A Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems

Laurent Itti and Christof Koch

California Institute of Technology, Computation and Neural Systems Program
MSC 139-74, Pasadena, California 91125, U.S.A.

ABSTRACT

Bottom-up or saliency-based visual attention allows primates to detect non-specific conspicuous targets in cluttered scenes. A classical metaphor, derived from electrophysiological and psychophysical studies, describes attention as a rapidly shiftable "spotlight". The model described here reproduces the attentional scanpaths of this spotlight: Simple multi-scale "feature maps" detect local spatial discontinuities in intensity, color, orientation or optical flow, 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 study the problem of combining feature maps, from different visual modalities and with unrelated dynamic ranges (such as color and motion), 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 non-linear normalization followed by summation, and (4) local non-linear competition between salient locations. 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 3 to 8-fold 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 4-fold improvement, while preserving generality.

Keywords: Attention, saliency, target detection, feature integration, learning

1. INTRODUCTION

Primates use saliency-based attention to detect, in real time, conspicuous objects in cluttered visual environments. Reproducing such non-specific target detection capability in artificial systems has important applications, for example in embedded navigational aids and in robot navigation. 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 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. At a given time, the location of the attentional spotlight is controlled by low-level feature extraction mechanisms, which crudely segment conspicuous locations from the entire visual scene, in a massively parallel fashion, and compete for attention. 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 sequence of circumscribed pattern recognition problems.[4]

A second, top-down, task-dependent and volitional component of attention can be superimposed onto the bottom-up and task-independent form of attention we study here.[5] We will disregard the top-down component.

Several models have been proposed to functionally account for visual attention in primates.[6–8,4,9–11] These models share similar general architecture. Multi-scale topographic "feature maps" detect local spatial discontinuities in intensity, color, orientation or optical flow. In biologically-plausible models, this is usually achieved by using a "center-surround" mechanism akin to biological visual receptive fields. Receptive field properties can be well approximated by difference-of-Gaussians filters (for non-oriented features) or Gabor filters (for oriented features).[8,11] 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 **(Figure 1)**.

Send correspondence to L.I. (itti@klab.caltech.edu) or C.K. (koch@klab.caltech.edu).

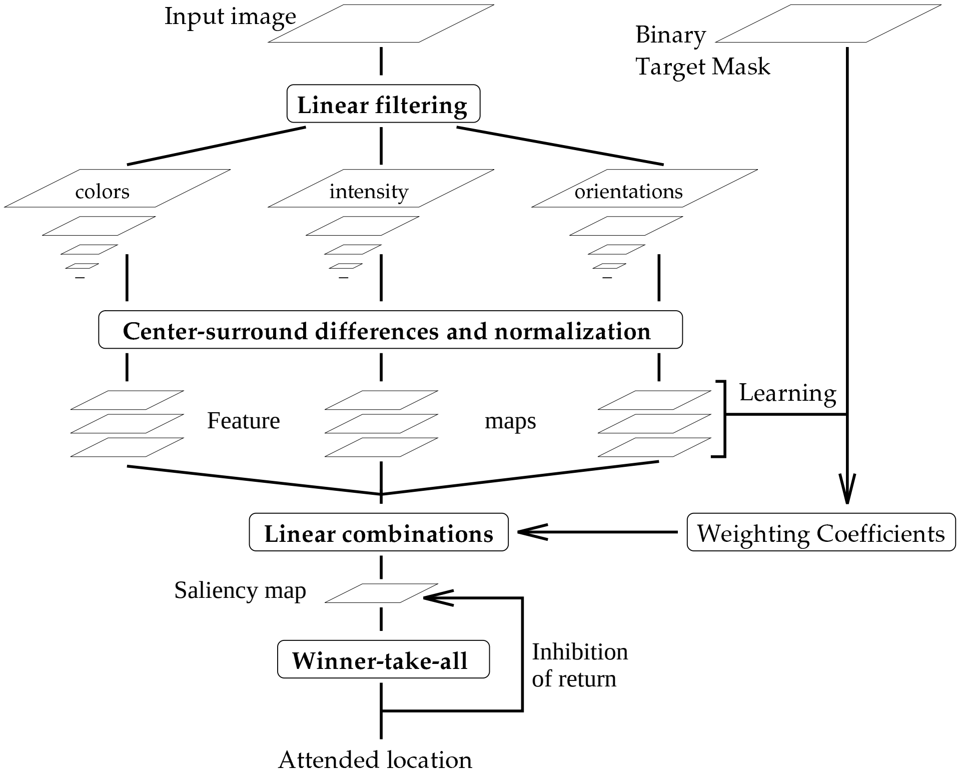

Figure 1. General architecture of the visual attention system studied here. Early visual features are extracted in parallel in several multiscale "feature maps", which represent the entire visual scene. Such feature extraction is achieved through linear filtering for a given feature type (e.g., intensity, 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.

