TARGET DETECTION USING SALIENCY-BASED ATTENTION

Laurent Itti and Christof Koch
Computation and Neural Systems Program
California Institute of Technology
Mail-Code 139-74 - Pasadena, CA 91125 - U.S.A.
Email: {itti, koch}@klab.caltech.edu

1. SUMMARY

Most models of visual search, whether involving overt eye movements or covert shifts of attention, are based on the concept of a "saliency map", that is, an explicit two-dimensional map that encodes the saliency or conspicuity of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the next attended target. Inhibiting this location automatically allows the system to attend to the next most salient location. We describe a detailed computer implementation of such a scheme, focusing on the problem of combining information across modalities, here orientation, intensity and color information, in a purely stimulus-driven manner. We have successfully applied this model to a wide range of target detection tasks, using synthetic and natural stimuli. Performance has however remained difficult to objectively evaluate on natural scenes, because no objective reference was available for comparison. We here present predicted search times for our model on the Search2 database of rural scenes containing a military vehicle. Overall, we found a poor correlation between human and model search times. Further analysis however revealed that in 3/4 of the images, the model appeared to detect the target faster than humans (for comparison, we calibrated the model's arbitrary internal time frame such that no more than 2-4 image locations were visited per second). It hence seems that this model, which had originally been designed not to find small, hidden military vehicles, but rather to find the few most obviously conspicuous objects in an image, performed as an efficient target detector on the Search2 dataset.

Keywords: Visual attention, saliency, preattentive, inhibition of return, model, winner-take-all, bottom-up, natural scene.

2. INTRODUCTION

Biological visual systems are faced with, on the one hand, the need to process massive amounts of incoming information (estimated at around 10^8 bits per second in the optic nerve of humans), and on the other hand, the requirement for nearly real-time capacity of reaction.

Surprisingly, instead of employing a purely parallel image analysis approach, primate vision systems appear to employ a serial computational strategy when inspecting complex visual scenes. Particular locations are selected based on their behavioral relevance or on local image cues. The identification of objects and the analysis of their spatial relationship usually involve either rapid, saccadic eye movements to bring the fovea onto the object, or covert shifts of attention. It consequently appears that the incredibly difficult problem of full-field image analysis and scene understanding is taken on by biological visual systems through a temporal serialization into smaller, localized analysis tasks.

Much evidence has accumulated in favor of a two-component framework for the control of where in a visual scene attention is focused to [1,2,3,4]: A bottom-up, fast, primitive mechanism that biases the observer towards selecting stimuli based on their "saliency" (most likely encoded in terms of center-surround mechanisms) and a second slower, top-down mechanism with variable selection criteria, which directs the "spotlight of attention" under cognitive, volitional control.

Koch and Ullman [5] introduced the idea of a saliency map to accomplish preattentive selection (see also the concept of a "master map" in [6]). This is an explicit two-dimensional map that encodes the saliency of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the most salient object, which constitutes the next target. If this location is subsequently inhibited, the system automatically shifts to the next most salient location, endowing the search process with internal dynamics.

We here describe a computer implementation of a preattentive selection mechanism based on the architecture of the primate visual system. We address the thorny problem of how information from different modalities - in the case treated here from 42 maps encoding intensity, orientation and color in a center-surround fashion at a number of spatial scales - can be combined into a single saliency map. Our algorithm qualitatively reproduces human performance on a number of classical search experiments.

Vision algorithms frequently fail when confronted with realistic, cluttered images. We therefore studied the performance of our search algorithm using high-resolution (6144x4096 pixels) photographs containing images of military vehicles in a complex rural background (Search2 dataset). Our algorithm shows, on average, superior performance compared to human observers searching for the same targets, although our system does not yet include any top-down task-dependent tuning.

3. THE MODEL

The model has been presented in more details in [8] and is only briefly described here (Fig. 1).

Input is provided in the form of digitized color images. Different spatial scales are created using Gaussian pyramids [7], which consist of progressively low-pass filtering and subsampling the input image. Pyramids have a depth of 9 scales, providing horizontal and vertical image reduction factors ranging from 1:1 (scale 0; the original input image) to 1:256 (scale 8) in consecutive powers of two. Each feature is computed by center-surround operations akin to visual receptive fields, implemented as differences between a fine and a coarse scale: the center of the receptive field corresponds to a pixel at scale c={2, 3, 4} in the pyramid, and the surround to the corresponding pixel at scale s=c+d, with d={3, 4}, yielding six feature maps for each type of feature. The differences between two images at different scales are obtained by oversampling

Paper presented at the RTO SCI Workshop on "Search and Target Acquisition", held in Utrecht, The Netherlands, 21-23 June 1999, and published in RTO MP-45.

the image at the coarser scale to the resolution of the image at the finer scale.

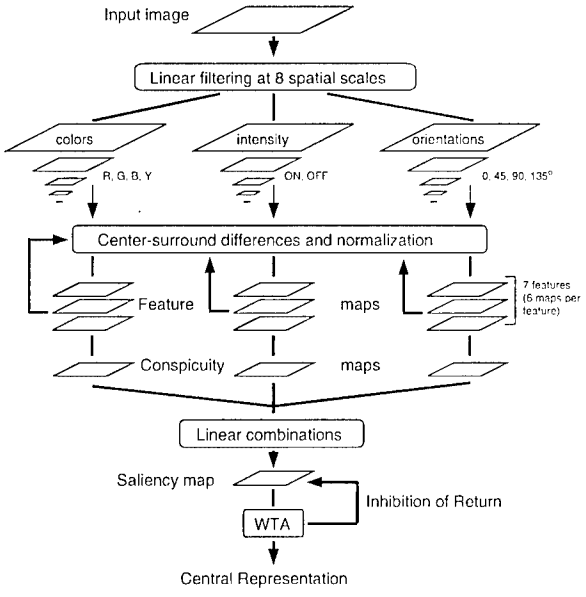

Figure 1: General architecture of the model. Low-level visual features are extracted in parallel from nine spatial scales, using a biological center-surround architecture. The resulting 42 feature maps are combined to yield three conspicuity maps for color, intensity and orientation. These, in turn, feed into a single saliency map, consisting of a 2D layer of integrate-and-fire neurons. A neural winner-take-all network shifts the focus of attention to the currently most salient image location. Feedback inhibition then transiently suppresses the currently attended location, causing the focus of attention to shift to the next most salient image location.

