Goal-directed search with a top-down modulated computational attention system

Simone Frintrop[1], Gerriet Backer[2], and Erich Rome[1]

[1]Fraunhofer Institut für Autonome Intelligente Systeme (AIS),
Schloss Birlinghoven, 53754 Sankt Augustin, Germany
[2]Krauss Software GmbH, Tiefe Str. 1, 38162 Cremlingen, Germany

Abstract. In this paper we present VOCUS: a robust computational attention system for goal-directed search. A standard bottom-up architecture is extended by a top-down component, enabling the weighting of features depending on previously learned weights. The weights are derived from both target (excitation) and background properties (inhibition). A single system is used for bottom-up saliency computations, learning of feature weights, and goal-directed search. Detailed performance results for artificial and real-world images are presented, showing that a target is typically among the first 3 focused regions. VOCUS represents a robust and time-saving front-end for object recognition since by selecting regions of interest it significantly reduces the amount of data to be processed by a recognition system.

1 Introduction and State of the Art

Suppose you are looking for your key. You know it to be somewhere on your desk but it still takes several fixations until your roaming view hits the key. If you have a salient key fob contrasting with the desk, you will detect the key with fewer fixations. This is according to the separation of visual processing into two subtasks as suggested by Neisser [9]: first, a fast parallel pre-selection of scene regions detects object candidates and second, complex recognition restricted to these regions verifies or falsifies the hypothesis. This dichotomy of fast localization processes and complex, robust, but slow identification processes is highly effective: expensive resources are guided towards the most promising and relevant candidates.

In computer vision, the efficient use of resources is equally important. Although an attention system generates a certain overhead in computation, it pays off since reliable object recognition is a complex vision task that is usually computationally expensive. The more general the recognizer – for different shapes, poses, scales, and illuminations – the more important is a pre-selection of regions of interest.

Concerning visual attention, most research has so far been done in the field of *bottom-up* processing (in psychology [13, 15], neuro-biology [2, 10] and computer vision [7, 6, 1, 11]). Bottom-up attention is merely data-driven and finds regions that attract the attention automatically, e.g., a black sheep in a white flock. Koch

& Ullman [7] described the first explicit computational architecture for bottom-up visual attention; it is strongly influenced by Treisman's *feature-integration theory* [13]. Many computational systems have been presented meanwhile [6, 1, 11, 14], most restricted to bottom-up computations.

While much less analyzed, there is strong neurobiological and psychophysical evidence for top-down influences modifying early visual processing in the brain due to pre-knowledge, motivations, and goals [17, 2, 16]. However, only a few computational attention models integrate top-down information. The earliest approach is the *guided search* model by Wolfe [15], a result of his psychological investigations of human visual search. Tsotsos' system considers feature channels separately and uses inhibition for regions of a specified location or those that do not fit the target features [14]. Hamker performs visual search on selected images but without considering the target background [5]. The closest related work is presented by Navalpakkam et al. [8]; however, the region to learn is not determined automatically and exciting and inhibiting cues as well as bottom-up and top-down cues are not separated. Furthermore, quality and robustness of the system are not shown. To our knowledge, there exists no complete, well investigated system of top-down visual attention comparable to our approach.

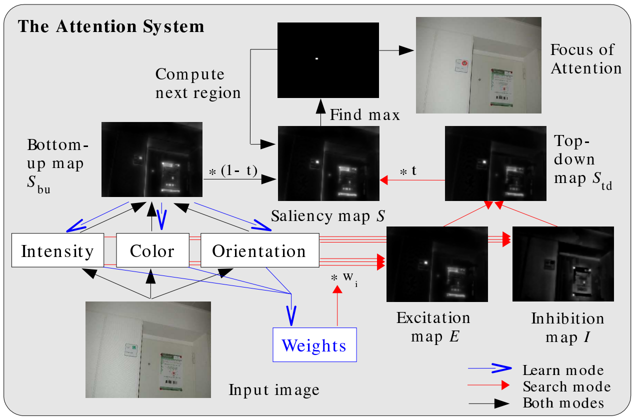
In this paper, we present the attention system VOCUS that performs goal-directed search by extending a well-known bottom-up system [6] by a top-down part. The bottom-up part computes saliencies for the features intensity, orientation, and color independently, weights maps according to the uniqueness of the feature, and finally fuses the saliencies into a single map. The top-down part uses previously learned weights to enable the search for targets. The weighted features contribute to a top-down saliency map highlighting regions with target-relevant features. The relative strengths of bottom-up and top-down influences are adjustable according to the task. Cues from both maps are fused into a global saliency map and the focus of attention is directed to its most salient region. The system shows good performance on artificial as well as on real-world data: typically, one of the first 3 selected regions contains the target, in many cases it is the first region. In the future, we will integrate the system into a robot control architecture enabling the detection of salient regions and goal-directed search.

