1

Local Energy Variability as a Generic Measure of Bottom-Up Salience

Antón Garcia-Diaz, Xosé R. Fdez-Vidal, Xosé M. Pardo and Raquel Dosil
Universidade de Santiago de Compostela
Spain

1. Introduction

In image analysis, complexity reduction by selection of regions of interest is considered a biologically inspired strategy. In fact, Human Visual System (HVS) is constantly moving away less relevant information in favour of the most salient objects or features, by means of highly selective mechanisms forming an overall operation referred to as visual attention. This is the evolutionary solution to the well known complexity reduction problem (Tsotsos, 2005), when dealing with the processing and interpretation of natural images; a problem that is a major challenge for technical systems devoted to the processing of images or video sequences in real time. Hence, attention seems to be an adequate bio-inspired solution which can be applied in a variety of computing problems. Along with available technical advances, this fact is key to explain why the description and computational modelling of the attentional function of the HVS has experienced an enormous increase in the last two decades. In fact, applications of computing visual conspicuity are already found in many different fields: image segmentation and object learning and recognition (Rutishauser et al., 2004); vision system for robots (Witkowski & Randell, 2004) and humanoid robots (Orabona et al., 2005); visual behaviour generation in virtual human animation (Peters & O'Sullivan, 2003); processing data from 3D laser scanner (Frintrop et al., 2003); content-based image retrieval (Marques et al., 2003), etc.

In models of attention it is common to differentiate between two types of attention, the bottom-up from an image-based saliency, which accounts for features that stand out from the context, and the top-down attention as task-dependent and knowledge-based. These two kinds of attention are widely assumed to interact each other, delivering a global measure of saliency that drives visual selection. In fact, neurophysiological results suggest that these two mechanisms of attention take place in separate brain areas which interact in a visual task (Corbetta & Shulman, 2002) (Buschman & Miller 2007).

Regarding bottom-up attention, there are both psychophysical and neurophysiological experiments supporting the existence of some kind of an image-based saliency map in the brain, and it can be also argued that understanding of bottom-up saliency should definitely help to elucidate the mechanisms of attention (Zhaoping, 2005).

Moreover, from a technical point of view, mainly concerned with a generic approach to active vision tasks, the modelling of bottom-up component of attention can play a crucial role in the reduction of the amount of information to process, regardless of the knowledge

Source: Pattern Recognition Techniques, Technology and Applications, Book edited by: Peng-Yeng Yin,
ISBN 978-953-7619-24-4, pp. 626, November 2008, I-Tech, Vienna, Austria

managed by a given system, providing salient locations (regions of interest) or salient features. But it can also be suitable to learn salient objects, to measure the low level salience of a given object in a scene, etc. Hence, improvements on generic approaches to the modelling of bottom-up, image-based saliency are of great importance for computer vision.

The feature integration theory by Treisman & Gelade (1980) marked the starting point for the development of computational models of visual attention. Its main contribution lies on the proposal of parallel extraction of feature maps representing the scene in different feature dimensions and the integration of these maps in a central one, which would be responsible for driving attention. As a remarkable result from this parallel processing of few features proposed and maintained by Treisman in several works, arises the explanation of pop-out effects observed in visual search experiments with humans. It is well known that a stimulus that is clearly different from a homogeneous surrounding in a single feature rapidly attract our glance without the need to search the scene, regardless of the number of nearby objects acting as distractors. In contrast, when distractors are clearly heterogeneous, or when the target differs from all of them in a combination of features rather than in only one, subjects need to examine the scene object by object to check for a match with the target, so the time wasted in search linearly grows with the number of distractors. Treisman held that this can be understood if parallel processing of features exhibiting pop-out effects is assumed, and thus the feature map corresponding to the unique different feature in the first case will strongly respond in the location of the target attracting attention to it. On the other hand, in the heterogeneous and in the conjunctive cases none or several maps in different locations will fire, without provide for a clear salient location, so explaining the need for a serial search.

These ideas were gathered by Koch & Ullman (1985), to conceive a saliency-based computational architecture, in which they also introduced a Winner Takes All (WTA) network to determine the next most salient region, combined with a mechanism of Inhibition Of Return (IOR) to allow for a dynamic selection of different regions of a scene in the course of time. This architecture is essentially bottom-up, although they pointed the possibility of introducing top-down knowledge through bias of the feature maps.

An important subsequent psychophysical model of attention trying to explain more results on visual search experiments is the Guided Search Model, hold by Wolfe, in which feature dimensions (colour and orientation) rather than features (vertical, green, horizontal, etc.) are assumed to be processed in parallel and so to have an independent map of salience (Wolfe, 1994). In this model also top-down influences are considered by means of top-down maps for each feature dimension. More recent psychophysical models of attention are focusing more on top-down than in bottom-up aspects of attention, introducing the reasoning on the gist of a scene and its layout as driving attention (Rensink, 2005) (Oliva, 2005).

We have already mentioned the Guided Search Model by Wolfe, but we can cite a number of examples of computational models of bottom-up visual attention, many incorporating also a top-down component. Some of them are conceived more to explain psychophysical and neurophysiological results than to reach a performance in machine vision or other technical applications dealing with natural images. This is the case of the FeatureGate model by Cave (1999), the adaptive resonance theory to model attention proposed by Grossberg (2005), the neurodynamical approach hold by Deco et al. (2005), the model of bottom-up saliency coded in V1 cells by Zhaoping (2005), etc. Other models are motivated by the study of attention from an information theoretical point of view, trying to catch and describe the strategy of information processing of the HVS with statistical and computational tools. This is the case of Tsotsos et al. (1995) who have hold the Selective Tuning Model exploiting the complexity analysis of the problem of viewing, and achieving by this way several predictions on the real behaviour of the HVS. It is also the case of Rajashekhar et al. (2006), who have studied the statistical structure of the points that attract the eye fixations of human observers in natural images, in surveillance and search task. From this study they have derived models for a set have modelled a set of low level gaze attractors, in the form of filter kernels.

Focusing in the computational models that are the most relevant for our work, we find two particular previous implementations of the Koch and Ullman architecture being of special interest. The first was made by Milanese and was initially only bottom-up (Milanese, 1993), employing colour (or intensity), orientation and edge magnitude, in a centre-surround approach, as low level conspicuity maps; and proposing a relaxation rule for the integration process in a final saliency map. In a later work (Milanese et al., 1993), a top-down component was added in the form of an object recognition system that, applied to a few small regions of interest provided by the bottom-up component, delivered a top-down map favouring regions of the recognized objects. This map was combined with the conspicuity maps to give a final saliency in which known objects were highlighted against unknown ones.

The second implementation of the Koch and Ullman architecture was hold by Itti et al. (1998) who similarly made use of contrast, colour and orientation as features, in a centre-surround approach, but introducing a simpler integration process of weighting and addition of maps at first and of iterative spatial competition and addition in a subsequent work (Itti & Koch 2000). These two approaches to integration were significantly faster than the relaxation rule proposed by Milanese. This model can be seen as the most developed and powerful among all models of bottom-up visual attention, considering the fact that its performance has been compared with human performance (Itti & Koch, 2000)(Itti, 2006)(Ouerhani et al., 2006)(Parkhurst & Niebur, 2005), and tested in a variety of applications (Walther, 2006)(Ouerhani & Hugli, 2006). Recently, Navalpakkam & Itti (2005) introduced a top-down module in the model, based on the learning of target features from training images. This produces a feature vector which is subsequently used to bias the feature maps of the bottom-up component, hence speeding up the detection of a known object, in relation to the plain bottom-up model.

Now turning back to the problem of modelling bottom-up attention, we still have to ask, as a first question to delimit, which guidelines or requirements are currently imposed to the modelling of early low level features?. An interesting and worthy approach to attentional relevant features can be found in a recent exhaustive review on psychophysical works dealing with pop-out generation in visual attention, where Wolfe & Horowitz (2004) have provided a list classifying a variety of features, from lowest level, like contrast, colour or orientation, to highest level, like words or faces, making the classification dependent on the evidence and probability of each feature being causing pop-out or not. Hence, there would be features with enough observed evidences of causing pop-out (as intensity contrast, orientation, colour, size), others with high probability, others with low probability and finally others without probability at all. Then, a model of visual attention should be able to account for at least those features which give rise to clear pop-out effects as deduced from all of these cumulated results.