3.1. Extraction of early visual features

With r, g and b being the red, green and blue channels of the input image, an intensity image I is obtained as I=(r+g+b)/3. From I is created a Gaussian pyramid I(s), where s={0..8} is the scale. The r, g and b channels are normalized by I, at the locations where the intensity is at least 10% of its maximum, in order to decorrelate hue from intensity. Four broadly tuned color channels are created: R=r-(g+b)/2 for red, G=g-(r+b)/2 for green, B=b-(r+g)/2 for blue, and Y=(r+g)/2-|r-g|/2-b for yellow (negative values are set to zero). Four Gaussian pyramids R(s), G(s), B(s) and Y(s) are created from these color channels. From I, four orientation-selective pyramids are also created using Gabor filtering at 0, 45, 90 and 135 degrees.

Differences between a "center" fine scale c and a "surround" coarser scale s yield six feature maps for each of intensity contrast, red-green double opponency, blue-yellow double opponency, and the four orientations. A total of 42 feature maps is thus created, using six pairs of center-surround scales in seven types of features.

3.2. The saliency map

The task of the saliency map is to compute a scalar quantity representing the salience at every location in the visual field, and to guide the subsequent selection of attended locations. The feature maps provide the input to the saliency map, which is modeled as a neural network receiving its input at scale 4.

3.2.1. *Fusion of information*

One difficulty in combining different feature maps is that they represent a priori not comparable modalities, with different dynamic ranges and extraction mechanisms. Also, because a total of 42 maps are combined, salient objects appearing strongly in only a few maps risk to be masked by noise or less salient objects present in a larger number of maps.

Previously, we have shown that the simplest feature combination scheme - to normalize each feature map to a fixed dynamic range, and then sum all maps - yields very poor detection performance for salient targets in complex natural scenes [9]. One possible way to improve performance is to learn linear map combination weights, by providing the system with examples of targets to be detected. While performance improves greatly, this method presents the disadvantage of yielding different specialized models (that is, sets of map weights) for each target detection task studied [9].

When no top-down supervision is available, we propose a simple normalization scheme, consisting of globally promoting those feature maps in which a small number of strong peaks of activity (conspicuous locations) is present, while globally suppressing feature maps which contain comparable peak responses at numerous locations over the visual scene. This "within-feature competitive" scheme coarsely ressembles non-classical inhibitory interactions which have been observed electrophysiologically [10].

The specific implementation of these interactions in our model has been described elsewhere [9] and can be summarized as follows (Fig. 2): Each feature map is first normalized to a fixed dynamic range (between 0 and 1), in order to eliminate feature-dependent amplitude differences due to different feature extraction mechanisms. Each feature map is then iteratively convolved by a large 2-D Derivative-of-Gaussians (DoG) filter. The DoG filter, a section of which is shown in Fig. 2, yields strong local excitation at each visual location, which is counteracted by broad inhibition from neighboring locations. At each iteration, a given feature map receives input from the preattentive feature extraction stages described above, to which results of the convolution by the DoG are added. All negative values are then rectified to zero, thus making the iterative process highly non-linear. This procedure is repeated for 10 iterations.

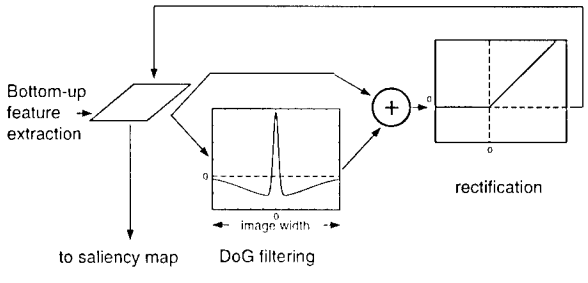

Figure 2: Illustration of the spatial competition for salience implemented within each of the 42 feature maps. Each map receives input from the linear filtering and center-surround stages. At each step of the process, the convolution of the map by a large Difference-of-Gaussians (DoG) kernel is added to the current contents of the map. This additional input coarsely models short-range excitatory processes and long-range inhibitory interactions between neighboring visual locations. The map is half-wave rectified, such that negative values are eliminated, hence making the iterative process non-linear. Ten iterations of the process are carried out before the output of each feature map is used in building the saliency map.

The choice of the number of iterations is somewhat arbitrary: In the limit of an infinite number of iterations, any non-empty map will converge towards a single peak, hence constituting only a poor representation of the scene. With very few iterations however, spatial competition is very weak and inefficient. Two examples showing the time evolution of this process are shown in Fig. 3, and illustrate that using of the order of 10 iterations yields adequate distinction between the two example images shown. As expected, feature maps with initially numerous peaks of similar amplitude are suppressed by the interactions, while maps with one or a few initially stronger peaks become enhanced. It is interesting to note that this within-feature spatial competition scheme resembles a "winner-take-all" network with localized inhibitory spread, which allows for a sparse distribution of winners across the visual scene.