2 The Visual Attention System VOCUS

The computational attention system VOCUS (Visual Object detection with a CompUtational attention System) consists of a bottom-up part computing data-driven saliency and a top-down part enabling goal-directed search. Global saliency is determined from both cues (cf. Fig. 1).

2.1 Bottom-up saliency

VOCUS' bottom-up part detects salient image regions by using image contrasts and uniqueness of a feature, e.g., a red ball on green grass. It was inspired by Itti et al. [6] but differs in several aspects resulting in considerably improved performance (see [3]). The feature computations are performed on 3 different scales



Feature	weights
intensity on/off	0.001
intensity off/on	9.616
orientation 0 °	4.839
orientation 45 °	9.226
orientation 90 °	2.986
orientation 135 °	8.374
color green	76.572
color blue	4.709
color red	0.009
color yellow	0.040
conspicuity I	6.038
conspicuity O	5.350
conspicuity C	12.312

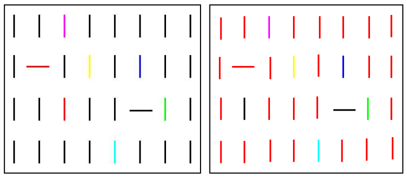
Fig. 1. The goal-directed visual attention system with a bottom-up part (left) and a top-down part (right). In learning mode, target weights are learned (blue arrows). These are used in search mode (red arrows). Right: weights for target name plate.

using image pyramids. The feature intensity is computed by *center-surround mechanisms* extracting intensity differences between image regions and their surroundings, similar to cells in the human visual system [10]. In contrast to [6], we compute on-off and off-on contrasts separately [3, 4]; after summing up the scales, this yields 2 intensity maps. Similar, 4 orientation maps (0 °, 45 °, 90 °, 135 °) are computed by Gabor filters and 4 color maps (green, blue, red, yellow) by first converting the RGB image into the Lab color space, second determining the distance of the pixel color to the prototype color (the red map shows high activations for red regions and small ones for green regions) and third, applying center-surround mechanisms. Each feature map X is weighted with the uniqueness weight $\mathcal{W}(X) = X/\sqrt{m}$, where m is the number of local maxima that exceed a threshold t. This weighting is essential since it emphasizes important maps with few peaks, enabling the detection of *pop-outs* (outliers). After weighting, the maps are summed up to the bottom-up saliency map S_{bu}.

2.2 Top-down saliency

To perform visual search, VOCUS first computes target-specific weights (learning mode) and, second, uses these weights to adjust the saliency computations according to the target (search mode). We call this target-specific saliency *top-down saliency.*

In **learning mode**, VOCUS is provided with a training image and coordinates of a *region of interest (ROI)* that includes the target. The region might be the output of a classifier specifying the target or determined manually by the user. Then, the system computes the bottom-up saliency map and the *most salient region (MSR)* inside the ROI. So, VOCUS is able to decide autonomously what is important in a ROI, concentrating on parts that are most salient and disregarding the background or less salient parts. Note that this makes VOCUS also robust to small changes of the ROI coordinates.



Feature	weights 1	weights 2
orientation 0 °	20.64	**29.84**
color red	**47.60**	10.29

Fig. 2. The weights for the target (red horizontal bar, 2nd in 2nd row) differ depending on the environment: in the left image (black vertical bars) color is more important than orientation (weights 1), in the other image (red vertical bars) vice versa (weights 2).

Next, weights are determined for the feature and conspicuity maps, indicating how important a feature is for the target. The weight w_i for map X_i is the ratio of the mean saliency in the target region $m_{(MSR)}$ and in the background $m_{(image-MSR)}$: $w_i = m_{(MSR)}/m_{(image-MSR)}$ where $i \in \{1, ..., 13\}$. This computation does not only consider which features are the strongest in the target region, it also regards which features separate the region best from the rest of the image (cf. Fig. 2).

