Journal of the Brazilian Computer Society, 2009; 15(3):3-17. ISSN 0104-6500

NLOOK: a computational attention model for robot vision

Milton Roberto Heinen, Paulo Martins Engel*

Informatics Institute, Federal University of Rio Grande do Sul
Av. Bento Gonçalves, 9500, Agronomia, P.O. Box 15064, 91501-970, Porto Alegre, RS, Brazil

Received: June 17, 2009; Accepted: August 27, 2009

Abstract: The computational models of visual attention, originally proposed as cognitive models of human attention, nowadays are being used as front-ends to some robotic vision systems, like automatic object recognition and landmark detection. However, these kinds of applications have different requirements from those originally proposed. More specifically, a robotic vision system must be relatively insensitive to 2D similarity transforms of the image, as in-plane translations, rotations, reflections and scales, and it should also select fixation points in scale as well as position. In this paper a new visual attention model, called NLOOK, is proposed. This model is validated through several experiments, which show that it is less sensitive to 2D similarity transforms than other two well known and publicly available visual attention models: NVT and SAFE. Besides, NLOOK can select more accurate fixations than other attention models, and it can select the scales of fixations, too. Thus, the proposed model is a good tool to be used in robot vision systems.

Keywords*: robot vision, visual attention, selective attention, focus of attention, biomimetic vision.*

1. Introduction

The amount of information coming down the optic nerve in the primate's visual system, estimated to be on the order of 10^8 bits per second, far exceeds what the brain is capable of fully processing and assimilating into conscious experience[35]. The strategy devised by biological systems to deal with this bottleneck is to completely process just some areas of the visual field, called interest regions, and keep the remainder relatively unprocessed[18]. According to Desimone and Duncan[6], the interest regions selection is driven by a competitive attention control mechanism, which facilitates the emergency of a winner among several potential targets. This mechanism allows the visual system to process relevant information to current tasks, while suppressing the irrelevant information that cannot be analyzed simultaneously[16]. The attention mechanism is modulated by two main kinds of cues[23]: bottom-up (visual scene elements that "pop-out")[17] and top-down (information from brain that changes the attention focus)[3]. The visual attention, together with other mechanisms, allows the human being to have a wide vision field and an accurate detail perception without exceeding the capacity to consciously assimilate it[35].

Inspired on biological attention systems, it is possible to develop computational visual attention systems that are able to select interest regions in the visual field to be completely processed, allowing to analyze complex scenes in real time with limited processing resources[30]. Although several visual attention models have been proposed and implemented[11,19,22,29,33,34,39], most of them are intended to understand the biological attention mechanisms, and thus have been validated just by cognitive science measures and by their biological plausibility[8]. However, for a computational visual attention model to be used in a robotic vision system, it is necessary that this model be relatively insensitive to 2D similarity transforms of the image, as in-plane translations, rotations, reflections, and scales. But according to Draper and Lionelle[8], the model called NVT[19], which is the best known visual attention model, is extremely sensitive to these kind of transformations, and thus is inadequate to be used as a front-end in robotic vision systems. Besides, an attention model should select not just the positions, but also the scales of the fixations[8].

This paper presents a new visual attention model, called NLOOK, which was intended to be used in robotic vision systems. This proposed model, which has a very good computational performance, is less sensitive to 2D similarity transforms than other attention models as SAFE[8] and NVT. Moreover, NLOOK can select the fixations in a more accurate way than the other attention models, and it can also select scales of the fixations as well as positions. Thus, NLOOK is a good tool for the registration problem in appearance-based matching systems and in other real time robotic vision tasks like object identification and landmarks detection. This paper is structured as follows: Section 2 describes some related work, in especial two publicly available visual attention models, NVT[19] and SAFE[8]; Section 3 describes the proposed model, called NLOOK; Section 4 describes the accomplished experiments and the obtained results; and Section 5 provides some final conclusions and remarks.

2. Related Work

The first computational model of visual attention was initially proposed in Koch and Ullman[22] and later improved in Itti et al.[19]. It is based on the feature integration theory[38] and on the hypothesis that an explicit two-dimensional topographic map is able to provide an efficient strategy to

*e-mail: mrheinen@inf.ufrgs.br

attention control using just bottom-up cues[16]. In this model, called NVT, the source image is decomposed into three sets of pre-attentive feature maps (sensitive to intensity, colors and orientations), which operate in parallel over the entire visual field. These three sets of feature maps feed into a unique saliency map, which codifies the most important stimulus of the entire visual field[18].

In NVT, the feature maps are computed using linear center-surround operations applied over Gaussian pyramids[2] with levels $\varphi \in [0, 8]$, where $\varphi = 0$ is the original image size. The center-surround operations were implemented in NVT as difference between a fine and a coarse level of the Gaussian pyramid, where the center is a pixel at level $c \in \{2, 3, 4\}$ and the surround is the corresponding pixel at level $s = c + \delta$, with $\delta \in \{3, 4\}$[19]. The resulting feature maps are reduced to level $\varphi = 4$ (the coarsest center level), normalized through a non-linear map normalization operator N(.), which promotes those feature maps with a small number of strong peaks of activity, and combined into a unique feature map.

To generate the intensity map $\bar{I}$ in NVT, the channels r (red), g (green) and b (blue) are extracted from a color source image and the grayscale image $I = (r + g + b)/3$ is used to create a Gaussian pyramid $I(\varphi)$, with $\varphi \in [0, 8]$. This Gaussian pyramid $I(\varphi)$ is used to compute the linear center-surround operations described above, thus generating the intensity map $\bar{I}$. The color maps $\overline{C}$, which codify spatial and chromatic opponency between red and green (RG opponency map) and between blue and yellow (BY opponency map), are created in a similar way. Initially the channels r, g and b are normalized by I to decouple hue from intensity, and four broadly-tuned color channels are created[19]. These channels are used to create Gaussian pyramids $R(\varphi)$, $G(\varphi)$, $B(\varphi)$, $Y(\varphi)$, and over these pyramids the linear center-surround operations are computed. The differences, however, are computed between different color pyramids, i.e., R(c) – G(s) and B(c) – Y(s). The orientation maps $O(\theta)$, where $\theta \in \{0°, 45°, 90°, 135°\}$ is the preferred orientation, are created using Gabor pyramids[12] and combined into a unique orientation map $\overline{O}$.

After their computation, the feature maps $\bar{I}$, $\overline{C}$ and $\overline{O}$ are combined into a unique saliency map $S = (N(\bar{I}) + N(\overline{C}) + N(\overline{O}))/3$. To select the focus of attention (FOA), NVT uses a "winner-take-all" (WTA) neural network with leaky integrate-and-fire neurons[26] and strong local inhibition. Besides, an inhibition of return (IOR)[21] mechanism avoids that the focus of attention be redirected immediately to a previously attended location, thus allowing the selection of the next most salient location[18].

Using this basic architecture, several machine and robot vision applications have been proposed recently. In[31] is described an action learning visual attention model which allow robots to recognize objects, movements and their associations. In Marfil et al.[28], a novel hierarchical framework for object-based visual attention is proposed and used in machine vision tasks. In Perko et al.[36], a probabilistic framework is used to integrate visual context and object detection in machine vision tasks. In Wang et al.[42] a visual brain chip based on selective attention is developed and used in robot vision applications.

Although NVT is probably the best known visual attention model, according to Draper and Lionelle[8] it is not appropriate to be used in robotic vision systems, because its highly sensitivity to 2D similarity transforms. In Draper and Lionelle[8] another visual attention model, called SAFE, is proposed, and this model is less sensitive to 2D similarity transforms than NVT. According to Draper and Lionelle[8], the main differences between SAFE and NVT are: i) SAFE implements center-surround operations using difference of Gaussians (DoG) applied at every level of the Gaussian pyramid; ii) unlike NVT, in SAFE feature maps are not combined across levels within a channel; instead, feature maps are combined across channels within each level, producing a pyramid of saliency maps; iii) SAFE selects scales in addition to spatial coordinates of FOAs; iv) fixations are selected in SAFE without a WTA network nor an IOR mechanism – it simply gets all local maxima above a threshold at every level of the saliency pyramid; v) feature maps are smoothed by a Gaussian kernel with standard deviation of 22.6; and vi) instead of Gabor filters, SAFE uses Ando's edge masks[1] to create orientation maps.

Analyzing its source code and visual results, SAFE has shown the following disadvantages: i) it uses DoG with large Gaussian kernels (the standard deviations are 14.12 and 22.6), which, followed by the Gaussian smoothing, removes fine details of images and increases execution times; ii) the execution times are very high (it is necessary 52 seconds to process an 512 × 512 image in a typical computer[I]; iii) Ando's edge masks make orientation maps very similar to intensity maps, because they detect object borders, not orientations; iv) fixations selected by SAFE are redundant, occurring several overlaps at different scales; v) the SAFE scale selection mechanism does not work appropriately (Section 4 describes the experiments in which this disadvantages are pointed out). Thus, although it has some advantages related to NVT, SAFE has many restrictions, and thus is not adequate to be used in real time robotic vision systems. Other attentional models were also analyzed[11,13,29,34], but their performance were similar to that of NVT in relation to the sensitivity to 2D similarity transforms. Thus, it was decided to propose and implement a new computational model of visual attention, called NLOOK, which is more adequate to be used in robotic vision task. The next section describes this proposed model in details.

3. Proposed model

Figure 1 shows the general architecture of the proposed model, called NLOOK[14,15]. This model is inspired in the NVT architecture[19], but has several improvements that makes it more adequate to be used in robotic vision systems. In the proposed model, input data can be provided either by static images or by a color video stream. At each time instant t, three sets of feature maps are generated for the current source

I Dell Optiplex 755 computer, Intel(R) Core(TM)2 Duo CPU 2.33GHz processor, 1.95GB of RAM and SO Debian Linux 64 bits.

