Visual Search in Static and Dynamic Scenes Using Fine-Grain Top-Down Visual Attention

Muhammad Zaheer Aziz and Bärbel Mertsching

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

Abstract. Artificial visual attention is one of the key methodologies inspired from nature that can lead to robust and efficient visual search by machine vision systems. A novel approach is proposed for modeling of top-down visual attention in which separate saliency maps for the two attention pathways are suggested. The maps for the bottom-up pathway are built using unbiased rarity criteria while the top-down maps are created using fine-grain feature similarity with the search target as suggested by the literature on natural vision. The model has shown robustness and efficiency during experiments on visual search using natural and artificial visual input under static as well as dynamic scenarios.

1 Introduction

Finding a robust, flexible, and efficient solution for visual search in real-life scenes has been a topic of significant interest for researchers in the field of machine vision. In the recent years emphasis has been increased on vision systems engineered according to the role model of human or natural vision in order to achieve generic solutions able to perform competently independent of the input complexity. Visual attention is one of the prominent attributes of natural vision that contributes into its efficiency and robustness. Computational models of this phenomenon has been built and applied to many vision-based problems.

A majority of the existing attention models have demonstrated visual search as a primary area of application for their models. Most of these models have utilized manipulation on bottom-up (BU) saliency maps in order to let the search target pop-out early. We argue that the top-down (TD) tasks of attention have a different nature and require a separate mechanism for computing saliency. The models of human vision such as [1] suggest target related feature processing in the V4 area of brain. Similarly the models on feature and conjunction search, for example [2], also presume excitation and inhibitions on particular feature magnitudes rather than whole channels. Results of psychophysical experiments reported by [3], [4], and [5] also support our argument. The work of [3] has shown that a population of neurons encoding the target color and/or orientation gets a gain while others get suppressed. According to [4], each feature channel can adopt many values that are evaluated by a specialized layer of neurons in the human brain. The experiments reported by [5] explicitly declare fine-grain nature

A. Gasteratos, M. Vincze, and J.K. Tsotsos (Eds.): ICVS 2008, LNCS 5008, pp. 3–12, 2008.

of TD attention showing that particular feature values are highlighted by human vision rather than the whole feature channel. These findings suggest that the TD saliency mechanism constructs task dependant maps to allow quick pop-out of the target rather than using the BU saliency maps, hence we propose to model the top-down pathway independent of the bottom-up process.

This paper introduces an approach that applies influence of the active attention behavior at the early stage of saliency map construction in contrast to the existing models that apply TD influences at a later stage. Processes for construction of BU and TD saliency maps are separated from each other in the proposed model. The major significance of this work is the proposal of a model of TD attention based upon fine-grain saliency maps, which has been done for the first time as per knowledge of the authors. Another highlight is the experiments on visual search using dynamic scenes carried out by active vision systems as the existing models of TD attention have mostly experimented with static images, rarely addressing true active vision in attentional search applications.

2 Related Work

Early computational models of visual attention such as [6] and [7] have proposed a comprehensive mechanism for determining BU saliency using some feature channels but they use the same BU saliency maps for search task as well. They apply high weight to the feature channel that facilitates highlighting the search target. Even the recent developments by the same group in this context [8][9] apply a similar strategy. The model of [10] determines weights for the feature maps that would highlight the target in a learning stage and applies them in the searching stage. Although [11] has separate components for BU and TD pathways in the model but the same saliency maps are used to deal with the TD pathway. The model presented in [12] also applies attentional bias towards the target by learning weights for the conspicuity maps that would make the required object prominent. Such approaches are likely to show inefficiency when distractors are also salient in the same feature channel.

The work presented in [13] has provided a search mechanism to detect the target by looking for its constituent parts. This approach can be considered close to fine-grain search but the methodology is inclined towards pure machine vision rather than following a biologically inspired approach. Using gist of the whole view to apply a TD influence to restrict search locations as proposed by [14] is also a useful concept that can accelerate biologically plausible visual search. This concept deals with signature of the whole image rather than individual items.