A starting issue underlying the selection of low level features lies in the assumption of a basis of “receptive fields”, suitable to efficiently extract all the information needed from an image. Therefore, an obliged reference should be the cumulated knowledge about visual receptive fields in five decades, from the seminal work of Hubel and Wiesel in the 60's. In this sense, there is a general agreement in viewing the region V1 region of the visual cortex as a sort of Gabor-like filter bank. However, we also should to have in mind the shadows threatening this sight, as have been pointed out in a recent review by Olshausen and Field (2005) on the emerging challenges to the standard model of V1, to the point of assessing that we only understand up to a 15% of the V1 function.
On the other hand, information theory has also provided a number of requirements for the construction and processing of early low level features. Hence many studies have oriented their work to discover the statistical structure of what we see and link it to the known neurological processing strategies of the HVS. The intrinsic sparseness of natural images has been pointed out by Olshausen & Field (1996) , who have demonstrated that an efficient coding maximizing sparseness is sufficient to account for neural receptive fields, because of the statistical structure of natural images. Likewise, Bell & Sejnowski (1997) found that the independent components of natural images were localised edge detectors, similar to neural receptive fields. Following this idea, Hoyer & Hyvärinen (2000) have applied the Independent Component Analysis (ICA) to the feature extraction on colour and stereo images, obtaining features resembling simple cell receptive fields, and thereby reinforcing this prediction.
This idea has been strongly supported by parallel neurophysiological works, showing increased population sparseness as well as decorrelated responses during experiments of observation of natural scenes, or when non classical receptive fields receive natural-like stimuli as input (Weliky et al. 2003) (Vinje & Gallant 2000).
Hence, what we can expect in a plausible, adapted to natural images, computing model of visual attention is that any representation of information to be processed, should be coded in a sparse way, and it should also lead to a decorrelation of the information captured by the vision system, in accordance with the structure of information in natural images and the results from neurophysiological experiments, as well as efficiency requirements.
Other important reference more directly related to attention is the work of Zetzsche, who, with basis on the analysis of the statistical properties of fixated regions in natural images, hold that i2D signals are preferred by saccadic selection in comparison to i1D and i0D signals, that is, regions containing different orientations (corners, curves, etc) do attract attention much more than regions with little structural content (simple edges, constant luminance, etc) (Zetzsche, 2005). We find this approach to low level conspicuity very enlightening, and pointing in the direction of a more formal approach to the definition of what is a low level feature.

1.1 Our approach

Intensity contrast, orientation, symmetry, edges, corners, circles,... all designate different but overlapping concepts. Then, a question arises: is there a formal and more general low-level measure capable of retaining and managing with all of the information related to them? We consider that local energy meets this condition, and we hold that its relative variability in a given region can produce a pop-out effect. Moreover, we expect early unguided attention to be driven by any pop-out stimulus present in the scene, and this is the basis for our working hypothesis: variability on local energy (as well as on colour) can be considered as driving attention by means of pop-out phenomena.

Local energy has proved to be a powerful tool for the extraction and segmentation of a variety of perceived features related to phase -from edges and corners to Mach bands or motion- and, in general, regions exhibiting phase congruency and phase symmetry, be in space or in spacetime (Kovesi 1993; 1996), (Morrone & Owens 1987), (Dosil et al. 2008).

In this chapter, exploiting the basic Koch and Ullman architecture, we present a saliency measure for the computational modelling of bottom-up attention, based on the detection of regions with maximum local energy variability, as a measure of local feature contrast and relative amount of structural content, which we have outlined in a previous brief paper (Garcia-Diaz et al. 2007).

We hold that this way, regions with maximum feature contrast and maximum structural content are extracted from a given image, providing a suitable map of salience to drive bottom-up attention.

We focus on local energy conspicuity computation in static scenes, while other relevant feature dimensions, like colour and motion, remain beyond the scope of this chapter. Likewise, we limit our study to the bottom-up component, without task or target constraints.

Qualitative and quantitative observations on a variety of results on natural images, suggest that our model ensures reproduction of both sparseness population increase, decorrelated responses and pop-out phenomena deployment of orientation, size, shape, and contrast singletons, widely observed in the human visual system (Vinje & Gallant 2000),(Weliky et al. 2003), (Zhaoping 2005), (Wolfe & Horowitz 2004).

To provide for results comparable with those found in literature, we carry out here the reproduction of several experiments already published by Itti & Koch (2000), improving the performance achieved by them in the deployment of orientation pop-out, and equalizing their results in the detection of military vehicles within cluttered natural scenes, in our case without the use of colour information.

Beyond the success in these tests of technical performance, other relevant contribution of this work lies on the new elements provided for the computational interpretation of different observed psychophysical pop-out phenomena (intensity contrast, edge, shape, etc.), as probably different faces or appearances of a pop-out effect bound to a unique low level feature dimension (local energy). Unlike the extended use of intuitive features conceived from natural language, we think that the results achieved by our model help to highlight the importance of tackling the modelling of feature dimensions in a more formal way, thereby, avoiding misleading conclusions when we assess the results from psychophysical experimental observations, with the aim of translating them in computational constraints or requirements.

This paper is organized as follows, in the section 2 we describe the model proposed; in section 3 we show the experimental results obtained and make a brief discussion of them; section 4 deals with conclusions; and finally an appendix offers a brief formal explanation of T^2 Hotelling statistic.

2. Extraction of salience and fixations

The model of bottom-up attention presented here involves the extraction of local energy variability as a measure of salience and the subsequent selection of fixations.

Thus, we extract initial local energy maps obtaining by this way a multi-scale and multi-oriented representation of the image. For each orientation we decorrelate the multi-scale information by means of a PCA. Next we fuse each of the new sets of *principal* scaled maps in corresponding oriented conspicuity measures, extracting variability with the computation of the statistical distance of each pixel from the centre of the distribution. Afterwards we locally excite and gather regions exhibiting maximum variability by a non-linear and centre-surround spatial competition. Therefore we reach a unique and final saliency map, on which we perform fixations. The following subsections detail the process.

2.1 Local energy from log Gabor receptive fields

As we have previously pointed out, one first question to tackle is related to the starting basis of receptive fields. A variety of elections have been made on the subject in previous models of bottom-up attention: Gabor functions (Itti et al., 1998) (Torralba, 2005), Difference of oriented gaussians (Milanese et al. 1995), Oriented derivative of Gaussians (Rao & Ballard, 1995), non linear i2D selective operators (Schill et al., 2001), etc...
We use, instead, a bank of log Gabor filters (Field 1987), which besides a number of advantages against Gabor filters, have complex valued responses. Hence, they provide in each scale and orientation a pair of filters in phase quadrature (Kovesi 1996), an even symmetric -real part- filter and its Hilbert transform, an odd, antisymmetric -imaginary part- filter, allowing us to extract local energy as the modulus (Morrone & Burr 1988) of this filter vector.

$$(r,g,b) = (R,G,B)/255; \\ I = (r+g+b)/3; \tag{1}$$

$$\mathrm{Resp}_{so}(x,y) = (I * \log \mathrm{Gabor}_{so})(x,y) = f_{so}(x,y) + h_{so}(x,y)i \tag{2}$$

$$e_{so}(x,y) = \sqrt{f_{so}^2(x,y) + h_{so}^2(x,y)} \tag{3}$$

All Gabor filters present a non-zero DC component, as well non-zero values for negative frequencies, which gives rise to artefacts. Field (1987) proposed to construct Gabors in a logarithmic frequency scale, the so called log Gabor filters, overcoming these pointed drawbacks. Besides this advantages the symmetric profile in a logarithmic frequency scale, characteristic of log Gabor filters, confers them one additional advantage: a long tail towards the high frequencies. Since natural images present scale invariance, this is, they present amplitude profiles that decay with the inverse of the frequency (Field, 1993), then a filter that presents a similar behaviour, should be able to properly encode those images (Kovesi, 1996). Moreover, they gain in biological plausibility respecto to Gabor, since they reproduce better the response of simple cells from cortex, logarithmic in the frequency domain.
The fact that log Gabors have no analytic representation in the spatial domain, forces us to construct the bank of filters in the frequency domain, performing the inner product between their transfer functions and the Fourier transform of the intensity of the image. This should not be seen as a problem, as the use of Fast Fourier Transform and Inverse Fast Fourier Transform algorithms, speed up a filtering process respect to a convolution operation. Anyway the log Gabor are given by the expression:

$$\log \text{Gabor}(f, \alpha; f_i, \alpha_i) = e^{-\frac{(\log(f/f_i))^2}{2(\log(\sigma_{fi}/f_i))^2}} e^{-\frac{(\alpha-\alpha_i)^2}{2(\sigma_\alpha)^2}} \tag{4}$$

We have used 6 scales and the central frequencies of the filters were spaced by one octave; other parameters were the minimum wavelength (λ_{min} = 2), the angular standard deviation (σ_α = 37.5º) or the frequency bandwidth (two octaves). This election of scales simply stretches the possible number of scales of the smallest images within the sets used in this work, and for simplicity it has not been modified for the rest of them since it has been observed to not significantly alter the results. In relation to the number of orientations the election accounts for the facts that pop-out effects are observed preferentially for deviations from four "canonical" orientations (Treisman 1993) -horizontal, vertical and right and left diagonal-, and is also needed a minimum difference of orientation angle of nearly 10º between distractors and target to generate a pop-out.

Once the initial receptive field responses have been extracted, the next step is necessarily related to the *feature* to extract from them. Again, a number of combined possibilities have been explored on the matter in previous models: intensity contrast, orientations, edges, predefined shapes, etc. But we put in question here the suitability of dividing the non-colour information in a number of feature dimensions in an early - low level- approach to attention. We hold instead the extraction of a low level, structurally meaningful, and multifaceted feature as local energy has proven to be. We obtain it as the modulus of the log Gabor responses.

2.2 Decorrelation and variability extraction

The next step to take is related to the integration of the initial feature maps in a final measure of saliency, and here we find again a variety of approaches in previous models. Focusing in the mentioned implementations of the Koch and Ullman architecture, Milanese et al. (1995) implemented a relaxation process by means of a non-linear updating rule which updates all the feature maps to satisfy a convergence criterion, and defining a heuristic energy function to minimize; in the other hand Itti & Koch (2000) have proposed an integration process based on the summation after the filtering of maps with iterative DoG filters, providing local within-feature and inter-feature competition.

Instead of convergence or summation for intra-feature integration we hold a *relative variability hypothesis*, by which one region is conspicuous as far as it contributes to the variability of responses in the ensemble of scales, leading to a measure of structural difference from the surround. So that, regarding local energy as a feature dimension split in oriented sub-dimensions, each characterized by a multi-scaled sub-feature vector, we propose a bottom-up attentional integration process based on the decorrelation of information and the subsequent extraction of the statistical distance from the average sub-feature vector.

