Visual Search: Structure from noise

Umesh Rajashekar
umesh@ece.utexas.edu
Dept. of Elec. and Comp. Eng.

Lawrence K. Cormack
cormack@psy.utexas.edu
Dept. of Psychology

Alan C. Bovik
bovik@ece.utexas.edu
Dept. of Elec. and Comp. Eng.

The University of Texas at Austin, Austin, TX 78712-1084, USA

Abstract

In this paper, we present two techniques to reveal image features that attract the eye during visual search: the *discrimination image paradigm* and *principal component analysis*. In preliminary experiments, we employed these techniques to identify image features used to identify simple targets embedded in $1/f$ noise. Two main findings emerged. First, the loci of fixations were not random but were driven by local image features, even in very noisy displays. Second, subjects often searched for a component feature of a target rather that the target itself, even if the target was a simple geometric form. Moreover, the particular relevant component varied from individual to individual. Also, principal component analysis of the noise patches at the point of fixation reveals global image features used by the subject in the search task. In addition to providing insight into the human visual system, these techniques have relevance for machine vision as well. The efficacy of a foveated machine vision system largely depends on its ability to actively select 'visually interesting' regions in its environment. The techniques presented in this paper provide valuable low-level criteria for executing human-like scanpaths in such machine vision systems.

CR Categories: I.5.0 [Computing Methodologies]: Pattern Recognition—General I.4.7 [Computing Methodologies]: Image Processing and Computer Vision—Feature MeasurementFeature representation

Keywords: Discrimination Images, Principal Component Analysis, Visual Search, Eye movements, $1/f$ noise

1 Introduction

The eyes are not like cameras in that, despite a large field of view, only a tiny central region is processed in detail. The decrease in resolution from the fovea towards the periphery is attributed to the distribution of the ganglion cells on the retina. The ganglion cells are packed densely at the center of the retina (i.e. the foveola), and the sampling rate drops almost quadratically as a function of eccentricity. In order to build a detailed representation of the image, the human visual system therefore uses a dynamic process of actively scanning the visual environment using discrete fixations linked by saccadic eye movements. The eye gathers most information during the fixations while little information is gathered during the saccades (due to saccadic suppression, motion blurring, etc.).

Not surprisingly, there has been significant interest in investigating image features that attract the human eye. A few reported studies on automatic visual search have examined fixation selection based on features like contrast, edges, object similarity [Moghaddamand and Pentland 1995] or combinations of randomized saliency and proximity factors [Klarquist and Bovik 1998]. These ideas however are based on high level intuition. [Privitera and Stark 2000] propose a computational model for human scan paths based on intelligent image processing of digital images. The crux of their methodology is to identify image-processing algorithms that mimic the eye in detecting points of interest. Their basic idea is to define algorithmic regions of interest (aROI) generated by the image processing algorithms and compare the results with human regions of interest (hROI). The comparison of the aROI and hROI is accomplished by analyzing their spatial/structural binding (location similarity) and temporal/sequential binding (order of fixations). The results indicate that the fixation point prediction coherence is about 0.54 for different subjects looking at the same image i.e. about half the predictions made are accurate. Another approach to analyze of regions-of-interest is to investigate the statistics of some simple image features like contrast and pixel intensity correlation at the point of gaze. Exploiting these statistics of images to predict fixation points seems to be a promising direction since the eye evolved using these statistics and the visual neurons may be optimized for their inputs. It has been demonstrated [Reinagel and Zador 1999] that subjects tend to fixate high-contrast regions and that the intensities of nearby image pixels at the fixation regions are less correlated than in image regions selected at random i.e. the eye fixates on regions rich in spatial structure. Another plausible reason is that this reflects the attempt of the eye to maximize the information it can gather at each fixation [Barlow 1961].

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. The development of foveated artificial vision systems, depends on the ability to model the eye movement mechanisms that automatically determine areas of interest in the image. Thus, a fundamental question in the emerging field of foveated, active artificial vision is therefore 'How do we decide where to point the cameras?' Early work [Zelinsky 1996; Kowler et al. 1995]on the determination of gaze emphasized cognitive factors and, while interesting, was not scientific in that it did not produce theories that could make accurate predictions in novel situations, and certainly did not provide the basis for camera movement algorithms in artificial visual systems. Obviously, such a theory is needed in order to understand biological vision and it is also, by definition, the most fundamental component of any foveated, active artificial vision system. The instantiation of automatic fixation models into the next generation of efficient, foveated, active vision systems can then be applied to a diverse array of problems including automated pictorial database query and data mining; image understanding; automated visual search in, for example, cancer detection, autonomous vehicle navigation; and real-time, foveated video compression [Lee 2000].

