# Integration of Bottom–Up and Top–Down Cues for Visual Attention Using Non–Linear Relaxation

Ruggero Milanese* Harry Wechsler
Sylvia Gil Jean–Marc Bost Thierry Pun

TR-94-014

March 1994

**Abstract**

Active and selective perception seeks regions of interest in an image in order to reduce the computational complexity associated with time–consuming processes such as object recognition. We describe in this paper a visual attention system that extracts regions of interest by integrating multiple image cues. Bottom–up cues are detected by decomposing the image into a number of feature and conspicuity maps, while a–priori knowledge (i.e. models) about objects is used to generate top–down attention cues. Bottom–up and top–down information is combined through a non–linear relaxation process using energy minimization–like procedures. The functionality of the attention system is expanded by the introduction of an alerting (motion–based) system able to explore and avoid obstacles. Experimental results are reported, using cluttered and noisy scenes.

*E-mail: milanese@icsi.berkeley.edu. H.W. is with the George Mason University, Fairfax, VA. SG, JMB, and T.P. are with the University of Geneva, Switzerland.

# Integration of Bottom–Up and Top–Down Cues for Visual Attention Using Non–Linear Relaxation

Ruggero Milanese[1,2*] Harry Wechsler[3] Sylvia Gil[1]
Jean–Marc Bost[1] Thierry Pun[1]

[1]Dept. of Computer Science
University of Geneva
Geneva, Switzerland

[2]Intl. Computer Science Inst.
1947 Center St.
Berkeley, CA 94704, USA

[3]Dept. of Computer Science
George Mason University
Fairfax, VA 22030, USA

**Abstract**

*Active and selective perception seeks regions of interest in an image in order to reduce the computational complexity associated with time–consuming processes such as object recognition. We describe in this paper a visual attention system that extracts regions of interest by integrating multiple image cues. Bottom–up cues are detected by decomposing the image into a number of feature and conspicuity maps, while a–priori knowledge (i.e. models) about objects is used to generate top–down attention cues. Bottom–up and top–down information is combined through a non–linear relaxation process using energy minimization–like procedures. The functionality of the attention system is expanded by the introduction of an alerting (motion–based) system able to explore and avoid obstacles. Experimental results are reported, using cluttered and noisy scenes.*

## 1 Introduction

Visual attention is the capability of biological visual systems to rapidly detect interesting parts of the visual input, in order to reduce the amount of data for complex processing tasks such as feature binding and object recognition [2] [6]. Low–level features such as color, orientation, and curvature are computed by specialized areas of the cortex, and allow to detect regions of interest according to bottom–up, data–driven criteria [6]. High–level features providing integrated, invariant representations for object recognition are computed by higher cortical areas, providing top–down cues for attention [2]. An additional, *alerting* strategy for the extraction of attention regions is represented by the collicular pathway, which detects moving objects entering the subject's field of view [6] [4].

The definition of a computational model of human attention has received considerable interest [5] [1] [11]. Some biologically–plausible systems have been proposed, which can be applied to synthetic images, or other simple images containing alphabetical characters [7] [10] [8] [3]. In most of these systems the selection of "locations" of interests is based on simple features, such as corners and edges. This paper proposes a strategy to extend the capabilities of previous models by extracting and integrating more complex information. This makes it suitable for applications to real images, containing noisy, textured objects.

Figure 1 outlines the main system components and their relations. Both cases of a static image and of a dynamic image sequence have been considered. In the static case, the current RGB color frame is first analyzed by the bottom–up subsystem, which extracts salient regions according to data–driven criteria. This is done in two stages: by extracting a number of *feature maps* $F^k_{x,y}$ $k = 1, ..., K$ (e.g. orientation, curvature, color contrast), and a corresponding number of *conspicuity maps* (C-maps) $C^k_{x,y}$, which enhance regions of pixels largely differing from their surround.

