An Evolutionary Programming Approach to the Simulation of Visual Attention

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

Abstract- 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. 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 an evolutionary programming approach (Michalewicz 1996) to derive a measure of visual attention that may be used to identify regions of interest in many categories of images. A population of individuals, or pixel neighbourhoods, is evolved that performs best at discriminating between salient and non-salient image features. A number of results are provided.

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. 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.

Grossberg's Adaptive Resonance Theory (Grossberg 1999) 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 (Osberger 1998) 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 (Luo 2000) 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. A number of other authors (Wong 2000, Chai 2000, Marichal 1996, Zhao 1996, Syeda-Mahmood 1997) also favour the pre-selection of criteria for determining regions of interest.

Itti et al (Itti 2000) 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 (Walker 1998) 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 (Mudge 1987) 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 (Desimone 1998) 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

0-7803-6657-3/01/$10.00 © 2001 IEEE

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.

Evolutionary algorithms have been shown to be very promising tools for the automatic design and optimisation of systems for image analysis (Stanhope 1998, Daida 1996). Some success has been achieved in applying evolutionary techniques to the segmentation of medical images (Cagnoni 1999).

This paper follows the thinking of Walker et al (Walker 1998) 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 (Pentland 1984). 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. The search is based upon an evolutionary process in which randomly generated constructs are tested and discarded or retained according to their discriminatory performance. 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 Visual Attention Framework

Let a set of measurements **a** correspond to a location **x** in an image where

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

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

Consider a neighbourhood N of **x** where

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

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

$$S_{\mathbf{x}} = \{\mathbf{x}'_1, \mathbf{x}'_2, \mathbf{x}'_3, \ldots, \mathbf{x}'_m\} \text{ and } \mathbf{F}(\mathbf{x}'_i) \text{ is defined.} \quad (1)$$

Select another location **y** for which **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 individual $S_{\mathbf{x}}$ matches $S_{\mathbf{y}}$ 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.

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

$C_{\mathbf{x}}$ is the number of mismatches between the individual $S_{\mathbf{x}}$ and t randomly selected locations **y**. Each time a match is obtained the individual $S_{\mathbf{x}}$ is allowed to 'die' and a new individual created as defined in (1). In this way individuals $S_{\mathbf{x}}$ that generate continuous sequences of mismatches are preserved with the highest performers lasting the longest. A population of the M highest performing individuals $S_{\mathbf{x}}$ is maintained over all the **x** and is drawn upon first before generating further random versions in (1). Individuals that generate high mismatch scores displace those with lower performances and survive to be selected again and again for their likely good discrimination in other parts of the image.

A location **x** will be worthy of attention if $C_{\mathbf{x}}$ is amongst the highest values associated with pixels in an image.

1 1 0 1 0 0 1 [1] 0 [1] 0 [0 1] 1 0 1 0 0 1 1 0 1 0 0 1 [1] 0 [1] 0 [1 1] 1 0 1

x y

Figure 1 Individual 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 (Stentiford 1973). Bits located within repetitive sequences (eg 110100110100) would obtain low attention estimates, but the vicinity of any occasional irregularities would be highlighted.

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

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.

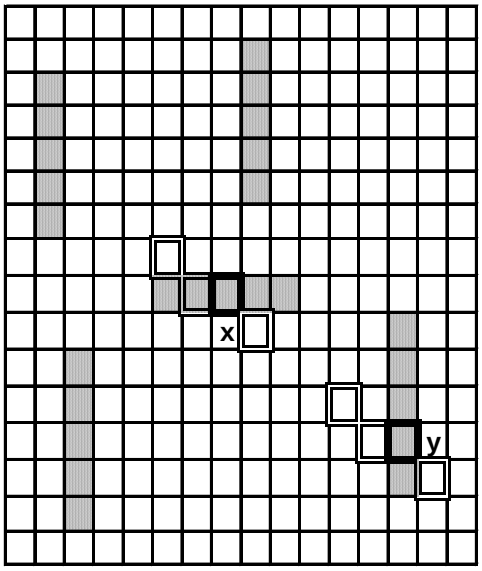

Figure 2 Individual at **x** mismatching at **y**.

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. 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 individual S_x if a mismatch is detected, and re-using S_x for comparison with the next of the t neighbourhoods. If however, S_x subsequently matches another neighbourhood, the score is not incremented, and a new individual S_x is created ready for the next comparison. In this way competing individuals 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.

Sometimes high performing individuals might be prematurely culled because a match is found early in the trials. However, the chances of this happening are least for the best performers and does not detract from the search for mismatches across the image.

