FIXATION SELECTION BY MAXIMIZATION OF TEXURE AND CONTRAST INFORMATION

Raghu G. Raj, Alan C. Bovik, and Lawrence K. Cormack

Center for Perceptual Systems, University of Texas at Austin, Austin, Texas, 78712, USA

ABSTRACT

We present information-theoretic underpinnings of a computation theory of low-level visual fixations in natural images. In continuation of our prior work on optimal contrast-based fixations [1], we develop an optimum texture-based fixation selection algorithm based on a recent theory of non-stationarity measurement in natural images [2]. Thereafter we propose a simple coupling of the optimal texture-based and contrast-based fixation features to produce a new algorithm called CONTEXT, which exhibits robust performance for fixation selection in natural images. The performance of the fixation algorithms are evaluated for natural images by comparison to randomized fixation strategies via actual human fixations performed on the images. The fixation patterns obtained outperform randomized, GAFFE-based [3], and Itti [4] fixation strategies in terms of matching human fixation patterns. These results also demonstrate the important role that contrast and textural information play in low-level visual processes in the Human Visual System (HVS).

Index Terms— Fixation Selection, Non-stationarity, Natural Scene Statistics, Textures, Contrast, MICA

1. INTRODUCTION

The bewildering complexity of natural scenes is rivaled only by the amazing ability of the Human Visual System (HVS) to comprehend it. Comprehension, from an operational point of view, entails, in part, the systematic analysis and integration of different types of visual information at various levels of processing performed by the HVS—from low-level vision (corresponding to the 'front-end' of the HVS) to high-level visual processing (i.e. the 'back-end' processing of HVS)—and, of course, the subsequent utilization of the resulting knowledge to yield intelligent behavior. From an image processing point of view, it seems very reasonable that understanding of the workings of this complex system should also involve understanding of the nature of the information that the HVS is 'designed to process' at various levels of abstraction from low- to high- level processing. This point of view of course makes the tacit assumption that the HVS is optimized in some way to process visual information.

Attneave [5] and Barlow [6] hypothesized back in the 1950's that information theory can provide a link between environmental statistics and the properties of neural responses, in that the retina and other stages of the early visual system have evolved to develop efficient codes (i.e. in the least number of bits) for the information processed at the respective stages (given biological constraints at each stage such as the available number of neurons etc). Verifying the hypothesis entails not only the discovery of rich Natural Scene Statistics (NSS) models but also establishing precise quantitative relationships to neural coding procedures that purportedly optimize certain aspects of NSS. Doing so would precisely establish the nature of the duality between NSS and low-level HVS processes.

Given the scope and generality of this hypothesis, various modified and restricted versions of this 'efficient coding hypothesis' have been proposed and verified by researchers. More recently, work in the above two-fold research program of developing powerful theoretical models for NSS coupled with investigations into their implications for information processing in the HVS have greatly advanced.

In this paper we, for the first time, explicitly propose and verify a Barlow-type hypothesis for fixation selection in natural images. Our general hypothesis is that low-level visual fixations performed by the HVS in natural scenes are driven by the goal of maximally extracting visual information from the scene. Specifically, we verify this hypothesis for the case of textural and contrast information. In continuation of our prior work on optimal contrast-based fixations [1], we develop, in Section 2, an optimum texture-based fixation strategy based on our recent theory of non-stationarity detection in natural images [2]. These two strands of work give us visual fixation patterns that optimally extract, respectively, contrast and textural information from natural scenes. We propose a simple coupling of these two fixation schemes and evaluate the performance of the resultant algorithm, named CONTEXT, in Section 3, by means of comparison to randomized fixation

Supported by NSF Grant 0427372

strategies via actual human fixations performed on the images. We find that the fixation patterns thus obtained outperform randomized and state-of-the-art fixation selection strategies [3-4] in terms of matching human fixation patterns.