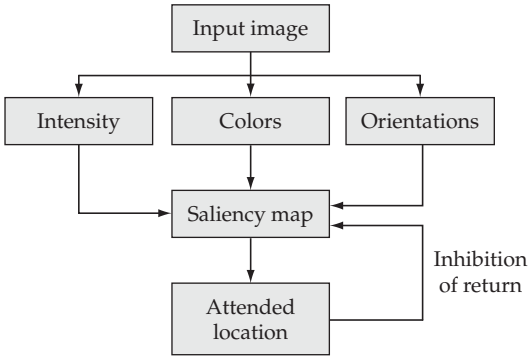


Figure 1. General architecture of the proposed model.

image: intensity (Subsection 3.2), colors (Subsection 3.3) and orientations (Subsection 3.4). Each feature is computed by a set of center-surround operations (Subsection 3.1) applied over scale-spaces[43]. The resulting maps are normalized through the subtraction of their mean and the division by their standard deviation. NLOOK uses this kind of normalization because it is less sensitive to 2D similarity transforms than the operator N(.) used in[19].

After their computation, all feature maps feed into a saliency scale-space (Subsection 3.5), and from this scale space is generated a unique saliency map $\overline{S}$, which topographically encodes the local relevance (conspicuity) of the entire visual field. A winner-take-all (WTA) neural network then selects the most salient point of this map and an inhibition of return (IOR) mechanism avoids that the focus of attention be redirected immediately to a previously attended location, thus allowing the selection of the next most salient location[18]. That whole process can be executed in real time at a rate of 30 images per second. The main differences of NLOOK in relation to other attention models are:

- NLOOK uses DoG filters applied over scale-spaces to compute the linear center-surround operations;
- The normalization procedure is more stable;
- NLOOK is less sensitive to 2D similarity transforms than other attention models;
- It can select scales as well as positions;
- The fixations selected by NLOOK are more accurate than those selected by other attention models;
- NLOOK uses a variable-size IOR mechanism;
- Its computational performance is very good.

The next subsections describe in detail all these improvements and the NLOOK modules.

3.1. Center-surround operations

In the mammalian visual system, many visual neurons are most sensitive in a small region of the visual space, while visual stimuli present in a broader, weaker antagonistic region concentric with the center inhibit the neuronal response[32]. This concept is generally implemented in visual attention models by a set of "center-surround" operations akin to visual receptive fields[19]. In NLOOK, these center-surround operations are implemented using difference of Gaussians (DoG) filters applied over scale-spaces[43] (this procedure is also used by Lowe on SIFT descriptors[27]), and this makes the proposed model less sensitive to 2D similarity transforms than other visual attention models. In fact, according to Draper and Lionelle[8], the best known visual attention model, called NVT[19], is highly sensitive to 2D similarity transforms because it does not use DoG filters – it simple implements the center-surround operations as differences between fine and coarse levels of a Gaussian pyramid[19].

More specifically, in NLOOK the center-surround operations are computed in the following way: initially the input image is sub-sampled into several octaves, and the initial image of each octave is the respective level of a Gaussian pyramid. Several scales are created for each octave through the successive convolution of the initial image with Gaussian kernels. Finally, differences of Gaussians (DoG) are generated through the absolute difference of adjacent scale levels. Figure 2, adapted from[27], illustrates this process.

The proposed model uses all possible octaves, i.e., it keeps generating octaves until the image is smaller than the lowest Gaussian kernel, which corresponds to five octaves for an 320 × 240 image and six octaves for an 416 × 416 image. As recommended by Lowe[27], NLOOK generates three scales per octave with standard deviations of 1.2263, 1.5450 and 1.9466. Figure 3 shows an example of scale-space before the differences of Gaussians be computed (the source image is shown later in Figure 8b). This scale-space has five octaves and three scales per octave.

3.2. Intensity maps

The first set of feature maps is concerned with intensity contrast, which in mammals is detected by neurons sensitive either to dark centers on bright surrounds, or bright center on dark surrounds[24]. To generate the intensity maps I (O, S) on NLOOK, where O is the number of octaves and S is the number of scales, the color input image is converted into a grayscale image I, and the DOGs are created from I using scale-spaces as described above. Thus, for a typical 416 × 416 image, twelve DoG images (two per octave) are created. Each DoG image is normalized through the subtraction of its mean and the division by its standard deviation. Like SAFE, in NLOOK different octaves and scales are not combined

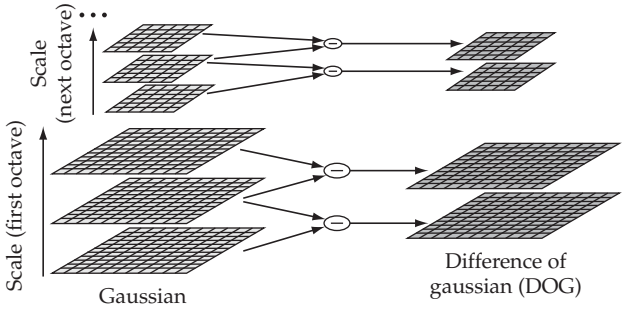


Figure 2. DoG construction using scale-spaces.

within a channel, i.e., all octaves and scales are maintained. Figure 4 shows an example of intensity scale space I (O, S), where O represents the octave and S represents the scale of each image. The bright points in Figure 4 represent the most conspicuous areas in terms of intensity contrast.

3.3. Color maps

According to Engel et al.[10], in the mammalian visual cortex, colors are represented using a so-called "color double-opponent" system: In the center of the receptive field, neurons are excited by one color and inhibited by another, while the converse is true in the surround. Such spatial and chromatic opponency exists for the red/green, green/red, blue/yellow and yellow/blue color pairs in human primary visual cortex[10,20]. In the proposed model, two scale-spaces, RG(O,S) and BY(O,S), are created to account for red/green and blue/yellow color double opponency. To create these color opponency scale-spaces, the r, g and b color channels of the source image are normalized by $I = (r + g + b)/3$ in order to decouple hue from intensity. Four broadly-tuned color channels are created for the red, green, blue and yellow colors[19]:

$$\begin{aligned} R &= r_n - (g_n + b_n)/2 \\ G &= g_n - (r_n + b_n)/2 \\ B &= b_n - (r_n + g_n)/2 \\ Y &= (r_n + g_n)/2 - |r_n - g_n|/2 - b_n \end{aligned} \quad (1)$$

Figure 3. Example of a scale-space before the DoG computation.

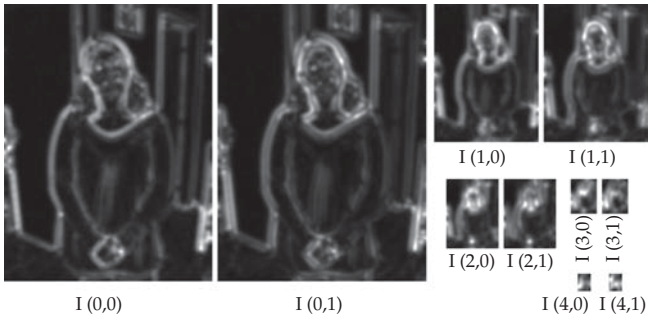


Figure 4. Example of an Intensity scale-space.

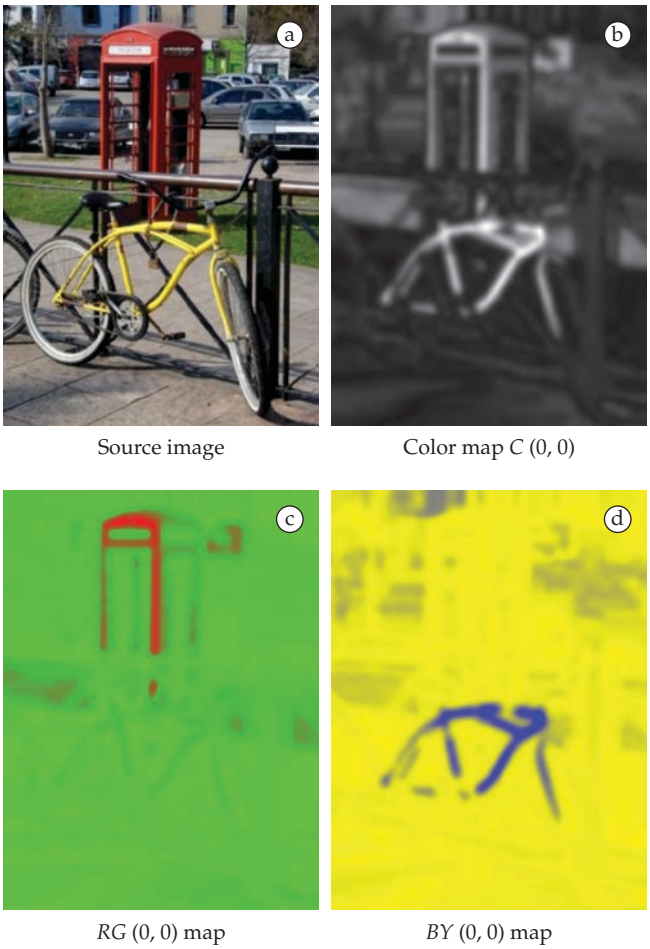
where negative values are set to zero. These four color channels are used to create the scale-spaces R (O, S), G (O, S), B(O, S) and Y(O, S). Therefore, the absolute differences are computed between different channels, i.e., for an octave o the DoGs are computed through: $|R(o, 0) - G(o,1)|$ and $|R(o, 1) - G(o, 2)|$ for RG scale-space; and $|B(o, 0) - Y(o, 1)|$ and $|B(o, 1) - Y(o, 2)|$ for BY scale-space. These scale-spaces are then normalized and combined into a color scale-space C (O,S) by:

$$C(o,s) = \tfrac{1}{2}\left(RG(o,s) + BY(o,s)\right) \quad (2)$$

Figure 5 shows some examples of color maps. Figure 5a shows the source image; Figure 5c shows the first octave and scale of the correspondent red/green scale-space (map RG (0, 0)); Figure 5d shows the first octave and scale of the blue/yellow map RG (0, 0); and Figure 5b shows the color map C (0, 0).