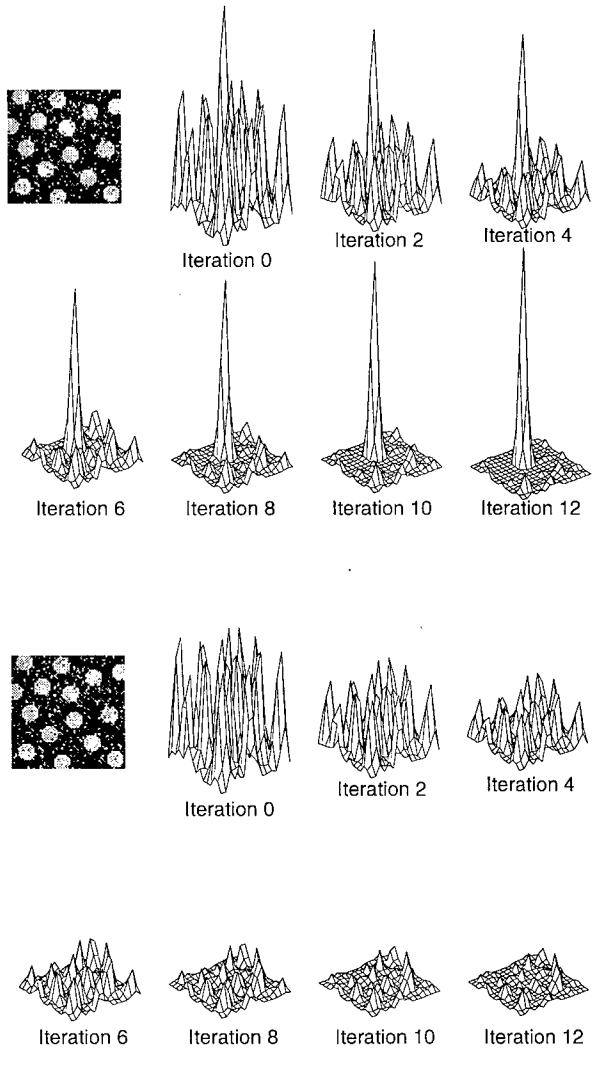

Figure 3: Example of operation of the long-range iterative competition for salience. When one (or a few) locations elicit stronger responses, they inhibit more the other locations than they are inhibited by these locations; the net result after a few iterations is an enhancement of the initially stronger location(s), and a suppression of the weaker locations. When no location is clearly stronger, all locations send and receive approximately the same amount of inhibition; the net result in this case is that all locations progressively become inhibited, and the map is globally suppressed.

After normalization, the feature maps for intensity, color, and orientation are summed across scales into three separate "conspicuity maps", one for intensity, one for color and one for orientation (Fig. 1).

Each conspicuity map is then subjected to another 10 iterations of the iterative normalization process. The motivation for the creation of three separate channels and their individual normalization is the hypothesis that similar features compete strongly for salience, while different modalities contribute independently to the saliency map. Although we are not aware of any supporting experimental evidence for this hypothesis, this additional step has the computational advantage of further enforcing that only a spatially sparse distribution of strong activity peaks is present within each visual feature, before combination of all three features into the scalar saliency map.

3.2.2. Internal Dynamics And Trajectory Generation

By definition, at any given time, the maximum of the saliency map's neural activity is at the most salient image location, to which the focus of attention (FOA) should be directed. This maximum is detected by a winner-take-all (WTA) network inspired from biological architectures [5]. The WTA is a 2D layer of integrate-and-fire neurons with a much faster time constant than those in the saliency map, and with strong global inhibition reliably activated by any neuron in the layer. In order to create dynamical shifts of the FOA, rather than permanently attending to the initially most salient location, it is necessary to transiently inhibit, in the saliency map, a spatial neighborhood of the currently attended location. This also prevents the FOA from immediately coming back to a strong, previously attended location. Such an "inhibition of return" mechanism has been demonstrated in humans [11]. Therefore, when a winner is detected by the WTA network, it triggers three mechanisms (Fig. 4):

1) The FOA is shifted so that its center is at the location of the winner neuron;

2) The global inhibition of the WTA is triggered and completely inhibits (resets) all WTA neurons;

3) Inhibitory conductances are transiently activated in the saliency map, in an area corresponding to the size and new location of the FOA. In order to slightly bias the model to next jump to salient locations spatially close to the currently attended location, small excitatory conductances are also transiently activated in a near surround of the FOA in the saliency map ("proximity preference" rule proposed by Koch and Ullman [5]).

Since we do not model any top-down mechanism, the FOA is simply represented by a disk whose radius is fixed to one twelvth of the smaller of the input image width or height. The time constants, conductances, and firing thresholds of the simulated neurons are chosen so that the FOA jumps from one salient location to the next in approximately 30-70ms (simulated time), and so that an attended area is inhibited for approximately 500-900ms, as it has been observed psychophysically [11]. The difference in the relative magnitude of these delays proved sufficient to ensure thorough scanning of the image by the FOA and prevent cycling through a limited number of locations.

Fig. 4 demonstrates the interacting time courses of two neurons in the saliency map and the WTA network, for a very simple stimulus consisting of one weaker and one stronger pixels in an otherwise empty map.

