A Proto-object Based Visual Attention Model*

Francesco Orabona[1], Giorgio Metta[1,2], and Giulio Sandini[2]

[1] DIST, University of Genoa, Viale Causa, 13 - Genoa 16145, Italy
[2] Italian Institute of Technology, Via Morego, 30 - Genoa 16163, Italy

Abstract. One of the first steps of any visual system is that of locating suitable interest points, 'salient regions', in the scene, to detect events, and eventually to direct gaze toward these locations. In the last few years, object-based visual attention models have received an increasing interest in computational neuroscience and in computer vision, the problem, in this case, being that of creating a model of 'objecthood' that eventually guides a saliency mechanism. We present here an model of visual attention based on the definition of 'proto-objects' and show its instantiation on a humanoid robot. Moreover we propose a biological plausible way to learn certain Gestalt rules that can lead to proto-objects.

1 Visual Attention

Spatial attention is often assimilated to a sort of 'filter' of the incoming information, a 'spotlight', an internal eye or a 'zoom lens'. In particular it is believed to be deployed as a spatial gradient, centered on a particular location. Even if supported by numerous findings (see [1] for a review), this view does not stress enough the *functional role* of the attentional system in an agent with a body.

The external world is sensed continuously and it is not necessarily mapped into some complicated internal model (although it is also clear that internal models are required to predict the future course of actions or to compensate specific dynamic effects of movement [2]). This idea has been summarized by O'Regan in the following statement:

The world as an outside memory [3].

This sentence remarks the fact that it is important to consider the problem of vision, and perception in general, deeply rooted in the physical world. Given that changes in the world seem to be easily detectable, it would be cheaper to memorize, for example, only a rough representation of the external world updating it when changes happen and directly accessing the sensory data when detailed information is needed. Moreover, it is not possible to model perception without simultaneously considering also action, so it is logical to think that perception is biased toward representations that are useful to act on the environment. To an extreme, Maturana and Varela [4] and the proponents of some of the dynamical

* This work was supported by EU project RobotCub (IST- 2004-004370) and CONTACT (NEST-5010).

L. Paletta and E. Rome (Eds.): WAPCV 2007, LNAI 4840, pp. 198–215, 2007.

approaches to the modeling of cognitive systems [5], define cognition as *effective action*. That is, cognition is the actions taken by the agent to preserve its coupling with the environment, where clearly, if action is not effective then it is likely that the agent dies (which ends the coupling with the environment).

In the specific instance of visual attention this corresponds to ask whether attention is deployed at the level of objects ('object-based') or at space locations ('space-based'). Object-based attention is equivalent to thinking that attention is geared to the use of the objects, that depends on the internal plan of the agents, its current status, and very importantly of its overall goal [6]. The idea of object-based attention is also supported by the discovery in the monkey of a class of neurons (*mirror neurons*) which not only fire when the animal performs an action directed to an object, but also when it sees another monkey or human performing the same action on the same object [7]. Indeed, this tight coupling of perception and action is present in in visual attention too: it has been shown in [8] that more object-based attention is present during a grasping action.

Object-based attention theories argue that attention is directed to an object or a group of objects, to process specific properties of the selection, rather than generic regions of space. There is a growing evidence both from behavioral and from neurophysiological studies that shows, in fact, that selective attention frequently operates on an object based representational medium in which the boundaries of segmented objects, and not just spatial position, determine what is selected and how attention is deployed (see [9] for a review). This reflects the fact that the visual system is optimized for segmenting complex scenes into representations of objects to be used both for recognition and action, since perceivers must *interact* with objects and not just with disembodied spatial locations.

But how can we attend to objects before they are recognized? To solve this contradiction Rensink [10,11] introduced the notion of 'proto-objects', that are volatile units of visual information that can be bound into a coherent and stable object when accessed by focused attention and subsequently validated as actual objects. In fact, it is generally assumed that the task of grouping pixels into regions is performed before selective attention is involved by perceptual organization and Gestalt grouping principles [12].

Guided by these considerations we developed a general proto-object based visual attention model and designed a biological motivated method to learn how to pre-segment images into proto-objects.

This article is organized as follows: Section 2 contains an introduction on the modeling of human visual attention. Section 3 details the robot's visual system and the proposed model, and in Section 4 some results are shown. In Section 5 a new method to build better proto-objects is described, with numerical validation in Section 6. Finally in 7 we draw some conclusions and future work.

2 Computational Models of Visual Attention

A dominant tradition in space-based theories of visual attention was initiated with a seminal paper by Treisman and Gelade [13]. They argued that some

primary visual properties allow a search in parallel across large displays of target objects. In such cases the target appears to 'pop out' of the display. For example there is no problem in searching for a red item amongst distractor items colored green, blue or yellow, while searching for a green cross is much more difficult when distractors include red crosses and green circles ('feature conjunction'). Treisman and Gelade proposed that in the pop-out tasks preattentional mechanisms permit a rapid target detection, in contrast to the conjunction task, which was held to require a serial deployment of attention over each item in turn. They suggested the division of the attention in two stages: a first 'preattentive' one that is traditionally thought to be automatic, parallel, and to extract relatively simple stimulus properties, and second stage 'attentive' serial, slow, with limited processing capacity, able to extract more complex features. They proposed a model called Feature Integration Theory (FIT) [13], in which a set of low-level feature maps extracted in parallel on the entire input image (preattentive stage) are then combined together by a spatial attention window operating on a master saliency map (attentive stage).

In the literature a number of attention models that follow this hypothesis have been proposed, *e.g.* [14,15] (for a complete review on this topic see [16]). An important alternative model is given by Sun and Fisher [17], which propose a combination of object- and feature-based theories. Presented with a manually segmented input image, their model is able to replicate human viewing behavior for artificial and natural scenes. The limit of the model is the use of human segmentation of the images, in practice, it employs information that is not available in the preattentive stage, that is before the objects in the image are recognized.