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. Here, we 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) non-linear combination which enhances feature maps with a few isolated peaks of activity, while suppressing feature maps with uniform activity; and (4) local non-linear iterative competition between salient locations.

2. MODEL

The details of the model used here have been presented elsewhere[11] and are briefly schematized in **Figure 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 multiscale "feature maps", and then need to be combined into a unique "saliency map", whose activity controls attention **(Figure 2)**.

2.1. Fusion of information

One difficulty in combining different feature maps into a unique scalar saliency map is that they represent *a priori* not comparable modalities, with different dynamic ranges and extraction mechanisms. Also, because a large number of feature maps are combined (6 for intensity computed at different spatial scales, 12 for color and 24 for orientation

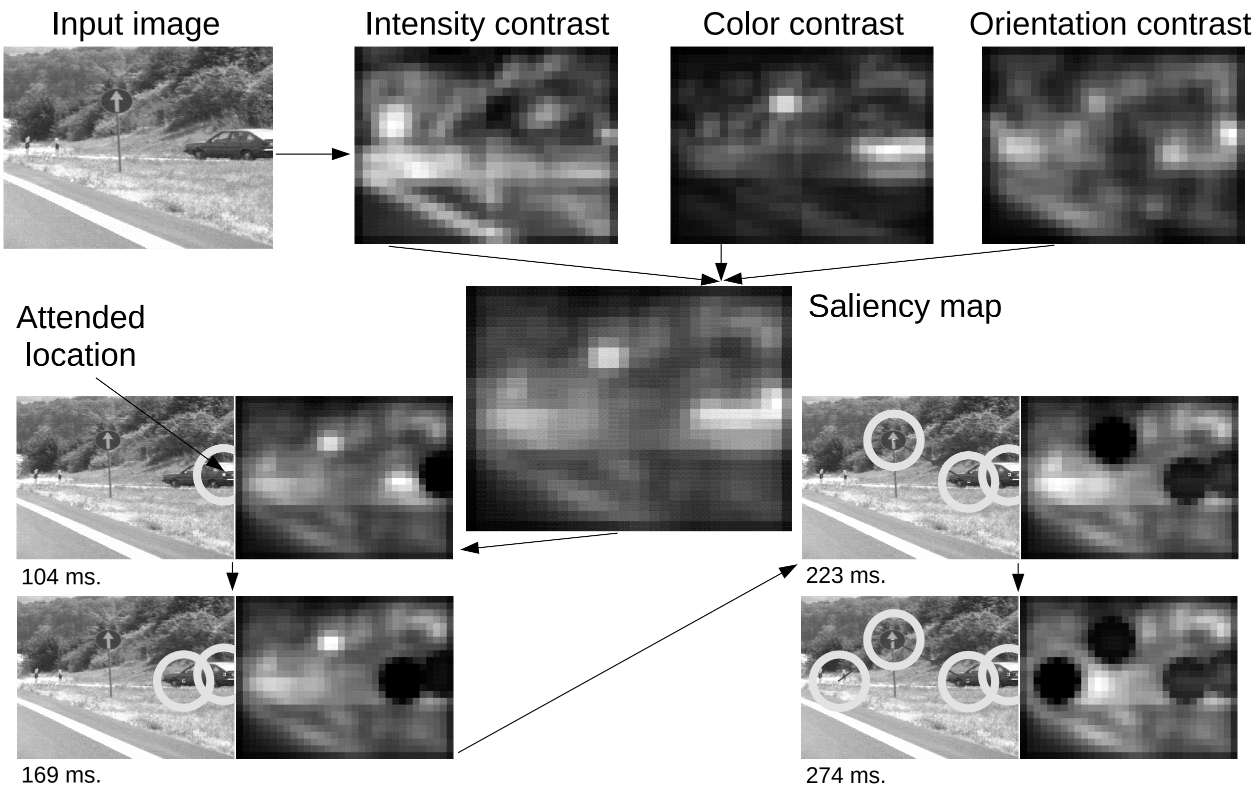

Figure 2. Example of operation of the model with a natural (color) image and the $\mathcal{N}(.)$ feature combination strategy (see Section 2.3). The most salient object is the (red) car, which appears strongly in both the Color Contrast and Orientation Contrast maps; it becomes the object of the first two attentional shifts (104 ms. and 169 ms. simulated time) and is subsequently suppressed in the saliency map by an "inhibition-of-return" mechanism. The next attended object is the (blue) traffic sign (223 ms.), and finally two smaller intersection indicators (274 ms.) More examples of model predictions on natural and synthetic images can be found at `http://www.klab.caltech.edu/~itti/attention/`

in our implementation), 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. 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 priory* 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. The final input to the saliency map is then the point-wise sum of all such feature maps.