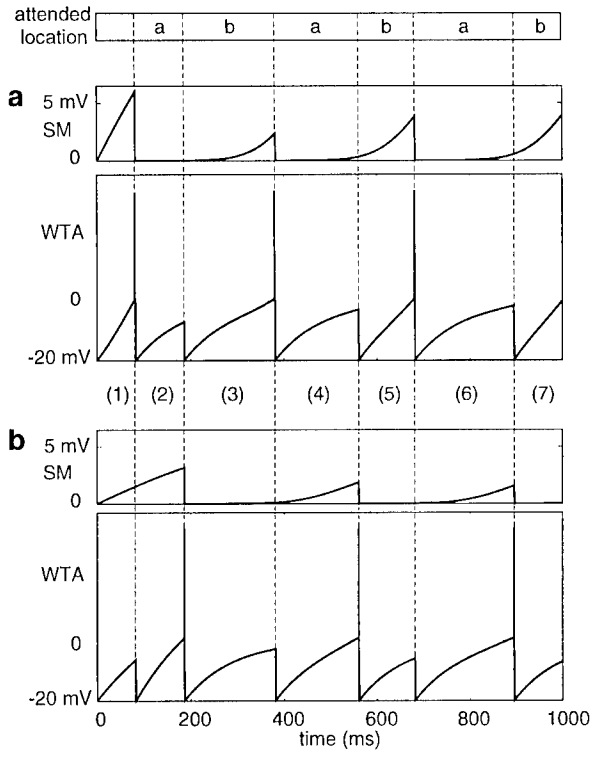

Figure 4: Dynamical evolution of the potential of some simulated neurons in the saliency map (SM) and in the winner-take-all (WTA) networks. The input contains one salient location (a), and another input of half the saliency (b); the potentials of the corresponding neurons in the SM and WTA are shown as a function of time. During period (1), the potential of both SM neurons (a) and (b) increases as a result of the input. The potential in the WTA neurons, which receive inputs from the corresponding SM neurons but have much faster time constants, increases faster. The WTA neurons evolve independently of each other as long as they are not firing. At about 80ms, WTA neuron (a) reaches threshold and fires. A cascade of events follows: First, the focus of attention is shifted to (a); second, both WTA neurons are reset; third, inhibition-of-return (IOR) is triggered, and inhibits SM neuron (a) with a strength proportional to that neuron's potential (i.e., more salient locations receive more IOR, so that all attended locations will recover from IOR in approximately the same time). In period (2), the potential of WTA neuron (a) rises at a much slower rate, because SM neuron (a) is strongly inhibited by IOR. WTA neuron (b) hence reaches threshold first. (3)-(7): In this example with only two active locations, the system alternatively attends to (a) and (b). Note how the IOR decays over time, allowing for each location to be attended several times. Also note how the amount of IOR is proportional to the SM potential when IOR is triggered (e.g., SM neuron (a) receives more IOR at the end of period (1) than at the end of period (3)). Finally, note how the SM neurons do not have an opportunity to reach threshold (at 20 mV) and to fire (their threshold is ignored in the model). Since our input images are noisy, we did not explicitly incorporate noise into the neurons' dynamics.

4. RESULTS

4.1. General performance

We tested our model on a wide variety of real images, ranging from natural outdoor scenes to artistic paintings. All images were in color, contained significant amounts of noise, strong local variations in illumination, shadows and reflections, large numbers of "objects" often partially occluded, and strong textures. Most of these images can be interactively examined on the World-Wide-Web, at:
http://www.klab.caltech.edu/~itti/attention/

Overall, the results indicate that the system scans the image in an order which makes functional sense in most behavioral situations.

It should be noted however that it is not straightforward to establish objective criteria for the performance of the system with such images. Unfortunately, nearly all quantitative psychophysical data on attentional control are based on synthetic stimuli. In addition, although the scan paths of overt attention (eye movements) have been extensively studied [12], it is unclear to what extent the precise trajectories followed by the attentional spotlight are similar to the motion of covert attention. Most probably, the requirements and limitations (e.g., spatial and temporal resolutions) of the two systems are related but not identical [13].

Although our model is mostly concerned with shifts of covert attention, and ignores all of the mechanistic details of eye movements, we attempt below a quantitative comparison between human and model target search times in complex natural scenes, using the Search2 database of images containing military vehicles hidden in a rural environment.

4.2. Search2 results

We propose a difficult test of the model using the Search2 dataset, in which target detection is evaluated using a database of complex natural images, each containing a military vehicle (the "target"). Contrary to our previous study with a simplified version of the model [8], which used low-resolution image databases with relatively large targets (typically about 1/10th the width of the visual scene), this study uses very-high resolution images (6144x4096 pixels), in which targets appear very small (typically 1/100th the width of the image). In addition, in the present study, search time is compared between the model's predictions and the average measured search times from 62 normal human observers [14].

4.2.1. Experimental setup

The 44 original photographs were taken during a DISSTAF (Distributed Interactive Simulation, Search and Target Acquisition Fidelity) field test in Fort Hunter Liggett, California, and were provided to us, along with all human data, by the TNO Human Factors Research Institute in the Netherlands [14]. The field of view for each image is 6.9x4.6 deg. Each scene contained one of nine possible military vehicles, at a distance ranging from 860 to 5822 meters from the observer. Each slide was digitized at 6144x4096 pixels resolution. Sixty two human observers aged between 18 and 45 years and with visual acuity better than 1.25 arcmin^-1 participated to the experiment (about half were women and half men).

Subjects were first presented with 3 close-up views of each of the 9 possible target vehicles, followed by a test run of 10 trials. A Latin square design [14] was then used for the randomized presentation of the images. The slides were projected such that they subtended 65x46 deg visual angle to the observers (corresponding to a linear magnification by about a factor 10 compared to the original scenery). During each trial, observers pressed a button as soon as they had detected the target, and subsequently indicated at which location on a 10x10 projected grid they had found the target. Further details on these experiments can be found in [14].

