Attentive mechanisms for dynamic and static scene analysis

MILANESE, Ruggero, GIL MILANESE, Sylvia, PUN, Thierry

MILANESE, Ruggero, GIL MILANESE, Sylvia, PUN, Thierry. Attentive mechanisms for dynamic and static scene analysis. *Optical engineering*, 1995, vol. 34, no. 8, p. 2428-2434

DOI : 10.1117/12.205668

Attentive Mechanisms for Dynamic and Static Scene Analysis

Ruggero Milanese
Sylvia Gil
Thierry Pun

Dept. of Computer Science
University of Geneva
24, rue Général–Dufour
1211 Geneva
Switzerland

Phone: +41 (22) 705-7628
Fax: +41 (22) 705-7780
E-mail: {milanese, gil, pun}@cui.unige.ch

Abstract

Attention mechanisms extract regions of interest from image data, in order to reduce the amount of information to be analyzed by time–consuming processes such as image transmission, robot navigation, and object recognition. In this paper two such mechanisms are described. The first one is an alerting system which extracts moving objects in a sequence through the use of multiresolution representations. The second one detects regions in still images which are likely to contain objects of interest. Two types of cues are used and integrated to compute the measure of interest. First, bottom–up cues result from the decomposition of the input image into a number of feature and conspicuity maps. The second type of cues is top–down, and is obtained from a–priori knowledge about target objects, represented through invariant models. Results are reported for both the alerting and the attention mechanisms, using cluttered and noisy scenes.

Subject terms: computer vision, visual attention, invariant object representations, alerting, multi–resolution.

1 Introduction

At the basis of all object recognition techniques is a correspondence problem between a set of features extracted from the image and a number of models. Given the exponential complexity of this problem, it is of fundamental importance that the features describing the input data be the most compact and discriminative. The use of image segmentation and grouping algorithms only partially solves this issue, since their results are very sensitive to noise and to the background underlying the targets. As a result, the high amount of image primitives that must be considered for the matching process still represents a major limitation to real-time applications.

For these applications, there is the need for further information selection mechanisms, that identify a limited number of regions in the image, containing the most relevant information. Since any measure of relevance highly depends on the application at hand, ad–hoc techniques must be designed for each problem. For most applications, however, a number of heuristics exist, that allow to identify relevant parts of the image in a general–purpose way. The first one is object's motion: moving objects represent for instance the major source of information for dynamic obstacle avoidance in robot navigation. The second one is object's *saliency*: regions containing information that statistically differs from the background is likely to identify objects on which the actions of a robot, such as grasping, are to be defined. These two relevance criteria are also applicable in other fields, such as automatic surveillance, and more generally, tasks based on human interaction such as image retrieval from data-bases and image transmission for teleconferencing [17].

Both relevance criteria have been widely studied in human vision. In dynamic scenes, the generation of ocular/attention shifts is highly influenced by an alerting signal corresponding to the detection of object's motion. This allows to foveate on the moving object, and to keep its image stabilized in the fovea through smooth pursuit (tracking) eye movements [6]. In static scenes, visual saliency represents the major information source for the generation of ocular saccades, as well as for shifts of internal attention [10].

Figure 1 shows the proposed implementation for these two mechanisms. The alerting subsystem, described in more detail in chapter 2, is responsible for detecting relevant regions in the time–varying case. It is based on a pyramidal representation of the input sequence, and it allows to rapidly locate moving objects, and to extract a compact approximation of their shape. The attention subsystem, described in chapter 3, is responsible for detecting regions of interest in static images. This is done by integrating bottom–up and top–down attentional cues into a single representation called the *saliency map* through a relaxation process. Bottom–up cues are obtained by extracting a number of feature (F-maps) and conspicuity maps (C-maps). Top–down cues are provided by comparing the images with objects of interest represented through an associative memory.