A relevant point (or region) is expected to have a scale composition vector (structure) far from the mean. Given the huge number of samples (pixels) as well as the high dimensionality (number of scales) to manage, we propose to perform an information decorrelation process and the further gathering of the T^2 value of each point, providing a measure of statistical distance in a space of decorrelated scales, as a measure of multi-scale relevance.

Going more into detail, we start from six local energy scale maps for each of the four orientations computed. From them we define at each point four sub-feature vectors, one for each orientation, with six components corresponding to the local energy values at each of the scales. We have as many sample vectors for each orientation as pixels are in the single local energy maps, that is

$$\mathbf{x_i} = (x_{i1}, x_{i2}, \cdots, x_{is})'; (i=1,2,\ldots,n) \tag{5}$$

Arranging these original vectors as columns -samples- in a matrix of data X for each orientation, we treat the rows -scales- as original -partially correlated- coordinates, and we perform a PCA on it. From the new -decorrelated- coordinates, we can extract the T^2 statistical distance of each sample -pixel- from the centre of the distribution.

$$\mathbf{X_o} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \vdots \\ x_{s1} & \cdots & x_{sn} \end{pmatrix} \rightarrow (PCA) \rightarrow T_o^2 = (T_{o1}^2, T_{o2}^2, \cdots, T_{on}^2) \tag{6}$$

where T^2 is defined as:

$$T_i^2 = (\mathbf{x_i} - \overline{\mathbf{x_i}})' \mathbf{S^1} (\mathbf{x_i} - \overline{\mathbf{x_i}}) \tag{7}$$

being **S** the covariance matrix of the samples -pixels-, and $\mathbf{x_i}$ a sample vector with the scale values as components.

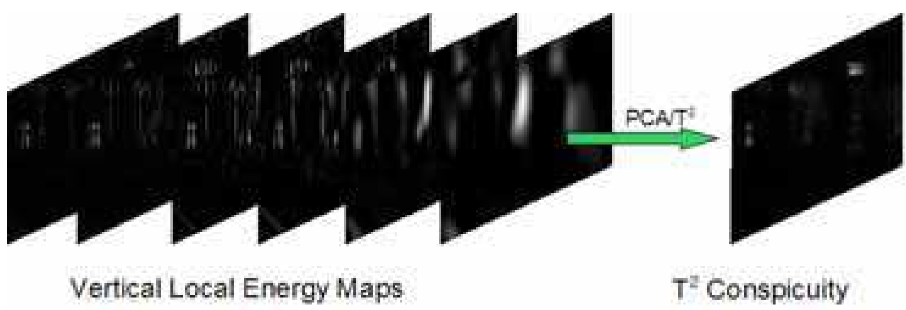

Fig. 1. T^2 conspicuity from PCA on local energy.

We take the resulting map of statistical distances as a relevance map for the analysed sub-feature. In figure 1 we can observe the high selectivity of this process, leading to an enhancement in population sparseness, synonym for code efficiency.

2.3 Local maxima excitation

We combine the previous procedure with a local maxima excitation to provide for a more robust and locally reinforced conspicuity. Next, to further compose a final saliency from oriented conspicuities, we gather the maximum values from the previous maps.

Local maxima excitation is addressed by means of a non-linear and centre-surround spatial competition. With preciseness, we apply iterative non-linear filtering of Difference of Gaussians (DoG), in a close but modified version of that implemented by Itti & Koch (2000). As well as minor differences in the inhibitory mechanisms, we reduce the number of iterations by modifying the last one. Thus, we take the excitatory signal (the response to the higher and narrower Gaussian) instead of the response to the difference (excitatory less inhibitory), avoiding the influence of the inhibitory signal, that is to say, achieving a strengthening of the regions with a high contribution to structure variability.

Hence, the excitatory and the inhibitory gaussian filters hold the following expressions:

$$Exc = \frac{c_{ex}^2}{2\pi\sigma_{ex}^2} e^{-(x^2+y^2)/2\sigma_{ex}^2} \tag{8}$$

$$Inh = \frac{c_{inh}^2}{2\pi\sigma_{inh}^2} e^{-(x^2+y^2)/2\sigma_{inh}^2} \tag{9}$$

Where we have used the values σ_{ex} = 2% and c_{ex} = 0.5 for the excitatory signal and σ_{in} = 25% and c_{in} = 1.5 for the inhibitory signal.

$$DoG = Exc - Inh = \frac{c_{ex}^2}{2\pi\sigma_{ex}^2} e^{-(x^2+y^2)/2\sigma_{ex}^2} - \frac{c_{inh}^2}{2\pi\sigma_{inh}^2} e^{-(x^2+y^2)/2\sigma_{inh}^2} \tag{10}$$

We perform a number of iterations (two in all of the experiments described here) for the following non-linear transformation:

$$\left| M_i + M_i * DoG - C_{inh} \right|_{>0} \rightarrow M_{i+1} \tag{11}$$

where the non-linear inhibitory term C_{inh} is established to the 2% of the maximum of the map. So far we have essentially followed the proposal made by Itti & Koch (2000) but in addition to these intermediate steps we finally impose the convolution of the map with the excitation signal, without any kind of inhibition:

$$M_{i+1} = M_i * Exc \tag{12}$$

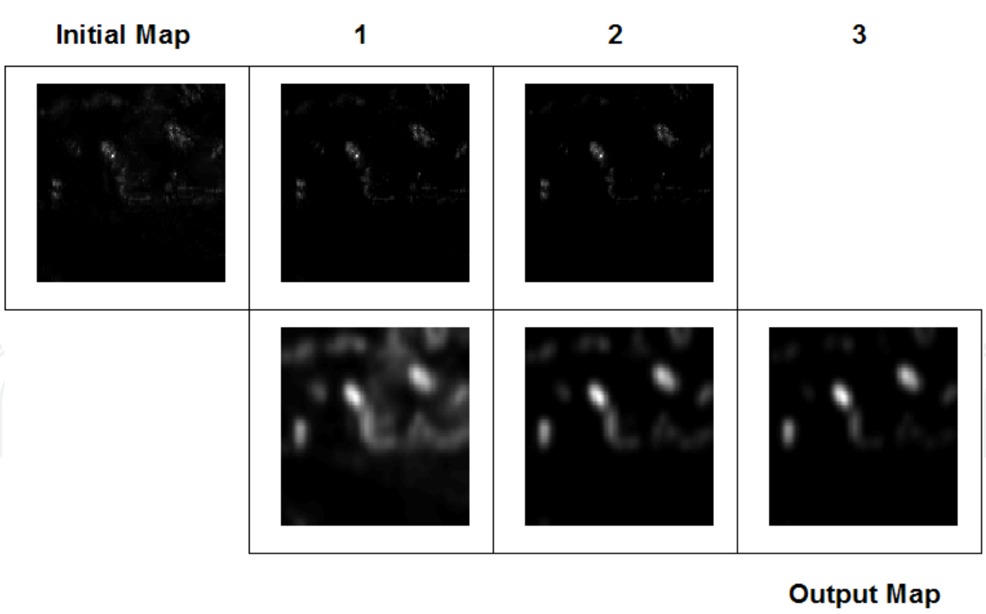

Fig. 2. Example of spatial Competition Process. First row: evolution of the map; second row: excitatory signal in each iteration and final conspicuity map

The overall effect of this spatial competition operation can be summarized in two main assessments, namely it favours small regions with many strong peaks, removing single isolated pixels and wide and constant regions, and it also reinforces the maxima, grading them in a conspicuity map. One example illustrates all of this in figure 2.

An example summarizing the whole integration process up to here can be seen in the figure 3, where we can check how the resulting relevance maps are actually a representation of regions with maximum contribution to structural variability, which have proven perceptually relevant and are supposed to strongly attract gaze. Furthermore, the highly competitive character of this procedure removes most noise and irrelevant regions, and reaches an important gain in population sparseness, retaining one or very few relevant regions, depending on the variability and thus low level significance of the feature considered. Hence, it seems to perform an efficient within feature competition, and also set the basis for a good inter feature competition.

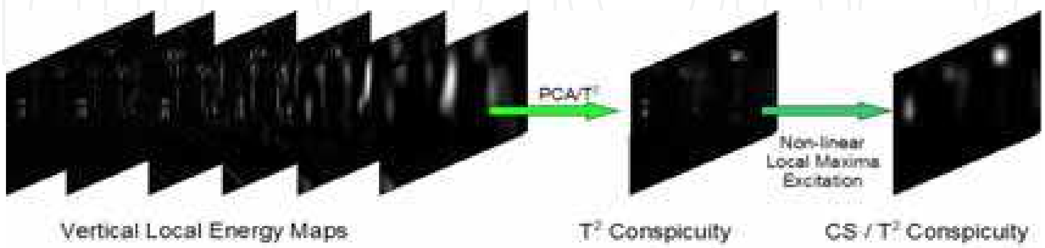

Fig. 3. Effect of non-linear local maxima excitation

After local maxima reinforcement, the next integration step is the obtaining of a final measure of saliency. This is done by taking the maximum values at each point from the previous conspicuity maps, giving rise to a horizontal competition between orientations, indeed reinforcing our strategy which aims to maximize the variability distance in structural content.