2. TEXTURE-BASED FIXATION FEATURES

We define texture as a 'roughly stationary' spatial process where the degree of non-stationarity decreases with increasing spatial scale of analysis [2]. The structure of natural images is the result of complicated non-linear interactions of such texture elements, where the non-linearities can be induced by occlusions, boundaries, spatial transients, and other phenomena. While contrast is a highly local image property, texture is a regional concept—requiring probabilistic descriptions on multi-dimensional spaces. However, these non-linearities usually induce non-stationarities containing considerable information about the structure of the image. Therefore we may pose that visual fixations that seek to extract textural information from natural images should be driven by image non-stationarities.

Clearly, if there are no significant non-stationarities present in an image, then it may be considered as a single texture, and so, performing multiple fixations will yield little textural information beyond the parameters of the texture model. Moreover, since statistical texture models generally assume that texture samples are drawn from stationary processes, recognizing stationary image regions is an important aspect of image information gathering. Texture-based segmentation is an obvious example of this [2]. Towards this end we have proposed a quantitative measure of non-stationarity called the *Natural Image Non-stationarity Index* (NANS Index), which we now describe, along with modifications towards developing a texture-based fixation-finding strategy.

A spatial random field is *stationary* if, for an arbitrary window, the joint distribution of the random variables associated with the window remains invariant with respect to translation across spatial coordinates. The size of the window defines the *scale* of image analysis [2].

Consider the case wherein the *NxN* non-stationarity analysis window consists of two non-overlapping regions that partition the window—one called the *center patch* and the other, the *surround patch*. This could consist of concentric circular and ring-shaped regions, for example, or square approximations to them. When such a geometry is used the non-stationarity measurement is called a center-surround or CS NANS Index, to distinguish it from indices computed using other geometries, such as side-by-side patches [2]. The center-surround window is then centered at every image coordinate (pixel) allowing computation of the CS NANS Index at every coordinate.

In order to measure non-stationarity, probability distributions must be associated with the center and surround patches. In [2, 7] we showed how the joint probability measures can be naturally defined via an *Multilinear ICA (MICA)* decomposition [7] of the center and surround patches. Given this the central idea of our *theoretical non-stationarity index* is to gauge the relative change of mutual information between center and surround patches [2]. Let *p* and *q* be probability densities associated with the center and surround patches respectively; and $\{p_i\}_{i=1}^{d}$ and $\{q_i\}_{i=1}^{d}$ be marginal distributions corresponding to the best ICA approximation of *p* and *q*. Then the relative change in mutual information between center and surround patches is:

$$\eta = \left| D\left(p \,\|\, \prod_i p_i \right) - D\left(q \,\|\, \prod_i q_i \right) \right| \Big/ \left| D\left(p \,\|\, \prod_i p_i \right) \right| \tag{1}$$

Where *H(p)* is the entropy of p and where:

$$D\left(q \,\|\, \prod_i q_i \right) - D\left(p \,\|\, \prod_i p_i \right) = \left[H(p) - H(q) \right] + \sum_i \left[H(q_i) - H(p_i) \right]$$
$$= \Delta H(p;q) + \Delta H\left[(q_i);(p_i) \right]$$

Whereas $\Delta H[(p_i),(q_i)]$ captures the entropy difference between the corresponding (MICA) filter responses of the center and surround patches, $\Delta H(p;q)$ measures the overall entropy change between the center and surround patches.

It turns out, however, that the numerical implementation of theoretical non-stationarity index in (1) is computationally very prohibitive and is thus, at present, impractical to deploy for fixation selection purposes. Nevertheless under special circumstances, it turns out that the numerator and denominator of (1) assume the form of a linear combination of correlations of filtered responses (of the center and surround patches respectively) between the different MICA channels [2]. The problem therefore reduces to find the optimal coefficients of the above correlations. To this end we effectively determine the weighting coefficients by computing the relative change in MSE when coding the two patches with respect to the MICA filters of the center patch—this results in a non-stationarity index, though sub-optimal, that has the form consisting of a linear combination of correlations as correlations as described above. It is this practical CS-NANS index that we deploy for our texture-based fixation selection algorithm below.