Insert Figure 1 about here

Several ways to connect the alerting and attention subsystems can be designed, depending on the task at hand. For autonomous robots, for instance, priority must be given to the alerting subsystem, since moving objects represent possible obstacles for the robot. If no moving objects are detected, the saliency map computed by the attention subsystem will be used, allowing the robot to explore partially unknown environments. In this case, it is important to prevent the attention system from repeatedly selecting the same regions. This can be done by storing each new saliency map in a long–term memory called the *history map* [10]. This map can then be used as an inhibitory input to the integration process, allowing new regions to be selected.

2 Alerting mechanism in dynamic scenes

The goal of the alerting subsystem is to rapidly detect and locate moving objects, and to represent them through compact masks approximating their shape. Differential methods are based on the substraction of subsequent frames in order to get rid of the static background and to process only the moving regions of the image. Examples of this method are proposed in [15] [3] [1]. In [15], after performing the difference between successive frames, a 2–D median filter is applied on the difference image in order to smooth the mask boundaries and eliminate small regions. Despite the action of the median filter, the resulting mask appears oversegmented. In [3], spatio-temporal derivatives of three subsequent images are used in order to label image pixels as either dynamic or static by means of maximum–a–posteriori (MAP) regularization. In order to speed up the convergence, deterministic relaxation is used rather than a stochastic one. However, this technique is still highly computationally intensive. The method proposed in [1] is based on global thresholds followed by a local refinement step based on MAP techniques, but the resulting masks do not match our requirements in terms of compactness and localization with respect to the moving objects. Another differential method is based on background substraction, but requires an updated model of the static background in order to isolate dynamic objects [7].

The proposed approach is based on simple temporal differences of subsequent frames and requires only two frames in order to obtain satisfactory results. The key data structure is a low–pass pyramid [8] [9]: $I^l_{x,y}(t), l = 0, ..., L$ (where $L + 1$ is the number of pyramid levels), which is built for each input image frame $I_{x,y}(t)$. The pyramid is computing by using a set of β–splines basis functions, given their compact support [16]. A corresponding number of temporal derivatives $D^l_{x,y}(t)$ are then computed, possibily through simple image differences (thus involving only 2 frames). From each temporal derivative, two complementary quantities are extracted: its magnitude, and the locations of sign changes. High magnitude values are located at moving objects boundaries, while significant sign changes occur in textured patches located at the interior of moving objects [5]. Significant sign changes are represented by binary images, whose values $\psi^l_{x,y} \in \{0, 1\}$ are defined as follows:

$$\psi^l_{x,y} = \begin{cases} 1 & \text{if } D^l_{x+m,y+n}(t) \cdot D^l_{x+u,y+v}(t) < -\vartheta_1 \\ 0 & \text{otherwise ,} \end{cases} \tag{1}$$

where ϑ_1 is a fixed threshold computed according to the sequence noise and $m, n, u, v \in \{-1, 0, 1\}$. Values of $\psi^l_{x,y} = 1$ thus express the presence of significant ($< -\vartheta_1$) sign changes in a 3x3 window centered at pixel x, y. Given the complementarity between the measures $D^l_{x,y}$ and $\psi^l_{x,y}$, these two factors are locally combined together through a *max* operator, in order to form primary motion–detection estimates:

$$E^l_{x,y} = \max\{|D^l_{x,y}(t)|, \psi^l_{x,y}\}. \tag{2}$$

High–resolution levels of $E^l_{x,y}$ detect temporal changes with a high spatial localization, but may only yield information at the object boundaries. Lower–resolution levels help to provide compact and unique masks for each moving object, by filling in regions of constant grey level.

Multiple–resolutions motion–detection 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 $l{+}1$ with that at locations $(2x{+}i, 2y{+}j),\ i, j \in \{0, 1\}$ at the higher resolution level l. The "horizontal" component consists of a diffusion process within each pyramid level, to fill in gaps and reduce noise.

