A context-based model of attention

Niek Bergboer and **Eric Postma** and **Jaap van den Herik** [1]

Abstract. Artificial visual systems need an attentional selection mechanism to constrain costly processing to relevant parts. An important design decision for such systems concerns the locus of selection. To guide the selection mechanism, traditional models of attention use either an early locus of selection based on low-level features (e.g., conspicuous edges) or a late locus of selection based on high-level features (e.g., object templates). An early locus is computationally cheap and fast but is an unreliable indicator of the objecthood. A late locus is computationally expensive and slow and requires the object to be selected to be known, rendering selection for identification useless. To combine the advantages of both loci, we propose the COBA (COntext BAsed) model of attention, that guides selection on the basis of the learned spatial context of objects. The feasibility of context-based attention is assessed by experiments in which the COBA model is applied to the detection of faces in natural images. The results of the experiments show that the COBA model is highly successful in reducing the number of false detections. From the results, we may conclude that context-based attentional selection is a feasible and efficient selection mechanism for artificial visual systems.

1 INTRODUCTION

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 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 [10]. Models of late selection rely on object templates that are matched to the contents of images [9].

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. In late selection, object-based guidance of attention requires the location (and identity) of the object to be known which renders the selective function of attention for identification useless.

[1] Universiteit Maastricht, Institute for Knowledge and Agent Technology (IKAT), Maastricht, The Netherlands email: n.bergboer@cs.unimaas.nl

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. The importance of spatial context for reliable object recognition is illustrated in 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 interpretation as faces is suppressed [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 [11]; this relation will be explored in more detail in the Discussion section.

The outline of the remainder of this paper is as follows. Section 2 describes the COBA model and how it is trained to build an object location saliency map. In section 3, the selection performance of the COBA model is evaluated on natural images containing faces. Section 4 discusses the results obtained in terms of efficiency and reliability. Finally, section 5 concludes that context-based selection is a feasible solution to the early-versus-late selection dilemma.

Figure 1. Examples of patterns that are similar to faces, but that are clearly not faces when viewed in their context.

2 THE COBA MODEL

The COBA model consists of three stages. In the first stage, the image is preprocessed using a biologically plausible transformation. The second stage involves the local context-based estimation of object locations. In the third stage, the local estimates are integrated into a global saliency map. Below, we discuss each of these stages in detail and provide an example of model training.

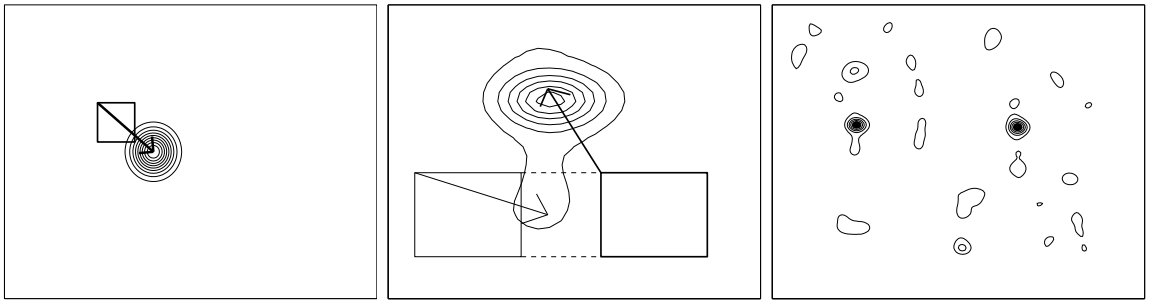

Figure 2. Graphical representation of the COBA attentional model.

