Picture Coding Symposium 2001, 25-27 April 2001, Seoul, Korea.

An estimator for visual attention through competitive novelty with application to image compression

Fred Stentiford
BTexaCT Research,
Adastral Park,
Martlesham Heath,
Ipswich, UK
fred.stentiford@bt.com

ABSTRACT

Existing models of visual attention have provided plausible explanations for many of the standard percepts and illusions and yet all have defied implementations that have led to generic applications. This paper describes a new measure of visual attention and its application to variable resolution compression mechanisms.

1. INTRODUCTION

Most higher animals in the world have an ability to sense danger by spotting anomalies in their environment and surviving by taking appropriate evasive action. Those organisms that have the benefit of vision are able to direct attention rapidly towards the unusual without any prior knowledge of the environment.

Grossberg's Adaptive Resonance Theory [1] holds that attention is strongly linked to resonances with previously learnt features. If input sense data is too different from any previously learned prototype then the learning of a new category is initiated. This means that the novelty contained in a scene is judged according to how well it fits a bank of stored 'templates'. Grossberg develops this theory and shows how it accords with many physiological observations.

Osberger et al [2] identified perceptually important regions by first segmenting images into homogeneous regions and then scoring each area using five intuitively selected measures. The approach was heavily dependent upon the success of the segmentation and in spite of this it was not clear that the method was able to identify important features in faces such as the eyes.

Luo et al [3] also devised a set of intuitive saliency features and weights and used them to segment images to depict regions of interest. The integration of features was not attempted. Itti et al [4] have defined a system which models visual search in primates. Features based upon linear filters and centre surround structures encoding intensity, orientation and colour, are used to construct a saliency map that reflects areas of high attention. Supervised learning is suggested as a strategy to bias the relative weights of the features in order to tune the system towards specific target detection tasks.

Walker et al [5] suggested that object features that best expose saliency are those which have a low probability of being mis-classified with any other feature. Saliency in an image is indicated in those areas that contain distinctive and uncommon features. The method relies upon deriving feature statistics from a training set of similar images such as faces. Mudge et al [6] also considered the saliency of a configuration of object components to be inversely related to the frequency that those components occur elsewhere.

Studies in neurobiology [7] are suggesting that attention is enhanced through a process of competing interactions among neurons representing all of the stimuli present in the visual field. The competition results in the selection of a few points of attention and the suppression of irrelevant material. It means that people and animals are able to spot anomalies in a scene no part of which they have seen before and attention is drawn in general to the anomalous object in a scene, not to the common-or-garden or the familiar. It seems unlikely that such novelty can be sharply identified by processes which are totally dependent upon comparisons with a growing dictionary of independent recognition codes. Indeed past a certain size of dictionary, sense data corresponding to novel events will almost always match elements already existing within the dictionary.

This paper follows the thinking of Walker et al [5] and proposes a measure of visual attention that depends upon the dissimilarity between neighbourhoods in the image. There are also links with the use of fractals to segment images according to their self similarity [8]. However, the approach taken in this paper makes no use of any a priori intuitively constructed features or any training process that depends upon the arbitrary selection of training data and learning algorithms. In this way it is hoped to avoid the exclusion of whole categories of images by the constraints imposed by specific rules of analysis or the use of preordained templates. An ideal visual attention system has to cope with anomalies and backgrounds that have never been seen before.

2. INFORMATION ATTENTION FRAMEWORK

Let a set of measurements $\mathbf{a}$ correspond to a location $\mathbf{x}$ in bounded n-space ($x_1, x_2, x_3, \ldots, x_n$) where

$$\mathbf{x} = (x_1, x_2, x_3, \ldots, x_n) \text{ and } \mathbf{a} = (a_1, a_2, a_3, \ldots, a_p)$$

Define a function $\mathbf{F}$ such that $\mathbf{a} = \mathbf{F}(\mathbf{x})$ wherever $\mathbf{a}$ exists. It is important to note that no assumptions are made about the nature of $\mathbf{F}$ eg continuity. It is assumed that $\mathbf{x}$ exists if $\mathbf{a}$ exists.

Consider a neighbourhood N of $\mathbf{x}$ where

$$\{\mathbf{x}' \in \text{ N iff } |x_i - x'_i| < \varepsilon_i \ \forall \ i\}$$

Select a set of m random points $S_\mathbf{x}$ in N where

$$S_\mathbf{x} = \{\mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_3, \ldots, \mathbf{x}'_m\} \text{ and } \mathbf{F}(\mathbf{x}'_i) \text{ is defined.}$$

Select another location $\mathbf{y}$ for which $\mathbf{F}$ is defined.