The updating rule of the vertical component is defined by an additive term $\zeta^l_{x,y} \cdot \Delta^l_{x,y}$. The first term $\zeta^l_{x,y}$ is a scaling factor which allows the image to remain in its dynamic range after the increment (cf. definition of $\gamma^k_{x,y}$ in eq. 4). The factor $\Delta^l_{x,y}$ is defined as a function of D^{l+1}. If $D^{l+1}_{x/2,y/2}$ is smaller than a threshold ϑ_2 (proportional the estimated image noise), then $\Delta^l_{x,y}$ is the quadratic function $-k_1 \cdot \left(D^{l+1}_{x,y} - \vartheta_2\right)^2$. Otherwise, $\Delta^l_{x,y} = g\left(D^{l+1}_{x,y} - k_2 \cdot \vartheta_2\right)$, where $g(\cdot)$ is a sigmoidal function, and k_1, k_2 are positive constants ensuring first and second order continuity of $\Delta^l_{x,y}$ at ϑ_2. This algorithm corresponds to pushing the values of the estimates $E^l_{x,y}$ further towards a bimodal distribution image, which is then staightforward to threshold.

At the end of this algorithm the pyramid contains multiple–resolution binary masks $M^l_{x,y}$ of the moving objects. Thanks to the diffusion component of the relaxation process, the shape of these regions tends to adapt to the shape of the underlying objects. However, given their dependency on temporal derivatives computed over multiple frames (≥ 2), and given the existence of non–zero values of $E^l_{x,y}$ on uncovered background, the shape of these regions is generally larger than the underlying objects. A refinement process is thus required, to extract a more accurate representation of the object shapes. To this end, the assumption is made that the shape of the objects is approximately convex. A convex polygonal approximation of the object's contour can then be computed for each region, in a coarse–to–fine way. Within each region R_i^l identified by a coarse–resolution mask $M_{x,y}^l, l > 0$, all points $K_i^l = \{(x,y) \in R^i, \mid D_{x,y}^l \mid > \vartheta_3, \nabla^2 I_{x,y}^l > \vartheta_4\}$ of high spatio–temporal gradients are selected, where ∇^2 is the Laplacian operator and ϑ_3, ϑ_4 are fixed thresholds. The polygonal approximation for the underlying object is then obtained through the convex hull of the set K_i^l. This representation obtained at a coarse level l is then propagated to higher–resolution levels of the pyramid. This can be done very efficiently by restricting the convex hull computation at level $l-1$ in a search window defined as the region enclosed in the convex hull at level l projected to the level $l-1$.

This method has been successfully tested on a variety of real indoors and outdoors image sequences (teleconference, traffic scenes, corridor scenes, etc.). The choice of the four parameters $\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4$ appears not to be critical. Although ad–hoc variations within a range ±5% may lead to slight improvements, a fixed set of values $(\vartheta_1 = 50, \vartheta_2 = 25, \vartheta_3 = 25, \vartheta_4 = 30)$ provided satisfactory results for all sequences.

Figure 2 reports the results obtained by the alerting system on two image sequences, showing for each sequence: the motion masks $M_{x,y}^l$ (for $l = 0$) and the polygonal approximations for the refined masks, respectively at a coarse level $(l = 3)$, and at the highest resolution level $(l = 0)$. It can be seen that, also for complex, non–convex objects such as the walking person, the final results correctly outline the shape of the moving object.

Insert Figure 2 about here

3 Visual attention in static scenes

The goal of the visual attention system is to select regions of interest from the analysis of static scenes. Previous work has been done in two directions. On one hand, biologically–plausible models have been proposed, which simulate human performance on synthetic test images [2] [12]. On the other hand, algorithms have been proposed for the extraction of salient locations in real images [14] [4]. However, salient locations are in these cases identified with simple features, such as corners and edges. In the proposed approach, regions of interest can be computed on the basis of more global properties. This makes it suitable for applications dealing with complex images, containing noisy, textured objects. As shown in figure 1, the attention system integrates two main components, called bottom–up and top–down, which are described in the following sections.

