Object-based Visual Attention: a Model for a Behaving Robot

Francesco Orabona, Giorgio Metta and Giulio Sandini

LIRA-Lab, DIST
University of Genoa, Genoa, ITALY 16145
{bremen,pasa,giulio}@liralab.it

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 the literature, the problem, in this case, being that of creating a model of "objecthood" that eventually guides a saliency mechanism. We propose here an object-based model of visual attention and show its instantiation on a humanoid robot. The robot employs action to learn and define its own concept of objecthood.

1. Introduction

Humans and many animals have a space-variant visual system that require them to move their eyes, three times a second on average, in order to position their foveae onto interesting locations of the visual space. This allows taking a series of small "snapshots" at very high-resolution. The fact that this is the only way that allows clear "vision" implies the existence of an attention system which, at any moment in time, selects the point to fixate next.

This leads to two sorts of questions: i) how to move the eyes efficiently to important locations in the visual scene, and ii) how to decide what is important and, as a consequence, where to look next. To answer these questions, we should note first 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" [1]. These are a step above the mere localized features, possessing some but not all of the characteristics of objects.

The visual attention model we propose 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 an instantiation of the model to control the fixation point of a robotic head. Moreover, through action, the attention system can go beyond proto-objects [2]. In fact, once an object is grasped, the robot can move and rotate it to build a statistical model of the features belonging to it, constructing a representation as a collection of proto-objects and their relative spatial locations. This internal representation then generates a top-down signal that bias attention toward known objects; as an example we will show how the top-down influence can be used to direct the attention of the robot to spot a specific object among other similar items lying on a table.

The rest of the paper is organized as follows. Section 2 contains an introduction on the modeling of human visual attention. Section 3 describes the experimental setup used in the experiments. Section 4 details the robot's visual system and the implementation. Section 5 introduces the probabilistic object model and shows how this is used for object recognition. Finally in sections 6 and 7 we show experimental results and we draw some conclusions.

2. Human visual attention

The attempts of modeling human visual attention traditionally rest on two assumptions. On the one hand, the *space-based* theory holds that attention is allocated to a region of space, with processing carried out only within a certain spatial window. This theory considers attention as a “spotlight”, an internal eye or a sort of “zoom lens”; attention is deployed as a spatial gradient, centered on a particular location.

On the other hand, *object-based* attention theories argue that attention is directed to an object or a group of objects, to process specific properties of the selected objects, rather than 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 [3] for a review). This reflects the fact that the visual system is optimized for segmenting complex scenes into representations of (often partly occluded) objects to be used both for recognition and action, since perceivers must interact with objects and not with disembodied spatial locations.

Finally, another classification can be made depending on which cues are actually used in modulating attention. Bottom-up information, which comes only from the input image, includes basic features such as color, orientation, motion, depth, and their conjunction thereof. A feature or a stimulus catches attention if it differs from its immediate surrounding in some dimensions and the surround is reasonably homogeneous in those same dimensions. However, in attention, higher-level mechanisms are involved as well. A bottom-up stimulus, for example, may be ignored if attention is already focused elsewhere [4]. In this case attention is also influenced by top-down information relevant to the particular task at hand which is not necessarily available in the image.

In the literature a number of attention models that follow the first hypothesis have been proposed [5-7]. Most of them are derived from Treisman’s Feature Integration Theory (FIT) [8], which employs a separate set of low-level feature maps that are combined together by a spatial attention window operating on a master saliency map. An important alternative model is given by Sun and Fisher [9], which propose an combination of object-and feature-based theories.

This paper presents an object-based model of visual attention that 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. Setup

The experiments reported in the paper were carried out on a robotic platform called Babybot. This is a humanoid upper torso which consists of a head, an arm and a hand. The head has 5 degrees of freedom, two of which control the neck pan and tilt, whereas the other three actuate two eyes to pan independently and tilt on a common axis. The arm is the well known Unimate PUMA 260, an industrial manipulator with 6 degrees of freedom; the hand has 5 fingers for a total of 6 degrees of freedom.

