Journal of Electronic Imaging 10(1), 161–169 (January 2001).

Feature combination strategies for saliency-based visual attention systems

Laurent Itti
Christof Koch
California Institute of Technology
Computation and Neural Systems Program
MSC 139-74, Pasadena, California 91125
E-mail: itti@klab.caltech.edu

Abstract. *Bottom-up or saliency-based visual attention allows primates to detect nonspecific conspicuous targets in cluttered scenes. A classical metaphor, derived from electrophysiological and psychophysical studies, describes attention as a rapidly shiftable "spotlight." We use a model that reproduces the attentional scan paths of this spotlight. Simple multi-scale "feature maps" detect local spatial discontinuities in intensity, color, and orientation, and are combined into a unique "master" or "saliency" map. The saliency map is sequentially scanned, in order of decreasing saliency, by the focus of attention. We here study the problem of combining feature maps, from different visual modalities (such as color and orientation), into a unique saliency map. Four combination strategies are compared using three databases of natural color images: (1) Simple normalized summation, (2) linear combination with learned weights, (3) global nonlinear normalization followed by summation, and (4) local nonlinear competition between salient locations followed by summation. Performance was measured as the number of false detections before the most salient target was found. Strategy (1) always yielded poorest performance and (2) best performance, with a threefold to eightfold improvement in time to find a salient target. However, (2) yielded specialized systems with poor generalization. Interestingly, strategy (4) and its simplified, computationally efficient approximation (3) yielded significantly better performance than (1), with up to fourfold improvement, while preserving generality.* © *2001 SPIE and IS&T.* [DOI: 10.1117/1.1333677]

1 Introduction

Primates use saliency-based attention to detect, in real time, conspicuous objects in cluttered visual environments. Reproducing such nonspecific target detection capability in artificial systems has important applications, for example, in embedded navigational aids, in robot navigation and in battlefield management. Based on psychophysical studies in humans and electrophysiological studies in monkeys, it is believed that bottom-up visual attention acts in some way akin to a ''spotlight.'' [1–3] The spotlight can rapidly shift across the entire visual field (with latencies on the order of 50 ms), and selects a small area from the entire visual scene. The neuronal representation of the visual world is enhanced within the restricted area of the attentional spotlight, and only this enhanced representation is allowed to progress through the cortical hierarchy for high-level processing, such as pattern recognition. Further, psychophysical studies suggest that only this spatially circumscribed enhanced representation reaches visual awareness and consciousness.[4]

Where in a scene the focus of attention is to be deployed is controlled by two tightly interacting influences: First, image-derived or ''bottom-up'' cues attract attention towards conspicuous, or ''salient'' image locations in a largely automatic and unconscious manner; second, attention can be shifted under ''top-down'' voluntary control towards locations of cognitive interest, even though these may not be particularly salient.[5] In the present study, we largely make abstraction of the top-down component and focus on the bottom-up, scene-driven component of visual attention. Thus, our primary interest is in understanding, in biologically plausible computational terms, how attention is attracted towards salient image locations. Understanding this mechanism is important because attention is likely to be deployed, during the first few hundred milliseconds after a new scene is freely viewed, mainly based on bottom-up cues. For a model which integrates a simplified bottom-up mechanism to a task-oriented top-down mechanism, we refer the reader to the article by Schill *et al.* in this issue and to Refs. 6 and 7.

A common view of how attention is deployed onto a given scene under bottom-up influences is as follows. Low-level feature extraction mechanisms act in a massively parallel manner over the entire visual scene to provide the bottom-up biasing cues towards salient image locations. Attention then sequentially focuses on salient image locations to be analyzed in more detail.[2,1] Visual attention hence allows for seemingly real-time performance by breaking down the complexity of scene understanding into a fast temporal sequence of localized pattern recognition problems.[8]

Several models have been proposed to functionally account for many properties of visual attention in primates.[6,8–13] These models typically share similar general architecture. Multi-scale topographic ''feature maps'' detect local spatial discontinuities in intensity, color, orientation and optical flow. In biologically plausible models, this is usually achieved by using a ''center-surround'' mecha-

Paper HVEI-10 received Aug. 9, 1999; revised manuscript received Aug. 25, 2000; accepted for publication Sep. 20, 2000.
1017-9909/2001/$15.00 © 2001 SPIE and IS&T.

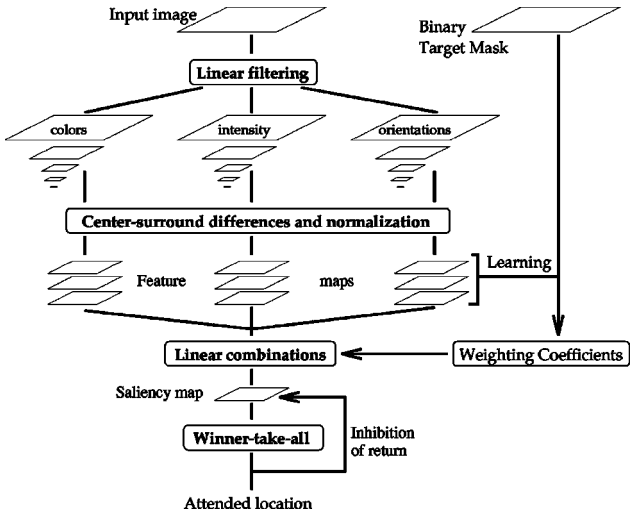

Fig. 1 General architecture of the visual attention system studied here. Early visual features are extracted in parallel in several multi-scale feature maps, which represent the entire visual scene. Such feature extraction is achieved through linear filtering for a given feature type (e.g., intensity, color or orientation), followed by a center-surround operation which extracts local spatial discontinuities for each feature type. All feature maps are then combined into a unique saliency map. We here study how this information should be combined across modalities (e.g., how important is a color discontinuity compared to an orientation discontinuity?). This can involve supervised learning using manually defined target regions (''binary target mask''). After such combination is computed, a maximum detector selects the most salient location in the saliency map and shifts attention towards it. This location is subsequently suppressed (inhibited), to allow the system to focus on the next most salient location.

nism akin to biological visual receptive fields, a process also known as a ''cortex transform'' in the image processing literature. Receptive field properties can be well approximated by difference-of-Gaussians filters (for nonoriented features) or Gabor filters (for oriented features).[10,13] Feature maps from different visual modalities are then combined into a unique ''master'' or ''saliency'' map.[1,3] In the models like, presumably, in primates, the saliency map is sequentially scanned, in order of decreasing saliency, by the focus of attention (Fig. 1).