3.1 The bottom–up subsystem

The bottom–up subsystem extracts salient regions according to data–driven criteria. This is done in three stages: (i) extraction of a number of *feature maps* $F^k_{x,y}$ $k = 1, ..., K$, representing the input image according to different criteria; (ii) computation of a corresponding number of *conspicuity maps* (C-maps) $C^k_{x,y}$ which enhance regions containing features that largely differ from their surround; and (iii) integration of the C–Maps into a single *saliency map* $S_{x,y}$, which identifies the selected regions.

From each RGB image, two chromatic and three achromatic feature maps are computed ($K = 5$). The chromatic ones are obtained through color–opponency filters, whose spatial profile is a 2-D Gaussian: $F_{x,y}^{red/green} = R'_{x,y} - G'_{x,y}$ and $F_{x,y}^{blue/yellow} = B'_{x,y} - \frac{R'_{x,y}+G'_{x,y}}{2}$, where $R'G'B'$ are the normalized RGB components of the image, convolved with a Gaussian operator.

The achromatic feature maps are obtained through differential operators applied on the intensity image $I_{x,y}$. These operators correspond to a bank of filters, defined by the oriented Gaussian 1st derivative:

$$\mathrm{GD}_1(x, y, \vartheta) = -\frac{u}{\sigma_x} \cdot \exp\left[\frac{-u^2}{2\sigma_x^2}\right] \cdot \exp\left[\frac{-v^2}{2\sigma_y^2}\right], \tag{3}$$

where $u = x\cos\vartheta + y\sin\vartheta$ and $v = -x\sin\vartheta + y\cos\vartheta$. This filter is used at 16 different orientations to provide both the *local orientation* feature map: $F_{x,y}^{orient} = \mathrm{argmax}_{\vartheta}\ \{I_{x,y} \star \mathrm{GD}_1(x, y, \vartheta)\}$, and the *edge magnitude* feature map: $F_{x,y}^{magn} = \max_{\vartheta}\ \{I_{x,y} \star \mathrm{GD}_1(x, y, \vartheta)\}$. The same derivatives are used to compute a third achromatic feature map defining the *local curvature*, obtained through the divergence operator on the normalized gradient of I: $F_{x,y}^{curv} = \mathrm{div}\left[\frac{\nabla I_{x,y}}{\|\nabla I_{x,y}\|}\right]$.

The five feature maps described above are processed by a "conspicuity" operator to assign a bottom–up measure of interest to each location. This measure is obtained by comparison of local values of the feature maps to their surround. To this end, another bank of multiscale, *difference of oriented Gaussians* (DOOrG) filters is used. Both Gaussians $G_{x,y}^{on}, G_{x,y}^{off}$ are elliptic rather than isotropic, with a fixed eccentricity factor $\frac{\sigma_y}{\sigma_x} = \frac{1}{3}$. This property defines a preferential direction ϑ for the filter which allows to better detect oriented blob–like regions from the feature maps. The coefficients of each Gaussian component are normalized by the constraint $\sum_{u,v} G_{u,v}^{on} = \sum_{u,v} G_{u,v}^{off} = 1$, so that the overall filter has zero DC component, yielding zero response to a constant feature map. The scale ratio of the two Gaussians is also fixed: $\frac{\sigma_{off}}{\sigma_{on}} = 3$. Three different values of σ_{on} are used for each filter, thus giving three classes of multiscale filters. As for the GD_1 filters, each DOOrG filter is also computed at multiple (8) orientations.

To get rid of the sign of the response, and to increase the contrast, the results of the convolution are rectified and squared. This corresponds to computing a bank of multiscale conspicuity maps, for three values of the scale parameter σ_i and eight orientations ϑ_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 σ_i, ϑ_j parameters are factored out by taking the local maximum: $C^k_{x,y} = \max_{i,j}\{C^k_{x,y}(\sigma_i, \vartheta_j)\}$.

