Predicting Visual Saliency and Saccade Probability

Bob Schafer, Boyko Kakaradov, Mindy Chang

Continually throughout the day, the brain must process incoming visual signals and transform them into appropriate actions, namely saccadic eye movements to the most behaviorally salient stimuli. The overarching problem that we address is how to use visual information to predict the parts of a scene that are most salient, and more specifically, to predict the targets of saccadic eye movements.

Predicting the targets of saccades is a difficult problem for several reasons. First, there are many different cognitive influences on the saliency of a visual object. For example, a set of keys sitting on a desk might go overlooked when scanning the scene for a coffee mug, but will immediately draw the gaze of a viewer who has been locked out of the next room. Secondly, the saliency of a target is history- and state-dependent: a nearby object will often be targeted by a saccade over a more salient, but more distant, alternative. Finally, although repeated viewings of the same visual stimuli are necessary for data collection and proper analyses, saliency changes as objects become more familiar.

Background

Because of the difficulty in predicting eye movements, the benchmarks described in the literature do not evaluate the ability to predict the single most likely saccade target at a given point in time. Instead, the metric used to evaluate a model of saccade prediction involves the creation of saliency maps, which are probability distributions over the entire visual scene. Itti et al. evaluated several models including their “surprise metric” by calculating the average saliency of thousands of randomly chosen pixels across all frames of a movie, and then calculating the saliencies of the pixels that were actually chosen as the endpoints of saccades [1]. The model evaluation, hereafter called the Itti metric, is defined as the fraction of saccade endpoints with saliency calculations greater than the average value across the saliency map. Additionally, Itti et al. used the Kullback-Leibler (KL) divergence to describe the difference in the distribution of saccade endpoint saliencies and the randomly chosen pixel saliencies. The best performing model, and the one that we use as a comparison in this study, results in an Itti metric of 72%, and a KL divergence of 0.24.

Research on visual saliency and active vision typically follows one of two approaches: a high-level analysis of objects, context, and the "gist" of a scene [2], or a low-level analysis of luminance, contrast, and local contours [1]. We chose to concentrate on the ability of low-level visual features to predict saliency. The hierarchical feed-forward model of the visual cortex proposed by Serre et al. for object recognition provides a framework for extracting features from a visual scene [3]. The proposed model uses biologically inspired filters that have been previously shown to be similar to the spatio-temporal receptive fields of visual cortical neurons [4,5]. The superior performance of this model on class-specific object recognition tasks suggests that the features it generates may be relevant to saliency prediction.

Many neural structures involved in planning and executing eye movements, including the frontal eye fields (FEF), superior colliculus, pulvinar nucleus of the thalamus, and the lateral intraparietal area, have been characterized as having neural activity that reflects the likelihood with which a saccade will be made to a certain region in visual space. Specifically, the FEF has been implicated as a visual salience map [6]. Activity in this cortical area underlies both covert visual attention and the preparation and execution of saccades. The firing rate of an FEF neuron is believed to describe the saliency of the represented area of the visual field. Thus the neural activity of a population of FEF neurons provides a representation of both the amount of attention allocated to specific spatial regions of a scene and the instantaneous probability of executing a saccade to each part of space as the scene changes over time. Our project includes the analysis of experimental neurophysiological data in addition to image and eye position analyses.

Methods

For our experimental setup, we collected data from 2 macaque monkeys, which viewed an LCD display during the presentation of a grayscale video. The two types of data acquired were eye position data during free viewing and neural data from the FEF during fixation. Five 5-10 second movie clips were taken of office scenes, outdoor foot traffic, and moving vehicles

around the Stanford campus. A single movie, of an office scene with moving people and chairs, was used for the analyses presented here. The movie's resolution was 640x480 pixels and was played at 30 frames/second for a total length of just under 6 seconds. Figure 1 illustrates the two types of data collected and the resulting saliency metrics.