A central problem, both in biological and artificial systems, is that of combining multi-scale feature maps, from different visual modalities with unrelated dynamic ranges (such as color and motion), into a unique saliency map. Models usually assume simple summation of all feature maps, or linear combination using *ad-hoc* weights. The object of the present study is to quantitatively compare four combination strategies using three databases of natural color images: (1) Simple summation after scaling to a fixed dynamic range; (2) linear combination with weights learned, for each image database, by supervised additive training; (3) nonlinear combination which enhances feature maps with a few isolated peaks of activity, while suppressing feature maps with uniform activity; and (4) local nonlinear iterative competition between salient locations within each feature map, followed by summation. The four strategies studied all involve a point-wise linear combination of feature maps into the scalar saliency map; the main difference between the four variants relies on the weights given to the various features. Indeed, there is mounting psychophysical evidence that different types of features do contribute additively to salience, and not, for example, through point-wise multiplication.[14] In the first three strategies, the different features are weighted in a nontopographic manner (one scalar weight for each entire map); in the fourth strategy, however, we will see that the weights are adjusted at every image location depending on its contextual surround.

2 Model

The details of the model used in the present study have been presented elsewhere[13] and are briefly schematized in Fig. 1. For the purpose of this study, it is only important to remember that different types of features, such as intensity, color or orientation are first extracted in separate multi-scale feature maps, and then need to be combined into a unique ''saliency map,'' whose activity controls attention (Fig. 2).

2.1 *Fusion of Information*

One difficulty in combining different feature maps into a single scalar saliency map is that these features represent *a priori* not comparable modalities, with different dynamic ranges and extraction mechanisms. Also, because many feature maps are combined (6 for intensity computed at different spatial scales, 12 for color and 24 for orientation in our implementation), salient objects appearing strongly in only a few maps risk being masked by noise or less salient objects present in a larger number of maps. The system is hence faced with a severe signal-to-noise ratio problem, in which relevant features, even though they may elicit strong responses in some maps, may be masked by the sum or weaker noisy responses present in a larger number of maps. The most simple approach to solve this problem is to normalize all feature maps to the same dynamic range (e.g., between 0 and 1), and to sum all feature maps into the saliency map. This strategy, which does not impose any *a priori* weight on any feature type, is referred to in what follows as the ''Naive'' strategy.

2.2 *Learning*

Supervised learning can be introduced when specific targets are to be detected. In such case, each feature map is globally multiplied by a weighting factor, which might correspond in biological systems to a simple change in the gain associated to a given feature type, under volitional control (such neuronal gain changes have been observed in awake behaving monkeys instructed, for example, to attend to a particular direction of motion[15]). The final input to the saliency map is then the point-wise weighted sum of all such feature maps.

Our implementation uses supervised learning in order to determine the optimal set of linear map weights for a given class of images. It seems reasonable to assume that such optimization may be carried out in biological systems while animals are trained to perform the desired target detection task. During the training phase, all feature weights are trained simultaneously, based on a comparison, for each feature type, of the map's response inside and outside manually outlined image regions which contain the desired targets. The learning procedure for the weight $w(\mathcal{M})$ of a feature map $\mathcal{M}$ consists of the following:

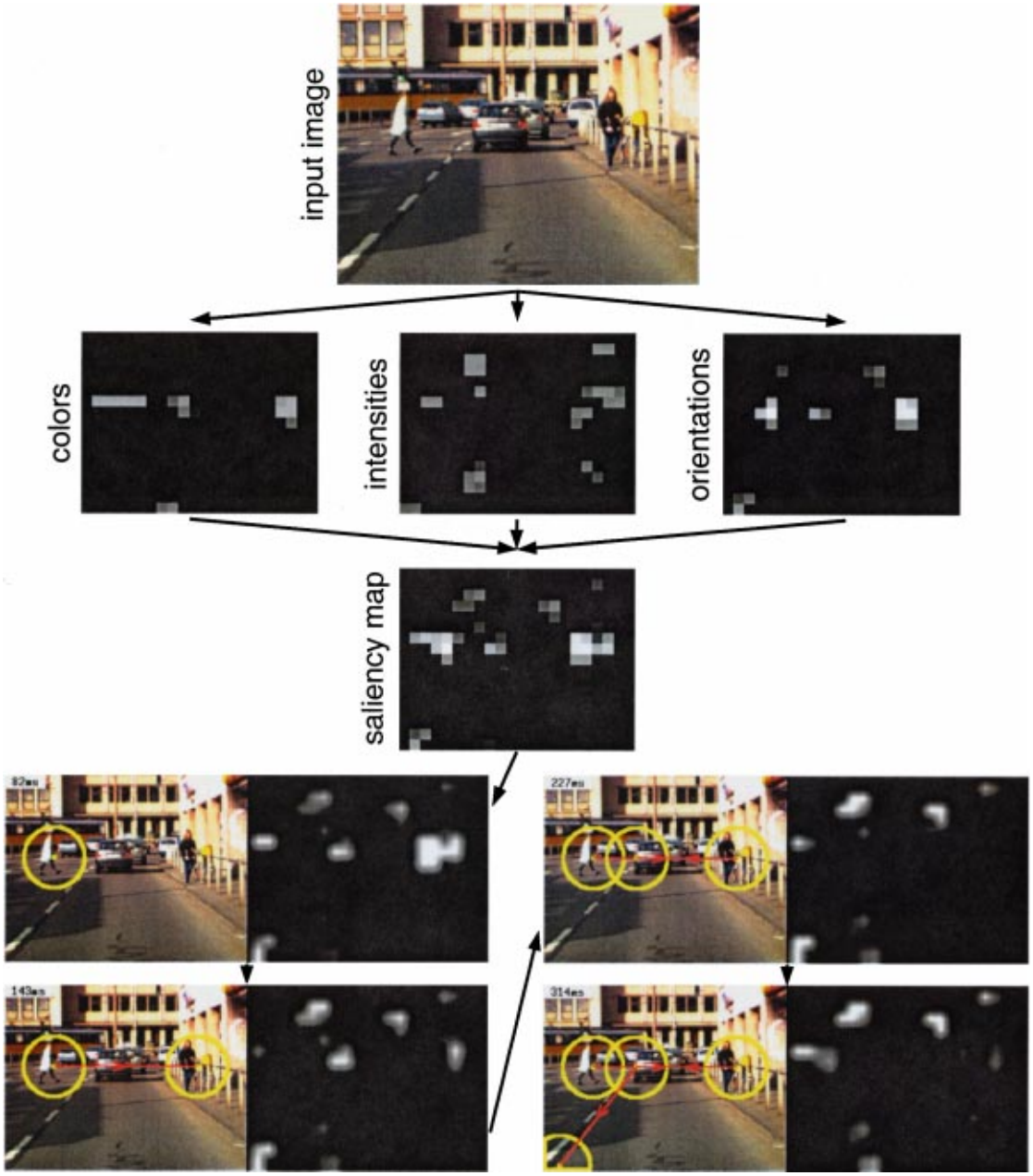

Fig. 2 Example of operation of the model with a natural (color) image and the iterative feature combination strategy (see Sec. 2.4). The most salient location is at one pedestrian, who appears strongly in the orientation maps; it becomes the object of the first attentional shift (82 ms simulated time) and is subsequently suppressed in the saliency map by an ''inhibition of return'' mechanism. The next attended location is at another pedestrian (143 ms) which appreaded strongly in the orientation and intensity maps, followed by a car (227 ms) and a street marking (314 ms). The inhibition of return is only transiently activated in the saliency map, such that the first attended location has regained some activity at 314 ms. More examples of model predictions on natural and synthetic images can be found at http://www.klab.caltech.edu/~itti/attention/

1. Compute the global maximum M_{glob} and minimum m_{glob} of the map $\mathcal{M}$;
2. compute its maximum M_{in} inside the manually outlined target region(s) and its maximum M_{out} outside the target region(s); and
3. update the weight following an additive learning rule independent of the map's dynamic range

$$w(\mathcal{M}) \leftarrow w(\mathcal{M}) + \eta(M_{\text{in}} - M_{\text{out}})/(M_{\text{glob}} - m_{\text{glob}}), \quad (1)$$

where $\eta > 0$ determines the learning speed. Only positive or zero weights are allowed.

This learning procedure promotes, through an increase in weights, the participation to the saliency map of those feature maps which show higher peak activity inside the target region(s) than outside; after training, only such maps remain in the system while others, whose weights have converged to zero, are computed no more. The initial saliency map (before any attentional shift) is then scaled to a fixed range, such that only the relative weights of the feature maps are important; with such normalization, potential divergence of the additive learning rule (explosion of weights) can hence be avoided by constraining the weights to a fixed sum.

We only consider local maxima of activity over various image areas, rather than the average activity over these areas. This is because local ''peak'' activity is what is important for visual salience: If a rather extended region contains only a very small but very strong peak of activity, this peak is highly salient and immediately ''pops out,'' while the average activity over the extended region may be low. This feature combination strategy is referred to in what follows as the ''Trained'' strategy.

2.3 *Contents-based Global Nonlinear Amplification*

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 (''odd man out'') are present, while globally suppressing feature maps eliciting comparable peak responses at numerous locations over the visual scene. The normalization operator, denoted $\mathcal{N}(.)$, consists of the following:

1. Normalize all the feature maps to the same dynamic range, in order to eliminate across-modality amplitude differences due to dissimilar feature extraction mechanisms;
2. for each map, find its global maximum M and the average $\overline{m}$ of all the other local maxima; and
3. globally multiply the map by

$$(M - \overline{m})^2. \quad (2)$$

Only local maxima of activity are considered such that $\mathcal{N}(.)$ compares responses associated with meaningful ''activation spots'' in the map and ignores homogeneous areas. Comparing the maximum activity in the entire map to the average over all activation spots measures how different the most active location is from the average. When this difference is large, the most active location stands out, and we strongly promote the map. When the difference is small, the map contains nothing unique and is suppressed. This contents-based nonlinear normalization coarsely replicates a biological lateral inhibition mechanism, in which neighboring similar features inhibit each other.[16] This feature combination strategy is referred to in what follows as the ''$\mathcal{N}(.)$'' strategy.

2.4 *Iterative Localized Interactions*

The global nonlinear normalization presented in the previous section is computationally very simple and is noniterative, which easily allows for real-time implementation. However, it suffers from several drawbacks. First, this strategy is not very biologically plausible, since global computations, such as finding the global maximum in the image, are used, while it is known that cortical neurons are only locally connected. Second, this strategy has a strong bias towards enhancing those feature maps in which a unique location is significantly more conspicuous than all others. Ideally, each feature map should be able to represent a sparse distribution of a few conspicuous locations over the entire visual field; for example, our $\mathcal{N}(.)$ normalization would suppress a map with two equally strong spots and otherwise no activity, while a human would typically report that both spots are salient.

Finally, the computational strategy employed in the previous section is not robust to noise, in the cases when noise can be stronger than the signal (e.g., speckle or ''salt-and-pepper'' noise); in such stimuli, a single pixel of noise so high that it is the global maximum of the map would determine the map's scaling. While such a problem is unlikely (since feature maps usually are built, from the noisy input image, using feature extraction mechanisms optimized to filter out the noise), it decreases the overall robustness of the system when using natural images.