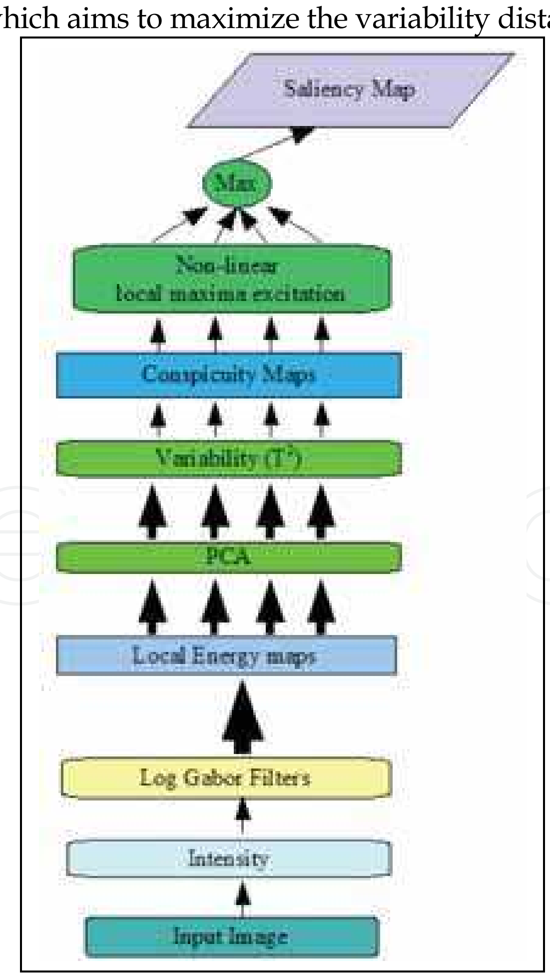

Figure 4. Extraction of Saliency from Local Energy

In the figure 4 is shown a complete scheme of our approach to the extraction of a bottom-up saliency map.

2.4 Fixations selection

Finally, making use of the extracted saliency map, the model should deploy a series of ordered fixations on the image. To do this, we have implemented a simplified version of the WTA neural network used by Itti & Koch (2000) in their experiments, but maintaining the basic assumptions for the focus of attention (FOA) size and considering the target detected when the FOA intersects its mask. This WTA is modelled by a two dimensional layer of integrate-and-fire neurons, with a mechanism of inhibition of return to prevent from attending always the same location. Therefore, neuron firing shifts the FOA to the correspondent location, and immediately afterwards a transient inhibitory feedback is applied to the surrounding region in the saliency map to allow for the subsequent selection of other salient locations.

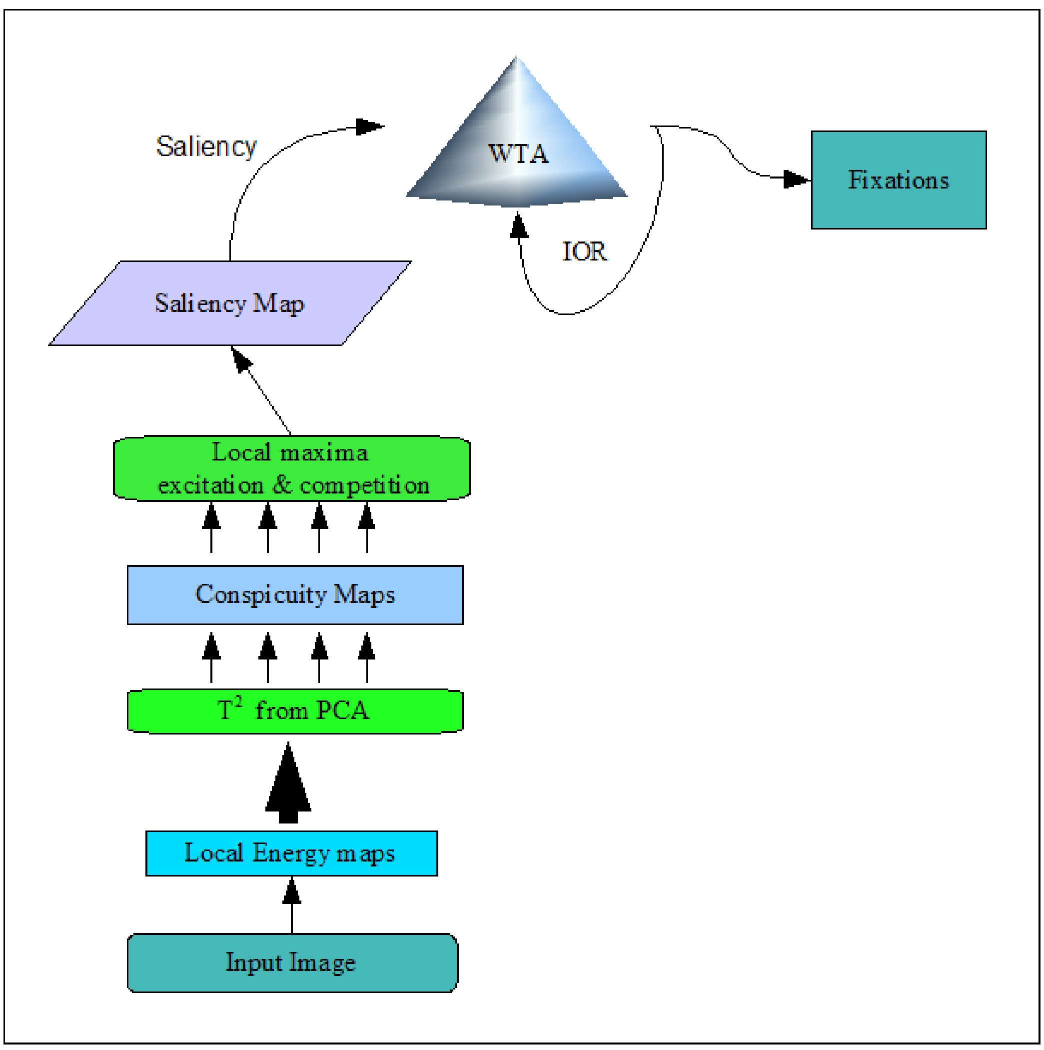

Fig. 5. Complete model of bottom-up attention based on local energy saliency

In the figure 5 is shown a scheme summarizing our local energy-based model of bottom-up attention.

3. Results

In this section we present the results obtained with the described model of bottom-up attention. We start, in the following subsection, with a qualitative analysis of its performance in psychophysically relevant situations. Other two subsections deal with the reproduction of quantitative experiments with public sets of images to evaluate both the capability to capture pop-out, as well as the search performance in a general purpose images dataset containing military vehicles in a landscape.

3.1 Reproduction of psychophysical phenomena and qualitative performance on natural scenes.

In this section we tackle the qualitative description of the behaviour of the model, showing the accordance with a variety of psychophysical results.
It is commonplace to relate the low-level saliency of a given target to the efficiency -in terms of wasted time- to find it. Thereby, a line for the qualitative analysis of a bottom-up attention model consists in checking the suitable reproduction of some relevant phenomena described in experiments of visual search.
The main and most characteristic of these phenomena is the pop-out, produced by an element differing in one unique feature from all of the others, that is to say, when a singleton is present in the image. Thus, one important aspect in a visual attention system consists in explaining saliency for singletons showing a pattern or feature unique in the image, be by the orientation, the size, the frequency content, etc. This kind of phenomena are the basis of the Treisman's FIT, which explains them by a privileged parallel processing of certain features.
Therefore, the pop-out of a target in a given image is strongly dependent on the context in which the target is present, and in the other hand it implies a behaviour highly non-linear: there is pop-out and the target is immediately found, or there is no pop-out and a serial search takes place, in which each of several elements with similar relevance are checked until the target is found. In this paper we show a wide range of pop-out phenomena successfully reproduced by our model, from local maxima in the variability of a structure descriptor as local energy is. In figure 6 we can see two first examples in which an element with a differing size fires a pop-out effect and rapidly attracts attention. It is not the size of the element itself, but the relative size respect to the others what causes a predominant salience.

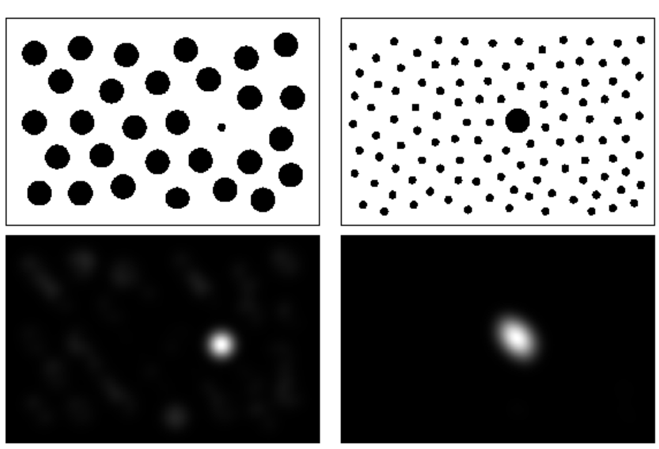

Fig. 6. Two examples of size pop-out

As we have seen, a singleton between many similar distractors shows a very high saliency, but what happens if distractors present differences themselves? As can be expected, differentiation of distractors leads to appearance of new singletons competing with the target, reducing its relative saliency. This effect of the distractor heterogeneity on target saliency, is well understood in our model as a reduction of the relative contribution to structure variability. In figure 7 we can see a meaningful example which illustrates well this question.

