Pop-out and IOR in Static Scenes with Region Based Visual Attention

Muhammad Zaheer Aziz and Bärbel Mertsching

GET Lab, University of Paderborn, Pohlweg 47-49, 33098 Paderborn, Germany,
<last name>@get.upb.de,
WWW home page: http://getwww.upb.de

Abstract. This paper proposes a novel approach to construct the saliency map by combining region-based maps of distinct features. The multiplication style feature fusion process in the natural visual attention is modelled as weighted average of the features under influence of the external top-down and the internal bottom-up inhibitions. The recently discovered aspect of feature-based inhibition is also included in the procedure of IOR along with the commonly implemented spatial and feature-map based inhibitions. Results obtained from the proposed method are compatible with the well known attention models but with the advantages of faster computation, direct usability of focus of attention in machine vision, and broader coverage of visually prominent objects.

1 Introduction

The models of artificial visual attention usually concentrate on modeling the natural process of attention. The proposed complex processes may closely simulate the biological system, which obviously has far superior parallel computing power and intelligence as compared to the artificial systems, but may not be efficient when executing on limited resources of mobile (or even static) vision systems. Moreover, the low resolution saliency maps representing the natural pre-attention computations are of course suitable for the biological vision that is able to compute many procedures parallel to the attention process but such maps demand a redundant routine for machine vision after determination of the focus of attention due to fuzziness in output of attention module. This poses an overhead on the limited computing resources available on vision systems. Hence there is a need to bridge the gap between the requirements of machine vision and true modeling of visual attention. The prerequisites include acceleration in computation speed and to make the output of attention directly usable by the machine vision algorithms. The internal processing of the attention mechanism has to be examined in detail and modifications for proper approximations are to be made so that the actual mimicking of attention is optimized against the said requirements.

This paper is concerned with the research on a region based model of attention in which feature maps are constructed using regions obtained from a

Proceedings of the ICVS Workshop on Computational Attention & Applications - WCAA 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other ICVS contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

segmentation process optimized to produce a good input for attention [1]. After successfully experimenting with the approach based upon convex hulls of regions [2], feature computation methods are further improved by using innovations in the concerned algorithms for making them independent of convex hulls. Some of these proposed methods can be seen in [3] and [4]. In this discussion we present the methods for performing feature map fusion, which leads to detection of pop-out, and inhibition of return (IOR) for static scenes. Dynamic scenes, in which either the camera moves around or the scene consists of moving objects (or both), requires more complex modeling and cannot be covered within the limited scope of this paper.

Methods used in other models for computation of the accumulated saliency map and implementation of IOR are summarized in the next section. Then the proposed techniques for the same purpose are presented in which some new aspects of inhibition are also introduced. Results obtained from the proposed system are comparable with the contemporary models. In absence of a real consensus on any quantitative assessment method for evaluation of attention models [5], the compatibility with the existing models that focus on factual simulation of natural visual attention is considered as a measure of success.

2 Related Work

In this section we mention the parts related to feature fusion and inhibition of return from some existing artificial visual attention models. Before doing so it is appropriate to refer to some literature on natural vision in order to have a brief insight into the background concepts used by these models. Most of the feature-based visual attention models are founded on the feature-integration theory [6] according to which features are automatically registered in parallel across the visual field in an early stage before the objects are identified. It is proposed that separable feature dimensions including color, orientation, spatial frequency, brightness, and direction of movement are coded and they are combined to form a single object in the focus of attention. The mathematical models to combine the feature channels in the pre-attention phase are proposed in [7] and [8]. Both agree on the fact that the features are combined in the visual cortex using a multiplication-style operation. The operation is modeled as square of sum in [7] whereas [8] models it as $J_{(x,y)} = k_1 \times (O_{(x,y)} + k_2) \times (C_{(x,y)} + k_3)$ for orientation map represented by $O_{(x,y)}$ and color map by $C_{(x,y)}$ where k_1, k_2 and k_3 are constants.

In context of inhibition of return, it has been established by experiments in psychophysics that inhibition takes place in terms of both location and object features [9] [10]. Evidence is provided for inhibition in the immediate vicinity of the attended location in [11] and a U-shaped function has been reported which strongly suppresses the immediate surroundings of the attended location and gradually fades to no suppression after a limited diameter. The work of [12] discovers the idea of feature based inhibition in which inhibition of the color of the recently attended object has been reported in human vision.