The model was presented with each image at full resolution. Contrary to the human experiment, no close-ups or test trials were presented to the model. The generic form of the model described above was used, without any specific parameter adjustment for this experiment. Simulations for up to 10,000 ms. of simulated time (about 200-400 attentional shifts) were done on a Digital Equipment Alpha 500 workstation. With these high-resolution images, the model comprised about 300 million simulated neurons. Each image was processed in about 15 minutes with a peak memory usage of 484 megabytes (for comparison, a 640x480 scene was typically processed in 10 seconds, and processing time approximately scaled linearly with the number of pixels). The focus of attention (FOA) was represented by a disk of radius 340 pixels (Figs. 5, 6, 7). Full coverage of the image by the FOA would hence require 123 shifts (with overlap); a random search would thus be expected to find the target after 61.5 shifts on average. The target was considered detected when the focus of attention intersected a binary mask representing the outline of the target, which was provided with the images. Three examples of scenes and model trajectories are presented in Figs. 5, 6, and 7. In the one image, the target was immediately found by the model, in another, a serial search was necessary before the target could be found, and in the last, the model failed to find the target.

4.2.2. Simulation results

The model immediately found the target (first attended location) in seven of the 44 images. It quickly found the target (fewer than 20 shifts) in another 23 images. It found the target after more than 20 shifts in 11 images, and failed to find the target in 3 images. Overall, the model consequently performed surprisingly well, with a number of attentional shifts far below the expected 61.5 shifts of a random search in all but 6 images. In these 6 images, the target was extremely small (and hence not conspicuous at all), and the model cycled through a number of more salient locations.

4.2.3. Tentative comparison to human data

The following analysis was performed to generate the plot presented in Fig. 8: First, a few outlier images were discarded, when either the model did not find the target within 2000ms of simulated time (about 40-80 shifts; 6 images), or when half or more of the humans failed to find the target (3 images), for a total of 8 discarded images. An average of 40ms per model shift was then derived from the simulations, and an average of 3 overt shifts per second was assumed for humans, hence allowing us to scale the model's simulated time to real time. An additional 1.5 second was then added to the model time to account for human motor response time. With such calibration, the fastest reaction times for both model and humans were approximately 2 seconds, and the slowest approximately 15 seconds, for the 36 images analyzed.

The results plotted in Fig. 8 overall show a poor correlation between human and model search times. Surprisingly however, the model appeared to find the target faster than humans in 3/4 of the images (points below the diagonal), despite the rather conservative scaling factors used to compare model to human time. In order to make the model faster than humans in no more than half of the images, one would have to assume that humans shifted their gaze not faster than twice per second, which seems unrealistically slow under the circumstances of a speeded search task on a stationary, non-masked scene. Even if eye movements were that slow, most probably would humans still shift covert attention at a much faster rate between two overt fixations.

4.2.4. Comparison to spatial frequency content models

In our previous studies with this model, we have shown that the within-feature long-range interactions are one of the key aspects of the model. In order to illustrate this point, we can compute a simple measure of local spatial fequency content (SFC) at each location in the input image, and compare this measure to our saliency map.

It could indeed be argued that the preattentive, massively parallel feature extraction stages in our model constitute a simple set of spatially and chromatically bandpass filters. A possibly much simpler measure of "saliency" could hence be based on a more direct measure of power or of amplitude in different spatial and chromatic frequency bands. Such simpler measure has been supported by human studies, in which local spatial frequency content (measured by Haar wavelet transform) was higher at the points of fixations during free viewing than on average, over the entire visual scene (see [8] for details).

We illustrate in Fig. 9, with one representative example image, that our measure of saliency actually differs greatly from a simple measure of SFC. The SFC was computed as shown previously [8], by taking the average amplitude of non-negligible FFT coefficients computed for the luminance channel as well as the red, green, blue and yellow channels.

While the SFC measure shows strong responses at numerous locations, e.g., at all locations with sharp edges, the saliency map contains a much sparser representation of the scene, where only locally unique such regions are preserved.

5. DISCUSSION

We have demonstrated that a relatively simple processing scheme, based on some of the key organizational principles of pre-attentive early visual cortical architectures (center-surround receptive fields, non-classical within-feature inhibition, multiple maps) in conjunction with a single saliency map performs remarkably well at detecting salient targets in cluttered natural and artificial scenes.

Key properties of our model, in particular its usage of inhibition-of-return and the explicit coding of saliency independent of feature dimensions, as well as its behavior on some classical search tasks, are in good qualitative agreement with the human psychophysical literature.

Using reasonable scaling of model to human time, we found that the model appeared to find the target faster than humans in 75% of the 36 images studied. One paradoxical explanation for this superior performance might be that top-down influences play a significant role in the deployment of attention in natural scenes. Top-down cues in humans might indeed bias the attentional shifts, according to the progressively constructed mental representation of the entire scene, in inappropriate ways. Our model lacks any high-level knowledge of the world and operates in a purely bottom-up manner.

This does suggest that for certain (possibly limited) scenarios, such high-level knowledge might interfere with optimal performance. For instance, human observers are frequently tempted to follow roads or other structures, or may consciously decide to thoroughly examine the surroundings of salient buildings that have popped-out, while the vehicle might be in the middle of a field or in a forest.

Although our model was not originally designed to detect military vehicles, our results also suggest that these vehicles where fairly "salient", according to the measure of saliency implemented in the model. This is also surprising, since one would expect such vehicles to be designed **not** to be salient.