Define the set $S_\mathbf{y} = \{\mathbf{y}'_1, \mathbf{y}'_2, \mathbf{y}'_3, \ldots, \mathbf{y}'_m\}$ where

$$\mathbf{x} - \mathbf{x}'_i = \mathbf{y} - \mathbf{y}'_i \text{ and } \mathbf{F}(\mathbf{y}'_i) \text{ exists.}$$

The neighbourhood of $\mathbf{x}$ is said to match that of $\mathbf{y}$ if

$$|F_j(\mathbf{x}) - F_j(\mathbf{y})| < \delta_j \text{ and } |F_j(\mathbf{x}'_i) - F_j(\mathbf{y}'_i)| < \delta_j \ \forall \ i,j.$$

In general δ_j is not a constant and will be dependent upon the measurements under comparison ie.

$$\delta_j = f_j(\mathbf{F}(\mathbf{x}), \mathbf{F}(\mathbf{y}))$$

A location $\mathbf{x}$ will be worthy of attention if a sequence of t neighbourhoods matches a small number of other neighbourhoods in the space.

$$1\,1\,0\,1\,0\,0\,1\,\underset{x}{\boxed{1}}\,0\,\boxed{1}\,0\,\boxed{0\,1}\,1\,0\,1\,0\,0\,1\,1\,0\,1\,0\,0\,1\,\underset{y}{\boxed{1}}\,0\,\boxed{1}\,0\,\boxed{1\,1}\,1\,0\,1$$

Figure 1 Neighbourhood at x mismatching at y.

For example, in the case of a one-dimensional sequence of binary digits, m bits are selected in the neighbourhood of a bit x [Figure 1]. Each of the bits possesses a single value of 1 or 0, so F(x') = a = 1 or 0. The neighbourhood of a second bit y elsewhere in the stream matches the first if the parity of all the m + 1 corresponding bits are equal ($\delta_j = 0$). It would be expected that high mismatching scores would be obtained for all x if the binary sequence was random [9]. Bits located within repetitive sequences (eg 110011001100) would obtain low attention estimates, but any occasional irregularities would be highlighted.

In the visual case of a two-dimensional still image, m pixels $\mathbf{x}'$ are selected in the neighbourhood of a pixel $\mathbf{x}$ [Figure 2]. Each of the pixels might possess three colour intensities, so $\mathbf{F}(\mathbf{x}') = \mathbf{a} = (r, g, b)$.

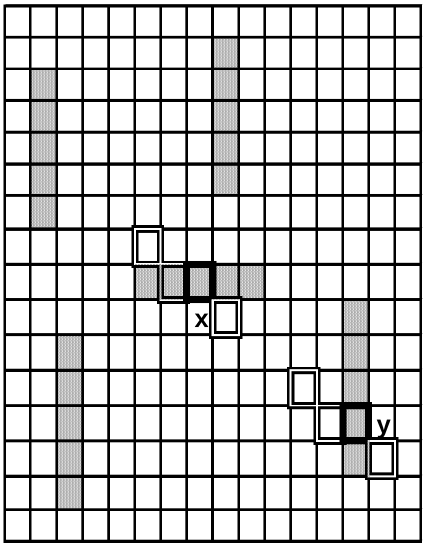

Figure 2 Neighbourhood at **x** mismatching at **y**.

The neighbourhood of a second pixel **y** matches the first if the colour intensities of all m + 1 corresponding pixels have values within δ of each other. Pixels **x** that achieve high mismatching scores over a range of t neighbouring pixel sets $S_\mathbf{x}$ and pixels **y** are assigned a high estimate of visual attention. This means that pixels possessing novel colour values that do not occur elsewhere in the image will be assigned high visual attention estimates. It also means that neighbourhoods that span different coloured regions (eg edges) are naturally given high scores if those colour adjacencies only occur rarely in the scene.

3. IMPLEMENTATION

The visual attention estimator has been implemented as a set of tools that process images and sets of images and produce corresponding arrays of attention scores and compressed versions of the image. The scores are thresholded and those above the threshold are displayed using a continuous spectrum of false colours with the maximum scores being marked with a distinctive colour.

