Foveated Analysis of Image Features at Fixations

Umesh Rajashekar [a], Ian van der Linde [a,b], Alan C. Bovik [a], Lawrence K. Cormack [a]

[a] *Center for Perceptual Systems, The University of Texas at Austin, USA*

[b] *Department of Computing, Anglia Ruskin University, Bishops Hall Lane, Chelmsford, Essex, CM1 1SQ, England*

Abstract

Analysis of the statistics of image features at observers' gaze can provide insights into the mechanisms of fixation selection in humans. Using a foveated analysis framework, in which image patches were analyzed at the resolution corresponding to their eccentricity from the prior fixation, we studied the statistics of four low-level local image features: luminance, RMS contrast, and bandpass outputs of both luminance and contrast, and discovered that the image patches around human fixations had, on average, higher values of each of these features at all eccentricities than the image patches selected at random. Bandpass contrast showed the greatest difference between human and random fixations, followed by bandpass luminance, RMS contrast, and luminance. An eccentricity-based analysis showed that shorter saccades were more likely to land on patches with higher values of these features. Compared to a full-resolution analysis, foveation produced an increased difference between human and random patch ensembles for contrast and its higher-order statistics.

Key words: Eye tracking, Foveation, Natural image statistics, Contrast, Point-of-gaze statistics
PACS:

Email addresses: umesh@cns.nyu.edu (Umesh Rajashekar), ianvdl@ece.utexas.edu (Ian van der Linde), bovik@ece.utexas.edu (Alan C. Bovik), cormack@psy.utexas.edu (Lawrence K. Cormack).

1 Introduction

By using a variable resolution sampling of the visual field, the human visual system has evolved to an efficient imaging system that allows for a wide field of view without the accompanying data glut. The resolution is highest at the center (fovea) and drops rapidly towards the periphery (Wandell, 1995). The human eyes interact actively with the visual environment to gather information efficiently from this multi-resolution visual input. The human visual system uses a combination of steady eye fixations linked by rapid ballistic eye movements called saccades (Yarbus, 1967). While the degradation of spatial resolution in the retina has been modeled accurately by measuring the contrast thresholds of transient stimuli (Banks, Sekuler, and Anderson, 1991; Geisler and Perry, 1998) and used in several applications (Geisler and Perry, 1998; Lee, Pattichis, and Bovik, 2001; Wang, Lu, and Bovik, 2003), the problem of selecting fixations in foveated systems is still an open research problem.

Despite the seemingly complex mechanisms that underly the process of active vision, human observers excel at visual tasks. Based simply on our own daily experience, the process of gathering visual information at the current fixation while simultaneously attending to the variable resolution visual periphery in search for potentially interesting regions seems effortless. Thus, an understanding of how the human visual system selects and sequences image regions for scrutiny is not only important to better understand biological vision, it is also the fundamental component of any foveated, active artificial vision system.

The human visual system has conceivably evolved multiple mechanisms for controlling gaze. The interplay of high-level cognitive and low-level image features influences eye movements in many intricate ways and makes the problem of modeling gaze a formidable task. Theories for automatic gaze selection can be broadly classified into top-down and bottom-up categories. Top-down approaches emphasize a high-level, cognitive or semantic understanding of the scene. Bottom-up approaches assume that eye movements are quasi-random and strongly influenced by low-level image features such as contrast and edge density. Given the rapidity and sheer volume of saccades during search tasks (over 15,000 each hour), it is also reasonable to suppose that there is a significant random component to selecting fixation locations. Thus, highlighting differences in the statistical properties of image features between observers' fixations and random fixations is a useful step towards gaze modeling.

Approaches supporting the bottom-up theory propose a computational model for gaze selection that is based on image processing techniques to accentuate features that are deemed visually relevant (Privitera and Stark, 2000), or derived from a biologically-inspired model of visual attention (Itti, Koch, and Niebur, 1998). The general framework of these approaches is to first high-

light several image primitives such as color, intensity, and orientation. Each of these features is then analyzed at various spatial scales to produce a saliency map. Fixations are deployed to regions in decreasing order of saliency with inhibition-of-return to discourage visiting previously fixated areas. Recent versions of this model also account for temporal flicker, and motion energy as motion primitives (Itti, 2004). Incorporating high-level contextual information into these low-level saliency-based models have been reasonably successful in emulating human fixation patterns in object detection tasks (Torralba, 2003; Hamker, 2005).

A recent version of the bottom-up approach of gaze modeling, is based on computing natural scene statistics directly at the *point of gaze* of observers, and extracting image features that are significant at these locations. The availability of relatively inexpensive, accurate eye trackers has made this approach feasible. In one reported work (Reinagel and Zador, 1999), the statistics of natural images at point of gaze were compared to the statistics of patches selected randomly from the same image sets. The results show that the regions around human fixations have higher spatial contrast and spatial entropy than the corresponding random fixation regions, indicating that the human eye may be trying to select image regions that help maximize the information content transmitted to the visual cortex by minimizing the redundancy in the image representation. In particular, the authors note that the RMS contrast at the point of gaze was on average 1.17 times the contrast obtained from an image shuffled set of patches. Even when the size of the patch around fixations was varied, local image contrast was found to be reliably higher (statistically significant) than those obtained from patches at random fixations, with a maximum difference occurring around patch sizes of 1° (Parkhurst and Niebur, 2003). These contrast results were also replicated by others (Mack, Castelhano, Henderson, and Oliva, 2003).

While these gaze-contingent measurements provide useful insight into visual features that are relevant for understanding and modeling gaze, the ensemble of image patches around fixations in the above mentioned studies were analyzed at maximum resolution (of the stimulus). Owing to the foveated nature of our visual system, image features that draw fixations are not encoded at full-resolution, but instead are extracted from the visual periphery whose resolution varies across the visual field. Parkhurst (Parkhurst, Law, and Niebur, 2002) tried to account for this by incorporating a variable resolution function in the model and discovered an improved correlation between points of high saliency and recorded fixations. However, in their work, the foveated structure was imposed on the extracted feature maps and not on the image stimulus. More recently, gaze contingent filtering in video sequences was found to provide improved model-predicted salience for some features such as orientation and flicker (Itti, 2006).

In this paper, we incorporate several enhancements to existing frameworks for gaze-contingent analysis of low-level image features at visual fixations. We recorded the eye movements of 29 observers as they viewed 101 calibrated natural images, and then attempted to quantify the differences in the statistics of image patches at observers' fixations and those selected at random. In addition to extending previously studied image features to include local patch luminance, RMS contrast, bandpass luminance, and bandpass contrast at full-resolution, we also incorporated a foveated analysis of these image features by first foveating the image at a fixation point, and analyzing the image patch at the subsequent fixation from this foveated stimulus. The foveated analysis is shown to differ significantly from the full-resolution analysis for contrast-related features. We use models of foveation that are more accurate for capturing the degradation of spatial resolution in human observers. This foveated model is applied directly to the image stimulus before any feature is extracted. A direct consequence of using a foveated analysis framework is the need to group patches of similar blur before computing any kind of statistics on them. We address this by using an eccentricity-based analysis that highlights the relevance of each of these features as a function of eccentricity from the previous fixation point. This study uses a much larger collection of human observers, a large, carefully selected dataset of high resolution natural calibrated images, accurate models of foveation, and eye movements that are recorded with very high spatial and temporal resolution than previously reported studies.

The rest of this paper is organized as follows. In Section 2, the experimental setup, data collection procedure, the image database, and the visual tasks are presented. As mentioned earlier, all image features are measured after taking into account the variable resolution periphery. The process of filtering the image based on an observer's current fixation is described here. The results of evaluating local image luminance, contrast, and bandpass statistics at human and randomly selected fixations are described in Section 3. The influence of each of these features as a function of saccade magnitude is presented. Finally, a discussion of how these results compare with prior gaze-contingent work, and some extensions of this work are discussed in Section 4.