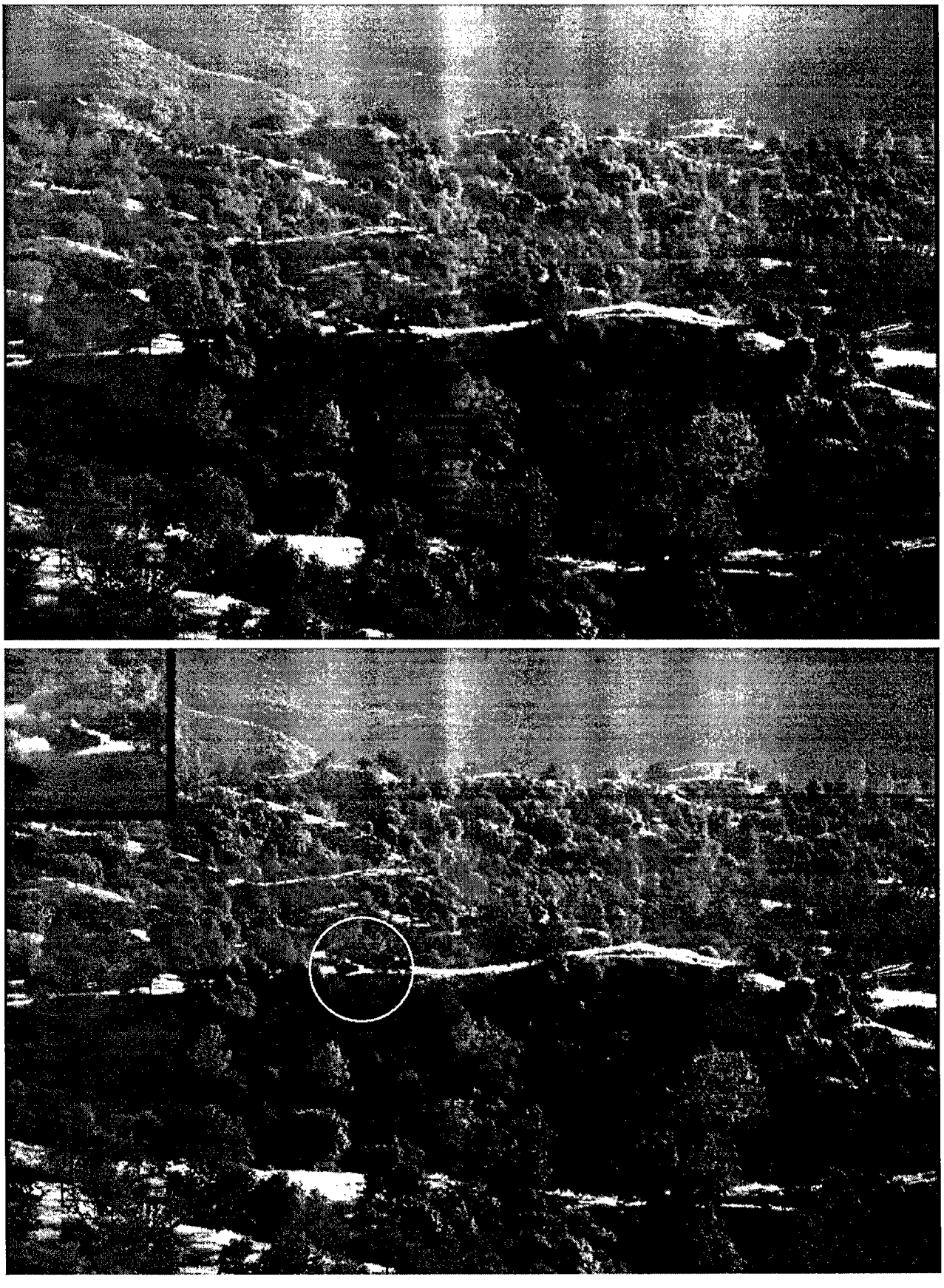

Figure 5: Example of image from the Search2 dataset (image 0018). The algorithm operated on the 24-bit color image. Top: original image; humans found the target in 2.8 sec on average. Bottom: model prediction: the target was the first attended location. After scaling of model time such that two to four attentional shifts occurred each second on average, and addition of 1.5 sec to account for latency in human motor response, the model found the target in 2.2 sec.

Figure 6: A more difficult example of image from the Search2 dataset (image 0019). Top: original image; humans found the target in 12.3 sec on average. Bottom: model prediction; because of its low contrast to the background, the target had lower saliency than several other objects in the image, such as buildings. The model hence initiated a serial search and found the target as the 10^{th} attended location, after 4.9 sec (using the same time scaling as in the previous figure).