The next stage is represented by the *integration process* which merges the C-maps into a single *saliency map*. This is obtained through a relaxation process, which modifies the values of the C-maps, until they identify a small number of convex regions of interest. An additional, source of information is generated by the top–down subsystem. An object recognition technique based on a *distributed associative memory* (DAM) is used to detect regions of the image which

*Electronic mail: milanese@icsi.berkeley.edu. This work was supported by the Swiss Fund for Scientific Research (NRP-23 4023-027036).

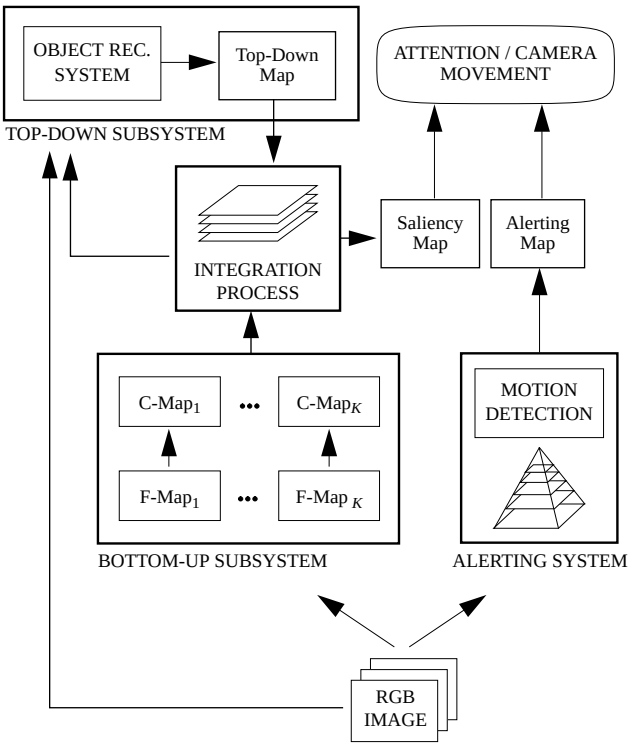


Figure 1: Overview of the attention system.

match with some stored models. The output of the DAM, called the *top–down* attention map, represents an additional input to the relaxation process which defines the saliency map.

In the time–varying case, the image sequence is analyzed by the *alerting subsystem*, which uses a pyramidal representation of the input to provide a fast, rough detection of objects moving against a static background. This pathway is normally ineffective, until an object eventually enters the field of view. In this case, it takes the control over the rest of the system and it directly elicits an attention/camera movement.

## 2 The bottom–up subsystem

Psychophysical experiments have shown that human subjects rapidly detect interesting regions according to multiple bottom–up criteria. For this reason, several feature maps are extracted. Two feature maps are obtained by color–opponency filters and are spatially defined by a 2-D Gaussian profile: $F^{red/green}_{x,y} = R'_{x,y} - G'_{x,y}$ and $F^{blue/yellow}_{x,y} = B'_{x,y} - \frac{R'_{x,y}+G'_{x,y}}{2}$, where $R'G'B'$ are the normalized $RGB$ components of the image, convolved with a Gaussian.

The achromatic component of the image, i.e. the intensity plane $I = R+G+B$, represents an additional information pathway for the attention system, and is convolved with a bank of oriented band–pass filters called the Gaussian 1st derivative: $\mathrm{GD}_1(x,y,\vartheta) = -A \cdot \frac{x\cos\vartheta + y\sin\vartheta}{\sigma_x} \cdot \exp\left[\frac{-(x\cos\vartheta+y\sin\vartheta)^2}{2\sigma_x^2}\right] \cdot \exp\left[\frac{-(-x\sin\vartheta+y\cos\vartheta)^2}{2\sigma_y^2}\right]$, where $A = 1/(2\pi\sigma_x\sigma_y)$. This filter approximates the receptive field profile of a large class of V1 cells [6], and is used at 16 different orientations to provide the *local orientation* feature map: $F^{orient}_{x,y} = \mathrm{argmax}_\vartheta \ (I_{x,y} \star \mathrm{GD}_1(x,y,\vartheta))$, and the *edge magnitude* feature map: $F^{magn}_{x,y} = \max_\vartheta \ (I_{x,y} \star \mathrm{GD}_1(x,y,\vartheta))$.