2.1 Proto-objects and Visual Attention

It is known that the human visual system extracts basic information from the retinal image in terms of lines, edges, local orientation etc. Vision though does not only represent visual features but also the *things* that such features characterize. In order to segment a scene in items, objects, that is to group parts of the visual field as coherent wholes, the concept of 'object' must be known to the system. In particular, there is an intriguing discussion underway in vision science about reference to entities that have come to be known as 'proto-objects' or 'pre-attentive objects' [10,11,18], since they need not to correspond exactly with conceptual or recognizable objects. These are a step above the mere localized features, possessing some but not all of the characteristics of objects. Instead, they reflect the visual system's segmentation of current visual input into candidate objects (*i.e.* grouping together those parts of the retinal input which are likely to correspond to parts of the same object in the real world, separately from those which are likely to belong to other objects). Hence the "objects" which we will be concerned with are segmented perceptual units.

The visual attention model we propose simply considers these first stages of the human visual processing, and employs a concept of salience based on proto-objects defined as blobs of uniform color in the image. Since we are considering an embodied system we will use the output of the model, implemented for

real-time operation, to control the fixation point of a robotic head. Then, through action, the attention system can go beyond proto-objects, discovering "true" physical objects [19,20]. The proposed object-based model of visual attention integrates bottom-up and top-down cues; in particular, top-down information works as a priming mechanism for certain regions in the visual search task.

3 The Model

In Figure 1 there is the block diagram of the model. We will describe in details in the following each block.

The input is a sequence of color log-polar images [21]. The use of log-polar images comes from the observation that the distribution of the cones, *i.e.* the retinal photoreceptors involved in diurnal vision, is not uniform. Cones have a higher density in the central region called fovea (approximately 2° of the visual field), while they are sparser in the periphery. This distribution influences the scanpaths during a visual search task [22] and so it has to be taken into account to better model overt visual attention. The log-polar mapping is in fact a model of the topological transformation of the primate visual pathways from the Cartesian image coming from the retina to the visual cortex, that takes also into account the space-variant resolution of the retinal images. This transformation can be well described as a logarithmic-polar (log-polar) mapping [21]. Figure 2 shows an example image and its log-polar counterpart.

One advantage of log-polar images is related to the small number of pixels and the comparatively large field of view. In fact the lower resolution of the periphery

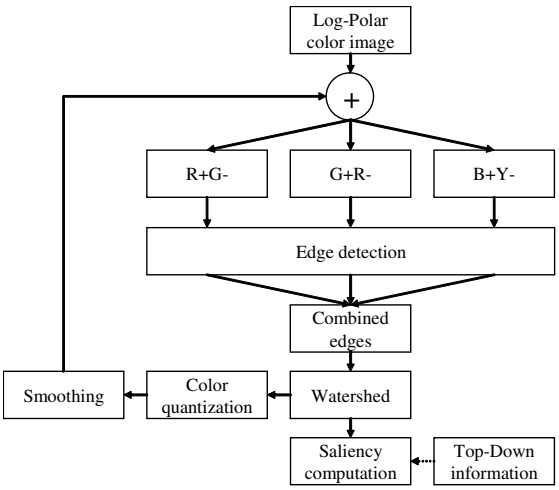


Fig. 1. Block diagram of the model. The input image is first separated in the three color opponency maps, then edges are extracted. A watershed transform creates the proto-objects on which the saliency is calculated, taking into account top-down biases.

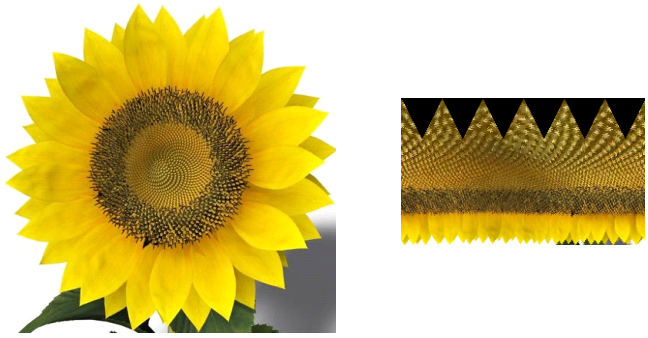


Fig. 2. Log-polar transform of an image. It is worth noting that the flower's petals, that have a polar structure, are mapped vertically in the log-polar image. Circles, on the other hand, are mapped horizontally. Furthermore, the stamens that lie in the center of the image of the flower, occupy about half of the corresponding log-polar image.

reduces the number of pixels and consequently the computational load of any processing, while standard algorithms can still be used on the high resolution central part (the fovea).

3.1 Feature Extraction

As a first step the input image at time t is averaged with the output of a color quantization procedure (see later) applied to the image at time $t-1$. This is to reduce the effect of the input noise. The red, green, blue channels of each image are then separated, and the yellow channel is constructed as the arithmetic mean of the red and green channels. Successively these four channels are combined to generate three color opponent channels, similar to those of the retina. Each channel, normally indicated as R^+G^-, G^+R^-, B^+Y^-, has a center-surround receptive field (RF) with spectrally opponent color responses. That is, for example, a red input in the center of a particular RF increases the response of the channel R^+G^- , while a green one in the surrounding will decrease its response. The spatial response profile of the two sub-regions of the RF, 'center' and 'surround', is expressed by a Gaussian, resulting in a Difference-of-Gaussians (DoG) response. A response is computed as there was a RF centered on each pixel of the input image, thus generating an output image of the same size of the input. This operation, considering for example the R^+G^- channel is expressed by:

$$R^+G^-(x,y) = \alpha \cdot R * g_c - \beta \cdot G * g_s \tag{1}$$

The two Gaussian functions, g_c and g_s, are not balanced: the ratio β/α is chosen equal to 1.5, consistent with the study of Smirnakis *et al.* [23]. The unbalanced ratio preserves achromatic information: that is, the response of the channels to a uniform gray area is not zero. The model does not need to process achromatic information explicitly since it is implicitly encoded, similarly to what happens

