A Visual Attention Operator Based on Morphological Models of Images and Maximum Likelihood Decision

Roman M. Palenichka

Université du Québec, Dept. of Computer Science
Hull, Québec, Canada
palenich@uqah.uquebec.ca

Abstract. The goal of the image analysis approach presented in this paper was two-fold. Firstly, it is the development of a computational model for visual attention in humans and animals, which is consistent with the known psychophysical experiments and neurology findings in early vision mechanisms. Secondly, it is a model-based design of an attention operator in computer vision, which is capable to detect, locate, and trace objects of interest in images in a fast way. The proposed attention operator, named image relevance function, is an image local operator that has local maximums at the centers of locations of supposed objects of interest or their relevant parts. This approach has several advantageous features in detecting objects in images due to the model-based design of the relevance function and the utilization of the maximum likelihood decision.

1 Introduction

Time-effective detection and recognition of objects of interest in images is still a matter of intensive research in computer vision community because the artificial vision systems usually fail to outperform the detection results by a human being. The detection problem is complicated when objects of interest have low contrast and various sizes or orientations and can be located on noisy and inhomogeneous background with occlusions. In many practical applications, the real-time implementation of object detection algorithms in such natural conditions is a matter of great concern. The results of numerous neurophysiological and psychophysical investigation of human visual system (HVS) indicate that the human vision can successfully cope with these complex situations because of using a visual attention mechanism associated with a model-based image analysis [1,2]. The goal of presented here investigation was not the simulation of human visual perception but the incorporation of its advantageous features into computer vision algorithms. Besides many remarkable properties of HVS like the mentioned model-based visual attention, the HVS has also some disadvantages such as visual illusions while detecting and identifying objects [3].

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 310-319, 2002.

Several models of attention mechanism in HVS in the context of reliable and time-effective object detection in static scenes have been proposed in the literature. They are mostly based on the generalization of edge and line detection operators and on the utilization of a multi-resolution image analysis including the wavelet theory [4-8]. Very good results of attention modeling have been reported by the application of symmetry operators to images [8]. There are known attention operators, which combine both the multi-resolution approach and the symmetry operators. Attention operators based on the wavelet image analysis also showed great potential, especially when integrating such novel types of wavelets as curvelets and ridgelets [7]. In contrast to the standard isotropic image analysis, they incorporate a multi-scale analysis of anisotropy of objects of interest in images.

The feature extraction approach *per se* is also a method for selecting regions of interest although it is a generic approach and requires explicit defining of relevant features. It can be considered as an intermediate stage between pre-attentive vision and post-attentive vision. Recently, a method for directed attention during visual search has been developed based on the maximum likelihood strategy [9]. It is suitable to detection of objects of interest of a particular class by pairing certain image features with the objects of interest but is restricted only to the detection task.

However, a few work has been done toward designing a model-based attention mechanism which is quite general and based on an image model of low and intermediate levels (description of object regions and their shape), and can yield an optimal detection and segmentation performance with respect to the underlying model. The low-level image modeling is requested because the attention mechanism in its narrow sense is a *bottom-up* image analysis process based on quite general intensity and shape properties in order to respond to various unknown *stimuli* as well as to provide a reasonable response when no object of interest is present.

In this paper, a new model of visual attention based on the concept of a multi-scale relevance function is proposed as a mathematical representation of some generally recognized results regarding the explanation of HVS mechanisms. The introduced relevance function is an image local operator that has local maxima at centers of location of supposed objects of interest or their parts if the objects have complex or elongated shapes. The visual attention mechanism based on the relevance function provides several advantageous features in detecting objects of interest due to the model-based approach used. While detecting objects, it provides a quick location of the objects of interest with various sizes and orientations. Besides some other advantages, the operating with the property map as an intermediate image representation enhances the possibility to treat images with inhomogeneous backgrounds and textured appearance of objects.