The learning of weights from one single training image yields good results if the target object occurs in all test images in a similar way, i.e., on a similar background and in a similar orientation. These conditions occur if the objects are fixed elements of the environment, e.g. fire extinguishers. Nevertheless, for movable objects it is necessary to learn from several training images which features are stable and which are not. This is done by determining the average weights from n training images using the geometric mean of the weights, i.e., $w_{i,(1..n)} = \sqrt[n]{\prod_{j=1}^{n} w_{i,j}}$. Instead of using all images from the training set, we choose the most suitable ones: first, the weights from one training image are applied to the training set, next, the image with the worst detection results is taken and the average weights from both images are computed. This procedure is repeated iteratively as long as the performance increases (details in [3, 4]).

In **search mode**, we determine a top-down saliency map that is integrated with the bottom-up map to yield global saliency. The top-down map itself is composed of an excitation and an inhibition map. The excitation map E is the weighted sum of all feature and conspicuity maps X_i that are important for the learned region, i.e., $w_i > 1$. The inhibition map I shows the features more present in the background than in the target region, i.e., $w_i < 1$:

$$\begin{aligned} E &= \textstyle\sum_i (w_i * X_i) \qquad && \forall i : w_i > 1 \\ I &= \textstyle\sum_i ((1/w_i) * X_i) \qquad && \forall i : w_i < 1 \end{aligned} \tag{1}$$

The top-down saliency map S_{td} results from the difference of E and I and a clipping of negative values: $S_{td} = E - I$. To make S_{td} comparable to S_{bu}, it is normalized to the same range. I, E, and S_{td} are depicted in Fig. 3, showing that the excitation map as well as the inhibition map have an important influence.

The global saliency map S is the weighted sum of S_{bu} and S_{td}. The contribution of each map is adjusted by the top-down factor $t \in [0..1]$: $S = (1 - t) * S_{bu} + t * S_{td}$. For $t = 1$, VOCUS considers only target-relevant features (pure

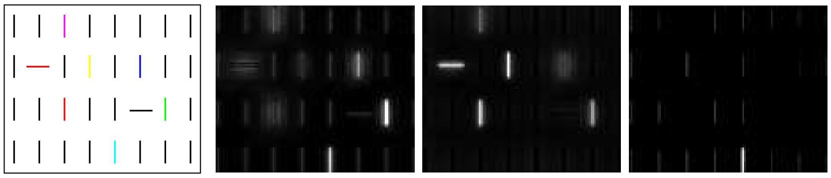


Fig. 3. Excitation and inhibition are both important: search target: cyan vertical bar (5th, last row). Left to right: test image, excitation map E, inhibition map I, top-down map S_{td}. E shows bright values for cyan, but brighter ones for the green bar (7th, 3rd row). Only the inhibition of the green bar enables the single peak for cyan in S_{td}.

top-down). For a lower t, salient bottom-up cues may divert the focus of attention, an important mechanism in human attention: a person suddenly entering a room immediately catches our attention. Also colored cues divert the search for non-colored objects as shown in [12]. Determining appropriate values for t depends on the system state, the environment and the current task; this is beyond the scope of this article and will be tackled when integrating our attention system into robotic applications.

After the computation of the global saliency map S, the most salient region is determined by *region growing* starting with the maximum of S. Finally, the focus of attention (FOA) is directed to this region. To compute the next FOA, this region is inhibited and the selection process is repeated.

3 Results

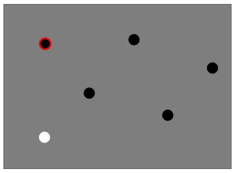
In this section, we present experimental results on artificial images to establish the link to human visual attention and on numerous real-world images. The quality of the search is given by the *hit number*, i.e., the number of the focus that hits the target (for several images the *average hit number*). Since we concentrate on computing the first n foci on a scene (usually $n = 10$), we additionally show the *detection rate*, i.e., the percentage of images in which the target was detected within the first n FOAs. Note that VOCUS works with the same parameters in all experiments; there is no target-specific adaptation.

Visual search in artificial images: first, VOCUS was trained on the image in Fig. 3 (left) to find different bars. Tab. 1 shows the hit number. The green, cyan, and yellow bar are not focused in bottom-up mode within the first 10 foci, since their saliency values are lower than those of the black vertical bars.