in the human retina's P-cells [24]. The ratio σ_s/σ_c, the standard deviation of the two Gaussian functions, is chosen equal to 3. To be noted that by filtering a log-polar image with a standard space-invariant filter leads to a space-variant filtered image of the original Cartesian image [25]. Edges are then extracted on the three channels separately using a generalization of the Sobel filter due to [26], obtaining $E_{RG}(x, y)$, $E_{GR}(x, y)$ and $E_{BY}(x, y)$. A single edge map is generated combining the tree outputs with a pixel-wise max(·) operator:

$$E(x, y) = \max \{|E_{RG}(x, y)|, |E_{GR}(x, y)|, |E_{BY}(x, y)|\} \tag{2}$$

3.2 Proto-objects

It has been speculated, that synchronizations of visual cortical neurons might serve as the carrier for the observed perceptual grouping phenomenon [27,28]. The differences in the phase of oscillation among spatially neighboring cells are believed to contribute to the segmentation of different objects in the scene. We have used a watershed transform (rainfalling variant) [29] on the edge map to simulate the result of this synchronization phenomenon and to generate the proto-objects. The intuitive idea underlying the watershed transform comes from geography: a topographic relief is flooded by water, watershed are the divide lines of the domains of attraction of rain falling over the region. In our model the watershed transform simulates the parallel spread of the activation on the image, until this procedure fills all the spaces between edges. Differently from other similar methods the edges themselves will never be tagged as blobs and the method does not require complex membership functions either. Moreover the result does not depend on the order in which the points are examined like in standard region growing [30]. As a result, the image is segmented into blobs which are either uniform or with a uniform gradient of color.

The definition of proto-objects is directly derived from the choice of the feature maps: i.e. closed areas of the image uniform in color.

A color quantized image is formed averaging the color inside each blob. The result is blurred with a Gaussian filter and stored: this will be used to perform temporal smoothing by simply averaging with the frame at time $t + 1$ to reduce the effect of noise and increase the temporal stability of the blobs. After an initial startup time of about five frames, the number of blobs and their shape stabilize. If movement is detected in the image then the smoothing procedure is halted and the bottom-up saliency map becomes the motion image.

A feature or a stimulus catches the attention if it differs from its immediate surrounding. To replicate this phenomenon in the system we compute a measure of bottom-up salience as the Euclidean distance in the color opponent space between each blob and its surrounding. However a constant size of the spot or focus of attention would not be very practical and rather it should change depending on the size of the objects in the scene. To account for this fact the greater part of the visual attention models in literature uses a multi-scale approach filtering with some type of 'blob' detector (typically a DoG filter) at various scales [16]. We reasoned that this approach lacks continuity in the choice of the size of the

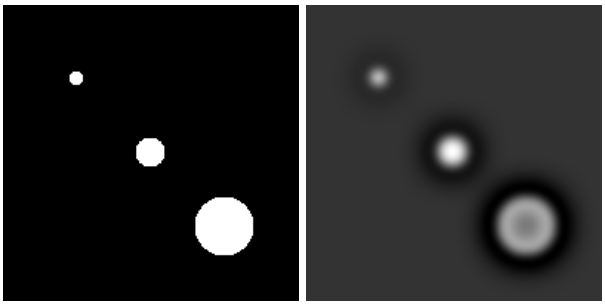


Fig. 3. The effect of a fixed size Difference-of-Gaussians filter. Blobs smaller of the positive lobe of the filter are depressed while larger ones are depressed in their centers.

focus of attention (see for example Figure 3). We propose instead to dynamically vary the region of interest depending on the size of the blobs. That is, the salience of each blob is calculated in relation to a neighborhood proportional to its size. In our implementation we consider a rectangular region 3 times the size of the bounding box of the blob as surrounding region, centered on each blob. The choice of a rectangular window is not incidental: filters over rectangular regions can be computed efficiently by employing the integral image as in [31].

The bottom-up saliency is thus computed as:

$$\begin{aligned} S_{bottom-up} &= \sqrt{\Delta RG^2 + \Delta GR^2 + \Delta BY^2} \\ \Delta RG &= \langle R^+G^- \rangle_{blob} - \langle R^+G^- \rangle_{surround} \\ \Delta GR &= \langle G^+R^- \rangle_{blob} - \langle G^+R^- \rangle_{surround} \\ \Delta BY &= \langle B^+Y^- \rangle_{blob} - \langle B^+Y^- \rangle_{surround} \end{aligned} \tag{3}$$

where $\langle\rangle$ indicates the average of the image values over a certain area (indicated in the subscripts). The top-down influence on attention is, at the moment, calculated in relation to the task of visually searching for a given object. In this situation a model of the object to search in the scene is given and this information is used to bias the saliency computation procedure. In practice, the top-down saliency map, $S_{top-down}$, is computed as the Euclidean distance in the color opponent space, between each blob's average color and the average color of the target, with a formula similar to (4). Blobs that are too small or too big in relation to the size of the images are discarded from the computation of salience with two thresholds. The blob in the center of the image (currently fixated) is also ignored because it cannot be the target of the next fixation. The total salience is simply calculated as the linear combination of the top-down and bottom-up contributions:

$$S = k_{td} \cdot S_{top-down} + k_{bu} \cdot S_{bottom-up} \tag{4}$$

The center of mass of the most salient blob is selected for the next saccade, in fact it has been observed that the first fixation to a simple shape that appears in the periphery tends to land on its center of gravity [32].

3.3 Inhibition of Return

In order to avoid being redirected immediately to a previously attended location, a local inhibition is transiently activated in the saliency map. This is called 'inhibition of return' (IOR) and it has been demonstrated in human visual psychophysics. In particular Tipper [33] was among the firsts to demonstrate that the IOR could be attached to moving objects. Hence the IOR works by anchoring tags to objects as they move; in other words this process seems to be coded in an object-based reference frame.