2 Representation of Planar Shapes of Objects

Analysis of images and detection of local objects of interest in images can be efficiently performed by using object-relevant image properties. There are considered properties of object planar shape as well as intensity properties within a region of interest containing an object on the background. Such properties have to be computed in each image point in order to be able to perform the image segmentation, which

provides object and background regions for the object recognition. This results in the computation of a *property map* as input data for further image analysis including the detection of objects of interest. It is assumed that in the general case the image intensity is represented by a vector of n primary features $\boldsymbol{x}=[x_1,\ldots,x_n]$. For example, pixels of a color image are three-component vectors. The primary features can be extracted from a gray-scale image on the basis of one feature vector per pixel. This includes the case of texture features when a set of local features is computed in each image point. Some examples of used primary features are given in the section of experimental results. In fact, the vector $\boldsymbol{x}=[x_1,\ldots,x_n]$ describes the parameters (properties) of an image intensity model.

On the second step, one final property z is computed by a linear clustering transformation:

$$z=a_1\cdot x_1+ a_2\cdot x_2+,\ldots,+ a_n\cdot x_n \ , \tag{2.1}$$

where the coefficient vector $\boldsymbol{a}=[a_1,\ldots,a_n]$ have to be computed in such a way that a separability measure between object and background pixels will be maximized in the new data space of z. One such reasonable choice is the well-known (in mathematical statistics) Fisher's linear discriminant based on sample mean value vectors within class and scatter vectors computed over object pixels and background pixels during the learning (estimation) of the property map transformation by Eq. (2.1).

In order to consider both the intensity and shape description of objects of interest in images simultaneously, a structural image model of the property map is used which is an intermediate image representation. This is a piecewise constant representation of the property map by the function $f(i,j)$, which is a certain linear function of components of the primary feature vector $\boldsymbol{x}=[x_1,\ldots,x_n]$ in point (i,j), i.e. $f(i,j)= z(i,j)$. It is also supposed that a zero-mean perturbation term $\lambda\cdot\nu(i,j)$ with a unit variance is present in the property map $f(i,j)$:

$$f(i,j) = h(i,j) * [\lambda\cdot\nu(i,j) + \sum_{l=0}^{1} \tau_l\cdot\varphi_l(i,j)] \ , \tag{2.2}$$

where $\{\tau_l,\ l=0,1\}$ are two constant intensity values of image plane segments corresponding to the background and objects of interest, respectively, $\varphi_1(i,j)$ is the binary map for objects, $h(i,j)$ is the smoothing kernel of a linear smoothing filter denoted by the convolution sign * , λ is the noise standard deviation. The function $\varphi_1(i,j)$ is equal to zero in the whole image plane Π except for the points belonging to objects of interest, whereas the function $\varphi_0(i,j)=1-\varphi_1(i,j)$ for $\forall\ (i,j)\in\Pi$.

The planar shape modeling is aimed at a concise shape representation of possible objects of interest whose property map satisfies the model by Eq. (2.1-2.2). It consists of a description of shape constraints for the representation of object binary map $\varphi_1(i,j)$ in Eq. (2.2). An efficient approach to describe the planar shape is the multi-scale morphological image modeling which defines objects of interest by using structuring elements and piecewise-linear skeletons [10]. In the underlying morphological model, one initial structuring element S_0 of minimal size as a set of points on the image grid is selected that determines the size and resolution of the objects. The structuring element at the scale m in a uniform scale system is formed as a consecutive binary

dilation (denoted by $\oplus$) by S_0, $S_m = S_{m-1} \oplus S_0$, m=1,2,...,K-1, where K is the total number of scales. The generation of the planar shape of a simple object can be modeled in the continuous case by a growth process along *generating lines* [10].

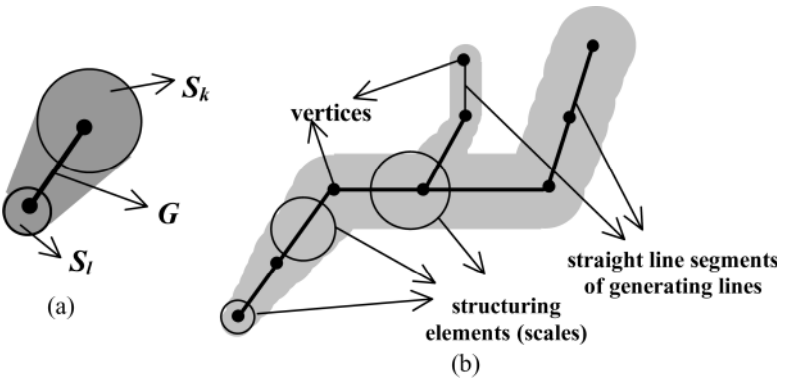

Fig. 1. Multi-scale formation of a simple object of interest (b) by the concatenation of blob-like objects (a)