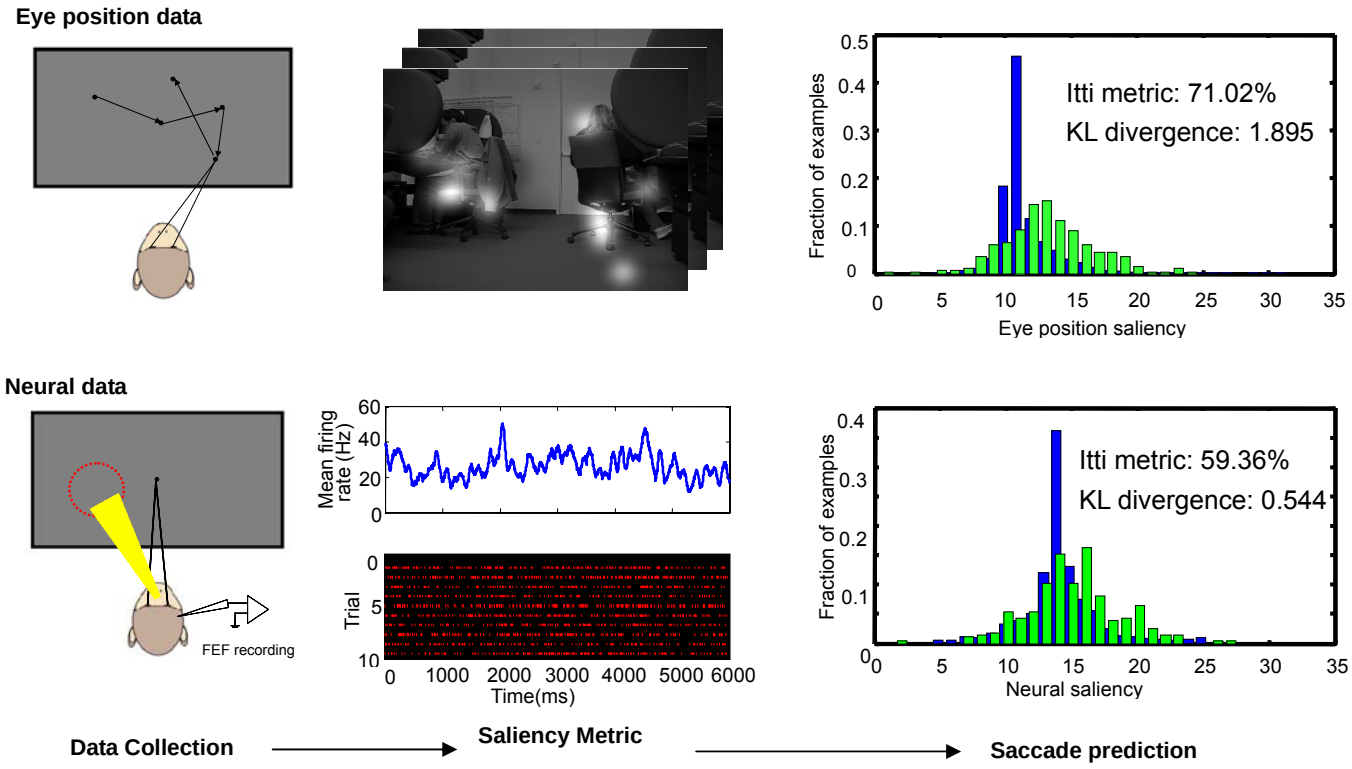

Figure 1: Experimental setup and saliency measures used for saccade prediction.

Eye position data: (left) Gaze targets were tracked as monkeys freely viewed a video. (center) Sample eye position saliency map (right) Distributions of eye position saliency for randomly selected points (blue) and saccade endpoints (green) **Neural data**: (left) The electrical activity of single neurons in FEF were recorded while monkeys fixated on a spot and the video was shown at different positions relative to the visual receptive field of the targeted neuron. (center) Sample raster plot of neuron spiking and corresponding mean firing rate in response to a single location in space throughout the course of the movie (right) Distributions of neural saliency for randomly selected points (blue) and saccade endpoints (green)

Eye Position Data

Eye movement data was collected from two macaque monkeys freely viewing a grayscale movie, which spanned visual angles of approximately 42 x 32°. Precise eye positions were recorded at 200 Hz using the scleral search coil technique. A coil of insulated, biocompatible wire was implanted into one eye of each monkey such that it moved with the eye. During the task, the monkey sat in a magnetic field, and the small currents induced by the changes in coil position were used to determine the direction of the monkey's gaze.