2 Methods

2.1 Observers

A total of 29 adult human volunteers (19 male and 10 female) participated in this study. All observers either had normal or corrected-to-normal vision. Observers consisted of members of the public, undergraduates, graduates stu-

dents, research fellows, and faculty from the University of Texas at Austin from a range of academic disciplines. Each observer visited for a single session, only 2 had seen the image stimuli previously; 24 were naïve as to the purpose of the experiment.

2.2 Natural Image Stimuli

101 static images of size $1024 * 768$ pixels (cropped from the center of the original $1536 * 1024$ images) were manually selected from a calibrated grayscale natural image database (van Hateren and van der Schaaf, 1998); images containing man-made structures and features such as animals, faces, and other items of high-level semantic interest that could have instinctively attracted attention were omitted. Images whose luminance statistics suggested saturation of the capture device, and thus exhibited non-linearity, were also omitted. Typical images are shown in Fig. 1.

The stimuli were displayed on a 21-inch, gamma corrected monitor at a distance of 134cm from the observer. The screen resolution was set at $1024 * 768$ pixels, corresponding to about 1 arc minute per pixel (or 60 pixels per degree of visual angle). The total spatial extent of the display was thus about $17^\circ \times 13^\circ$ of visual angle. The MATLAB psychophysics toolbox (Pelli, 1997; Brainard, 1997) was used for stimulus presentation. Since the range of brightness varied drastically across the image database, each image was scaled so that the brightest point in each image corresponded to the brightest output level of the monitor without affecting the image contrast.

Before displaying each stimulus image, a Gaussian noise image was displayed to help suppress after-images corresponding to the previous image that may otherwise have attracted fixations. Each image was displayed for 5 seconds. The ambient illumination in the experiment room was kept constant for all observers, with a minimum of 5 minutes luminance adaptation provided while the eye-tracker was calibrated.

2.3 Visual Task

Observers were instructed to free view each of the 101 images. To discourage observers from fixating at only one location in the image and to insure a somewhat similar cognitive state across observers, they were given a simple memory task: following the display of each image, observers were shown a small image patch ($1^\circ * 1^\circ$) and asked to indicate (via a numeric keypad) whether the image patch was from the image they just viewed or not. Auditory feedback (via a sampled voice) was provided to indicate a correct or incorrect

response. Before starting the main experiment, each observer went through a training session of 10 trials to ensure that the observer became familiar with the handheld control box, dark adapted, and comfortable in the experimental environment prior to data collection. Images for the practice session were selected from the same database as the images used for the experiment proper.

2.4 Eye Tracking

As the observers viewed the scene, their eye movements were recorded using an SRI Generation V Dual Purkinje eye tracker. It has an accuracy of $< 10'$ of arc, precision of about $1'$ of arc, a response time of under 1 ms, and bandwidth of DC to $>$ 400Hz. The output of the eye tracker (horizontal and vertical eye position signals) was sampled at 200 Hz by a National Instruments data acquisition board in a Pentium IV host computer, where the data were stored for offline data analysis.

Monocular eye tracking was used to reduce calibration time. A bite bar and forehead rest were used to restrict the subject's head movement. The subject was first positioned in the eye tracker and a system lock established onto the subject's eye. A linear interpolation on a $3 * 3$ calibration grid was then done to establish the transformation between the output voltages of the eye tracker and the position of the subject's gaze on the computer display. The calibration also accounted for crosstalk between the horizontal and vertical voltage measurements.

This calibration routine was repeated compulsorily every 10 images, and a calibration test run after every image. This was achieved by requiring that the observer fixate for 500ms within a 5 s time limit on a central square region ($0.3^\circ \times 0.3^\circ$) prior to progressing to the next image in the stimulus collection. If the calibration had drifted, the observer would be unable to satisfy this test, and the full calibration procedure was re-run. The average number of calibrations per observer for the 101 images was 16.5, i.e. between 6 and 7 images were typically viewed before the calibration test was failed. Average calibration error for passed calibration tests was 5.48 pixels horizontally and vertically. The requirement for a central fixation prior to displaying the next image also ensured that all observers commenced viewing the image stimuli from the same location.

The average duration for the experiment was approximately 1 hour, including the initial calibration. Observers who became uncomfortable during the experiment were allowed to take a break of any duration they desired. Post-experimental debriefing revealed that most observers rated the eye-tracker as only mildly uncomfortable. Plotting the mean performance of the observers

over time in the patch detection task does not suggest a prevailing fatigue factor, with performance slope remaining constant throughout (Fig. 2).

2.5 Image Data Acquisition

The sampled voltages corresponding to the eye movements of the observers for each trial were converted to gaze coordinates (i.e. position of gaze on the image in pixels). Next, the path of the subject's gaze was divided into fixations and the intervening saccadic eye movements using spatio-temporal criteria derived from the known dynamic properties of human saccadic eye movements. Stated simply, a sequence of eye position recordings was considered to constitute a fixation if the recorded gaze co-ordinates remained within a stimulus diameter of 1° visual angle for at least 100ms. The exact algorithm (adapted from (ASL, 1998)) accommodates for drifts, blinks and micro-saccadic eye movements. The resulting pattern of fixations for a single trial is shown by the dots in Fig. 3. The lines show the eye movement trajectories linking the fixations. The first fixation is indicated by a square in the center of the image.

We then extracted circular patches of diameters $32, 64, 96, 160, 192$ pixels centered at each fixation. This corresponded to patches of diameter ranging from 0.5° to 3.2°. Image patches around fixations that extended beyond the boundary of the image were discarded. If we simply extracted patches around each fixation, fewer smaller sized patches would be discarded at image boundaries than larger sized patches. Figure 4 shows a plot of the percentage of fixations that were used in the analysis as a function of patch size. We decided to use 192 pixels (3.2°) as the maximum patch diameter because it provided a trade-off between a fairly large patch while still retaining 94% of all the recorded fixations. Finally, we ensured that the number of image patches analyzed under each of the patch sizes was always the same by first creating a bank of image patches of diameter 192 pixels, and extracting image patches of smaller diameter from the center of this image set.

2.6 The Image-Shuffled Database

The ensemble of the image patches around the fixation points was then analyzed to determine if it contained any features that had statistically significant differences from an ensemble of image patches that were picked randomly. The ensemble of randomly selected patches was obtained by replacing the fixations of an observer for a particular image with those of a different image. The image shuffled database therefore simulates a human observer who is not driven by the image features of that particular image, but otherwise satisfies all criteria of human eye movement statistics. Further, this methodology of simulating

random fixations accounts for both known potential biases of human eye movements (such as the tendency of observers to fixate at the image center, and the log-normal distribution of saccade magnitudes), and unknown biases (such as possible correlations between magnitude and the angle of the saccades). Tatler *et. al* (Tatler, Baddeley, and Gilchrist, 2005) provide a discussion of how such biases might influence the statistics of image features.

Other researchers (Reinagel and Zador, 1999; Parkhurst and Niebur, 2003; Mack et al., 2003) have also simulated a random observer by uniformly distributing fixations in an image, and extracting local image patches around these fixations. However, since we propose to foveate images at the current fixation, n, and then extract image statistics at the subsequent fixation point, $n+1$, it is important that image patches within each ensemble (random and human-selected) be blurred to similar extents overall. To illustrate this point, Fig. 5 shows the distribution of saccade magnitudes (distance between fixations n and $n+1$) for both human fixations (solid line) and by uniformly distributing fixations (dashed line) in the image. Unlike the plot for human saccade magnitudes which peaks at around 1.5°, the saccade magnitude plot for the uniformly distributed fixations peaks at a higher value of 7°. Since the low-pass filtering applied to a patch is proportional to the magnitude of the saccade leading to a fixation upon that patch, all the image patches in the database obtained using the uniformly distributed fixations would be blurred more than the image shuffled database, which will bias the final results. For this reason, all comparisons with the random observer in this paper correspond to this image shuffled database, and not to the uniformly distributed fixations.