The size of neighbourhoods is specified by the maximum distance of 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. It is possible to isolate attention features at different scales by insisting that neighbourhood pixels are all selected such that

$$|x_i - x'_i| > \eta_s.$$

Pixels are matched if their colour values are separated by less than a certain threshold in the chosen colour space (Sharma 1997). 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. A version has been successfully implemented in parallel on separate machines although each machine still needs access to the whole image.

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 black and white patterns show pop out effects [Figure 3a, 3b] and illusions

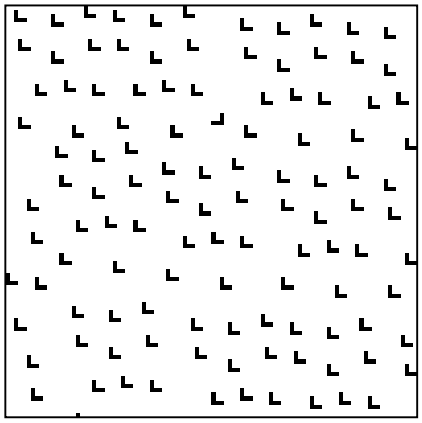

Figure 3a

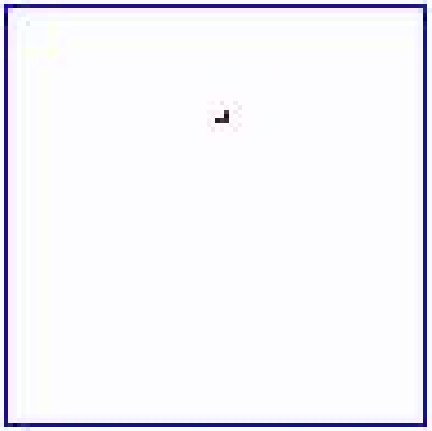

Figure 3b

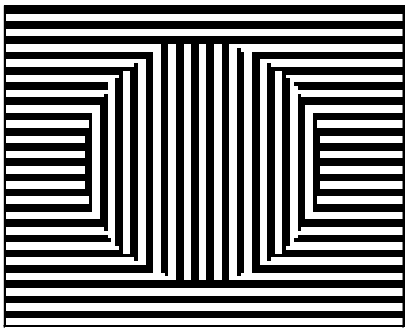

Figure 4a

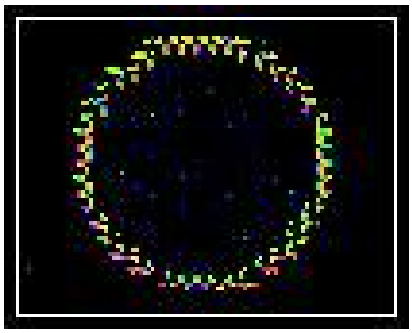

Figure 4b

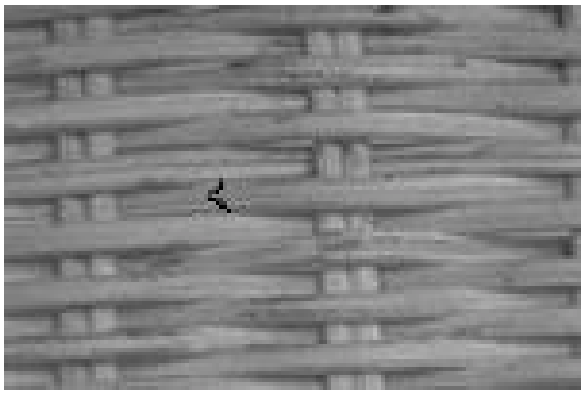

Figure 5a

Figure 5b

Figure 6a

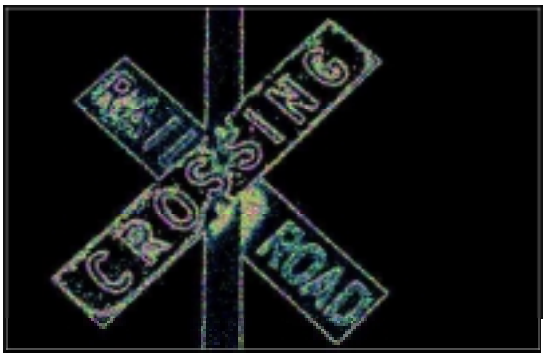

Figure 6b

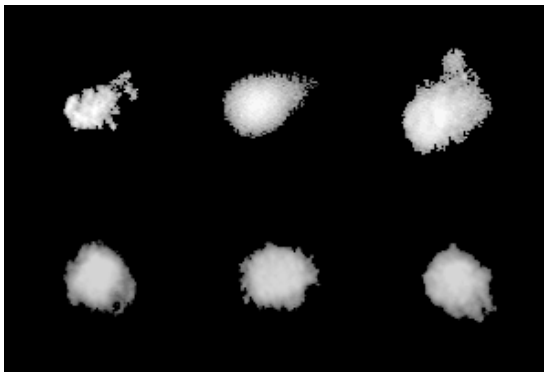

Figure 7a

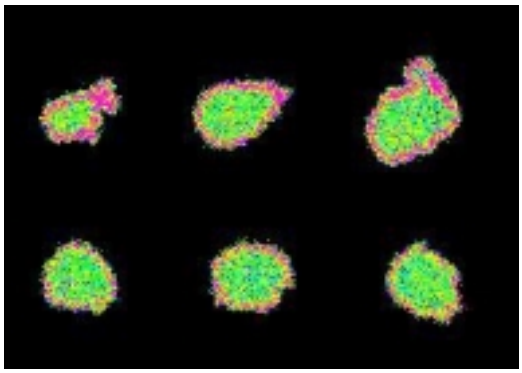

Figure 7b

Figure 8a

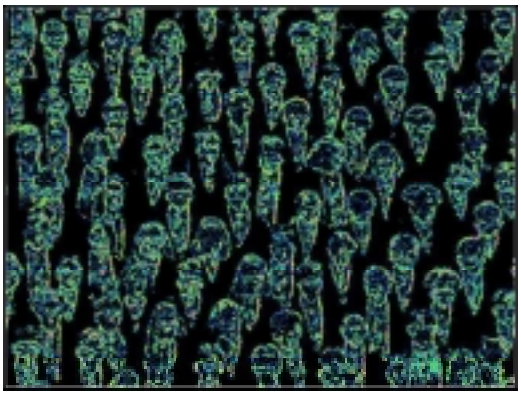

Figure 8b

Figure 9a

Figure 9b

Figure 10a

Figure 10b

Figure 11a

Figure 11b

[Figure 4a, 4b] that bear some similarity with human visual perception. Defects in textures [Figure 5a, 5b] are highlighted very strongly depending upon their frequency of occurrence against a regular background. Warning signs [6a, 6b] almost always possess high visual attention measures by virtue of their relatively rare colouration and colour adjacencies.

The ability of the system to detect artefacts by comparison with a background of normal examples is illustrated in an image of skin cells some of which have been exposed to ultra violet light [7a, 7b]. Cells damaged by UV are characterised by an uneven cell structure whereas normal cells possess a more predictable cell wall (Curnow 2001). Here the upper row of damaged cells is highlighted by larger areas of maximum visual attention estimates.