A *local scale* value is assigned to each vertex point and generating lines are represented as concatenations of straight-line segments. A *blob-like* object defined by its two vertices is formed by two structuring elements S_k and S_l corresponding to the end vertices of a given straight-line segment G (see Fig. 1a) The domain region U of a blob-like object is formed by using the operation of dilation of a generating straight line segment (set) G with a variable in size structuring element (scale), $S(G)$:

$$U = G \oplus S(G) = \bigcup_{(i,j)\in G} S_{m(i,j)}(i,j), \text{ and } m(i,j) = \alpha_k(i,j)\cdot k + \alpha_l(i,j)\cdot l, \quad (2.3)$$

where $S_m(i,j)$ is the structuring element with a variable size m, k and l are the sizes of the structuring elements S_k and S_l, $\alpha_k(i,j)$ and $\alpha_l(i,j)$ are the two ratios of distances of the current point (i,j) to the end points of the segment G. A simple model is adopted for multi-scale object formation using the blob-like objects at different scales: an object of interest is formed from blob-like objects by a concatenation of their vertices, start and end points (see Fig. 1b). Finally, this morphological planar shape model is coupled with the model of image property map by Eq.(2.2) in such a way that the function $\varphi_1(i,j)$ in Eq. (2.2) satisfies the described morphological model.

3 Multi-scale Relevance Function of Images

3.1 Definition of the Relevance Function

Here, an improved model-based relevance function is presented as a modification of the relevance function approach that was initially described in [11]. First of all, it is considered as applied not to the initial image $g(i,j)$ but to the property map $f(i,j)$ represented by Eq. (2.1-2.2). The point on the image plane located on an object generating line, which corresponds to the maximal value of the likelihood function,

allows optimal localization of the object of interest. Two basic local characteristics (constraints) of the image property $f(i,j)$ are involved in the definition of the relevance function: local object-to-background contrast, x, and homogeneity of object, y. Considering a single scale S_k , let the object sub-region $O(i,j)$ be a symmetric structuring element centered at point (i,j) and the sub-region $B(i,j)$ be a ring around it generated by the background structuring element, i.e. $O=S_k$ and $B=S_{k+1}\setminus S_k$ (see Fig. 2). The local contrast can be defined as the difference between mean value of object with a disk structuring element and background intensity within a ring around it:

$$x = \frac{1}{|O|}\sum_{(m,n)\in O(i,j)} f(m,n) - \frac{1}{|B|}\sum_{(m,n)\in B(i,j)} f(m,n).$$

The homogeneity of object y is measured by the difference between an object *intensity of reference, a,* and local (current) estimated intensity. The two constraints x and y take into account all object's potential scales in the definition of the multi-scale relevance function:.

$$\frac{1}{|O|}\sum_{(m,n)\in O(i,j)} f(m,n) \leftrightarrow \frac{1}{K}\sum_{k=0}^{K-1}\left(\frac{1}{|S_k|}\sum_{(m,n)\in S_k} f(m,n)\right),$$

where the object mean intensity is averaged over all K scales $\{S_k \subseteq O(i,j)\}$, $|S_k|$ denotes the number of points in S_k (see Fig. 2a). Similarly, the multi-scale estimation of the background intensity is made by averaging over K single-scale background regions (see Fig. 2b).

The object position, a *focus of attention* (i_f, j_f), is determined as the point in which the joint probability $P(x,y/\text{object})$ will be maximal provided the object point is being considered:

$$(i_f, j_f) = \arg\max_{(m,n)\in A} \{P(x(m,n)/object)P\{(y(m,n)/object)\} \qquad (3.1)$$

where A is a region of interest, which might be the whole image plane. In the proposed definition of the relevance function, it was supposed that $P(y/\text{object})$ follows a Gaussian distribution $\mathbf{N}(0;\sigma_y^2)$ and $P(y/\text{object})$ is also approximated by a normal distribution law $\mathbf{N}(h;\sigma_x^2)$, where h is the mean value of the object local contrast. It can be easily proved that the maximization of joint probability by Eq. (3.1) in the conditions of the assumed model is reduced to the maximization of the image relevance function:

$$R(i,j) = \left(\frac{1}{|O|}\sum_{(m,n)\in O(i,j)} f(m,n) - \frac{1}{|B|}\sum_{(m,n)\in B(i,j)} f(m,n)\right)^2 - \alpha\left(a - \frac{1}{|O|}\sum_{(m,n)\in O(i,j)} f(m,n)\right)^2 \qquad (3.2)$$