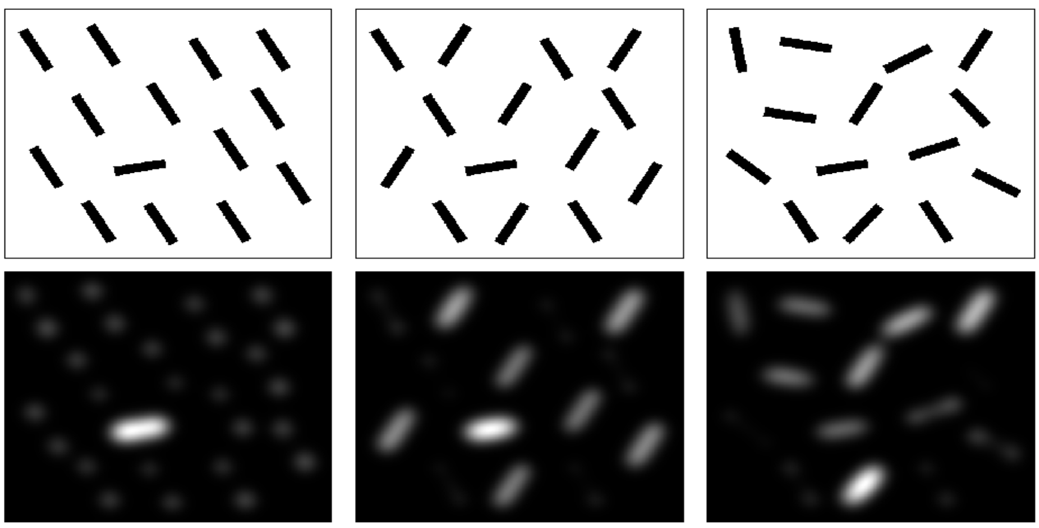

Fig. 7. Distractor heterogeneity reduces saliency of the target up to prevent for pop-out. In the three images the unique feature of the target is the same, but in the left image the distractor heterogeneity makes less relevant the orientation than other structural features.

Another factor reducing saliency is related to resemblance between the searched element and the surrounding ones, usually referred as the target-distractor similarity. Here it is this similarity with surrounds what threatens the status of singleton of the target. In the frame of our model, this can be explained again as a reduction in the local contribution to structure variability in the image. In figure 8 is shown how our model reproduces well this behaviour observed in psychophysical experiments. As can be seen there is no a linear relation between difference in size and relative saliency, since the model collects local variability maxima in a non-linear approach, aiming decorrelated and sparse responses.
There is another important set of phenomena observed in visual searches, commonly denoted as "search asymmetries". The related to visual search experiments designed on a given feature space, where target and distractors exchange its characterization, giving rise to different behaviours and therefore "asymmetric" attentional performance in such feature spaces. Disregarding the common discussion on the feature definitions involved, and on the suitability in talking of such asymmetries, we show the behaviour of our model in two typical situations and how it coincides basically with that described in psychophysical observations, providing with an simple explanation for them. The first of these cases has to see with the so called presence/absence asymmetry, in which target and distractors are the same element except by the presence or absence of an additional simple feature. What typically happens in these experiments is that the presence of the additional feature generates a pop-out while its absence remains unnoticed and does not fire any pop-out.

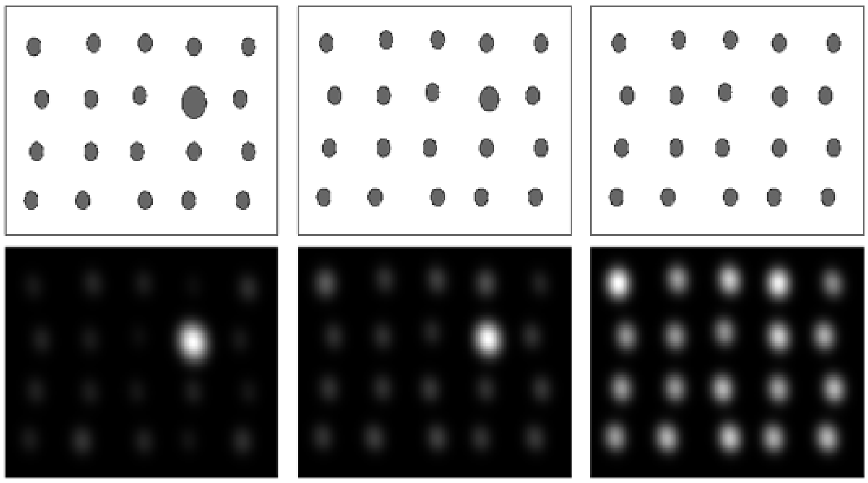

Fig. 8. Target-Distractor similarity reduces saliency of the target abruptly

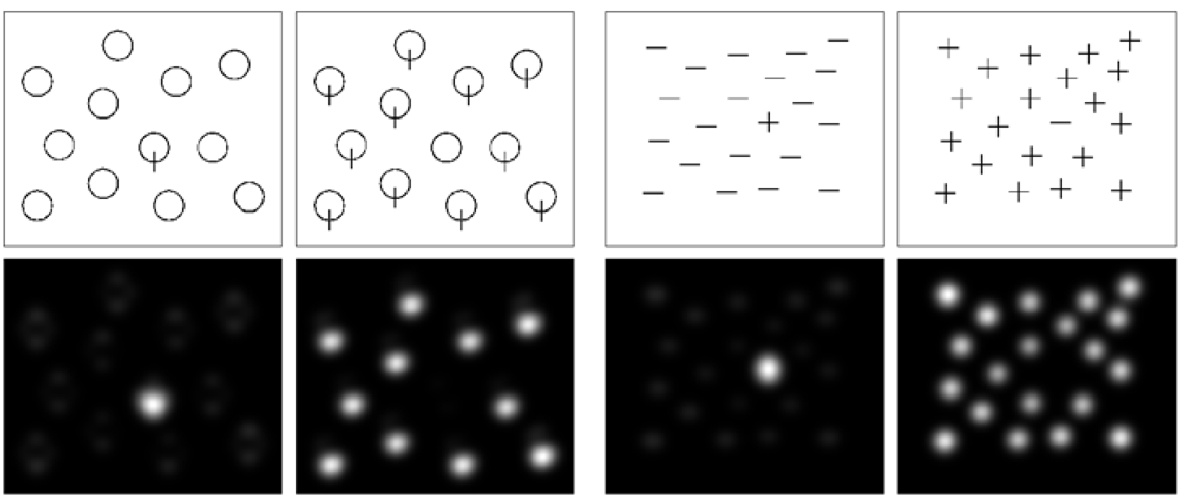

Fig. 9. Reproduction of the so called presence/absence asymmetry

In figure 9 we can see two examples of this asymmetry, in the left all the elements are circles which can have or not a vertical bar, in the rigth elements are horizontal bars and the aditional feature is again a vertical bar. As can be seen in both cases the model reproduces well psychophysical observations. In other hand the explanation is simple: the structure of the element(s) labelled by the "absence" is present in all the elements, so this contribution to structure variability in the image is equalled by all the other stimuli, while the element(s) with the additional feature present an additional contribution to structure variability, increasing its relative salience. Thus, such asymmetry is not an asymmetry in our model and the observed behaviour is perfectly understandable.

Another classical example of search assymetry is found in experiments in which target and distractor differing only in orientation exchange the value of this feature. It has been observed in such cases that the threshold in orientation difference needed to fire pop-out varies in a significant amount between the two possibilities. Treisman & Gormican (1988) have explained this phenomenon with basis in a privileged treatment of certain "canonical" orientations which would break the expected symmetry. As figure 10 shows this is well reproduced by our model. This should not result surprising, as the model computes four different orientations, which in spite of gathering all the possible orientations existing in the image, they make it in a unequal way. So, our low level descriptor makes use in fact of a set of canonical orientations and thus, is not symmetric respect to any orientation. But this seems to be in accordance with the performance observed for the HVS.