The visual attention estimator is of little help in those images where higher level mental processes and memories usually provide the stimulus for visual attention. If the image is largely self-similar and no neighbourhood is significantly different from others then the estimator will not generate high visual attention scores. For example, in a sea of faces [8a, 8b] nothing particular stands out unless a friend is spotted in which case attention is drawn to that individual to the exclusion of all others. However, the algorithm as described in this paper does not make use of *a priori* information except at the very lowest level (eg colour perceptual modelling) and is not able to simulate higher levels of cognitive behaviour.

The technique identifies the most striking aspects of images whether it be a boat on the sea or the shape of a mountain [9a, 9b]. In the case of a flower it is the central portion that attracts the greatest attention [10a, 10b], although the outline of the petals is also important. The central areas of the petals do not produce high visual attention scores because they occupy a much greater area of the image that either the central structure or the outer edge of the petals. Providing textual material does not predominate in an image, the pixels making up the characters are highlighted by the attention mechanism [11a, 11b]. The visual structures associated with printed text do not occur in nature and so mismatching neighbourhoods are relatively easily found in such images.

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* or Lamarckian 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 huge volumes of unseen input data space. Darwinian evolution has the advantage that it proceeds without preordained heuristics and relies upon the power of diversity to encompass good solutions and select for the best ones (Stentiford 2000). 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.

The current method produces a high performing population by selecting from individuals that are generated completely randomly. It should be possible to increase performance by mutating competent individuals, testing for better discrimination, and retaining those that perform better. Individuals may be altered by adding, deleting, or moving pixels in the neighbourhood. Whether this speeds the search process remains to be seen.

The visual attention map of an image forms part of the metadata of that image; it adds value in that the extra

information may be used to segment the image for a variety of purposes including compression. Areas of low visual attention can be compressed at a much higher rate than the rest of the image without affecting perceptual quality. As discussed above, small quantities of text would be unaffected by the compression.