In Figure 5c, the red areas represent more conspicuous regions, and in the Figure 5d the blue areas represent more conspicuous regions[II]. In Figure 5b, the bright points represent the most conspicuous areas.

II In grayscale versions of this paper, the dark points in Figures 5c and 5d represent the most conspicuous areas.



Source image — Color map *C* (0, 0)

RG (0, 0) map — *BY* (0, 0) map

Figure 5. Examples of color maps.

3.4. Orientation maps

The mammalian visual cortex has neurons which are sensitive to spatial orientation, and according to[5], the receptive field sensitivity profile of these neurons is approximated by Gabor filters, which are the product of a cosine grating and a 2D Gaussian envelope[32]. In NLOOK, orientation maps are created in a similar way as the intensity maps, but the grayscale scale-space generated from I is convolved with a Gabor filter before the DoG computation. Thus, for each pixel (x, y) of the intensity maps I (o, s) (Subsection 3.2) the $\psi_\theta(x, y)$ values are calculated:

$$\psi_\theta(x,y) = \frac{1}{2\pi} e^{(x^2+y^2)/2} \times \Theta_\theta(x,y) \quad (3)$$

where $\Theta_\theta(x, y)$, with $\theta \in \{0°, 45°, 90°, 135°\}$, are complex sinusoids obtained by the equations:

$$\Theta_{0^\circ}(x,y) = e^{i(\pi/2)x}; \quad \Theta_{45^\circ}(x,y) = e^{i(\pi\sqrt{2}/4)(x+y)};$$
$$\Theta_{90^\circ}(x,y) = e^{i(\pi/2)y}; \quad \Theta_{135^\circ}(x,y) = e^{i(\pi\sqrt{2}/4)(y-x)}. \quad (4)$$

These four orientation scale-spaces are normalized and combined into a unique scale-space of orientations O (O, S) by the equation:

$$O(o,s) = \frac{1}{4}\left(\sum_{\theta\in\{0^\circ,45^\circ,90^\circ,135^\circ\}} N(O(o,s,\theta))\right) \quad (5)$$

Figure 6 shows examples of orientation maps (the bright points represent the most conspicuous areas). Figure 6a shows the source image. Figures 6c, 6b, 6e and 6f show, respectively, the orientation maps of the first octave and scale (O (0, 0, θ) maps) with preferred orientations of 0, 45, 90 and 135. Figure 6b shows the orientation the first octave and scale of the orientation scale-space O (O, S).

3.5. Saliency map

After generating all feature scale-spaces, NLOOK combines them into a unique saliency scale-space S (O, S) through the normalization and point-by-point addition of the corresponding octaves and scales. More specifically, for each octave o and scale s the saliency map S (O, S) is computed by:

$$S(o,s) = \tfrac{1}{3}\left(I(o,s) + C(o,s) + O(o,s)\right) \quad (6)$$

Figure 7 shows a saliency scale-space S (O, S) computed from the source image shown in Figure 8b.

After its computation, the saliency scale-space S (O, S) is combined into a unique saliency map $\overline{S}$ by an expansion of every octave/scale to the original size (level 0), normalization and point-by-point addition. Unlike NVT, which creates the saliency map at the coarsest level, NLOOK creates the unique saliency map $\overline{S}$ at the finest level to avoid information losses.

The unique saliency map $\overline{S}$, that does not exist in SAFE, acts like a "summary" of S (O, S), and also allows for the use of a unique IOR mechanism. The existence of a unique saliency map, also called "master map"[37], has been suggested by[22], in which a topographically organized map encodes information on where salient (conspicuous) objects are located in the visual field[32]. Figure 8a shows the unique saliency map $\overline{S}$ computed from saliency scale-space shown in Figure 7, and Figure 8b shows the corresponding source image.

3.6. Scale selection

Apart from computing the positions of the most interesting image locations, NLOOK is able to find out the approximate dimensions of this locations, also called the characteristic scale. According to Lindeberg[25], the characteristic scale of a pixel within an image can be determined by locating the extrema of the Laplacian jet of this pixel. The Laplacian jet is

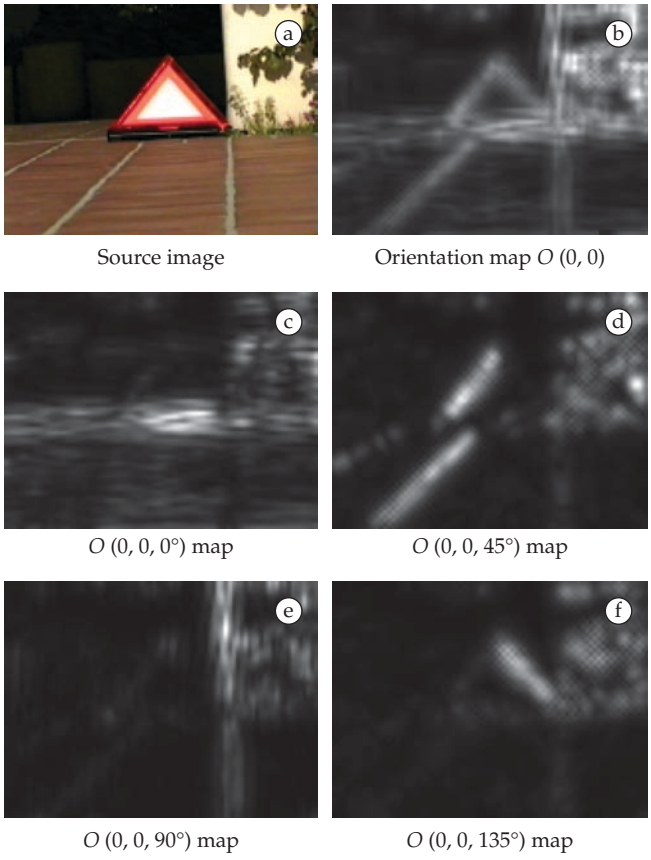


Figure 6. Examples of orientation maps.

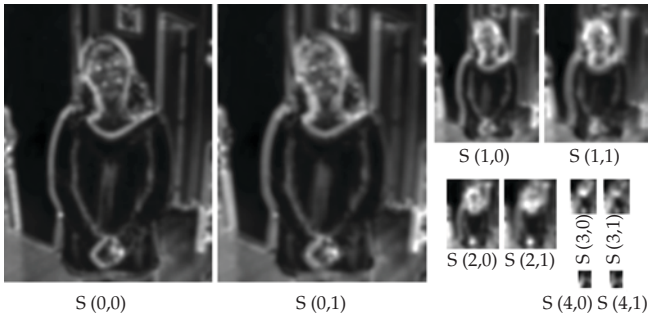
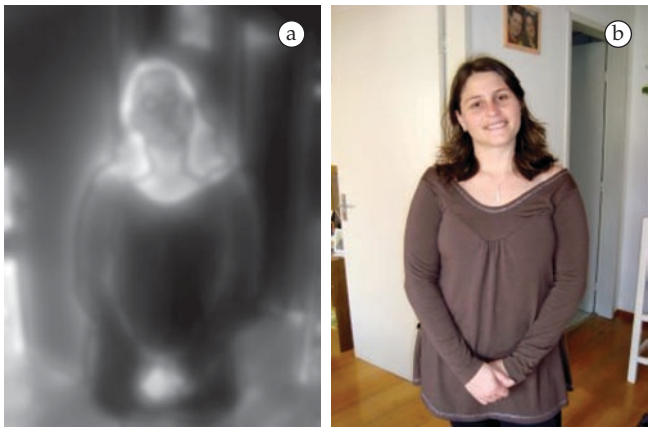


Figure 7. Example of a saliency scale-space.



Saliency map Source image

Figure 8. Example of a unique saliency map $\overline{S}$.

a function across the levels of a difference of Gaussian scale-space at the coordinates of a given pixel, and the outcomes of the Laplacian will be the highest at the scale in which the contrast between close neighboring pixels is maximal. This corresponds, by definition, to the characteristic scale of that location[41].

The procedure described above can be applied over a saliency scale-space to find out the approximate dimensions of the most interesting regions (FOAs) in the visual field. Once the location of an interest region is found, the Laplacian jet profile at this location is analyzed to find out a maximum or minimum extremum. According to[40], a more precise location in scale is determined by interpolation using a second-order Taylor expansion:

$$\hat{s} = -\frac{L_s}{L_{ss}} = \frac{L(s-1) - L(s+1)}{L(s+1) - 2L(s) + L(s-1)} \quad (7)$$

where s is the scale of the octave in which the extremum was found, and L_s and L_{ss} are the first and second partial derivatives of the Laplacian function L relative to the level s, respectively. The offset $\hat{s}$ is added to the extrema in order to determine more accurately the characteristic scale. According to[4], the radius rroi of the region of interest can be computed from the interpolated octave by:

$$r_{roi} = 2^{(o-1)} \times k_s \times b^{(s+\hat{s})} \quad (8)$$

where o and s are the octave and the scale of the maxima, the constant k_s = 1.6 is an empirical correction factor for the scale, which is given by a geometric progression with base b = √2.[41] Thus, Equation 8 is used in NLOOK to find out the dimensions of the FOA and the inhibition of return radius, as described in the next section.

3.7. Inhibition of return

The unique saliency map $\overline{S}$, described above, defines the most salient image location at any given time to which the focus of attention should be directed. The $\overline{S}$ map feeds into a biologically-plausible "winner-take-all" (WTA) neural network, in which synaptic interactions among units ensure that only the most active locations remain, while all other locations are suppressed.