2.1 Preprocessing

The COBA model operates on natural images that contain objects of interest. The task is to detect the objects adequately. Therefore, we introduce the concept of *context windows*, i.e., parts of the natural image that are of the same size as the object, and that lie in the spatial vicinity of the object. The square depicted in the left panel of Figure 2 represents a context window (e.g., in Figure 3, this is the upper-left solid square containing one eye and a part of the nose). In the preprocessing stage of the COBA model, the contents of context windows is transformed into feature vectors by means of a standard multi-scale wavelet transformation. In the early stages of the natural visual system a similar transformation on retinal images is performed [7]. In the experiments reported here we use an overcomplete Haar-wavelet basis for transforming the contents of the context windows [8]. We extract quadruple-resolution wavelets at wavelet scale 4 from a window of size 19×19 pixels [8]. In this way, relations between large-scale features on a fine spatial resolution are incorporated in the feature-vector representation. Quadruple-resolution scale-4 wavelets yield 17×17 coefficients for a given orientation for a context window of size 19×19. As we use 3 orientations (horizontal, vertical, and diagonal), this yields a 867-dimensional feature vector for each context window. To facilitate further processing, the raw feature vectors are projected onto 16-dimensional reduced feature vectors v using principal component analysis.

2.2 Context-based estimation of object locations

In the second stage of the COBA model the feature vectors (representing the contents of the context windows) are transformed into estimates of object locations. The aim is to estimate the location of the object *relative* to the context window, i.e., to estimate the vector $\overrightarrow{x_r} = (x_r, y_r)$ by means of the window features v, where x_r represents the horizontal relative location and y_r represents the vertical relative location.

The transformation is learned using a training set consisting of the estimated $\widehat{\overrightarrow{x_r}}$ (acquired from context windows) and the associated true object position $\overrightarrow{x_r}$. As a learning algorithm we use cluster-weighted modelling [3], because it is straightforward and efficient.

The function that minimises the mean square error between $\widehat{\overrightarrow{x_r}}$ and $\overrightarrow{x_r}$ is the conditional expected value [6, p. 247]:

$$\widehat{\overrightarrow{x_r}} = \int \overrightarrow{x_r} f(\overrightarrow{x_r}, v) d\overrightarrow{x_r}, \tag{1}$$

where the joint probability density function (PDF) $f(\overrightarrow{x_r}, v)$ describes the relation between the two random variables $\overrightarrow{x_r}$ and v. It is given by:

$$f(\overrightarrow{x_r}|v) = \frac{f(\overrightarrow{x_r}, v)}{f(v)}. \tag{2}$$

After training, the preprocessed contents of context windows is translated into a PDF for the relative object location. The left panel of Figure 2 is an illustration of such a PDF (the contour lines). The arrow represents the estimate of the object location relative to the context window (the square). Figure 3 illustrates the position and extent of a typical context window for the object class of faces. The large dotted square denotes the region from which the context windows (training samples) are taken, the top left corner of the dashed square in the centre corresponds to the true face location, and the solid square in the upper left corner of the training region represents the context window for relative displacement of -8 pixels in both the horizontal and vertical direction in the down-sampled image[2].

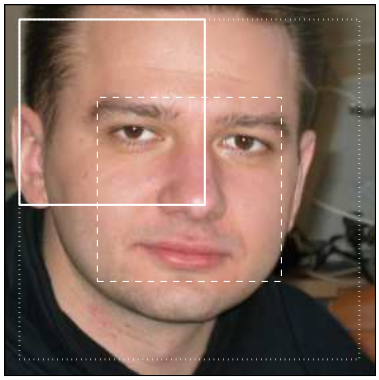

Figure 3. The training region (the large dotted square) used to obtain samples for the learning algorithm.

2.3 Integration of local estimates

The third stage of the COBA model is the addition of the PDFs obtained at all locations and scales in the image to yield a global object saliency map. The integration is illustrated in the middle panel of Figure 2 for a given scale and two locations. PDFs are obtained at a grid of window locations. In order to obtain a larger-scale saliency map for the object, all individual PDFs are added. This adding process is performed by moving each individual PDF to its *absolute* location

[2] We show the training region overlaid on a *high-resolution* face in Figure 3 to illustrate clearly what portion of the head is used as context. When extracting features from a facial context, the high-resolution image is first downscaled such that the face has a size of 19×19 pixels.

in the image, which is obtained by adding the expectation value of the relative location to the current location of the window. In the Figure, the expectation value for the position based on features from the leftmost context window lies slightly below and to the right of the current location. The expectation value for the position based on features from the rightmost context window lies above and to the left of the current position. The integration stage results in an image-wide object salience map. The right panel of Figure 2 is an illustration of the object saliency map. The contour lines demarcate the saliency.

