IMAGE FEATURES THAT DRAW FIXATIONS

Umesh Rajashekar[1], *Lawrence K. Cormack*[2], *Alan C. Bovik*[1]

[1] Laboratory for Image and Video Engineering, Dept. of Elec. and Comp. Eng.
[2] Department of Psychology
The University of Texas at Austin, Austin, TX 78712-1084, USA

ABSTRACT

The ability to automatically detect 'visually interesting' regions in an image has many practical applications especially in the design of active machine vision systems. This paper describes a data-driven approach that uses eye tracking in tandem with principal component analysis to extract low-level image features that attract human gaze. Data analysis on an ensemble of image patches extracted at the observer's point of gaze revealed features that resemble derivatives of the 2D Gaussian operator. Dissimilarities between human and random fixations are investigated by comparing the features extracted at the point of gaze to the general image structure obtained by random sampling in monte-carlo simulations. Finally, a simple application where these features are used to predict fixations is illustrated.

1. INTRODUCTION

Despite a large field of view, the human visual system (HVS) processes only a tiny central region (the fovea) with very great detail while the resolution drops rapidly towards the periphery. As a result, to build a detailed representation of a scene from this multi-resolution input, the human eye uses rapid eye movements called *saccades* to actively scan the environment. To assimilate visual information, the human eye therefore uses a combination of steady eye *fixations* linked by ballistic saccades. This active interaction with the visual scene clearly has promising advantages in both speed and storage requirements when extended to active machine vision systems.

Artificial active vision systems find applications in a diverse array of problems such as automatic pictorial database query, image understanding, and automated visual search in, for example, cancer detection and autonomous vehicle navigation. However, 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 addressed. This paper presents an approach to answer the question by attempting to understand and model human eye fixations in natural viewing tasks.

Competing theories for gaze prediction can be classified into two general categories - top-down and bottom-up. Top-down approaches for gaze prediction emphasize a high-level understanding of the scene and has been popular in task specific experiments (visual search). The rapidity and sheer volume of eye movements, however, pose serious challenges to this theory. The human eye makes an average of 18,000 fixations in an hour. It seems implausible that the HVS uses semantic scene information to make a majority of these fixations. Further, cognitive interpretation of scenes is far from being sufficiently mature to warrant the use of these theories for gaze prediction in natural viewing tasks.

The second theory of gaze prediction is an empirical bottom-up approach. Here, eye movements are believed to be quasi-random and driven by low-level image structure. Since the HVS evolved in a natural environment and natural images occupy a relatively small subspace of all possible images, it is theorized [1] that early visual processing may exploit the statistics inherent in its environment to represent the input as efficiently as possible. Approaches supporting this theory propose a computational model for human gaze prediction based on image processing that accentuates image features that are deemed relevant. In an interesting study, Privitera & Stark [2] used a suite of algorithms such as detecting the presence of symmetry, center surround regions in images that resemble receptive field profiles, wavelets, contrast, and edges- per-unit-area to predict points on interest in an image and compared these predictions with human eye fixations. They report that $43\% - 54\%$ of their predictions match human fixations.

A more recent version of the bottom-up approach is based on natural scene statistics. In one reported work [3], the statistics of natural images at point of gaze were compared to the statistics of random patches from the same image sets. Their results show that the human fixation regions have higher spatial contrast and spatial entropy than the corresponding random fixation regions indicating that the human eye may be trying to select image regions that help maximize the information content transmitted to the visual cortex by minimizing the redundancy in the image represen-

0-7803-7750-8/03/$17.00 ©2003 IEEE

tation. These results seem to indicate that eye movements are possibly drawn to image features that are signatures of naturally occurring scenes and motivates the approach used in this paper.