2.7 Foveation

An important contribution of this paper is the foveated analysis of the low-level features of image patches at the resolution at which they were actually encoded. To achieve this, the given image was first foveated at the observer's current fixation before the patch at the subsequent fixation was extracted for analysis. There are many ways of creating a foveated version of an image given a fixation point such as band-pass filtering (Lee, Pattichis, and Bovik, 2001), DCT-domain (Sheikh, Evans, and Bovik, 2003), and multi-resolution (Chang and Yap, 1997; Geisler and Perry, 1998), approaches. Since neither speed nor storage was an issue for our offline analysis, we used the spatially-varying bandpass filtering approach, where every pixel in the foveated image was obtained by blurring its grayscale value with a low pass filter of appropriate cut-off frequency (obtained from models of the contrast sensitivity function). The contrast sensitivity (CSF) measured as function of spatial frequency, f (cycles per degree), and retinal eccentricity, e (degrees), is modeled (Geisler

and Perry, 1998) as:

$$CSF(f, e) = C_0 \exp\left(-\alpha f \frac{e + e_2}{e_2}\right) \quad (1)$$

where C_0 , $\alpha = 0.106$, and $e_2 = 2.3^0$ are constants that provide an adequate fit to experimentally recorded contrast threshold values. Since we are mainly interested in retaining the relative magnitudes of the sinusoidal amplitudes, the value of C_0 is not relevant, and is set to 1.0.

The CSF can be considered to be a two dimensional transfer function that can be used to blur an image at various eccentricities, e. For implementation purposes, we compute the 2D discrete fourier transform of the image, pad it appropriately, and perform a point-wise multiplication with the CSF described above (for every possible eccentricity), perform an inverse Fourier Transform, and crop out the valid image area. Ideally, we will have to create enough blurred versions of the original image to account for the largest possible eccentricity (length of the image diagonal). In our analysis, we found that the maximum saccade magnitude seldom exceeded 12 degrees (720 pixels). Therefore, we created 720 blurred versions of each image in increments of one pixel (arc minute).

To create an image that is foveated around a fixation point, we simply needed to find the appropriately blurred version of every pixel. We begin by computing the eccentricity of a pixel location (in the foveated image) from the fixation point, select the blurred image corresponding to this eccentricity (from the 720 blurred versions), and select the grayscale value from the exact same location as the pixel from this blurred image. The process was then repeated for every pixel to obtain the corresponding foveated image. Figure 6 shows the original image at full-resolution and a foveated version of the same image with the fixation point indicated by the white dot.

2.8 Eccentricity-based analysis of image statistics

One of the direct consequences of evaluating statistics of foveated patches is that, if care is not taken, image patches that have been blurred to different extents will be grouped and analyzed together, thus possibly resulting in erroneous statistics due to this cross-resolution comparison of image patches. In our analysis, since the patches around human fixations and those in the image shuffled ensemble have the exact same saccade magnitudes (and hence the same average blur), the statistics of image features will be underestimated just as often as they will be overestimated, and will not bias these statistics significantly.

However, in order to alleviate the effect of comparing statistics of patches that were blurred to different extents, we performed an eccentricity-based analysis where patches of similar blur were grouped together and analysis of the relevant statistic was repeated for each group (saccade bin). This eccentricity-based analysis also provides insight into the influence of the location (and hence resolution) of the image feature in the visual periphery on influencing saccades. Cross-resolution issues do not arise for the full-resolution analysis because all patches have the same resolution as the image.

To perform the eccentricity-based analysis of our image features, each patch in the database was first associated with the eccentricity of the saccade magnitude that was executed to get to that particular patch - i.e. the eccentricity of the fixation point from the previous fixation. (The first fixation was ignored for this analysis.) Figure 5 shows the histograms of saccade magnitudes of all observers (solid line), and all images in this experiment. These saccade magnitudes were then partitioned into 5 bins such that each bin contained the same number of patches. The vertical lines in Fig. 5 show the boundaries of the 5 bins that was used for the analysis. We decided to use 5 bins to achieve a trade-off between the number of patches per bin and the total number of bins. After binning, the number of image patches per bin was found to be around $6,000$. A uniform binning of the saccade magnitudes is not recommended since the number of patches within each bin in a uniformly size saccade bin would have been very different.

2.9 Bootstrapping

To evaluate the statistical significance of the image feature under consideration, we used bootstrapping to obtain the sampling distribution of the mean as follows. For each bootstrap trial, the ensemble of image patches at the observer's fixations (and from the image shuffled fixations) for each image was sampled with replacement. The feature of interest was computed for these patches, and averaged across the 101 images in the database. This process was repeated 200 times to obtain the sampling distribution of the average image feature. The error bars in all the figures in this paper correspond to a 95% confidence interval obtained using this bootstrap procedure.

3 Results

We now present the result of evaluating four local image features: the patch luminance, root-mean-squared contrast, bandpass luminance, and bandpass contrast on image patches centered at human fixations and patches from the

image shuffled database.

3.1 Luminance Statistics

To verify if luminance is a feature that significantly influences fixations, we computed the average luminance of patches at human fixations and compared them to the luminance of image patches from the image shuffled database. The average luminance for each image patch was computed using a circular raised cosine weighting function (Raj, Geisler, Frazor, and Bovik, 2005). The raised cosine function w is expressed as:

$$w(i) = 0.5 * \left[\cos\left(\frac{\pi r_i}{R}\right) + 1\right] \tag{2}$$

where $r_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}$ is the radial distance of a pixel location (x_i, y_i) from the center of the patch, (x_c, y_c), and R is the patch radius. The mean luminance for a given image patch weighted using the raised cosine window was computed as:

$$\bar{I} = \frac{1}{\sum_{i=1}^{M} w_i} \sum_{i=1}^{M} I_i w_i \tag{3}$$

where M is the number of pixels in the patch, I_i is the grayscale value of pixel at location (x_i, y_i).

The absolute values of the patch luminance will, of course, depend on the database on images and will also vary across images. Since we were mainly interested in the differences between the image statistics at observers' fixation and randomly selected fixations, and not the absolute values, we simply computed the ratio of average patch luminance at the observers' fixations to the average patch luminance for image patches from the image shuffled database for each image, and then averaged this ratio across the $N(= 101)$ images in the database. Further, using the saccade binning idea from Section 2.8, we computed the average luminance ratios of image patches within each saccade bin as follows.

$$\overline{I}_{ratio}(e) = \frac{1}{N} \sum_{n=1}^{N} \frac{\overline{I}_{pog}(e, n)}{\overline{I}_{rand}(e, n)} \tag{4}$$

where $\overline{I}_{pog}(e, n)$ and $\overline{I}_{rand}(e, n)$ correspond to average luminance for the patches around observers' fixations and the image shuffled database respectively for image number n. The eccentricity of the patch with respect to the prior fixation is denoted by e.

This eccentricity-based analysis of the luminance ratio is plotted as a function of saccade magnitude in Fig. 7. The error bars correspond to 95% confidence intervals on the mean obtained via bootstrapping. Each panel corresponds to the patch size (indicated by the title) that was used for the analysis. From this plot, we see that the ratios take values that are consistently well above 1.0 indicating that saccades executed by observers tend to land on regions with higher luminance. However, given that the maximum ratio is only around 1.04, the effect does not seem to be very strong and diminishes with increasing patch size. The tendency to fixate at brighter regions is more pronounced at lower saccade magnitudes and with smaller patch sizes. Finally, since foveation does not affect the mean patch luminance, there is no statistically significant difference between the foveated (dashed) and full-resolution (solid) curves.