For $t = 1$, all targets are focused immediately with one exception (magenta vertical). Magenta has a lot of blue portions so the blue regions are also enhanced during search for magenta. This leads to focusing the blue before the magenta bar. Note that also black bars are found, considering the lack of color by inhibiting colored objects. For $t = 0.5$, bottom-up and top-down cues are both regarded. It shows that in most cases the hit number is the same as for $t = 1$, except for the red vertical bar. Here, the bottom-up saliency of the red horizontal bar diverts the focus. It looks as if the bottom-up cues have less influence than the top-down cues, but note that the saliency values in the bottom-up

	Hit number for several target bars (and saliency value)								
	red	blue	black	magenta	red	black	green	cyan	yellow
top-down factor	horiz	vert	horiz	vert	vert	vert	vert	vert	vert
t = 0.0	1 (24)	2 (23)	3 (21)	4 (18)	5 (17)	6 (15)	- (12)	- (8)	- (7)
t = 0.5	1	1	1	2	2	1	1	1	1
t = 1.0	1	1	1	2	1	1	1	1	1

Table 1. Search performance for different target bars for the training image of Fig. 3, left; test image is the horizontally flipped training image. The first 10 FOAs are computed. $t = 0$ is pure bottom-up search, $t = 1$ pure top-down search. The performance is given by the *hit number* on the target. The numbers in parentheses show the saliency value at the target region.



Intensity of background in %	10	20	30	40	50	60	70	80	90
Required t to override pop-out	0	0	0	0.3	0.4	0.5	0.6	0.9	-

Fig. 4. When searching for a black dot (left), the required value of the top-down factor t increases with rising background intensity (right).

saliency map (values in parentheses) do not differ a lot, so little influence by the top-down map is enough to change the order of the foci. In a *pop-out experiment* [15], the contrasts between target and distractor values are much larger.

To test the influence of the top-down factor t systematically, we use an image with one white and 5 black dots on a grey background (Fig. 4). We vary the intensity of the background between 10% (nearly white) and 90% (nearly black) and determine the top-down factor required to override the white pop-out and find a black dot. It shows that with increasing intensity, the value of t required to override the pop-out increases too, up to $t = 0.9$ for 80% intensity. For 90%, the pop-out is so strong that it cannot be overriden anymore.

Visual search in real-world images: In Tab. 2 and Fig. 5, we show some search results obtained from more than 1000 real-world images. We chose targets with high and with low bottom-up saliency, fixed at the wall as well as movable ones. The objects occupy about 1% of the image and occur in different contexts, with different distractors, and on differently colored and structured backgrounds.

We chose two kinds of targets fixed in an office environment: name plates and fire extinguishers. The largest test set was available for the name plates: we took 906 test and 54 training images of 3 different halls of our institute, showing about 50 different doors and name plates in different contexts due to differently colored and structured posters and photos at walls and doors. The movable objects are a key fob and a highlighter. The highlighter was placed on two different desks: a dark and a bright one. For each kind of target, one was highly salient by itself (fire extinguishers and highlighters are designed to attract attention), while the bottom-up saliency for the name plates and the key fob was much smaller.

Target	# test	Average hit number and detection rate [%]		
	im.	t = 0 (d.r.)	t = 0.5 (d.r.)	t = 1 (d.r.)
fire extinguisher	46	2.69 (94%)	1.09 (100%)	1.06 (100%)
key fob	28	4.42 (80%)	1.27 (100%)	1.23 (100%)
name plate	906	3.94 (48%)	2.48 (85%)	2.06 (89%)
highlighter	60	2.54 (90%)	1.73 (98%)	1.48 (100%)

Table 2. Search performance on real world data. Left: the targets and the extracted region for learning. Note that this is not the training data, training images contain a whole scene instead of an extracted object. The training images are chosen from training sets of 10 to 40 images with an algorithm described in [3]. For each target, 10 FOAs are computed. The table shows the average hit number for different top-down factors *t* and, in parentheses, the detection rate (d.r.) within 10 FOAs.

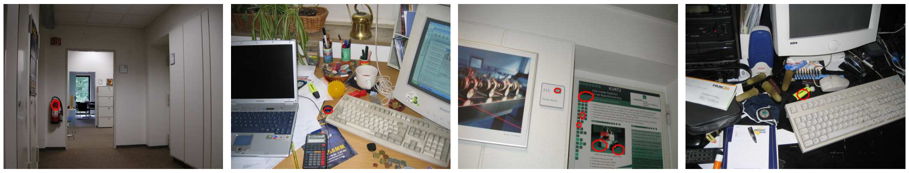


Fig. 5. Some of the results from Tab. 2. Search for a a fire extinguisher, a key fob, a name plate, and a highlighter. The FOAs are depicted by red ellipses. All targets detected with 1st FOA, only in the 3rd image with the 6th FOA.