3.2 The integration process

The next stage of bottom–up attention requires the integration of the C-maps into a single saliency map S. This is done through a non–linear relaxation process which reduces noise, and increases the coherence of the different C-maps in an incremental way. The saliency map is then obtained by thresholding the average value of the C-maps, once a convergence criterion is satisfied.

At each iteration of the relaxation process, the value of each $C^k_{x,y}(t)$ is updated by an additive factor: $\gamma^k_{x,y}(t) \cdot \Delta^k_{x,y}(t)$. The term $\gamma^k_{x,y}$ is a scaling coefficient defined by:

$$\gamma^k_{x,y} = \begin{cases} M - C^k_{x,y} & \text{if } \Delta^k_{x,y} \geq 0 \\ C^k_{x,y} - m & \text{otherwise }, \end{cases} \tag{4}$$

where m, M are respectively the minimum and maximum values of all C-maps. This coefficient guarantees that the update will keep new values of $C^k_{x,y}(t+1)$ within the original range $[m, M]$.

The quantity $\Delta^k_{x,y}$ represents the most important part of the increment; it is obtained by minimizing an energy functional $E(t)$ through a gradient–descent procedure: $\Delta^k_{x,y} = -\frac{\partial E}{\partial C^k_{x,y}}$. The energy function is the linear combination of four different functions: $E = \sum_{i=1}^{4} \lambda_i E_i$, where each $E_i, i = 1, ..., 4$ represents a measure of "incoherence" of the configuration of the C-maps, and $\lambda_i, i = 1, ..., 4$ are weighting coefficients used to normalize $\Delta^k_{x,y}$ in range $[0, 1]$.

E_1 represents the local *inter–map* incoherence, i.e. the fact that different C-maps enhance different, conflicting regions of the image. This energy term is computed through the sum of local "variances" across different C-maps: $E_1 = \sum_{x,y} \sum_k (C^k_{x,y} - \frac{1}{K} \sum_h C^h_{x,y})^2$. 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: $E_2 = \sum_k \sum_{x,y} \left(\nabla^2 C^k_{x,y}\right)^2$. To avoid the fact that the regions of attention may 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, through a competitive relation between 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): $E_3 = \sum_k \sum_{x,y} (C^k_{x,y} - m) \cdot \sum_{(u,v) \notin N(x,y)} (C^k_{u,v} - m)$. 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: $E_4 = \sum_{x,y} (C^k_{x,y} - m) \cdot (M - C^k_{x,y})$.

The updating term $\gamma^k_{x,y}(t) \cdot \Delta^k_{x,y}(t)$ computed through this algorithm depends on the values of the coefficients $\lambda_i, i = 1, ..., 4$. By appropriately assigning these values it is possible to force the updating term in directions which favor specific energy components. However, this requires a-priori knowledge on the image which is not always available. For this reason, these parameters are assigned values that give equal importance to each energy component, i.e. $\lambda_i = \frac{1}{4} \cdot (\max_{x,y,k} \mid \frac{\partial E_i}{\partial C^k_{x,y}} \mid)^{-1}$.

This method has been used for a large number of different images (currently about one hundred). For most of them, a dozen iterations are sufficient to rapidly converge, i.e. to reduce the absolute value of the updating terms below a fixed threshold, set to 0.01. At convergence, the average sum of the C-maps $\frac{1}{K} \sum_{k=1}^{K} C^k_{x,y}$ is taken as the saliency map S. Thanks to the contribution of the fourth energy component to the updating term, the values of the saliency map are almost binary. For this reason, even if the convergence criterion is not perfectly satisfied, the relaxation process is always stopped after only 12 iterations, and the results of S are binarized by thresholding at the middle of the range $[m, M]$.