The human visual system has evolved multiple mechanisms for controlling gaze. These mechanisms differ in the amount of image processing and interpretation they require, and the relative importance of each of them is situation-dependent. Since mechanisms that require relatively little image interpretation are likely to be most relevant for current work in artificial vision, our goal is to develop an image-based theory of human eye movements to isolate and understand the data-driven mechanisms that guide eye movements. In this paper we present novel applications using two statistical techniques: the discrimination image paradigm and principal component analysis, to extract fundamental image features used in a search task. In our approach, we record human eye movements in a visual search task in which subjects look for targets embedded in noise. Image patches at the subject's point of gaze are then extracted from the noise background to create a bank of image patches that the subject found 'interesting.' We then use the statistical image analysis techniques mentioned earlier to extract image properties that are most common in these interesting patches. Our approach is unique in that we exploit statistics inherent in the noise image patches to reveal what the eye finds interesting. This approach has two fundamental advantages. First, it extracts low-level features derived directly from the linear contribution of each stimulus pixel in attracting gaze. Second, since the stimuli are composed of random noise, there are no high-level features of cognitive or emotional interest to interfere with the image-based mechanisms determining gaze position.

The paper is organized as follows. In Section 2 we discuss in detail the discrimination image paradigm and the principal component analysis techniques. Section 3 discusses the experimental methodology. Section 4 describes the results obtained using the data analysis routines and finally Section 5 concludes by summarizing the results obtained .

2 Algorithms for Data Analysis

2.1 Discrimination Image Paradigm (DIP)

The discrimination image paradigm was originally developed to determine exactly what information was being used in simple visual discriminations [Beard and Ahumada, Jr. 1998]. The idea is to embed in a discrimination task a sufficient amount of visual noise so that the overall signal-to-noise ratio, and hence the outcome of the discrimination task, is largely determined by the external added noise. This task is repeated many times with different added noise on each trial. The noise from each trial when the observer makes a given response is then saved and averaged together. Over many trials, the resulting 'discrimination image' represents the linear contribution (or weight) of each pixel in determining that particular response from the subject.

For illustration purposes, assume that two bars should ideally be in vertical alignment but, are in practice always offset one way or the other as shown in the left panel of the Fig. 1. On each trial, the bars are embedded in noise to limit performance, randomly offset top-leftward or -rightward (center panel of figure), and then briefly presented to the subject. The discrimination image paradigm is designed to reveal image features the human visual system uses to decide if the bars are shifted right or to the left. If the subject responded 'rightward' or 'leftward,' the noise for that trial is averaged into the 'right' or 'left' image, respectively. At the end of the experiment (generally 10,000 trials run over several sessions), definite filter properties begin to emerge in the two average images. Finally, the images are differenced and thresholded for statistical significance (pixels within 2 standard deviations of the mean can be set to gray, for example). A discrimination image from a nearly identical experiment (one using a windowed-sinusoid instead of a sharp bar) is shown in the rightmost panel. The subject seems to be using elongated, vertical, odd-symmetric 'filters' sensitive to the horizontal shift of the bars to make the decision, which clearly reflects a plausible stratagem for this task. We extend this basic methodology to eye movements for the first time.