For the case of the assumed model without noise the relevance function takes the maximum at the start or end point of a blob-like object. Insignificant shift in the location might be introduced by the present noise depending on the noise variance.

The relevance function $R\{f(i,j)\}$ have to be computed within a region of interest A and takes its maximal value in the focus of attention (i_f, j_f).

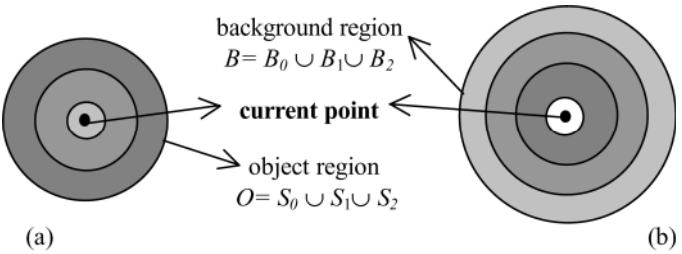

Fig. 2. An illustration to the definition of a three-scale relevance function. Kernel functions for the estimation of object intensity (a) and background intensity (b) are shown as gray levels

3.2 Robust Anisotropic Estimation of Object Intensity

The approach of relevance function is more suitable for large in size objects and low level of noise in the model of property map by Eq. (2.2). Often, thin and elongated low-contrast objects to be detected appear in real images. The simple estimation of average object intensity yields poor results since the object-to-background contrast x in Eq. (3.1) will be low. The remedy to such a situation is the anisotropic estimation of object intensity at certain expenses of the computational complexity. It is based on the morphological image model and the notion of so-called *object structuring regions*. The lth object structuring region V_l^k, l=1,..,L, at scale k is a sub-region of a dilation of a straight-line segment of generating lines with slope θ_l by the scale S_k. Object structuring region V_0^k at scale k coincides with the kth structuring element, i.e. it is a disk of radius r_k. Some examples of object structuring regions are given in Fig. 3. The concept of structuring regions and their derivation from the object morphological model was first introduced in the context of adaptive intensity estimation and image filtering [10]. The object intensity is estimated adaptively depending on the object orientation for the case of elongated object parts and edges.

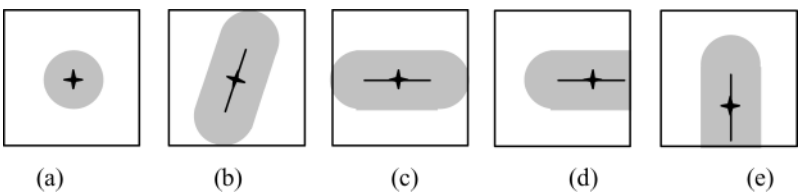

Fig. 3. Examples of object structuring regions (shaded areas) used in the robust estimation of object intensity

This approach can be successfully applied for the robust parameter estimation when computing the relevance function. First, average object intensities $\{q_l\}$ and local variances $\{s_l\}$ are computed inside all the object structuring regions. The average intensity value in region V_μ^k is selected as the result of intensity estimation, where μ

is the structuring region with minimal variances among all L regions. It is clear that such a decision coincides with the maximum likelihood estimation of intensity when assuming Gaussian distributions for point-wise deviations of intensities from the mean value inside respective structuring regions.

3.3 Estimation of Local Scales and Extraction of Planar Shapes

The location of an object of interest as the focus of attention point is followed by the determination of its potential scale and orientation in order to ensure a size-invariant recognition. On the other hand, such a preliminary estimation of scale simplifies the further image analysis provided the estimation is computationally simple. For example, the potential object scale is determined by the maximal value of absolute difference of intensities within a disk S_k and a ring around it, $R_k=S_{k+1}\setminus S_k$, for all k=0,1,...,K-1.