Using a combination of eye tracking and image analysis, this paper presents an image-based theory of human eye movements to isolate and understand the data-driven mechanisms that guide eye movements. In our approach, we recorded human eye movements in a free-viewing scenario where observers viewed a database of natural and man-made images. Image patches at the observer's point of gaze were then extracted to create a bank of image patches that the observer found 'interesting.' We then subjected this bank to Principal Component Analysis (PCA) to extract the most 'common' image features that drew the observer's fixation. Such a generic data-driven approach is powerful because, unlike the top-down and bottom-up approaches, it reveals low-level fixation attractors without making any assumptions about what image features should be analyzed.

This paper is organized as follows. In Section 2 we describe the experimental methodology. Section 3 presents the results and Section 4 presents the conclusions and some future directions.

2. EXPERIMENTAL METHODS

2.1. Observers

Six observers, three of them familiar with the experiments and three naive observers, were used for the experiment. All observers either had normal or corrected-to-normal vision.

2.2. Stimuli and Tasks

The stimuli consisted of images from the '*Natural*' database consisting of landscapes and the '*Man-made*' database with scenes consisting of urban scapes. Figs. 1(a) and 1(b) illustrate some of the images from the '*Man-made*' and '*Natural*' sets respectively. All images were $640 * 480$ pixels and displayed on a 21 inch monitor with screen resolution at $640 * 480$ pixels at a distance of $180cm$ from the observer. This set up corresponded to about 52 pixels/degree of visual angle.

Observers were presented with about 100 images from each database in separate sessions and instructed to view the displayed image until they were confident of being able to describe the scene.

2.3. 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 and a response time of under $1ms$. This spatio-temporal resolution is much better than the popular video-based eye trackers and is essential for resolving any phase-sensitive image components in subsequent analysis. 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.

2.4. Data Analysis-PCA

The eye movement trajectories (hereafter referred to as scanpaths) were first classified into fixations and saccadic eye movements using spatio-temporal criteria derived from the known dynamic properties of human saccadic eye movements. A typical eye scanpath for a single trial is shown in Fig. 6(a) as a dotted line with white dots indicating fixations. A region of interest (ROI) of $32 * 32$ pixels around each fixation was extracted from the image. To avoid any edge effects, each image patch was masked with a radially symmetric mask that tapered to zero rapidly towards the edges. The set of all ROIs for an observer for a given image database ('*Natural*'/'*Man-made*') was used to create an image patch data bank which was analyzed using PCA.

PCA (also referred to as the Hotelling transform or the Karhunen - Loeve transform) [4] is a technique for extracting inter-pixel relationships in a data set. PCA is often used in dimensionality reduction to represent maximum information (in the minimum mean square sense) about a given data set using the least number of uncorrelated linear descriptors: the principal components (computed as the eigenvectors of the covariance matrix of the data set). Each eigenvector has a corresponding eigenvalue that represents the variance (in the data set) captured by that vector. Once computed, these orthonormal eigenvectors are ordered according to their eigenvalues 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.

3. RESULTS

Fig. 2 illustrates the results of applying PCA to an image bank consisting of about 3000 image patches extracted from the fixations of an observer using the '*Man-made*' database. Fig. 2(a) shows the first 15 eigenvectors ordered from left to right and top to bottom in descending order of their corresponding eigenvalues (Fig. 2(b)). The first eigenvector simply represents the average of all the image patches and is not of much significance. The first bar on the eigenvalue

plot therefore corresponds to the second eigenvector shown.

To compare and contrast the PCA results obtained by actual human fixations to those obtained by random image sampling, PCA was performed again on image patches selected at random from the image database taking care to keep the number of patches the same as that obtained from the human fixations. Statistical differences in results are thus indicators of how observers' fixations differ from those obtained by randomly sampling the image. Fig. 3 shows the results corresponding to the random fixation scenario for the '*Man-made*' database. The error bars on the eigenvalues represent 1 standard deviation variation obtained by monte-carlo simulations where different random image patches were selected for each simulation. Figs. 4 and 5 illustrate similar results for the '*Natural*' database.