3.2 Local Contrast Statistics

We now discuss the statistics of another low-level image feature, the local image contrast at fixations. Similar to the luminance computation, the contrast of image patches around observers' fixations was compared to the contrast of image patches from the image-shuffled database. For each image patch, a weighted root-mean-squared (RMS) contrast using a circular raised cosine weighting function (2) was computed as follows:

$$C = \sqrt{\frac{1}{\sum_{i=1}^{M} w_i} \sum_{i=1}^{M} w_i \frac{(I_i - \bar{I})^2}{(\bar{I})^2}} \tag{5}$$

where M is the number of pixels in the patch, I_i is the grayscale value of pixel at location (x_i, y_i) and $\bar{I}$ is the mean of the patch (3).

We then computed the ratio of average RMS contrast at the observers' fixations to the average RMS contrast for image patches from the image shuffled database for each image, and then averaged this ratio across the $N = 101$ images in the database as follows:

$$\overline{C}_{ratio}(e) = \frac{1}{N} \sum_{n=1}^{N} \frac{\overline{C}_{pog}(e, n)}{\overline{C}_{rand}(e, n)} \tag{6}$$

where $\overline{C}_{pog}(e, n)$ and $\overline{C}_{rand}(e, n)$ correspond to average RMS contrast for all patches at eccentricity e around observers' fixations and the image shuffled database respectively, for image number n.

The eccentricity-based RMS contrast ratios for various patch sizes are shown in Fig. 8. From these plots, we note that the curves for both the full-resolution

and foveated analysis are significantly higher than 1.0 for all eccentricity values (and patch sizes) indicating that observers select patches that are of higher contrast than the random observer. The higher values of the contrast ratios (within any patch size) suggest that the effect of contrast is stronger than local patch luminance. Second, the RMS ratio for the foveated analysis are significantly higher than those for the full-resolution analysis for all patch sizes that were used in this analysis. This is evident in Fig. 9, where the contrast ratios (averaged across eccentricity bins) are plotted as a function of the patch size that was used to compute contrast. While the full-resolution analysis is in agreement with previous reported results on the RMS contrast statistics (Reinagel and Zador (1999)), *it underestimates the influence of RMS contrast in drawing fixations*. In Fig. 9, for example, the foveated statistics at a patch size of 1° shows that the average contrast at human fixations is 1.1 times the contrast at random fixations - a result which is significantly different from the value of 1.07 that is obtained from the full-resolution analysis. Finally, as in the case of luminance ratios, increasing the patch size reduces the ratios towards 1.0.

3.3 Bandpass Luminance Statistics

Thus far, we have found that both the mean luminance and RMS contrast are significantly higher for image patches at human fixations than those obtained from the image-shuffled database, with RMS contrast having a stronger effect. The next image feature that we investigated was the output of center-surround-like filters on these image patches. The motivation for using this feature is based on the intuition that it is not necessarily regions of higher luminance or contrast, but regions that differ from their surroundings that will draw fixations. For example, in Fig. 10, it is very likely that human observers will fixate on the central region in both the images. The central square, despite having lower luminance (in the left image) and lower contrast (in the right image) than any other region in the image, draws attention because it differs from its surroundings. Such features can be detected by the outputs of center-surround or more generally, bandpass kernels.

Given a collection of image patches at human and randomly selected fixations, one could resort to a brute force approach and vary various parameters of a bandpass kernel such as its size (full width at half-max) and shape (as defined by aspect ratio) to find the bandpass filter whose outputs are maximally different when applied to the two sets of image patches. In other words, the optimal bandpass filter is the one that selects the bands of spatial frequencies whose energies differ maximally between human and random patches. Thus, instead of the brute force approach described earlier, we used a simple alternative where the filter was designed in the spatial frequency domain.

To locate the spatial frequencies that are most relevant for separating the two patch ensembles in a given image n, we first compute the ratio of the average discrete fourier transform (DFT) magnitudes of patches at point of gaze to that of the patches selected randomly:

$$F_{ratio}(e,n) = \frac{\frac{1}{P(e,n)} \sum_{p=1}^{P(i,e)} |\tilde{I}(e,p)_{pog}|}{\frac{1}{R(e,n)} \sum_{r=1}^{R(e,n)} |\tilde{I}(e,r)_{rand}|} \tag{7}$$

where $\tilde{I}(e,\cdot)_{pog}$ and $\tilde{I}(e,\cdot)_{rand}$ are the DFTs of an image patch at eccentricity, e, at point-of-gaze and random fixations respectively. $P(e,n)$ and $R(e,n)$ correspond to the number of image patches at human and random fixations respectively in image n at eccentricity e. The average value of this ratio across the $N = 101$ images was then computed to yield: $\overline{F}_{ratio}(e) = (1/N) \sum_{n=1}^{N} F_{ratio}(e,n)$.

Figure 11 shows the plots of $\overline{F}_{ratio}(e)$ for a patch size of $1.6^\circ \times 1.6^\circ$. Each panel corresponds to a ratio of centered discrete fourier transforms at a particular eccentricity bin (indicated by the title). The top row shows plots of these ratios for the full-resolution analysis, and the bottom row for the foveated analysis. The white circle in each panel of Fig. 11 corresponds to the maximum visible spatial frequency at a given eccentricity as derived from (1). We notice that while the full-resolution analysis selects several spatial frequencies beyond what the observer can see (outside the white circle), the foveated analysis, by design, ignores spatial frequencies beyond the cut off frequency limit. The white circle does not always fit the resulting $F_{ratio}(e)$ plots snugly for the foveated analysis, because each saccade bin represents several eccentricities, and the mean eccentricity within each bin was used to select the cut off frequency.

Since we are looking at ratios of magnitudes of DFTs of patches selected by human fixations to those from the image-shuffled fixation, the selection of the optimal bandpass kernel amounts to selecting the spatial frequencies that are significantly greater than 1.0. One could, for example, simply select the spatial frequency corresponding to the maximum ratio within each eccentricity bin in Fig. 11. We selected those spatial frequencies whose ratios were greater than 0.98 times the maximum ratio value at any particular eccentricity. This allows for the selection of a band of frequencies instead of a single spatial frequency as the relevant bandpass kernel.

Once the bandpass kernels were identified, the final step involved computing the average energy at these spatial frequencies in the image patches as the feature of interest. In particular, given an image patch $I(p,e)$, located at an eccentricity e from the previous fixation, we computed the energy of the patch only at the relevant spatial frequencies as highlighted by $F_{ratio}(e)$. The ratio

of the average value of this energy for human fixations to that of the patches from the image-shuffled random fixations was computed for each image, and averaged across the N images.

The resulting ratios of bandpass luminance for a patch size of $1.6^\circ \times 1.6^\circ$ is shown in Fig. 12 as a function of saccade magnitude. Due to computational issues, the bandpass analysis was performed for only this patch size. We notice that the values of the ratios are higher (with a maximum value of 1.25) than those obtained by evaluating just the luminance values of the local image patches (a maximum ratio value of 1.04). The average luminance and bandpass luminance ratio value across all saccades for this patch size was 1.01 and 1.16 respectively. In other words, observers are more likely to fixate on image regions that have a bandpass luminance profile than patches that are brighter on average. Further, we also notice that there is no statistical significance between the foveated and the full-resolution analysis. Since foveation does not alter the mean luminance of a region, the bandpass luminance statistic - a measure of difference between the means of two regions - is not affected significantly by foveation.