Insert Figure 3 about here

Figure 3 shows the results on some synthetic images, used as visual search experiments on human vision. The selected regions allow to reproduce well–known *pop–out* phenomena. Figure 4 shows the results of the integration process on some real images. The attention regions are correctly located at some of the major foreground objects. It should be noticed that only a limited number of regions can be detected in two of these images. This is a consequence of the 3rd energy component, which penalizes an excessive total size in the regions of attention. Since these two images contain several foreground objects, only a few of them could be selected at once. One technique for the retrieval of the remaining ones is to use the *history map* introduced in section 1. This map stores the results of previous saliency maps, and can be used as a further input to the relaxation process, which penalizes locations belonging to previous attention regions. In this way, the system can select an unlimited number of attention regions in an iterative way (cf. [10] for more details).

Insert Figure 4 about here

3.3 The top–down subsystem

The top–down attention subsystem uses knowledge about the task to select the regions of the image most likely to contain objects of interest. This is done by learning descriptions of target objects through distributed associative memories (DAM) [13]. The top–down measure of interest at a location (x, y) is then computed in terms of the similarity of the image contents at that location with the stored models.

In order to ensure some degree of invariance to the representation of the targets, a preprocessing step is required, based on the complex–log (or log–polar) transform of the image [13]. Given a center point (x_0, y_0) of the transform, a complex number is used to represent it in a compact way in the polar–log domain $z_0 = x_0 + jy_0$. 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 along the real and imaginary axes respectively. These shifts can be factored out by considering the energy spectrum $|\,\mathcal{F}(u, v)\,|$ of the complex–log image.

The components of $|\,\mathcal{F}(u, v)\,|$ 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 $\mathbf{X}$.

Once the matrix $\mathbf{M}$ has been constructed, it can be used on a novel stimulus vector $\mathbf{x}'$ to produce a classification through an output vector $\mathbf{y}'$. Through a statistical interpretation of DAMs in terms of multiple linear regression, a *coefficient 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 [13]. 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:

$$T_{x,y} = \sum_{i=1}^{Q} R^2(x_i, y_i) \cdot \exp\left[-\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma_T^2}\right]. \tag{5}$$

The top–down map T can directly be integrated with the bottom–up system by modifying the updating rule of the relaxation process (cf. previous section). The updating term of the modified rule is given by the product between the scaling coefficient $\gamma^k_{x,y}(t)$ defined similarly to the previous section, and a new term $\Theta^k_{x,y}(t) = \left[\alpha\Delta^k_{x,y}(t) + (1-\alpha)(2T_{x,y}(t) - 1)\right]$. 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.

Insert Figure 5 about here

4 Conclusions

In this paper two types of attention mechanisms have been described. The first one analyzes a multi–resolution representation of the spatio–temporal derivatives of an image sequence in order to extract the location and shape of moving objects. The second one processes still images, and discriminates between their features to extract regions containing objects of interest. Although both mechanisms are mostly based on a data–driven approach, it has been shown that it is possible to customize the system through the use of a–priori knowledge of target objects.

Both mechanisms use highly distributed, though iterative computations (cf. pyramidal relaxation and integration of C-maps). However, the number of iterations required for both of them is very limited, being set to a fixed value. The remaining steps are based on simple filtering operations. The overall system can thus be easily implemented using specialized hardware, providing an effective tool to reduce data and computation time for further processes.

Applications of these mechanisms are currently being done in two directions. The alerting system is used for automatic highway–control problems [6]. It allows to count the number of vehicles, providing the initial data for a tracking system which computes vehicle kinetic functions such as trajectory, velocity and acceleration. The attention system is used in several contexts: for defect detection from natural surface images, and for applications of object recognition of man–made objects. The availability of the attention system improves the performance of object recognition in multiple ways. Obviously, it reduces the amount of data to process. Most importantly, however, it allows to hypothesize the presence of a single object within each region of attention, which leads to a huge reduction in the complexity of the matching process. This can be exploited in the use of efficient recognition schemes, such as geometric hashing.