Tab. 2 shows that the performance depends on the kind of target and on its environment. We identified 3 scenarios: 1) The object is very salient and often detected at once in bottom-up mode, e.g., fire extinguisher and highlighter. Here, VOCUS is also very successful: the target is in average detected with the 1st or 2nd FOA. 2) The object is not very salient so the bottom-up value is low, but there are few regions in the scene with similar features. This enables VOCUS to separate the target well from the environment, resulting in a high detection rate (e.g. the key fob). 3) The target is not salient and there are a lot of similar regions in the environment that divert the focus. Even for humans, these are difficult conditions for visual search. An example is the name plate: in some of the test images, there are posters on walls or doors with colors similar to the logo of the name plate making the search difficult (cf. Fig. 5).

The results show convincingly that VOCUS is successful in finding targets. Salient objects are nearly always detected with the 1st or 2nd FOA, and even in difficult settings, the amount of regions to be investigated by a classifier is drastically reduced. The results also reveal that the system is highly robust: the images were taken under natural conditions, i.e., illumination and viewpoint of the target vary (details in [3]). Most difficult are images with several regions having nearly the same saliency values as the target; there, small variations of the image data may lead to a changed order of the foci. We also compared VOCUS to Itti's attention system NVT on the data presented in [8]; in [3] we showed that VOCUS clearly outperforms the NVT.

4 Conclusion

We introduced the new attention system VOCUS that uses previously learned target information to perform goal-directed search. During learning, it considers not only the properties of the target, but also of the background. In search mode, bottom-up and top-down influences compete for global saliency. The biologically plausible system has been thoroughly tested on artificial and real-world data and its suitability for detecting different targets was extensively demonstrated. In our examples, the target usually was among the first three selected regions.

In future work, we plan to utilize the system for robot control. Therefore, first the runtime (currently 1.7 sec on a 1.7 GHz Pentium IV for 300×300 pixel images) has to be improved to enable real-time performance. Then, VOCUS will determine salient regions in the robot's environment and search for previously learned objects. Directing the attention to regions of potential interest will be the basis for efficient object detection and manipulation.

References

1. Backer, G., Mertsching, B. and Bollmann, M. Data- and model-driven Gaze Control for an Active-Vision System. *IEEE Trans. on PAMI* **23(12)** (2001) 1415–1429.
2. Corbetta, M. and Shulman, G. L. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews* **3** (3, 2002) 201–215.
3. Frintrop, S. VOCUS: A Visual Attention System for Object Detection and Goal-directed Search. PhD thesis University of Bonn Germany (to appear 2005).
4. Frintrop, S., Backer, G. and Rome, E. Selecting what is Important: Training Visual Attention. In: Proc. of KI (accepted), LNCS, Springer (2005).
5. Hamker, F. Modeling Attention: From computational neuroscience to computer vision. In: Proc. of WAPCV'04 (2004) 59–66.
6. Itti, L., Koch, C. and Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. on PAMI* **20** (11, 1998) 1254–1259.
7. Koch, C. and Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology* **4** (4, 1985) 219–227.
8. Navalpakkam, V., Rebesco, J. and Itti, L. Modeling the influence of task on attention. *Vision Research* **45** (2, 2005) 205–231.
9. Neisser, U. *Cognitive Psychology* Appleton-Century-Crofts New York 1967.
10. Palmer, S. E. *Vision Science, Photons to Phenomenology* The MIT Press 1999.
11. Sun, Y. and Fisher, R. Object-based visual attention for computer vision. *Artificial Intelligence* **146** (1, 2003) 77–123.
12. Theeuwes, J. Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review* **11** (2004) 65–70.
13. Treisman, A. M. and Gelade, G. A feature integration theory of attention. *Cognitive Psychology* **12** (1980) 97–136.
14. Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N. and Nuflo, F. Modeling Visual Attention via Selective Tuning. *AI* **78** (1-2, 1995) 507–545.
15. Wolfe, J. Guided Search 2.0: A Revised Model of Visual Search. *Psychonomic Bulletin & Review* **1** (2, 1994) 202–238.
16. Wolfe, J. M., Horowitz, T., Kenner, N., Hyle, M. and Vasan, N. How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research* **44** (2004) 1411–1426.
17. Yarbus, A. L. *Eye Movements and Vision* Plenum Press (New York) 1969.