Proceedings of the ICVS Workshop on Computational Attention & Applications - WCAA 2007
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany
This document and other ICVS contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Now we consider the feature combination techniques as implemented by some models of artificial visual attention for determining the pop-out and then inhibiting it. The model presented in [13] and [14] first normalize the feature maps of color contrast C, intensity I, and orientation O using a normalization function N and then apply a simple weighted sum to obtain the input S for the saliency map as follows:

$$S = [N(I) + N(C) + N(O)]\,/3$$

The saliency map is implemented as a 2D layer of leaky integrate-and-fire neurons that takes S as input and feeds into a Winner Take All (WTA) neural network. The WTA network ensures only one occurrence of most active location. In this model the inhibition of return is implemented by spatially suppressing an area in the saliency map around the current focus of attention while feature-based inhibition is not considered. Another recent effort [15] by the same group includes the task driven top-down influence during the bottom-up saliency map construction. The elementary units of computation are pixels or small image neighborhoods arranged in a hierarchical structure.

The model proposed in [16] uses a weighted sum of feature maps to obtain a combined saliency map. They use Independent Component Analysis algorithm for unsupervised learning to determine relative importance of features and to reduce redundancy. An adaptive mask is used to suppress the recently attended object for performing the inhibition of return. The model of [17] also computes a weighted sum of individual feature maps for obtaining an integrated attention map but introduce a manipulator map which is multiplied to the sum. The output map C is obtained by applying a threshold function θ on the weighted sum of the feature maps M_i and multiplying it to the manipulator map M_m. Hence

$$C_{(x,y)} = \sum_{i=1}^{N} \theta(w_i \times M_{i(x,y)}) \times \prod_{m=1}^{l} M_{m(x,y)}$$

The maximum in C is taken as the point of attention. No inhibition function was used as it was not needed in their application.

The model presented in [18] includes the aspect of tracking multiple objects while focusing attention in a dynamic scene. They first determine the features that lead to activation of the neural fields that are in turn responsible for determination of pop-out and then adapt the weights of these feature maps so that a pop-out emerging due to a specific feature receives the main support from that particular feature map. A separate map is used for IOR where the visited location is marked as highly active. This activity inhibits the master map of attention to avoid immediate revisiting of the attended location. The activity of the inhibition map decays slowly in order to allow revisiting of the location after some time.

The method proposed in [19] implements a hierarchical selectivity process using a winner-take-all neural network. They apply a top-down influence to increase or decrease the baseline of neural activity of the most prominent feature channel. As their model deals with so called 'groupings' of pixels, the IOR process

works on siblings of the current focus of attention in the hierarchy of groupings and sub-groupings. Another recent model in [5] uses the direct sum of the feature channels to compute a two-dimensional saliency map but they introduce an anisotropic Gaussian as the weighting function centered at the middle of the image. They have not reported any indigenous inhibition mechanism.

3 Proposed Techniques

Although pop-out detection and IOR are named as two different processes but they are very much interdependent on each other. The IOR greatly influences the process of pop-out by dictating what not to attend in the consequent attention cycle. We consider two types of inhibitions in our model. First is the top-down influence that can be regarded as an external stimulus from outside of the core attention mechanism. This inhibition factor may come from long term knowledge, recent experiences, and current needs. For example when a subject is asked to search for a red pen on a table, this top-down requirement forces to inhibit other features and highlight the features of color and eccentricity for the attention system. The other type of inhibition occurs within the attention mechanism to avoid repeatedly focusing on the same object. As both of these inhibition factors affect the weights of different features maps while accumulating them into a combined saliency map, we recommend to model both of these factors separately for clear demarcation of the two effects.

The top-down inhibition influence depends upon entities that are external to the scope of attention itself hence the provision to incorporate this influence is included here but it is not discussed in detail. On the other hand, we consider three types of internal inhibition mechanisms namely spatial, feature based, and feature-map based. Most of the existing models of attention implement either the spatial inhibition, in which a specific area around the point of attention is inhibited, or the feature-map based inhibition in which the weight of the wanted feature channel is adjusted to obtain required results. In some recent studies in psychophysics the feature based inhibition has also been reported [10] for example the color of the focus of attention is ignored in the successive attempts of attention [12]. This aspect of bottom-up inhibition is also included in the proposed approach.