From the point of view of the sensors, the head is equipped with two cameras and two microphones for visual and auditory feedback. Proprioceptive information is provided to the robot by optic and magnetic encoders mounted on all joints of the head, arm and hand. More details about the Babybot can be found for example in [10].

4. Model

A block diagram of the model is shown in Figure 1; the input is a sequence of color log-polar images [11]. The use of log-polar images comes from the observation that the distribution of the cones, i.e. the photoreceptors of the retina involved in diurnal vision, is not uniform. This distribution seems to influence the scanpaths during a visual search task and so it has to be taken into account to better model overt visual attention [12]. In addition, the lower resolution of the periphery of the field of view reduces the images’ size and thus reduces the computational load.

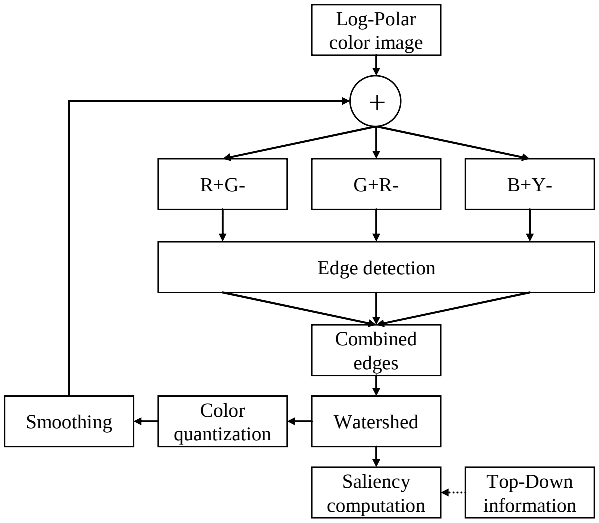

Figure 1. Block diagram of the model.

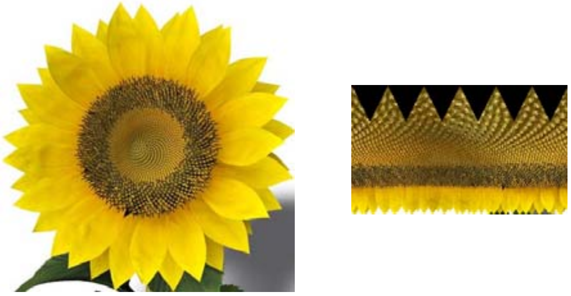

Figure 2. Log-polar transform of an image.

4.1. Log-polar images

The log-polar mapping is a model of the topological transformation of the primate visual pathways from the retina to the visual cortex. Cones have a higher density in the central region called fovea (approximately 2° of the visual field), while they are sparser in the periphery. Consequently, the resolution is higher and uniform in the center while it decreases in the periphery, moving away from the fovea.

From the mathematical point of view the log-polar mapping can be expressed as a transformation between the polar plane *(ρ,θ)* (retinal plane), the log–polar plane (ξ,η) (cortical plane) and the Cartesian plane *(x,y)* (image plane), as follows [11]:

$$\begin{cases} \eta = q \cdot \theta \\ \xi = \log_a \dfrac{\rho}{\rho_0} \end{cases} \qquad (1)$$

where ρ_0 is the radius of the innermost circle, *1/q* is the minimum angular resolution of the log-polar layout and *(ρ,θ)* are the polar co-ordinates. These are related to the conventional Cartesian reference system by:

$$\begin{cases} x = \rho \cdot \cos\theta \\ y = \rho \cdot \sin\theta \end{cases} \qquad (2)$$

Figure 2 shows a Cartesian image and its log-polar counterpart as derived from Equations (1) and (2).

4.2. 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 RF is expressed by a Difference-of-Gaussians (DoG) over the two sub-regions of the RF, ‘center’ and ‘surround’. 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:

$$R^+G^-(x,y) = \alpha \cdot R * g_c - \beta \cdot G * g_s \qquad (3)$$

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. [13]. The unbalanced ratio preserves the achromatic information: that is, the response of the channels to a uniform gray area is not zero. Hence 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 [14]. 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 [15].