An additional achromatic map representing high–frequency information is the *local curvature*, computed by considering the intensity image $I$ as a surface, and by applying the divergence operator to the normalized gradient of $I$: $F^{curv}_{x,y} = \mathrm{div}\left[\frac{\nabla I}{\|\nabla I\|}\right](x,y)$. Finally, when no color information is available, the intensity image $I$ is also used as a further achromatic feature map.

The feature maps described above are analyzed by a "conspicuity" operator to assign a bottom–up measure of interest to each location. This measure compares local values of the feature maps to their surround. To this end, another bank of multiple–scale, *difference of oriented Gaussians* (DOOrG) filters is used. Both Gaussians are elliptic rather than isotropic, with an eccentricity factor $r_2 = \frac{\sigma_y}{\sigma_x}$. This property defines a preferential direction $\vartheta$ for the filter which allows to better detect oriented blob–like regions from the feature maps. The scale ratio of the two Gaussians is also fixed: $r_1 = \frac{\sigma_{off}}{\sigma_{on}}$. Each DOOrG filter is computed at 8 orientations, i.e. half the number of orientations defined for the $\mathrm{GD}_1$ filters. For a certain orientation $\vartheta$ this corresponds to: $\mathrm{DOOrG}_{x,y}(r_1,r_2,\vartheta,\sigma) = B_1\Phi_{x,y}(\sigma,\sigma r_2,\vartheta) - B_2\Phi_{x,y}(\sigma r_1,\sigma r_1 r_2,\vartheta)$, where each function $\Phi_{x,y}(\sigma_x,\sigma_y,\vartheta)$ is a 2-D oriented Gaussian, and the constants $B_1$ and $B_2$ are defined so that the sum of the coefficients of each component is normalized to 1. This also implies that the DOOrG filter has zero dc component, yielding zero response to a constant feature map.

To get rid of the sign of the response, and to increase the contrast, the results of convolution are rectified and squared. This corresponds to computing a bank of multiscale conspicuity maps, for 3 values of the scale parameter $\sigma_i$ and eight orientations $\vartheta_j$: $C^k_{x,y}(\sigma_i,\vartheta_j) = (F^k_{x,y} \star \mathrm{DOOrG}_{x,y}(\sigma_i,\vartheta_j))^2$. In order to obtain a unique conspicuity map for each feature, the

$\sigma_i, \vartheta_j$ parameters are factored out by taking the local maximum: $C^k_{x,y} = \max_{i,j}\left(C^k_{x,y}(\sigma_i, \vartheta_j)\right)$.

## 3 The integration process

In order to combine the C-maps into a single saliency map $S$, their average value can be used: $S_{x,y} = \frac{1}{K}\sum_{k=1}^{K} C^k_{x,y}$. However, in virtually all practical cases, this provides noisy, and ambiguous results. For this reason, a relaxation process is applied to the C-maps, so that $S$ will finally approach a binary map, containing a limited number of convex regions.

The relaxation process is defined by a non–linear updating rule: $C^k_{x,y}(t+1) = C^k_{x,y}(t) + \gamma^k_{x,y}(t) \cdot \Delta^k_{x,y}(t)$, for each element: $x, y = 1, .., W,\ k = 1, .., K$. The quantity $\Delta^k_{x,y}$, representing the most important part of the increment, is obtained by minimizing an energy functional $E$ through a gradient–descent procedure. The term $\gamma^k_{x,y}$ is a scaling coefficient depending on the values of both $C^k_{x,y}$ and $\Delta^k_{x,y}$, and is described below.