For each frame of the movie, a saliency map was obtained by overlaying the eye trajectories from all trials and representing each gaze target as a small Gaussian (σ = 20 pixels). Gaussians, rather than single points, were used to account for a) any imprecision in the eye position readings, and b) the assumption that direction of gaze often indicates salient objects or regions, rather than simply salient pixels. Because it takes time for the brain to process an image, plan an eye movement, and execute the saccade, the eye position at any time t reflects the movie frame on the screen at approximately t -150 ms. We therefore shifted the eye position maps back in time by 5 movie frames, or 165 ms, so that eye movements would reflect the images that influenced them, rather than the images occurring after their completions. Saccade endpoints were defined as the eye positions during movie frames following drops in eye velocity below a threshold of 20 deg/sec.

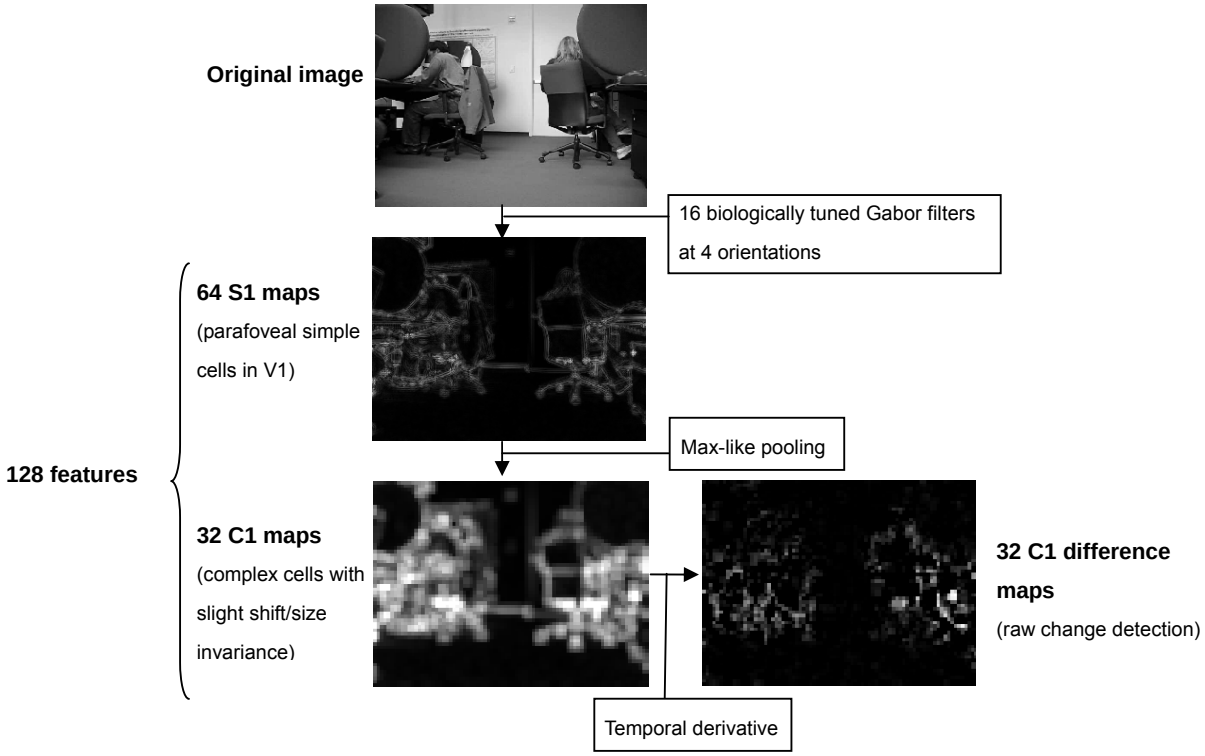

Figure 2: Feature Selection. Sample features from a single frame of the video.

