Features That Draw Visual Attention: An Information Theoretic Perspective

Neil D. B. Bruce[1,2]

[1] Dept. of Computer Science, York University, Toronto, Canada
[2] Centre for Vision Research, York University, Toronto, Canada

Abstract

A novel image operator is proposed for the purpose of predicting the focus of visual attention in arbitrary natural scenes based on local statistics. The proposed method is based on the hypothetical premise that attention proceeds by way of sampling a scene in a manner that maximizes the information acquired from the scene. A tractable means of computing the joint likelihood of local statistics in a low-dimensional space is presented and shown to have a close relationship to the representation of retinal image stimulus existing in the primary visual cortex of primates.

Keywords: Visual Attention, Information Theory, Feature Extraction, Independent Components.

1. Introduction

There exist numerous models intended for predicting the focus of visual attention that range from purely biological to almost purely mathematical. A large body of literature exists establishing that attention appears to be directed on the basis of properties of the input constituted by the response of photoreceptors at the retina in conjunction with the goals of the observer. It appears then, that attention is stimulus driven with some task dependent bias. A question that naturally follows from this observation is: "What sort of stimulus tends to draw attention from a human observer?" This question is typically considered independent of any particular task, and measured quantitatively by examining the correlation between certain local statistical properties of an image and coordinate locations fixated by a set of human observers when asked to examine the image with no particular instructions given. Privitera and Stark present a review of various algorithmic means of detecting regions of interest as well as comparing predictions of such algorithms with human eye tracking data[1]. Features investigated include edges per unit area based on the Canny operator[2], High curvature masks incorporating varying orientation acute angles and "X" masks, a 7x7 positive-centre/negative-surround operator, Gabor masks to measure grey-level orientation differences[3], the discrete wavelet transform[4], a measure of local symmetry, Michaelson Contrast[5], an intensity based point entropy operator, coefficients of the discrete cosine transform[7], and a Laplacian of Gaussian. Analysis was carried out on a data set produced using 7 human subjects and 15 images. Generally each of the operators investigated showed some correlation to fixation points in the context of some images and virtually no correlation for others. None of the algorithms was suggested as being a universal predictor of regions of interest. It is worth noting that the translation from a continuous feature map to a discrete set of fixation points was performed by selecting only local maxima in the feature maps and clustering of the resulting loci. Such an operation seems somewhat questionable since one is discarding significant information concerning the output of the operator in the vicinity of local maxima (like the height of such peaks) and in areas where no obvious local maxima lie. It also remains unclear whether a data set based on

7 subjects and 15 images is sufficient for a quantitative comparison. These issues are addressed in some capacity in section 4. Other features implicated in attention include contrast, shape, location, and size[7], contrast, curvature, length, and orientation of contours and perimeter, intensity, area and elongation of regions[8] and centre-surround masks applied in luminance, red-green and blue-yellow feature planes[9]. Topper introduced an interesting addition to the attention literature rooted in information theory[10]. The foundation of his work lies in the assertion that it is the uniqueness of the response in a feature domain that determines salience. For example, in a scene composed almost entirely of edges and containing a small flat region, one may be more likely to attend to the small homogeneous region. It is the uniqueness of this region in the edge domain that determines salience and not the raw measure of edge strength. Owing to the close relationship between this observation, and some ideas that arise from information theory, Topper suggested Shannons self information measure[11] as a suitable transformation from a feature plane into a dimension that more closely corresponds to perceptual salience. In the context of this problem, Shannon's self information measure may be described as follows: The information I(**x**) conveyed by an event **x** is inversely proportional to the likelihood of observing **x**. Shannon's proposed relationship may be written explicitly as $I(\mathbf{x}) = -\log(p(\mathbf{x}))$. The reason for choosing this particular relationship lies primarily in the fact that the information conveyed by two independent events is given by the sum of the information afforded by each individual event when using the log operator. In a feature domain, each value of edge strength may be mapped to a value of edge information by considering the likelihood distribution of edge strength over the entire image, which may be determined through a histogram or kernel density estimate[12]. Topper conducted a comparison study along the same lines as that carried out by Privitera and Stark, with a particular emphasis on comparing correlation between feature and information maps and eye tracking data. Topper found that in all cases, the transformation to the information domain improved correlation to the set of eye tracking data. Given the aforementioned considerations and in particular Topper's findings, we explore the premise that visual attention proceeds entirely by way of maximizing the information sampled from an image. That is, in this scheme we are not interested in particular feature measures, but rather, the likelihood of observing the raw set of local statistics. A means of computing this likelihood is outlined which allows the evaluation of the information content of such statistics in a classic information theoretic sense and in a manner that is tractable. We will also demonstrate that the proposed operator is both biologically plausible, and more closely correlated to fixation density than features investigated in each of the aforementioned studies, and in particular, some of the more promising information domain operators considered by Topper.

2. Measuring the Joint Likelihood of Local Statistics

As mentioned, the premise of the proposed approach is that the informativeness of local statistics is a quantity more closely related to attention than many other feature operators that are typically employed in attention models. It is important to stress that unlike Topper, we are not considering the informativeness in a particular feature domain, and on a pixel by pixel basis. Instead, the more general measure of the informativeness of the raw set of statistics in a local neighborhood is considered. Informativeness in a feature domain requires the estimation of a one dimensional probability density function. In the more general case that requires computing a joint likelihood measure based on local statistics, estimation of a $3*M*N$ dimensional probability density function is required for a local window size of $M*N$ in RGB space. In practice, estimation of such high dimensional probability functions proves unfeasible requiring prohibitively large degrees of computation and data. A natural question to consider is whether there exists a feasible method for reducing the dimensionality of this representation to the degree that available data and computational resources allow its estimation. Dimensionality reduction is a problem that arises in many statistical applications and one for which a rich body of literature exists. In particular, principal components and independent components have been applied in a variety of image processing and machine vision applications in the past decade most often aimed at image compression. In this work, we have employed a representation based on independent component analysis for dimensionality reduction based on two considerations: i. A representation based on independent components reduces the problem of estimating a *3*M*N* dimensional probability density function to the problem of estimating *3*M*N* one-dimensional probability density functions. This is accomplished through a transformation of the data that is lossless, and invertible. ii. As is discussed in section 3. this choice yields a representation that has some close ties to neural circuitry of the human visual cortex.

2.1 Independent Representation of Local Statistics

Independent component analysis assumes that the data under consideration was formed from a number of independent sources, combined in an additive manner. To simplify discussion of local statistics, let $x=[x_1,...,x_N]^T$, represent the local statistics in column vector form with x_{3*j*k}, $x_{3*j*k+1}$ and $x_{3*j*k+2}$ coding the red, green and blue values for pixel *j,k* in the original patch. Mathematically, we can express the mixing model for the above setup as $\boldsymbol{x} = \sum_{i=1}^{N} \boldsymbol{a}_i \boldsymbol{s}_i = \boldsymbol{A}\boldsymbol{s}$, where the basis functions $\boldsymbol{a}_i$, $i=1,...,N$ constitute the columns of the *nxN* matrix $\boldsymbol{A}$, and $s=[s_1,...,s_N]^T$, a random vector. To determine the coefficients of $\boldsymbol{A}$, it is necessary to find a linear transformation $\boldsymbol{V}$ of the input data $\boldsymbol{x}$ such that, for some vector $\boldsymbol{w}$, $\boldsymbol{w}=\boldsymbol{V}\cdot\boldsymbol{x}$ with the components of $\boldsymbol{w}$ as statistically independent as possible. $\boldsymbol{w}$ allows an estimate of $\boldsymbol{s}$ for a particular data sample, and $\boldsymbol{A}$ is the pseudo-inverse of $\boldsymbol{V}$. Each $\boldsymbol{w}_i$ is then given by the dot product $\langle V_i, x\rangle = \sum_{m,n} V_i(m,n)x(m,n)$. There are a variety of unsupervised learning algorithms for estimating the set of basis functions given a number of examples of $\boldsymbol{x}$. In our implementation, we have employed Sejnowski and colleagues infomax ICA algorithm[13]. It has been demonstrated that such a representation provides a reasonable means of representing very general non-Gaussian statistics[14] and in a manner that parallels the representation existent in the human primary visual cortex[15]. This last consideration is elaborated on in section 3. The ICA algorithm was applied to a set of 360,000 7x7 image patches chosen randomly from a set of 3600 natural images and at four spatial scales. A linear combination of the resulting basis functions may be used to describe each local neighborhood of any arbitrary image. For a particular basis function $\mathbf{a}_k$, with a value of $\mathbf{v}_k$ corresponding to a particular local neighborhood, it is possible to measure the likelihood $p(\mathbf{a}_k=\mathbf{v}_k)$. This is accomplished through a Parzen window probability density estimate[16] for $p(\mathbf{a}_k)$ based on the coefficients arising from every local neighborhood of the image and corresponding to the basis function $\mathbf{a}_k$. Since the coefficients corresponding to each $\mathbf{a}_k$ are independent variables, $p(\mathbf{a}_1=\mathbf{v}_1 \cap \mathbf{a}_2=\mathbf{v}_2 \cap \ldots \mathbf{a}_k=\mathbf{v}_k \cap \ldots\ \mathbf{a}_n=\mathbf{v}_n) = p(\mathbf{a}_1=\mathbf{v}_1)p(\mathbf{a}_2=\mathbf{v}_2)\ldots p(\mathbf{a}_k=\mathbf{v}_k)\ldots p(\mathbf{a}_n=\mathbf{v}_n) = \prod_{i=1}^{n} p(\mathbf{a}_i = \mathbf{v}_i)$. This likelihood is readily converted to an information measure by considering –log($\prod_{i=1}^{n} p(\mathbf{a}_i = \mathbf{v}_i)$). The overall aforesaid process to derive an information map based on a given image is depicted in figure 1.

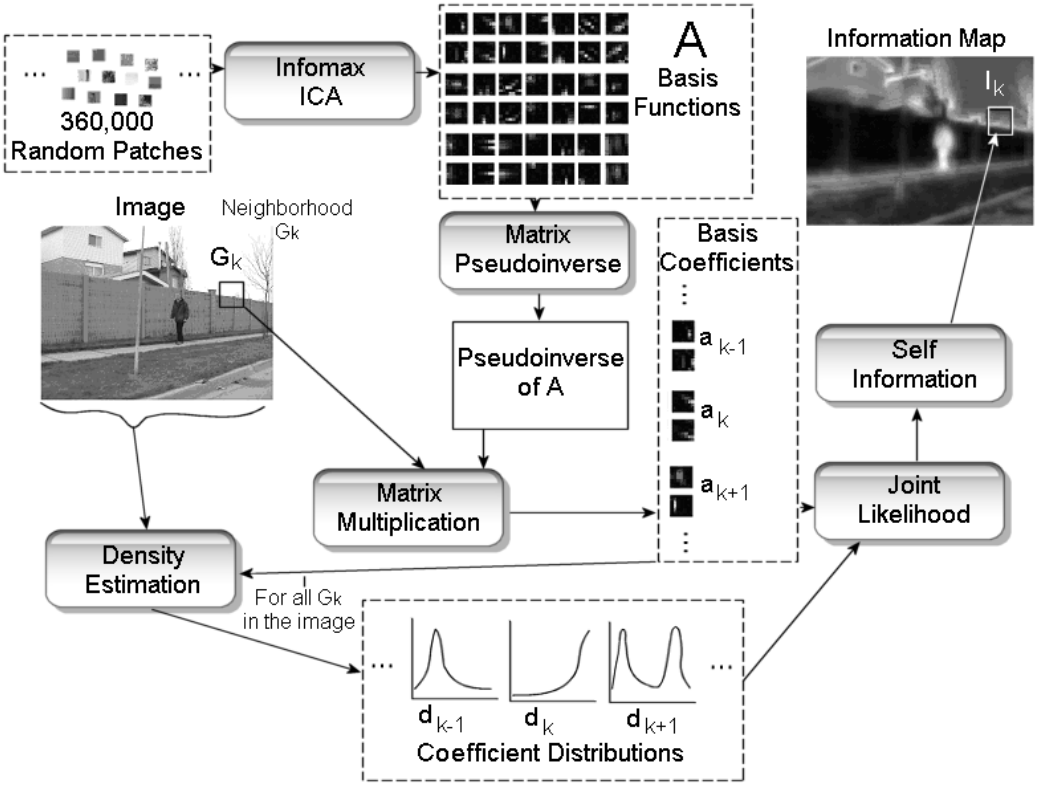

Figure 1. The overall framework for measuring the information afforded by a set of local statistics.

3. Independent Components and V1

An important consideration in designing a framework for simulating a sensory system, is that of biological plausibility. Thus far we have established that independent components allow a representation of local statistics in a manner that allows the joint likelihood to be computed in a manner that is tractable. Although computationally the proposed approach could not be realized without such a representation, there is another more fundamental reason for choosing a representation based on independent components. Olhausen and Field[15] have demonstrated that the response of simple cells in the primary visual cortex may be explained in terms of a strategy for producing a sparse distribution of activity in response to gray-level values in natural images. In particular, they demonstrated that independent component analysis applied to natural gray-level image patches gives rise to localized and oriented edge detectors whose responses closely match spatial receptive field properties of simple cells in V1[17]. Their work has inspired further investigation of minimum redundancy sparse coding of cell responses to other forms of visual input. van Hateren and Ruderman[18] produced basis functions that resemble simple cells in V1 through independent component analysis of spatiotemporal data. Wachtler and Sejnowski[19] and Tailor et al.[20] examined the problem in relation to spatiochromatic data. In each case, they found that ICA of spatiochromatic patches gives rise to independent component filters that consist of luminance and color filters. The luminance filters were very similar to those obtained in the gray-level case examined by Olhausen and Field. Color filters resembled either red-green or blue-yellow opponent cells in V1. All of these studies provide strong evidence that neural circuitry appears to be built on the basis of minimizing redundancy of neuron population responses with regard to the input they process. It also follows that the proposed architecture is built on a representation that is fundamentally very similar to the representation existent in the human primary visual cortex, and as such, is more closely tied to neural circuitry than existing attentional feature measures.

4. On the Interpretation of Fixation Data

Eye tracking data was collected for twenty subjects with normal vision, for a set of 120 images. Each image was presented for four seconds in random order, with a mask presented in between each pair of images for one second. Any image location upon which the eyes of the subject rested for more than 200 ms was deemed a fixation point and recorded. An important consideration in comparing the output of feature detectors with experimental eye tracking data is the manner in which the fixation data is represented. Data from eye tracking experiments typically comes in the form of coordinates of fixations. As mentioned, in this raw form there is no obvious metric for measuring the difference between feature extractor output and discrete fixation coordinates. Further, attention is not necessarily directed to single points but rather, more realistically modeled as directed upon extended regions with visual acuity a maximum at the discrete fixation points. Koesling et al. suggest a means of transforming raw fixation data to fixation density maps to allow direct comparison with feature detector output[21] as follows: For a given image, all of the fixation points from 20 subjects were merged into a single data set. A distribution that approximately models visual acuity dropoff in the human visual system, modeled by a Gaussian distribution with appropriately chosen parameters, is placed at each fixation point. To calculate the fixation density map, the two-dimensional Gaussian distributions as described, are summed over the entire image. This approach provides for each image, a continuous fixation density map based on 20 subjects. In this case, the parameters of the Gaussian were such that one standard deviation lie 20 pixels from the center of a fixation point in each direction. Such a representation allows comparison with feature maps without the need to discard all information outside of local maxima.

4.1 Comparing Fixation Density and Feature Maps

To compare the derived eye tracking density maps with feature maps, a suitable metric is required to measure the difference between the two. The most obvious means of computing the difference between the maps is that of summing the absolute value of the difference between each pixel in the density map and corresponding feature map. This operation is analogous to computing the volume between two probability density surfaces in the continuous case. For simplicity, allowing straightforward comparison with other studies and avoiding the introduction of *ad hoc*, or parametric difference metrics, the aforementioned scheme has been applied in all cases as a natural quantitative measure of the difference between feature extractor output and fixation density. In effect, this metric affords a comparison between the output of each feature operator, and the degree to which each local neighborhood of the image is sampled on average by a human observer.

5. Correlation Between Image Features and Fixation Density

Information maps were produced through the proposed information measure for the full set of 120 images. The difference between resulting information maps and fixation density was measured as outlined in section 4.1 and an average measure recorded for each spatial scale. The same operation was carried out for 6 of the features examined by Topper showing the strongest correlation to eye tracking density. The resulting average difference values are displayed in table 1. Each difference is shown in its raw form and below the raw value in normalized form. Normalization in this case is given by (raw-random)/(2-random) where random corresponds to the score resulting from the average difference between a random probability distribution each corresponding density map, resulting in a value between 0 for a case that is no better than a random probability distribution and 1 for a perfect match. It is clear in examining these values that the information operator based on the joint likelihood stands out in its similarity to fixation density. What is not obvious from the numbers, is how these measures translate to qualitative similarity between feature and density maps and in particular, response in non-salient regions and where local maxima lie. Figures 2 and 3 demonstrate for a number of images, the original image (left), the average of scale 1,2, and 3 information maps (center), and experimental density map (right). Figure 2 contains images that tend to have qualitatively homogeneous backgrounds and figure 3 the converse. It is evident in viewing these figures that there exists significant similarity between the information maps and experimental density maps in terms of response in regions of interest, response elsewhere, and the position of local maxima.

Filter/Scale	**Scale 1**	**Scale 2**	**Scale 3**	**Scale 4**
Random Map	1.4121	1.4121	1.4121	1.4121
	0.0000	0.0000	0.0000	0.0000
Sobel Magnitude	1.3494	1.3353	1.3411	1.3469
	0.1067	0.1306	0.1208	0.1109
Sobel Orientation	1.3485	1.3551	1.3616	1.3616
	0.0636	0.0970	0.0859	0.0859
Intensity	1.3471	1.3524	1.3532	1.3528
	0.1106	0.1015	0.1002	0.1009
Variance	1.3613	1.3697	1.3748	1.3747
	0.0864	0.0721	0.0634	0.0636
Hue	1.3485	1.3490	1.3468	1.3620
	0.1082	0.1073	0.1111	0.0852
ICA Information	1.2875	1.2864	1.2874	1.3389
	0.2119	0.2138	0.2121	0.1245

Table 1: Average difference between various feature maps and their corresponding experimental density maps. This value is presented in raw form (top of each row) and in normalized form (bottom of each row).

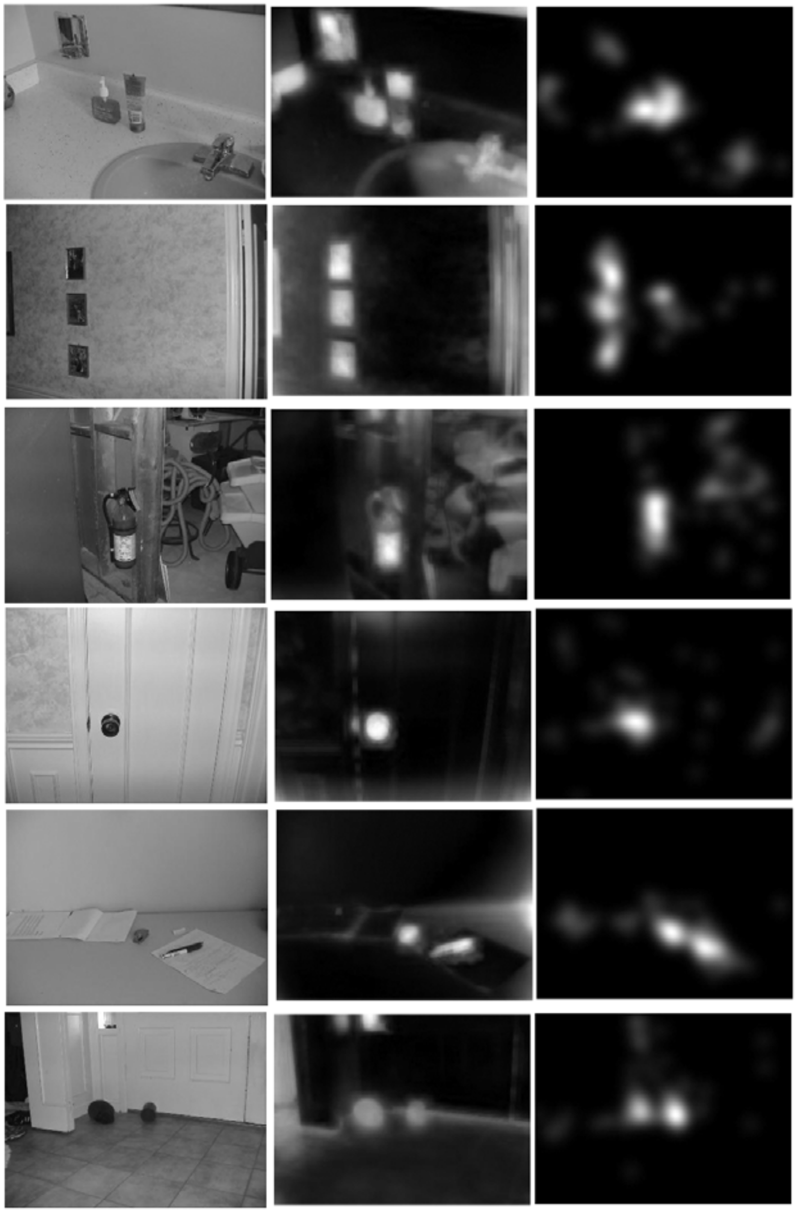

Figure 2: Images from the data set with relatively flat background elements(left), corresponding information maps using the proposed operator(center), fixation density produced from experimental data(right).

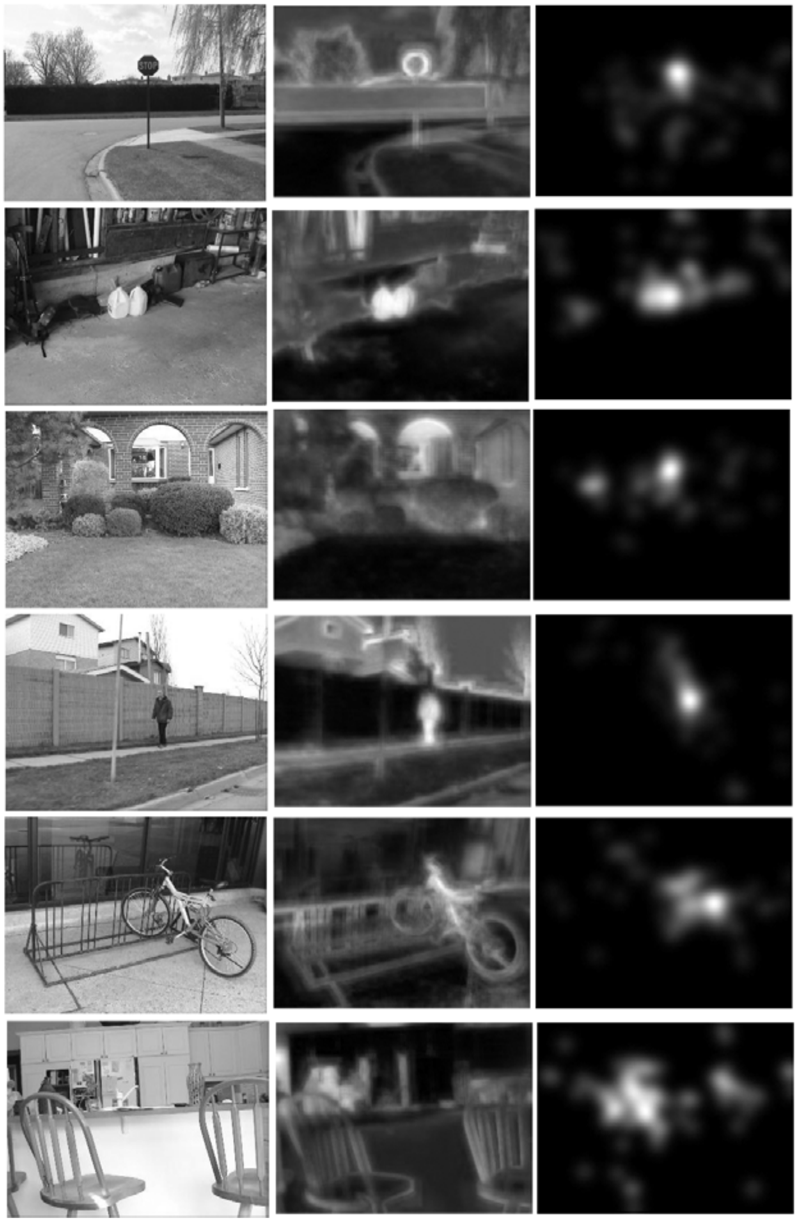

Figure 3: Images from the data set with many edges or significant clutter(left), corresponding information maps using the proposed operator(center), fixation density produced from experimental data(right).

6. Conclusion

A novel means of quantifying the joint likelihood of local statistics was presented through representation based on independent components. The self-information based on this joint likelihood was demonstrated to bear a close resemblance quantitatively and qualitatively to experimental fixation density. In particular, the proposed information measure appears to be more closely tied to fixation density than image operators previously implicated in the focus of attention. The proposed information measure appears to afford a reasonable predictor of regions of interest for any given image using a representation fundamentally connected to the primary visual cortex of primates.

References

[1] C. M. Privitera and L.W. Stark (2000), Algorithms for defining visual regions of interest: Comparison with eye fixations, IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 22, pp. 970-981.

[2] J. Canny, A computational approach to edge detection, IEEE Trans. PAMI (1986), no. 8, pp. 679-698.

[3] D. Dunn and W. E. Higgins, Optimal Gabor filters for texture segmentation, IEEE Trans. Image Processing (1995), vol. 4, pp. 947-964.

[4] I. Debauchies, "The wavelet transform, time-frequency localization and signal analysis," IEEE Trans. Information Theory (1990), vol. 36, pp. 961-1005.

[5] M. D. Levine, Vision in Man and Machine, McGraw-Hill, New York, 1985.

[6] D.J. Bailey and N. Birch, "Image compression using a discrete cosine transform image processor". Electronic Engineering (1999), pp. 39-44.

[7] W. Osberger and A. J. Maeder, Automatic Identification of Perceptually Important Regions in An Image, Proceedings IEEE conference on ICPR (1998), no.1, pp. 701-704.

[8] R. Milanese, J.M. Bost, T. Pun, A bottom-up attention system for active vision, 10th European Conference on Artificial Intelligence (1992), pp. 808-810.

[9] L. Itti, C. Koch, E. Niebur, A model of saliency based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Analysis (1998), no. 11, pp. 1254-1259.

[10] T. N. Topper, Selection Mechanisms in Human and Machine Vision, University of Waterloo, Ph.D. Thesis, 1991.

[11] C. E. Shannon, A mathematical theory of communication, The Bell Systems Technical Journal (1948), no. 27, pp. 93-154.

[12] D. Scott, Nonparametric Probability Density Estimation, Ph.D. Thesis, Department of Mathematical Sciences, Rice University, 1996.

[13] T.W. Lee, M. Girolami, T.J. Sejnowski, Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. Neural Computation (1999), 11, 417-441.

[14] M.S. Bartlett, S. Makeig, A.J. Bell, T.P. Jung, T.J. Sejnowski (1995). ICA of EEG data. Soc. for Neuroscience Abs., 21:437.

[15] D.J. Field, and B. A. Olshausen (1996), Emergence of simple-cell receptive field properties by learing a sparse code for natural images. Nature, 381:607—609.

[16] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, 1973.

[17] B. A. Olhausen, and D. J. Field, Natural image statistics and efficient coding, Network: Computation in Neural Systems (1996), No. 7, pp. 333-339.

[18] J.H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in the primary visual cortex. Proc. R. Soc. Lond. B (1998), 265:359--366.

[19] T. W. Lee, T. Wachtler, T. J. Sejnowski, Color opponency is an efficient representation of spectral properties in natural scenes.Vision Res (2002), 42(17):2095-103.

[20] D. Tailor, L. Finkel, G. Buchsbaum, Color Opponent Receptive Fields Derived from Independent Component Analysis of Natural Images, Vis. Res. (2000), no. 40.

[21] H. Koesling, E. Carbone, H. Ritter, Tracking of Eye Movements and Visual Attention, University of Bielefeld Technical Report, 2002.