GAFFE: A Gaze-Attentive Fixation Finding Engine

Umesh Rajashekar, *Member, IEEE,* Ian van der Linde, Alan C. Bovik, *Fellow, IEEE* Lawrence K. Cormack

***Abstract*— The ability to automatically detect visually interesting regions in images has many practical applications, especially in the design of active machine vision and automatic visual surveillance systems. 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, we studied the statistics of four low-level local image features: luminance, contrast, and bandpass outputs of both luminance and contrast, and discovered that image patches around human fixations had, on average, higher values of each of these features than image patches selected at random. Contrast-bandpass showed the greatest difference between human and random fixations, followed by luminance-bandpass, RMS contrast, and luminance. Using these measurements, we present a new algorithm that selects image regions as likely candidates for fixation. These regions are shown to correlate well with fixations recorded from human observers.**

***Index Terms*— Eye tracking, Point-of-gaze, Foveation, Fixation selection**

I. Introduction

The human visual system is constantly bombarded with a slew of visual data, from which it actively selects and assimilates relevant visual information in an efficient and seemingly effortless manner. Despite a large field of view, the human visual system processes only a tiny central region (the fovea) with great detail while the resolution drops rapidly towards the periphery [1]. Such a *foveated* visual encoding provides for a large field of view without the accompanying data glut. To assimilate visual information and build a detailed representation from this multi-resolution visual input, the human visual system uses a dynamic process of actively scanning the visual environment using fixations linked by rapid, ballistic eye movements called saccades [2]; most visual information is acquired during a fixation and little or no information is gathered during a saccade [3].

The active nature of looking, as instantiated in the human visual system, promises to have advantages in both speed and reduced storage requirements in artificial vision systems as well. In fact, several foveated vision sensor arrays have been designed and used in real-time imaging systems [4]–[6]. The next generation of efficient, foveated, active vision systems [7] could potentially be applied to a diverse array of problems such as automated pictorial database query, image understanding, image quality assessment [8], automated object detection, autonomous vehicle navigation, and real-time, foveated video compression [9], [10]. Also, the ability to understand and reproduce an expert radiologist's eye movements could be used in semi-automated detection of lesions in digital mammograms [11] - a problem of life-saving significance. Machine vision systems that can actively select visually interesting regions in an image also find applications in the area of planetary exploration [12]. It is conceivable that planetary rovers in the future will not need to wait for signals to move its cameras from an operator on earth who is several light seconds (or years) away. Many other significant applications can be envisioned.

While the degradation of spatial resolution in the retina has been modeled accurately by measuring the contrast thresholds of transient stimuli [13], [14], the fundamental question in the area of foveated, active artificial vision of 'How do we decide where to point the cameras next?' remains poorly understood. Despite the seemingly complex mechanisms that seem to underly the process of active vision, human observers seem to 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.

Research into the general area of how humans deploy eye movements in visual tasks has received significant attention for many decades [2], [15], [16]. Competing theories for gaze selection can be broadly classified into two general categories: top-down (cognitive/high-level) and bottom-up (precognitive/low-level). Top-down approaches for gaze prediction emphasize a high-level understanding of the scene and has been popular in task-specific experiments. Yarbus, in his pioneering work on eye movements [2], demonstrated that human eye movements are strongly influenced by high-level mechanisms such as the specific visual task given to the observer. Top-down implementations of gaze-selection have incorporated spatial relationships of object [17] and scene schema representations [18] and shown significant improvements in search times in visual search tasks. While such top-down implementations provide possible directions of exploration in gaze selection, cognitive interpretation of scenes is far from being sufficiently mature to generalize to natural viewing tasks. Given the rapidity and sheer volume of saccades during search tasks (over 15,000 each hour), it is reasonable to suppose that there is a significant bottom-up, computationally inexpensive component to selecting fixation locations. The

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

The authors are affiliated with the Center for Perceptual Systems at The University of Texas at Austin (UT-Austin), Austin, TX-78712. U. Rajashekar is currently at the Laboratory for Computational vision at New York University, I. van der Linde is with the Department of Computing, Anglia Ruskin University, UK., A. C. Bovik is with the Dept. of Electrical and Computer Engineering at UT-Austin, and L. K. Cormack is with the Dept. of Psychology at UT-Austin. (email: umesh@cns.nyu.edu; ianvdl@ece.utexas.edu; bovik@ece.utexas.edu; cormack@psy.utexas.edu)

goal of this paper is to investigate bottom-up, image-based mechanisms that guide eye fixations. Moreover, we believe that the development of future high-level visual search systems may benefit from the insights gained from successful low-level search strategies.

Bottom-up approaches to gaze selection assume that eye movements are quasi-random and driven by low-level image features. They propose a computational model for human gaze selection based on image processing to accentuate certain image features that are deemed relevant for drawing gaze. The influence of certain low-level image features such as edges and areas of high curvature in drawing fixations was established as early as 1935 [15], [16]. Williams [19], studied the influence of color, shape, and size in visual search and concluded that, among the attributes studied, the color of the target was the most important image feature in influencing saccades. More recently, Privitera & Stark [20] used a suite of algorithms such as detecting symmetry, center-surround regions in images that resemble receptive field profiles, wavelets, contrast, and edges-per-unit-area to select points of interest in an image and found that $43\% - 54\%$ of their fixation selections overlapped with actual human eye fixations. In another model inspired by mammalian visual systems [21], an image is first decomposed into its intensity, color, and orientation channels. Each feature is then represented by Gaussian pyramids which are used to compute center-surround responses to enhance features that differ from their neighbors. Using a normalization operator, these feature maps are combined across scales and features to result in conspicuity or saliency maps, whose peaks identify visually interesting regions. Several modifications to this general model that include motion parameters [22], novel combinations of the feature maps, and modulation by high-level contextual priors have been shown to provide superior gaze selection results. Torralba [23] proposed a statistical framework for incorporating high-level contextual information into such low-level saliency-based models for predicting gaze in object detection. The use of scene context in conjunction with saliency maps is shown to correlate better with human fixations than using only the saliency map to select fixations in visual search tasks.

Since the human visual system evolved in a natural environment and natural images occupy a relatively small subspace of all possible images, it is theorized [24]–[26] that early visual processing may exploit the statistics inherent in its environment to represent the input as efficiently as possible. With the availability of inexpensive, accurate eye trackers, a recent trend in the bottom-up approach to understanding gaze has been to directly measure and quantify the differences in the statistics of image patches at the *point of gaze* of observers and those selected at random. Reinagel *et al.* [27] show that human fixation regions have higher spatial contrast and spatial entropy than randomly fixated regions, indicating that the human eye may be trying to select image regions that maximize the information content transmitted to the visual cortex. Recently, Parkhurst *et al.* [28] replicated these results with various sizes of the patch used to compute the local image contrast, and found that local image contrast is reliably higher (statistically significant) than those obtained from patches at random fixations. They found that the difference in the contrast statistics between human and random fixations was larger for intermediate patch sizes, with a maximum difference occurring around patch sizes of 1°. While these gaze-contingent approaches have provided insight into the visual features that are useful for understanding and hence modeling gaze, the ensemble of image patches at observer's fixations have always been analyzed at the native resolution of the stimulus. A moment of introspection suggests that analysis of bottom-up fixation attractors must actually involve a foveated framework, where low-level image features that attract subsequent fixations are extracted from the visual periphery whose resolution varies across the visual field. Parkhurst *et al.* [29] 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 [30].

In this paper, we present a gaze-attentive fixation finding engine (GAFFE) that uses a bottom-up modality for fixation selection in natural scenes. GAFFE uses a data-driven framework where eye tracking was first used to evaluate the contributions of four *foveated* low-level image features in drawing fixations of observers. In particular, as described in Section II, we recorded the eye movements of 29 observers as they viewed 101 calibrated natural images, and attempted to quantify the differences in the statistics of four image features (described in Section III): luminance, contrast, and bandpass outputs of luminance and contrast at observers' fixations and fixations selected at random. Following a discussion of the image analysis at point of gaze, a foveated fixation selection algorithm that selects image regions in novel scenes as likely candidates for fixation based upon a linear combination of the relevant low-level features is presented in Section IV. Finally, we evaluate the performance of GAFFE by computing the correlation between the predicted and recorded fixations.

GAFFE introduces several new techniques to gaze selection as described below. As mentioned before, all previous approaches to evaluating image statistics at the point-of-gaze have ignored the foveated sampling of the human visual system. We address this issue by first foveating the stimulus at the observer's current fixation point (using established models of resolution fall-off in the periphery [13]), and then analyze the statistics of the various image features using appropriately blurred versions of image patches centered on the subsequent fixation. A direct consequence of this is an eccentricity-based analysis where every image patch around a fixation is analyzed based on its eccentricity from the previous fixation. We also introduce contrast-bandpass as a new low-level feature that is shown to correlate very well with human fixations. While point-of-gaze analyses have been used before to quantify the differences in low-level images at human and randomly selected fixations, this information has not been used to actually select fixations in novel scenes.

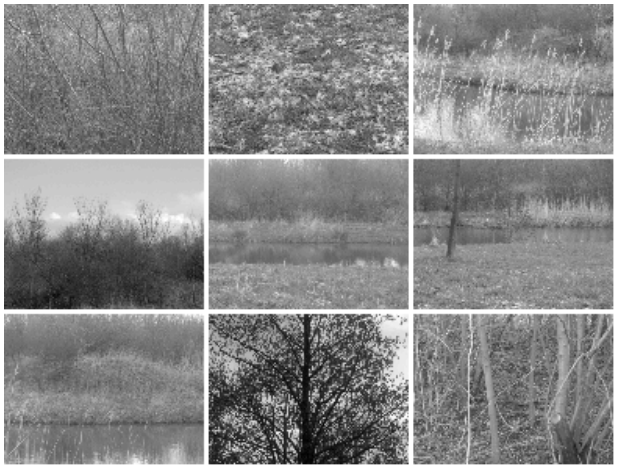

Fig. 1. Examples of images used for the experiment

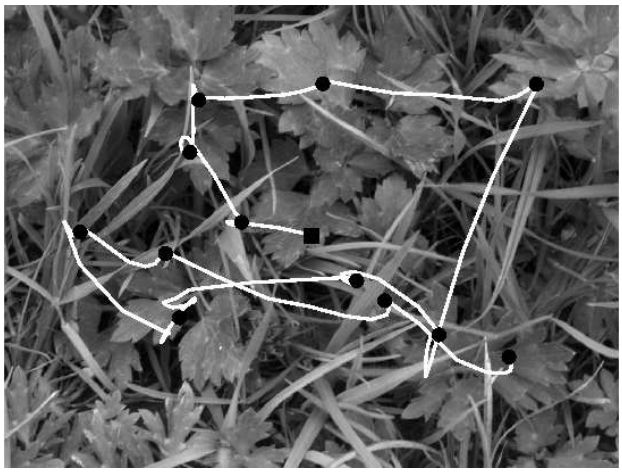

Fig. 2. Example of an observer's eye movement trace superimposed on the image stimulus. The dots are the computed fixations. The square in the center of the image is the first fixation.

II. Eye Tracking Methods

GAFFE is based on a gaze-attentive framework; this means that the features used for fixation selection are those that are statistically significant at recorded human gaze locations (when compared to features at randomly selected fixations). This section describes the experimental procedure that was used to record human eye movements in a natural viewing task.

A. Stimuli and Tasks

101 static images of size $1024 * 768$ pixels were manually selected from a calibrated gray scale natural image database [31]. Since we were interested in developing a bottom-up framework for fixation selection, 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. 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 corresponded to about 1 arc minute per pixel. Each image was displayed for 5 seconds in a fixed order for all observers.

Observers were instructed to view each of the images as they desired. All observers began viewing the image stimuli from the center of the screen. Following the display of each image, observers were shown a small image patch and asked to indicate whether the image patch was from the image they just viewed or not. This task was used to encourage observers to scan the entire scene. A total of 29 (24 naïve) adult human volunteers participated in this study. All observers either had normal or corrected-to-normal vision.

B. Eye Tracking

Human eye movements were recorded using an SRI Generation V Dual Purkinje eye tracker. It has an accuracy of < 10 arc minute, and a precision of ~ 1 arc minute. A bite bar and forehead rest was used to restrict the observer's head movements. The observer was first positioned in the eye tracker and a positive lock established onto the observer's eye. A linear interpolation on a 3×3 calibration grid was then done to establish the linear transformation between the output voltages of the eye tracker and the position of the observer's gaze on the computer display. The output of the eye tracker (horizontal and vertical eye position signals) was sampled at $200Hz$ and stored for offline data analysis. This calibration routine was repeated every 10 images, and a calibration test run after every image.

C. Image Data Acquisition

The gaze coordinates corresponding to the eye movements of the observers for each trial were divided into fixations and saccades using spatio-temporal criteria derived from the known dynamic properties of human saccadic eye movements [32]. The resulting pattern of fixations for a single trial is shown by the dots in Fig. 2. The lines show the eye movement trajectories linking the fixations. As mentioned earlier, we propose a foveated framework to analyze the statistics of low-level features of image patches at the resolution at which they were encoded by the observer. To achieve this, the image was first foveated at the observer's current fixation, say n, and a patch centered at the subsequent fixation, $n + 1$, was extracted for analysis. Thus all image patches were analyzed at the resolution at which they were encoded *prior* to fixating the patch. 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°.

A consequence of using such a foveated analysis framework is that the ensemble of patches extracted around fixations contain image patches that have been blurred to different extents. Further, it is also possible that saccades of different magnitudes are driven by different features. Thus, there arises a need to perform an eccentricity-based analysis of local image features, where patches of similar blur are grouped together and the relevant image feature is analyzed separately for each blur. Tatler *et al.* [33] have observed that the influence of image features are not uniform across saccade magnitudes and note that by ignoring the dependence of image features on saccade

magnitudes, prior work in this area ([27], [28], [34]) generally tends to estimate the influence of visual features incorrectly. In our study, since we use a foveated analysis framework, we analyze patches over the range of spatial frequencies at which they were processed by the human visual system, and thus incorporate both saccade and spatial frequency dependence of image patches into our analysis.

To perform the eccentricity-based analysis of our image statistics, each patch in the database was first associated with the length of the saccade, e (in degrees), that was executed to reach that particular patch. The distribution of these saccade magnitudes were quantized into 5 bins such that each bin contained the same number of patches (around 6000) and the patches in each bin were analyzed separately. Patches with small eccentricity values were blurred less than patches with larger eccentricity values in accordance with established models for foveation [13]. The location of the saccade bin boundaries were: $0.03, 1.68, 2.45, 3.45, 4.98,$ and 14.99 degrees.

III. Computing Local Image Features

The image patches around observers' fixation points were then analyzed to determine if the statistics of the four image features: luminance, contrast, luminance-bandpass, and contrast-bandpass were statistically different from image patches that were picked randomly. The randomly selected patches were obtained by shuffling the fixations of an observer for a particular image with that of a different image. Thus this image shuffled database simulates a random human observer whose fixations are not influenced by features of the underlying image, but otherwise captures all the statistics of human eye movements. 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).

For any image feature, S, we were interested in the differences (and not the absolute values) in the image statistics at observers' fixation and randomly selected fixations. Therefore, for each image, n, we computed the ratio of the average patch feature at eccentricity, e, at the observers' fixations, $\overline{S}(e,n)_{obs}$ to the average patch feature for image patches from the image shuffled database, $\overline{S}(e,n)_{rand}$, and then averaged this ratio across the $N(=101)$ images in the database:

$$\overline{S}(e)_{ratio} = \frac{1}{N}\sum_{n=1}^{N}\frac{\overline{S}(e,n)_{obs}}{\overline{S}(e,n)_{rand}}. \quad (1)$$

Finally, to evaluate the statistical significance of the image statistic under consideration, we used bootstrapping [35] to obtain the sampling distribution of this mean ratio as follows. Given an image, a set of patches around observers' fixations was extracted. From this set, a new collection of image patches was obtained by sampling with replacement. The feature of interest (such as patch luminance or RMS contrast) was then computed for this set of patches. The above process was repeated for the image-shuffled fixations for that image. The ratio (1) was then computed across the 101 images in the database to constitute one bootstrap replication. This process was repeated 200 times to obtain the sampling distribution of the average ratio (1) for that feature and used to identify the confidence intervals.

The rest of this section describes how each of the four image features: luminance, contrast, luminance-bandpass, and contrast-bandpass was computed for an image patch.

A. Luminance Computation

The mean luminance, $\bar{I}$, for an image patch was computed using a circular raised cosine weighting function, w as follows:

$$\bar{I} = \frac{1}{\sum_{i=1}^{M} w_i}\sum_{i=1}^{M} I_i w_i \quad (2)$$

where, M is the number of pixels in the patch, I_i is the grayscale value of the pixel at location i and the raised cosine function w is expressed as:

$$w(i) = 0.5\left[\cos\left(\frac{\pi r_i}{R}\right) + 1\right] \quad (3)$$

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.

B. RMS Contrast Computation

For an image patch, a weighted root-mean-squared contrast using a circular raised cosine weighting function, w, was computed as:

$$C = \sqrt{\frac{1}{\sum_{i=1}^{M} w_i}\sum_{i=1}^{M} w_i \frac{(I_i - \bar{I})^2}{(\bar{I})^2}} \quad (4)$$

where M is the number of pixels in the patch, I_i is the grayscale value of pixel at location i, and $\bar{I}$ is the mean luminance of the patch from (2).

C. Bandpass of Patch Luminance

Attention often seems to be drawn to regions that differ from their surroundings in some aspect. Such regions can be detected by the outputs of center-surround or, more generally, bandpass (Gabor) kernels (which have been popular models for the receptive fields of simple cells in the primary visual cortex). Thus, the next image feature that we investigated was the output of Gabor filters operating on the patch luminance. Of the many Gabor kernels that can be used to filter an image patch, we used the kernel that best modeled (in a least squares sense) the spatial frequencies where the human patches differed significantly from the random patches. In particular, we computed the ratio of the average FFT magnitudes of the image patches at point-of-gaze to those at random fixations and modeled the significant spatial frequencies using least square fits of Gabor functions (Fig. 6). The technique is described

in more detail in the Appendix. Having found the bandpass kernels, the final step involved filtering the image patches using the kernels. Given an image patch $I(e)$, located at an eccentricity e from the previous fixation, we select the Gabor kernel, $Gab_{lum}(e)$, corresponding to this eccentricity bin, and computed the maximum absolute value of the result of filtering this image patch with the Gabor kernel as our feature: $G_{lum} = \max|Gab_{lum}(e) * I(e)|$, where $*$ corresponds to the convolution operator.

D. Bandpass of Patch Contrast

Finally, bandpass outputs of local image contrast (i.e. contrast of contrasts) was used to capture higher order image structure that is ignored by the luminance Gabors described in Section III-C. For example, regions whose central and surrounding regions have the same mean luminance, but different contrast profiles can be captured by this feature. Computing the contrast-bandpass Gabor kernel is more complicated than the luminance-bandpass kernels because we first have to compute local image contrast - which itself depends on the size of neighborhood used to compute the contrast - and then find 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 first computed the magnitude of the local image gradient for each pixel and used this as a measure of an extremely local (pixel-level) measure of image contrast. The goal of designing the contrast bandpass kernel now amounts to determining the spatial scales at which these local image gradients vary. We then computed the ratios of the average FFT magnitudes of the *gradient patches* at point-of-gaze to those at random fixations and modeled the significant spatial frequencies using least square fits of Gabor functions. With the bandpass kernels designed, we repeated the Gabor filtering as before with the exception that the filtering was applied to the local patch gradient instead of the patch itself: $G_{grad} = \max|Gab_{grad}(e) * |\nabla I(e)||$, where $*$ corresponds to the convolution operator, $|\nabla I(e)|$ is the magnitude of the gradient of an image patch at eccentricity e, and $Gab_{grad}(e)$ is the Gabor kernel at this eccentricity.

IV. Gaze-attentive Fixation Selection

Luminance and contrast statistics were computed for all the patch sizes mentioned earlier. However, the bandpass ratios were computed only for a single patch size of $1.6^\circ \times 1.6^\circ$ due to the computational constraints of finding the optimal bandpass kernels. This patch size was selected because it provided the maximum contrast ratio between human and random fixations. The value of the feature ratio, $\overline{S}(e)_{ratio}$, was computed as described in (1) for the four image features described in Section III and is plotted as a function of saccade magnitude, e, in Fig. 3 for a patch size of $1.6^\circ \times 1.6^\circ$. The error bars represent a 90% confidence interval obtained from the 200 bootstrap replications. First, we note that for all features, the mean value of $\overline{S}(e)_{ratio}$ is significantly higher than 1.0, which implies that the image patches around human fixations had, on average, higher values for each of these features than the image patches selected at random *at all eccentricities*.

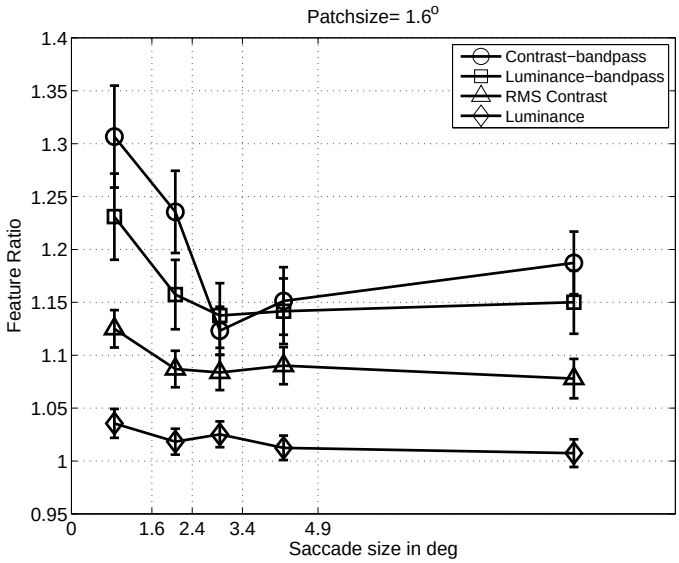

Fig. 3. Plots of the average feature ratios, $\overline{S}(e)_{ratio}$, as a function of saccade magnitude. Error bars denote 90% confidence intervals obtained via bootstrapping. The patch size used for computing the features was 96×96 pixels ($1.6^\circ \times 1.6^\circ$).

Second, by examining the actual values of the ratios, we found that contrast-bandpass showed the greatest difference between human and random fixations (maximum ratio of 1.3, average of 1.2), followed by luminance-bandpass (maximum of 1.23, average of 1.16), RMS contrast (max of 1.12, average of 1.09), and finally luminance (max of 1.04, average of 1.01). Contrast-bandpass (or contrast of contrasts) could correspond to regions with a clear distinction of foreground and background, and thus instinctively draw human fixations and produce a very high value for the $\overline{S}(e)_{ratio}$. Our results agree with Tatler *et al.*'s [33] findings that short saccades are more image feature dependent than long saccades. In summary, the point-of-gaze analysis shows that image patches selected by human observers have higher luminance, contrast, and stronger bandpass profiles than randomly selected patches. In a related study [36], we have also discovered that a full-resolution analysis for these features produces similar results, but underestimates the influences of contrast-related features; the resulting ratios were found to be higher (statistically significant) for the foveated patches. In Section IV-B, we also show that the foveated framework perfoms better than the full-resolution analysis in gaze selection.

Since these statistics were obtained directly from the fixations of human observers, these findings can also be used to select fixations in new scenes in a manner that mimics the fixation pattern of human observers. The remainder of this section presents a simple algorithm that uses these visually important image features to select fixations in a new scene. Given an image, the algorithm begins by selecting the center of the image as the first fixation point. This selection is consistent with previous findings that observers tend to first fixate at the middle of the image stimulus [37]. To simulate the foveated encoding of the human visual system, the image is then foveated around this central fixation point. The foveated image

is then filtered to create a saliency map for each of the four features discussed earlier. Saliency maps for luminance and contrast are computed using a fixed kernel size of $1.6^\circ \times 1.6^\circ$ pixels. Saliency maps for the bandpass kernels are obtained using the five Gabor kernels (one per saccade bin) obtained as described in the Appendix. The filtering process for the bandpass kernels is space-variant - i.e. the type of kernel that is used at a certain location in the image depends on the distance of that location from the current fixation point. Therefore, image regions that are nearest to the current fixation point are filtered with the kernel corresponding to the small magnitude saccade bins in Fig. 6, and points that are farther are filtered using the corresponding kernel from a large magnitude saccade bin. Since the kernels in Fig. 6 were computed for 5 saccade bins, the resulting filtered image has 5 circular regions of filtered outputs. The filtered output can be interpreted simply as a likelihood map in which regions with large values are more likely to draw a fixation than those regions with lower values. The four feature maps were then linearly combined using a weighted average where the weights for each of the feature maps were selected to be proportional to the maximum value of the ratio values they generated in the comparison against randomly selected patches. Thus the weights for the luminance, contrast, luminance-bandpass, and contrast-bandpass from Fig. 3 were 1.04, 1.12, 1.23, and 1.30 respectively. The weights were normalized to sum to unity. The algorithm uses a greedy criterion in selecting the maximum value from this weighted selection map as the next fixation point, foveates the image around this point, and repeats this process. The resulting selection map was also weighted using an inverted Gaussian mask centered on each selected fixation point. This masking simulates an inhibition-of-return mechanism [38] and prevents the future fixation selections from landing very close to previously selected fixations. At each stage, to alleviate boundary artifacts of filtering, the selection map was also weighted with a rectangular mask that had a value of ones in the center and tapered sharply towards zero at the image boundaries.

A. Qualitative Comparison of Fixation Selections

Figure 4 qualitatively illustrates the performance of the fixation selection algorithm. For visualization purposes, the fixations of 29 observers on these images were clustered using a density-constrained clustering algorithm, wherein the growth of the cluster is constrained by a minimum density requirement. In other words, the cluster is allowed to grow in size only if the new cluster contains a minimum number of fixations per unit area. Details of the implementation of a density-based algorithm, DBSCAN, can be found in [39]. A density constraint that required at least four fixations in a 1° region of a cluster produced reasonable clusters in these tests. Ten clusters with the maximum density of fixations are shown as ellipses in Fig. 4. The fixation selection algorithm was used to select a sequence of 10 fixations, each of which was represented by a 2D Gaussian window, illustrated by the bright regions in Fig. 4. The full width at half-max of the Gaussian roughly equaled the diameter of the human foveola (about 1° visual angle). The degree of overlap between the ellipse and the bright regions is a subjective measure of the performance of GAFFE. An objective measurement is presented in the following section.

B. Quantitative Comparisons of Fixation Selections

In section IV-A, we demonstrated qualitatively that fixations can be selected using a linear combination of low-level image features. Quantifying the similarity between recorded fixations and those selected by an algorithm generally involve clustering human fixations into regions that are then compared with fixations selected by the algorithm using string-matching algorithms [20]. Other methods of comparing human fixations to predictions are discussed in [40]. In our experiments, since there were many fixations per image (about 300), we opted to extrapolate this human eye fixation data to a pseudo-dense fixation selection map similar to the method used in [40]. First, given an image, each recorded fixation for that image was represented by a 2D Gaussian window whose full-width at half max was selected to be a 1° of visual angle as in Section IV-A. Then, the fixation selection algorithm was used to select 10 fixation points, each of which was again represented by a 2D Gaussian window. The resulting maps, when normalized to sum to unity, can be viewed as two dimensional probability density maps, where peaks correspond to regions with a high probability of drawing an observer's fixation. We then computed a zero-lag correlation between these two maps to quantify the degree of overlap between the fixations selected by the algorithm and the recorded fixations. As mentioned earlier, the first fixation for GAFFE was manually selected at the image center to match the oberservers' task. To avoid spurious correlations due to this set up, we ignored the first fixation from both recorded and predicted fixations before computing the correlation coefficients.

Figure 5 shows the average correlation values between the recorded fixations and the fixations generated by the four image features (luminance, RMS contrast, luminance-bandpass, contrast-bandpass) discussed earlier. The error bars represent standard errors. Since the 'combined' feature weights the contrast-bandpass most heavily, its correlation is only marginally higher than the contrast-bandpass. Finally, we also computed the correlation coefficient for a full-resolution model that uses the same four image features as above, but without a foveated framework, and for another popular saliency model for fixation selection [21]. We note that, in general, the fixations selected by the foveated analysis correlates better with the recorded fixations than those generated by the full-resolution models. The lower bound on the correlation coefficient is obtained by randomly selecting the same number of fixation locations as the algorithm. We see from the correlation plots, that all image features perform better than a random fixator, with the combined feature map producing the best correlation to recorded fixations. The upper bound on the correlation coefficient is determined by the variability in fixation locations across observers. This inter-observer variability was measured by separating observers into two groups and computing the correlation in the fixation maps between the

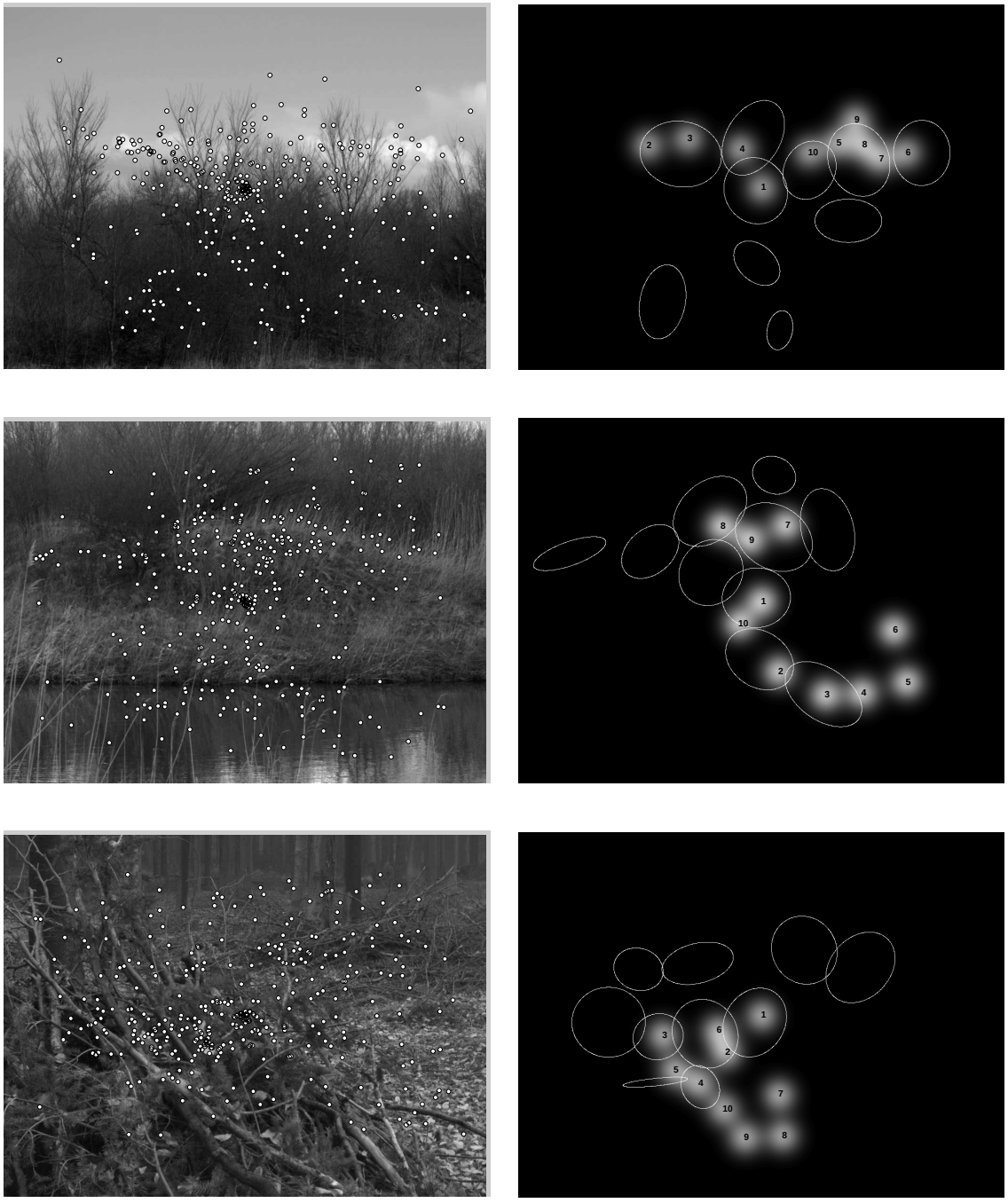

Fig. 4. Examples of fixation selection using a combination of image features. The left column shows the original images with fixations superimposed. The right column shows fixations selected using a linear combination of four image features. The numbers denote the order in which fixations are selected. Each fixation is represented by a 2D Gaussian window. The ellipses denote clusters of human fixations on these images.

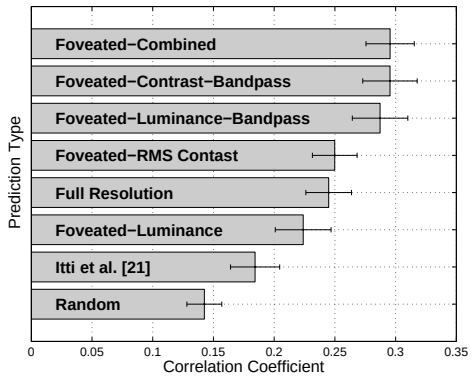

Fig. 5. Quantitative comparison of selected fixations with recorded fixations. The bars show the correlation between recorded fixations and fixations selected using image features.

two groups. The upper bound was found to be around 0.75 for our database, suggesting further room for improvement from other bottom-up or a combination of bottom-up and top-down features.

For our simulations, we used 10 fixations since it represented the average number of fixations executed by an observer for an image in our experiment. In an another analysis, we gradually increased the number of fixations selected by GAFFE (from 1 to 10) and found that the difference in correlation coefficients between the various features decreased. It is likely that with a large number of fixations, the inhibition-of-return simply forces future fixations to span the entire image, thereby resulting in a similar value of the correlation coefficient. To evaluate the influence of fixation durations on the correlation analysis, we also computed the correlation coefficient between fixations selected by GAFFE and a pseudo-dense map of recorded fixations where each fixation was replaced by a Gaussian whose amplitude was scaled in proportion to the duration of the corresponding fixation. The correlation coefficients were found to be lower when the fixation durations were included. This decrease can be attributed to the fact that GAFFE weights all fixations equally, whereas in reality, some fixations are more salient than the others.

V. Conclusion

The interplay of top-down (high-level/cognitive) mechanisms such as image understanding and bottom-up (low-level/ pre-cognitive) image features (such as edges, contrast and motion) influence eye movements in many intricate ways, making the task of accurately modeling gaze a formidable task. However, analysis of stimuli at observers' point of gaze can provide an understanding of strategies used by observers in visual tasks. In this paper, we presented GAFFE: a bottom-up, data-driven procedure wherein eye tracking is first used to measure the influence of four foveated low-level image features (luminance, contrast, luminance-bandpass, and contrast-bandpass) in drawing the fixation of human observers. The foveated image features thus computed are used to select fixations in novel scenes and are shown to correlate well with the fixations selected by human observers. Contrast-bandpass is shown to provide the best correlation amongst the four features studied. The matlab code for GAFFE can be downloaded from `http://live.ece.utexas.edu/research/gaffe`. In the near future, as a service to the 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.

While this paper presented the selection of visual fixations in the absence of any particular visual task, we have also used a similar gaze-contingent analysis framework to discover strategies used by human observers in a visual search task where observers searched for simple geometric targets such as horizontal edges and triangles embedded at very low signal-to-noise ratios in noise stimuli that had the spectral characteristics of natural images. By analyzing properties of the noise stimulus at observers' fixations, we were able to reveal idiosyncratic, target-dependent features used by observers in the visual search task [41], [42]. The extracted features were also found to be effective in selecting potential locations of targets that matched human fixations in novel noise stimuli [43].

The performance of GAFFE has been tested only on a specific collection of outdoor natural scenes, with features extracted at a single scale that was based on a particular viewing scenario. A natural extension of this work would involved a multi-scale analysis of the relevant features, which can include higher order features such as orientation, texture, and structure amongst others [44]. Recently, it has been shown that luminance and contrast are statistically independent features in natural images [45]. Thus, it is reasonable to assume that the saliency maps contributed by these two features to the fixation selection map are not overly redundant. Further analysis is required to evaluate the contributions of the other features to the saliency map, and the effect of these features on different datasets of images [28]. A related issue is one of feature combination. As seen in Fig. 5, a simple linear combination of the features is only marginally better (in terms of correlation) than using the contrast-bandpass as a fixation predictor. We can envision an image-dependent weighting of the features, where weights are computed dynamically based on the distribution of the four foveated features in the image. The inhibition-of-return mechanism in GAFFE is rather unnatural in that it never allows future fixations to land near previously fixated regions. A decay factor can be incorporated into this mechanism to allow GAFFE to return to previous fixation locations after a certain number of fixations. Additionally, one could incorporate well-known observer idiosyncrasies such as the tendency to fixate at the image center. Incorporating such a return-to-center mechanism after a certain number of fixations would allow the algorithm to 'reset' and explore previously ignored regions of the image. As mentioned earlier, it might also be useful to incorporate fixation dwell times in the analysis of image features for fixation selection. Another bugaboo in gaze prediction is the design of a quantitative metric for comparing predicted fixations with recorded fixations. We chose the correlation coefficient over other approaches such as string-

editing and Kullback-Leibler divergence owing to its ease of interpretation. However, it is possible that the quantitative comparisons between recoded and predicted fixations can be greatly simplified by assuming that the pseudo-dense map of recorded fixations represents a true probability of fixation locations for the image, and comparing the log-likelihood for the different sets of predicted fixations. This procedure obviates the need to estimate a dense probability distribution from only 10 predicted fixations. Finally, we note that GAFFE does not predict the sequence of fixations of observers; instead it selects regions that will, on average, be selected by human observers. Modeling the sequence of fixations, however, is much more challenging than predicting their locations [20].

APPENDIX
BANDPASS KERNEL DESIGN

To compute the bandpass kernel that provides maximum separation between human and random patches, we could resort to a brute force approach by changing various parameters of the bandpass kernel (such as full width at half-max, shape, and orientation). The following is an alternative approach that involves designing the kernels in the Fourier domain. We begin by locating the spatial frequencies where the human patches differ significantly from the random patches as follows. Given a patch p in image i located at an eccentricity e from the previous fixation, the ratio of the average discrete Fourier transforms (DFT) of image patches at point of gaze, $FFT(p,e)|_{PoG}$, to the discrete Fourier transform of patches selected randomly $FFT(p,e)|_{Rand}$ was computed:

$$F_{ratio}(i,e) = \frac{\frac{1}{P(i,e)} \sum_{p=1}^{P(i,e)} abs(FFT(p,e))|_{PoG}}{\frac{1}{R(i,e)} \sum_{r=1}^{R(i,e)} abs(FFT(r,e))|_{Rand}} \quad (5)$$

where $P(i,e)$ and $R(i,e)$ correspond to the number of image patches at human and random fixations respectively. The average value of this ratio across all images is computed as

$$F_{ratio}(e) = \frac{1}{N} \sum_{i=1}^{N} F_{ratio}(i,e) \quad (6)$$

where N is the number of images in the database. Prior to computing the DFT, each image patch was first windowed using a raised cosine window to avoid edge effects.

Figure 6 shows the plots of $F_{ratio}(e)$ for a patch size of $1.6^\circ \times 1.6^\circ$ for various saccade eccentricities, e. Each panel in the top row of Fig. 6 corresponds to a ratio of centered DFTs, and thus the central regions in each plot corresponds to low spatial frequencies with spatial frequency increasing away from the center.

Since we are looking at ratios of magnitudes of DFTs of patches selected by human fixations to those from the image-shuffled fixation, spatial frequencies with ratio-values greater than 1.0 (shown in white) have higher energy in the point-of-gaze patches (and therefore influence observers' gaze). Similarly, spatial frequencies with values close to 1.0 (shown by dark regions) do not play an important role in drawing fixations because their energy is similar to the image shuffled patches. Finally, to locate the statistically significant spatial frequencies, the FFT ratios were bootstrapped and 100 bootstrap estimates of $F_{ratio}(e)$ were computed. Spatial frequencies that were statistically different from 1.0 were selected and modeled using Gabor kernels using numerical optimization routines in Matlab. The resulting fits are shown in the bottom row of of Fig. 6 are indeed a coarse approximation, and better models for relevant spatial frequencies can be used. The effect of foveation manifests itself by highlighting the lower frequencies at larger saccade magnitudes. This is expected because, for large saccades the patches are foveated to a greater extent, and therefore the region of relevant frequencies gets smaller.

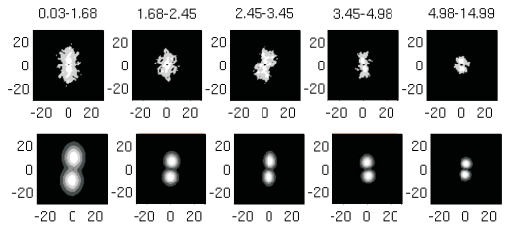

Fig. 6. Design of Bandpass kernels. The top row shows plots of $F_{ratio}(e)$ as a function of saccade magnitude for a patch size of $1.6^\circ \times 1.6^\circ$ pixels. Each column corresponds to the saccade bin in which the DFT analysis was performed (the bins are indicated on the title). The x and y axis on these plots correspond to cycles per degree. All plots have been plotted using the same colormap. The bottom row shows the corresponding best fitting Gabors.

REFERENCES

[1] B. A. Wandell, *Foundations of Vision*. Sinauer Associates, 1995.
[2] A. L. Yarbus, *Eye movements and vision*. New York: Plenum Press, 1967.
[3] D. Burr, M. Concetta Morrone, and J. Ross, "Selective suppression of the magnocellular visual pathway during saccadic eye movements," *Nature*, vol. 371, no. 6497, pp. 511–513, Oct. 1994.
[4] A. Moini, "Vision chips or seeing silicon," http://www.eleceng.adelaide.edu.au/Personal/moini, March 1997.
[5] S. Xia, R. Sridhar, P. Scott, and C. Bandera, "An all CMOS foveal image sensor chip," in *ASIC Conference 1998. Proceedings. Eleventh Annual IEEE International*, 1998, pp. 409–413.
[6] R. Etienne-Cummings, J. Van der Spiegel, P. Mueller, and M.-Z. Zhang, "A foveated silicon retina for two-dimensional tracking," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 47, no. 6, pp. 504–517, 2000.
[7] W. Klarquist and A. Bovik, "Fovea: a foveated vergent active stereo vision system for dynamicthree-dimensional scene recovery," *Robotics and Automation, IEEE Transactions on*, vol. 14, no. 5, pp. 755–770, 1998.
[8] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporatinghigher level perceptual factors," in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, 1998, pp. 414–418 vol.3.
[9] S. Lee, M. Pattichis, and A. Bovik, "Foveated video compression with optimal rate control," *Image Processing, IEEE Transactions on*, vol. 10, no. 7, pp. 977–992, 2001.
[10] Z. Wang, L. Lu, and A. Bovik, "Foveation scalable video coding with automatic fixation selection," *Image Processing, IEEE Transactions on*, vol. 12, no. 2, pp. 243–254, 2003.
[11] G.-Z. Yang, L. Dempere-Marco, X.-P. Hu, and A. Rowe, "Visual search: psychophysical models and practical applications," *Image and Vision Computing*, vol. 20, no. 4, pp. 273–287, Apr. 2002.

[12] C. Privitera and L. Stark, "Human-vision-based selection of image processing algorithms for planetary exploration," *Image Processing, IEEE Transactions on*, vol. 12, no. 8, pp. 917–923, 2003.

[13] W. Geisler and J. Perry, "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, 1998, pp. 294–305.

[14] M. S. Banks, A. B. Sekuler, and S. J. Anderson, "Peripheral spatial vision: limits imposed by optics, photoreceptors, and receptor pooling." *J Opt Soc Am A*, vol. 8, no. 11, pp. 1775–1787, Nov. 1991.

[15] T. Buswell G., *How People Look at Pictures: A Study of The Psychology of Perception in Art*. Chicago, USA: The University of Chicago Press, 1935.

[16] N. H. Mackworth and A. J. Morandi, "The gaze selects informative details within pictures," *Perception and Psychophysics*, vol. 2, pp. 547–552, 1967.

[17] L. E. Wixson and D. H. Ballard, "Using intermediate objects to improve the efficiency of visual search," *International Journal of Computer Vision (Historical Archive)*, vol. 12, no. 2 - 3, pp. 209–230, Apr. 1994.

[18] J. M. Henderson, "Object identification in context: the visual processing of natural scenes." *Can J Psychol*, vol. 46, no. 3, pp. 319–341, Sept. 1992.

[19] L. G. Williams, "The effects of target specification on objects fixated during visual search." *Acta Psychol (Amst)*, vol. 27, pp. 355–60, 1967.

[20] C. Privitera and L. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 9, pp. 970–982, 2000.

[21] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 11, pp. 1254–1259, 1998.

[22] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *Image Processing, IEEE Transactions on*, vol. 13, no. 10, pp. 1304–1318, 2004.

[23] A. Torralba, "Modeling global scene factors in attention," *Journal of the Optical Society of America A (Optics, Image Science and Vision)*, vol. 20, no. 7, pp. 1407–1418, July 2003.

[24] H. B. Barlow, *Sensory Communication*. MIT Press, 1961, ch. Possible principles underlying the transformation of sensory messages, pp. 217–234.

[25] D. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of the Optical Society of America A (Optics and Image Science)*, vol. 4, no. 12, pp. 2379–2394, Dec. 1987.

[26] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 1193–1216, 2001.

[27] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," *Network: Computation in Neural Systems*, vol. 10, no. 4, pp. 341–350, 1999.

[28] D. J. Parkhurst and E. Niebur, "Scene content selected by active vision," *Spatial Vision*, vol. 16, no. 2, pp. 125–154, June 2003.

[29] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," *Vision Research*, vol. 42, no. 1, pp. 107–123, Jan. 2002.

[30] L. Itti, "Quantitative modelling of perceptual salience at human eye position," *Visual Cognition*, vol. 14, no. 4 - 8, pp. 959–984, Aug. 2006.

[31] J. H. van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex." *Proc Biol Sci*, vol. 265, no. 1394, pp. 359–366, Mar. 1998.

[32] ASL, "Applied science laboratories, eye tracking system instruction manual, ver 1.2." 1998.

[33] B. W. Tatler, R. J. Baddeley, and B. T. Vincent, "The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task." *Vision Res*, vol. 46, no. 12 (Print), pp. 1857–1862, June 2006.

[34] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, no. 5, pp. 643–659, Mar. 2005.

[35] B. Efron, *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

[36] U. Rajashekar, I. van der Linde, A. C. Bovik, and L. K. Cormack, "Foveated analysis of image features at fixations," *Vision Research*, vol. 47, no. 25, pp. 3160–3172, Nov. 2007.

[37] S. K. Mannan, K. H. Ruddock, and D. S. Wooding, "The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images." *Spat Vis*, vol. 10, no. 3, pp. 165–188, 1996.

[38] M. Posner, R. Rafal, L. Choate, and J. Vaughan, "Inhibition of return: Neural basis and function," *Cognitive Neuropsychology*, vol. 2, no. 3, pp. 211–228, 1985.

[39] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. of 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96)*, 1996, pp. 226–231.

[40] D. S. Wooding, M. D. Mugglestone, K. J. Purdy, and A. G. Gale, "Eye movements of large populations: Ii. deriving regions of interest, coverage, and similarity using fixation maps," *Behavior Research Methods, Instruments, & Computers*, vol. 34, no. 4,1, pp. 509–517, November 2002.

[41] U. Rajashekar, A. Bovik, and L. Cormack, "Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis," *Journal of Vision*, vol. 6, no. 4, pp. 379–386, 2006.

[42] A. Tavassoli, I. van der Linde, A. C. Bovik, and L. K. Cormack, "An efficient technique for revealing visual search strategies with classification images." *Percept Psychophys*, vol. 69, no. 1, pp. 103–112, Jan 2007.

[43] U. Rajashekar, L. Cormack, and A. Bovik, "Point-of-gaze analysis reveals visual search strategies," in *Human Vision and Electronic Imaging IX*, vol. 5292, no. 1. SPIE-Int. Soc. Opt. Eng, 2004, pp. 296–306.

[44] J. Wolfe and T. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Reviews Neuroscience*, vol. 5, no. 6, pp. 495–501, 2004.

[45] V. Mante, R. A. Frazor, V. Bonin, W. S. Geisler, and M. Carandini, "Independence of luminance and contrast in natural scenes and in the early visual system." *Nat Neurosci*, vol. 8, no. 12 (Print), pp. 1690–1697, Dec. 2005.