The energy function $E$ is the linear combination of four different functions: $E = \sum_{i=1}^{4} \lambda_i E_i$, each representing a measure of "incoherence" of the configuration of the C-maps. $E_1$ represents the local *inter–map* incoherence, i.e. the fact that the C-maps enhance different, conflicting regions of the image. This is computed through the sum of local "variances" across different C-maps: $E_1 = A_1 \cdot \sum_{x,y}\sum_k (C^k_{x,y} - \frac{1}{K}\sum_h C^h_{x,y})^2$, where $A_i, i = 1..4$ are scaling constants. The second energy component represents the *intra–map* incoherence, i.e. the inadequacy of each C-map as a representation of a few convex regions of attention. This is evaluated through the overall response of the Laplacian operator: $E_2 = A_2 \cdot \sum_k \sum_{x,y} \left(\nabla^2 C^k_{x,y}\right)^2$. To avoid that the regions of attention grow to include an excessive portion of the image, the third energy component penalizes a configuration of C-maps whose overall activity is too high. This forces the C-maps to share a limited amount of global activity. This is obtained through a competitive relation between the each local value $C^k_{x,y}$ and the average value of all pixels which are located outside a local neighborhood $N(x,y)$ centered on $(x,y)$: $E_3 = A_3 \cdot \sum_k \sum_{x,y} (C^k_{x,y} - m) \cdot \sum_{(u,v)\notin N(x,y)} (C^k_{u,v} - m)$, where $m, M$ are the minimum and maximum values of all the C-maps. The fourth energy measure is introduced to force the values of the C-maps to either one of the extrema of the range $[m, M]$. $E_4$ is thus proportional to the distance of each $C^k_{x,y}$ to both extrema: $E_4 = A_4 \cdot \sum_{x,y} (C^k_{x,y} - m) \cdot (M - C^k_{x,y})$.

The values of the constants $A_i$ are chosen so that $|\ \frac{\partial E_i(t)}{\partial C^k_{x,y}}\ | \leq 1,\ \forall i$. In addition, $\lambda_i \in [0,1]$, and $\sum_i \lambda_i = 1$. To obtain the actual increment to $C^k_{x,y}(t)$, $\Delta^k_{x,y}$ is multiplied by the scaling coefficient $\gamma^k_{x,y}$, defined as: $M - C^k_{x,y}$ if $\Delta^k_{x,y} \geq 0$, and by $C^k_{x,y} - m$ otherwise. This guarantees that the new value $C^k_{x,y}(t+1)$ will remain in the allowed range $[0,1]$.

The convergence criterion for the relaxation process is defined by: $\max_{x,y,k} |\ \gamma^k_{x,y}(t) \cdot \Delta^k_{x,y}(t)\ | < \varepsilon$, where $\varepsilon$ is an appropriately small constant. For most images, a value of $\varepsilon = 0.01$ requires about a dozen iterations. At convergence, the binary *attention map* is obtained by thresholding the saliency map $S(t)$ in the middle of the range $[m, M]$.

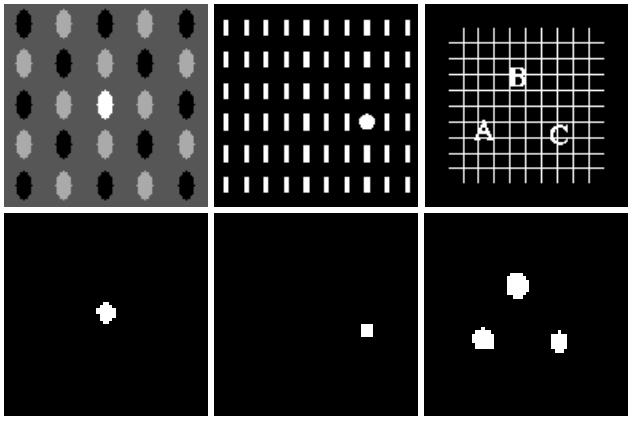


Figure 2: Results on some synthetic images.