The object-based attention models such as [15] seem to have similarity with the fine-grain nature of attention because objects are defined by particular feature values. Existing object-based models have concentrated on finding only BU saliency using objects as a fundamental unit. Hence TD saliency maps based upon fine-grain concept still remains untried.

The proposed region-based methodology for attention modeling has developed as an evolutionary process. The earlier model from our group [16] introduced

attentional tracking in dynamic scenes but it had high computation time and lacked robustness in visual search. The first prototype [17] for the region-based approach used convex hulls of the segmented regions. After enabling the segmentation algorithm to produce an optimized input for use of attention [18], new methods were developed to compute BU saliency using channels of color [19] and other features [20]. Methods for applying inhibition of return (IOR) and determining pop-out in the region-based paradigm were proposed in [21], groundwork for using fine-grain saliency using color channel was established in [22], and solution for handling bottom-up attention and IOR in dynamic scenarios was proposed in [23]. Here we extend the model by introducing other feature channels in the TD pathway and propose methods for TD map fusion and IOR on both TD and BU saliency maps.

3 Proposed Region-Based Approach

The proposed model groups pixels of the visual input possessing similar color attributes into clusters using a robust segmentation method [18] before starting attention related processes. Assigning fine grain attributes to these regions allows using them as units to be processed by attention procedures. Some models such as [16] perform a clustering step in the final saliency map but such late clustering becomes less effective and inefficient because most of the feature magnitudes related to the actual objects get faded away at this stage because of processing on fine and coarse scales of input.

The proposed model separates the steps of feature magnitude computation and saliency evaluation as shown in figure 1. The primary feature extraction function F produces a set of regions $\Re$ consisting of n regions each represented as R_i and feeds each R_i with data regarding location, bounding rectangle, and magnitudes of each feature ϕ_i^f $(f \in \Phi)$. As five channels of color, orientation, eccentricity, symmetry and size are considered in the current status of our model hence we have $\Phi = \{c, o, e, s, z\}$.

Computation of the BU saliency using the rarity criteria is performed by the process S whose output is combined by W that applies weighted fusion of these maps to formulate a resultant BU map. The function G consideres the given TD conditions to produce fine grain saliency maps that are combined by the function C into a resultant TD map. The function P applies appropriate weights to the resultant saliency maps according to the active attention behavior, combines them into a master conspicuity map, and applies a peak selection mechanism to choose one pop-out at a time. The focus of attention at a particular time t is stored in the inhibition memory using which the process of IOR suppresses the already attended location(s) at time $t+1$ in order to avoid revisiting of the same location. The memory management function M decides whether to place the recent focus of attention in inhibition memory or excitation memory according to the active behavior and sets the weights of inhibition.

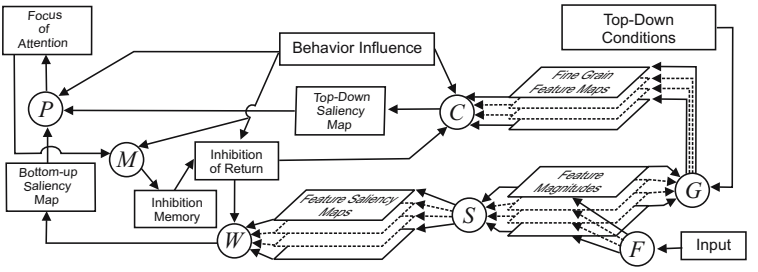

Fig. 1. Architecture of the proposed region-based attention model

3.1 Fine-Grain Top-Down Saliency Maps

The process G from the architecture diagram shown in figure 1 is responsible for construction of fine-grain saliency maps for each feature channel f considered in the model. The search target is defined as a set of top-down feature values F_{td} in which the individual features are referred as F_{td}^f. For constructing the saliency map with respect to color ($f = \{c\}$), we define D^h as the difference of hue that can be tolerated in order to consider two colors as similar, D^s as the tolerable saturation difference, D^I as the allowed intensity difference for equivalent colors, and ϕ_i^c as the magnitude of the color feature for R_i. Now, the TD color saliency γ_i^c of each region R_i is determined as follows:

$$\gamma_i^c = \begin{cases} \frac{a(D^h - \Delta_i^h)}{D^h} + \frac{b(D^s - \Delta_i^s)}{D^s} + \frac{c(D^I - \Delta_i^I)}{D^I} & \text{for } \Delta_i^h < D^h \text{ \& chromatic } \phi_i^c, F_{td}^c \\ \frac{(a+b+c)(D^I - \Delta_i^I)}{D^I} & \text{for } \Delta_i^I < D^I \text{ \& achromatic } \phi_i^c, F_{td}^c \\ 0 & \text{otherwise} \end{cases}$$

where a, b, and c are weighting constants to adjust the contribution of each color component into this process. Δ_i^h, Δ_i^s, and Δ_i^I are magnitudes of the difference between ϕ_i^c and F_{td}^c in terms of hue, saturation, and intensity respectively. We take $a = 100$, $b = 55$, and $c = 100$ because the saliency values of a region lie between the range of 0 and 255 in our model. The value of b is kept smaller in order to keep more emphasis on the hue and intensity components. Hence a perfect match would result in a saliency value equal to 255.

The color map had specific requirements being a composite quantity whereas the other feature channels consist of single-valued quantities; hence they can be processed using a simpler procedure. Having Θ^f as the normalized ratio of the feature magnitudes ϕ_i^f and F_{td}^f (for $f \neq \{c\}$) defined as

$$\Theta^f = \begin{cases} \phi_i^f / F_{td}^f & \text{for } \phi_i^f < F_{td}^f \\ F_{td}^f / \phi_i^f & \text{otherwise} \end{cases}$$

which always keeps $1 \geq \Theta^f \geq 0$. Now the TD saliency γ_i^f of a region R_i with respect to a feature f ($f \in \Phi, f \neq \{c\}$) will be computed as

$$\gamma_i^f = \begin{cases} k\Theta^f & \text{for } \Theta^f > D^\Theta \\ 0 & \text{otherwise} \end{cases}$$

where k is a scaling constant and D^{Θ} is the ratio above which the two involved quantities may be considered equivalent. We take $k = 255$ because the maximum amount of saliency can be 255 in our implementation and $D^{\Theta} = 0.91$.

3.2 Map Fusion and Pop-Out

In this paper we are concerned with the TD portion of the model hence we explain the map fusion function C that produces a resultant TD saliency map. We take W_{td}^{f} as the TD weight for the map of feature channel f that gets a value depending upon the active behavior of the vision system. Under search behavior, high weights are set for color channel while keeping low weights for other shape-based features because the target could be in an arbitrary size or orientation in the given input. Under track behavior other feature channels also gain high weight because the target has to match strict criteria. The resultant TD saliency $\gamma_i(t)$ of a region R_i at time t is computed as follows

$$\gamma_i(t) = \sum^{\forall f \in \Phi} \left(W_{td}^{f} \gamma_i^{f} \right) / \sum^{\forall f \in \Phi} W_{td}^{f}$$

Resultant of BU saliency is obtained as $\beta_i(t)$ for which details can be seen in [21]. The function P combines the BU and TD saliency maps to produce the final conspicuity map. The active behavior again plays an important role at this step by adjusting weights of these two maps. Under explore behavior the major emphasis remains on the BU map while during other behaviors, like search or track, high weight goes to the TD channel. Denoting the behavior dependant weight for TD map as W_{td}^{b} and for BU map as W_{bu}^{b}, the final saliency $S^{i}(t)$ of each R_i at time t is given as

$$S^{i}(t) = \left(W_{bu}^{b} \beta_i(t) + W_{td}^{b} \gamma_i(t) \right) / \left(W_{bu}^{b} + W_{td}^{b} \right)$$