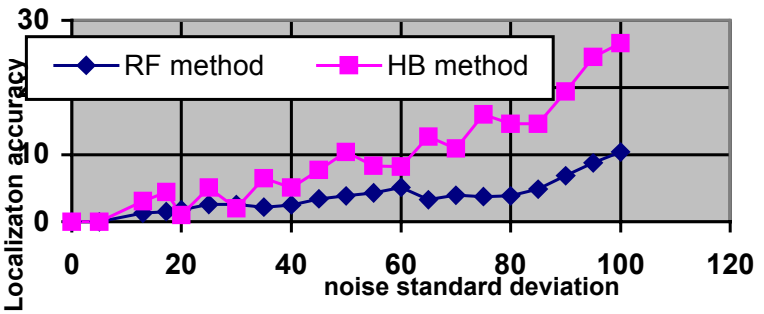

Fig. 4. Localization accuracy (in pixels) vs. noise deviation for two methods: relevance function (RF) and histogram-based binarization (HB)

The proposed model of visual attention mechanism can be successfully applied to time-efficient detection of objects of interest and its shape description by binarization and piecewise-linear skeletonization. In this framework, the object detection consists of several (many) consecutive stages of a multi-scale local image analysis while each of them is aimed at the determination of the next salient maximum of the relevance function [11]. A statistical hypothesis, the so-called *saliency hypothesis*, is first formulated and tested concerning whether an object is present or not with respect to the current local maximum of the relevance function. Statistically, the estimated value of actual object contrast x in Eq. (3.1) is tested on its significance. For this purpose, the result of scale estimation is used in order to estimate the contrast value in a better way. If the hypothesis testing result is positive then the current point is selected as a vertex of object skeleton. The image fragment in the neighborhood (region of attention) of the current attention point is binarized in order to have local binary shape of detected object of interest [11]. If using the property map as an input image, the threshold value is the mean value of object intensity and the background intensity.

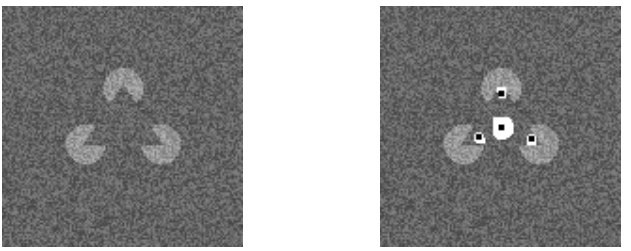

Fig. 5. Experiments with the illusion of Kanisza triangle: (a) - initial noisy image of an imaginary triangle; (b) - result of attention mechanism (maximum points of the relevance function) starting at large scales

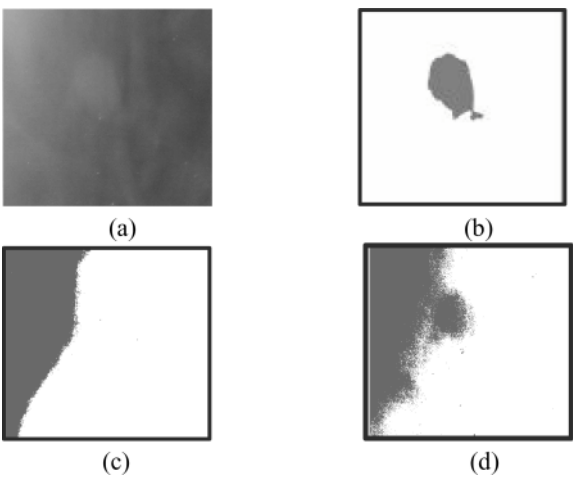

Fig. 6. Results of lesion detection and segmentation in an X-ray image fragment of lungs with a significant slope of intensity: (b) - using the multi-scale relevance function; (c) and (d) using the histogram-based binarization [12]

4 Experimental Results

The relevance function approach to the detection of objects has been tested on synthetic and real images from industrial and medical diagnostic imaging. The main purpose of testing on synthetic images was the performance evaluation during the localization of low-contrast and noisy images. For example, the graph in Fig. 4 shows the experimental dependence of the location bias on the noise level for the noisy image of a bar-like object of known position. For comparison, the object center has been determined in the result of a wavelet transform [7] followed by a histogram-based binarization [12] with subsequent computation of the image centroid.