Neural Data

Monkeys were previously set up for chronic neural recording from the FEF in accordance with National Institutes of Health and Society for Neuroscience guidelines. Before each recording, a single tungsten electrode was lowered with a micromanipulator into the cortex until neurons were detected and eye movements could be evoked with 100 ms trains of low (< 50μA) current pulses at 200 Hz. The vector of these evoked saccades defined the site's response field (RF), and determined the part of visual space represented by the neurons that were then isolated and recorded. While the monkey fixated on one spot, the movie was played at different positions on the screen such that different patches of the video fell within the RF of the targeted neuron.

The neural saliency metric of a location of the movie was defined as the average firing rate of the neuron whose RF was centered on that region. We were interested in the neural activity produced by the characteristics of individual frames of the movie clip. Therefore, spike trains were shifted back in time by 100 ms, which is an estimate of the visual latency of neurons in the FEF, and were aligned to the image on the screen at that time.

Feature selection

For feature selection, we used the MATLAB code from Serre et al. available online. Figure 2 shows sample features obtained for a single frame of the movie. The first layer (S1) emulates parafoveal simple cells in V1 via 64 biologically-tuned Gabor filters, paired in 8 size bands (ranging from 7 & 9 pixels to 35 & 37 pixels) each containing 4 orientations (0°, 45°, 90°, and 135°). The second layer (C1), representing complex cells with slight shift size tolerance, generates 32 responses using a max-like pooling operation. The change in C1 response between consecutive frames was also taken as an additional 32 features. Thus for a given pixel, a total of 128 features were extracted by taking the values of the S1, C1, and C1 difference responses corresponding to that pixel.

Results

SVM

As an initial pass at saccade prediction, we treated saccade occurrence as a binary categorical variable. Support vector machine classification was used to distinguish between saccade endpoints and randomly selected non-gaze targets using the 128 features [7]. An equal number of pixels were selected for each of the two classes, and 70% were used for training, while 30% were used for testing. The optimal cost (C) and gamma (γ) parameters for an RBF kernel were found using a grid search across several orders of magnitude to select the parameters yielding the highest 10-fold cross validation value. The overall accuracy was 72.4 ± 5.2%, but only 54% of actual saccades were predicted as saccades. By adjusting the weightings of the error costs to penalize false negatives more heavily (i.e.: penalize actual saccades that were not predicted), we achieved an overall accuracy of 60.3 ± 10.4% but were able to predict 67% of actual saccades.

Regression

Saccades are discrete and therefore a binary categorization between saccade targets and non-targets is accurate, but the underlying saliency of a visual target is continuous. Our data sets provided us with two continuous saliency measures of regions across the movie clip, which allowed us to use regression as a complement to the SVM analysis. The first set of saliency values came from the pixel intensities of the eye position maps for each frame of the movie (Figure 1, top, center). The second set of continuous saliency labels was produced by the neural data: the instantaneous firing rate of an FEF neuron in response to the presentation of a movie frame is thought to indicate the saliency of the part of the image within the cell's RF (Figure 1, bottom, center).

In the first regression, a vector *x* of 128 features was calculated (as described above) for each of 20 evenly-spaced pixels across each frame of the movie clip. The label *y* was the luminance intensity of the corresponding pixel in the eye position saliency map, which was in the interval [0,1]. In the second regression, the training data *x* was confined to the pixels at the centers of the FEF RFs, and the labels *y* were the firing rates, normalized to the interval [0,1]. After learning the feature weights *θ* using each of these two training sets and their respective labels, these weights were used to calculate the predicted saliencies of pixels in two different test sets: a grid of pixels across all frames of the movie clip, and the pixels that were actually endpoints of saccades.

$$\begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_N^{(1)} \\ \vdots & \vdots & & \vdots \\ 1 & x_1^{(m)} & \cdots & x_N^{(m)} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_N \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

where $x^{(i)}$ = S1, C1, C1 difference map values for pixel i

$y^{(i)}$ = saliency value for pixel i

Using the weights learned with the eye position saliency map labels, 71% of saccade endpoint pixels had calculated saliencies greater than the average pixel saliency (Figure 1, top, right). This result was similar to the number published by Itti et al. (72%). The Kullback-Leibler divergence comparing the distribution of saccade endpoint saliencies to the randomly chosen pixel saliencies was 1.90, much greater than Itti's value of 0.24. Using the weights trained on the neural data, 59% of saccade endpoints had greater calculated saliencies than the average pixel, and the Kullback-Leibler divergence was 0.54 (Figure 1, bottom, right).

