EVOLUTIONARY DESIGN OF CONTEXT-FREE ATTENTIONAL OPERATORS

Neil D. B. Bruce and M. Ed Jernigan

Department of Systems Design Engineering
University of Waterloo, Ontario, Canada, N2L 3G1

ABSTRACT

A framework for simulating the visual attention system in primates is presented. Each stage of the attentional hierarchy is chosen with consideration for both psychophysics and mathematical optimality. A set of attentional operators are derived that act on basic image channels of intensity, hue, and orientation to produce maps representing perceptual importance of each image pixel. The development of such operators is realized within the context of a genetic optimization. The model includes the notion of an information domain where feature maps are transformed to a domain that more closely represents the response one might expect from the human visual system. The model is applied to a number of natural images to assess its efficacy in predicting guidance of attention in arbitrary natural scenes.

1. INTRODUCTION

The visual attention system in primates appears to employ a serial computational strategy when processing complex visual scenes [1]. Certain areas of interest in the scene are selected based on behavioral significance or on local image characteristics. Despite the perception that we see everything around us, a relatively small portion of information gathered by the human visual system actually influences human behavior [2]. The human visual system has been well studied but remains far from being fully understood. That said, a great deal of evidence now exists in support of a two-component system consisting of a bottom-up primitive component where attention is guided purely by image stimuli and a slower top-down component where attentional selection is guided under cognitive control [1]. In this paper, we present a computational strategy to simulate fast bottom-up attentional selection in primates. The existence of such a system has been suggested as a necessary component in making general vision tractable in real-time [3].

One of the first neurally credible frameworks for simulating human visual attention was proposed by Koch and Ullman [4]. Their model focused on the idea of a 'saliency map', defined as a two-dimensional topographic representation of conspicuity. Their proposed model consisted of 4 key steps: Low-level feature extraction, centre-surround differences to produce feature maps, combination of those feature maps to produce a unique topographical saliency map, and finally, attentional selection and inhibition of return. Futher investigation of this model has taken place in the last 15 years including close examination of various components of the model by Koch, Ullman and additionally Niebur and Itti [5]. Some of the ideas that come out of the Koch and Ullman framework contribute to the work presented in this paper and are discussed in more detail in the section that follows.

Another well-known study on the issue of computational visual attention is that of Privitera and Stark [6]. Privitera and Stark evaluated numerous algorithmic approaches to detecting regions of interest by comparing the output of such algorithms to eye tracking data captured using standard eye tracking equipment. Privitera and Stark compared 10 different algorithmic methods for detecting regions of interest. The measures that they investigated include such measures as edge strength, high curvature, center surround response, gabor masks, wavelet transforms, symmetry, and contrast. Privitera and Stark found that each of the 10 operators showed a strong correlation to measured fixations for some of the images but performed quite poorly for others.

Topper introduced an interesting addition to the visual attention literature rooted in information theory [7]. The premise of his work is as follows: Strength of a particular feature in an image locality does not in itself guarantee that ones attention will be drawn to that image region. For example, in an image that has a high degree of variance throughout most of the image, one may well be more likely to attend to more homogenous regions of the image. Detectors based on strength in variance or edges would fail miserably in such a case. A more realistic approach would involve detecting parts of the scene that are most different from the rest of the scene. Topper's idea was to transform feature maps to a more perceptually relevant domain through an operator that quantifies the uniqueness of measured feature strengths. Owing to the close ties between this premise and ideas that come out of information theory, Topper suggested Shannon's measure of self information as an appropriate transform for this purpose. Shannon's measure of self information may be

described in the context of the visual attention framework as follows:

Let F be a given feature strength that has a probability of occurrence $P(F)$ in the image, and let $I(F)$ represent the amount of information gained when one learns that F has occurred. Then $I(F) = -\log(1/P(F))$ where $P(F)$ is given by a histogram density estimate on F. The information operator ensures unique feature strengths (a localized region with unusual hue for example) receive a large confidence value in the information domain.

Topper performed a set of experiments along the same lines as those of Privitera and Stark. He measured the correlation of information maps to eye tracking density maps following the application of Shannon's self information measure to feature maps. As in the case of Privitera and Stark, the correlation for each operator was substantial in some cases and worse in others. A key difference though, was that the addition of the information operator made for a more robust detector, able to deal with images where strength in a particular feature was not the primary indicator of where attention might be drawn.

Tompa introduced an approach to computational visual attention based on a subset of the measures employed by Topper for which the correlation to density maps was seen to be particularly strong [8]. The information maps derived from this feature subset were then integrated by means of a few elementary operators to derive an overall perceptual importance map.

Tompa's model involves three key components: The first component is the derivation of feature maps from the original image. The 6 operators used in Tompa's approach are Sobel edge magnitude, Sobel edge orientation, intensity, hue, variance, and moment of inertia. These measures were observed to have the strongest correlation to eye tracking results in Topper's work when combined with the self-information operator. The second stage consists of combining the information maps to arrive at a final importance map. Tompa evaluated various simple approaches at this stage including taking the average, sum of squares, minimum, and maximum of the 6 maps. The sum of squares operator was found on average to provide the best results.

It is clear that a variety of different approaches have been taken to deal with simulating the human visual attention system. One might notice that all of these models seem to have common elements. All of them involve some form of low-level extraction of features on the image. Most involve some transformation from these measured feature maps to a domain that more closely resembles a representation of perceptual relevance. Another common component is the combination of maps representing importance to produce an overall saliency map. There appears to be a fundamental similarity between many of the existing models regardless of whether they were derived through psychophysical principles or under strictly mathematical considerations. This observation provides the motivation for the model that is developed in this work. The proposed model may be viewed as an abstraction of existing approaches with choices for various components made bearing in mind both psychophysical and mathematical considerations.

2. THE MODEL

The proposed framework encompasses ideas from a number of the approaches mentioned in the previous section and consists of 4 key components:

1. An early feature extraction phase in which the initial RGB image is divided into an intensity channel, a hue channel and 4 orientation channels using oriented Gabor filters as is the case in the Koch and Ullman model. The choice of these channels allow us to arrive at any of the operators employed in Tompa's study by applying a relatively simple nonlinear operator to one of the 3 channels. Tompa's approach may then be viewed as a specific case of the model presented in this paper. Koch and Ullman also provide strong psychophysical evidence in support of using these basic channels.

2. Nonlinear filtering with operators intended to respond (when coupled with the Shannon information operator) to signal patterns that tend to draw attention from human observers. These operators are found through stochastic search of a function space consisting of quadratic Volterra filters of local extent. The intention of searching the function space is to locate unknown operators in the space that exhibit even stronger correlation to eye tracking density maps under the premise that such operators exist. It is clear why a measure such as variance might hint at areas that will draw attention but it is expected that some other operators chosen from a function space that includes many of Tompa's choices and designed specifically for focusing attention might do far better. The structure of a quadratic Volterra filter is as follows:

$$g(x,y) = h_o + \sum_{i,j \in S} h_1(i,j) f(x-i, y-j) + \sum_{i,j,k,l \in S} h_2(i,j,k,l) f(x-i, y-j) f(x-k, y-l)$$

with S the local extent support region of the filter. The h coefficients determine the nature of the filter and are the parameters that are chosen through the course of the GA optimization. The function that measures the effectiveness of a particular operator is:

$$C = \sum_{1}^{n} SI(g * I_n) - D_n$$

where C represents cost, g the local extent quadratic Volterra filter being assessed, I_n image n in the test set,

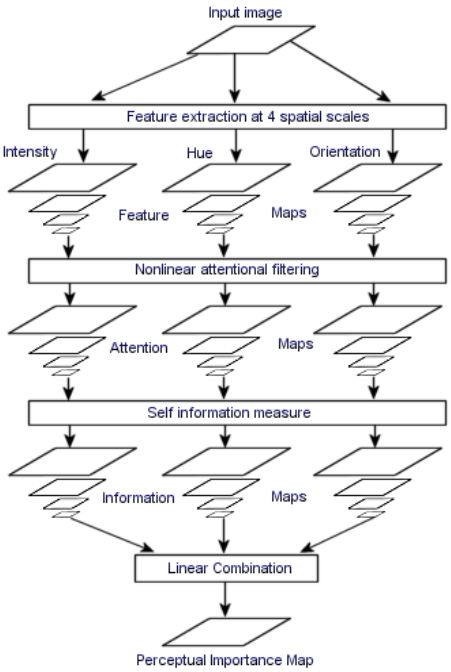

Fig. 1. Overview of the proposed attentional hierarchy

SI Shannon's self-information measure, and D_n the experimental density map corresponding to the image I_n. The optimization then seeks to find the function g that minimizes C. This optimization is performed for one channel and at one resolution at a time to produce an attentional transformation at each scale and for each channel.

Experimental density maps were produced for 120 images based on measured fixations of 20 subjects. Images were presented in random order and shown for 4 seconds each. Attention is not focused on distinct mathematical points, but rather on extended regions. A density map was computed for each of the images as the sum of 2-D Gaussian distributions centered at each fixation across all subjects as in [9].

3. The information operator employed in Topper's work that takes each higher-level map to a domain that more accurately reflects the response one might expect from the human visual system. This stage is similar to the centre surround difference in the Koch and Ullman model.

4. Combination of the information maps derived in step 3 to arrive at an overall perceptual importance map. The overall importance map is produced by averaging the individual information maps. The prediction in one of the 3 channels is generally quite good. Also, a strong peak in one of the 3 channels results in a corresponding strong peak in the combined map. For this reason, it is expected that not much would be gained from employing a more complicated fusion procedure. This scheme is also consistent with the psychophysical observation that attention is guided by within feature spatial competition [10]. An overview of the proposed architecture is shown in figure 1. It is clear that there exist similarities between this approach and existing models. Key distinguishing characteristics include the training of custom attentional operators and inclusion of the information operator that Topper proposed.

3. RESULTS

The proposed model was applied to a wide variety of test images consisting of indoor and outdoor scenes and including images ranging from a few areas of interest to a larger number. Fig. 2. demonstrates the working of the model on an image of a storefront. In this particular example, correct prediction of areas of interest relies primarily on the hue and orientation information maps.

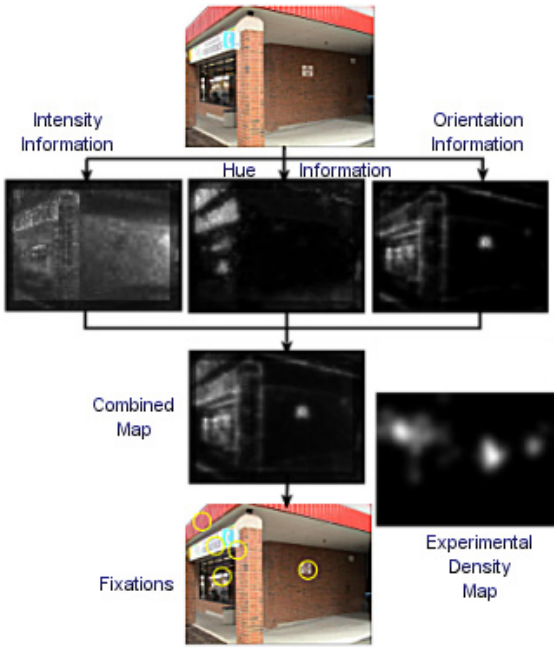

Fig. 2. Example application of the model. Shown are (Top to bottom) the original image, information maps, the combined map and the original with fixations superimposed

Fig. 3. Fixations selected by the proposed model for a variety of images. Fovea sized regions that contain the most confidence are selected and inhibited in the combined map until at least 50 percent of the confidence is suppressed.

It is seen here that the averaging preserves the peaks that exist in the individual maps. Fovea sized regions that contain the most confidence in the combined map are selected and inhibited (initialized) as in [5] to determine a sequence of fixations. These fixations are indicated by yellow circles superimposed on the image. In each case shown in figures 2 and 3, fixations are selected until at least 50 percent of the confidence in the combined map has been inhibited. Figure 3 presents the final result indicating the fixations predicted by our model for a number of different images. In each image tested, most of the key distractors in the image were selected by our model.

4. DISCUSSION

In this paper, we have proposed a new framework for simulating the visual attention system in primates. Unlike existing approaches, we have designed nonlinear operators explicitly for the purpose of responding to image stimulus that might draw attention from human observers. The model is demonstrated to afford detection of features based on a variety of different types of stimuli as well as contending with the issue of scale. Predictions of the trained operators correlate closely with fixation points present in experimental results. Future work will include closer analysis of the problem of combining information maps.

5. REFERENCES

[1] J.R. Bergen, and B. Lulesz, "Parallel versus serial processing in rapid pattern discrimination," *Nature*, pp. 696-698, 1983.

[2] D. J. Simmons, and D.T. Levin, "Failure to detect changes to attended objects", *Investigative Opthalmology and Visual Science*, *38*, *3273*, 1997.

[3] J.K. Tsotsos, S.M. Culhane, W.Y. Wai, Y.H. Lai, N. Davis, and F. Nuflo, "Modeling visual-attention via selective tuning", *Artificial Intelligence*, *78*, pp. 507-545, 1995.

[4] C. Koch, and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, 4, 219-227, 1985.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency based visual attention for rapid scene analysis," *IEEE Transactions PAMI*, 20, pp. 1254-1259, 1998.

[6] C. Privitera, and L. Stark, "Algorithms for Defining Visual Region-of-Interest: Comparison with Eye Fixations," *IEEE PAMI Transactions. 22* (9): pp. 970-982, 2000.

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

[8] D. Tompa, "Perceptual Importance Maps for Visual Attention," M.A.Sc. Thesis, University of Waterloo, 2002.

[10] M. Pomplun, H. Ritter, B. Velichkovsky, "Disambiguating Complex Visual Information: Towards Communication of Personal Views of a Scene", *Perception*, 25(8), 931-948, 1996.

[11] A.M. Sillito, and H.E. Jones, "Context-dependent interactions and visual processing in v1," *Journal of Physiology Paris*, *90*, pp. 205-209, 1996.