Figure 2 shows the results on some synthetic images defining visual search problems. The selected regions allow to reproduce well–known *pop–out* phenomena. Figure 3 shows the results on some real images. The attention regions are correctly located at some of the major foreground objects. It should be noticed that the limited number of regions is a by–product of the 3rd energy component, which penalizes high amounts of global activity.

## 4 The alerting subsystem

Data–driven attention regions are also produced by the alerting subsystem, which detects the shape of objects moving on a still background through a low–pass pyramidal representation, built for each image frame by using a set of $\beta$–splines basis functions [4]. At each level $l$ of the pyramid ($l = 0, ..., \log_2 W$), first estimates of motions are obtained by computing temporal image differences $D^l_{x,y}(t) = I^l_{x,y}(t) - I^l_{x,y}(t-1)$. Local differences $D^l_{x,y}(t)$ provide two motion parameters,

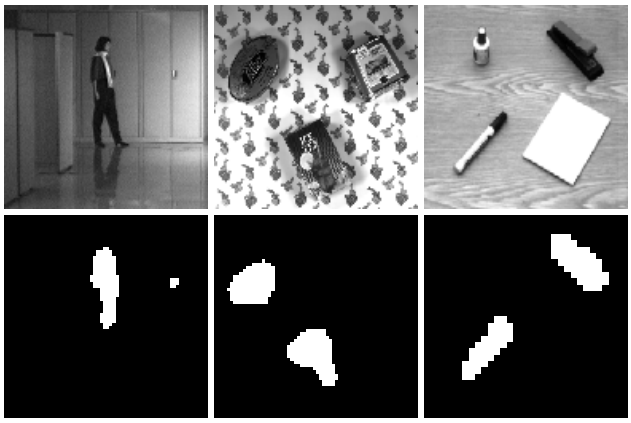

Figure 3: Results on some real images.

through their magnitude, and through the locations of sign changes. These two factors are locally combined together to form the first motion estimates $E^l_{x,y}$ (cf. figure 4.b). Highest–resolution levels have a better spatial localization, but may only yield information at the object boundaries. Lower–resolution levels help solve the aperture problem, by filling in the interior of moving objects having constant grey level.

Multiple–resolutions motion estimates $E^l_{x,y}$ are combined through a coarse–to–fine pyramidal relaxation process. Its goal is to locally propagate the pixel values *horizontally* within each level as well as *vertically*, across contiguous levels of the pyramid. The "vertical" component of the relaxation process combines information at location $(x, y)$ of level $l$ with that at locations $(2x+i, 2y+j)$, $i, j \in \{0, 1\}$ at the higher resolution level $l-1$. The "horizontal" component consists of a diffusion process within each pyramid level, to fill in gaps and reduce noise.

In a similar way to the relaxation process of the C-maps, the updating rule of the vertical component is defined by a term $\gamma^l_{x,y} \cdot \Delta^l_{x,y}$, where $\gamma^l_{x,y}$ is a scaling coefficient. The increment $\Delta^l_{x,y}$ is defined as a function of $D^{l+1}$. If $D^{l+1}_{x/2,y/2}$ is smaller than a threshold $\xi$ (proportional the estimated image noise), then $\Delta^l_{x,y}$ is the quadratic function $-k_1 \cdot \left(D^{l+1}_{x,y} - \xi\right)^2$. Otherwise, $\Delta^l_{x,y} = g\left(D^{l+1}_{x,y} - k_2 \cdot \xi\right)$, where $g(\cdot)$ is a sigmoidal function, and $k_1$ and $k_2$ are two positive constants. This algorithm corresponds to pushing the values of the estimates $E^l_{x,y}$ further towards either 0 or 1.