We consequently propose a fourth feature combination strategy, which relies on simulating local competition between neighboring salient locations. The general principle is to provide self-excitation and neighbor-induced inhibition to each location in the feature map, in a way coarsely inspired from the way long-range cortico-cortical connections (up to 6–8 mm in cortex) are believed to be organized in primary visual cortex.[17,18]

Each feature map is first normalized to values between 0 and 1, in order to eliminate modality-dependent amplitude differences. Each feature map is then iteratively convolved by a large two-dimensional (2D) difference of Gaussians (DoG) filter, and negative results are clamped to zero after each iteration. The DoG filter, a one-dimensional (1D) section of which is shown in Fig. 3, yields strong local excitation at each visual location, which is counteracted by broad inhibition from neighboring locations. Specifically, such filter $\text{DoG}(x)$ is obtained by

$$\text{DoG}(x,y) = \frac{c_{\text{ex}}^2}{2\pi\sigma_{\text{ex}}^2} e^{-(x^2+y^2)/2\sigma_{\text{ex}}^2} - \frac{c_{\text{inh}}^2}{2\pi\sigma_{\text{inh}}^2} e^{-(x^2+y^2)/2\sigma_{\text{inh}}^2}. \quad (3)$$

In our implementation, $\sigma_{\text{ex}} = 2\%$ and $\sigma_{\text{inh}} = 25\%$ of the input image width, $c_{\text{ex}} = 0.5$ and $c_{\text{inh}} = 1.5$ (Fig. 3). At each

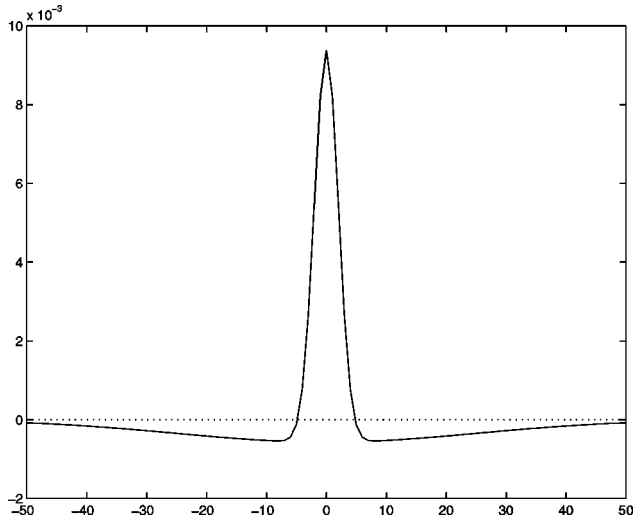

Fig. 3 One-dimensional (1D) section of the 2D difference of Gaussians (DoG) filter used for iterative normalization of the feature maps. The central excitatory lobe strongly promotes each active location in the map, while the broader negative surround inhibits that location, if other strongly activated locations are present nearby. The DoG filter represented here is the one used in our simulations, with its total width being set to the width of the input image.

iteration of the normalization process, a given feature map $\mathcal{M}$ is then subjected to the following transformation:

$$\mathcal{M} \leftarrow |\mathcal{M} + \mathcal{M} * \mathrm{DoG} - C_{\mathrm{inh}}|_{\geq 0}, \tag{4}$$

where DoG is the 2D difference of Gaussian filter described above, $|.|_{\geq 0}$ discards negative values, and C_{inh} is a constant inhibitory term ($C_{\mathrm{inh}} = 0.02$ in our implementation with the map initially scaled between 0 and 1). C_{inh} introduces a small bias towards slowly suppressing areas in which the excitation and inhibition balance almost exactly; such regions typically correspond to extended regions of uniform textures (depending on the DoG parameters), which we would not consider salient.

The 2D DoG filter, which is not separable, is implemented by taking the difference between the results of the convolution of $\mathcal{M}$ by the separable excitatory Gaussian of the DoG, and of the convolution of $\mathcal{M}$ by the separable inhibitory Gaussian. One reason for this approach is that two separable 2D convolutions (one of which, the excitatory Gaussian, has a very small kernel) and one subtraction are computationally much more efficient than one inseparable 2D convolution. A second reason is boundary conditions; this is an important problem here since the inhibitory lobe of the DoG is slightly larger than the entire visual field. Using Dirichlet (wraparound) or ''zero-padding'' boundary conditions yields very strong edge effects which introduce unwanted nonuniform behavior of the normalization process (e.g., when using zero padding, the corners of an image containing uniform random noise invariably become the most active locations, since they receive the least inhibition). We circumvent this problem by truncating the separable Gaussian filter $\mathcal{G}$, at each point during the convolution, to its portion which overlaps the input map $\mathcal{M}$ (Fig. 4). The truncated convolution is then computed as, using the fact that $\mathcal{G}$ is symmetric around its origin

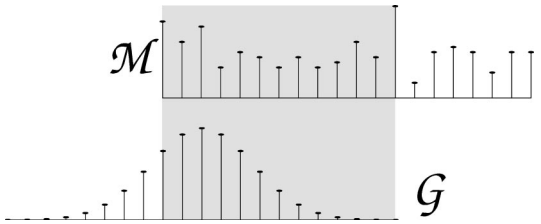

Fig. 4 Truncated filter boundary condition consists of only computing the dot product between filter $\mathcal{G}$ and map $\mathcal{M}$ where they overlap (shaded area), and of normalizing the result by the total area of $\mathcal{G}$ divided by its area in the overlap region.

$$\mathcal{M} * \mathcal{G}(x) = \frac{\Sigma_i \mathcal{G}(i)}{\Sigma_{i \in \{\mathrm{overlap}\}} \mathcal{G}(i)} \sum_{i \in \{\mathrm{overlap}\}} \mathcal{M}(i)\mathcal{G}(i). \tag{5}$$

Using this ''truncated filter'' boundary condition yields uniform filtering over the entire image (see, e.g., Figs. 5 and 6), and, additionally, presents the advantage of being more biologically plausible than Dirichlet or zero-padding conditions: A visual neuron with its receptive field near the edge of our visual field indeed is not likely to implement zero padding or wrap around, but is likely to have a reduced set of inputs, and to accordingly adapt its output firing rate to a range similar to that of other neurons in the map.