2.4 Model training

Before evaluating the COBA model, we have to train the cluster-weighted model to achieve reliable context-based estimates of object location. The main design parameter for the cluster-weighted model is the number of clusters that are used. We train a model using 8 clusters, as preliminary results have shown that further increasing the number of clusters does not improve detection. The algorithm is trained on a dataset of $1,885$ faces. Within the facial vicinity shown in Figure 3, training samples are obtained from context windows at relative displacements $-8, -6, -4, -2, 0, +2, +4, +6, +8$ in both the horizontal and vertical directions. Thus, 81 relative displacements for each face are used, which yields a total of $152,685$ samples in the training set.

Figure 4 shows the trained attention model in terms of learned cluster centres. Each cluster has a centre in the input space (i.e., a combination of visual features) that is linked to a centre in the output space (i.e., a relative face location). Each row in the left halve of the figure corresponds to a cluster centre in the input space. The three columns show horizontal, vertical, and diagonal wavelet details, respectively. The grey value represents the magnitude of the brightness gradient in the given orientation. The right halve of the figure shows the cluster centres in the output space; each ellipse shows a one standard-deviation confidence interval of the relative location with respect to the location of the face (indicated by an X). For instance, the particular pattern of brightness gradient magnitudes given in the top row is indicative of a position slightly to the upper-right of the actual face location.

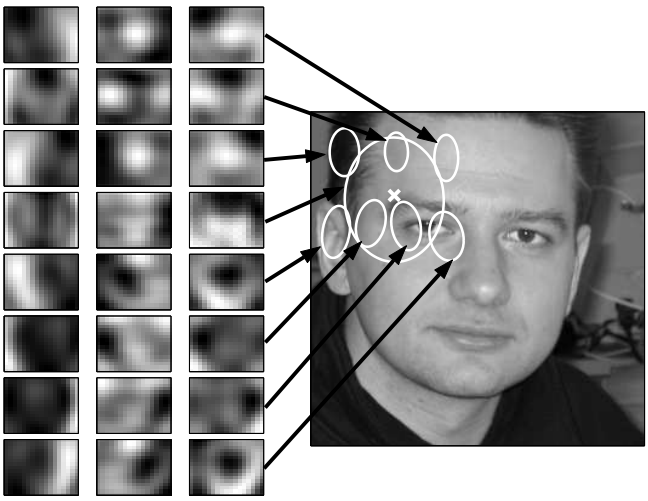

Figure 4. The cluster centres for the Gaussian mixture models, with their relative position and confi dence intervals.

The Figure reveals that spatially large features such as the eyes, the nose, the mouth, and the edges of the head, are used to locate the face. For instance, the second row of Figure 4 shows that the presence of the eyes in the centre of the window, indicated by the two bright spots in the vertical details, provides a strong contextual clue that the current position is slightly above the location of the face. However, in practice, there will not be merely one cluster that determines the location of the face; the perceived features are projected onto a linear combination of the eight cluster centres, and the estimated face location will also be a linear combination of the relative positions of each of the clusters.

3 EXPERIMENTAL EVALUATION

Below, we evaluate the COBA model of attentional selection on a face-detection task. In active regions (i.e., regions with a high object saliency) a face detector is applied to detect the presence of a face. We use a face detector based on the work of Viola and Jones [12] and Lienhart et al. [5].