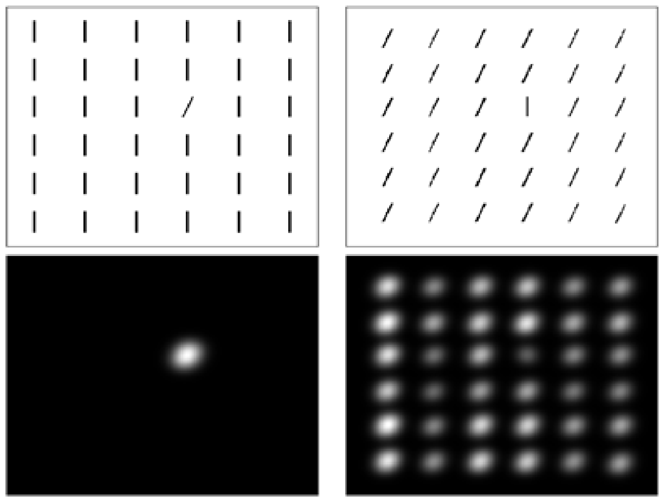

Fig. 10. Asymmetry in orientation pop-out threshold

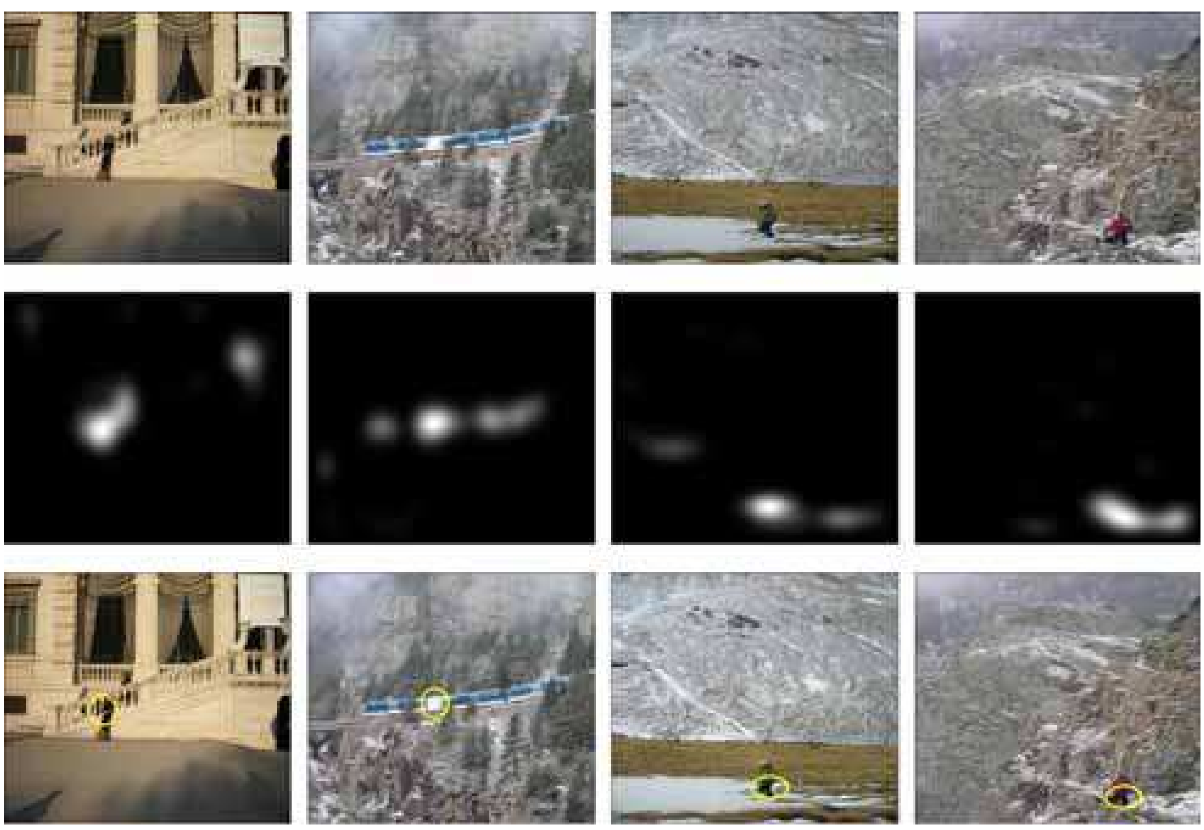

Fig. 11. Performance on natural scenes. Top: original image; centre: saliency map; bottom: first fixation.

Finally, to complete this qualitative description of the performance of the model figure 11 shows the saliency map and the first fixation for five cluttered natural scenes with different relevant objects, different visibilities, and different contexts. We should remember at this point that colour information was discarded in this work. As we can appreciate saliency maps are sparse, present few concentrated salient regions, corresponding in all cases to elements of obvious relevance. Therefore, the capability of the model to reproduce pop-out phenomena is not limited to artificial stimuli in synthetic images but it is also confirmed with different targets in natural scenes.

3.2 Performance on orientation pop-out

In this section we dealt with the reproduction of the orientation pop-out effects observed in the human visual system, parallel to that already carried out by Itti & Koch (2000). All the images, and their respective binary versions with the masks for target detection, are public and can be found in (http://ilab.usc.edu).

In the figure 12 we can see twenty examples of the obtained results, with the saliency map obtained and the correspondent fixations performed, they can give an idea of the robustness of the model in capturing orientation pop-out.

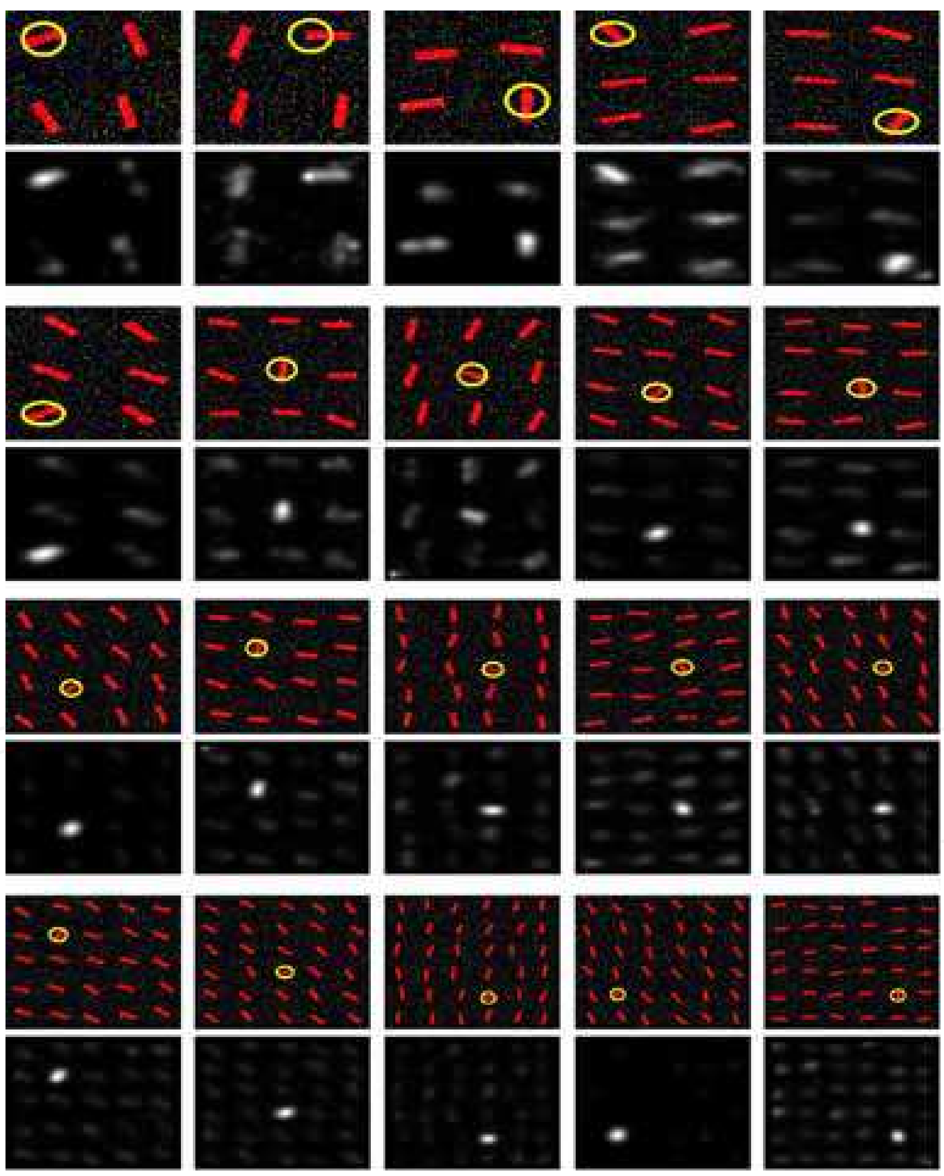

Fig. 12. Twenty examples of results in the pop-out orientation experiment.. For each case, fixations (top) and saliency (bottom) are shown.

In figure 11, the overall results are shown as the mean number of false fixations before target detection faced to the number of distractors present in the image. The dashed line represents the chance value corresponding to a pure serial search without any orientation pop-out effect (supposed half of distractors visited before the detection), and the blue points connected by a solid line, the average performance of our model, with the error bars being one standard deviation.

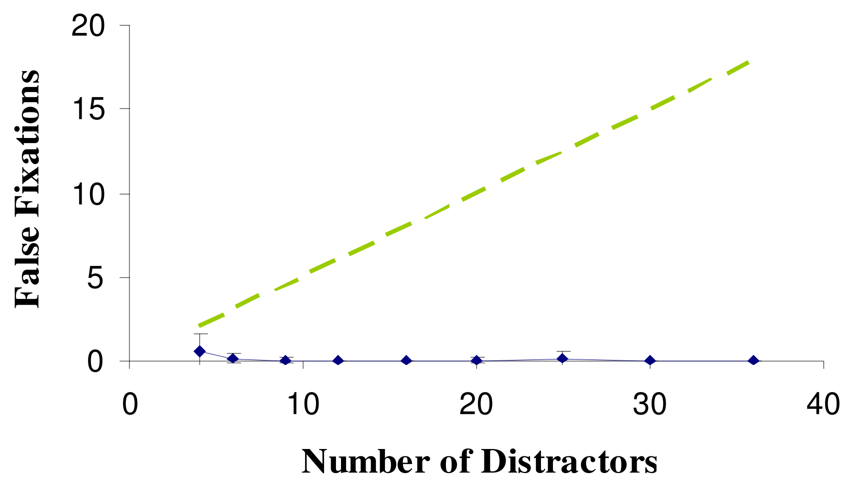

Fig. 13. Number of fixations against number of distractors. The dashed green line represents what could be expected in a serial search (half of distractors attended before target detection). Blue points connected by a blue solid line, and black bars, show the correspondent average and standard deviation values obtained.

We can assess that the model assures for a robust capture of orientation pop-out, independently of the number of distractors, since we obtained a flat slope in the number of false fixations. The performance became slightly poorer when the number of distractors was very small, which can be explained in terms of a reduction of the pop-out effect by the expectable reduction in the relative contribution of the target to structure variability in the image.