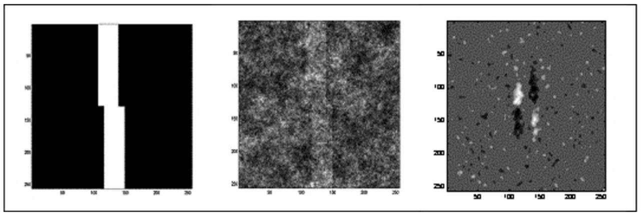

Figure 1: Discrimination Images for vertical bar set up

2.2 Principal Component Analysis (PCA)

PCA [Duda et al. 2000] is a technique for extracting inter-pixel relationships. It is also referred to as the Hotelling transform [Hotelling 1933] or the Karhunen-Loeve transform [Jayant and Noll 1984]. The main idea behind PCA is to represent maximum information (in the minimum mean square sense [Duda et al. 2000]) about a given data set using the least number of uncorrelated linear descriptors: the principal components. The principal components are found by projecting the data set onto a new set of orthogonal bases vectors. Given a set of observations of the random column vector $\vec{x}$, it can be shown [Jayant and Noll 1984] that the orthogonal basis vectors are given by the eigenvectors obtained by the eigenvalue decomposition [Strang 1988] of the correlation matrix $C = \vec{x}\vec{x}^t$ where $\vec{x}^t$ is the transpose of $\vec{x}$. The eigenvalues corresponding to the eigenvectors represent the variance captured by each vector. The new orthonormal basis vectors thus found can be ordered according to the variance captured by each basis vector so that the component that accounts for the most variation in the data is represented first and hence captures the fundamental structure of the data set. PCA has been used for image analysis in face recognition [Turk and Pentland 1991b] and natural image statistics [Hancock et al. 1992].

To better understand the use of PCA in image feature extraction consider the following illustration. In Fig. 2, the left hand panel shows four synthetic images from a set of 40. Each has a vertical Gabor patch of fixed phase and a lower-amplitude horizontal Gabor patch of variable phase embedded in noise. The middle panel shows the first four components generated by the PCA, and the right panel shows their associated weights. As can be readily seen, the PCA was quite effective at extracting out the underlying functions, with the phase-varying Gabor represented by the second and third components in roughly quadrature phase such that a linear combination could yield a Gabor of any phase. The fourth and remaining 36 components are basically noise and have correspondingly low weights.

While there are many techniques to compute PCA, one of the simplest is to compute the eigenvalue decomposition as described before. Assume that we have T observations of an N- dimensional random variable $\vec{x}$: $X = [\vec{x}_1, \vec{x}_2, \vec{x}_3, ...\vec{x}_T]$. The $N*N$ covariance matrix $C = XX^t$ can become intractable for vector dimensions that we are concerned with. For example, a $64*64$ image patch produces a covariance matrix with 2^{20} entries. Hence, a simplified way of calculating the eigenvalue decomposition is adopted [Turk and Pentland 1991a]. Assuming that the number of observations T is usually less than the dimensions of the sample, there will be only T, instead of N meaningful eigenvectors. Therefore, the principal components are computed by first finding the

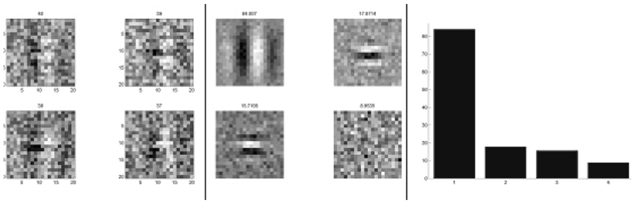

Figure 2: PCA for Gabor patches example

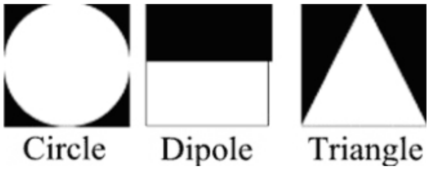

Figure 3: Examples of targets used for DIP

eigenvectors $V = [\vec{v_1}, \vec{v_2}, \vec{v_3}, ...\vec{v_T}]$ of the $T * T$ covariance matrix $L = X^t X$. The eigenvectors $U = [\vec{u_1}, \vec{u_2}, \vec{u_3}, ...\vec{u_T}]$ corresponding to C are represented as a linear combination of the input vectors given by $U = XV$.

3 Methods

3.1 Observers

Three observers, two of them familiar with the experiments and one naive subject, were used for the experiment. Two of the subjects were corrected for normal vision.

3.2 Stimuli and Tasks

The experiments used synthetic images of targets embedded in noise, and the subject's task was simply to find the target. In our preliminary experiments, we have been using simple targets such as circles, dipoles and triangles as show in Fig. 3. The noise we used had a Fourier amplitude that was inversely proportional to the frequency, since this mimics the average spectrum of natural images [Field 1987] and thus making it an effective type of noise for obscuring (or 'masking') targets. Such noise is generally referred to as '$1/f$ noise.' The size of the target was $64 * 64$ pixels and that of the noise matrix was $640 * 480$ pixels. The MATLAB psychophysics toolbox [Brainard 1997; Pelli 1997] was used for stimulus presentation.