6 Conclusions

An evolutionary programming technique for estimating visual attention has been described. Pixels that are surrounded by novel structure are distinguished from those that are embedded in neighbourhoods that have many similar counterparts elsewhere in the image. Populations of individuals, or pixel neighbourhoods, are evolved that characterise the regions of interest in images. 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.

It has been shown that objects worthy of visual attention can be located in a number of different scenes. Potential applications include the detection of surface defects in manufacturing processes. This could be extended to the location of anomalies in more general images such as those of airborne objects, or man-made intrusions in country or ocean scenes. The approach also has potential for providing an objective measure of the extent of cellular damage as reflected by optical appearance. Artefacts may be characterised by the level of attention they attract against a background of normal controls. Finally, 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.

The computational demands can be high, but the algorithm is capable of parallel implementation making the processing time independent of the image size. Moreover serial implementations are speeded up by concentrating processing on parts of the image likely to yield areas of high visual attention as indicated by sample pixel scores.

7 Acknowledgements

Acknowledgements are due to Volker Typke who wrote the software and produced some of the results in this paper.

Bibliography

Cagnoni, S., Dobrzeniecki, A. B., Poli, R., and Yanch, J. C. (1999) “Genetic algorithm-based interactive segmentation of 3D medical images”, Image and Vision Computing, 17:881-895.

Chai, D., Ngan, K. N. and Bouzerdoum (2000) "Foreground/background bit allocation for region-of-interest coding", IEEE Conf. On Computer Vision and Pattern Recognition, June.

Curnow, A., Salter, L., Morley, N., and Gould, D. (2001), “A preliminary investigation of the effects of arsenate on irradiation-induced DNA damage in cultured human lung fibroblasts”, to be published in the Journal of Toxicology and Environmental Health.

Daida, J. M., Bersano-Begey, T. F., Ross, S. J., and Vesecky, J. F. (1996) “Evolving feature-extraction algorithms: adapting genetic programming for image analysis in geoscience and remote sensing”, Proc 1996 Int Geoscience and Remote Sensing Symposium, pp 2077 – 2079.

Desimone, R. (1998) “Visual attention mediated by biased competition in extrastriate visual cortex”, Phil. Trans. R. Soc. Lond. B, 353, pp 1245 – 1255.

Grossberg, S. (1999) “The link between brains, learning, attention, and consciousness”, Consciousness & Cognition”, 8, 1-44.

Itti, L. and Koch, C. (2000) “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.

Luo, J. and Singhal, A. (2000) “On measuring low-level saliency in photographic images”, IEEE Conf. On Computer Vision and Pattern Recognition, June.

Marichal, X., Delmot, T., De Vleeschouwer, C., Warscotte, V. and Macq, B. (1996) "Automatic detection of interest areas of an image or of a sequence of images", IEEE Int. Conf. on Image processing.

Michalewicz, Z. and Michalewicz, M. (1996) “Evolutionary computation: main paradigms and current directions”, Appl. Math. And Comp. Sci., vol 6, No 3, pp 393 – 413.

Mudge, T. N., Turney, J. L. and Volz (1987) “Automatic generation of salient features for the recognition of partially occluded parts”, Robotica, Vol 5, pp 117-127.

Osberger, W. and Maeder, A. J. (1998) “Automatic identification of perceptually important regions in an image”, 14th IEEE Int. Conference on Pattern Recognition, 16-20th August.

Pentland, A. P. (1984) “Fractal-based descriptions of natural scenes”, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6, pp 661-674.

Sharma, G. and Trussell, H. J. (1997) “Digital Color Imaging”, IEEE Trans on Image Processing, vol. 6, No. 7, July, pp 901-932.

Stanhope, S. A. and Daida, J. M. (1998) “Genetic programming for automatic target classification and recognition in synthetic aperture radar imagery”, Lecture Notes in Computer Science, vol 1447, pp 735 – 744.

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

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

Syeda-Mahmood, T. F. (1997) "Data and model-driven selection using color regions", Int. J. Computer Vision, Vol 21, No 1, pp 9-36.

Walker, K. N., Cootes, T. F. and Taylor, C. J. (1998) “Locating Salient Object Features”, British Machine Vision Conference.

Wong, H. and Guan, L. (2000) "Characterization of perceptual importance for object-based image segmentation", IEEE Conf. On Computer Vision and Pattern Recognition, June.

Zhao, J., Shimazu, Y., Ohta, K., Hayasaka, R., and Matsushita, Y. (1996) "An outstandingness oriented image segmentation and its application", Int. Symposium on Signal Processing and its Applications, August.