A context-based model of attention

(extended abstract)[1]

Niek Bergboer Eric Postma Jaap van den Herik

Institute for Knowledge and Agent Technology, Universiteit Maastricht,
P.O. Box 616, 6200 MD Maastricht

It is well known that natural visual systems rely on attentional mechanisms that select and process relevant objects in an efficient way. Similarly, artificial visual systems need attentional-selection mechanisms to reduce the computational burden of processing entire images. So, their aim is to focus on the parts containing the object of interest. In the domain of natural vision the locus of selection has been debated for many years (see [1] for an overview). The two extreme views are (1) that selection takes place at an early stage of visual processing (i.e., early selection), and (2) that it takes place at a late stage (i.e., late selection). In early selection, attention is guided by conspicuous changes in elementary features, such as colour, texture, or spatial frequency. Models of early selection contain so-called saliency maps that contain the response respond to conspicuous changes in a single feature, e.g., [4]. The activities in these maps represent locations to be attended. In late selection, attention is guided by complex feature combinations or even objects [7]. Models of late selection rely on object templates that are matched to the contents of images [6].

From a computational point of view, both early and late selection pose considerable problems. In early selection, the likelihood of mistakes is large, since in natural images many changes of elementary features occur. As a result, the attentional mechanism has to visit many locations of which only a few correspond to objects of interest. The object-based guidance of attention in late selection renders the selective function useless, as late selection requires the location (identity) of the objects to be known in advance.

Several models have attempted to combine saliency maps with template matching (see, e.g., [4]). Below, we propose a novel approach, the COBA (COntext BAsed) model of attentional selection. The main idea underlying the COBA model is that the spatial context of an object is important for its localisation, as illustrated in the left panel of Figure 1. The two small square images (left in the figure) are enlarged versions of the square regions indicated by boxes in the large images. Considered in isolation, both small images are highly similar to faces. When considered in their natural context, the small images will not be interpreted as faces [2].

In the COBA model, attentional selection is guided by an object saliency map. Active locations on the map indicate likely locations of objects. Using automatic learning, the object saliency map is generated from feature combinations that form a likely spatial context for objects. In this paper we focus on applying the COBA model to spatial contexts and on the detection of faces in natural images. Our method is related to the more global selection method proposed by Torralba and Sinha [8].

The COBA model consists of three stages. In the first stage, windows are taken from natural images at a grid of scales and locations. The contents of these windows are pre-

[1]This is a summary of a paper accepted for the European Conference on Artificial Intelligence ECAI 2004, August 2004, Valencia, Spain.

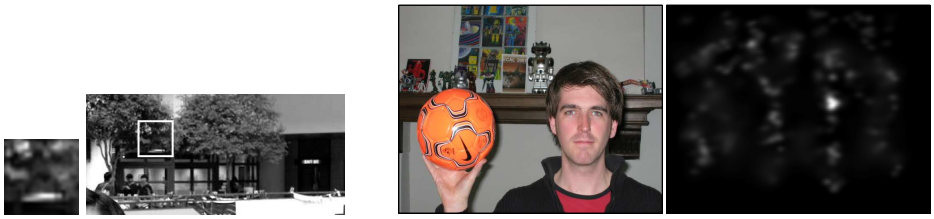

Figure 1: *Left:* Example of a pattern that is similar to a face, but that is clearly not faces when viewed in its context. *Right:* Example image and obtained object saliency map.

processed by transforming them into feature vectors by means of a standard multi-scale wavelet transformation [5].

In the second stage, the feature vectors (representing the contents of the context windows) are transformed into estimates of object locations. A cluster-weighted modelling technique [3] is used to obtain a probability density function for the location of the object.

The third stage is the addition of the PDFs obtained at all locations and scales in the image to yield a global object saliency map; an example is shown in the right panel of Figure 1 The COBA model of attentional selection has been evaluated on a face-detection task. In active regions (i.e., regions with a high object saliency) a face detector [9] is applied to detect the presence of a face. Results on a 775-image test set containing $1,885$ labelled faces indicate that the false-positive rate can be reduced from 20 false positives per image to 1 false positive per image, while the detection rate drops only slightly from 91.4% to 80.4%.

An analysis of the trained models shows that context-based attentional selection is an efficient and viable way of dealing with the early-versus-late selection dilemma. From the results, we may conclude that context-based selection reduces the number of false detections and the size of the search space. As a consequence, it can be readily applied in artificial visual systems.

References

[1] G. Backer, B. Mertsching, and M. Bollmann. Data- and model-driven gaze control for an active vision system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1415–1429, December 2001.

[2] N. H. Bergboer, E. O. Postma, and H. J. van den Herik. Context-based object detection in still images. Submitted elsewhere.

[3] N. Gershenfeld. *The Nature of Mathematical Modeling*. Cambridge University Press, Cambridge, MA, 1999.

[4] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, November 1998.

[5] C. Papageorgiou and T. Poggio. A trainable system for object detection. *International Journal of Computer Vision*, 38(1):15–33, 2000.

[6] E. O. Postma, H. J. van den Herik, and P. T. W. Hudson. SCAN: A scalable neural model of covert attention. *Neural Networks*, 10(6):993–1015, 1997.

[7] B. J. Scholl. *Objects and attention*. Elsevier Sciences Publishers, Amsterdam, 2002.

[8] A. Torralba and P. Sinha. Statistical context priming for object detection. In *Proceedings of the International Conference on Computer Vision*, Vancouver, Canada, 2001.

[9] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2001.