3.3 Inhibition Using Saccadic Memory

After having attended a region at time $t-1$, the saliency value of that region with respect to each feature f is inhibited for use at time t. Instead of using an inhibition map as done by existing methods we use a memory oriented mechanism. As our application area is mobile active vision systems, previously attended locations may get relocated in subsequent frames of input. We propose to put the attended regions into a spatial inhibition memory M_{inh}^{s} able to remember p regions. An item is inserted into M_{inh}^{s} as M_{k}^{s} where the age k is set to 1 for freshly arrived item and the older entries get an increment in their values of k on arrival of a new item. In order to deal with the problem of relocation of regions in context of the view-frame, we use the world coordinates of the regions calculated using the head angles and the position of regions within the view-frame.

We apply the inhibition right after formulation of region saliency in order to make the model efficient. We take the time at which the freshly arriving regions get their saliency as $t-1$ and the time after going through the inhibition process

as t. Hence, at time t, for each R_i with BU and TD saliency values with respect to a feature f, represented as $\beta_i^f(t)$ and $\gamma_i^f(t)$ respectively, are updated from $\beta_i^f(t-1)$ and $\gamma_i^f(t-1)$ as follows:

$$\beta_i^f(t) = \delta_1^k \beta_i^f(t-1) \text{ when } D^s(R_i, M_k^s) < r^{inh} \forall k \in \{1..p\}$$
$$\gamma_i^f(t) = \delta_2^k \gamma_i^f(t-1) \text{ when } D^s(R_i, M_k^s) < r^{inh} \forall k \in \{1..p\}$$

where r^{inh} is the radius in which inhibition takes effect and $D^s(R_i, M_k^s)$ is the spatial distance between the considered region R_i and the region in the memory location M_k^s. δ_1^k and δ_2^k are inhibition factors both having a value between 0 and 1. The value of δ_1^k becomes closer to 1 as the age of M_k^s increases, hence suppression on recently attended items is stronger than the older ones. δ_2^k remains the same for all items in the memory because under TD attention, such as search, once the target is found at a location then that location has to be strongly inhibited during the next few saccades of further search.

4 Results

Experiments were performed to test the search capabilities of the proposed method by using three scenarios. In each scenario the search target was given to the system in form of an image containing the isolated target over a blank background. In the current status the system is able to work with single regions at a time rather than composite objects hence the system picks the largest region from the picture of the target as the region to search. The first scenario of experiments was the search in static scenes in which the attention mechanism was allowed to mark as many occurrences of the target as possible. These experiments tested the ability of the system to select all relevant locations. Figure 2(a) reflects this scenario with the search field as a still scene viewed through the camera of the mobile vision system available in our laboratory and four occurrences the target (a dull blue box) in the scene. In the second scenario a simulated vision system was set into motion and it was required to mark the locations matching the search target one location per frame. This scenario was useful to test the ability of inhibition of return in dynamic scenes. Figure 2(b) represents this scenario in the simulation framework developed in our group [24]. In the third scenario the attention mechanism was required to perform overt attention

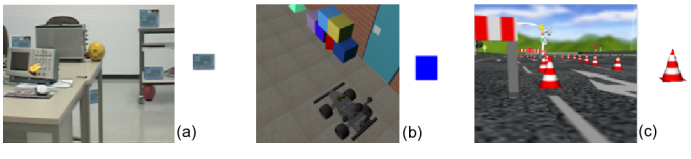

Fig. 2. Samples from visual input used in experiments. (a) Search field and target used as static scenario (b) Search environment and target used for dynamic scene scenario (c) Search environment and target used to test overt attention.