After selection of the most active location, the saliency scale-space S (O, S) is analyzed to find out the octave/scale most important, i.e., the octave/scale with the most effective contribution to the saliency at this point. If a draw occurs, the finer octave/scale is selected. An inhibition of return (IOR) mechanism[21] is then applied over the unique saliency map $\overline{S}$ in order to avoid that the focus of attention be redirected immediately to a previously attended location. In NLOOK, the diameter of IOR varies according to the most important octave/scale selected, that is, the standard deviation σ of the inverted Gaussian kernel is given by Equation 8, described above. Thus, Equation 8 allows large IOR sizes at coarser levels and small IOR sizes at finer levels.

Figure 9 exemplifies the IOR mechanism in the saliency map shown previously in Figure 8, and Figure 10 shows the corresponding selected FOAs. Gray arrows were added in Figure 9 to facilitate the visualization of the inhibited regions.

3.8. Hardware and software

The prototype of the proposed model was implemented in C ANSI programming language and it uses the Open Source Computer Vision Library[III] (OpenCV), that is a free software

III OpenCV – http://www.intel.com/technology/computing/opencv/

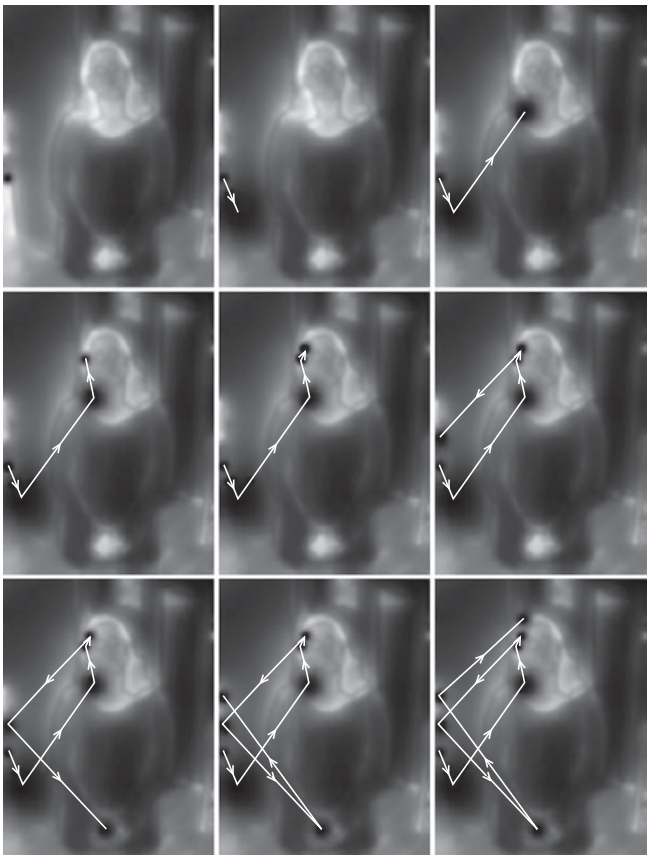


Figure 9. Inhibition of return mechanism

Figure 10. Selected FOAs.

Figure 11. Pioneer 3-DX robot.

library for C and C++ programming languages. This library implements, in an efficient and parallel way, several routines for image processing and machine vision. Beside this, the implemented prototype uses POSIX threads to allow the parallel execution in multi-processed platforms.

The computer used in the experiments was a Dell Optiplex 755 equipped with an Intel(R) Core(TM)2 Duo CPU 2.33GHz processor, 64 bits architecture, 1.95GB of RAM memory, GPU Intel and operating system Debian Linux 4 of 64 bits. The CCD device used was a Creative WebCam NX Ultra, that is able to provide 320 × 240 color images. Figure 11 shows the Pioneer 3-DX robot, the robotic platform to be used with the proposed model in tasks like object recognition, landmarks detection, obstacle avoidance, localization and mapping.

4. Experiments and Results

In this section, three sets of experiments have been accomplished to validate the proposed model, and also to compare it with two other publicly available attention models: NVT[19] and SAFE[8]. The first set of experiments, described in Subsection 4.1, aim to assess the sensitivity of the attention models to 2D similarity transforms. The second set of experiments, described in Subsection 4.2, are devised to verify the accuracy of all attention models, and also the scale selection mechanism, using synthetic images. The third set of experiments, described in Subsection 4.3, compares the attention models using the same criteria of the second set of experiments, but using natural color images instead. Although NLOOK is able to analyze color video streams in real time at a rate of 30 frames per second, in all experiments just static images were used, because this allows to compare the proposed model with the other attention models in a more accurate way (SAFE, by instance, works just with static images).

4.1. Similarity transforms experiments

This subsection describes experiments to assess the sensitivity of NLOOK to 2D similarity transforms, and also to compare it with SAFE and NVT. Thus, fifteen 320 × 240 images, shown in Figure 12, were selected (some of these images were originally used in[18]), and over them the following transformations were applied:

- Vertical and horizontal reflection;
- Rotations from 45 to 315 in 45 intervals;
- Vertical and horizontal translation of 1, 4, 9, 16 and 27 pixels;
- Rescaling by 1.4, 1.2, 1.1, 0.9, 0.8 and 0.6 factors.

This sums 25 distinct transformations per image. An attention model insensitive to 2D similarity transforms will find the same fixations in the original and in the transformed images -- these fixations will be just displaced according to the applied transformation. To avoid information losses, the original images were added with gray borders before the transformations, and the edges among the source image and borders were blurred to avoid an increment of saliency at these edges. The size of gray-bordered images was 416 × 416 pixels, and thus six octaves were used in NLOOK.

To quantify the performance of the three models, two measures (adapted from[8]) were used: gross error rate (GE) and mean drift (MD). GE counts the rate of fixations in the original image that are not within a threshold radius of any fixation in the transformed image, once the geometric transformation is compensated for. In the experiments, the radius

Figure 12. Images used to test all attention models.

Table 1. Mean drift.

Transformation	NVT	SAFE	NLOOK
Reflect V	13.4288	5.6574	0.5766
Reflect H	16.0495	1.8356	0.3387
Rotate 45	22.2108	4.5208	3.5377
Rotate 90	14.4942	4.9963	1.9867
Rotate 135	21.6493	6.6386	4.6888
Rotate 180	21.4476	5.9266	2.8976
Rotate 225	22.9832	6.4488	4.4418
Rotate 270	16.3473	4.6999	1.9933
Rotate 315	23.4663	4.4876	3.3868
Translate V 1	2.7220	0.7161	0.5014
Translate V 4	9.4484	0.4649	0.5998
Translate V 9	12.6907	0.8594	1.6793
Translate V 16	7.2120	2.0024	0.0000
Translate V 27	16.2587	2.3876	2.0367
Translate H 1	2.9999	0.6460	0.7285
Translate H 4	8.4032	0.5569	0.8655
Translate H 9	13.4400	0.9468	1.4291
Translate H 16	6.3246	1.5716	0.1344
Translate H 27	16.3116	2.6540	1.4920
Scale 1.4	28.7113	18.9905	4.7260
Scale 1.2	22.9721	16.6508	3.5227
Scale 1.1	16.9717	13.0330	2.6105
Scale 0.9	16.9837	10.0380	3.0198
Scale 0.8	20.1752	12.6799	3.7399
Scale 0.6	28.9947	21.8215	8.4385
Mean	16.1079	6.0492	2.3749
Standard dev.	7.2060	6.1126	1.9416

threshold was 17 pixels, what matches the IOR radius used by NVT (this value was also used in Draper and Lionelle[8]). MD measures the average distance in pixels from desired and actual FOA positions. The desired positions are the FOA positions at the original image compensated by the applied geometric transformation, and the actual positions are the FOA positions in the transformed image. Thus, the mean drift values are computed by:

$$MD = \frac{1}{N}\sum_{n=1}^{N} D(Fd(n), Fo(n)) \quad (9)$$

where N is the number of extracted FOAs per image (10 in the experiments), Fd(n) is the desired position of FOA n, Fo(n) is the actual position of FOA n and D(.) is the Euclidean distance between Fd(n) and Fo(n). Thus, NVT, SAFE and NLOOK were tested with the images of Figure 12 and their respective transformed versions, and for each image version these two performance measures were computed. The results obtained in these experiments are shown in Tables 1 and 2.

Table 1 shows the obtained mean drift values. The first column describes the applied transformation. The following columns show the mean drift values, averaged over all images, for NVT, SAFE and NLOOK. The last two rows show the mean and the standard deviation computed over all experiments (15 images × 25 transforms = 375 experiments). Table 2 shows the gross error rate values obtained in these experiments, averaged over all images.

In[8] similar experiments evaluating NVT and SAFE are described, but in these previous experiments just four images

Table 2. Gross error rate.

Transformation	NVT (%)	SAFE (%)	NLOOK (%)
Reflect V	22.00	4.67	0.00
Reflect H	26.00	0.67	0.00
Rotate 45	26.67	6.00	5.33
Rotate 90	23.33	3.33	0.00
Rotate 135	32.67	6.00	5.33
Rotate 180	34.00	3.33	0.00
Rotate 225	35.33	7.33	4.67
Rotate 270	27.33	2.67	0.00
Rotate 315	29.33	4.67	4.67
Translate V 1	3.33	0.00	0.00
Translate V 4	7.33	0.00	1.33
Translate V 9	6.67	0.00	3.33
Translate V 16	8.00	2.67	0.00
Translate V 27	22.00	3.33	2.67
Translate H 1	4.00	0.00	0.67
Translate H 4	10.00	0.00	1.33
Translate H 9	9.33	0.67	1.33
Translate H 16	10.67	1.33	0.00
Translate H 27	20.00	4.00	2.00
Scale 1.4	41.33	30.00	7.33
Scale 1.2	29.33	23.33	6.67
Scale 1.1	18.67	15.33	2.00
Scale 0.9	14.67	16.00	3.33
Scale 0.8	26.00	16.67	4.00
Scale 0.6	45.33	42.00	14.00
Mean	21.33	7.76	2.80
Standard dev.	11.82	10.62	3.28

have been used, two fractal and two natural images, but one of the natural images was very dark. Figures 13a and 13b show, respectively, the boxplot graphs of mean drift (MD) and gross error rate (GE) computed over all experiments.