All feature map 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:

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);
3. Update the weight following an additive learning rule independent of the map's dynamic range:

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

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 converged to zero, are no more computed. 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; potential divergence of the additive learning rule (explosion of weights) is hence avoided by constraining the weights to a fixed sum.

Note that here 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 non-linear 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;
3. Globally multiply the map by:
$$(M - \overline{m})^2. \tag{2}$$

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 non-linear normalization coarsely replicates a biological lateral inhibition mechanism, in which neighboring similar features inhibit each other.[12] This feature combination strategy is referred to in what follows as the "$\mathcal{N}(.)$" strategy.

2.4. Iterative localized interactions

The global non-linear normalization presented in the previous section is computationally very simple and is non-iterative, 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.

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 6-8mm in cortex) are believed to be organized in primary visual cortex.[13,14]

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 2D "difference of Gaussians" (DoG) filter, and negative results are clamped to zero after each iteration. The DoG filter, a 1D section of which is shown in **Figure 3**, yields strong local excitation at each visual location, which is counteracted by broad inhibition from neighboring locations. Specifically, such filter $\mathcal{DoG}(x)$ is obtained by:

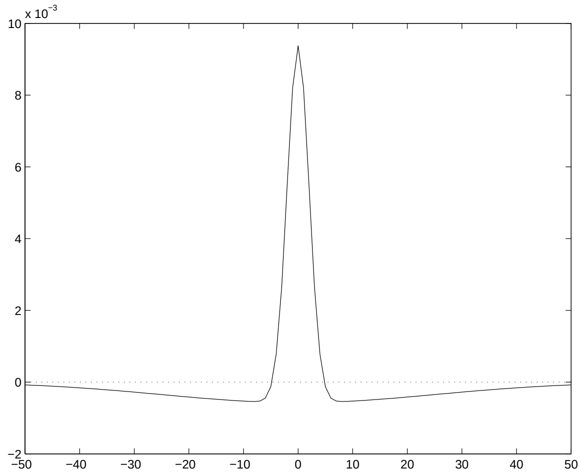

Figure 3. One-dimensional section of the 2D difference of Gaussians (DoG) filter used for iterative normalization of the feature maps. The central excitatory lobe strongly promotes each active location in the map, while the broader negative surround inhibits that location, if other strongly activated 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.

$$\mathcal{DoG}(x,y) = \frac{c_{ex}^2}{2\pi\sigma_{ex}^2} e^{-\frac{x^2+y^2}{2\sigma_{ex}^2}} - \frac{c_{inh}^2}{2\pi\sigma_{inh}^2} e^{-\frac{x^2+y^2}{2\sigma_{inh}^2}} \tag{3}$$

In our implementation, $\sigma_{ex} = 2\%$ and $\sigma_{inh} = 25\%$ of the input image width, $c_{ex} = 0.5$ and $c_{inh} = 1.5$ **(Figure 3)**. At each iteration of the normalization process, a given feature map $\mathcal{M}$ is then subjected to the following transformation:

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

where $\mathcal{DoG}$ is the 2D Difference of Gaussian filter described above, $|.|_{\geq 0}$ discards negative values, and C_{inh} is a constant inhibitory term ($C_{inh} = 0.02$ 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 (wrap-around) or "zero-padding" boundary conditions yields very strong edge effects which introduce unwanted non-uniform 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 $\mathcal{M}$ **(Figure 4)**. The truncated convolution is then computed as, using the fact that $\mathcal{G}$ is symmetric around its origin:

$$\mathcal{M} * \mathcal{G}(x) = \frac{\sum_i \mathcal{G}(i)}{\sum_{i \in \{\text{overlap}\}} \mathcal{G}(i)} \sum_{i \in \{\text{overlap}\}} \mathcal{M}(i)\mathcal{G}(i) \tag{5}$$

Using this "truncated filter" boundary condition yields uniform filtering over the entire image (see, e.g., **Figures 5, 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

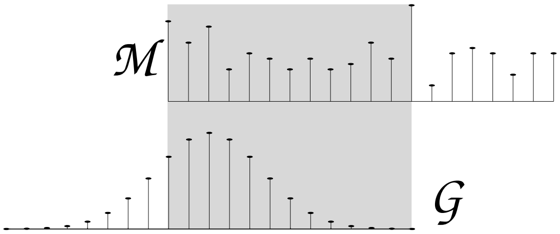

Figure 4. "Truncated filter" boundary condition consists of only computing the dot product between filter $\mathcal{G}$ and map $\mathcal{M}$ where they overlap (shaded area), and of normalizing the result by the total area of $\mathcal{G}$ divided by its area in the overlap region.