Two examples of operation of this normalization scheme are given in Figs. 5 and 6, and show that, similar to $\mathcal{N}(.)$, a map with many comparable activity peaks is suppressed while a map where one (or a few) peak stands out is enhanced. The dynamics of this new scheme are, however, much more complex than those of $\mathcal{N}(.)$, since now the map is locally altered rather than globally (nontopographically) multiplied; for example, a map such as that in Fig. 5 converges to a single activated pixel (at the center of the initial

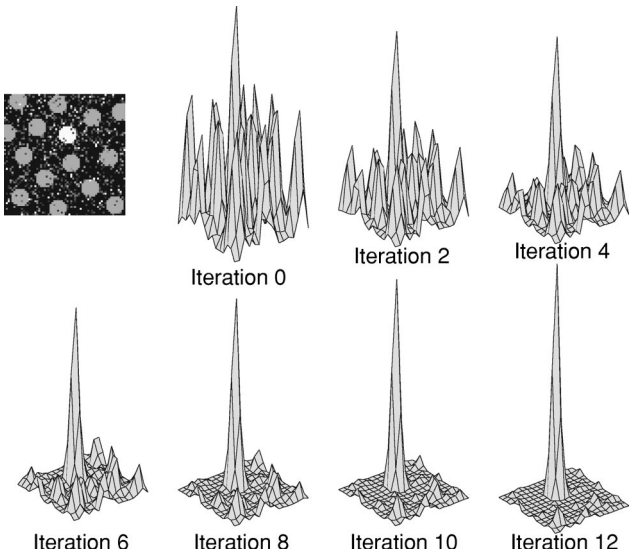

Fig. 5 Iterative normalization of a feature map containing one strong activation peak surrounded by several weaker ones. After a few iterations, the initially stronger peak has gained in strength while at the same time suppressing weaker activation regions. Note how initially very strong speckle noise is effectively suppressed by the ietrative rectified filtering.

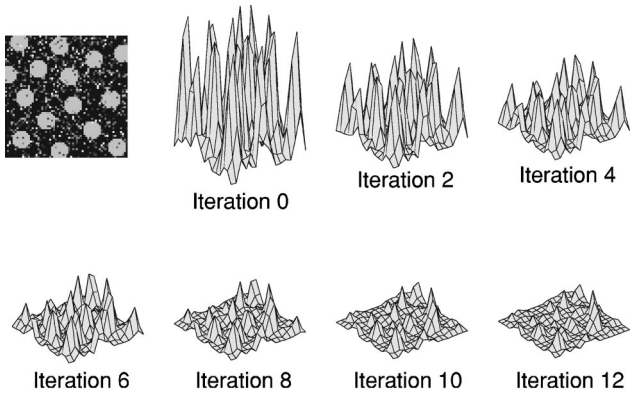

Fig. 6 Iterative normalization of a feature map containing numerous strong activation peaks. This time, all peaks equally inhibit each other, resulting in the entire map being suppressed.

strong peak) after a large number of iterations. Note finally that, although the range of the inhibitory filter seems to far exceed that of intrinsic cortico-cortical connections in primates,[18] it is likely that such inhibition is fed back from higher cortical areas where receptive fields can cover substantial portions of the entire visual field, to lower visual areas with smaller receptive fields. In terms of implementation, the DoG filtering proposed here is best carried out within the multi-scale framework of Gaussian pyramids.[13] Finally, it is interesting to note that this normalization scheme resembles a ''winner-take-all'' network with localized inhibitory spread, which has been implemented for real-time operation in Analog-VLSI.[19] This normalization scheme will be referred to at the ''Iterative'' scheme in what follows.

3 Results and Discussion

We previously have applied our model to a variety of search tasks, including psychophysical pop-out tasks,[20] visual search asymmetries,[21] images containing a military vehicle in a rural background,[20] various test patterns,[13] images containing pedestrians,[22] and various magazine covers, scientific posters, and advertising billboards. The remarkable performance of our model at reproducing or exceeding human search performance in such a diverse variety of tasks seems to indicate that the model indeed is able to find salient objects irrespectively of their nature. Here, we use new sets of test images, which contain targets of increasing complexity and variability.

We used three databases of natural color images to evaluate the different feature combination strategies proposed above (Fig. 7). The first database consisted of images in which a red aluminum can is the target. It was used to demonstrate the simplest form of specialization, in which some feature maps in the system specifically encode for the main feature of the target (red color, which is explicitly detected by the system in a red/green feature map[13]). The second database consisted of images in which a vehicle's emergency triangle was the target. A more complicated form of specialization is hence demonstrated, since the target is unique in these images only by a conjunction of red color and of 0° (horizontal), 45° or 135° orientations. These four feature types are represented in the system by four separate and independent feature maps.[13] The third database consisted of 90 images acquired by a video camera mounted on the passenger side of a vehicle driven on German roads, and contained one or more traffic signs. Among all 90 images, 39 contained one traffic sign, 35 contained two, 12 contained three, 2 contained four, and 1 contained five traffic signs. This database contained a wide variety of targets, of various colors (red, blue, yellow, white, black, orange), shapes (circular, triangular, square, rectangle), textures (uniform, striped, with lettering, dull, luminous); in addition, these signs (and the targets in the other two databases as well) could be fully visible or partially occluded, shiny or dull, in the shadow or showing specular reflections, light or dark, large or small, and viewed frontally or at an angle, in scenes which also demonstrated high degrees of variability (please see `http://www.klab.caltech.edu/~itti/attention/`).

What characterizes the image databases used here is that we chose the target patterns to be ''perceptually salient.'' Since this is not a trivial property of an object, we used the simplification that traffic signs have been designed, optimized, and strategically placed in the environment to be perceptually salient. The exact nature of the targets used here, however, is not our main focus; the present study indeed aims at comparing the four proposed feature combination strategies for the computation of salience.