Edges are then extracted on the three channels separately using a generalization of the Sobel filter due to [16], obtaining $E_{RG}(x,y)$, $E_{GR}(x,y)$ and $E_{BY}(x,y)$. A single edge map is generated combining the tree outputs:

$$E(x,y) = \max\left\{\left|E_{RG}(x,y)\right|, \left|E_{GR}(x,y)\right|, \left|E_{BY}(x,y)\right|\right\} \qquad (4)$$

The log-polar transform has the side effect of sharpening the edges near the fovea due to the magnification factor of the mapping; this is compensated multiplying each pixel by a factor which is exponential on the eccentricity.

4.3. Proto-objects

It has been speculated, that synchronizations of visual cortical neurons might serve as the carrier for the observed perceptual grouping phenomenon [17, 18]. The differences in the phase of oscillation among spatially neighboring cells is believed to contribute to the segmentation of different objects in the scene.

We have used a watershed transform (rainfalling variant) [19] on the edge map to simulate the result of this synchronization phenomenon and to generate the proto-objects.

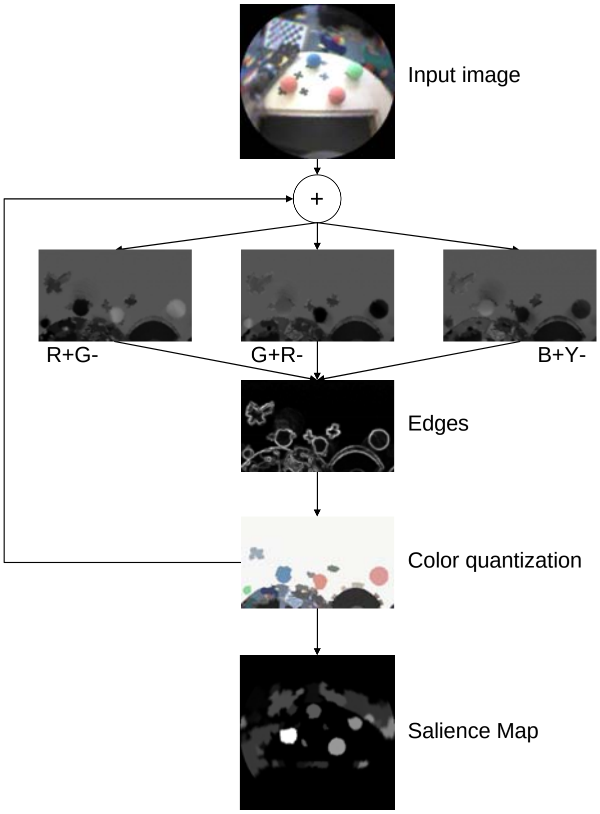

Figure 3. Example of model maps.

The intuitive idea underlying this method 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 view 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 [20]. As a result, the image is segmented into blobs with either uniform or uniform gradient of color.

Each blob is tagged with the average of the color of the pixels within its area (this leads to a sort of color quantizated image). The result is blurred with a gaussian filter and stored: this will be used to perform a time-smoothing by simple 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 (as difference between two consecutive frames) then the smoothing procedure is halted and the bottom-up saliency map becomes the motion image.

As already mentioned above, a feature or a stimulus catches the attention of the system if it differs from its immediate surrounding. We chose to compute the bottom-up salience as the Euclidean distance in the color opponent space between each blob and its surrounding. The size of the spot or focus of attention is not constant: it changes 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 difference of Gaussian filter) at various scales [21]. We reasoned that this approach lacks continuity in the choice of the size of the focus of attention. 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, rather it was chosen because filters over rectangular regions can be computed efficiently by employing the integral image as in [22]. The bottom-up saliency is thus computed as:

$$\begin{aligned} S_{bottom-up} &= \sqrt{\Delta RG^2 + \Delta GR^2 + \Delta BY^2} \\ \Delta RG &= \left\langle R^+G^- \right\rangle_{blob} - \left\langle R^+G^- \right\rangle_{surround} \\ \Delta GR &= \left\langle G^+R^- \right\rangle_{blob} - \left\langle G^+R^- \right\rangle_{surround} \\ \Delta BY &= \left\langle B^+Y^- \right\rangle_{blob} - \left\langle B^+Y^- \right\rangle_{surround} \end{aligned} \tag{5}$$