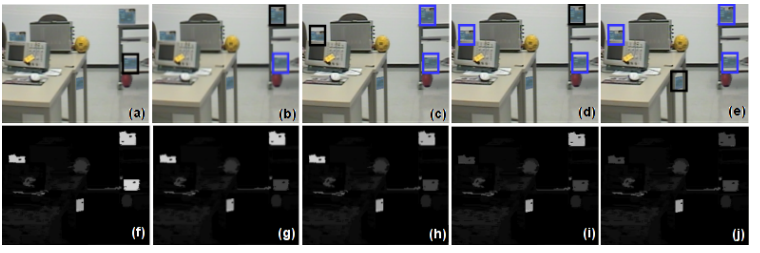

Fig. 3. Top row: Fixated (black) and inhibited (blue) locations for static scenario given in figure 2(a). Bottom row: Top-down saliency maps at time of each fixation.

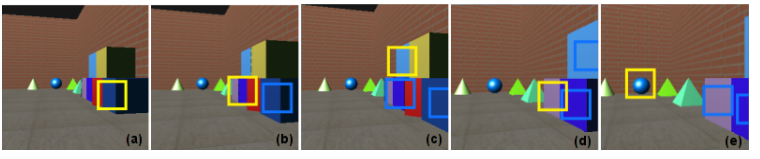

Fig. 4. Fixated (yellow) and inhibited (blue) locations for dynamic scenario given in figure 2(b)

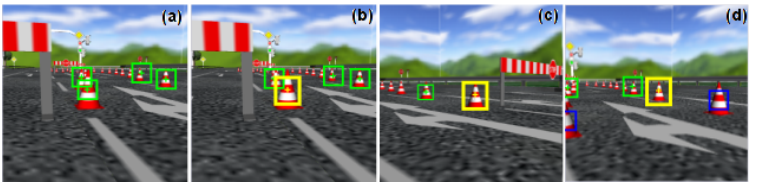

Fig. 5. Target locations brought into center of camera view (yellow), salient locations (green), and inhibited locations (blue) for overt attention scenario given in figure 2(c)

to the best matching location by bringing the target into center of camera view. Hence one selection per saccade was allowed. These experiments tested the ability of the system to locate the (estimated) position of the search target in three dimensional space. Figure 2(c) shows a sample input for this scenario.

Figure 3 demonstrates output of attentional search for the test case given in figure 2(a). Results of first five fixations by the attention system ($t = 1$ to $t = 5$) are reported here. The current focus of attention is marked with a black rectangle while blue rectangles are drawn at the inhibited locations. It may be noted that the four target locations are marked in the first five fixations in which the extra fixation is due to a repeated saccade on an object that had such a high top-down saliency that it still remained higher than the fourth object, which had relatively less similarity with the target, even after inhibition. This aspect can be noticed in the saliency maps provided in the second row of figure 3. Results of search in a dynamic scene performed by the vision system in motion are shown in figure 4.

The occurrences of the target in the environment, both shown in figure 2(b), are marked by the vision system working in the simulation framework. After fixating on the best matches, the system tries to pick target locations even when they have less similarity with the target, for example, the later fixations are done based only upon color similarity. Figure 5 demonstrates the results of overt attention in which the vision system maneuvers its camera head to bring the search target into to fovea area (center of view). Salient locations are marked with green rectangles, the attended locations brought into the center of view with yellow, and inhibited ones with blue.

5 Evaluation and Conclusion

In order to quantitatively evaluate the performance of our model, we carried out experiments using some specially designed visual data apart from the visual input consisting of natural and virtual reality images. Five occurrences of a predefined search target were embedded in each test image that contained distractors offering quantified amount of complexity. In the simplest case, as shown in images labeled as 1 and 1-D in figure 6, the distractors possessed high difference of features (Color, orientation, and size) from the target. The rest of the samples were created using different combinations of feature differences as shown in table 1 where H represents a high difference from target and L stands for low difference, hence the inputs labeled 8 and 8-D offer the maximum amount of complexity. The samples labeled as 1-D to 8-D contain extra distractors possessing high bottom-up saliency and the occurrences of the targets were distorted by introducing gradually rising blur (increasing from right to left in each image).