Our system implements a simple object-based IOR. A list of the last 5 positions visited is maintained in a head-centered coordinate system and updated with a FIFO (First In First Out) policy. The position of the tagged blob is stored together with the information about its color. When the robot gaze moves — for example by moving the eyes and/or the head — the system keeps track of the blobs it has visited. These locations are inhibited only if they show the same color seen earlier: so in case an inhibited object moves or its color changes, the location becomes available for fixation again.

4 Results on sample images

Even if our model is inherently built not to work on static images, we have tried a comparison with the model of Itti *et al.* [15], using the same database of images they use [34]. It consists of 64 color images with an emergency triangle and relative binary segmentation masks of the triangle[1]. First, the original images and segmentation masks are cropped to a square and transformed to the log-polar format (see Figure 4 (*a*) and (*b*) for the Cartesian remapped images). To simulate the presence of a static camera, the images are presented to the system continuously and, after five 'virtual' frames, the bottom-up saliency map is confronted with the mask. In this way we measure the ability of the system to spot the salient object in the images, simulating the pop-out phenomenology. The obtained result is that in 49% of the images a point inside the emergency triangle is selected as the most salient (see an example in Figure 4 (*c*)). However a direct comparison with the results of Itti and Koch in [34], by counting the number of false detection before the target object is found, is not possible since after each saccade the log-polar image changes considerably.

Other experiments were carried out on a robotic platform called Babybot [35]. This is a humanoid upper torso which consists of a head, an arm and a hand. From the point of view of the sensors, the head is equipped with two log-polar cameras and two microphones for visual and auditory feedback. The attentional system were used to guide the object recognition system and to guide the robot in manipulation tasks

[1] `http://ilab.usc.edu/imgdbs/`, last access 30/05/2007.

Fig. 4. Result on a sample image taken from [34]. (a) is the log-polar input image and (b) the corresponding taget binary mask. (c) is the bottom-up saliency map.

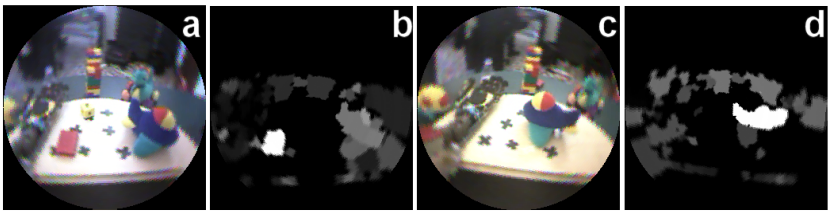


Fig. 5. Example saliency maps. (b) is the bottom-up saliency map of the image (a). (d) is the top-down saliency map of (c) while searching for the blue airplane.

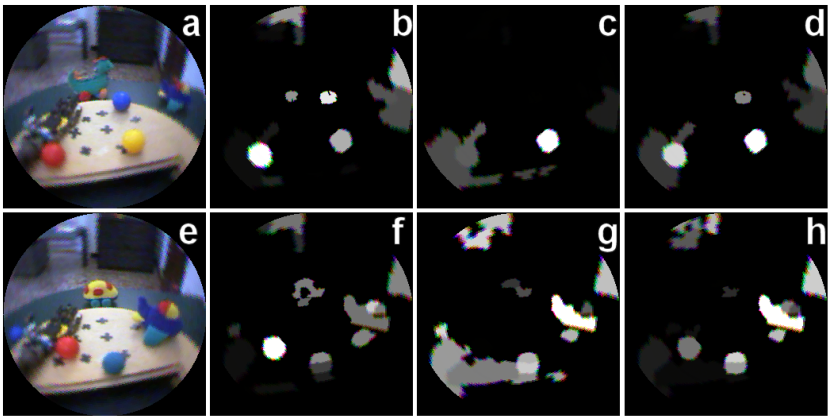


Fig. 6. Combining top-down and bottom-up maps. (b) and (f) are the bottom-up saliency maps of (a) and (e). (c) and (g) are the top-down ones, while searching respectively for the yellow ball and the blue airplane. In (d) and (h) the bottom-up and top-down contributions are equally weighted; this can result in clearer maps.

[35,20]. Two examples of saliency maps from the input images of the robot are shown in Figure 5: in (b) there is a purely bottom-up ($k_{td} = 0$, $k_{bu} = 1$ in Equation (7)) map which is the result of the processing of the scene in (a); in (d) there is a purely top-down ($k_{td} = 1$, $k_{bu} = 0$) map output after the processing of (c). In Figure 6 there are the saliency maps of two images with different settings of k_{td} and k_{bu}.

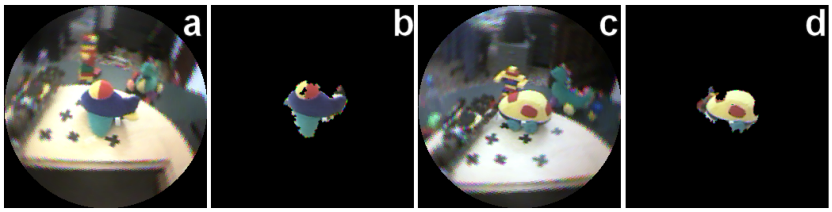


Fig. 7. Example segmentations of objects. (*b*) and (*d*) are obtained from (*a*) and (*c*) using the proto-objects that are estimated to belong to the target objects.

Moreover using any learning procedure it is possible to estimate which proto-objects compose a particular object and use this information to attempt a figure-ground segmentation [20]. An example of these segmentations is shown in Figure 7. Note that even if the result is not visually perfect, it has all the information to guide a manipulation task [35].