At the end of this algorithm the full–resolution image at the bottom of the pyramid contains a binary *alerting map* of the moving objects. Thanks to the diffusion component of the relaxation process, the shape of these regions tends to be "convex", and to adapt to

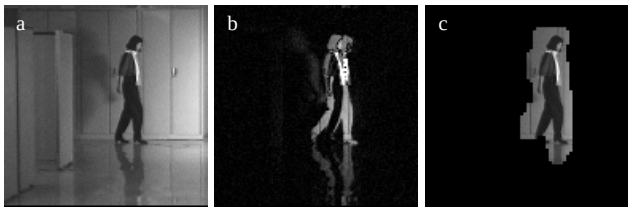


Figure 4: The alerting system. (a) second frame; (b) motion estimates $E^0$; (c) final alerting map $D^0$.

the shape of the underlying objects. Figure 4 shows the results obtained when a persons is walking through a corridor.

## 5 The top–down subsystem

Distributed associative memories (DAM) are a simple but effective technique to learn object categories from training samples. During the recognition phase, a DAM can be used to recognize target objects, and hence, to generate top–down measures of interest. However, a preprocessing step is required to provide some degree of invariance to the representation of the input image.

This preprocessing step is based on the complex–log (or log–polar) transform of the input image [9]. Given a center point represented by a complex number $z_0 = x_0 + jy_0$, this transform maps a point $(x, y)$ of the image into the coordinates $z = \log(\sqrt{(x-x_0)^2 + (y-y_0)^2}) + j\text{atan}(\frac{y-y_0}{x-y_0})$. This transformation allows to simulate the focal/peripheral fields of an image, and maps scalings and rotations into translations in the real and imaginary axes. These shifts can be factored out by considering the energy spectrum $\mid \mathcal{F}(u, v) \mid$ of the complex–log image.

The components of $\mid \mathcal{F}(u, v) \mid$ are ordered in a vector $\mathbf{x}$ representing the input stimulus to the DAM. During the learning phase, the DAM finds an association matrix $\mathbf{M}$ between a set of input stimuli $\mathbf{x}_h$ and their class $\mathbf{y}_h$. If all stimulus and response vectors are written in two matrices $\mathbf{X}$ and $\mathbf{Y}$, $\mathbf{M}$ is defined by $\mathbf{Y} = \mathbf{MX}$, and is solved by minimizing $\|\mathbf{MX} - \mathbf{Y}\|^2$. This corresponds to $\mathbf{M} = \mathbf{YX}^+$, where $\mathbf{X}^+ = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is the Moore–Penrose generalized inverse of the matrix $\mathbf{X}$.

During the recognition phase, an unknown stimulus vector $\mathbf{x}'$ is presented to the memory matrix $\mathbf{M}$, and the estimated class can be recovered from the output vector $\mathbf{y}'$. Through a statistical interpretation of DAMs in terms of multiple linear regression, a *coeffi-*

*cient of determination* $R^2 = (\mathrm{var}(\mathbf{x}') - \mathrm{RSS})\,/\mathrm{var}(\mathbf{x}')$, is obtained for each association produced by the DAM on an unknown stimulus $\mathbf{x}'$, where RSS is the residual sum of squares [9]. The value of $R^2 \in [0,1]$ evaluates the quality of the association: it is 1 for a perfect association, and 0 when no correlation exists between the stimulus and the produced response.

The top–down measure of interest is given by the $R^2$ measure, representing the "quality" of the recognition. In order to avoid the application of the DAM to all vectors $\mathbf{x}_{u,v}$ centered at each location $(u,v)$ of the input image, a number of relevant "indexing" points is required. These points are given by the bottom–up subsystem, and are obtained by detecting a limited number of peaks $\{(x_i,y_i), i = 1,...,Q\}$ in the saliency map $S$, after just two iterations of the relaxation process. In order to spread the results of the $R^2$ measures over a neighborhood centered on each point $(x_i,y_i)$, and to obtain a distributed representation for the top–down map $T$, the values $R^2(x_i,y_i)$ are convolved with an isotropic Gaussian filter: $T_{x,y} = \sum_{i=1}^{Q} R^2(x_i,y_i) \cdot \exp\left[-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma_T^2}\right]$.