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 a given object. In this situation a model of the object to search in the scene is given (see Section 5) and this information is used to bias the saliency computation procedure. In practice, the top-down saliency map is computed as the Euclidean distance in the color opponent space, between each blob’s average color and the average color of the target:

$$\begin{aligned} S_{top-down} &= \sqrt{\Delta RG^2 + \Delta GR^2 + \Delta BY^2} \\ \Delta RG &= \left\langle R^+G^- \right\rangle_{blob} - \left\langle R^+G^- \right\rangle_{object} \\ \Delta GR &= \left\langle G^+R^- \right\rangle_{blob} - \left\langle G^+R^- \right\rangle_{object} \\ \Delta BY &= \left\langle B^+Y^- \right\rangle_{blob} - \left\langle B^+Y^- \right\rangle_{object} \end{aligned} \tag{6}$$

with a notation similar to the one above.

Blobs that are too small (1/550 of image area) or too big (1/4 of the image area) are discarded from the computation of salience and will not be considered as possible candidates to be part of objects. The blob in the center of the image (currently fixated) is ignored also 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:

$$S = k_{td} \cdot S_{top-down} + k_{bu} \cdot S_{bottom-up} \quad (7)$$

and normalized in the range 0-255. The center of mass of the most salient blob is selected for the next saccade. An example of the intermediate and final maps of bottom-up salience is shown in Figure 3. All the computations are done on log-polar images, but input and output images are shown remapped to cartesian for clarity.

4.4. 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. IOR does not seem to function in retinal coordinates but it is instead attached to environmental locations. It has been proposed that the IOR is required to allow an efficient visual search by discouraging shifting the attention toward locations that have already been inspected, and it seems to be working also in the case of moving objects (for a review see [23]).

All these findings lead to the conclusion that the human visual system works by tagging objects and moving tags as objects move, hence the IOR seems to be coded in an object-based frame of reference.

Our system implements a simple object-based IOR. A list of the last five positions visited [24] 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.

5. Learning about objects

We assume the robot has already grasped the object; this can happen because a collaborative human has given the object to the robot or because it has autonomously grasped the object (even by chance initially). Both solutions are valid bootstrapping behaviors for the acquisition of an internal model of the object. When the robot holds the object it can explore it by moving and rotating it.

Objects are represented by the blobs generated by the visual attention system and their relative positions (neighboring relations). The model is created statistically by looking at the same object for some time from different points of view. A histogram of the number of times a particular blob is seen is used to estimate the probability that the blob belongs to the grasped object.

In the following, we use the probabilistic framework proposed by Schiele and Crowley [25]. We want to calculate the probability of the object *O* given a certain local measurement *M*. This probability *P(O|M)* can be calculated using Bayes’ formula:

$$P(O \mid M) = \frac{P(M \mid O)P(O)}{P(M)} \quad (8)$$

where: *P(O)* the *a priori* probability of the object *O*, *P(M)* the *a priori* probability of the local measurement *M*, and *P(M|O)* is the probability of the local measurement *M* when the object *O* is fixated. In the following experiments we only carried out a detection experiment for a single object, there are consequently only two classes, one representing the object and another representing the background. *P(O)* and *P(~O)* are simply set to 0.5 because they do not affect the order of the maxima of *P(O|M)*. Since a single blob is not discriminative enough, we considered the probabilities of observing pairs of blobs instead. To simplify the probability estimation (the number of possible combinations) we have chosen to observe only pairs composed of the central blob (taken as reference) and one surrounding blob as the local measurement *M*:

$$P(M \mid O) = P\left(B_i \mid B_c \text{ and } \left(B_i \text{ adiacent } B_c\right)\right) \quad (9)$$

where B_i is the *i-th* blob that surrounds the central blob B_c that belongs to the object *O*. That is, we exploit the fact the robot is fixating the object and assume the central blob will be constant across fixations. The color of the central blob will be stored and used to bias the visual search (see Section 4.3). The probabilities *P(M|~O)* are estimated during the exploration phase by considering the blobs not adjacent to the central blob. The local measurements are considered independent because they refer to different blobs, so we factorize the total probability $P(M_1,\dots,M_N|O)$ in the product of the probabilities $P(M_i|O)$. An object is considered ‘found’ if the probability $P(O|M_1,\dots,M_N)$ is greater than a fixed threshold. When the object is found after visual search, a figure-ground segmentation is attempted: each blob is selected if it is adjacent to the central recognized blob and if its probability to belong to the object is greater of 0.5.

In practice, we estimate the probability of all blobs adjacent to the central blob to belong to the object. This procedure, although requiring the “active participation” of the robot (through gazing) is faster than estimating all probabilities for all possible pairs of blobs of the fixated object. Estimation of the full joint probabilities would require a larger training set than the one we were able to use in our experiments. Our experimental scenario required the construction of the object model on the fly with the shortest possible exploration procedure, which naturally leads to estimating probabilities with few samples. It is likely that many bins in the histograms, used to estimate probabilities, are empty. To overcome this problem we have used a probability smoothing method. In particular we employed as zero count smoothing the Lidstone’s law of succession:

$$P\left(M \mid O\right)=\frac{count(M \wedge O)+\lambda}{count(O)+v\lambda} \qquad (10)$$

for a v valued problem. With λ=1 and a two valued problem (v=2), we obtain the well-known Laplace’s law of succession. Following the results of Kohavi et al. [26], we choose $\lambda=1/n$ where n is equal to the number of images utilized during the training phase.

A first use of the system is to create a visual model of the hand of the robot (a special object). By relying on this model the robot can distinguish the grasped object from parts of the hand that might still be visible.

6. Results

The behavior of the robot during the learning phases is shown in Figure 5: all the blobs bordering the central one (blue) are used for learning the visual appearance of the object.

Two examples of the saliency map are shown in Figure 4: in 4.4 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 4.1; in 4.5 there is a purely top-down (k_{td}=1, k_{bu}=0) map output after the processing of 4.2. In the latter the robot was instructed to search for the toy airplane. After a saccade on the object and a successfully recognition the figure-ground segmentation is shown in Figure 4.6. The center of mass of the segmented object is used to guide the grasping action of the robot.

We have tested the attention system while guiding the recognition and grasping of objects in the Babybot. In order to qualitatively evaluate the performance, we have done a comparison test of the bottom-up attention using the database of images by Itti et al. [27] (color images with an emergency triangle and relative binary segmentation masks of the triangle), which is freely available on the Internet (http://ilab.usc.edu/imgdbs/). First, the original images and segmentation masks are cropped to a square and transformed to the log-polar format (252x152 pixels) (see Figure 6.1 and Figure 6.2 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 49% of the images a point inside the emergency triangle was selected as the most salient (see an example in Figure 6.3). It is worth noting that a direct comparison with the results of Itti et al., by counting the number of false detection before the target object is found, was not possible since after each saccade the log-polar image should change completely.