The estimator is based upon the principle of detecting and measuring differences between neighbourhoods in the image. Such differences are recorded and the scores incremented when pixel configurations are selected that do not match identical positional arrangements in other randomly selected neighbourhoods in the image. The gain of the scoring mechanism is increased significantly by retaining the pixel configuration $S_\mathbf{x}$ if a mismatch is detected, and re-using $S_\mathbf{x}$ for comparison with the next of the t neighbourhoods. If however, $S_\mathbf{x}$ subsequently matches another neighbourhood, the score is not incremented, and an entirely new configuration S_x is generated ready for the next comparison. In this way competing configurations are selected against if they contain little novelty and turn out to represent structure that is common throughout the image. Indeed it is likely that if a mismatching pixel configuration is generated, it will mismatch again elsewhere in the image, and this feature in the form of S_x once found, will accelerate the rise of the visual attention score provided that the sequence is not subsequently interrupted by a match.

The size of neighbourhoods is specified by the maximum distance components ε_i to the pixel being scored. The neighbourhood is compared with the neighbourhoods of t other randomly selected pixels in the image that are more than a distance epsilon from the boundary. Typically $\varepsilon_i = 3$ and $t = 100$, with $m = 3$ neighbouring pixels selected for comparison. Larger values of m and the ε_i are selected according to the scale of the patterns being analysed. Increasing the value of t improves the level of confidence of the detail in the attention estimate display.

Pixels are matched if their colour values are separated by less than a certain threshold in the chosen colour space [10]. The results reported here have been derived using a modified form of the HSV colour space, and further work is necessary using spaces that provide colour difference formulae that more closely match the performance of the human visual system in visual attention tasks.

Computational demand increases linearly with t and the area of the image. An image of size 640 x 480 with $t = 100$ and $m = 3$ takes about 40 seconds on a 330MHz Pentium. It is not felt that processing time is a major drawback because the visual attention score for each pixel is not dependent on the scores of others and the algorithm is therefore capable of parallel implementation.

Processing time is considerably reduced at the risk of missing small details by means of a focusing strategy. Pixels are sampled for scoring according as they lie on a relatively sparse grid placed on the image under analysis. Any pixels obtaining a score greater than a certain threshold trigger the scoring of all pixels in that neighbourhood of the grid. This has the effect of attributing most processing power to areas of high visual attention whilst ignoring expanses of background material (eg expanses of sky).

4. RESULTS

A number of images have been processed that illustrate the performance of the algorithm. The examples in this section are presented as an original image together with a false colour representation of the visual attention estimates. The results on some standard illusions [Figure 3a, 3b] bear some similarity with human visual perception.

In the following examples the visual attention map has been used to derive a variable resolution version of JPEG coding to yield significant compressions beyond the original, and without affecting the perceptual quality of the image. The technique identifies the most striking aspects of images whether it be a boat on the sea or the shape of a mountain [4a, 4b]. In the case of a flower it is the central portion that attracts the greatest attention [5a, 5b], although the outline of the petals is also important. Providing textual material does not predominate in an image, the pixels making up the characters are highlighted by the attention mechanism [6a, 6b]. The visual structures associated with printed text do not occur in nature and so mismatching neighbourhoods are relatively easily found in such images. The compressed versions of figures 4 and 5 are shown in Figures 7 and 8.

5. DISCUSSION

The results described in this paper lend support to the conjecture that visual attention is to a certain extent dependent upon the disparities between neighbourhoods in the image. Eye-tracking experiments are planned to test and refine this hypothesis using results generated by the algorithm.

Although it cannot be proved, it is believed that the performance of the approach across a diverse set of visual content is due to the insistence that no *a priori* guidance is introduced into the estimation mechanism. Any form of heuristic filtering, quantisation, or normalisation will certainly enhance the performance on specific categories of input data, but it will also limit performance on potentially large volumes of unseen input data space [11]. However, if it is known for sure that attention must only be directed at objects possessing a certain colouring, for example that of human skin, then it would be reasonable to modify the colour space model to reflect this and encourage matching to take place between all colours except those possessing a hue close to that of skin.

6. CONCLUSIONS

A new technique for estimating visual attention has been proposed. Results indicate some similarity with the behaviour of the human visual system, although eye-tracking experiments need to be carried out over a range of images to test the validity of the model and to establish more clearly the part played by the higher level mental processes of the observer.

Knowing the important areas in an image means that differential compression techniques can be applied in such a way that perceived quality is not impaired whilst achieving potentially higher levels of compression than existing techniques that are unable to discriminate within images. This approach complements the JPEG 2000 image compression standard [12] by providing a region of interest profile to set the differential compression rates.