3.4 Bandpass Contrast Statistics

Finally, we extend the analysis to the center-surround (or bandpass) statistics of local image contrast. Intuitively, this is a measure of the contrast of local image contrast. The motivation for computing this statistic is that it captures higher order structure that are missed by the other three features. For example, as in Fig. 10 (right), the human eye might land on regions whose central and surrounding regions have the same mean luminance (and hence not captured by the bandpass luminance kernels), but different contrast profiles. One way to evaluate this feature is by computing the difference between the local image contrast in a central region of an image patch and the local image contrast in an area surrounding that image patch. The problem of computing the optimal bandpass kernel is more involved than before because we first have to compute local image contrast - which itself depends on the size of neighborhood used to compute the contrast - and then optimize the size of the bandpass kernel that maximally separates human and random patches in the sense of this particular statistic. To address this issue, we compute the magnitude of the local image gradient for each pixel and use this as a measure of an extremely local (pixel-level) measure of image contrast. The goal of designing the optimal contrast bandpass kernel now amounts to determining the spatial scales at which these local image gradients vary. Similar to the approach used in Section 3.3, we computed the average DFT magnitudes of the gradient patches at point-of-gaze to those at random fixations and detected the significant spatial frequencies.

Having located the significant spatial frequencies, we repeated the process of computing the energy of each patch within the relevant spatial frequency bands as before (Section 3.3) except that the filtering was applied to the local patch gradient instead of the patch itself. Figure 13 shows the value of this ratio as a function of saccade magnitude for a patch size of $1.6^\circ \times 1.6^\circ$. The ratio values for this feature is the highest of all the ratios we have computed thus far (with a maximum of 1.3 and average of 1.2). We also notice that the foveated analysis produces ratios that are statistically higher than the full-resolution analysis (for three of the five saccade bins).

4 Discussion

Analysis of the statistics of image features at observers' gaze can provide insights into the mechanisms of fixation selection in humans. Using a foveated analysis framework, in which features were analyzed at the spatial resolution at which they were encoded, we studied the statistics of four low-level local image features: luminance, RMS contrast, bandpass outputs of luminance and contrast, and discovered that the image patches around human fixations had, on average, higher values of each of these features than the image patches selected at random. Second, by examining the actual values of the ratios (Fig. 14), we found that bandpass contrast showed the greatest difference between human and random fixations (maximum ratio of 1.3), followed by bandpass luminance (1.25), contrast (1.12), and luminance (1.04). The results are consistent with the intuition that it is not necessarily local luminance or contrast, but rather the variation in the features with respect to its surroundings that seem to be stronger around human fixations. In fact, automatic fixation selection algorithms have incorporated center-surround mechanisms in their design to capture luminance variations (Itti and Koch, 2001). Finally, if we interpret bandpass contrast as a feature that highlights regions whose texture is different from the surrounding textures, the high value produced by the feature could be attributed to top-down mechanisms such as observers fixating on objects (that are distinct from their surroundings) in the real world scenes; randomly-selected fixations on the other hand are less likely to land often on such objects in the scene.

Incorporating foveation into saliency models for video sequences has been reported to improve the predictive ability of some features such as flicker and orientation, but not luminance contrast (Itti, 2006). The study suggests that contrast measures on grayscale static images might not be influenced by lower simulation realism. However, in this study, we found that contrast measures such as local RMS contrast and bandpass contrast are indeed affected by foveation. One possible reason for this contradiction in findings could be that motion is a very strong cue and dominates contrast cues. It is also possible that

the measure of contrast used in that study was the output of center-surround filters. As mentioned earlier (and seen in Fig. 12), we do not expect contrast measures defined by center-surround mechanisms to be affected significantly since foveation does not alter the mean luminance of the center and surround patch.

Tatler *et al.* (Tatler, Baddeley, and Vincent, 2006) have observed that the influence of image features is not uniform across saccade magnitudes and note that by ignoring this dependence, prior work in this area (Reinagel and Zador, 1999; Parkhurst and Niebur, 2003; Tatler, Baddeley, and Gilchrist, 2005) tends to estimate the influence of visual features incorrectly. They also study the influence of various image features as a function of spatial frequency. Itti (Itti, 2005) performed a similar eccentricity analysis of low-level features in drawing fixation in video sequences and found that short and long saccades had increased saliency at human fixations. In our study, since we use a foveated analysis framework, we analyze patches at the spatial frequency at which they were processed by the human visual system, and so incorporate both saccade and spatial frequency dependence into our analysis. Our results agree with the findings that short saccades are more image feature dependent than long saccades. Long range saccades, in the case of foveated analysis, land of patches that are blurred so strongly, that eventually the ratios should tend to 1.0. Using increasingly larger patch sizes reduces the ratios towards 1.0 (as seen for luminance and RMS contrast) suggesting that the effect of the image features might be local around the fixation point. Increasing the patch size can also results in a greater overlap between image content at human and random fixations, thus creating smaller differences in feature values in the two image ensembles.

These eccentricity-dependent statistics could be incorporated into the design of an automatic Bayesian foveated fixation algorithms that uses low-level image features to guide fixations as follows. Given a novel scene, the algorithm would begin by selecting the first fixation point at the center of image and foveating the scene around this point. Using local image features and the empirical distribution of foveated image features that actually drew fixations at various eccentricities, the algorithm would then create a saliency map that captures the probability that a region in the image will be the next to attract the fixation can be created. A greedy algorithm can be used to select the peak in the resulting saliency map as the next fixation point. The image will then be foveated at the new fixation point and the process repeated. As more image features that differentiate human and random fixations are discovered, the selection mechanism would be able to pool the saliency maps from each of these feature layers and select peaks from this combined saliency map. While it has been shown that luminance and contrast in natural scenes are statistically independent of each other (Mante, Frazor, Bonin, Geisler, and Carandini, 2005), further analysis is needed before we can consider the other features to con-

tribute independently to the saliency maps. Finally, in addition to image features, the model will also incorporate the statistics of human eye movements, such as the distributions of saccade magnitudes and saccade orientation, inhibition of return, and the tendency of observers to fixate at the image center. Initial simulations using bandpass contrast as the relevant feature for fixation selection highlights regions that correlate well with fixations recorded from observers Rajashekar et al (2007). A Matlab implementation of this algorithm will be made available at `http://live.ece.utexas.edu/research/gaffe` shortly. Our analysis for bandpass contrast in this paper quite simplistic. Instead of using patch gradients for contrast measures, one could analyze the local image contrast at a given patch size, and evaluate the frequency variations at that patch size to design a better bandpass contrast kernel. We also note that the bandpass kernels could also be designed by bootstrapping the DFT ratios, and selecting those spatial frequencies that are statistically different from 1.0. The resulting band of frequencies could also be modeled as Gabor filters if necessary. In this analysis, we use a single patch size across all eccentricities. It would be interesting to vary the patch size to match the size of receptive fields at various eccentricities and recompute the scene statistics.

In conclusion, analysis of the statistics of image features at point-of-gaze must incorporate foveation to facilitate a better understanding of their impact on gaze prediction. We use a foveated framework for analyzing the statistics of image patches at the resolution at which they were perceived and showed that foveated analysis reveals an influence of RMS contrast and bandpass contrast that is statistically more significant than that obtained using the full-resolution analysis. The large number of subjects and images, the high accuracy of recorded eye movements, and the careful selection of natural calibrated images in this experiment makes this dataset a very useful tool to evaluate the influences of other low-level fixation attractors in still images. In the near future, as a service to the vision community, we will be providing free access to the entire collection of eye movements. The accompanying manuscript, DOVES: A Database of Visual Eye Movements, is currently under review. In addition to evaluating the statistics of disparity and motion primitives at observers' fixations, our group is also looking into information-theoretic approaches to selecting visual fixations (Raj et al., 2005).

Acknowledgements
This research was supported by a grant from the National Science Foundation (ITR-0427372) and by ECS-0225451.

References

ASL, 1998. Applied science laboratories, eye tracking system instruction manual, ver 1.2.

Banks, M. S., Sekuler, A. B., Anderson, S. J., Nov. 1991. Peripheral spatial vision: limits imposed by optics, photoreceptors, and receptor pooling. J Opt Soc Am A 8 (11), 1775–1787.