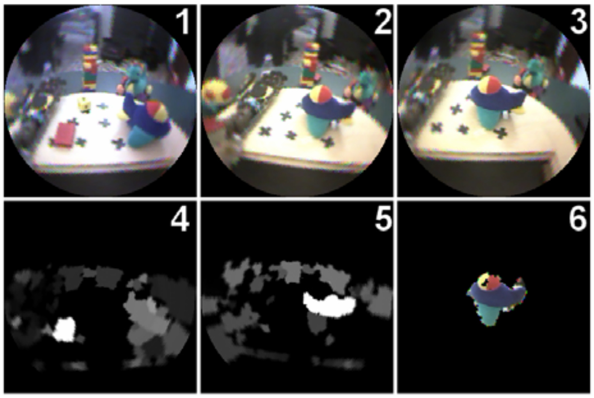

Figure 4. Example saliency maps. In (4) there is the bottom-up saliency map of the image (1). In (5) the top-down saliency map of (2), while searching for the blue toy airplane. Image (6) is the figure-ground segmentation of the image in (3), after having recognized the object.

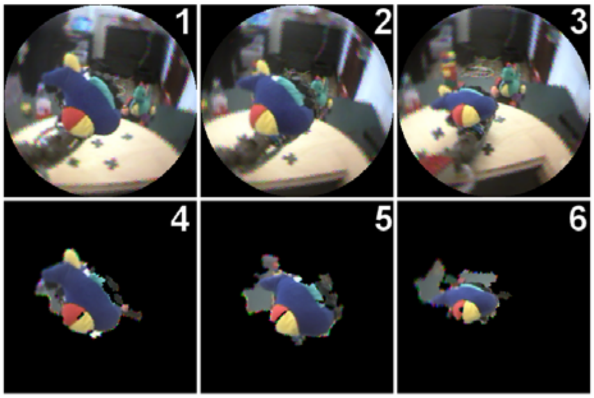

Figure 5. Some example images during exploration phase (1-3) and related segmentations (4-6) used to build the statistical model of the object.

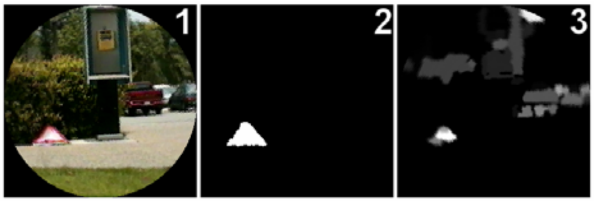

Figure 6. Result on a static example image.

7. Conclusion

We have presented the 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 mechanism of a humanoid robot. This eventually gives rise to a different sort of problems compared to the more typical implementations that only generate scan paths on static images.

The algorithm divides the visual scene in color blobs; each blob is assigned a bottom-up saliency 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 segmented object enter of mass (see Section 4.3 and Section 6), similarly to the behavior observed in humans. 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 [28].

When the camera moves, a new blob will appear in the image center. This active behavior simplifies the segmentation and the recognition task since there will always be a blob in the center that will be segmented from the background.

A similar approach has been taken by Sun and Fisher [9] 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.

On the other hand, our system has been shown in practice to be useful in guiding a humanoid robot in selecting objects to be grasped, by helping the visual search and recognition task.

As a possible extension, the watershed transform could be extended to additional dimensions in feature space (e.g. local orientation) thus providing new ways of both segmenting and recognizing objects.

Acknowledgments

The research described here has been funded by the project MIRROR (IST-2000-28159), ADAPT (IST-2001-371173), and ROBOTCUB (IST-004370).

The authors wish to thank the anonymous reviewers for the fruitful comments and discussion on the earlier version of this manuscript.

References

[1] Z. Pylyshyn, "Visual indexes, preconceptual object, and situated vision," Cognition, vol. 80, pp. 127-158, 2001.

[2] G. Metta and P. Fitzpatrick, "Early Integration of Vision and Manipulation," Adaptive Behavior, vol. 11, pp. 109-128, 2003.

[3] B. J. Scholl, "Objects and attention: the state of the art," Cognition, vol. 80, pp. 1-46, 2001.

[4] S. Yantis, "Control of visual attention," in Attention, H. Pashler, Ed.: Psychology Press, 1998, pp. 223-256.

[5] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 1254-1259, 1998.

[6] G. Sela and M. D. Levine, "Real-Time Attention for Robotic Vision," Real-Time Imaging, vol. 3, pp. 173-194, 1997.