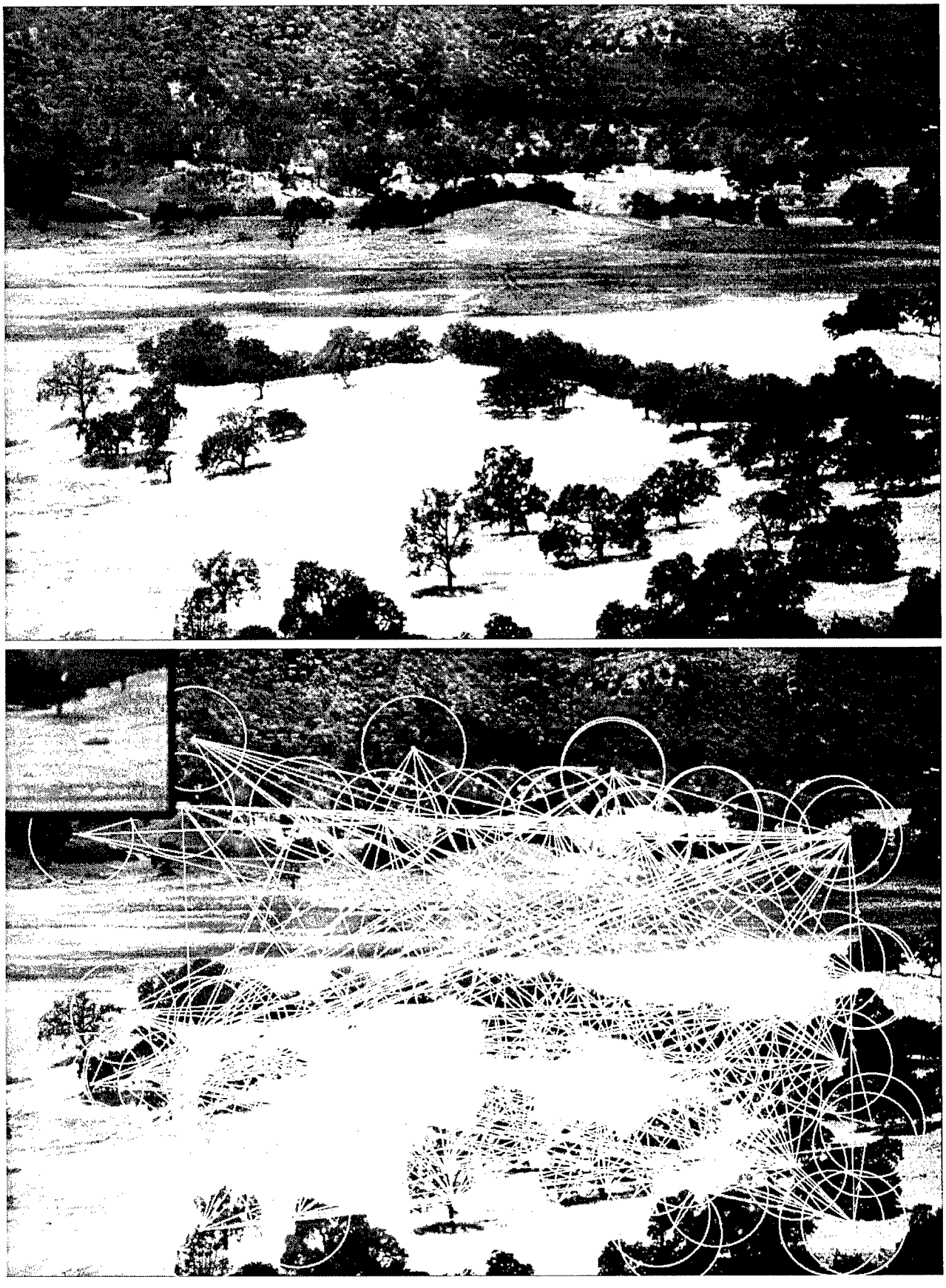

Figure 7: Example of image from the Search2 dataset (image 0024) in which the model did not find the target. Top: original image; humans found the target in 8.0 sec on average. Bottom: model prediction; the model failed to find the target, whose location is indicated by the white arrow. Inspection of the feature maps revealed that the target yielded responses in the different feature dimensions which are very similar to other parts of the image (foliage and trees). The target was hence not considered salient at all.

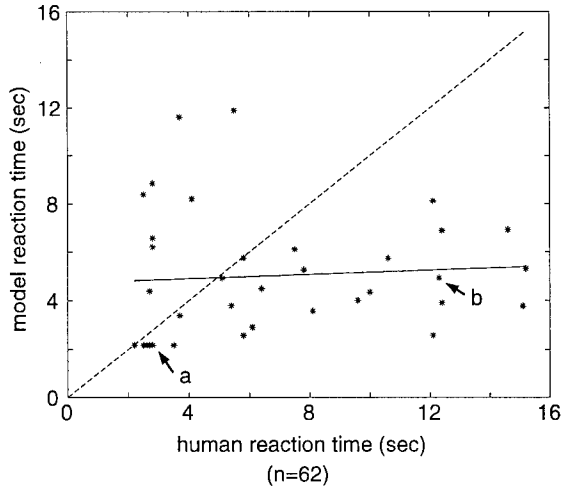

Figure 8. Mean reaction time to detect the target for 62 human observers and for our deterministic algorithm. Eight of the 44 original images are not included, in which either the model or the humans failed to reliably find the target. For the 36 images studied, and using the same scaling of model time as in the previous two figures, the model was faster than humans in 75% of the images. In order to bring this performance down to 50% (equal performance for humans and model), one would have to assume that no more than two visual locations can be visited each second. Arrow (a) indicates the "pop-out" example of Fig. 5, and arrow (b) the more difficult example presented in Fig. 6.

Looking at the details of individual feature maps, we realized that in most cases of quick detection of the target by the model, the vehicle was salient due to a strong, spatially isolated peak in the intensity or orientation channels. Such peak usually corresponded to the location of a specular reflection of sunlight onto the vehicle. Specular reflections were very rare at other locations in the images, and hence were determined to pop-out by the model. Because these reflections were often associated with locally rich SFC, and because many other locations also showed rich SFC, the SFC map could not detect them as reliably. Because these regions were spatially unique in one type of feature, they however popped-out for our model. Our model would hence have shown much poorer performance if the vehicles had not been so well polished.

6. CONCLUSION

In conclusion, our model yielded respectable results on the Search2 dataset, especially considering the fact that no particular adjustment was made to the model's parameters in order to optimize its target detection performance.

One important issue which needs to be addressed however is that of the poor correlation between model and human search times. We hypothesized in this study that top-down, volitional attentional bias might actually have hurt humans with this particular dataset, because trying to understand the scene and to willfully follow its structure was of no help in finding the target. A verification of this hypothesis should be possible once the scanpaths of human fixations during the search become available for the Search2 dataset.