Acknowledgements. We would like to thank Prof. Harry Wechsler (George Mason University) for his contribution to the design of the system. I.R.B.S.A. (Switzerland) kindly provided some of the image sequences. This work was supported by grant no. 4023-027036 from the Swiss Fund for Scientific Research.

References

[1] R. Aach, A. Kaup, R. Mester "Statistical Model-Based Change Detection in Moving Video". *Signal Processing*, Vol. 31, 1993, pp. 165-180.

[2] S. Ahmad, "VISIT: An Efficient Computational Model of Human Visual Attention". Ph.D. Thesis, University of Illinois at Urbana–Champaign, 1991.

[3] P. Bouthémy and P. Lalande, "Detection and Tracking of Moving Objects Based on a Statistical Regularization Method in Space and Time". Proc. First European Conference on Computer Vision, Antibes, France, April 1990, pp. 307-311.

[4] G.-J. Giefing, H. Janssen and H. Mallot, "Saccadic Object Recognition with an Active Vision System". 10th Eur. Conf. on Artificial Intelligence, 1992, pp. 803-805.

[5] S. Gil and T. Pun, "Non-linear Multiresolution Relaxation for Alerting". Eur. Conf. on Circuit Theory and Design, Davos, Switzerland, Elsevier Science, 1993, pp. 1639-1644.

[6] S. Gil, R. Milanese, and T. Pun, "Feature Selection for Object Tracking in Traffic Scenes". SPIE Conference on Smart Highways, Boston, MA, Oct. 31-Nov. 4, 1994.

[7] M. Kilger, "A Shadow Handler in a Video-based Real-time Traffic Monitoring System". IEEE Workshop on Applications of computer Vision, Palm Springs, CA, 1992, pp. 1060-1066.

[8] T. Lindeberg, “Scale-Space for Discrete Signals”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, 1990, pp. 234-254.

[9] S.G. Mallat, “A Theory for Multiresolution Signal Decomposition: The Wavelet Representation”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, 1989, pp. 674-693.

[10] R. Milanese, “Detecting Salient Regions in an Image: from Biological Evidence to Computer Implementation”. Ph.D. thesis, Univ. of Geneva, 1993. (Available through anonimous ftp to: cui.unige.ch, cd pub/milanese/thesis).

[11] R. Milanese, H. Wechsler, S. Gil, J.-M. Bost and T. Pun, “Integration of Bottom-Up and Top-Down Cues for Visual Attention Using Non-Linear Relaxation”. IEEE Conf. on Computer Vision and Pattern Recognition, Seattle, 1994, pp. 781-785.

[12] B. Olshausen, C. Anderson, and D. Van Essen, “A Neurobiological Model of Visual Attention and Invariant Pattern Recognition Based on Dynamic Routing of Information”. *Journal of Neuroscience*, Vol. 13, 1993, pp. 4700-4719.

[13] W. Pöltzleitner and H. Wechsler, “Selective and Focused Invariant Recognition Using Distributed Associative Memories”. *IEEE Trans. PAMI*, Vol. 12, No. 8, 1990, pp. 809-814.

[14] P. A. Sandon, “Simulating Visual Attention”. *Journal of Cognitive Neuroscience*, Vol. 2, No. 3, 1990, pp. 213-231.

[15] R. Thoma, M. Bierling, "Motion Compensating Interpolation Considering Covered and Uncovered Background". *Signal Processing: Image Communication*, Vol. 1, 1989, pp. 191-212.

[16] M. Unser, A. Aldroubi, M. Eden, "The L2 Polynomial Spline Pyramid". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, 1993, pp. 364-380.

[17] H. Zabrodsky and S. Peleg, "Attentive transmission". *Journal of Visual Communications and Image Representation*, Vol. 1, 1990, pp. 189-198.