The above NANS Index can be implemented using ICA filters, or other similar decomposition (since MICA can be generalized with respect to arbitrary basis functions [7]). However the direct computation of even the practical MICA-based CS-NANS index suffers from considerable computational complexity, since the MICA filters must be computed from every patch. Therefore, for the problem at hand, we have developed a sub-optimal approach based on computation of relative coding distortion with respect to the $d<M^2$ dominant filters [2] from a *MxM* bank of Gabor filters (where *M<<N*). It turns out that the resulting non-stationarity index has the same form as (1) i.e. linear combination of correlations of filtered responses. Throughout this paper we employ this Gabor-based CS NANS index to quantify the non-stationarity structure of

images. As an example, Fig. 1(*a*) shows a natural image containing two primary substances: grass and water. The corresponding non-stationarity map is shown in Fig. 1(*b*).

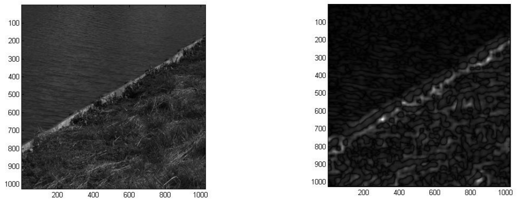
We now formulate a greedy algorithm for determining the optimum non-stationarity-based fixations, which can be stated in terms of the following simple rule: *The next optimum fixation point is simply the point in the image corresponding to the maximum non-stationarity.*

3. SIMULATIONS: THE CONTEXT ALGORITHM

Having developed optimal contrast and texture based fixations above, the question is what is the best way to combine these visual cues to yield optimal performance. A natural approach would be to formulate a joint optimization problem for extracting contrast and textural information. However, we use a simpler approach for cue combination which is that of a simple alternation of contrast and texture based fixation patterns. As it turns out, this simple strategy, which we call the *CONTEXT algorithm*, performs remarkably well in modeling human fixations that in many cases outperforms both the contrast- and texture-based fixations performed separately. Furthermore we point out that since the HVS operates on contrast images, in the simulations we performed the non-stationarity measurement on the corresponding contrast image.

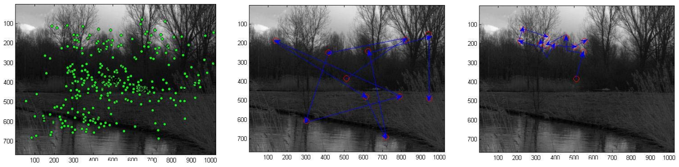
The experimental set up used to generate the fixation points is consistent with that described in [8] for the gathering of the human fixation patterns wherein the pixel resolution is 1 arc-minute per pixel. In all the simulations, N_{fix} = 10 fixation points were generated which were then compared to the aggregate of all human fixation patterns performed for the image under consideration. Comparison between two sets of fixation points was achieved by firstly forming probability maps by dropping Gaussians of a certain with a each fixation point, combining them by means of a *max* operator and finally normalizing to obtain a probability map. The probability maps were then compared by means of both Average and Harmonic Mean KLS measures (which are both forms of symmetric KLD). In order to get a complete picture of the performance of the various algorithm, Gaussians of one-, two-, and three- foveal width were employed.

Figure 2 shows a representative example of the fixation performance of CONTEXT, the GAFFE algorithm [3] together with actual human fixations recorded for the image. Figures 3-8 show the performance of the various fixation strategies (including both texture and contrast performed separately) for a larger collection of images from the DOVES database [8] of natural image fixations. For benchmarking we also evaluated the performance of randomized fixation strategies. In true randomized fixations the fixation coordinates were chosen randomly. In HVS-random fixations, human fixations from another randomly selected image were placed on the current image. From the results we can see that the CONTEXT algorithm overall outperforms both randomized and state-of-the-art fixation strategies.