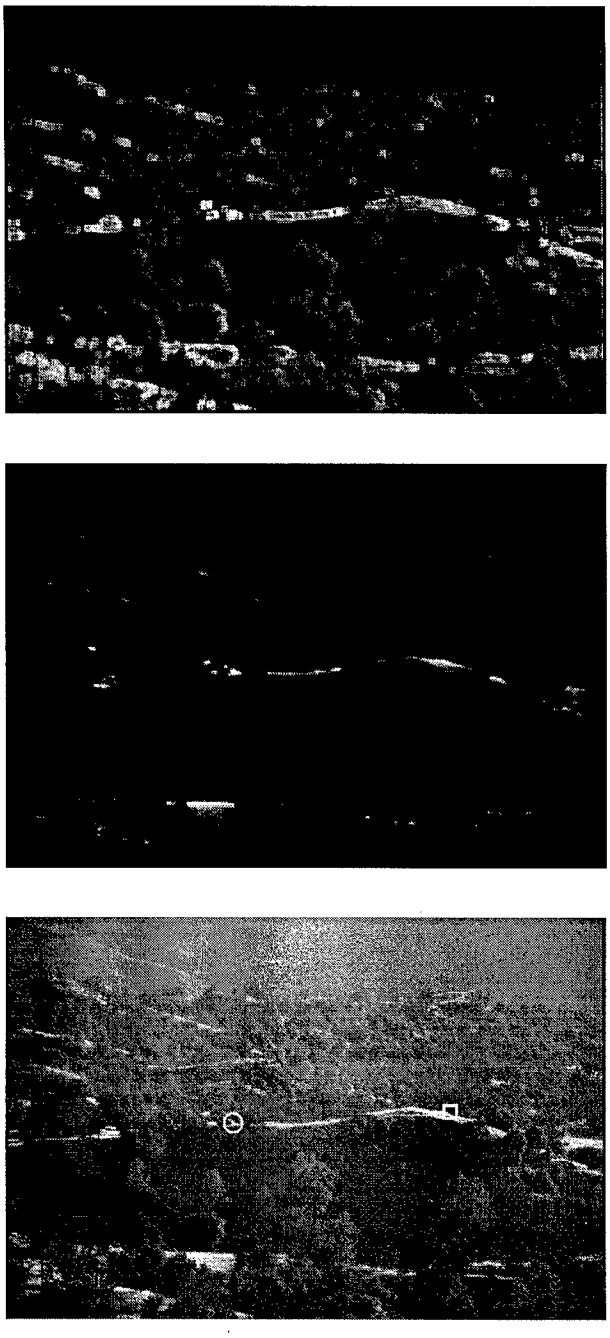

Figure 9: Comparison of SFC and saliency maps for image 0018 (shown in Fig. 5). Top: the SFC map shows strong response at all locations which have "rich" local textures; that is almost everywhere in this image. Middle: The within-feature, spatial competition for salience however demonstrates efficient reduction of information by eliminating large areas of similar textures. Bottom: The maximum of the saliency map (circle) is at the target, which appeared as a very strong isolated object in a few intensity maps because of the specular reflection on the vehicle. The maximum of the SFC map is at another location on the road.

7. ACKNOWLEDGMENTS

We thank Dr. A. Toet from TNO-HFRI for providing us with the search2 dataset and all human data. This work was supported by NSF (Caltech ERC), HIMH, ONR and NATO.

8. REFERENCES

1. James, W., *The Principles of Psychology*. Harvard Univ Press, Cambridge, MA, USA, 1890.
2. Treisman, A.M., Gelade, G., "A feature-integration theory of attention", *Cognit Psychol* 12, pp. 97-136, 1980.
3. Bergen, J.R., Julesz, B., "Parallel versus serial processing in rapid pattern discrimination", *Nature*, 303, pp. 696-8, 1983.
4. Barun, J., Julesz, B., "Withdrawing attention at little or no cost: detection and discrimination tasks. *Percept Psychophys*, 60, pp. 1-23, 1998
5. Koch, C., Ullman, S., "Shifts in selective visual attention: towards the underlying neural circuitery". *Hum Neurobiol*, 4, pp. 219-27, 1985.
6. Treisman, A., "Features and objects: the fourteenth Barlett memorial lecture", *Q J Exp Psychol [A]*, 40, pp. 201-37, 1998.
7. Burt, P., Adelson, E., "The laplacian pyramid as compact image code", *IEEE Trans on Comm*, 31, pp. 532-40, 1983.
8. Itti, L., Koch, C., Niebur, E., "A model of saliency-based visual attention for rapid scene analysis", *IEEE Trans Patt Anal Mach Intell*, 20, pp. 1254-9, 1998.
9. Itti, L., Koch, C., "A comparison of feature combination strategies for saliency-based visual attention", in *SPIE Human Vision and Electronic Imaging IV*, San Jose, CA, 1999 (in press).
10. Sillito, A.M., Grieve, K.L., Jones, H.E., Cudeiro, J., Davis, J., "Visual cortical mechanisms detecting focal orientation discontinuities", *Nature*, 378, pp. 492-6, 1995.
11. Posner, M.I., Cohen, Y., Rafal, R.D., "Neural systems control of spatial orienting", *Philos Trans R Soc Lond B Biol Sci*, 298, pp. 187-98, 1982.
12. Yarbus, A.L., *Eye movements and vision*, Plenum Press, New York, USA, 1967.
13. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K, Lai, Y.H., Davis, N., Nuflo, F. "Modeling visual-attention via selective tuning", *Artif Intell*, 78, pp.507-45, 1995.
14. Toet, A., Bijl, P., Kooi, F.L., Valenton, J.M., *A high-resolution image dataset for testing search and detection models* (Report TNO-TM-98-A020), TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1998.