Our results clearly improve those obtained by Itti & Koch (2000), regarding the visible reduction of the mean value and even more important, the remarkable reduction of the standard deviation values; giving account for the fact that with our procedure there are not cases with a large number of false detections, thereby achieving a more robust performance. They haven't published numerical results so we can't carry out a numerical comparison.

3.3 Search performance on natural scenes

In this section we handle with target detection within natural scenes in a set of images containing a military vehicle in a landscape, again parallel to that already carried out by Itti & Koch (2000). The images from this set were sub sampled versions of images belonging to the search_2 database described by Toet et al. (2001). They correspond to 44 natural scenes containing a military vehicle of variable relative dimensions, identical to those used by Itti & Koch, except for the resolution of the images: ours had the fourth part size (1536x1024 pixels). Some of these images can be seen in figure 14.

We also assumed the same relative size for the FOA, which would imply a mean result of 61.5 fixations for random target detection.

One interesting feature of this database is the availability of the search time distribution curves obtained for human observers for each of the 44 images, allowing for a comparison with human performance in a search task in natural scenes. Simulated search times have been calibrated so that a mean of 330 ms elapse between two fixations, and an additional time of 1.5 s has then added to account for human motor response time.

Fig. 14. Examples of natural images from the military vehicles set. As can be seen many of them present a very low visibility of the target.

The main results take the very close overall values than Itti & Koch (2000), and so the model found the target with the first fixation in seven of the 44 images, and with fewer than twenty fixations in 23 images. The model has failed in two cases. Figure 13 shows the saliency map and the correspondent fixations performed for five images of high visibility of the target.

As Itti's model did, our model reached a poor correlation with human, and can also be considered faster than them finding the target, under the exposed search time calibration assumptions. In any case, this is a very inaccurate approach, because of the fact that search time distributions for humans are not well represented by the mean value.

But in other hand, qualitative analysis allows to see an agreement on the classification of an image, when it is considered as one with a high visibility of the target. This classification in humans has been done by comparison of the curves of search time distribution, and in the model selecting those images with less of 6s of search time. This comparison yields the result of eleven images classified like showing high visibility both by humans and the model, and only two images classified with medium visibility by the model while showing high visibility for humans.

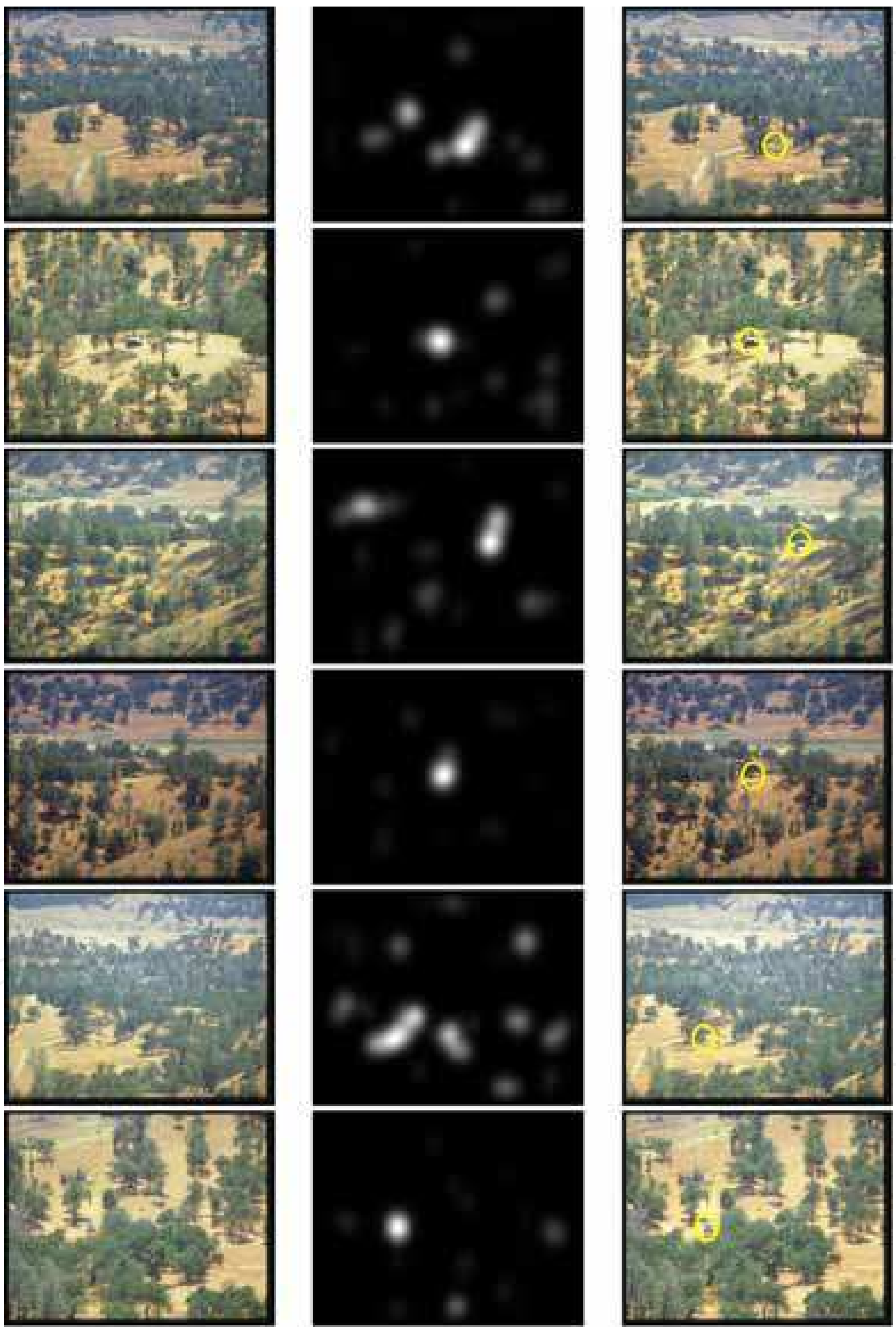

Fig. 15. Six examples of high visibility of the target, where a bottom-up approach makes sense. Left: original image; centre: saliency map; right: fixations performed.

In images with poor visibility for humans the agreement didn't exist, but these images, without a conspicuous target are expected to be processed by humans in a top-down manner, not bottom-up, the only one that is being modelled here.
Another comment to be made is related to the sparse maps obtained, particularly when a pop-out effect is clear. In the other hand, when the visibility of the target is lower, more elements from the landscape gain relative salience.

4. Conclusions

In this chapter we have exposed a particular approach to model bottom-up saliency, based in the hypothesis that variability of local energy is capable of capture the pop-out produced by the local contrast in a variety of non-colour features.
Hence, we have employed local energy as a suitable general descriptor of non-colour structure in the image, combined with information decorrelation, statistical distance computation and non-linear maxima excitation to detect local maxima of structure variability. The biological plausibility of the model arises from the combination of known features from V1 behaviour with a highly non-linear and collective performance, as well as assumptions based on psychophysical considerations (e.g. election of orientations).
The model is implemented in a meaningful and understandable fashion, thereby providing for a complete and robust computational frame to reproduce and formally explain the main observed features of static and non-colour bottom-up attention in humans.
Tested in synthetic as well as natural images, this approach gives rise to a simple model of bottom-up attention with a high performance, which accounts for pop-out effects and other psychophysical phenomena, and also solves conspicuity-driven search tasks more efficiently and robustly than a powerful state of art approach to bottom-up attention as it is that hold by Itti and colleagues. All of this is achieved with a simple scheme: while other models need for the separate use of intensity contrast and orientation (Itti et al. 2000), edges and orientation (Milanese 1993), and other combinations, we only make use of local energy as low-level descriptor to characterize non-colour structure.
It is important to remark that the model makes a generic approach to bottom-up saliency, without the use of any kind of knowledge or feature constraints, related to the target nor the task, and it is expected to reproduce human performance in corresponding situations, as unguided surveillance or conspicuity-driven visual search, on non-colour scenes or when relevance does not lie in colour.
Although local energy is not an intuitive feature from common language, it can account for many of these perceived intuitive features in a more reliable way for computational modelling purposes. In fact, it is underlying them. Moreover, conceived as a descriptor of structure is a powerful tool to understand a variety of features and the phenomena related to them without loss of meaning.
Furthermore, this approach takes into account and incorporates important features of HVS as expected and observed increased population sparseness and response decorrelation in comparison to previous Gabor-like and feature extraction models of saliency computation.
In progress and future work will deal with other feature dimensions, like colour and motion, in order to allow the model to work with real dynamic scenes; and also with a more depth study on the comparison with human performance.

5. Acknowledgements

This work has been financially supported by the Ministry of Education and Science of the Spanish Government through the research project AVISTA (TIN2006-08447), and by the Ministry of Industry and Innovation of the Galician Government through the project (PGIDIT07PXIB206028PR).

6. References

Bell, A. J. & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. Vision Research, 37,23, 3327-3338

Buschman, T. J. & Miller, E. K. (2007). Top-Down Versus Bottom-Up Control of Attention in the Prefrontal and Posterior Parietal Cortices Science Vol. 315. no. 5820, pp. 1860 - 1862

Cave K.R. The FeatureGate Model of Visual Selection Psychological Research 1999 62 182-194

Corbetta M. & Shulman G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. Nature Reviews, 3(3); 201–215.