(*a*) (*b*)

Fig. 1. Non-stationary analysis of a natural



(a) (b) (c)

Fig. 2. Comparison of texture-contrast with human fixations on image #245. (*a*) Human fixations; (*b*) CONTEXT fixations; (c) GAFFE fixations.

These results point towards the eventual construction of a unified information-theoretic understanding of low-level visual fixation processes in the HVS, which in turn can yield insights into the deeper questions of visual understanding of natural scenes.

4. REFERENCES

[1] R.G. Raj, W.S. Geisler, R.A. Frazor, and A.C. Bovik, "Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy , " *J. Opt Soc Amer A,* vol. 22, pp. 2039-2049, Oct 2005.

[2] R.G. Raj and A.C. Bovik, "Non-stationarity measurement in natural images," *IEEE Trans Image Process*, November 2007, submitted.

[3] U. Rajashekar, I. van der Linde, A.C. Bovik, and L.K. Cormack, "GAFFE: A gaze-attentive fixation finding engine," *IEEE Trans Image Processing*, to appear, 2008.

[4] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE PAMI,* vol. 20, pp. 1254-1259, 1998.

[5] F. Attneave, "Some informational aspects of visual perception," *Psy. Review*, vol. 61, pp. 183–93, 1954.

[6] H.B. Barlow, "Possible principles underlying the transformation of sensory messages," *Sensory Comm.* WA Rosenblith, pp. 217–34. MIT Press, 1961.

[7] R.G. Raj and A.C. Bovik, "MICA: A multilinear ICA decomposition for natural image modeling," *IEEE Trans Image Process*, to appear, 2008 (vol. 17, issue 3).

[8] I. van der Linde, U. Rajashekhar, A.C. Bovik and L.K. Cormack, *DOVES: A Database of Visual Eye Movements,* July 2007. Available: http://live.ece.utexas.edu/research/doves.

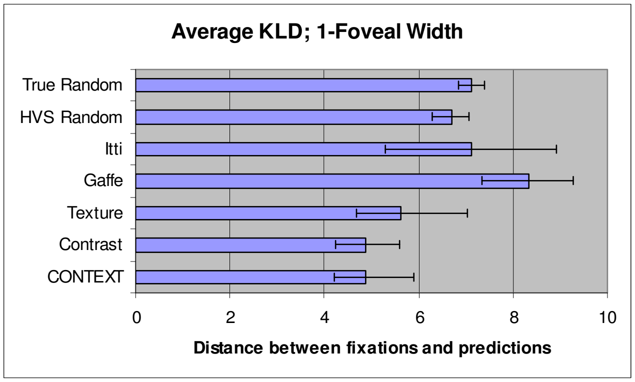


Fig. 3. Average KLD, 1-Foveal Width. Error bars indicate standard deviations.

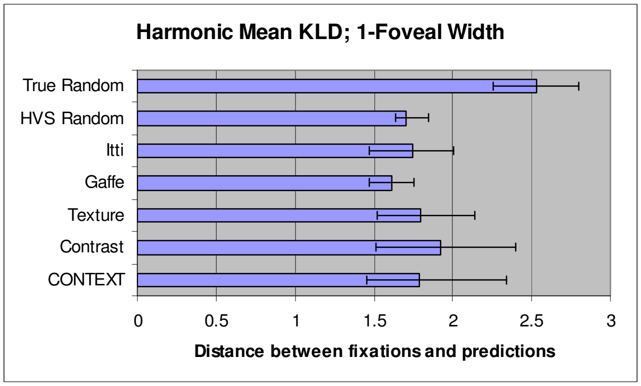


Fig. 4. Harmonic Mean KLD, 1-Foveal Width. Error bars indicate standard deviations.