3.1 Data

The performances of the context-selection method is assessed on 775 images from the Internet that together contain $1,885$ labelled faces (henceforth referred to as the "web set"). The images contain labelled faces that are at least 30×30 pixels in size. To ensure that all faces are found, the images are classified for face sizes of 24.5×24.5 pixels and up. To this end, a scale-space pyramid of the image is calculated in which subsequent scales differ by a factor 1.1.

3.2 Experimental methodology

In order to obtain statistically valid results, we perform a 10-fold cross-validation procedure. Therefore we split the dataset into 10 parts. In the model used for the localisation of faces in images belonging to set i, we leave out part i from the training set. In addition, the model is used for different "step sizes"; for a given "step size" s, a PDF is calculated only at every s pixels in both the horizontal and vertical direction.

To assess the attentional selection performance, we employ three performance measures: *true positives* (i.e., correctly detected faces), *false negatives* (i.e., falsely rejected faces), and *false positives* (i.e., falsely detected faces).

From all object windows that are classified positively, at least one should overlap sufficiently with a labelled face in order for that face to be a *true positive*; the overlap criteria are: (1) the size of the window should be within a factor 1.5 of the size of the labelled face, and (2) the window's centre should be within a distance to the labelled face's centre that is not larger than 30 per cent of that face's size. If none of the detected windows satisfies the two overlap criteria for a given labelled face, that labelled face is a *false negative*. Detected windows that do not satisfy the overlap criteria for any of the labelled faces are regarded as *false positives*.

To assess the benefits of context-based attentional selection with the COBA model, the results obtained are compared to results obtained without a selection mechanism. To this purpose, all images were scanned pixel-wise in their entirety using the face detector. This brute-force method yields a detection rate of 91.3% at 2.60 false positives per image on the web set.

3.3 Results

Figure 5 shows detection results for COBA on the web set. Both graphs are obtained by varying the fraction of the search space in which the object classifier is used. The range is based on the region-selection results of the first stage of COBA; only the given fraction of point with the highest saliency are selected as sufficiently salient to be further examined by the object classifier. The left graph

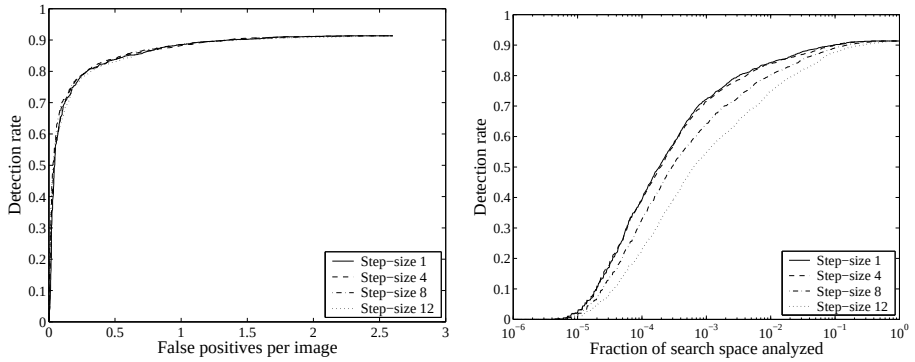

Figure 5. The detection results for the COBA model of attentional selection.

of Figure 5 shows a receiver operating characteristic (ROC) curve (which shows the trade-off between detection rate and the false-positive rate). The right graph shows the trade-off between the detection rate and the fraction of the brute-force search space that is actually searched after the spatial-context selection. In both curves, the fraction of the search space is varied from 10^{-6} to 1 (no selection, equivalent to the brute-force method). In each figure, results are shown for the step sizes 1, 4, 8, and 12. The graphs show that a stricter context selection causes a lower false-positive rate and a smaller search space, at the cost of a lower detection rate.