A glance at these eigenvectors indicate that they resemble derivatives of 2D Gaussians. The lower order vectors (2 and 3) resemble vertical and horizontal edge detectors while components (4-6) resemble bar detectors at different orientations (the actual sign of the eigenvectors being irrelevant). However, there are many interesting differences across these cases as described below. The following discussions focus on the second and third vectors for ease of comparison.

3.1. Discussion for '*Man-made*' database

To get an idea of the general image structure in the '*Man-made*' database, we will first make some inferences from Fig. 3 which was obtained by random sampling of the images. Observing the second eigenvector in Fig. 3(a), it is clear that horizontal edges are the most dominant image features in the '*Man-made*' database followed by the vertical edge. Comparing their corresponding eigenvalues (Fig. 3(b)), the contribution of the horizontal edges is almost twice that of the vertical edge. However, in contrast, the second eigenvector for the human fixations in Fig. 2(a) is vertical indicating that this particular observer actually preferred vertical edges despite the abundance of horizontal edges in the database. Another interesting difference is that unlike the random fixation case, second and third eigenvectors for human fixations have similar weights as seen in Fig. 2(b). That is, the bank of image patches labelled by the human eye as 'interesting' have similar contributions from both the horizontal and vertical edges.

Compared across observers, the second and third eigenvectors take orientations rotated away from the horizontal axis by different amounts indicating a desire to fixate at regions that are not commonplace in the image database. A separate analysis of the average patch orientations confirmed these results indicating that the orientation specificity of eigenvectors indeed reflect image content.

3.2. Discussion for '*Natural*' database

Comparing the PCA results in Fig. 4 and Fig. 5, the eigenvectors and the relative magnitudes of the eigenvectors for the human versus random fixation are very similar. Horizontal structures therefore seem to be the most common features in natural images and the ones that draw human fixation (unlike the vertical edges for '*Man-made*' scenes). It is probably incorrect at this point to deduce that human fixations are random because, comparing the absolute magnitudes of the eigenvalues in Fig. 4(b) and Fig. 5(b) the strength of the eigenvalues for human fixations are statistically much larger (greater than 1 standard deviation) than that for the random indicating that the human eye samples more horizontal and vertical edges than that selected by randomly sampling the image. The results appear to be consistent across observers for the '*Natural*' database.

3.3. Application - Predicting visually interesting regions

Since the eigenvectors obtained by PCA of human fixations capture the 'visually interesting' image features they can be used as filter kernels to predict regions of interest in an image. Fig. 6(a) shows an image with an observer's eye scanpath superimposed as a white dashed line and the white dots indicating fixations. Fig. 6(b) shows the results of convolving the image using the second to fifth eigenvectors from Fig. 2 weighted according to their respective eigenvalues and then adding the absolute value of the result of each of the kernels. This result can be interpreted as a likelihood map that reflects the probability (represented as intensity) that a pixel will draw a human fixation. A relatively good overlap between the bright regions and the true fixations (indicated by white dots) in Fig. 6(b) suggests that the PCA kernels are able to pick out some visually interesting regions in the image.

4. CONCLUSIONS

In this paper we use a combination of eye tracking and principal component analysis to extract low-level image features that attract human fixations in a scene. The results of the analysis indicate that humans are not random in their decision of where to fixate next but rather seem to have preferences to certain derivative of Gaussian-like low-level image structures that are idiosyncratic across observers but clearly deviant from the general image structure. One of the major drawbacks of PCA is that it exploits only linear correlation. We are extending our analysis to include the more powerful Independent Component Analysis [5] which can exploit even non-linear dependencies in the data. Also, we are currently investigating quantitative measures to compare the predictions of interesting regions by the kernels with the fixations of a human observer.