Several shape and intensity illusions can be modeled (i.e. explained) by the above described visual attention mechanism. Such known examples of illusions connected to the planar shape of objects are the Kanizsa figures (see Fig. 5) [3]. The application of

the relevance function at larger scales yields the focus of attention at the centers of the illusionary triangle in Fig. 5a. The next three local maximums of the relevance function are located at the corners of the Kanisza figures (Fig. 5b). After the local binarization, the local fragments in the respective regions of attention are then identified as corners and the whole object as an illusionary triangle.

The proposed object detection method using relevance function has been tested on real images from diagnostic imaging where the visual attention model has its suitable application areas. The objects of interest are defect indications (quality inspection) or abnormalities of a human body (medical diagnostics), which are usually small in size, low-contrast and located on inhomogeneous backgrounds. One such example is related to lesion detection in radiographic images of lungs (see Fig. 6). Here, the property map has been obtained by a linear clustering transformation of such primary features as three polynomial coefficients of linear polynomial intensity. Such polynomial model is an adequate representation of image intensity when a significant slope value is presenting the background intensity. The result of lesion detection and binarisation is shown in Fig. 6b. The application of the method of histogram-based binarisation [12] gives poor results of shape extraction because of the significant slope in the background intensity even after making a correction to the threshold position on the histogram (see Fig. 6c and 6d for comparison).

5 Conclusions

A model for visual attention mechanisms has been proposed in the context of object detection and recognition problems. A multi-scale relevance function has been introduced for time-effective and geometry-invariant determination of object position. As compared to known visual attention operators based on the standard multi-resolution analysis and wavelet transform, this method has several distinctive features. Firstly, it is a model-based approach, which incorporates some structural features of sought objects in the design of the relevance function. Secondly, it provides a tracking capability for the case of large and elongated objects with complex shape due to the constraint of object homogeneity. The third advantage of this approach is the possibility to treat images with inhomogeneous backgrounds and textured appearance of objects because of working with the property map as an intermediate image representation. It exhibits a high localization accuracy at the same computation time as compared to the multi-resolution approach.

References

1. V Cantoni, S. Levialdi and V. Roberto, Eds., *Artificial Vision: Image Description, Recognition and Communication,* Academic Press, (1997).
2. L. Yarbus, *Eye movement and vision*, Plenum Press, N. Y., (1967).
3. M. D. Levine, *Vision in Man and Machine*, McGraw-Hill, (1985).

4. T. Lindeberg, "Detecting salient blob-like image structures and their scale with a scale-space primal sketch: a method for focus of attention", *Int. Journal of Computer Vision*, Vol. 11, (1993) 283-318.
5. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", *IEEE Trans., Vol. PAMI-20,* No. 11, (1998) 1254-1259.
6. J. K. Tsosos *et al*., "Modeling visual attention via selective tuning", *Artificial Intelligence*, Vol. 78, No. 1-2, (1995) 507-545.
7. J. L. Starck, F. Murtagh, and A. Bijaoui, *Image Processing and Data Analytsis: the Multiscale Approach*, Cambridge University Press, Cambridge, (1998).
8. D. Reisfeld *et. al.,* "Context-free attentional operators: the generalized symmetry transform", *Int. Journal of Computer Vision*, Vol. 14, (1995) 119-130.
9. H. D. Tagare, K. Toyama, and J.G. Wang, "A maximum-likelihood strategy for directing attention during visual search", *IEEE Trans., Vol. PAMI-2,* No. 5, (2001) 490-500.
10. R. M. Palenichka and P. Zinterhof, "A fast structure-adaptive evaluation of local features in images", *Pattern Recognition*, Vol. 29, No. 9, (1996) 1495-1505.
11. R. M. Palenichka and M. A. Volgin, "Extraction of local structural features in images by using multi-scale relevance function", *Proc. Int. Workshop MDML'99*, LNAI 1715, Springer, (1999) 87-102.
12. P. K. Sahoo *et al*., "A survey of thresholding techniques", *Computer Vision, Graphics and Image Process.,* Vol. 41, (1988) 233 -260.