The input for the proposed techniques is a list of n regions represented as R_i each having data about the feature values and the saliency magnitudes regarding m features, namely color, orientation, eccentricity, symmetry, and size ($m = 5$). Before summing up the feature saliencies, their weights $W_f^{inh}(t)$ are first initialized ($t = 0$ at initialization) such that the color map gets the highest weight, the size map get the lowest weight, and the others a medium one. Then the weights are adjusted such that the feature map offering the sharpest peak contributes more in the accumulated saliency map. It is done by finding the distance between the maximum and the average saliency value in each map. The feature map with the highest distance is considered as most prominent one and its weight is increased by a multiplicative factor δ (we take $\delta = 2$). Hence for

each feature f the distance between its maximum and average is given by:

$$\Delta_f = \left[max\left(S_f^t(R_i)\right) - \sum_{k=1}^{n} S_f^t(R_k)/n \; \forall R_i, \, 1 \leq i \leq n \right] \; \forall f, \, 1 \leq f \leq m$$

where $S_f^t(R_i)$ extracts the saliency value of feature f from the region R_i at time t. For the next attention cycle, the increment to the weight of the feature map offering the sharpest peak is applied as follows:

$$W_f^{inh}(t+1) = \begin{cases} W_f^{inh}(t) \times \delta & \text{for } \Delta_f = max(\Delta_j), \, 1 \leq j \leq m \\ W_f^{inh}(t) & \text{otherwise} \end{cases}$$

Now the total saliency ρ_i^t of a region R_i at time t is computed as:

$$\rho_i^t = \frac{\sum_{f=1}^{m} \left(W_f^{inh}(t) \times W_f^{td}(t) \times S_f^t(R_i) \right)}{\sum_{f=1}^{m} \left(W_f^{inh}(t) \times W_f^{td}(t) \right)}$$

where $W_f^{td}(t)$ is the weight of the feature map f according to the top-down influence. For this discussion we keep all $W_f^{td}(t) = 1$. After having attended a region at time t, the saliency value of each feature f is inhibited for use at time $t+1$. The inhibition function works on saliency map of each feature by influencing the saliency magnitude $S_f^t(R_i)$ of each feature in every region as follows:

$$S_f^{t+1}(R_i) = \begin{cases} S_f^t(R_i) \times \xi_s^i \times \xi_c^i & \text{for } f = 1 \text{(color feature)} \\ S_f^t(R_i) \times \xi_s^i \times \xi_f^i & \text{otherwise} \end{cases}$$

where ξ_s^i is the inhibition factor in spatial context for region R_i, ξ_c^i is the color inhibition factor, and ξ_f^i is the factor to inhibit due to similarity of features other than color. All three factors have values in the range between 0 and 1. ξ_s^i has the lowest magnitude when a region is at minimum distance from the focus of attention. It gradually grows to 1 after a particular radius r^{inh} around the center of attention. In other words, the decay is strongest near the center and it weakens to no decay as the boundary of inhibition circle is approached. Hence having the spatial distance between a given region R_i and the region under focus R_{foa} represented by $D^s(R_i, R_{foa})$ we define

$$\xi_s^i = \phi \times D^s(R_i, R_{foa})/r^{inh} + \phi_{min}, \text{ where } \phi_{min} + \phi = 1$$

where $D^s(R_i, R_{foa})$ is set to r^{inh} for $D^s(R_i, R_{foa}) > r^{inh}$ in order keep the outcome within a unit amount. ϕ is the weight of the actual contribution from the distance of R_i from the FOA while ϕ_{min} is the minimum value of ξ_s^i when $D^s(R_i, R_{foa})$ is close to zero. We take $\phi = 0.67$ and $\phi_{min} = 0.33$.

The second inhibition factor ξ_c^i inhibits in context of color. Regions close to the FOA having similar color get a strong suppression of saliency while beyond the radius r^{inh} this suppression reduces. As the color similarity has different criteria for chromatic and achromatic colors hence ξ_c^i has different formats for both situations. For an achromatic region

$$\xi_c^i = \phi \times \frac{D^s(R_i, R_{foa})}{r^{inh}} \times \frac{D^{int}(R_i, R_{foa})}{\tau^{int}} + \phi_{min}$$

where $D^{int}(R_i, R_{foa})$ is the intensity difference between R_i and the FOA region which is set to the intensity difference threshold τ^{int} when $D^{int}(R_i, R_{foa}) > \tau^{int}$. On the other hand for regions with chromatic colors

$$\xi_c^i = \phi \times \frac{D^s(R_i, R_{foa})}{r^{inh}} \times \frac{D^{hue}(R_i, R_{foa})}{\tau^{hue}} \times \frac{D^{sat}(R_i, R_{foa})}{\tau^{sat}} \times \frac{D^{int}(R_i, R_{foa})}{\tau^{int}} + \phi_{min}$$