All targets were outlined manually, and binary target masks were created. A target was considered detected when the focus of attention (FOA) intersected the target. The images were 640×480 (red can and triangle) and 512×384 (traffic signs) with 24 bit color, and the FOA was a disk of radius 80 (red can and triangle) and 64 (traffic signs) pixels. Complete coverage of an entire image would consequently require the FOA to be placed at 31 different locations (with overlap). A system performing at random would have to visit an average of 15.5 locations to find a unique, small target in the image.

Each image database was split into a training set (45 images for the can, 32 for the triangle, 45 for the traffic signs) and a test set (59, 32 and 45 images, respectively). Learning consisted, for each training set, of five randomized passes through the whole set with halving of the learning speed η after each pass.

We compared the results obtained on the test image sets with the four proposed feature combination strategies:

1. Naive model with no dedicated normalization and all feature weights set to unity;
2. model with the noniterative $\mathcal{N}(.)$ normalization;
3. model with 12 iterations of the Iterative normalization; and
4. trained model, i.e., with no dedicated normalization but feature weights learned from the corresponding training set.

We retained in the test sets only the most challenging images, for which the target was *not* immediately detected by at least one of the four versions of the model (easier images in which at least one version of the model could

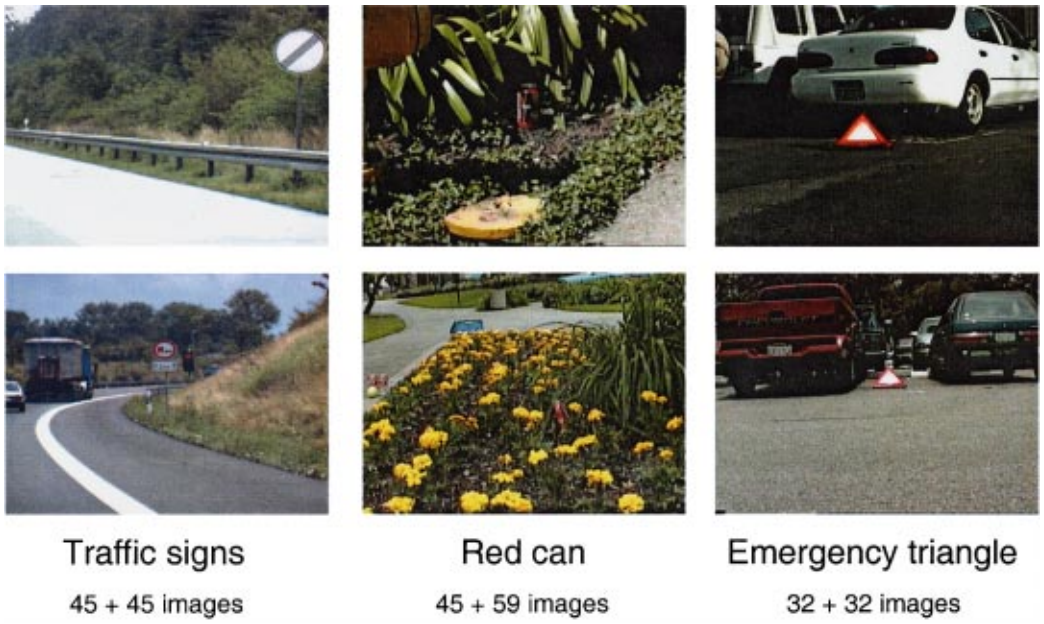

Fig. 7 Example images from the three image databases studied. The number of images for training+test sets is shown for each database.