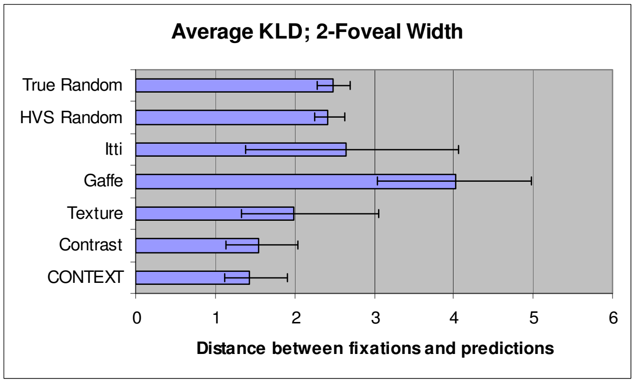


Fig. 5. Average KLD, 2-Foveal Width. Error bars indicate standard deviations.

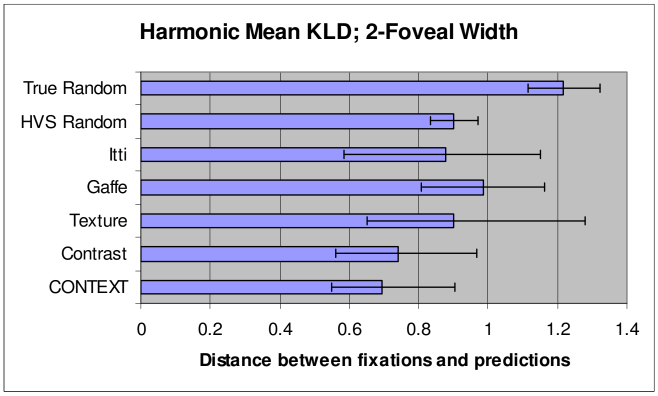


Fig. 6. Harmonic Mean KLD, 2-Foveal Width. Error bars indicate standard deviations

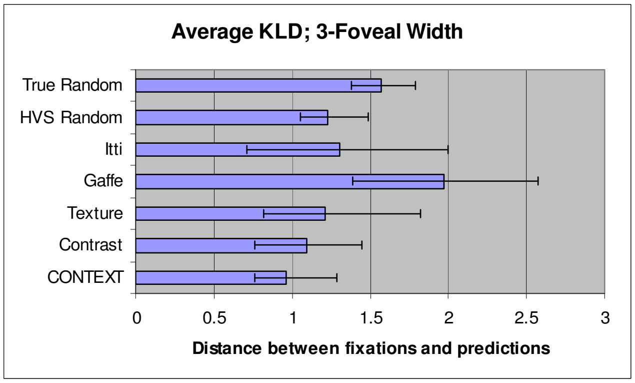


Fig. 7. Average KLD, 3-Foveal Width. Error bars indicate standard deviations

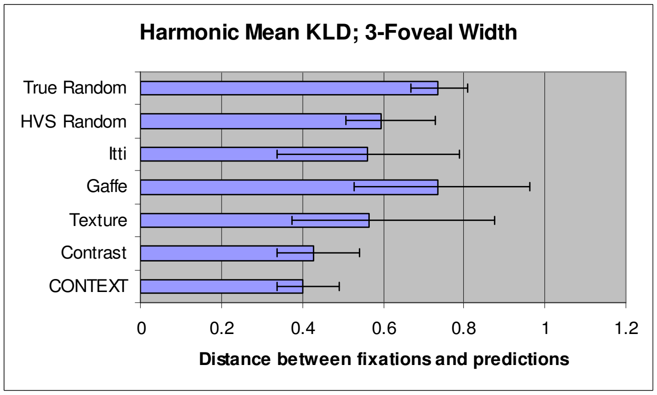


Fig. 8. Harmonic Mean KLD, 3-Foveal Width. Error bars indicate standard deviations