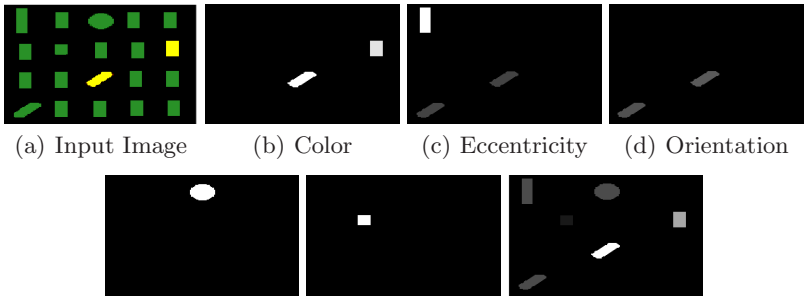
where $D^{hue}(R_i, R_{foa})$ and $D^{sat}(R_i, R_{foa})$ are the hue and saturation differences between the color of R_i and the FOA rectified above the related thresholds τ^{hue} and τ^{sat} respectively. Now the third inhibition factor ξ_f^i inhibits saliency of those regions that have similarity to the FOA in terms of features other than color. We take $\xi_f^i = 0.75$ when $D^f(R_i, R_{foa})$, the difference with respect to the feature f, is below a threshold τ^f.

Yet another inhibition is modeled for preventing a feature map to gain extra ordinary weight. When a weight $W_f^{inh}(t)$ becomes equal to $max(W_f^{inh}(t) \forall f, 1 \leq f \leq m)$ then it is set back to the original value that was assigned to it during the initialization step. This mechanism keeps the weights of feature maps in a cycle as they keep rising when the concerned feature map contains a sharp peak until this peak gets attended or gets inhibited due to attention to some neighboring region.

4 Results

In order to demonstrate the feature combination and pop-out detection we present the output of the system using a synthetic benchmark image shown in figure 1(a). The bar at the top left has saliency due to its eccentricity, the circle in the top row is the most symmetric, the square in the second row has a unique size, the rightmost yellow rectangle in the same row is prominent due the its color, the third tilted region in the third row is different in its orientation in addition to being yellow, and the similar region at the left bottom is oriented differently compared to most of the regions but has a green color similar to the majority. The feature maps resulting from the feature computation routines of the proposed model are shown in figures 1(b) to (f). The color map clearly shows the two regions having the contrasting color (yellow) as salient, the orientation map highlights the differently oriented regions, the eccentricity map shows the elongated regions as prominent, the circle emerges in the symmetry map, and the size map shows the uniquely sized small square as outstanding. The weighted

combination of these feature maps into a master map is shown in figure 1(g). The tilted region having saliency with respect to color, eccentricity, and orientation gets the highest saliency and it will be the first pop-out to attract visual attention. Other regions with saliency due to only one feature show less prominence according to the initial weights set for the related feature maps, for example due to the small initial weight for size the small square shows the least saliency.

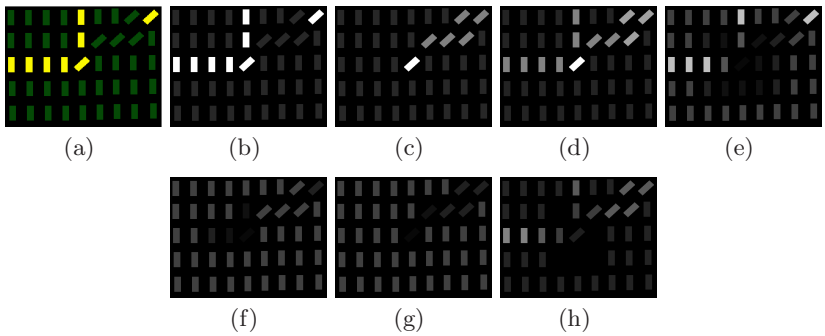


(a) Input Image (b) Color (c) Eccentricity (d) Orientation

(e) Symmetry (f) Size (g) Saliency map

Fig. 1. Feature maps and combined saliency map using a synthetic benchmark image

Figure 2 demonstrates the effect of first IOR on the color map, orientation map, and the master map in order to show the results of spatial and feature based inhibition. The input image shown in figure 2 (a) contains objects that are salient either due to their color contrast or orientation. After inhibition on the first pop-out located at the image center, which is salient due to both features, the regions having similar orientation and color are inhibited even at far distances along with the spatial inhibition within a certain radius.