The subject was shown a target and instructed to search for the target in each subsequent stimulus display. Blocks of 50 trials with the target embedded randomly in $1/f$ noise backgrounds were used. 10 different patterns of $1/f$ noise were selected randomly during each block of trials to discourage the subject from remembering the structure from previous noise stimuli presentations. The signal-to-noise ratio was set such that the subject generally made many fixations (~ 20) to find the target. On finding the target, the subject pressed a button and proceeded to the next image. Periodic verifications (every 10 trials) of the calibration was done by displaying a dot on the display at the position of gaze in real-time and, if necessary, recalibration was done (although this was rarely required).

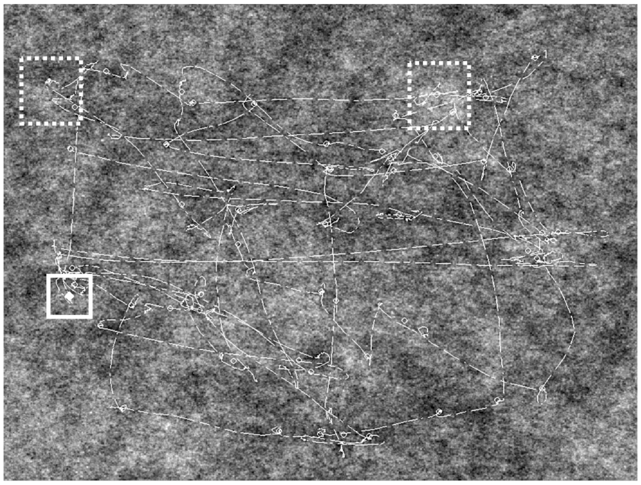

Figure 4: Example scan path while searching in $1/f$ noise

3.3 Eye Tracking

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

A bite bar and forehead rest was used to restrict the subject's head movement. A 21-inch monitor with a gamma corrected display was used to display the stimulus at a distance of 180cm from the subject. The screen resolution was set to $640 * 480$ corresponding to about 34 pixels/degree of visual angle.

The subject was first positioned in the eye tracker and a positive 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.

3.4 Image data acquisition

The sampled voltages from each trial were converted to gaze position on the image. 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 [Applied Science Laboratories 1998]. The resulting patterns for a single trial are shown in Fig. 4. Eventually, the subject found (or thinks they found) the target, an example of which is outlined by the solid box in Fig. 4 (It was a dipole, and is very difficult to see on this particular trial). We defined a 'region of interest' (ROI) of $128 * 128$ pixels around each fixation, two examples of which are shown by the dashed boxes. To avoid edge effects, each region was masked by a radially symmetric Butterworth filter shaped window whose fall off was chosen so that it tapered to zero rapidly near the edges of each region. The ensemble of these ROIs around the fixation points were then subjected to DIP and PCA algorithms as discussed in Section 2. MATLAB was used for all offline analysis.