Similar to Draper and Lionelle[8], the results (Table 1) show that NVT is very sensitive to 2D similarity transforms, and although SAFE results were better than NVT results, its performance was not so good as the performance obtained in Draper and Lionelle[8]. On the other hand, NLOOK performance was the best in almost all transformations, and the differences are statistically significant.

Figure 14 shows the visual output of all attention models using an original image and its respective 45 rotated version (a small part of the borders was removed to improve visualization). It is noted that the FOAs selected by NLOOK are the same in both image versions, showing that it is insensitive to 45 rotation of this image. SAFE has a similar performance (the FOAs are the same in both image versions), but some fixations are overlapped, i.e., SAFE has chosen the same regions using several scales. On the other hand, NVT was sensitive to 45 rotation – it chose different fixations at each image version.

As mentioned above, the FOAs selected by SAFE are usually redundant, i.e., it selects the same objects in the image several times. On the other hand, the FOAs selected by NLOOK are less redundant due to the variable-size IOR mechanism, smaller Gaussian kernels and the high-detailed saliency map $\bar{S}$, generated at the finest scale. To quantify the redundancy of the analyzed models the following measure at original images (without transformations) was used: for each selected FOA the Euclidean distance to its nearest neighbor was computed, and then the mean of these distances was computed. A model which generates few overlaps will have larger distances than a model which generates several overlaps. The computed distances are:

- NVT: Mean = 45.92; Standard dev. = 27.95;
- SAFE: Mean = 30.12; Standard dev. = 26.17;
- NLOOK: Mean = 49.15; Standard dev. = 15.59.

These results show that SAFE fixations are really more redundant than NLOOK fixations, i.e., the FOAs selected by SAFE are more overlapped than the FOAs selected by NLOOK. In relation to NVT, the mean distances were similar to those obtained with NLOOK.

Although this redudancy criterion is useful to measure the overlapping degree of fixations, it is not useful to measure the quality of the selected FOAs. The next subsection describes some experiments devised to compare the capacity of each attention model to select the most salient objects (random circles) in synthetic images.

4.2. Experiments using synthetic images

According to Draper et al.[7] and Draper and Lionelle[8], it is very difficult to evaluate attention systems, especially when

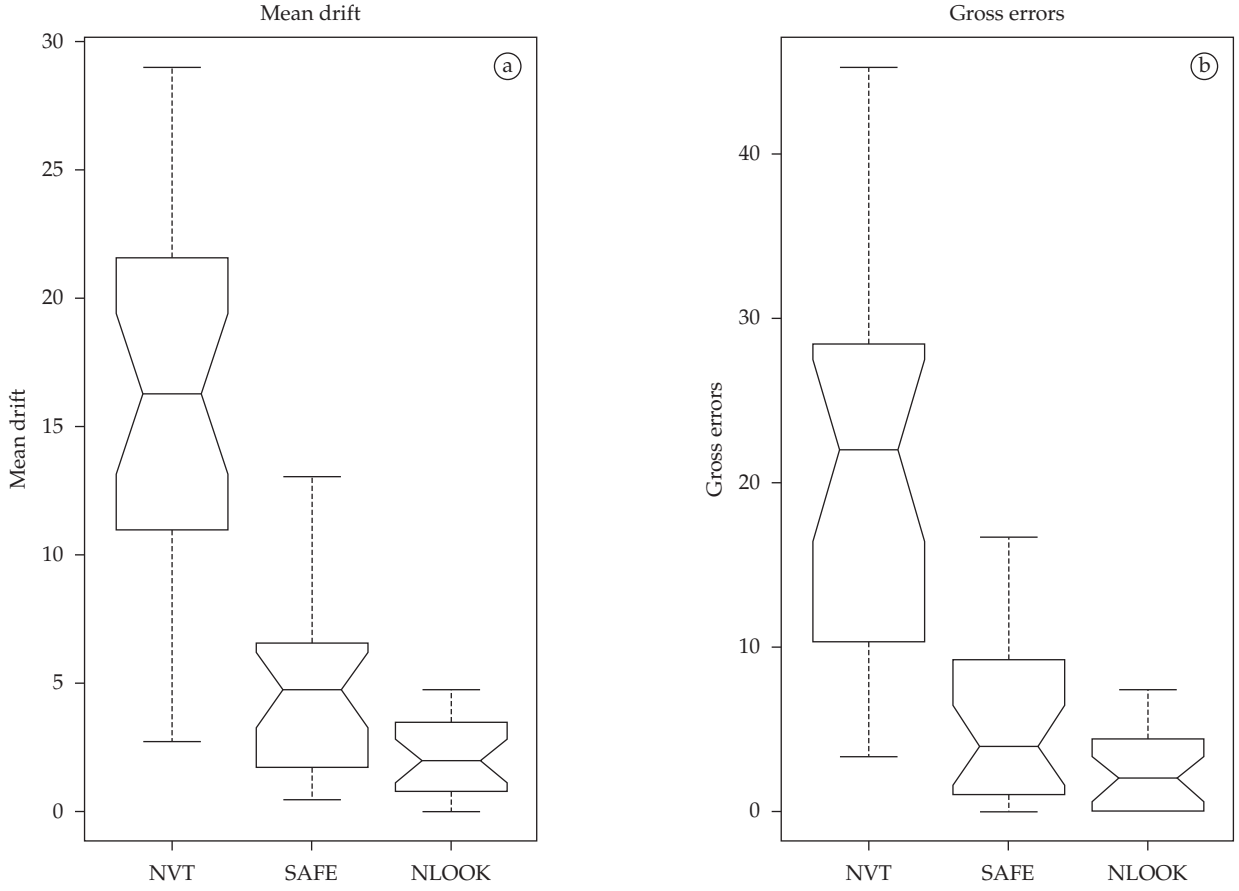


Figure 13. Boxplot of mean drift and gross errors.

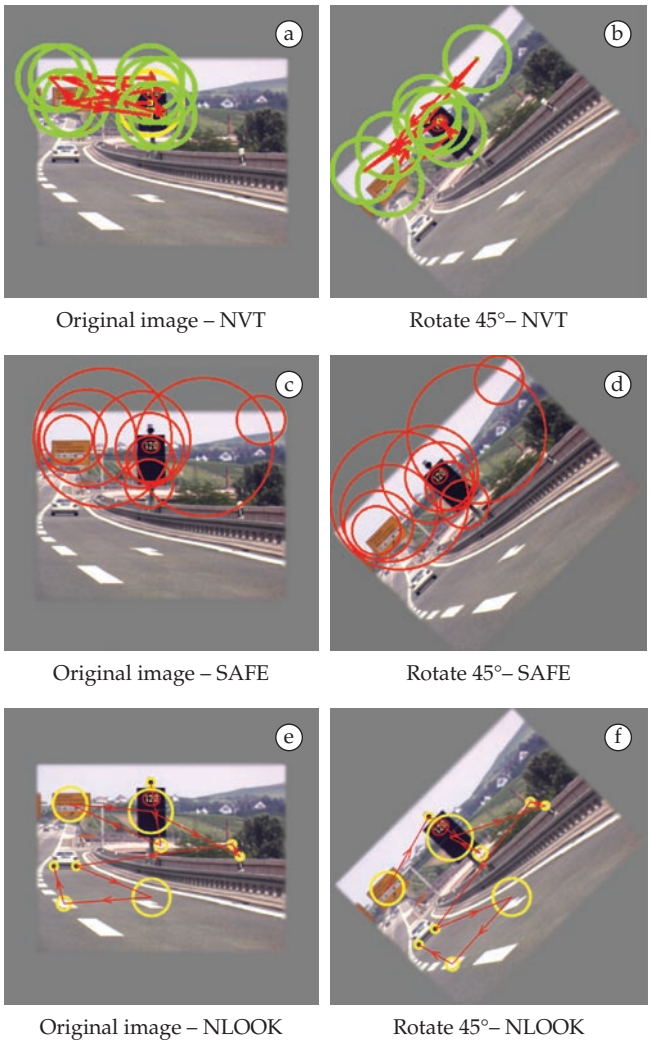


Figure 14. Visual outputs using original and 45°rotated images

they are used as a front-end in robotic vision systems. Eye tracking systems, for example, are able to record eye movements and thus measure overt attention[18], but they cannot measure covert attentional fixations or selected scales. In this section, instead of comparing the attention models with data recorded by eye tracking systems, the following methodology was used: to generate synthetic images in which the interesting points are known a priori (filled circles with random colors and sizes), and to compare the selected FOAs with the positions and scales of the random circles. The main advantage of this methodology is the controllability of the experiments, what makes possible to compare the accuracy of the different attention models using images in which the salient regions are evident.

Thus, to validate the attention models 100 synthetic color images were randomly generated. Each synthetic image has 10 non-overlapped filled circles in different positions and with different sizes and colors. The color of each circle, and also the background color, was generated using random values in the [0,255] interval for the R, G and *B* color channels. Figure 15 shows some of the generated images. To quantify the accuracy of all attention models (NVT, SAFE and NLOOK) in selecting the generated circles, two measures were used:

- Position drift: distance in pixels between each circle and the nearest FOA;
- Radius difference: difference in pixels between the radius of each circle (actual radius) and the radius of the nearest FOA (obtained radius).

Each attention model was configured to generate 15 fixations. Table 3 shows the results obtained in these experiments. The first column describes the attention model. The following columns show the mean (μ) and the standard deviation (σ) of the position drift and radius difference measures.