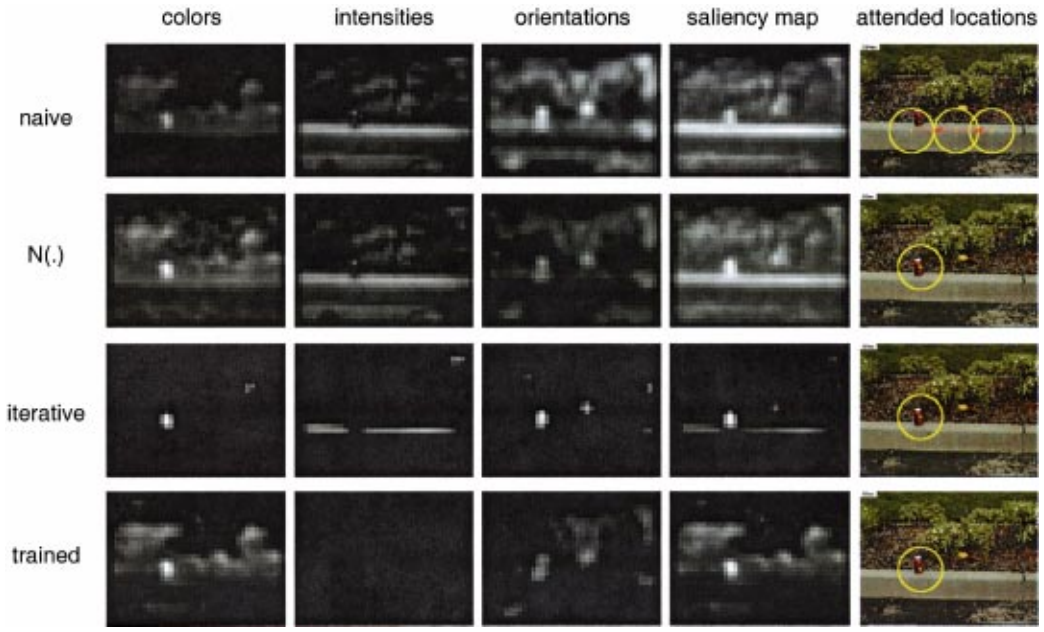

Fig. 8 Comparison of the internals of the four versions of the model, for one image from the ''red can'' test set, in which a red aluminum can is the most salient object. The can appears with medium strength in the color maps, due to its color contrast with the background (the response is not the strongest possible because the background is not green, and only red/green and blue/yellow color contrasts are computed). The curb, however, appears very strongly in all intensity maps, and also less strongly in the horizontal orientation maps. In the naive version of the model, the color activity from the can is outnumbered by the activity elicited by the curb in a larger number of intensity and orientation maps. As a result, detection of the can is accidental, while the model is scanning the curb. The $\mathcal{N}(.)$ strategy yields strong suppression of the horizontal orientation, because more localized activation peaks exist in the vertical orientation, as well as some suppression of the extended curb in the intensity channel. The color channel, with its strong singularity, is, however, globally enhanced and yields correct detection of the can. The iterative strategy yields complete suppression of the horizontal orientation as well as overall much suppression of all regions which are not among the few strongest in each feature map. The red can clearly becomes the most salient location in the image. Finally, training using other images with similar views of this red target of vertical orientation has entirely suppressed the intensity and horizontal orientations, such that the saliency map is dominated by the color channel. The trained model hence easily finds the can as the most salient object.

Table 1 Average number of false detections (mean±standard deviation) before target(s) found, for the red can test set ($n=59$), emergency triangle test set ($n=32$) and traffic signs test set ($n=45$; 17 images with 1 sign, 19 with 2, 6 with 3, 2 with 4 and 1 with 5). For the traffic sign images which could contain more than one target per image, we measured both the number of false detections before the first target hit, and before all targets in the image had been detected.

	Naive	$\mathcal{N}(.)$	Iterative	Trained
Red can	2.90±2.50	1.67±2.01	1.24±1.42	0.35±1.03
Triangle	2.44±2.20	1.69±2.28	1.42±1.67	0.87±1.29
Traffic[a]	1.84±2.13	0.49±1.06	0.52±1.05	0.24±0.77
Traffic[b]	3.26±2.80	1.27±2.12	0.70±1.18	0.77±1.93

[a]Before *first* sign found.

[b]Before *all* signs found.

immediately find the targets had been previously discarded to ensure that performance was not at ceiling). Results are summarized in Table 1.

The Naive model, which represents the simplest solution to the problem of combining several feature maps into a unique saliency map (and had the smallest number of free parameters), performed always worse than when using $\mathcal{N}(.)$. This simple contents-based normalization proved particularly efficient at eliminating feature maps in which numerous peaks of activity were present, such as, for example, intensity maps in images containing large variations in illumination. Furthermore, the more detailed, iterative implementation of spatial competition for salience (which has the highest number of free parameters) yielded comparable or better results, in addition to being more biologically plausible.

The additive learning rule also proved efficient in specializing the generic model. One should be aware, however, that only limited specialization can be obtained from such global weighting of the feature maps: Because such learning simply enhances the weight of some maps and suppresses others, poor generalization should be expected when trying to learn for a large variety of objects using a single set of weights, since each object would ideally require a specific set of weights. Additionally, the type of linear training employed here is limited, because *sums* of features are learned rather than *conjunctions*. For example, the model trained for the emergency triangle might attend to a strong oblique edge even if there was no red color present or to a red blob in the absence of any oblique orientation. To what extent humans can be trained to preattentively detect learned conjunctive features remains controversial.[12] Nevertheless, it was remarkable that the trained model performed best of the four models studied here for the database of traffic signs, despite the wide variety of shape (round, square, triangle, rectangle), color (red, white, blue, orange, yellow, green) and texture (uniform, striped, lettered) of those signs in the database.

In summary, while the Naive method consistently yielded poor performance and the Trained method yielded specialized models for each task, the iterative normalization operator, and its noniterative approximation $\mathcal{N}(.)$, yielded reliable yet nonspecific detection of salient image locations. We believe that the latter two represent the best approximations to human saliency among the four alternatives studied here. One of the key elements in the iterative method is the existence of a nonlinearity (threshold) which suppresses negative values; as we can see in Figs. 5 and 6, in a first temporal period, the global activity over the entire map typically decreases as a result of the mutual inhibition between the many active locations, until the weakest activation peaks (typically due to noise) pass below threshold and are eliminated. Only after the distribution of activity peaks has become sparse enough can the self-excitatory term at each isolated peak overcome the inhibition received from its neighbors, and, in a second temporal period, the map's global activity starts increasing again. If many comparable peaks are present in the map, the first period of decreasing activity will be much slower than if one or a few much stronger peaks efficiently inhibit all other peaks. In Fig. 8, we show a comparison of the internal maps for the four versions of the model on a test image. This figure demonstrates, in particular, how the Iterative scheme yields much sparser maps, in which most of the noisy activity present in some channels (such as the intensity channel in the example image) is strongly suppressed.

Note that our model certainly does not represent the most efficient detector for the type of targets studied here. One could indeed think of much simpler dedicated architectures to detect traffic signs or soda cans (e.g., algorithms based on template matching). However, as mentioned earlier, what characterizes our model is that it finds salient objects, vehicles, persons, or other image regions in a manner which is largely independent of the nature of the targets. For the purpose of the present study, the good but imperfect performance of the model allowed us to compare the four feature comparison strategies using a set of very varied natural scenes in which target detection performance was not at ceiling.

The proposed iterative scheme could be refined in several ways in order to mimic more closely what is known of the physiology of early visual neurons. For example, in this study, we have not applied any nonlinear ''transducer function'' (which relates the output firing rate of a neuron to the strength of its inputs), while it is generally admitted that early visual neurons have a sigmoidal transducer function.[23,24] Also, we have modeled interactions between neighboring regions of the visual field by simple self-excitation and subtractive neighbor-induced inhibition, while more complicated patterns of interactions within the ''nonclassical receptive field'' of visual neurons have been reported.[25,26] Finally, the scale of the excitatory lobe of our iterative filter should be adaptive, and change depending on object size, type of image, type of image area, or top-down influences. This problem (as well as the development of an object-based rather than circular focus of attention) is currently under study in our laboratory.