The top–down map $T$ can directly be integrated with the bottom–up system by modifying the updating rule of the relaxation process (cf. sect. 3). The modified rule is given by: $C^k_{x,y}(t+1) = C^k_{x,y}(t) + \gamma^k_{x,y}(t) \cdot \left[\alpha\Delta^k_{x,y}(t) + (1-\alpha)(2T_{x,y}(t)-1)\right]$. The parameter $\alpha \in [0,1]$ determines the relative importance assigned to the bottom–up and top–down subsystems.

Figure 5 shows the results obtained for a DAM trained to recognize instances of the pen and the white–ink bottle. The top–down map shows a very low $R^2$ value at one peak of the saliency map, corresponding to an unknown object (the cup). The final saliency map obtained by integrating the top–down map with the relaxation process is shown in fig. 5.d. For comparison, the saliency map obtained from the bottom–up system alone is shown in 5.e. The top–down information forces the relaxation process to suppress the region containing the unknown object, although this would have been selected by the bottom–up process, to the expense of the white–ink bottle.

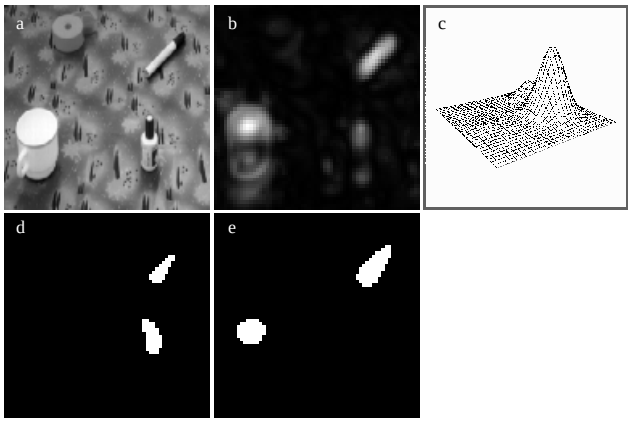


Figure 5: Integration of the top–down map. (a) input image; (b) saliency map $S(2)$; (c) top–down map; (d) results with top–down; (e) results without top–down.

## References


[1] S. Ahmad, *VISIT: An Efficient Computational Model of Human Visual Attention.* Ph.D. Thesis, University of Illinois at UC, 1991.

[2] R. Desimone, *Neural Circuits for Visual Attention in the Primate Brain.* G.A. Carpenter and S. Grossberg (eds), *Neural Networks for Vision and Image Proc.* MIT Press, 1992, 343-364.

[3] G.-J. Giefing, H. Janssen and H. Mallot, *Saccadic Object Recognition with an Active Vision System.* 10th ECAI, 1992, 803-805.

[4] S. Gil and T. Pun, *Non–Linear Multiresolution Relaxation for Alerting.* Eur. Conf. on Circ. Th. and Design, 1993, Elsevier Science, 1639-1644.

[5] C. Koch and S. Ullman, *Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry.* Human Neurob., 4, 1985, 219-227.

[6] R. Milanese, *Detecting Salient Regions in an Image: from Biological Evidence to Computer Implementation.* Ph.D. thesis, Univ. of Geneva, 1993. (Anonim. ftp: cui.unige.ch, pub/milanese).

[7] M.C. Mozer, *The Perception of Multiple Objects: a Connectionist Approach.* MIT Press, 1991.

[8] B. Olshausen, C. Anderson, and D. Van Essen, *A Neural Model of Visual Attention and Invariant Patt. Rec.* Caltech, CNS Memo 18, 1992.

[9] W. Pöltzleitner and H. Wechsler, *Selective and Focused Invariant Recognition Using Distributed Associative Memories.* IEEE PAMI, 12 (8), 1990, 809-814.

[10] P. A. Sandon, *Simulating Visual Attention.* Journal of Cog. Neurosc., 2 (3), 1990, 213-231.

[11] M.J. Swain and M. Stricker (eds), *Promising Directions in Active Vision.* IJCV, 11 (2), 1993, 109-126.