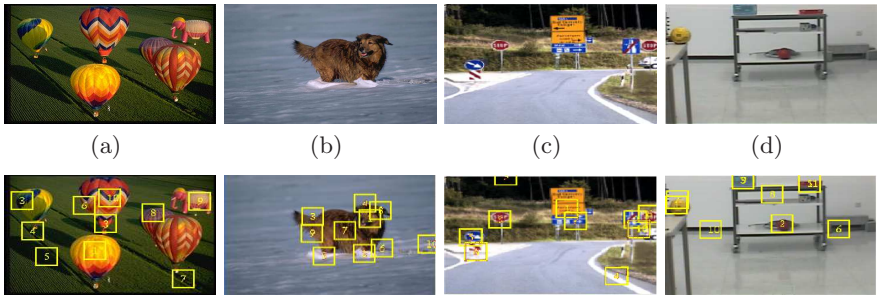


(a) (b) (c) (d) (e)

(f) (g) (h)

Fig. 2. Results of inhibition of return (a) Input image (b) Color map (c) Orientation map (d) Master map (e) Spatial inhibition around current FOA (center) (f) Inhibition in color map (g) Inhibition in orientation map (h) Master map after first IOR

Figure 3 presents the output of the proposed mechanism for attending first ten salient location in some real-life static images. The samples shown here include a balloon image provided on the internet resource of [13], an animal image used as test case by the model of [20], a traffic scene used as sample by [18] and a lab scene viewed through the camera head of a mobile vision system. It can be qualitatively assessed the proposed system has marked all those areas that are usually important for a human observer.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 3. First ten foci of attention on real-life images (a) - (d) Input images. (e) - (h) Output of the proposed model

5 Discussion

We compare the results from the proposed model with the existing models of [13] and [18] as the source code of the earlier is available on internet and the later was developed in the same research group formerly. Figure 4 displays the output of first ten foci of attention using the said two models. It may be observed that the locations attended by the proposed model mostly resemble the reference model of [13]. Some points that are important for attention, such as the sign boards at the right side of the traffic scene and the box lying at the left on top of the table in the lab scene, are ignored by the reference model. The attention by the proposed model to these locations may be considered as an positive aspect. Some locations that are skipped by the proposed method, such as the balloon at the right bottom in the balloons image, are mainly due to strong feature based inhibition as the red areas on the said balloon are suppressed after attending a red region on the neighboring balloon.

In order to compare the working speed of the three models under discussion they were executed on the same machine and the CPU time spent in milliseconds for attending the first ten foci of attention was recorded. Because of the fairly high amount of time taken by the existing models, the graphs of time shown in figure 5 is plotted after taking the logarithm of values. The advantage of

speed in the proposed model is clearly visible. Along with this gain in time, the direct usability of the FOA in machine vision applications can lead to further improvement in performance of the overall vision system.

The main disadvantage of the region based technique appears in presence of noise in the input causing poor quality of segmentation output that leads to decline in performance of attention mechanism. The proposed method has shown considerable robustness even in presence of noise. For example the lab scene is one of the samples taken from the live input from the robot camera head that contains a significant amount of transmission distortions and the balloon image is also a very noisy sample. The results are still of acceptable quality.

