Saccadic guidance for information gathering in natural scenes

Kirsty McNaught (kirstym@gatsby.ucl.ac.uk), **Maneesh Sahani** (maneesh@gatsby.ucl.ac.uk)
Gatsby Computational Neuroscience Unit, 25 Howland Street London, United Kingdom, W1T 4JG

Abstract

Human vision is foveated, with much higher resolution at the centre of gaze compared to peripheral areas. When viewing a scene, humans move their eyes several times a second to bring different parts of the scene into their foveal vision. This necessitates an active sampling mechanism for optimal parsing of scene content. Previous work has shown natural image statistics have higher contrast in locations that are fixated compared to other image locations. We propose a normative explanation for these observations by calculating the expected information gain associated with a particular fixation based on natural image statistics at different resolutions. We train a model to predict the expected foveal retinal ganglion cell responses for an image patch given an existing observation at peripheral resolution. Our model outputs both a mean prediction and an uncertainty. We predict that patches of the image for which the prediction has high uncertainty offer the most information gain and should therefore be strong contenders for the next fixation. We analyse human gaze data to show that fixated image patches are associated with a higher conditional entropy than a reference ensemble, and fixation durations are positively correlated with conditional entropy (expected surprise).

Keywords: vision; saliency; information; active sampling

Introduction

Several studies have investigated the differences in statistics of parts of images that are chosen for fixations compared to the marginal distribution of natural image statistics. For example it has been shown that at fixation locations, images exhibit higher local contrast and lower local pixel correlations (Reinagel & Zador, 1999) and contain higher frequency edges (Baddeley & Tatler, 2006). Others have shown differences in higher order statistics (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000).

It has also been shown that subjects are more likely to fixate on parts of an image that are considered informative by some metric — either by an subjective rating (e.g. Mackworth & Morandi, 1967) or by a measure of self information (Bruce & Tsotsos, 2006). However, such approaches tend to only account for peripheral drop-off in visual acuity by introducing a global Gaussian blur across the resulting saliency map.

Raj, Geisler, Frazor, and Bovik (2005) propose that an active strategy for a foveated visual system might be to choose fixation locations that are likely to minimise uncertainty about foveal contrasts. In this work we extend the idea to minimise uncertainty about all retinal ganglion cell responses in the fovea, and show that this may provide a normative explanation for why high contrast parts of an image are likely fixation candidates.

Methods

We take natural images from the Van Hateren dataset (Hateren & Schaaf, 1998) and pass them through the retinal model of Bradley, Abrams, and Geisler (2014) at different peripheral eccentricities. At higher eccentricities the retinal ganglion cell responses are at lower resolution and reflect a larger receptive field with a lower frequency difference-of-Gaussians filter.

We train a convolutional neural network to predict higher resolution (foveal) retinal responses from lower resolution (peripheral) responses. The input low resolution patch is padded to avoid edge effects, and the output of the network is a higher resolution image patch with a mean and a variance for each pixel. The network was trained by minimising a negative log likelihood loss based on a Gaussian output.

For gaze data, we use the Doves dataset (Van Der Linde, Rajashekar, Bovik, & Cormack, 2009) which contains fixation data from 29 human observers as they viewed 101 images from the Van Hateren dataset. In order to minimise top-down influences, we consider only the first saccade for each image.

For 'peripheral' samples, we use a retinal model at 3° eccentricity (the average length of a first saccade was 2.5°). We take a 1°x 1° image patch at the peripheral locations, and predict the central 2/3 of the patch at foveal resolution.

Where fixation patches are compared with non-fixation patches, we use a reference ensemble of all fixation locations from other images, including first saccades from all subjects.

Results

We compute the conditional entropy at high resolution for each low-resolution image patch according to our model and compare the average entropies for fixation patches and reference patches for each image. We find that the entropies of fixation locations are on average higher than reference patches ($p < 0.0001$, Wilcoxon two-sided signed rank test).

We then consider the duration of each fixation. If the goal of the saccade is information-gathering, we might expect that more surprising foveal patches would warrant a longer fixation time. The conditional entropy given by our model is a measure of the expected surprise (negative log likelihood) before the saccade is carried out. The actual negative log likelihood of the observed foveal response under this posterior is a measure of the true surprise. We show that fixation duration is positively correlated with predicted entropy ($p < 0.01$) but that after correcting for entropy, observed negative likelihood is not correlated with duration.

Acknowledgements

This work was supported by the Gatsby Charitable Foundation. Thanks to Kevin Li, Eszter Vértes, Aapo Hyvärinen and Ricardo Monti for suggestions and feedback.

References

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis. *Vision research*, *46*(18), 2824–2833.

Bradley, C., Abrams, J., & Geisler, W. S. (2014). Retina-v1 model of detectability across the visual field. *Journal of vision*, *14*(12), 22–22.

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In *Advances in neural information processing systems* (pp. 155–162).

Hateren, J. H. v., & Schaaf, A. v. d. (1998, Mar). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings: Biological Sciences*, *265*(1394), 359-366.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial vision*, *13*(2), 201–214.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & psychophysics*, *2*(11), 547–552.

Raj, R., Geisler, W. S., Frazor, R. A., & Bovik, A. C. (2005). Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy. *JOSA A*, *22*(10), 2039–2049.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, *10*(4), 341–350.

Van Der Linde, I., Rajashekar, U., Bovik, A. C., & Cormack, L. K. (2009). Doves: a database of visual eye movements. *Spatial vision*, *22*(2), 161–177.