Discussion

Natural vision is an active process which involves two concurrent, interdependent subprocesses: decoding of visual information, and decision of the most relevant, or salient, parts of the visual world. Our results corroborate and extend recent studies on visual saliency that have explored the sufficiency of biologically plausible low-level visual features to explain what is salient. This suggests that high-level, global image analysis and saliency prediction is not necessary to account for and predict a significant level of visual saliency. The results of our eye position and saccade analyses can be directly compared to the recent

results published by Itti et al. Using various machine learning techniques, we have achieved comparable, and sometimes superior, results in predicting the salient parts of a visual scene. Similarly, because saliency drives the guidance of saccadic eye movements, we are able to predict the endpoints of saccades at least as well, and in some cases better than, previous studies.

We do not suggest that the brain is able to execute the types of machine learning algorithms described here. We as researchers use these algorithms to determine the relationship between image features, neural activity, and eye movements, but these relationships do not necessarily need to be learned from scratch by the brain. Neuroanatomical connections between visual cortex and the FEF are present from birth, and thus the only learning necessary in the brain is a tuning of the weights of the synapses underlying these hard-wired connections. If moving the eyes appropriately to salient objects leads to a reward in one form or another, the learning of synaptic weights might be accomplished through a form of reinforcement learning. Considerable recent work has shown that reinforcement learning does in fact occur in the brain, and that the neurotransmitter dopamine might underlie this process.

Our results bolster the argument that the primate frontal eye field (FEF) serves as a saliency map in the brain, and that neural activity in this area appears to be informed by low-level visual cortex, not purely by cognitive processes, complex objects, and global features. According to several computational models of visual cortex (such as the dynamic routing model), the visual cortical hierarchy receives feedback from a high-lever structure reminiscent of a saliency map. It is known that neurons from FEF synapse directly onto visual cortex in a spatially-specific manner, suggesting feedback from the saliency map onto the incoming feature maps. To our knowledge, no existing model includes this type of feedback in the calculation of visual saliency or prediction of eye movements. Our model can be extended to include feedback from the saliency map onto the feature detectors, modulating the signals on future frames. It is possible that including such biologically-inspired dynamic control over the feature detectors could improve saliency prediction. Additionally, several models of object recognition and eye movements include the importance of high-level visual features, such as geometric shapes, complex objects, and global features. By including additional layers of processing in our model of visual cortex, and by implementing feedback from these layers onto the lower levels, it would be possible to model the influence of these high-level features on the formation of the saliency map.

An alternate approach to predicting visual saliency considers a hidden Markov model (HMM), in which hidden states are the true saliencies of the pixels of an image and the observations are 1) the visual features produced by the image, 2) the neural data, and 3) the eye position data. By using expectation maximization, we could simultaneously determine the optimal state transition probabilities and the mean observations produced by each state. These results would provide further insight into the relationship between saliency and neural activity, eye movements, and visual processing.

References

[1] Itti, L. and Baldi, P. Bayesian Surprise Attracts Human Attention, In: Advances in Neural Information Processing Systems, Vol. 19 (NIPS 2005), pp. 1-8, Cambridge, MA: MIT Press, 2006.

[2] Torralba, A., Oliva, A., Castelhano, M., & Henderson, J.M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological Review, 113, 766-786.

[3] Serre, T., L. Wolf and T. Poggio. Object Recognition with Features Inspired by Visual Cortex. In: Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Press, San Diego, June 2005.

[4] Bell, A.J. and Sejnowski, T.J (1997) The 'Independent Components' of Natural Scenes are Edge Filters. Vision Research. 37(23):3327-38.

[5] van Hateren, J.H. and Ruderman, D.L. (1998) Independent Component Analysis of Natural Image Sequences Yields Spatio-temporal Filters Similar to Simple Cells in Primary Visual Cortex. Proc.R.Soc.Lond. B 265:2315-2320.

[6] Thompson, K.G., Bichot, N.P. (2005) A visual salience map in the primate frontal eye field. Progress in Brain Research. 147.

[7] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm