FOVEATED ANALYSIS AND SELECTION OF VISUAL FIXATIONS IN NATURAL SCENES

Umesh Rajashekar[1], *Ian van der Linde* [2], *Alan C. Bovik*[1], *Lawrence K. Cormack*[1]

[1] Center for Perceptual Systems, The University of Texas at Austin, USA
[2] Department of Computing, Anglia Ruskin University, UK

ABSTRACT

The ability to automatically detect visually interesting regions in images has practical applications in the design of active machine vision systems. Analysis of the statistics of image features at observers gaze can provide insights into the mechanisms of fixation selection in humans. Using a novel foveated analysis framework, in which features were analyzed at the spatial resolution at which they were perceived, we studied the statistics of four low-level local image features: luminance, contrast, center-surround 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. Center-surround contrast showed the greatest difference between human and random fixations, followed by contrast, center-surround luminance, 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 observers.

Index Terms— Visual system, Active vision

1. INTRODUCTION

Despite a large field of view, the human visual system (HVS) processes only a tiny central region (the fovea) with great detail while the resolution drops rapidly towards the periphery. To assimilate visual information and build a detailed representation from this multi-resolution visual input, the HVS uses a dynamic process of actively scanning the visual environment using steady fixations linked by rapid, ballistic eye movements called saccades. Such a *foveated* visual perception provides for a large field of view without the accompanying data glut and has excellent potential for use with artificial vision systems.

While the degradation of spatial resolution in the retina has been modeled accurately by measuring the contrast thresholds of transient stimuli [1], the fundamental question in the area of foveated, active artificial vision of 'How do we decide where to point the cameras next?' has not been thoroughly understood. An understanding of how the HVS 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 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 so many intricate ways that it makes the problem of modeling gaze a formidable task. Since mechanisms that require relatively little image interpretation are likely to be most relevant for current work in artificial vision and automatic visual search, the goal of this paper is to investigate bottom-up, image-based mechanisms that guide eye fixations.

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. In an interesting study, Privitera & Stark [2] 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 neurobiologically inspired model [3], 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. These maps are normalized and combined across scales and features to result in conspicuity maps, whose peaks indentify visually interesting regions. Several modifications that include motion parameters, novel combinations of the feature maps, and modulation by high-level contextual priors have also been developed.

With the availability of inexpensive, accurate eye trackers, a recent trend in the bottom-up approach to understanding gaze has been to quantify the differences in the statistics of image patches at the *point of gaze* of observers and those selected at random. Reinagel *et al.* [4] 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. Other studies

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

have corroborated these findings [5].

While the 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 maximum resolution (of the stimulus). A moment of introspection suggests that, analysis of fixation attractors *must* involve a foveated framework, where low-level image features that attract subsequent fixations are derived solely based on the information obtained by the visual system from its periphery (whose resolution varies across the visual field).

In this paper, we sought insight into the influence of four local image features: luminance, contrast, and center-surround outputs of luminance and contrast in the selection of image regions by the HVS. In particular, 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 these features at observers' fixations and fixations selected at random. In contrast to previous work, a foveated framework was used to analyze the statistics of image patches at the spatial resolution at which the patch was perceived by the HVS. A simple algorithm that selects image regions as likely candidates for fixation based upon a linear combination of these features is presented.

2. EXPERIMENTAL METHODS

2.1. Stimuli and Tasks

101 simages (1024×768 pixels) containing natural habitats of trees, grass, and water (Fig. 1) were selected from the van Hateren database of calibrated grayscale images. 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 free view each of the images as they desired. All observers commenced viewing the image stimuli from the center of the screen. To encourage observers to scan the entire scene, 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. A total of 29 (24 naïve) adult human volunteers with normal or corrected-to-normal vision participated in this study.

2.2. 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 compulsorily every 10 images, and a calibration test run after every image.

2.3. 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. 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 perceived by the observer. To achieve this, the image was first foveated at the observer's current fixation and a patch centered at the 'next' fixation was extracted for analysis. We extracted circular patches of diameter 96 pixels ($1.6°$). We have also tested our simulations for other patch diameters ranging from 32 to 192 pixels.

3. COMPUTING IMAGE FEATURES

3.1. Luminance Computation

The mean luminance for an image patch was computed using a circular raised cosine weighting function, w : $\bar{I} = (1/\sum_{i=1}^{M} w_i)\sum_{i=1}^{M} I_i w_i$ where, M is the number of pixels in the patch, I_i is the grayscale value of pixel at location i. The weighting function $w_i = 0.5 * (\cos(\frac{\pi r_i}{R}) + 1)$, where R was the patch radius and r_i, the radial distance of pixel i from the patch center.

3.2. RMS Contrast Computation

For an image patch, a weighted root-mean-squared (RMS) contrast using a circular raised cosine weighting function, w, was computed: $C = \sqrt{(1/\sum_{i=1}^{M} w_i)\sum_{i=1}^{M} w_i \frac{(I_i-\bar{I})^2}{(\bar{I})^2}}$.

3.3. Center-Surround of Luminance

The next image feature that we investigated was the output of center-surround filters operating on the 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, alternatively, Gabor kernels. Given an image patch, $I(\cdot)$, and a Gabor kernel, $Gab(\cdot)$ we computed $G = \max|Gab(\cdot) * I(\cdot)|$, where $*$ corresponds to the convolution operator. Of the many Gabor kernels that can be used to filter an image patch, we used the kernel which best modeled (in a least squares sense) the spatial frequencies where the human patches differed significantly from the random patches.

3.4. Center-Surround of Contrast

Finally, center-surround differences of local image contrast (i.e. contrast of contrast) was used to capture some higher order image structure that is ignored by the luminance Gabors. For example, regions whose central and surrounding regions have the same mean luminance, but different contrast profiles can be captured by this feature. Since we are now interested in the spatial frequency distribution of local image contrast, we used the magnitude of the local image gradient for each pixel as a measure of an extremely local (pixel-level) measure of image contrast. The design of the Gabor kernel and the computation of the filter output proceeds similar to that in Section 3.3 with the difference that the input to the analysis routine is the magnitude of the local image gradient (instead of the patch luminance).

4. RESULTS

The ensemble of image patches around observers' fixation points was then analyzed to determine if the statistics of the four image features discussed above were statistically different from that of an ensemble of image patches that were picked randomly. The ensemble of randomly selected patches was 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.

A consequence of using a foveated analysis framework is that the human and the random ensembles contain image patches that have been blurred to different extents. Thus, there arises a need to perform an eccentricity-based analysis, where patches of similar blur are grouped together and the relevant image feature is analyzed separately for each blur. 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 get to that particular patch. The distribution of these saccade magnitudes were partitioned into 5 bins such that each bin contained the same number of patches (around 6000) and the patches in each bin were analyzed separately.

For each image feature, S, since we were interested in the differences (and not the absolute values) in the image statistics at observers' fixation and randomly selected fixations, we computed the ratio of average patch feature at the observers' fixations ($\overline{S}(e,n)_{pog}$) to the average patch feature for image patches from the image shuffled database ($\overline{S}(e,n)_{rand}$) for each image, n, and then averaged this ratio across the $N(=101)$ images in the database as follows: $\overline{S}(e)_{ratio} = \frac{1}{N}\sum_{n=1}^{N}\frac{\overline{S}(e,n)_{pog}}{\overline{S}(e,n)_{rand}}$. Finally, to evaluate the statistical significance of the image statistic under consideration, we used bootstrapping to obtain the sampling distribution of the mean statistic of interest.

The value of this ratio, $\overline{S}(e)_{ratio}$, for the four image features is plotted as a function of saccade magnitude, e, in Fig. 2. The error bars represent a 95% confidence interval from 200 bootstrap resamples. 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*. While our findings for RMS contrast are in agreement with previously reported results [4, 5], in a related study, we also discovered that using foveated image patches produces significantly higher (statistically) contrast ratios than a non-foveated analysis. Second, by examining the actual values of the ratios, we found that center-surround contrast showed the greatest difference between human and random fixations (maximum ratio of 1.29), followed by contrast (1.12), center-surround luminance (1.11), and luminance (1.04).

5. SELECTING VISUALLY INTERESTING REGIONS

Our analysis shows that image patches selected by the HVS have higher luminance, contrast, and stronger center-surround profiles than randomly selected patches. 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 and creates a foveated image around this 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 96 pixels. Saliency maps for the center-surround outputs are obtained using five Gabor kernels (one per saccade bin) obtained using the procedure described in Section 3.3. The filtering process 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. The filtered output is interpreted simply as a likelihood map in which regions with large values are more likely to draw a fixation than 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 from Fig. 2. The algorithm selects the maximum value from this weighted selection map as the next fixation point, foveates the image around this point, and repeats this process. Inhibition of return was incorporated to avoid selecting previously selected regions.