Figure 16 shows the boxplot graphs of these experiments. Figure 16a shows the boxplot graph of the position drift, and Figure 16b shows the boxplot graph of the radius difference measure. These results show that the proposed model performs better than NVT and SAFE according to both measures. In fact, the FOA locations selected by NLOOK are nearer to the desired locations (center of the drawn circles), what means that NLOOK is able to locate the center of the circles in a more accurate way than the other two attention models. Besides, the scales selected by NLOOK are very good (the difference is just two pixels in average), which shows that NLOOK is able to select the FOA dimensions in a very effec-

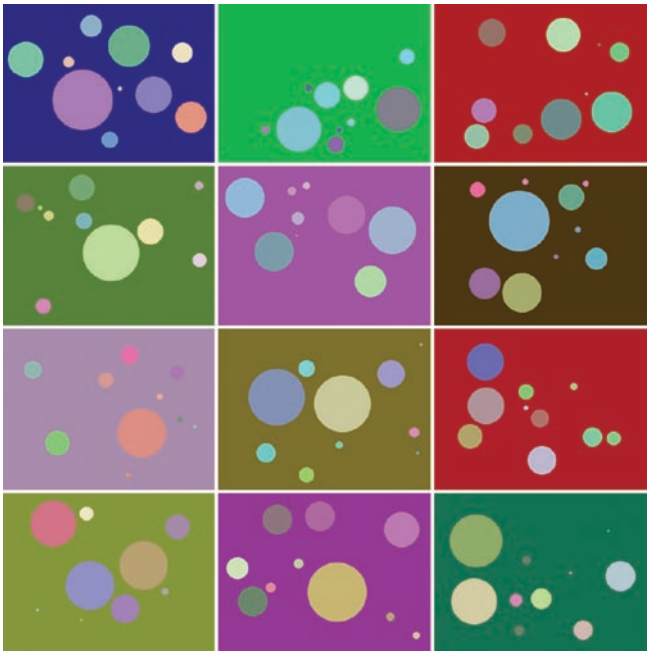


Figure 15. Example of the generated synthetic images.

Table 3. Position drift and radius difference using synthetic images.

	Position drift		Radius diff.	
Model	μ	σ	μ	σ
NVT	32.0517	10.8693	9.9410	1.7929
SAFE	31.1990	10.9083	22.7104	7.0741
NLOOK	6.8291	6.8666	2.6750	2.1480

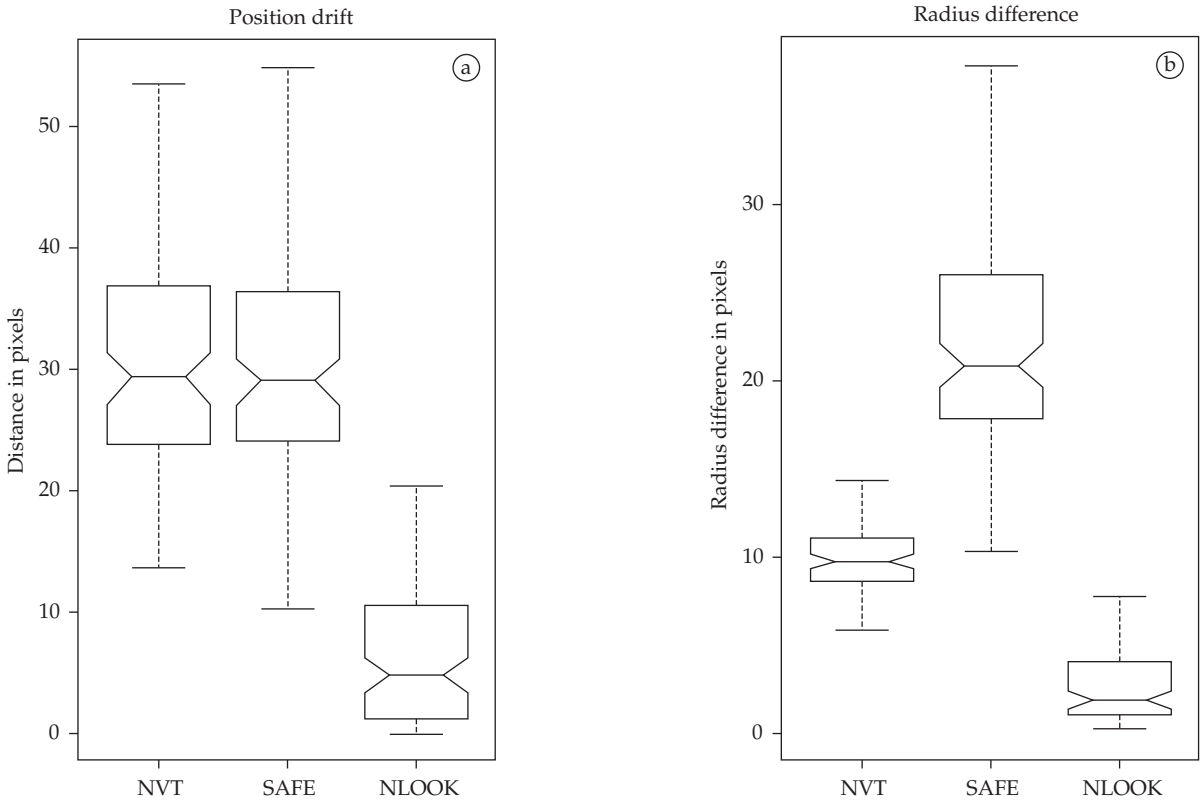


Figure 16. Boxplot graphs of the synthetic images experiments.

tive way. On the other hand, the radius differences obtained by SAFE are even worse than NVT radius differences, which is a very impressive, because NVT uses always the same radius size for all FOAs (17 pixels).

Figure 17 shows the FOAs selected by the three attention models in two generated images. It can be noticed that NVT selects the same image regions several times, not exploring other regions with lower conspicuity. SAFE was able to select the most conspicuous image regions, too, but it selects the same regions several times with different scales. On the other hand, NLOOK was able to select all circles in Figure 17 images, and the selected scales are very close to the actual circle dimensions. Therefore, NLOOK had also selected some image regions without circles (regions among other circles), but this is not a problem, because for a visual attention system these regions may be very interesting, too.

Although these results are very impressive, for an accurate evaluation of an attention model is necessary to evaluate it using natural color images, which is done in the experiments described in the next subsection.

4.3. Experiments using natural images

This subsection describes experiments similar to those described in the earlier subsection, but using natural color images instead. Thus, some images, shown in Figure 18, were selected to be used in these experiments[IV]. Each selected

IV The red arrows in Figure 18 were drawn just to show the target locations. The images used in the experiments do not have these arrows.

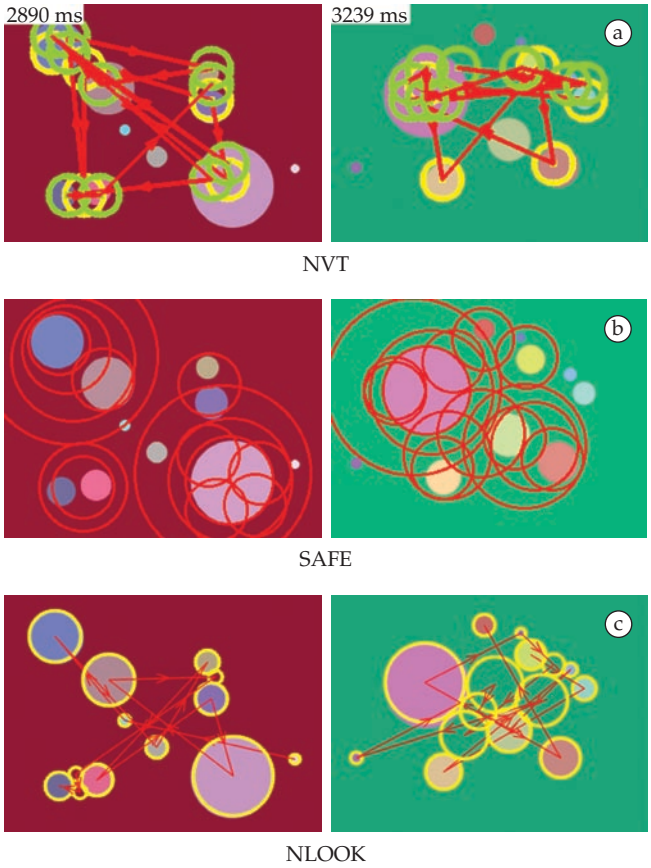


Figure 17. Visual outputs of all models using synthetic images.

Figure 18. Natural color images used in the experiments.

image contains one or more elements that ``pop-out'', that is, are very salient for the human being. Some images have traffic signs, which are specially designed to be salient -- they must catch the driver's attention. Some images have flowers, which were designed by nature to be salient – they need to attract insects like bees to reproduce. Other images, like the yellow slipper among green weeds, have objects which are also very salient.

In order to compare the performance of the three attention models, the most salient locations of each image, marked in Figure 18 with red arrows, were manually measured, and their positions and scales were used to compute the position drift and radius difference measures (described previously in Subsection 4.2). Table 4 shows the results obtained in this experiments, and Figure 19 shows the corresponding boxplot graphs.

It can be noticed in Table 4 and Figure 19 that the obtained results with natural color images are similar to those obtained with synthetic images, that is, NLOOK performs better than

Table 4. Position drift and radius difference using natural images.