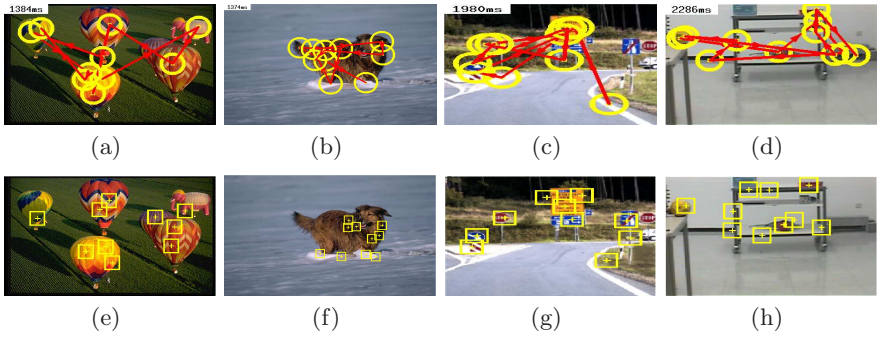


Fig. 4. (a) - (d) First ten foci of attention on by model of [13] using input images of figure 3 (a) to (d). (e) - (h) First ten foci of attention on by model of [18] using the same input

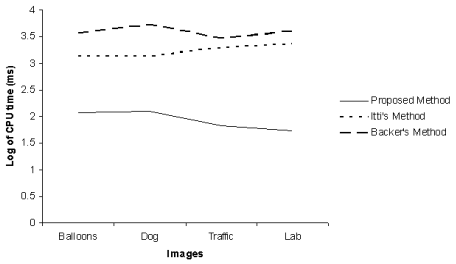


Fig. 5. Comparison of run time for attending first ten foci of attention

References

1. Aziz, M.Z., Mertsching, B.: Color segmentation for a region-based attention model. In: Farbworkshop 2006, Ilmenau - Germany (2006) 74–83
2. Aziz, M.Z., Mertsching, B., Shafik, M.S., Stemmer, R.: Evaluation of visual attention models for robots. In: ICVS 06, New York - USA, IEEE (2006) index–20
3. Aziz, M.Z., Stemmer, R., Mertsching, B.: Region-based depth feature map for visual attention in autonomous mobile systems. In: AMS 2005, Stuttgart - Germany, Informatic Actuell, Springer (2005) 89–95
4. Aziz, M.Z., Mertsching, B.: Color saliency and inhibition in region based visual attention. In: WAPCV 2007, Hyderabad - India (2007) 95–108
5. Meur, O.L., Callet, P.L., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. Transactions on Pattern Analysis and Machine Intelligence **28** (2006) 802–817
6. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Congnitive Psychology **12** (1980) 97–136
7. Koch, C.: Biophysics of Computation. Oxford University Press, New York (1999)
8. Neri, P.: Attentional effects on sensory tuning for single-feature detection and double-feature conjunction. Vision Research (2004) 3053–3064
9. Gibson, B.S., Egeth, H.: Inhibition of resturn to object-based and environment-based locations. Perception and Psychophysics **55** (1994) 323–339
10. Weaver, B., Lupianez, J., Watson, F.L.: The effects of practice on object-based, location-based, and static-display inhibition of return. Perception & Psychophysics **60** (1998) 993–1003
11. Cutzua, F., Tsotsos, J.K.: The selective tuning model of attention: Psychophysical evidence for a suppressive annulus around an attended item. Vision Research (2003) 205–219
12. Law, M.B., Pratt, J., Abrams, R.A.: Color-based inhibition of return. Perception & Psychophysics (1995) 402–408
13. Itti, L., Koch, U., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. Transactions on Pattern Analysis and Machine Intelligence **20** (1998) 1254–1259
14. Itti, L., Koch, C.: A saliency based search mechanism for overt and covert shifts of visual attention. Vision Research (2000) 1489–1506
15. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vision Research (2005) 205–231
16. Park, S.J., Ban, S.J., Sang, S.W., Shin, J.K., Lee, M.: Implementation of visual attention system using bottom-up saliency map model. In: ICANN/ICONIP 03, LNCS 2714, Springer (2003) 678–685
17. Heidmann, G., Rae, R., Bekel, H., Bax, I., Ritter, H.: Integrating context-free and context-dependant attentional mechanisms for gestural object reference. In: ICVS 03, LNCS2626, Springer (2005) 22–33
18. Backer, G., Mertsching, B., Bollmann, M.: Data- and model-driven gaze control for an active-vision system. Transactions on Pattern Analysis and Machine Intelligence **23** (2001) 1415–1429
19. Sun, Y., Fischer, R.: Object-based visual attention for computer vision. Artificial Intelligence **146** (2003) 77–123
20. Stentiford, F.: An estimator for visual attention through competitive novelty with application to image compression. In: Picture Coding Symposium, Seoul - Korea (2001) 101–104