Fig. 3 illustrates, qualitatively, the performance of the fixation selection algorithms for two images from Fig. 1 (row 1, columns 1 and 2). For visualization purposes, the fixations of 29 observers on these images were clustered using a density-constrained clustering algorithm. The top ten clusters with the maximum density of fixations are shown as ellipses in Fig.

3. 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. 3. The full width at half-max (FWHM) of the Gaussian roughly equaled the diameter of the human foveola (about 1° visual angle). A good overlap between the ellipse (observers' true fixations) and the bright regions in the selection map (points selected by the algorithm) shows that the fixation selection algorithm is able to select fixations with reasonable success. We also used the Kullback-Leibler Distance (KLD) to quantify the distance between the selected and recorded fixations. First, each algorithmically selected fixation was replaced by a 2D Gaussian (FWHM = 1°) and each fixation cluster by a 2D Gaussian whose shape was determined by the shape of the ellipse. The KLD between these two maps was computed and is shown in Fig. 4 for 15 images. The top bar represents the KLD between randomly generated fixations and the recorded fixations, and indicates the minimum performance expected from any fixation selection algorithm. The remaining bars show the performance of the linearly combined feature maps and each individual features. All image features perform better than a random searcher, with the combined feature map producing the best correlation to recorded fixations.

In summary, despite their rapidity and sheer volume, fixations of human observers are not deployed randomly. We found that image regions selected by human fixations tend to have, on average, higher luminance, contrast, and center-surround profiles than patches selected at random. The technique presented here can be extended to analyze the distribution of higher order features such as orientation,texture, and structure.

6. REFERENCES

[1] 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.

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

[3] 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.

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

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

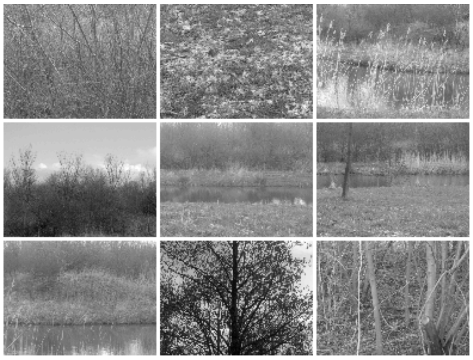

Fig. 1. Examples of images used for the experiment

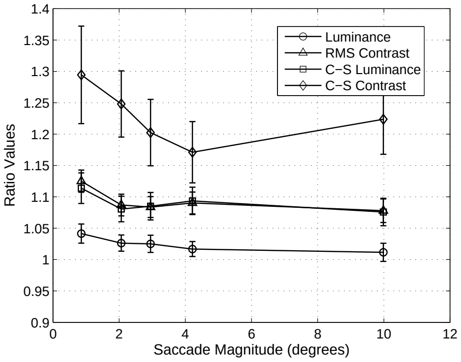

Fig. 2. Plots of $\overline{S}(e)_{ratio}$ as a function of saccade magnitude

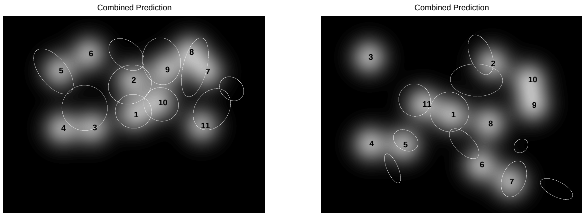

Fig. 3. Comparing algorithmically selected fixations (bright regions) with clusters of human fixations (ellipses) for two images from Fig. 1 (row 1, columns 1 and 2).

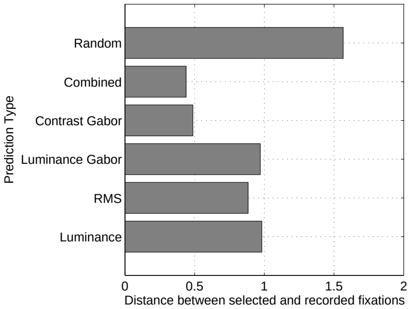

Fig. 4. KLD between selected and recorded fixations.