Brainard, D., 1997. The psychophysics toolbox. Spatial Vision 10 (4), 433–436.

Chang, E.-C., Yap, C. K., 1997. A wavelet approach to foveating images. In: SCG '97: Proceedings of the thirteenth annual symposium on Computational geometry. ACM Press, New York, NY, USA, pp. 397–399.

Geisler, W., Perry, J., 1998. A real-time foveated multiresolution system for low-bandwidth video communication. In: Human Vision and Electronic Imaging III. Vol. 3299. SPIE-Int. Soc. Opt. Eng, pp. 294–305.

Hamker, F. H., 2005. The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. Computer Vision and Image Understanding 100 (1-2), 64–106.

Itti, L., 2004. Automatic foveation for video compression using a neurobiological model of visual attention. Image Processing, IEEE Transactions on 13 (10), 1304–1318.

Itti, L., Aug. 2005. Quantifying the contribution of lowlevel saliency to human eye movements in dynamic scenes. Visual Cognition 12 (6), 1093–1123.

Itti, L., Aug. 2006. Quantitative modelling of perceptual salience at human eye position. Visual Cognition 14 (4 - 8), 959–984.

Itti, L., Koch, C., Mar. 2001. Computational modelling of visual attention. Nat Rev Neurosci 2 (3), 194–203.

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on 20 (11), 1254–1259.

Lee, S., Pattichis, M., Bovik, A., 2001. Foveated video compression with optimal rate control. Image Processing, IEEE Transactions on 10 (7), 977–992.

Mack, M. L., Castelhano, M. S., Henderson, J. M., Oliva, A., November 2003. What the visual system "sees": The relationship between fixation positions and image properties during a search task in real-world scenes. In: -. OPAM Annual Workshop, Vancouver.

Mante, V., Frazor, R. A., Bonin, V., Geisler, W. S., Carandini, M., Dec. 2005. Independence of luminance and contrast in natural scenes and in the early visual system. Nat Neurosci 8 (12 (Print)), 1690–1697.

Parkhurst, D., Law, K., Niebur, E., Jan. 2002. Modeling the role of salience in the allocation of overt visual attention. Vision Research 42 (1), 107–123.

Parkhurst, D. J., Niebur, E., Jun. 2003. Scene content selected by active vision. Spatial Vision 16 (2), 125–154.

Pelli, D., 1997. The videotoolbox software for visual psychophysics: transforming numbers into movies. Spatial Vision 10 (4), 437–442.

Privitera, C., Stark, L., 2000. Algorithms for defining visual regions-of-interest:

comparison with eye fixations. Pattern Analysis and Machine Intelligence, IEEE Transactions on 22 (9), 970–982.

Raj, R., Geisler, W., Frazor, R., Bovik, A., Oct. 2005. Contrast statistics for foveated visual systems: fixation selection by minimizing contrast entropy. Journal of the Optical Society of America A (Optics, Image Science and Vision) 22 (10), 2039–2049.

Rajashekar, U., van der Linde, I., Bovik, A. C., Cormack, L. K., 2007. GAFFE: A Gaze-Attentive Fixation Finding Engine. Accepted for publication in IEEE Transactions on Image Processing

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

Sheikh, H. R., Evans, B. L., Bovik, A. C., Feb. 2003. Real-time foveation techniques for low bit rate video coding. Real-Time Imaging 9 (1), 27–40.

Tatler, B. W., Baddeley, R. J., Gilchrist, I. D., Mar. 2005. Visual correlates of fixation selection: effects of scale and time. Vision Research 45 (5), 643–659.

Tatler, B. W., Baddeley, R. J., Vincent, B. T., Jun. 2006. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. Vision Res 46 (12 (Print)), 1857–1862.

Torralba, A., Jul. 2003. Modeling global scene factors in attention. Journal of the Optical Society of America A (Optics, Image Science and Vision) 20 (7), 1407–1418.

van Hateren, J. H., van der Schaaf, A., Mar. 1998. Independent component filters of natural images compared with simple cells in primary visual cortex. Proc Biol Sci 265 (1394), 359–366.

Wandell, B. A., 1995. Foundations of Vision. Sinauer Associates.

Wang, Z., Lu, L., Bovik, A., 2003. Foveation scalable video coding with automatic fixation selection. Image Processing, IEEE Transactions on 12 (2), 243–254.

Yarbus, A. L., 1967. Eye movements and vision. Plenum Press, New York.

Fig. 1. Examples of images used in the experiment.

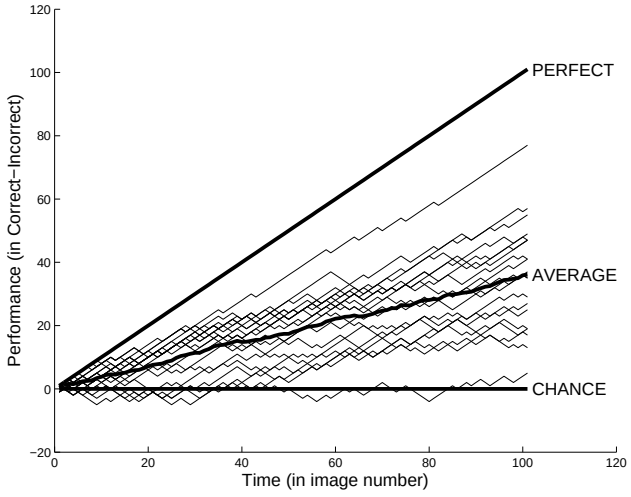


Fig. 2. Subject performance as a function of number of images viewed. Performance was measured as the number of correct responses minus the number of incorrect responses to the patch detection task.

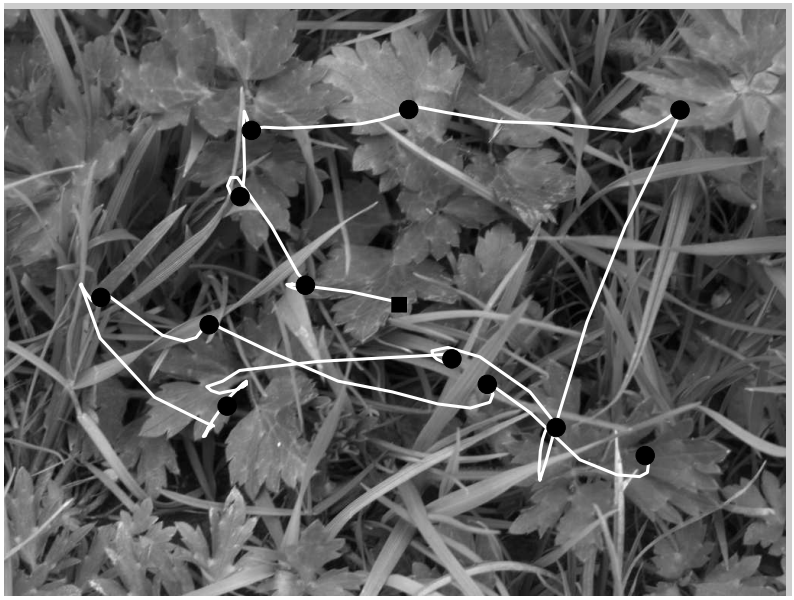


Fig. 3. Example of an observer's eye movement trace superimposed on the image stimulus. The dots are the computed fixations. The square in the center of the image is the first fixation.