In conclusion, we compared four simple strategies for combining multiple feature maps from different visual modalities into a single saliency map. The introduction of a simple learning scheme proved most efficient for detection of specific targets, by allowing for broad specialization of the generic model. Remarkably, however, good performance was also obtained using a simple, nonspecific normalization which coarsely replicates biological within-

feature spatial competition for saliency. Both the additive learning and the nonlinear (iterative or not) normalization strategies can provide significant performance improvement to models which previously used *ad-hoc* weighted summation as a feature combination strategy.

Acknowledgments

This work was supported by ONR, NIMH and NSF (Caltech ERC). The authors thank Daimler-Benz for providing them with some of the test images used in this study.

References

1. C. Koch and S. Ullman, ''Shifts in selective visual attention: Towards the underlying neural circuitry,'' *Hum. Neurobiol.* **4**, 219–297, 916 (1985).
2. A. M. Treisman and G. Gelade, ''A feature-integration theory of attention,'' *Cogn. Psychol.* **12**, 97–136 (1980).
3. J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg, ''The representation of visual salience in monkey parietal cortex,'' *Nature (London)* **391**, 481–4 (1998).
4. J. K. O'Regan, R. A. Rensink, and J. J. Clark, ''Change-blindness as a result of 'mudsplashes' (letter),'' *Nature (London)* **398**, 34 (1999).
5. R. Rarasuraman, *The Attentive Brain*, MIT Press, Cambridge, MA (1998).
6. J. M. Wolfe, ''Guided search 2.0: A revised model of visual search,'' *Psychon. Bull. Rev.* **1**, 202–38 (1994).
7. C. K. R. , ''The featuregate model of visual selection,'' *Psychol. Res.* **62**, 182–194 (1999).
8. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. H. Lai, N. Davis, and F. Nuflo, ''Modeling visual attention via selective tuning,'' *Artif. Intel.* **78**, 507–45 (1995).
9. B. A. Olshausen, C. H. Anderson, and D. C. V. Essen, ''A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information,'' *J. Neurosci.* **13**, 4700–1947 (1993).
10. R. Milanese, S. Gil, and T. Pun, ''Attentive mechanisms for dynamic and static scene analysis,'' *Opt. Eng.* **34**, 2428–34 (1995).
11. S. Balnja and D. A. Pomerleau, ''Expectation-based selective attention for visual monitoring and control of a robot vehicle,'' *Rob. Auton. Syst.* **22**, 329–44 (1997).
12. E. Niebur and C. Koch, ''Computational architectures for attention,'' in Ref. 5.
13. L. Itti, C. Koch, and E. Niebur, ''A model of saliency-based visual attention for rapid scene analysis,'' *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259 (1998).
14. H. Nothdurft, ''Salience from feature contrast: Additivity across dimensions (in process citation),'' *Vision Res.* **40**, 1183–1201 (2000).
15. S. Treue and J. C. M. Trujillo, ''Feature-based attention influences motion processing gain in macaque visual cortex,'' *Nature (London)* **399**, 575–579 (1999).
16. M. W. Cannon and S. C. Fullenkamp, ''A model for inhibitory lateral interaction effects in perceived contrast,'' *Vision Res.* **36**, 1115–25 (1996).
17. C. D. Gilbert, A. Das, M. Ito, M. Kapadia, and G. Westheimer, ''Spatial integration and cortical dynamics,'' *Proc. Natl. Acad. Sci. U.S.A.* **93**, 615–22 (1996).
18. C. D. Gilbert and T. N. Wiesel, ''Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex,'' *J. Neurosci.* **9**, 2432–2442 (1989).
19. T. K. Horiuchi, T. G. Morris, C. Koch, and S. P. DeWeerth, ''Analog vlsi circuits for attention-based, visual tracking,'' in *Neural Information Processing Systems (NIPS) 9*, T. P. M. C. Mozer, M. I. Jordan, Eds., pp. 706–712, MIT Press, Cambridge, MA (1997).
20. L. Itti and C. Koch, ''A saliency-based search mechanism for overt and covert shifts of visual attention (in process citation),'' *Vision Res.* **40**, 1489–1506 (2000).
21. F. Tehrani, L. Itti, and C. Koch, ''Visual search asymmetries reproduced by simple model,'' *Invest. Ophthalmol. Visual Sci.* **41**, S423 (2000).
22. L. Itti, C. Papageorgiou, T. Poggio, and C. Koch, ''A cooperative vision system for detecting pedestrians in natural images'' (submitted).
23. H. R. Wilson, ''A transducer function for threshold and suprathreshold human vision,'' *Biol. Cybern.* **38**, 171–178 (1980).
24. D. J. Heeger, ''Normalization of cell responses in cat striate cortex,'' *Vis. Neurosci.* **9**, 181–197 (1992).
25. A. M. Sillito and H. E. Jones, ''Context-dependent interactions and visual processing in v1,'' *J. Physiol. (Paris)* **90**, 205–209 (1996).
26. A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro, and J. Davis, ''Visual cortical mechanisms detecting focal orientation discontinuities,'' *Nature (London)* **378**, 492–496 (1995).

Laurent Itti received his MS in Image Processing from the Ecole Nationale Superieure des Telecommunications in 1994, and PhD in Computation and Neural Systems from Caltech in 2000. In September 2000, he joined the faculty of the Department of Computer Science at the University of Southern California. His primary research interest is in biologically plausible computational brain modeling, and in the comparision of model simulations to empirical measurements from living systems. Of particular interest in his laboratory is the development of computational models of biological vision, and the application of such models to computer vision problems. Dr. Itti teaches in the areas of computational neuroscience, neuroinformatics, vision, and the mathematical foundations of computer science.

Christof Koch received his PhD in Physics from the University of Tubingen, Germany. His research focuses on understanding the biophysical mechanisms underlying information processing in individual nerve cells as well as the neuronal operations underlying spatial vision, motion, shape perception and visual attention in the primate's visual system using electrophysiological, brain imaging, psychophysical and computational tools. Together with Dr. Francis Crick, Professor Koch works on the neuronal basis of visual consciousness.