5. REFERENCES

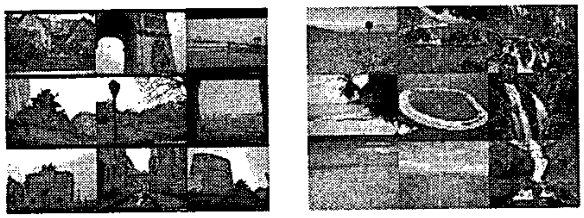
[1] Horace B. Barlow, *Possible principles underlying the transformation of sensory messages*, pp. 217–234, M.I.T. Press, Cambridge MA, 1961.

[2] Claudio M. Privitera and Lawrence W. Stark, "Algorithms for defining visual regions-of-interest: comparison with eye fixations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. Volume: 22, no. Issue:9, pp. 970–982, Sept 2000.

[3] Pamela Reinagel and Anthony M. Zador, "Natural scene statistics at the center of gaze," *Network: Computation in Neural Systems*, vol. 10, no. 1-10, 1999.

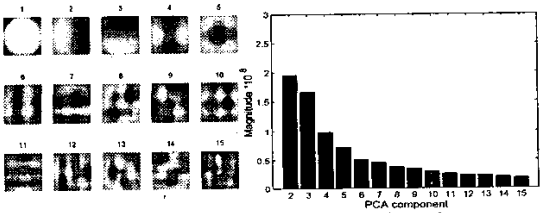
[4] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, chapter 3, pp. 114–117, Harcourt Brace Jovanovich, San Diego, Second edition, November 2000.

[5] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja, *Independent Component Analysis*, John Wiley & Sons, 1 edition, May 2001.



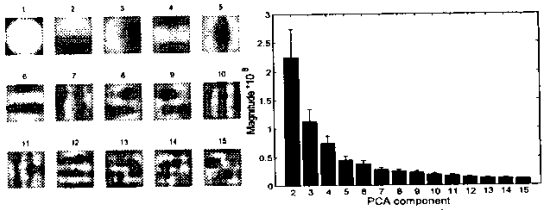
(a) '*Manmade*' (b) '*Natural*'

Fig. 1. Sample Images



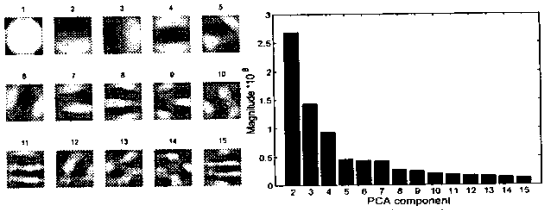
(a) Eigenvectors (b) Eigenvalues

Fig. 2. PCA on '*Man-made*' at fixations



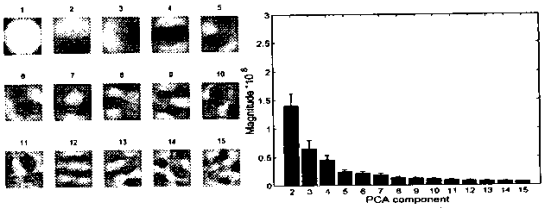
(a) Eigenvectors (b) Eigenvalues

Fig. 3. PCA on '*Man-made*' at random locations



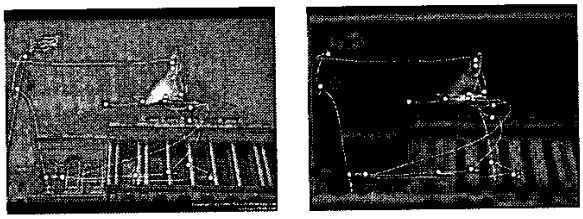
(a) Eigenvectors (b) Eigenvalues

Fig. 4. PCA on '*Natural*' at fixations



(a) Eigenvectors (b) Eigenvalues

Fig. 5. PCA on '*Natural*' at random locations



(a) Image with observer's eye scanpath superimposed (b) Prediction of interesting regions by PCA kernels

Fig. 6. Comparing true fixations with PCA predictions. White dots indicate fixations