5 A Better Definition of Proto-objects

As said above, object-based theories of attention stress the importance of the segmentation of the visual input in coherent regions. The term 'grouping' (or 'segmentation') is a common concept in the long research history of perceptual grouping by the Gestalt psychologists. Back at the beginning of the last century they described, among other things, the ability of the human visual system to organize parts of the retinal stimulus into 'Gestalten', that is, into organized structures. They also formulated the so-called Gestalt laws (proximity, common fate, good continuation, closure, *etc.*) that are believed to govern our perception.

Nowadays the more typical view of such grouping demonstrations would be that they reflect non-arbitrary properties within the stimuli, which the visual system exploits heuristically because these properties are likely to reflect divisions into distinct objects in the real world. In this sense it should be possible to learn these heuristic properties and hence to learn from the image statistics better rules to build the proto-objects [12].

5.1 Learning the Association Fields

A first step in the implementation of the Gestalt laws are the 'association fields' [36]. These fields are supposed to resemble the pattern of excitatory and inhibitory lateral connection between different orientation detector neurons as found, for instance, by Schmidt *et al.* [37]. Schmidt has shown that cells with an orientation preference in area 17 of the cat are preferentially linked to iso-oriented cells. The coupling strength decrease with the difference in the preferred orientation of pre- and post-synaptic cell.

In the literature, association fields are often hand-coded and employed in many different models with the aim to reproduce the human performance in

(*a*) (*b*)

Fig. 8. (*a*) Sample input image from the Berkeley Segmentation Database. (*b*) Complex cells output to the image in (*a*) for 0° filter of formula (5).

contour integration. Models typically consider variations of the co-circular approach [38,39,40], which states that two oriented elements are very likely part of the same curve if they are tangent to the same circle. Our approach is instead to try to learn these association fields directly from natural images. Starting from the output of a simulated layer of complex cells, without any prior assumption, we want to estimate the mean activity around points with given orientations.

The extension of the fields is chosen to be of 41x41 pixels taken around each point, and the central pixel of the field is the reference pixel. We have chosen to learn 8 association fields, one for each discretized orientation of the reference pixel. Despite this quantization, to cluster the different fields, the information about the remaining pixels in the neighbor is not quantized, differently from other approaches, *i.e.* [41]. There is neither a threshold nor a pre-specified number of bins for discretization and thus we obtain a precise representation of the association fields. In the experiments we have used the images of the Berkeley Segmentation Database [42], that consists of 300 images of 321x481 and 481x321 pixels (see Figure 8 (*a*) for an example).

For mathematical convenience and to represent orientation precisely, we have chosen to use a tensor notation. Hence for each orientation of the reference pixel, we calculate the mean tensors associated with the surrounding pixels, from the 41x41 patches densely collected from 200 images of the database. These mean tensors will represent our association fields.

5.2 Feature Extraction Stage

There are several models of the complex cells of V1, but we have chosen to use the classic energy model [43]. The response at orientation θ is calculated as the sum a quadrature pair of even- and odd-symmetric filters:

$$E_\theta = \sqrt{(I * f_\theta^e)^2 + (I * f_\theta^o)^2} \tag{5}$$

Our even-symmetric filter is a Gaussian second-derivative, the corresponding odd-symmetric is its Hilbert transform. In Figure 8 (*b*) there is an example of the output of the complex cells model for the 0° orientation. Then the edges are

thinned using a standard non-maximum suppression algorithm. The outputs of these filters are used to construct our local tensor representation.

Second order symmetric tensors can capture the information about the first order differential geometry of an image. Each tensor describes both the orientation of an edge and its confidence for each point. In practice a second order tensor is denoted by a 2x2 symmetric matrix and can be visualized as an ellipse, whose major axis represents the estimated tangential direction and the difference between the major and minor axis the confidence of this estimate. Hence a point on a line will be associated with a thin ellipse while a corner with a circle. The tensor at each point is constructed by direct summation of three quadrature filter pair output magnitudes as in [44]:

$$T = \sum_{k=1}^{3} E_{\theta_k} \left(\frac{4}{3} \hat{n}_k^T \hat{n}_k - \frac{1}{3} I \right) \tag{6}$$

where I is the 2x2 identity matrix, E_{θ_k} is the filter output as calculated in (5) with θ_k corresponding to the direction of $\hat{n}_k$:

$$\hat{n}_1 = (1, 0)\,,\ \hat{n}_2 = \left(1/2, \sqrt{3}/2\right),\ \hat{n}_3 = \left(-1/2, \sqrt{3}/2\right) \tag{7}$$

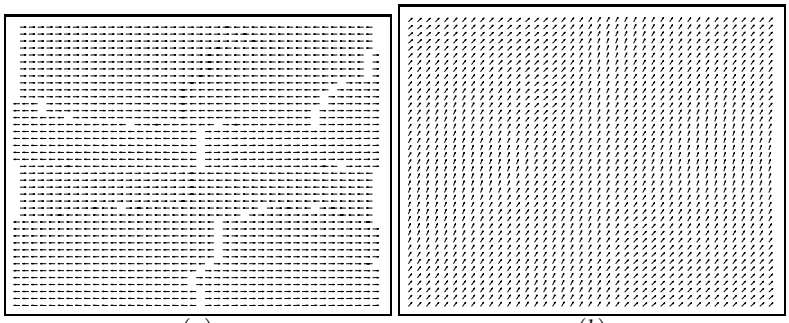
The greatest eigenvalue λ_1 and its corresponding eigenvector e_1 of a tensor associated to a pixel represent respectively the strength and the direction of the main orientation. The second eigenvalue λ_2 and its eigenvector e_2 have the same meaning for the orthogonal orientation. The difference $\lambda_1 - \lambda_2$ is proportional to the likelihood that a pixel contains a distinct orientation.