Our results are instrumental to arrive at three observations. The main observation is that with context-based attentional selection, the false positive rate can be reduced considerably while still retaining a high detection rate. Although one could in principle choose any point on the graphs to compare the results, we chose to compare detection results under conditions at which the method is practically applicable. The criterion we used for practical applicability is that the method should yield a detection rate of at least 80%. An advantage of the COBA model of attentional selection is that, after the region-selection stage, we can choose in which fraction of the search space we use the object classifier. Table 1 lists the fractions of the search space that must be searched to obtain an 80% detection rate, together with the number of false positives per image obtained at that fraction. The best results at one false positive per image are obtained with step size 1: the false-positive rate is reduced by a factor 9 while the detection rate drops only slightly from 91.3% to 80.0%.

Table 1. Detection results with a search space fraction chosen such that the detection rate is 80%.

Step size	*Fraction of search space*	*False positives per image*
1	$3.26 \cdot 10^{-3}$	0.288
4	$4.16 \cdot 10^{-3}$	0.281
8	$9.66 \cdot 10^{-3}$	0.290
12	$2.22 \cdot 10^{-2}$	0.320

Additional experiments have been performed in which a richer set of raw visual features were used. In this multiscale feature set, double-resolution scale-2 wavelets have been extracted from the window, in addition to the quadruple-resolution scale-4 wavelets. This yields a raw feature vector $v_R \in \mathbb{R}^{1,734}$. Results for this feature set are listed in Table 2. The results are slightly better than when using the original raw features; the false-positive rate can be reduced by a factor of 11.

Table 2. Detection results for the multiscale feature set, with a search space fraction chosen such the detection rate is 80%.

Step size	*Fraction of search space*	*False positives per image*
1	$2.21 \cdot 10^{-3}$	0.240
4	$2.37 \cdot 10^{-3}$	0.217
8	$7.22 \cdot 10^{-3}$	0.272
12	$1.77 \cdot 10^{-2}$	0.284

A second important observation is that the search space for the selection mechanism can be reduced by a large factor while still retaining an acceptable detection rate; e.g., when using our region-selection model with a step size of 1 based on the multiscale feature set, only a fraction $2.21 \cdot 10^{-3}$ of the search space has to be searched. This implies a search-space reduction of a factor 452. Given the experiments listed in Tables 1 and 2, the search space reduction to obtain a 80% detection rate lies between 45 and 452.

Our third observation is that the performance of the COBA model of attentional selection is relatively insensitive to the choice of a step size. Detection and false-positive rates are hardly affected by increasing the step size from 1 to 4. Choosing a step size of 4 means that our region-selection model needs to be evaluated at only $1/16$th of the locations compared to a step size of 1, thereby yielding a speed increase of roughly a factor 15. Using larger step sizes of 8 and 12 reduces the number of region-selection model evaluations even further, but deteriorates the detection rates. Moreover, these large step sizes require the object classifier to be used in a larger portion of the search space.

Figure 6 shows a typical example of attentional selection with the COBA model. The left panel shows the original image. The centre panel shows the object (face) saliency map in which grey values represent probabilities. The right panel shows the final detection results: the solid squares represent object detections found in the first 10^{-3} fraction of the search space, whereas the dashed squares represent object detections found in the remainder of the search space (exhaustive search). It is clear that by using context-based attentional selection, the COBA model has lowered the number of false detections and reduced the search space considerably.

4 DISCUSSION

Inspired by the importance of spatial context for object recognition, we presented the COBA model of attentional selection. Our results show that the COBA model is effective in reducing the search space while retaining a high detection rate. This means that it is able to

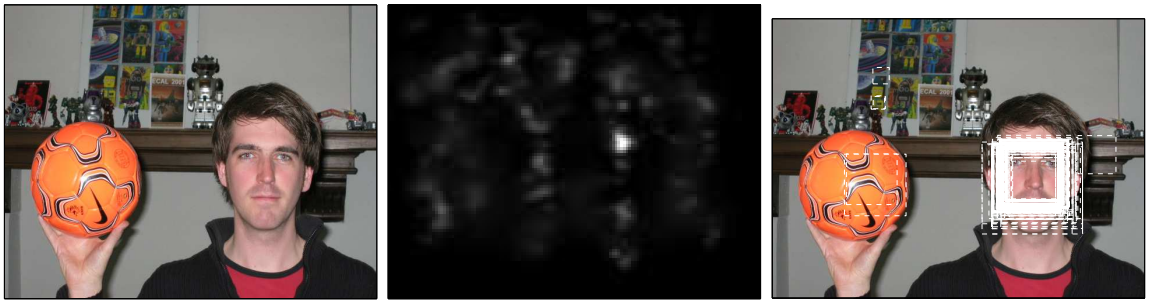

Figure 6. Typical example of context-based attentional selection with the COBA model.