Table 1. Feature differences between target and distractors used for figure 6

Image label	1 / 1-D	2 / 2-D	3 / 3-D	4 / 4-D	5 / 5-D	6 / 6-D	7 / 7-D	8 / 8-D
Color	H	H	H	H	L	L	L	L
Orientation	H	H	L	L	H	H	L	L
Size	H	L	H	L	H	L	H	L

A comparison of attentional search models is given in [12] using the criteria of time taken and number of fixations to reach the target. Similarly [10] uses the average hit number to reach the target as a measure of search efficiency. Evaluation of our model in terms of these two metrics using the test cases given in figure 6 is shown in figure 7. The proposed model was able to locate the search target in the first fixation in all experiments (hence, 1 fixation per search). Average time to fixate on the first target location was 23.6 milliseconds in these evaluation experiments while average search time in the natural images was 69.3 ms on Linux based 3 Ghz Pentium 4 machine. This time includes segmentation and feature computation processes. None of the distractors possessing high BU saliency were fixated in all experiments. The time reported by [12] for an average search is 1.1 seconds on Linux based dual Opteron machine with minimum 2.2

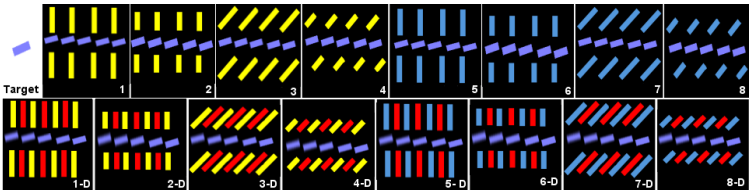

Fig. 6. Search target and sample input, having distractors offering different levels of complexity, used for quantitative evaluation

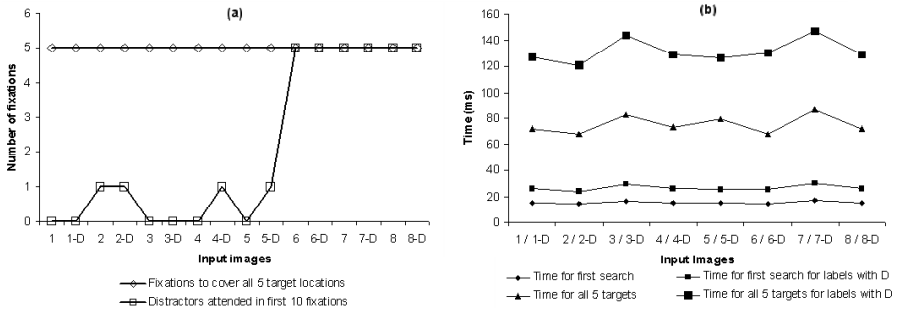

Fig. 7. Results of experiments using evaluation data given in figure 6

GHz clock speed while they have reported the average search time of the model of [7] to be 1.43 seconds on the same machine. In terms of fixations per search, the model of [7] has an average of 4.03, [12] has 2.73, and [10] has reported an average of 1.45 in best cases and 3.39 in worst cases.

It may be concluded that the region-based methodology with the innovation of constructing the fine-grain saliency maps separate from the bottom-up maps, using the concepts taken from the recent literature on research in natural visual attention, is an efficient and robust alternative to the existing approaches. The proposed method is also immune to bottom-up saliency of distractors in every feature channel and does not require any tuning of parameters or adjusting of weights. The memory based inhibition mechanism has also shown success in static as well as dynamic scenarios.

References

1. Lanyon, L., Denham, S.: A model of object-based attention that guides active visual search to behaviourally relevant locations. In: Paletta, L., Tsotsos, J.K., Rome, E., Humphreys, G.W. (eds.) WAPCV 2004. LNCS, vol. 3368, pp. 42–56. Springer, Heidelberg (2005)
2. Laar, P., Heskes, T., Gielen, S.: Task-dependent learning of attention. Neural Networks 10, 981–992 (1997)
3. Hamker, F.H.: Modeling attention: From computational neuroscience to computer vision. In: Paletta, L., Tsotsos, J.K., Rome, E., Humphreys, G.W. (eds.) WAPCV 2004. LNCS, vol. 3368, pp. 118–132. Springer, Heidelberg (2005)