	Position drift		Radius diff.	
Model	μ	σ	μ	σ
NVT	16.9232	11.4335	11.2460	10.4780
SAFE	18.8858	14.9485	19.8754	13.0765
NLOOK	2.2444	1.2886	2.1373	1.2855

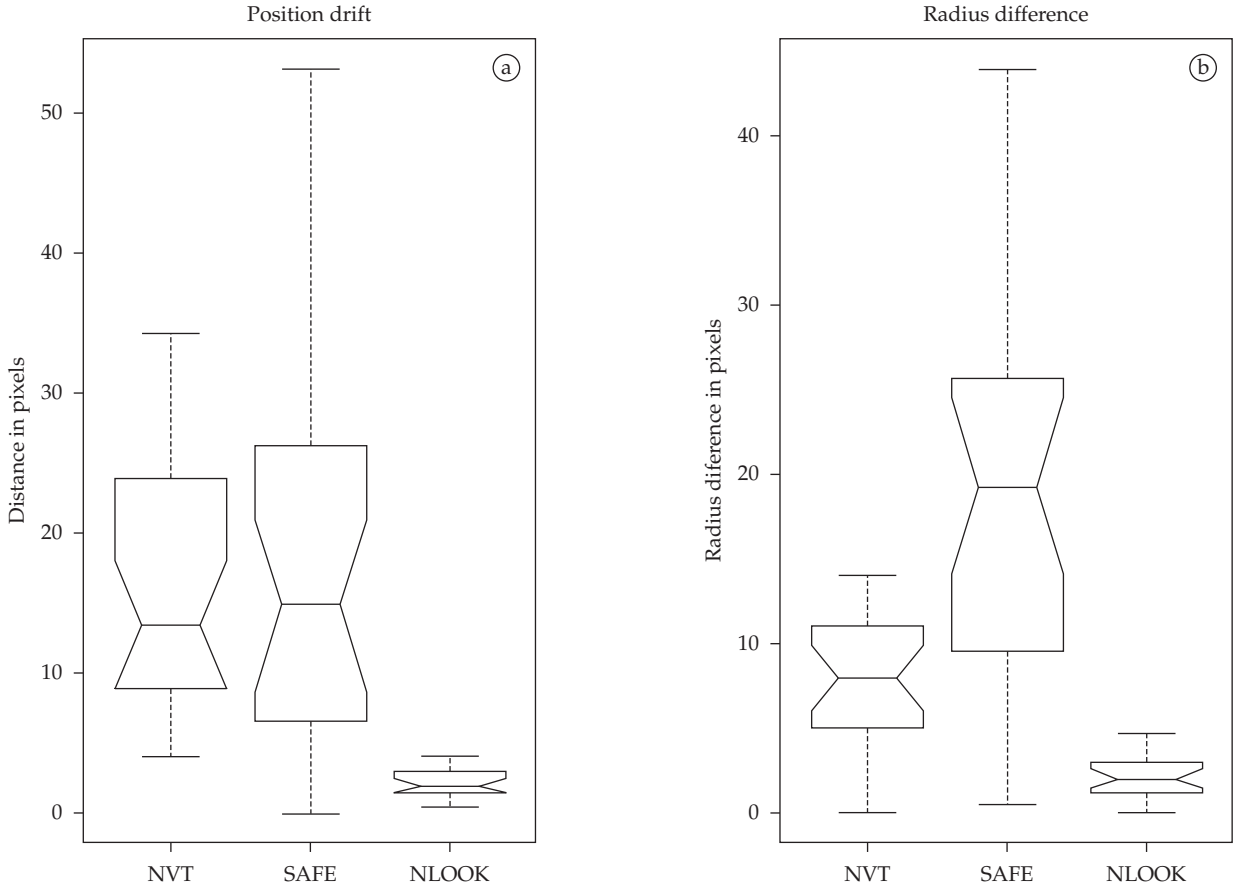


Figure 19. Boxplot graphs of the experiments using natural images.

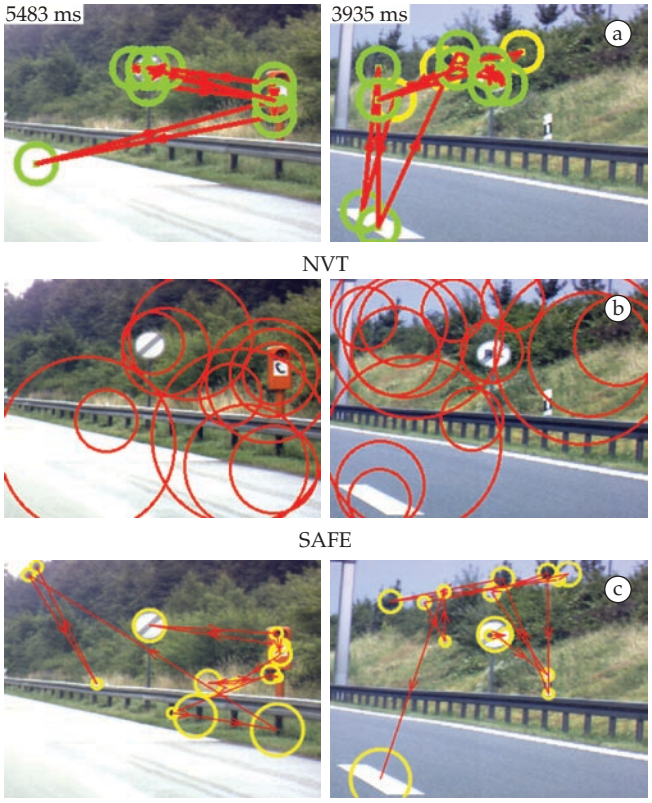


Figure 20. Visual outputs of all models using images with traffic signs.

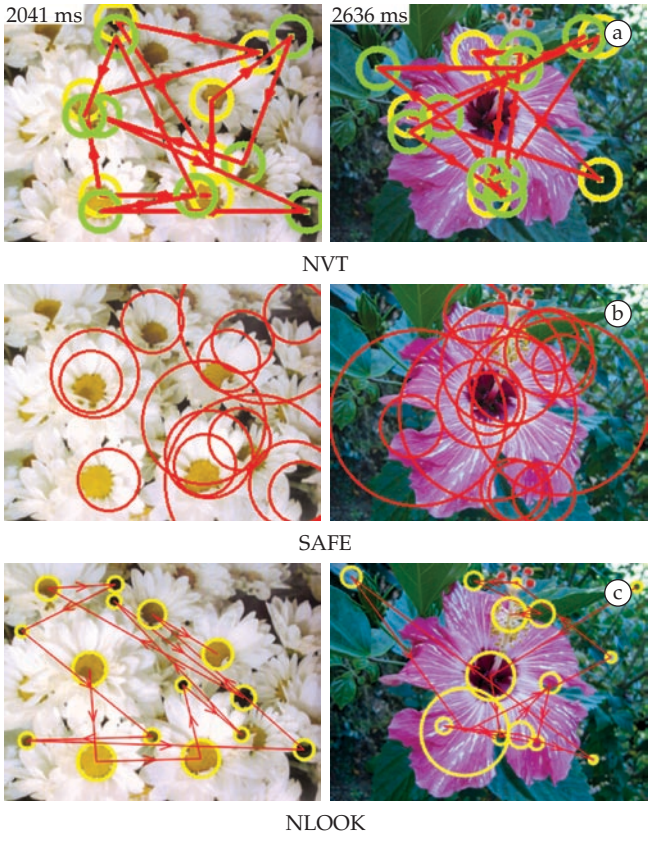


Figure 21. Visual outputs of all models using images with flowers.

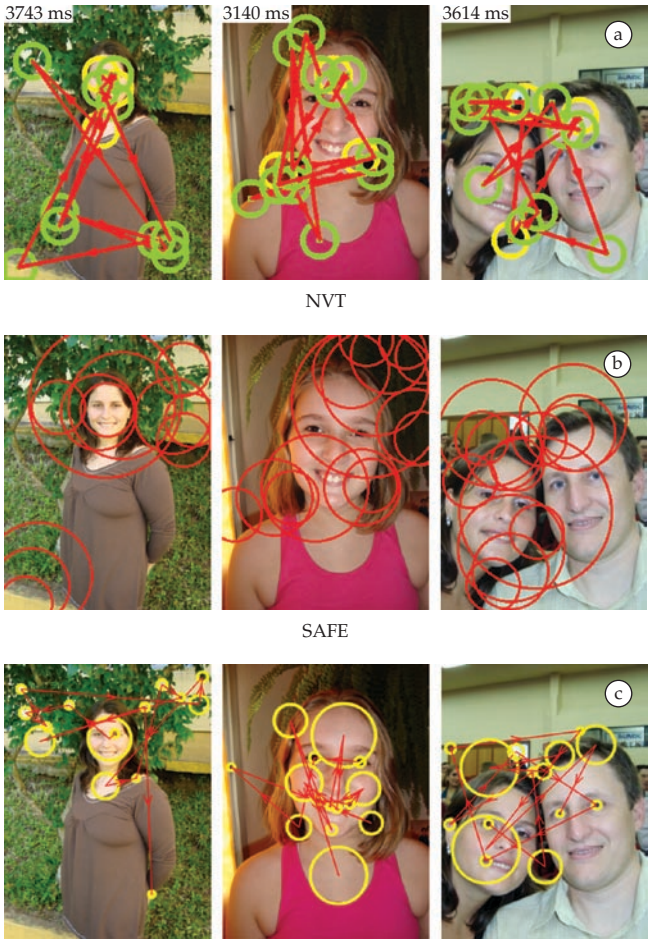


Figure 22. Visual outputs of all models using images with people.

the other two attention models. Moreover, the position drift values obtained using natural images are even smaller than those obtained with synthetic images, which shows that NLOOK can select the most salient elements with more precision than NVT and SAFE. In relation to the selected scales, the NLOOK performance is also very good (a difference of just two pixels in average), and SAFE was again worse than NVT. Figure 20 shows the fixations selected by each attention model (NVT, SAFE and NLOOK) using traffic signs images (these images were originally shown in Itti and Koch[18]).