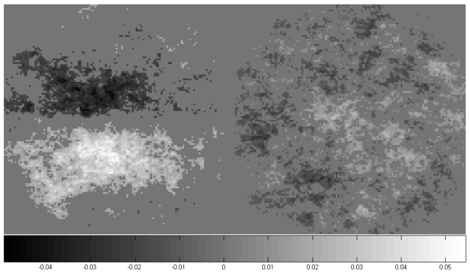

Figure 5: Discrimination Images for Dipole search

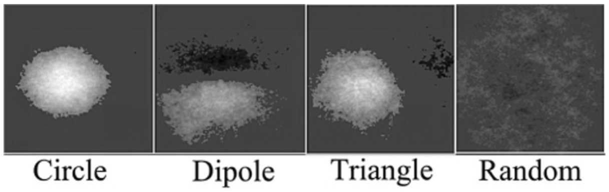

Figure 6: Discrimination images for all targets

4 Results

4.1 DIP on fixation regions

To form a discrimination image, all the images in the ROI ensemble were averaged and thresholded for statistical significance. The resulting discrimination image for a dipole search is shown in the left panel in Fig. 5. Gray denotes a value of zero, white corresponds to positive values and black to negative values. The right panel shows the result of selecting an equal number of randomly positioned ROIs for comparison. Clearly, this image is tending towards an image with no specific structure. This discrimination image represents the feature that, when seen in the periphery of the visual field, draws the gaze for closer inspection. The observer, unlike the random fixation case, seemed to attend to a small, central portion of dipole, perhaps weighting the lower white portion more. Fig. 6 describes the discrimination images for additional targets. For the case of a circle, the subject seemed to be fixating at points which have a bright region with a dark background while for the triangle, the subject seemed to be searching for a white region and the sharp diagonal right edge of the triangle. What makes this technique truly intriguing is that the structure discovered by the DIP algorithm is obtained from the noise structure alone. Since the target features vary in their position in each fixation patch (the subject need not fixate exactly at the same spot in the feature), it is possible that many interesting features in the images are getting swamped in the averaging process. In the following section, we describe the results of applying PCA, which looks for global image properties and hence can potentially reveal more image structure.

4.2 PCA on fixation regions

Before computing the PCA, the columns of each ROI were concatenated to convert the matrix into a column vector. The algorithm described in 2.2 was used to compute the basis vectors. Shown in Fig. 7 are the results of a PCA for two different targets on the same visual search task described before. The eigenvalues shown in the bottom panel of Fig. 7 correspond to the eigenvectors shown in he upper panel and were used to select and order the first 15 significant principal components. In addition to being interesting in their own right, they provide a good illustration of the usefulness of the eigenvalues. The data on the left were generated by the subject searching for the circle in noise, while those on the right were generated by searching for the dipole. The first principal component reflects the structure of the mask used and is of no computational significance to us. The first bar on the eigenvalue plot therefore corresponds to the second eigenvector and not that of the mask extracted by the first component. A glance at the eigenvalues for the circle reveals that the second and third components are about evenly weighted, indicating that edges at all orientations (i.e. linear combinations of the first two components) were about equally attractive to the subject. The eigenvalues for the dipole, however, show a marked preference for horizontal edge information which means that, even in the periphery where visual acuity is poor, the visual system was actively seeking out potential edges, rather than just searching for a bright or dark blob, or casting the eyes about randomly in hopes of fortuitously acquiring the target.

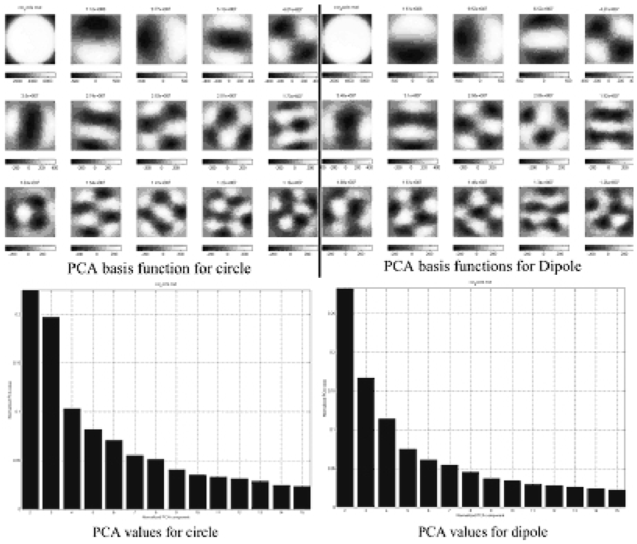

Figure 7: Comparing PCA results for circle vs. dipole

5 Discussion

We have clearly shown, as a proof of principle, the effectiveness of discrimination images as a novel and powerful way of investigating visual search tasks. PCA was also used to illustrate the use of this familiar statistical technique at points of fixation. The emergence of structure from noise is truly intriguing and gives an insight into what an observer might be looking for while searching for targets. The selection of $1/f$ noise is instrumental in this experiment. While most of the DIP type of experiments [Beard and Ahumada, Jr. 1998] need about 12,000 or so trials, the structure in the $1/f$ noise made it possible to reveal structure in a matter of a few thousand fixations.

Both of the above techniques, discrimination image and PCA, share the following feature: their outputs can be used as linear kernels with which to filter input images. The result of this filtering can be considered a likelihood map in image space that reflects the probability of the eye fixating on any given pixel. This likelihood can be used to probabilistically predict human fixation patterns, both alone and in conjunction with other known rules of viewing, and these predictions will be tested in further experiments as we to continue to refine our models.