5.3 The Path Across a Pixel

We have run our test only for a single scale, choosing the σ of the Gaussian filters equal to 2, since preliminary tests have shown that a similar version of the fields is obtained with other scales as well. Two of the obtained fields are in Figure 9. It is clear that they are somewhat corrupted by the presence of horizontal and vertical orientations in any of the considered neighbors and by the fact that in each image patch there are edges that are not passing across the central pixel. On the other hand we want to learn association field for curves that *do* pass through the central pixel.

We believe that this is the same problem that Prodöhl *et al.* [45] experienced using static images: the learned fields supported collinearity in the horizontal and vertical orientations but hardly in the oblique ones. They solved this problem using motion to implicitly tag only the important edges inside each patch.

Once again the neural way to solve this problem can be the synchrony of the firing between nearby neurons (see Section 3.2). We considered for each image patch only pixels that belong to any curve that goes through the central pixel. In this way the dataset contains only information about curves connected to the central pixel. Note that we select curves inside each patch, not inside the entire image. The simple algorithm used to select the pixels is the following:



(*a*) (*b*)

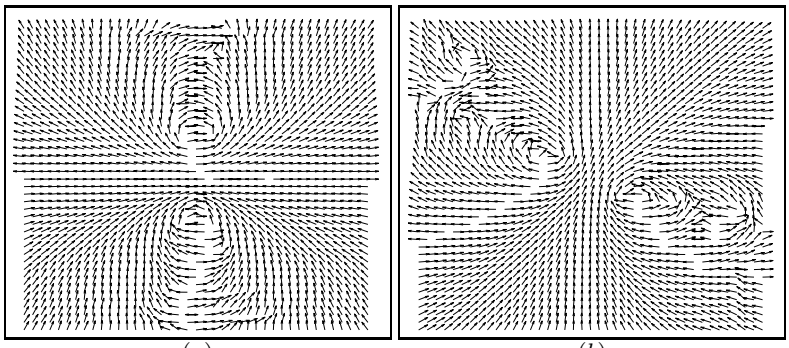
Fig. 9. Main directions for the association fields for the orientations of 0° (*a*) and 67.5° (*b*) in the central pixel

1. put central pixel of the patch in a list;
2. tag the first pixel in the list and remove it from the list. Put surrounding pixels that are active (non-zero) in the list;
3. if the list is empty quit otherwise go to 2.

This procedure removes the influence of horizontal and vertical edges that are more present in the images and that are not removed by the process of averaging. On the other hand, we are losing some information, for example about parallel lines, that in any case are not useful for the enhancement of contours. Note that this method is completely "parameter free"; we are not selecting the curves following some specific criterion, instead we are just pruning the training set from noisy or biased inputs. It is important to note that this method will learn the natural image bias toward horizontal and vertical edges [46], but it will not be biased to learn these statistics *only*, as in Prodöhl *et al.* [45] when using static images. A similar approach that uses self-caused motion has been developed in [47] to disambiguate the edges of a target object from those in the background.

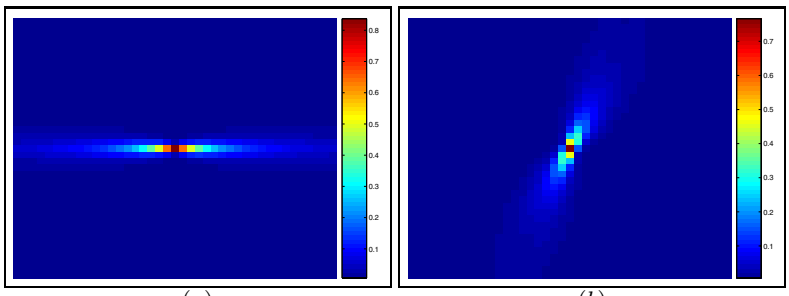
6 Validating the Association Fields

Figures 10 and 11 show the main orientations and strengths (eigenvalues) of the mean estimated tensors for the orientations of 0° and 67.5° of the central pixel, obtained with the modified procedure described in Section 5.3. The structure of the obtained association fields closely resembles the fields proposed by others based on collinearity and co-circularity. While all the fields have the same trend, there is a clear difference in the decay of the strength of the fields. To see this we have considered only the values along the direction of the orientation in the center, normalizing the maximum values to one. Figure 12 (*a*) shows this decay. It is clear that fields for horizontal and vertical edges have a wider support, confirming the results of Sigman *et al.* [41].



(*a*) (*b*)

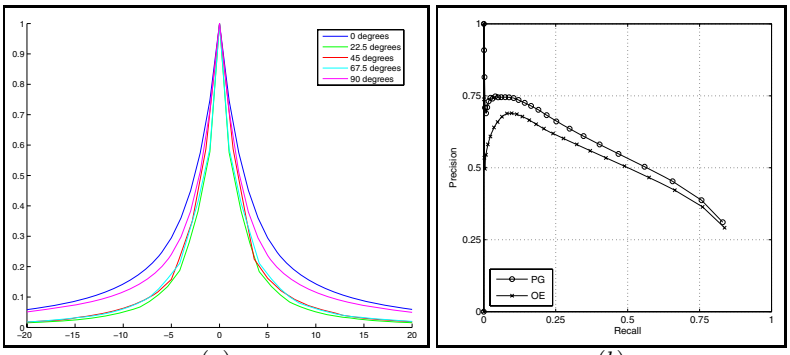
Fig. 10. Main directions for the association field for orientation of 0° (*a*) and 67.5° (*b*), with the modified approach. Compare them with the results in Figure 9.