It can be noticed, observing the fixations in Figure 20, that NVT was able to select the most interesting objects in both images, but with low precision. SAFE was also able to select most of the target objects, but the positions and scales were very imprecise. NLOOK, on the other hand, was able to select the most important objects in the visual field with very accurate positions and scales. Besides, the fixations selected by NLOOK have lower dimensions than those selected by SAFE, which allows a more accurate segmentation and inspection of these fixations. Figure 21 shows the obtained fixations in images containing flowers, which also shows that the fixations selected by NLOOK are much more accurate than the fixations selected by NVT and SAFE.

Just for visualization purposes, Figure 22 shows the fixations selected by all attention models using images containing people (these images were not used in the Table 4 experiments). It can be noticed that the precision of the fixations selected by NLOOK is far superior in this images, too – NLOOK was able to select with high precision visual elements like eyes, teeth and faces. These results show that NLOOK is a very good tool to be used in robotic vision systems, because it allows the automatic selection of the position and scales of the most important elements of the visual field, which is very useful in robotic tasks like landmark location and object recognition[13].

5. Conclusion

This paper presents a new visual attention model, called NLOOK, specially developed to be used as a front-end in robotic vision systems. The proposed model, which has a very good computational performance, is less sensitive to 2D similarity transforms than other well-known attention models like SAFE[8] and NVT[19], and can select fixation points in scale as well as position. Besides, NLOOK can select more accurate fixations than SAFE and NVT, which allows a better exploration of the visual field. The future perspectives include: i) using top-down cues to create feature maps; ii) to use the proposed model and a new probabilistic clustering algorithm, called INBC (Incremental Naïve Bayes Clustering)[9], in object categorization and landmark identification tasks; and iii) using the proposed model as a front-end in robotic vision tasks by a real robot Pioneer 3-DX.

Acknowledgements

This work is supported by CNPq, an entity of the Brazilian government for scientific and technological development.

References

1. Ando S. Image field categorization and edge/corner detection from gradient covariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence 2000;* 22(2):179-190.
2. Burt PJ, Hong T and Adelson EH. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications* 1983; 31(4):532-540.
3. Connor CE, Egeth HE and Yantis S. Visual attention: bottom-up versus top-down. *Current Biology* 2004; 14(19):850-852.
4. Crowley JL, Riff O and Piater J. Fast computation of characteristic scale using a half octave pyramid. In: *Proceedings of International Workshop on Cognitive Vision; 2002;* Zurich, Switzerland. Berlin, Germany: Springer-Verlag; 2002. p. 1-8.
5. Daugman JG. Complete discrete 2-d gabor transforms by neural networs for image analysis and compression. *Proceedings of IEEE Transactions on Acoustics, Speech, and Signal 1988;* 36(7):1169-1179.
6. Desimone R and Duncan J. Neural mechanisms of selective visual attention. *Annual Reviews Neuroscience* 1995; 18(1):193-222.
7. Draper BA, Baek K and Boody J. Implementing the expert object recognition pathway. In: *Proceedings of International Conference on Computer Vision Systems;* 2003; Graz, Austria. Berlin, Germany: Springer-Verlag; 2003. p. 1-11.
8. Draper BA and Lionelle A. Evaluation of selective attention under similarity transformations. *Computer Vision and Image Understanding* 2002; 100(1):152-171.
9. Engel PM. *INBC*: an incremental algorithm for dataflow segmantation based on a probabilistic approach. Porto Alegre: Universidadade Federal do Rio Grande do Sul; 2009. (Technical Report RP-3690)
10. Engel S, Zhang X and Wandell B. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature* 1997; 388(6637):68-71.
11. Frintrop S. *VOCUS*: a visual attention system for object detection and goal-directed search. [PhD thesis] . Bonn: *Universität Bonn*; 2006.
12. Greenspan S, Belongie S, Goodman R, Perona P, Rakshit S and Anderson CH. Overcomplete steerable pyramid filters and rotation invariance. In: *Proceedings of IEEE Computer Vision and Pattern Recognition*; 1994; Seattle, WA. Los Alamitos, CA: IEEE Press; 1994. p. 222-228.
13. Harel J and Koch C. On the optimality of spatial attention for object detection. In: *Proceedings of 5 International Workshop on Attention in Cognitive Systems*; 2009; Santorini, Grécia. Berlin, Germany: Springer-Verlag; 2009. p. 1-14. (v. 5395).
14. Heinen MR and Engel PM. Visual selective attention model for robot vision. In: *Proceedings of 5 IEEE Latin American Robotics Symposium;* 2008; Salvador, Brazil. Los Alamitos, CA: IEEE Press; 2008. p. 1-6.
15. Heinen MR and Engel PM. Evaluation of visual attention models under 2d similarity transformations. In: *Proceedings of 24 ACM Symposium on Applied Computing*; 2009; Honolulu, Hawaii. New York, NY: ACM press; 2009. (Special Track on Intelligent Robotic Systems).
16. Indiveri G, Mürer R and Kramer J. Active vision using an analog VLSI model of selective attention. *IEEE Transactions on Circuits and Systems* 2001; 48(5):492-500. (parte II, Analog and digital signal processing).
17. Itti L. Models of bottom-up attention and saliency. San Diego: Elsevier Press; 2005. p. 576-582.
18. Itti L and Koch C. Computational modeling of visual attention. *Nature Reviews*. 2001; 2(3):194-203.
19. Itti L, Koch C and Niebur E. A model of saliency-based visual attention for rapid scene nalysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998;* 20(11):1254-1259.
20. Kentridge R, Heywood C and Davidoff J. Color perception. In: Arbib MA. (Ed.). *The handbook of brain theory and neural networks*. 2 ed. Cambridge: MIT Press; 2003. p. 230-233.
21. Klein RM. Inhibition of return. *Trends in Cognitive Sciences*. 2000; 4(4):138-147.

22. Koch C and Ullman S. Shifts in selective visual attention: toward the underlying neural circuitry. *Human Neurobiology* 1985; 4(4):219-227.

23. Lee KW, Buxton H and Jianfeng F. Cue-guided search: a computational model of selective attention. *IEEE Trans. Neural Networks* 2005; 16(4):910-924.

24. Leventhal AG. *The neural basis of visual function*. Boca Raton: CRC Press; 1991. (v. 4, Vision and visual dysfunction).

25. Lindeberg T. Feature detection with automatic scale selection. *International Journal of Computer Vision* 1998; 30(2):79-116.

26. Liu YH and Wang XJ. Spike-frequency adaptation of a generalized leaky integrate-and-fire model neuron. *Journal of Computational Neuroscience* 2001; 10(1):25-45.

27. Lowe DG. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 2004; 60(2):91-110.

28. Marfil R, Bandera A, Rodríguez JA and Sandoval F. A novel hierarchical framework for object-based visual attention. In: *Proceedings of 5 International Workshop on Attention in Cognitive Systems*; 2009; Santorini, Grécia. Berlin, Germany: Springer-Verlag; 2009. p. 27-40. (v. 5393).

29. Marques O, Mayron L, Borba G and Gamba H. An attention-driven model for similar images with image retrieval applications. *EURASIP Journal on Advances in Signal Processing* 2007; (1):1-17.

30. Mozer MC and Sitton M. Computational modeling of spatial attention. In: Pashler H. (Ed.). *Attention*. London: Psychology Press, London; 1998. p. 341-395.

31. Nagai Y. From bottom-up visual attention to robot action learning. In: *Proceedings of 8 IEEE International Conference on Development and Learning*; 2009; Shanghai, China. Los Alamitos, CA: IEEE Press.

32. Niebur E and Koch C. Control of selective visual attention: modeling the "where" pathway. *Neural Information Processing System* 1996; 8(1):802-808.

33. Orabona F, Metta G, and Sandini G. Object-based visual attention: a model for a behaving robot. In: *Proceedings of 3 Attention and Performance in Computational Vision*; 2005; San Diego, CA. Los Alamitos, CA: IEEE Press; 2005.

34. Ouerhani N, Bur A and Hügli H. Visual attention-based robot self-localization. In: *Proceedings of European Conference on Mobile Robots*; 2005; Ancona, Italy. Los Alamitos, CA: IEEE Press; 2005. p. 8-13.

35. Pashler, H. *The Psycology of Attention*. Cambridge: MIT Press; 1997.

36. Perko R, Wojek C, Schiele B and Leonardis A. Integrating visual context and object detection within a probabilistic framework. In: *Proceedings of 5 International Workshop on Attention in Cognitive Systems*; 2009; Santorini, Grécia. Berlin, Germany: Springer-Verlag; 2009. p. 54-68. (v. 5395).

37. Treisman AM. Features and objects: the fourteenth bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology* 1988; 40(2):201-237.

38. Treisman AM and Gelade G. A feature integration theory of attention. *Cognitive Psychology* 1980; 12(1):97-136.

39. Tsotsos JK, Culhane SM, Wai WYK, Lai Y, Davis N, and Nuflo F. Modeling visual attention via selective tuning. *Artificial Intelligence* 1995; 78(1/2):507-545.

40. Vieira-Neto H. *Visual novelty detection for autonomous inspection robots*. [PhD thesis] . Essex: *University* of *Essex*; 2006.

41. Vieira-Neto H and Nehmzow U. Visual novelty detection with automatic scale selection. *Robotics and Autonomous Systems* 2007; 55(9):693-701.

42. Wang T, Zheng N and Mei K. A visual brain chip based on selective attention for robot vision application. In: *Proceedings of IEEE International Conference on Space Mission Challenges for Information Technology*; 2009. Los Alamitos, CA: IEEE Press; 2009. p. 93-97.

43. Witkin AP. Scale-space filtering. In: *Proceedings of International Joint Conference on Artificial Intelligence*; 1983; Karlsruhe, Germany. San Fransisco, CA: Morgan Kaufman; 1983. p. 1019-1022.