R. Dosil; X. R. Fdez-Vidal; X. M. Pardo & A. Garcia (2008) Motion Representation Using Composite Energy Features Pattern Recognition, 41(3):1110-1123, 2008

Deco G.; Rolls E.T. & Zihl J. A (2005). Neurodynamical Model of Visual Attetion In L. Itti, G. Rees, J.K. Tsotsos (eds) Neurobiology of Attention (Elsevier Academia Press, 2005, 593-599).

Field D. J. (1987) Relations Between the Statistics of Natural Images and the Response Properties of Cortical Cells. Journal of the Optical Society of America A, 4(12), 1987, 2379–2394.

Field D. J.(1993). Scale-invariance and self-similar ’wavelet’ transforms: an analysis of natural scenes and mammalian visual systems. In M. Farge, J. Hunt & J. Vassilicos(eds.), Wavelets, Fractals and Fourier Transforms (Oxford, Clarendon Press, 1993, 151–193).

Frintrop S.; Rome E.; Nüchter A. & Surmann H. (2005) A Bimodal Laser-Based Attention System Journal of Computer Vision and Image Understanding (CVIU), 100(1-2), 2005,

Garcia-Diaz, A.; Fdez-Vidal, X.R.; Dosil, R.; Pardo, X.M. (2007) Local energy saliency for bottom-up visual attention. Proceedings of the IASTED International Conference on Visualization, Imaging and Image Processing.

Grossberg S. (2005). Linking attention to learning, expectation, competition and consciousness. n L. Itti, G. Rees, J.K. Tsotsos (eds) Neurobiology of Attention (Elsevier Academia Press, 2005, 652-662).

Hoyer P. O. & HyvärinenA. (2000). Independent component analysis applied to feature extraction from colour and stereo images. Network: Computation in Neural Systems, 11, 2000 191-210.

Itti L.; Koch C. & Niebur E. (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (11), 1998,1254-59.

Itti L. & C. Koch (2000), A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research, 40, 2000, 1489-1506.

Itti L. (2006). Quantitative modelling of perceptual salience at human eye position. Visual Cognition, 14(4/5/6/7/8), 2006, 959-984.

Jackson J.E. (1991). A User's Guide to Principal Components., Wiley-Interscience, John Wiley & Sons, New York,.USA. 1991.

Koch C. & Ullman (1985) S. Shifts in selective visual attention: towards the underlying neural circuitry. Human neurobiology, 4(4), 1985, 219-227.

Kovesi P. (1996). Invariant Measures of Image Features from Phase Information. Ph.D. (The University or Western Australia, 1996).

Kovesi P (1999).Image Features from Phase Congruency. Videre: Journal of Computer Vision Research The MIT Press, 1(3), 1999, 1-26

Marques O. ; Mayron L. M., Borba G. B. & Gamba H. R. (2006). Using visual attention to extract regions of interest in the context of image retrieval Proceedings of the ACM SE'06, Melbourne, FL, 2006, 638-643.

Milanese R. (1993). Detecting Salient Regions in an Image: from Biological Evidence to Computer Implementation. Ph.D. thesis, Univ. of Geneva,1993.

Milanese R. , Gil S. & Pun T. (1995). Attentive mechanisms for dynamic and static scene analysis. Optical engineering, 34(8), 1995 2428-34.

Morrone, M.C. & Owens, R.A. (1987) Feature detection from local energy. Pattern Recognition Letters 6, 303-313, 1987.

Morrone M.C. & Burr D.C.(1988). Feature Detection in Human Vision: A Phase-Dependent Energy Model. Proceedings of the Royal Society of London B, 235, 1988, 221–245.

Navalpakkam V. & Itti L. (2005) Modeling the influence of task on attention, Vision Research, 45(2), 2005, 205-231.

Nothdurft H.C. (2005) Salience of Feature Contrast In L. Itti, G. Rees, J.K. Tsotsos (eds) Neurobiology of Attention (Elsevier Academia Press, 2005, 233-239)

Oliva A. (2005). Gist of the scene. In L. Itti, G. Rees, J.K. Tsotsos (eds) Neurobiology of Attention (Elsevier Academia Press, 2005, 251-256).

Olshausen B. A. & Field D.J. (1996). Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. Nature, 381, 1996, 607-609.

Olshausen B. A. & Field D.J. (2005) How close are we to understanding V1? Neural Computation 17, 2005, 1665-1699

Orabona F.; Metta G. & Sandini G. (2005). Object-based Visual Attention: a Model for a Behaving Robot Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops – 03, 2005, 89.

Ouerhani N. & Hugli H. (2003) Maps: Multiscale attention-based presegmentation of color images, Lecture Notes in Computer Science, 2695, 2003, 537-549.

Ouerhani N. , R.Wartburg, , H.,Hugli R. Mueri (2004) Empirical validation of the saliency-based model of visual attention, Electronic letters on computer vision and image analysis, 3(1), 2004, 13-24.

Parkhurst D.J. & Niebur E. (2005) Stimulus-Driven Guidance of Visual attention in Nattural Scenes In L. Itti, G. Rees, J.K. Tsotsos (eds) Neurobiology of Attention (Elsevier Academia Press, 2005, 240-245).

Peters C. & O'Sullivan C. (2003). Bottom-Up Visual Attention for Virtual Human Animation CASA, 2003, 111-117.

Rajashekhar U.; Bovik A.C., & Cormack L.K. (2006). Visual search in noise. Journal of Vision, 6, 2006, 379-386

Rao R. P. N. & Ballard D. (1995) An Active Vision Architecture Based on Iconic Representations Artificial Intelligence 78, 1995, 461-505

Rensink R. A. (2000). The dynamic representation of scenes. Visual Cognition, 7; 2000; 17–42.

Rutishauser U.; Walther D.; Koch C. & Perona P. (2004). Is bottom-up attention useful for object recognition? IEEE Conference on Computer Vision and Pattern Recognition, 2, 2004, 37-44.

Schill K.; Umkehrer E.; Beinlich S.; Krieger G. & Zetzsche C. Scene analysis with saccadic eye movements: Top-Down and Bottom-Up Modelling. J. Electron. Imaging 10, 2001, 152-160.

Toet A. , Bijl P. & Valeton J.M. (2001), Image dataset for testing search and detection models. Optical Engineering 40(9), 2001, 1760-1767.

Torralba A. (2005). Contextual Influences on Saliency In L. Itti, G. Rees, J.K. Tsotsos (eds) Neurobiology of Attention (Elsevier Academia Press, 2005, 586-592).

Treisman A. & Gelade G. A (1980), Feature-Integration Theory of Attention. Cognitive Psychology, 12, 1980, 97–136.

Treisman A. & Gormican, S. (1988) Feature analysis in early vision: Evidence from search asymmetries. Psychological Review, 95(1), 15-48

Treisman A. (1993). The perception of features and objects In Baddeley, A. and Weiskrantz, L. (eds) Attention: Selection, awareness, and control, Clarendon Press, Oxford, 1993 5–35..

Tsotsos J. K.; Culhane S.; Wai. W.; Lai Y.; Davis N.& F. Nuflo, Modeling visual attention via selective tuning, Artificial Intelligence, 8(1-2), 1995, 507 – 547.

Tsotsos J. K. (2005) Computacional foundations for attentive Processes. In L. Itti, G. Rees, J.K. Tsotsos (eds) Neurobiology of Attention (Elsevier Academia Press, 2005, 3-7).

Walther D. , (2006) Interactions of visual attention and object recognition: computational modelling, algorithms, and psychophysics. PhD. (California Institute of Technology, 2006).

Weliky M. , J. Fiser, R. H. Hunt & D. N. Wagner, Coding of natural scenes in primary visual cortex. Neuron 37, 2003, 703-718.

Witkowski M. & Randell D. (2006) Modes of Attention and Inattention for a Model of Robot Perception. Proc. Towards Autonomous Robotic Systems, 2006, 246-253.

Wolfe J. M. (1994). Guided Search 2.0 A Revised model of visual search. Psychonomic Bulletin, 1(2), 1994, 202-238.

Wolfe J. M. & Horowitz T. S. (2004). What attributes guide the deployment of visual attention and how do they do it?. Nature Reviews. Neuroscience, 5(6), 2004, 495-501.

Zetzsche C. (2005). Natural Scene Statistics and Salient Visual Features In L. Itti, G. Rees, J.K. Tsotsos (eds) Neurobiology of Attention (Elsevier Academia Press, 2005, 226-232).

Zhaoping L. (2005). The primary visual cortex creates a bottom-up saliency map In L. Itti, G. Rees, J.K. Tsotsos (eds) Neurobiology of Attention (Elsevier Academia Press, 2005, 570-575)

Pattern Recognition Techniques, Technology and Applications

Edited by Peng-Yeng Yin

ISBN 978-953-7619-24-4
Hard cover, 626 pages
Publisher InTech
Published online 01, November, 2008
Published in print edition November, 2008

A wealth of advanced pattern recognition algorithms are emerging from the interdiscipline between technologies of effective visual features and the human-brain cognition process. Effective visual features are made possible through the rapid developments in appropriate sensor equipments, novel filter designs, and viable information processing architectures. While the understanding of human-brain cognition process broadens the way in which the computer can perform pattern recognition tasks. The present book is intended to collect representative researches around the globe focusing on low-level vision, filter design, features and image descriptors, data mining and analysis, and biologically inspired algorithms. The 27 chapters coved in this book disclose recent advances and new ideas in promoting the techniques, technology and applications of pattern recognition.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821