(*a*) (*b*)

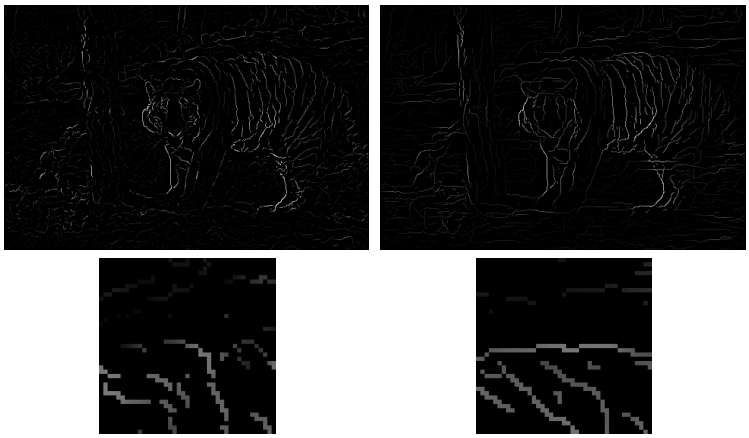
Fig. 11. Difference between the two eigenvalues of the association fields of Figure 10

The obtained fields can be used with any existing model of contour enhancement, but to test them we have used the tensor voting scheme proposed by Guy and Medioni *et al*. [39]. The choice is somewhat logical considering to the fact that the obtained fields are already tensors. In the tensor voting framework points communicate with each other in order to refine and derive the most preferred orientation information. We compared the performances of the tensor voting algorithm using the learned fields versus the simple output of the complex cell layer, using the Berkeley Segmentation Database and the methodology proposed by Martin *et al*. [48,42]. In the databes for each image a number of different human segmentations is available. The methodology proposed by Martin *et al*. aims at measuring with ROC-like graphs the distance between the human segmentations and the artificial ones. We can see the results on 100 test images and relatives human segmentations in Figure 12 (*b*), better result are associated with curves that are located higher in the graph. We can see that there is always an improvement using the tensor voting and the learned association fields instead of just using the outputs of the complex cells alone. An example of the results on the test image in Figure 8 (*a*), after the non-maximum suppression procedure, are shown in Figure 13.



(a) (b)

Fig. 12. (*a*) Comparison of the decay for the various orientations. On the y axis there are the first eigenvalues normalized to a maximum of 1, on the x axis is the distance from the reference point along the main field direction. (*b*) Comparison between tensor voting with learned fields (PG label) and the complex cell layer alone (OE label).



(a) (b)

Fig. 13. (*a*) Test image contours using the complex cell layer alone. (*b*) Test image contours using tensor voting with the learned fields. Notice the differences with the (*a*): the contours are linker together and the gaps are reduced. Especially on the contour of back of the tiger the differences are evident (bottom images).

7 Conclusion

We have presented the general implementation of a visual attention system employing both top-down and bottom-up information. It runs in real time on a standard Pentium class processor and it is used to control the overt attention system of a humanoid robot. Running an attention system on a robotic platform

generates a set of problems which are not apparent when only generating scan paths on static images. Although not discussed in details here, the robot implementation requires, for example, a complex management of the IOR together with a body-centered coordinate system (for representing object locations).

Our algorithm divides the visual scene in color blobs; each blob is assigned a bottom-up saliency value depending on the contrast between its color and the color of the surrounding area. The robot acquires information about objects through active exploration and uses it in the attention system as a top-down primer to control the visual search of that object. The model directs the attention on the proto-object's or center of mass, similarly to the behavior observed in humans (see Sections 3.2 and 4). In [35,20] the proposed visual attention system was also used to guide the grasping action of a humanoid robot.

A similar approach has been taken by Sun and Fisher [17] but the main difference with this work is that they have assumed that a hierarchical set of perceptual groupings is provided to the attention system by some other means and considered only covert attention. In this sense we have tried to address this problem directly presenting a method to learn precise association fields from natural images. An unsupervised bio-inspired procedure to get rid of the non-uniform distribution of orientations is used, without the need of the use of motion [45]. The learned fields were used in a computer model and the results were compared using a database of human tagged images which helps in providing clear numerical results.

Moreover the framework introduced is general enough to work with other additional feature maps, extending the watershed transform to additional dimensions in feature space (e.g. local orientation) thus providing new ways of both segmenting and recognizing objects. As future work we want to integrate the associative fields learnt from natural images with the proposed visual attention model. We are also looking to an extension of the associative fields to a hierarchical organization to develop even more complex image features.

References

1. Cave, K., Bichot, N.: Visuospatial attention: beyond a spotlight model. Psychonomic Bulletin & Review 6, 204–223 (1999)
2. Kawato, M.: Internal models for motor control and trajectory planning. Current Opinion in Neurobiology 9, 718–727 (1999)
3. O'Regan, J.: Solving the "real" mysteries of visual perception: the world as an outside memory. Canadian Journal of Psychology 46, 461–488 (1992)
4. Maturana, R., Varela, F.: Autopoiesis and Cognition: The Realization of the Living. D.Reidel Publishing Co., Dordecht (1980)
5. van Gelder, T., Port, R.: It's about time: An overview of the dynamical approach to cognition. In: van Gelder, T., Port, R. (eds.) Mind as motion - Explorations in the Dynamics of Cognition, MIT Press, Cambridge, MA (1995)
6. Craighero, L., Fadiga, L., Rizzolatti, G., Umilta', C.: Action for perception: a motor-visual attentional effect. J. Exp. Psychol. Hum. Percept. Perform. 25, 1673–1692 (1999)