[7] R. Milanese, S. Gil, and T. Pun, "Attentive Mechanisms for Dynamic and Static Scene Analysis," Optical Eng., vol. 34, pp. 2,428–2,434, 1995.

[8] A. M. Treisman and G. Gelade, "A feature integration theory of attention," Cognitive Psychology, vol. 12, pp. 97-136, 1980.

[9] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," Artificial Intelligence, vol. 146, pp. 77-123, 2003.

[10] L. Natale, "Linking action to perception in a humanoid robot: a developmental approach to grasping," PhD Thesis in DIST. Genova: University of Genoa, 2004.

[11] G. Sandini and V. Tagliasco, "An Anthropomorphic Retina-like Structure for Scene Analysis," Computer Vision, Graphics and Image Processing, vol. 14, pp. 365-372, 1980.

[12] J. M. Wolfe and G. Gancarz, "Guided Search 3.0 Basic and Clinical Applications of Vision Science," in Basic and Clinical Applications of Vision Science, V. Lakshminarayanan, Ed. Dordrecht, Netherlands: Kluwer Academic, 1996, pp. 189-192.

[13] S. M. Smirnakis, M. J. Berry, D. K. Warland, W. Bialek, and M. Meister, "Adaptation of retinal processing to image contrast and spatial scale," Nature, vol. 386, pp. 69-73, 1997.

[14] V. A. Billock, "Cortical Simple Cells Can Extract Achromatic Information from the Multiplexed Chromatic and Achromatic Signals in the Parvocellular Pathway," Vision Research, vol. 35, pp. 2359-2369, 1995.

[15] W. von Seelen and H. A. Mallot, "Neural Mapping and Space-Variant Image Processing," presented at IJCNN International Joint Conference on Neural Networks, San Diego, CA, 1990.

[16] X. Li, T. Yuan, N. Yu, and Y. Yuan, "Adaptive color quantization based on perceptive edge protection," Pattern Recognition Letters, vol. 24, pp. 3165-3176, 2003.

[17] R. Eckhorn, R. Bauer, W. Jordan, M. Brosch, M. Kruse, W. Munk, and H. J. Reitboeck., "Coherent oscillations: A mechanism of feature linking in the visual cortex?," Biological Cybernetics, vol. 60, pp. 121-130, 1988.

[18] C. M. Gray, P. König, A. K. Engel, and W. Singer, "Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties," Nature, vol. 338, pp. 334-336, 1989.

[19] P. D. Smet and R. Pires, "Implementation and analysis of an optimized rainfalling watershed algorithm," presented at IS&T/SPIE's 12th Annual Symposium Electronic Imaging 2000, San Jose, California, USA, 2000.

[20] S. Y. Wan and W. E. Higgins, "Symmetric region growing," IEEE Transactions on Image Processing, vol. 12, pp. 1007-1015, 2003.

[21] L. Itti and C. Koch, "Computational modelling of visual attention," Nature Reviews Neuroscience, vol. 2, pp. 194-203, 2001.

[22] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," International Journal of Computer Vision, vol. 57, pp. 137-154, 2004.

[23] R. M. Klein, "Inhibition of return," Trends in Cognitive Sciences, vol. 4, pp. 138-145, 2000.

[24] J. M. Wolfe, "Moving towards solutions to some enduring controversies in visual search," Trends in Cognitive Sciences, vol. 7, pp. 70-76, 2003.

[25] B. Schiele and J. L. Crowley, "Where to look next and what to look for," presented at IROS '96, Osaka, 1996.

[26] R. Kohavi, B. Becker, and D. Sommerfield, "Improving simple Bayes," presented at European Conference on Machine Learning, 1997.

[27] L. Itti and C. Koch, "Feature Combination Strategies for Saliency-Based Visual Attention Systems," Journal of Electronic Imaging, vol. 10, pp. 161-169, 2001.

[28] D. Melcher and E. Kowler, "Shapes, surfaces and saccades," Vision Research, vol. 39, pp. 2929-2946, 1999.