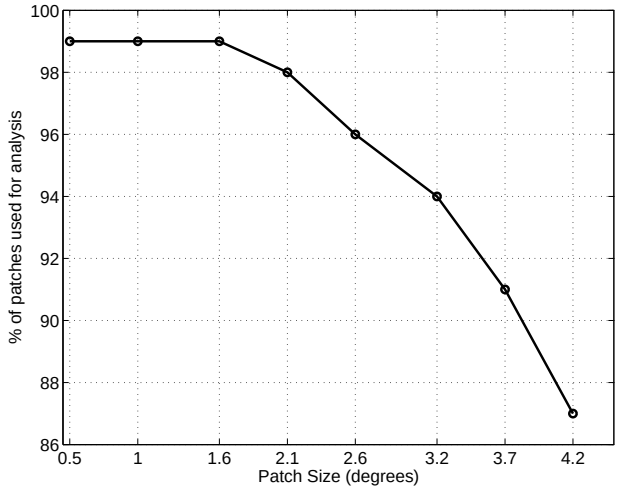


Fig. 4. Effect of patch size on the percentage of total fixations used for analysis.

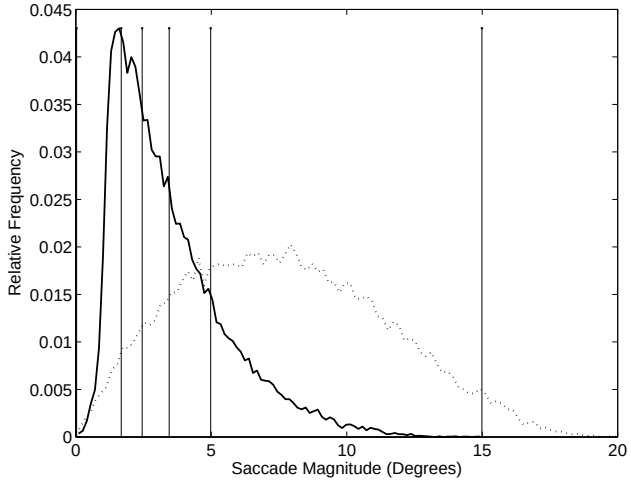
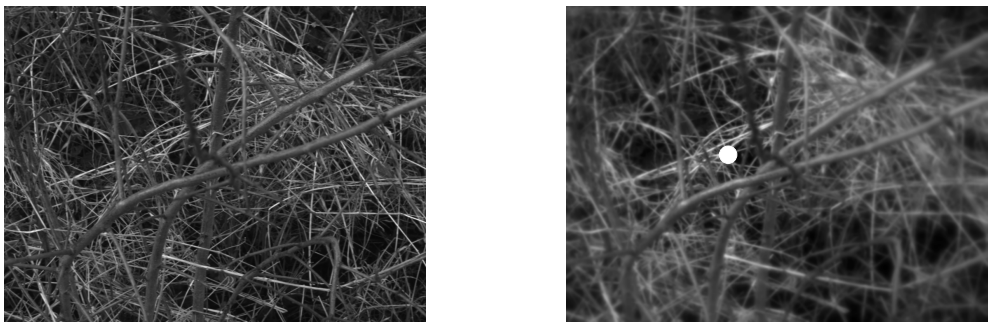


Fig. 5. Distribution of saccade magnitudes for human observers (solid line) and uniformly distributed fixations (dotted line). The vertical lines indicated the boundary of saccade bins used for the eccentricity-based analysis. Each bin contains approximately 6000 fixations.



(a) Original Image (b) Foveated Image

Fig. 6. Example of a full-resolution image (a) that has been foveated (b) about the fixation point indicated by the white dot.

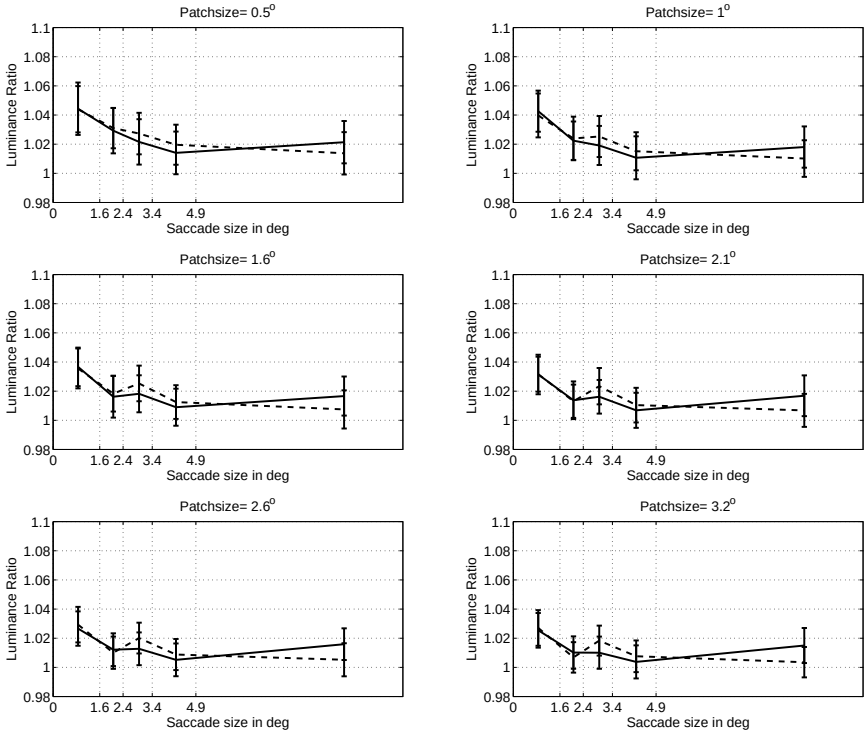


Fig. 7. Eccentricity-based analysis of ratios of mean luminance at observers' fixations to random fixations. Each panel plots the values of the ratio as a function of saccade eccentricity. Solid lines denote full-resolution analysis and the dashed lines indicate foveated analysis. Error bars signify 95% confidence intervals. Each panel corresponds to the patch diameter (indicated in degrees by the title for the panel) that was used to compute the mean luminance.

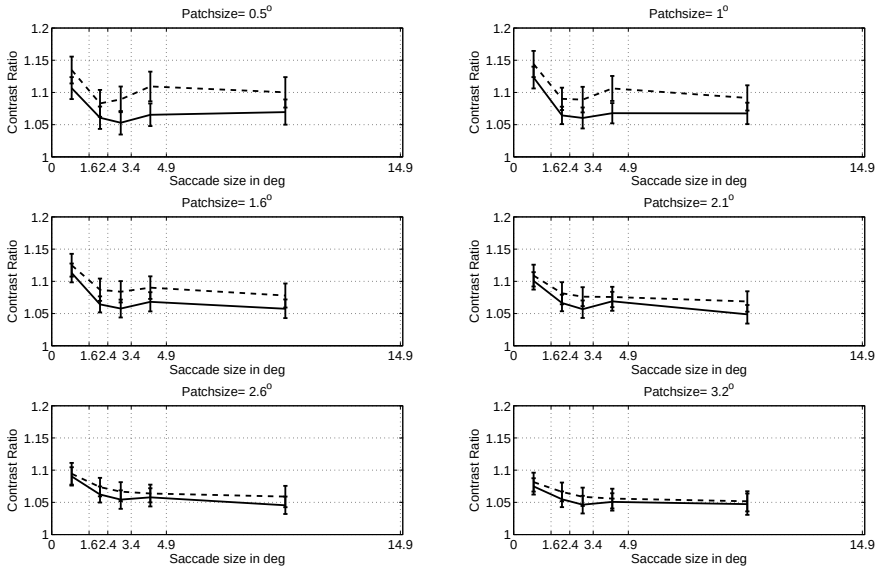


Fig. 8. Eccentricity-based analysis of ratios of RMS contrast at observers' fixations to random fixations. Each panel plots the values of the ratio as a function of saccade eccentricity. Solid lines denote full-resolution analysis and the dashed lines indicate foveated analysis. Error bars signify 95% confidence intervals. Each panel corresponds to the patch diameter (indicated in degrees by the title for the panel) that was used to compute the RMS contrast.