It should be emphasized again that in the discrimination image paradigm, we need not confine ourselves to averaging in the pixel domain. Much more could be learned by, for example, deriving 'discrimination spectra.' Consider the case in which the subject was searching for a complex target or, alternatively, either of two targets which, if averaged, would cancel each other out (the above dipole and it's negative for example.) In this case, averaging the Fourier amplitude spectra rather than the pixels themselves would probably yield more informative results.

Principal component analysis, while elegant in its own right, does not capture local image features [Bell and Sejnowski 1996]. We are investigating the application of a more recent tool: Independent Component Analysis [Hyvärinen et al. 2001] to extract fundamental structure at points of gaze both in search tasks like the one described in this paper and in free gazing of natural scenes.

Overall, we feel that even though we are just beginning to apply the PCA and DIP type analysis to the specific search task described above the results are very promising. With a unique combination of eye tracking capability and image analysis tools we have been able generate some very interesting preliminary results, which may reflect low-level features used in search tasks. This line of research with more controlled experiments might help reveal results that will be fundamental to the design of active artificial foveated machine vision systems.

References

APPLIED SCIENCE LABORATORIES. 1998. Eye tracking system instruction manual. Ver 1.2.

BARLOW, H. B. 1961. *Possible principles underlying the transformation of sensory messages*. M.I.T. Press, Cambridge MA, 217–234.

BEARD, B. L., AND AHUMADA, JR., A. J. 1998. A technique to extract relevant image features for visual tasks. *SPIE Proc. Human Vision and Electronic Imaging III Vol. 3299*, 79–85.

BELL, A. J., AND SEJNOWSKI, T. J. 1996. Learning the higher-order structure of a natural sound. *Network: Computation in Neural Systems 7*, 2.

BRAINARD, D. H. 1997. The psychophysics toolbox. *Spatial Vision 10*, 433–436.

DUDA, R. O., HART, P. E., AND STORK, D. G. 2000. *Pattern Classification*, Second ed. Harcourt Brace Jovanovich, San Diego, November, ch. 3, 114–117.

FIELD, D. J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12), 2379–2394.

HANCOCK, P. J. B., BADDELEY, R. J., AND SMITH, L. S. 1992. The principal components of natural images. *Network 3*, 61–70.

HOTELLING, H. 1933. Analysis of a complex of statistical variables into principal components. *J. Educational Psychology 27*, 417–441.

HYVÄRINEN, A., KARHUNEN, J., AND OJA, E. 2001. *Independent Component Analysis*, 1 ed. John Wiley & Sons, May.

JAYANT, N. S., AND NOLL, P. 1984. *Digital coding of waveforms : principles and applications to speech and video*. Prentice-Hall, Englewood Cliffs, New Jersey, ch. 12, 535–546.

KLARQUIST, W., AND BOVIK, A. C. 1998. Fovea: a foveated vergent active stereo system for dynamic three-dimensional scene recovery. *IEEE Tran. on Robotics and Automation 14*, 5 (October), 755–770.

KOWLER, E., ANDERSON, E., DOSHER, B., AND BLASER, E. 1995. The role of attention in the programming of saccades. *Vision Research 35*, 1897–916.

LEE, S. 2000. *Foveated Video Compression and Visual Communications over Wireless and Wireline Networks*. PhD thesis, The University of Texas at Austin, Austin,TX.

MOGHADDAMAND, B., AND PENTLAND, A. 1995. Probabilistic visual learning for object detection. *Fifth Int. Conf. Computer Vision* (June), 786–793.

PELLI, D. G. 1997. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision 10*, 437–442.

PRIVITERA, C. M., AND STARK, L. W. 2000. Algorithms for defining visual regions-of-interest: comparison with eye fixations. *IEEE Trans. on Pattern Analysis and Machine Intelligence Volume: 22*, Issue:9 (Sept), 970–982.

REINAGEL, P., AND ZADOR, A. M. 1999. Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems 10*, 1-10.

STRANG, G. 1988. *Linear Algebra and its Applications*. Harcourt Brace Jovanovich, San Diego.

TURK, M., AND PENTLAND, A. 1991. Eigen faces for recognition. *J. Cognitive Neuroscience 3* (March), 71–86.

TURK, M., AND PENTLAND, A. 1991. Face recognition using eigenfaces. *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, 586–591.

ZELINSKY, G. J. 1996. Using eye saccades to assess the selectivity of search movements. *Vision Research 36*, 14 (July), 2015–2228.