4. Deco, G.: The computational neuroscience of visual cognition: Attention, memory and reward. In: Paletta, L., Tsotsos, J.K., Rome, E., Humphreys, G.W. (eds.) WAPCV 2004. LNCS, vol. 3368, pp. 100–117. Springer, Heidelberg (2005)
5. Navalpakkam, V., Itti, L.: Top-down attention selection is fine-grained. Journal of Vision 6, 1180–1193 (2006)
6. Itti, L., Koch, U., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. Transactions on PAMI 20, 1254–1259 (1998)
7. Itti, L., Koch, C.: A saliency based search mechanism for overt and covert shifts of visual attention. Vision Research, pp. 1489–1506 (2000)
8. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vision Research, pp. 205–231 (2005)
9. Navalpakkam, V., Itti, L.: Optimal cue selection strategy. In: NIPS 2006, pp. 1–8. MIT Press, Cambridge (2006)
10. Frintrop, S., Backer, G., Rome, E.: Goal-directed search with a top-down modulated computational attention system. In: Kropatsch, W.G., Sablatnig, R., Hanbury, A. (eds.) DAGM 2005. LNCS, vol. 3663, pp. 117–124. Springer, Heidelberg (2005)
11. Michalke, T., Gepperth, A., Schneider, M., Fritsch, J., Goerick, C.: Towards a human-like vision system for resource-constrained intelligent cars. In: ICVS 2007, Bielefeld University eCollections, Germany, pp. 264–275 (2004)
12. Hawes, N., Wyatt, J.: Towards context-sensitive visual attention. In: Second International Cognitive Vision Workshop (ICVW 2006) (2006)
13. Tagare, H.D., Toyama, K., Wang, J.G.: A maximum-likelihood strategy for directing attention during visual search. Transactions on PAMI 23, 490–500 (2001)
14. Peters, R.J., Itti, L.: Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In: CVPR 2007, IEEE, Los Alamitos (2007)
15. Sun, Y., Fischer, R.: Object-based visual attention for computer vision. Artificial Intelligence 146, 77–123 (2003)
16. Backer, G., Mertsching, B., Bollmann, M.: Data- and model-driven gaze control for an active-vision system. Transactions on PAMI 23, 1415–1429 (2001)
17. Aziz, M.Z., Mertsching, B., Shafik, M.S., Stemmer, R.: Evaluation of visual attention models for robots. In: ICVS 2006, IEEE, New York (2006) index–20
18. Aziz, M.Z., Mertsching, B.: Color segmentation for a region-based attention model. In: 12. Workshop Farbbildverarbeitung (FWS 2006), pp. 74–83 (2006)
19. Aziz, M.Z., Mertsching, B.: Color saliency and inhibition in region based visual attention. In: WAPCV 2007, Hyderabad, India, pp. 95–108 (2007)
20. Aziz, M.Z., Mertsching, B.: Fast and robust generation of feature maps for region-based visual attention. In: IEEE Transactions on Image Processing (2008)
21. Aziz, M.Z., Mertsching, B.: Pop-out and IOR in static scenes with region based visual attention. In: WCAA-ICVS 2007, Bielefeld University eCollections (2007)
22. Aziz, M.Z., Mertsching, B.: Region-based top-down visual attention through fine grain color map. In: 13 Workshop Farbbildverarbeitung (FWS 2007), pp. 83–92 (2007)
23. Aziz, M.Z., Mertsching, B.: Color saliency and inhibition using static and dynamic scenes in region based visual attention. In: Attention in Cognitive Systems. LNCS (LNAI), vol. 4840, pp. 234–250 (2007)
24. Kutter, O., Hilker, C., Simon, A., Mertsching, B.: Modeling and Simulating Mobile Robots Environments. In: 3rd International Conference on Computer Graphics Theory and Applications (GRAPP 2008) (2008)