distinguish between faces and face-like patterns in unlikely contexts, such as those displayed in Figure 1. Purely object-based methods, like most current object-detection methods, are not able to distinguish between face patterns and face-like non-face patterns.

The histograms in Figure 7 provide insight into the composition of the search space of context-based attentional selection. The histograms show the ranking distribution of the true positives (top histogram) and false positives (bottom histogram) in the attentional search space (for step size 4). The histograms are obtained by using the image-wide saliency map to determine what fraction of the search space would have to be searched to find each true-positive or false-positive detection. It is clear that the majority of the true positives are located in the first $1/1000$th part of the search space, whereas most of the false positives are not found until almost the entire search space is analyzed.

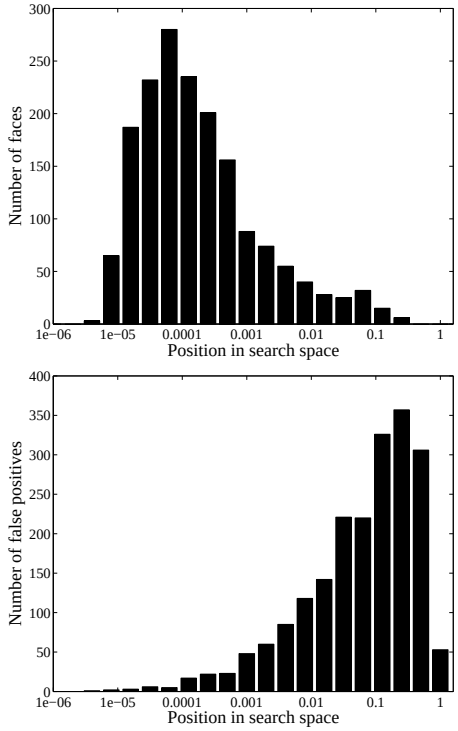

Figure 7. Attentional ranking of true and false positives using COBA

Our work is related to the work by Torralba and Sinha [11], who predict object locations based on general image statistics. Their method selects a rather large portion of the image as a region that might contain the target object, and is thus less specific than the COBA model in locating objects. Their method can be used as a global pre-selection mechanism for the COBA model.

5 CONCLUSIONS

In this paper, we proposed a context-based model COBA of attentional selection that learns to recognize likely object contexts to generate a saliency map of object locations. Our results show 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.

ACKNOWLEDGEMENTS

The authors are grateful for the comments made by the referees. This research is carried out within the ToKeN 2000 project EIDETIC (grant number 634.000.001) of the Netherlands Organisation for Scientific Research (NWO).

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] R. Lienhart, L. Liang, and A. Kuranov, 'An extended set of haar-like features for rapid object detection', Technical report, Intel Research, (June 2002).
[6] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, John Wiley & Sons, Inc., New York, NY, 1994.
[7] S. E. Palmer, *Vision Science, Photons to Phenomenology*, MIT Press, Cambridge, MA, 1999.
[8] C. Papageorgiou and T. Poggio, 'A trainable system for object detection', *International Journal of Computer Vision*, **38**(1), 15–33, (2000).
[9] 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).
[10] B. J. Scholl, *Objects and attention*, Elzevier Sciences Publishers, Amsterdam, 2002.
[11] A. Torralba and P. Sinha, 'Statistical context priming for object detection', in *Proceedings of the International Conference on Computer Vision*, Vancouver, Canada, (2001).
[12] 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).