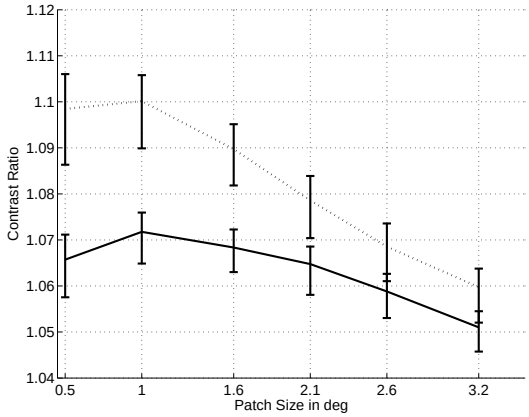


Fig. 9. Ratios of RMS contrast (at observers' fixations to random fixations) as a function of patch size. Solid lines denote full-resolution analysis and the dashed lines indicate foveated analysis. Error bars signify 95% confidence intervals.

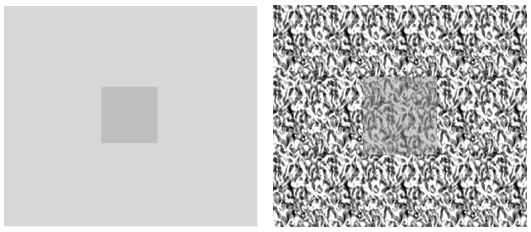


Fig. 10. Center-surround (bandpass) kernels can be used to detect luminance (left) and contrast (right) variations

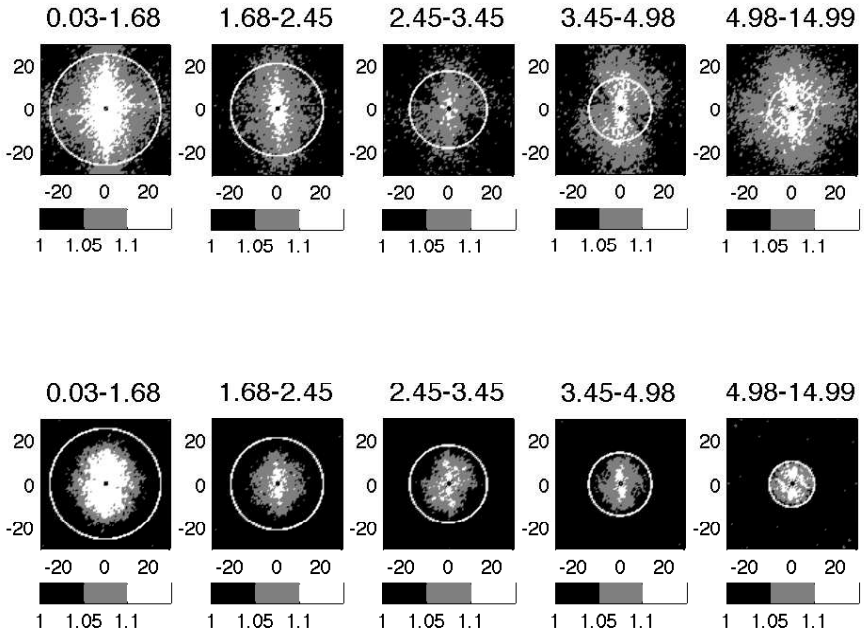


Fig. 11. Design of Bandpass kernels. The figure shows plots of F_{ratio} for full-resolution (top row) and foveated (bottom row) patches as a function of saccade magnitude for a patch size of $1.6^\circ \times 1.6^\circ$ pixels. Each column corresponds to the saccade bin in which the DFT analysis was performed (the bins are indicated on the title). The x and y axis on these plots correspond to cycles per degree. All plots have been plotted using the same colormap.The white circle in each panel indicates the spatial frequency cut off given by (1).

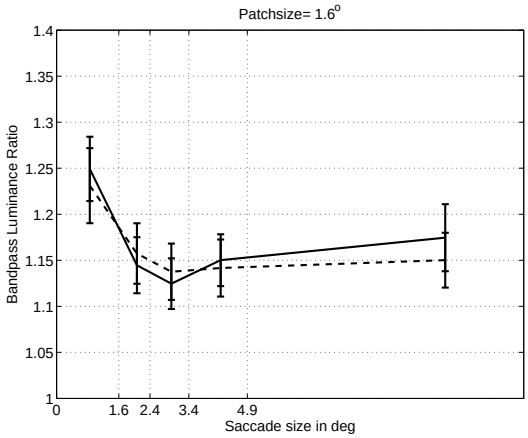


Fig. 12. Eccentricity-based analysis of ratios of bandpass luminance at observers' fixations to random fixations. The panel shows the values of the ratio as a function of saccade eccentricity. Solid lines denote full-resolution analysis and the dashed lines indicate foveated analysis. Error bars signify 95% confidence intervals. The patch diameter was $1.6^\circ \times 1.6^\circ$.

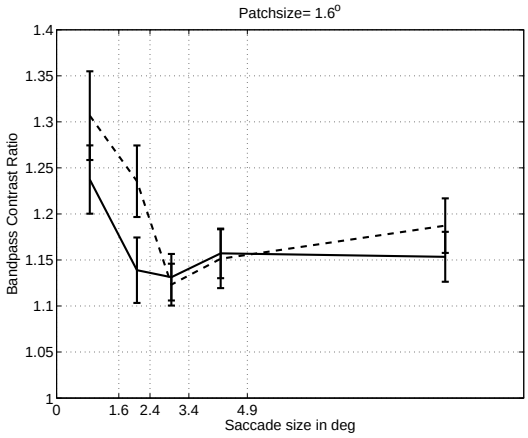


Fig. 13. Eccentricity-based analysis of ratios of bandpass contrast of image patches at observers' fixations to random fixations. The panel shows the values of the ratio as a function of saccade eccentricity. Solid lines denote full-resolution analysis and the dashed lines indicate foveated analysis. Error bars signify 95% confidence intervals. The patch diameter was $1.6^\circ \times 1.6^\circ$.

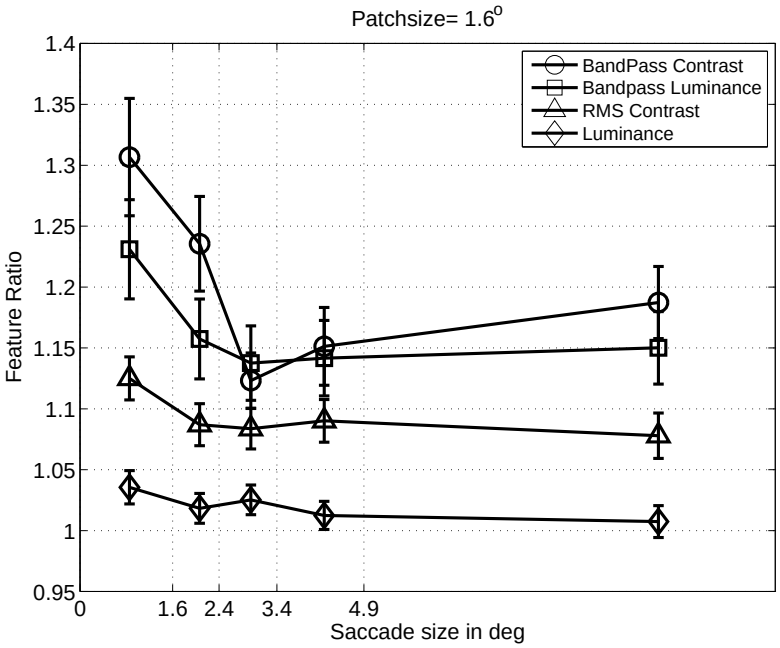


Fig. 14. Plots of the four foveated local image features as a function of saccade magnitude for a patch diameter of $1.6° \times 1.6°$. Error bars signify 95% confidence intervals on the sample mean.