7. REFERENCES

[1] S. Grossberg, "The link between brains, learning, attention, and consciousness", Consciousness & Cognition", 8, 1-44, 1999.

[2] W. Osberger, and A. J. Maeder, "Automatic identification of perceptually important regions in an image", 14th IEEE Int. Conference on Pattern Recognition, 16-20th August 1998.

[3] J. Luo, and A. Singhal, "On measuring low-level saliency in photographic images", IEEE Conf. On Computer Vision and Pattern Recognition, June 2000.

[4] L. Itti, and C. Koch, "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention", http://www.klab.caltech.edu/~itti/attention/publications/00_VR also in Vision Research 2000.

[5] K. N. Walker, T. F. Cootes, and C. J. Taylor, "Locating Salient Object Features", British Machine Vision Conference, 1998.

[6] T. N. Mudge, J. L. Turney, and Volz, "Automatic generation of salient features for the recognition of partially occluded parts", Robotica, Vol 5, pp 117-127, 1987.

[7] R. Desimone, "Visual attention mediated by biased competition in extrastriate visual cortex", Phil. Trans. R. Soc. Lond. B, 353, pp 1245 – 1255, 1998.

[8] A. P. Pentland, "Fractal-based descriptions of natural scenes", IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, pp 661-674, 1984.

[9]F. W. M. Stentiford, "An evolutionary approach to the concept of randomness", British Computer Journal, vol. 16, no. 2, May, 1973.

[10]G. Sharma, and H. J. Trussell, H. J., "Digital Color Imaging", IEEE Trans on Image Processing, vol. 6, No. 7, July, pp 901-932, 1997.

[11] F. W. M. Stentiford, "Evolution, the best possible search algorithm?", BT Technology Journal, Vol. 18, No. 1, January 2000.

[12] http://www.jpeg.org/JPEG2000.htm#commlinks

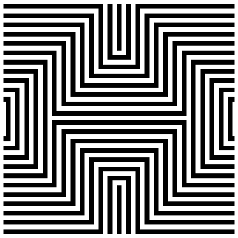

Figure 3a

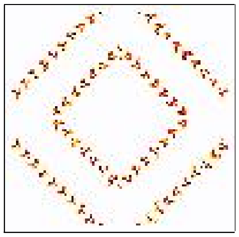

Figure 3b

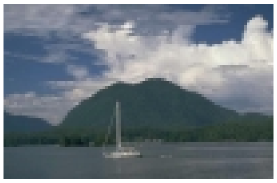

Figure 4a (21433 bytes)

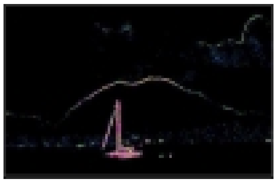

Figure 4b

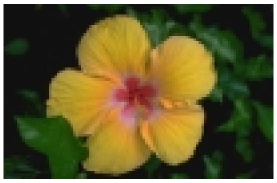

Figure 5a (20672 bytes)

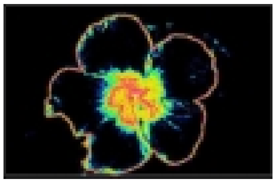

Figure 5b

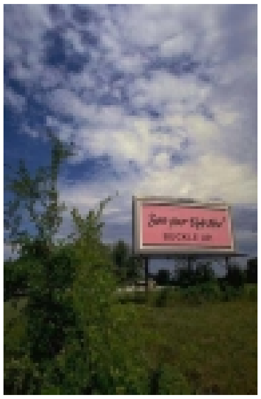

Figure 6a

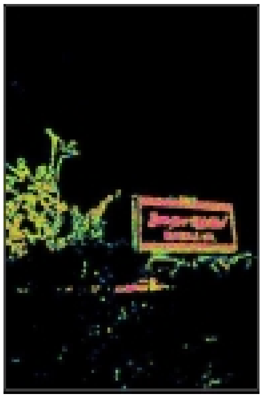

Figure 6b

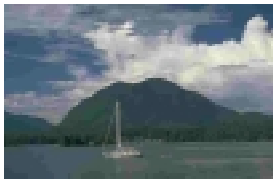

Figure 7 (7706 bytes)

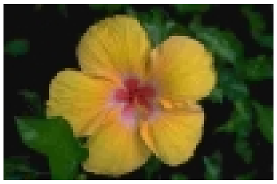

Figure 8 (10916 bytes)