7. Fadiga, L., Fogassi, L., Gallese, V., Rizzolatti, G.: Visuomotor neurons: ambiguity of the discharge or 'motor' perception? Int. J. Psychophysiol. 35, 165–177 (2000)
8. Fischer, M.H., Hoellen, N.: Space- and object-based attention depend on motor intention. The Journal of General Psychology 131, 365–378 (2004)
9. Scholl, B.J.: Objects and attention: the state of the art. Cognition 80, 1–46 (2001)
10. Rensink, R.A., O'Regan, J.K., Clark, J.J.: To see or not to see: The need for attention to perceive changes in scenes. Psychological Science 8(5), 368–373 (1997)
11. Rensink, R.A.: Seeing, sensing, and scrutinizing. Vision Research 40(10–12), 1469–1487 (2000)
12. Palmer, S., Rock, I.: Rethinking perceptual organization: the role of uniform connectedness. Psychonomic Bulletin & Review 1(1), 29–55 (1994)
13. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology 12(1), 97–136 (1980)
14. Milanese, R., Gil, S., Pun, T.: Attentive mechanisms for dynamic and static scene analysis. Optical Engineering 34, 2428–2434 (1995)
15. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 20, 1254–1259 (1998)
16. Itti, L., Koch, C.: Computational modeling of visual attention. Nature Reviews Neuroscience 2(3), 194–203 (2001)
17. Sun, Y., Fisher, R.: Object-based visual attention for computer vision. Artificial Intelligence 146, 77–123 (2003)
18. Pylyshyn, Z.W.: Visual indexes, preconceptual objects, and situated vision. Cognition 80(1-2), 127–158 (2001)
19. Metta, G., Fitzpatrick, P.: Early integration of vision and manipulation. Adaptive Behavior 11, 109–128 (2003)
20. Orabona, F.: Learning and Adptation in Computer Vision. PhD thesis, University of Genoa (2007)
21. Sandini, G., Tagliasco, V.: An anthropomorphic retina-like structure for scene analysis. Computer Vision, Graphics and Image Processing 14, 365–372 (1980)
22. Wolfe, J.M., Gancarz, G.: Guided search 3.0. In: Lakshminarayanan, V. (ed.) Basic and Clinical Applications of Vision Science, pp. 189–192. Kluwer Academic, Dordrecht, Netherlands (1996)
23. Smirnakis, S.M., Berry, M.J., Warland, D.K., Bialek, W., Meister, M.: Adaptation of retinal processing to image contrast and spatial scale. Nature 386, 69–73 (1997)
24. Billock, V.A.: Cortical simple cells can extract achromatic information from the multiplexed chromatic and achromatic signals in the parvocellular pathway. Vision Research 35, 2359–2369 (1995)
25. Mallot, H.A., von Seelen, W., Giannakopoulos, F.: Neural mapping and space-variant image processing. Neural Networks 3(3), 245–263 (1990)
26. Li, X., Yuan, T., Yu, N., Yuan, Y.: Adaptive color quantization based on perceptive edge protection. Pattern Recognition Letters 24, 3165–3176 (2003)
27. Eckhorn, R., Bauer, R., Jordan, W., Brosch, M., Kruse, M., Munk, W., Reitboeck, H.J.: Coherent oscillations: A mechanism of feature linking in the visual cortex? Biological Cybernetics 60, 121–130 (1988)
28. Gray, C.M., König, P., Engel, A.K., Singer, W.: Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature 338, 334–336 (1989)
29. De Smet, P., Pires, R.L.V.: Implementation and analysis of an optimized rainfalling watershed algorithm. In: Proc. of SPIE, VCIP'2000, vol. 3974, pp. 759–766 (2000)
30. Wan, S., Higgins, W.: Symmetric region growing. IEEE Trans. on Image Processing 12(9), 1007–1015 (2003)

31. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
32. Melcher, D., Kowler, E.: Shapes, surfaces and saccades. Vision Research 39, 2929–2946 (1999)
33. Tipper, S.P.: Object-centred inhibition of return of visual attention. Quarterly Journal of Experimental Psychology 43A, 289–298 (1991)
34. Itti, L., Koch, C.: Feature combination strategies for saliency-based visual attention systems. Journal of Electronic Imaging 10(1), 161–169 (2001)
35. Natale, L., Orabona, F., Berton, F., Metta, G., Sandini, G.: From sensorimotor development to object perception. In: Proc. of the 5th IEEE-RAS International Conference on Humanoid Robots, Tsukuba, Japan, pp. 226–231 (2005)
36. Field, D.J., Hayes, A., Hess, R.F.: Contour integration by the human visual system: evidence for local "association field". Vision Research 33(2), 173–193 (1993)
37. Schmidt, K., Goebel, R., Löwel, S., Singer, W.: The perceptual grouping criterion of collinearity is reflected by anisotropies of connections in the primary visual cortex. European Journal of Neuroscience 5(9), 1083–1084 (1997)
38. Grossberg, S., Mingolla, E.: Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations. Percept. Psychophys. 38, 141–171 (1985)
39. Guy, G., Medioni, G.: Inferring global perceptual contours from local features. Int. J. of Computer Vision 20, 113–133 (1996)
40. Li, Z.: A neural model of contour integration in the primary visual cortex. Neural Computation 10, 903–940 (1998)
41. Sigman, M., Cecchi, G.A., Gilbert, C.D., Magnasco, M.O.: On a common circle: Natural scenes and gestalt rules. PNAS 98(4), 1935–1940 (2001)
42. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. of ICCV 2001, vol. 2, pp. 416–423 (2001)
43. Morrone, M., Burr, D.: Feature detection in human vision: A phase dependent energy model. Proc. Royal Soc. of London B 235, 221–245 (1988)
44. Knutsson, H.: Representing local structure using tensors. In: Proceedings 6th Scandinavian Conference on Image Analysis, Oulu, Finland, pp. 244–251 (1989)
45. Prodöhl, C., Würtz, R.P., von der Malsburg, C.: Learning the gestalt rule of collinearity from object motion. Neural Computation 15, 1865–1896 (2003)
46. Coppola, D.M., Purves, H.R., McCoy, A.N., Purves, D.: The distribution of oriented contours in the real world. PNAS 95, 4002–4006 (1998)
47. Fitzpatrick, P., Metta, G.: Grounding vision through experimental manipulation. Philos. trans. - Royal Soc., Math. phys. eng. sci. 361(1811), 2185–2615 (2003)
48. Martin, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries using local brightness, color and texture cues. IEEE Trans. Pattern Anal. Mach. Intell. 26(5), 530–549 (2004)