.
- s_{z}: the saliency of point z

The probability of interest is:

\[ s_z = p(C=1 | f, L = l_z) \]
\[ = p(F=f | L = l_z, C=1)p(F=f | C=1) \]

We assume simplicity that features and location are independent and conditionally independent given C = 1:

\[ p(F=f | l_z) = p(F=f | l_z, C=1) \]
\[ p(F=f | C=1) = p(F=f | C=1, L = l_z) \]

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

\[ s_z = \frac{p(F=f | C=1, L = l_z)p(F=f | C=1)}{p(F=f | C=1)p(F=f | C=1)} \]

Learning the scales of objects
The scale at which objects appear varies significantly so there is no single optimal scale. However the distribution of scales within each object class is distinct.

We cluster the percent increase in size of the resized features for each training object using one Gaussian mixture model per class with three clusters each.

The three cluster centers found for each class are used to re-scale the filters when computing:

\[ R(C=1 | F=f, t_d) \]

Conclusions
- We hypothesize that one important goal of the visual system is to locate interesting objects.
- Bottom-up saliency emerges naturally from our model as the self-information of the visual features in previous experience.
- SUN’s appearance model for a target, which combines bottom-up saliency with top-down knowledge, emerges from our model as the pointwise mutual information between the visual features and the presence of the target.
- During our testing, appearance proved a comparable predictor to Contextual Guidance, providing guidance even in target-absent conditions.

Quantitative Performances
- Percentage of fixations falling in top 20% most salient region.
- The time at which objects appear varies significantly so there is no single optimal scale. However the distribution of scales within each object class is distinct.

Experimental Design
- Evaluation was performed similarly to (Torralba et al., 2006).
- Training: 329 images with cups, 284 with paintings, and 699 with people from the annotated LabelMe dataset.
- Testing: Eye-tracking data from subjects told to count the number of targets in each image.

Features Used for SUN
- 362 linear features are learned by applying a complete independent component analysis (ICA) algorithm to 11 x 11 patches of color natural images from the Kyoto dataset (Wachter et al., 2007).
- When used this way, ICA permits us to learn the statistical structure of the visual world’s basic features.

Saliency is information
- Bottom-up saliency is the self-information (-log p(F=f)) of the features observed at a point. In the absence of a known target, this term dominates.
- Rare or novel features attract attention.

Learning appearance
- For ease of computation, we combine our bottom-up and appearance terms.
- Recall, this combination is pointwise mutual information.

\[ \log p(F=f | C=1) \]
\[ \log p(F=f | C=1) - \log p(F=f | C=1) \]

Comparing predictions
- The two instances shown were chosen because they represented the greatest disagreement between the Contextual Guidance model and SUN’s appearance model.

Looking at the first two terms, concerning features:

\[ \log p(F=f | L = l_z) \]
\[ \log p(F=f | C=1) \]

\[ \log p(F=f | C=1) - \log p(F=f | C=1) \]

To learn p(C=1 | F=f), we find images from the LabelMe dataset (Russell et al., 2008) that contain the current class of interest, either people, cups, or paintings.

We then applied the learned filters to the set of patches, rescaling the filters to the patch size to get one response from each feature per image patch and normalizing the responses appropriately.

The set of patches is used to train a support vector machine (SVM) modified to give probability estimates (Chin-Chung & Chin-Jan, 2001).