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 **Figures 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 (non-topographically) multiplied; for example, a map such as that in **Figure 5**

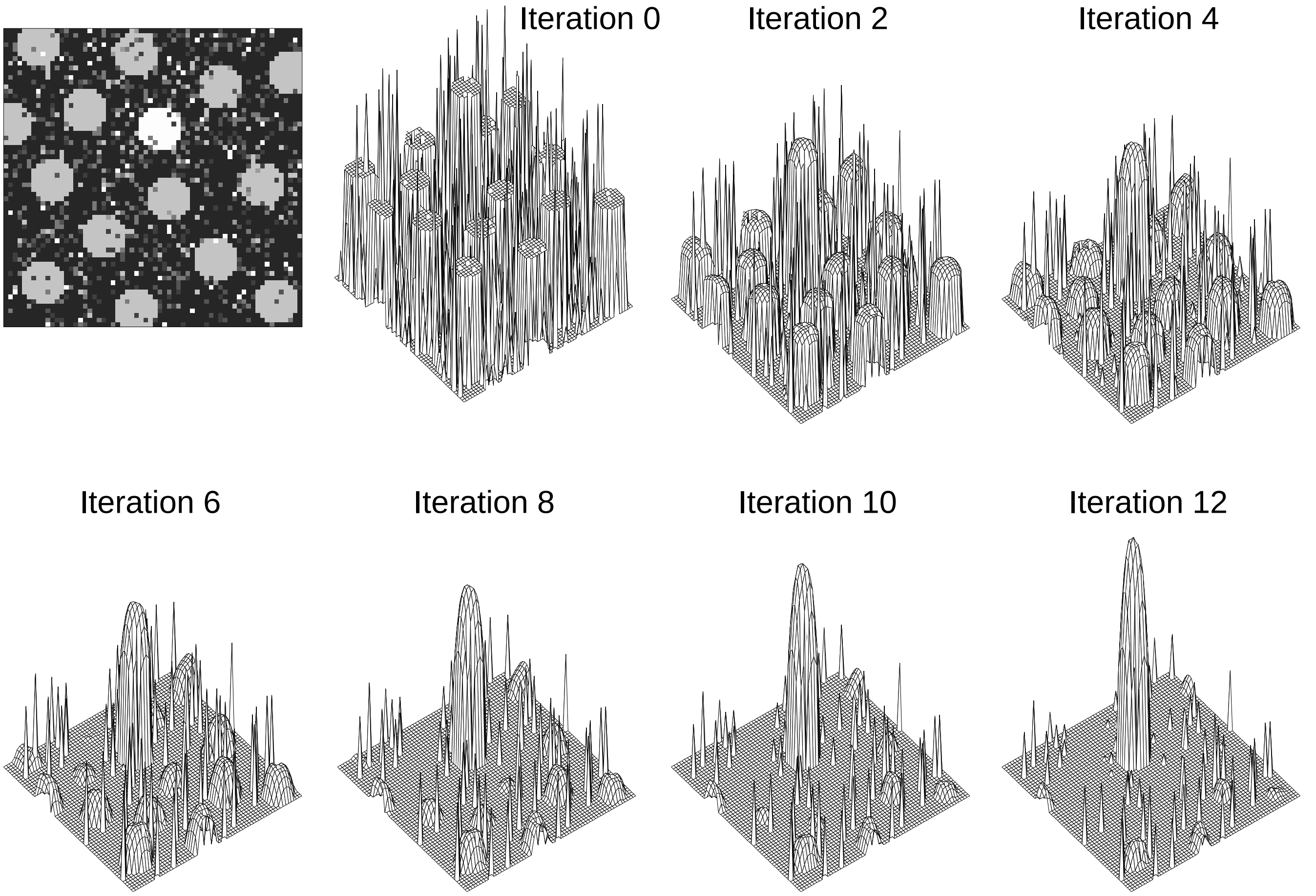

Figure 5. Iterative normalization of a feature map containing one strong activation peak surrounded by several weaker ones. After a few iterations, the initially stronger peak has gained in strength while at the same time suppressing weaker activation regions.

converges to a single activated pixel (at the center of the initial 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,[14] 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 multiscale framework of Gaussian Pyramids.[11] 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.[15] This normalization scheme will be referred to at the "Iterative" scheme in what follows.

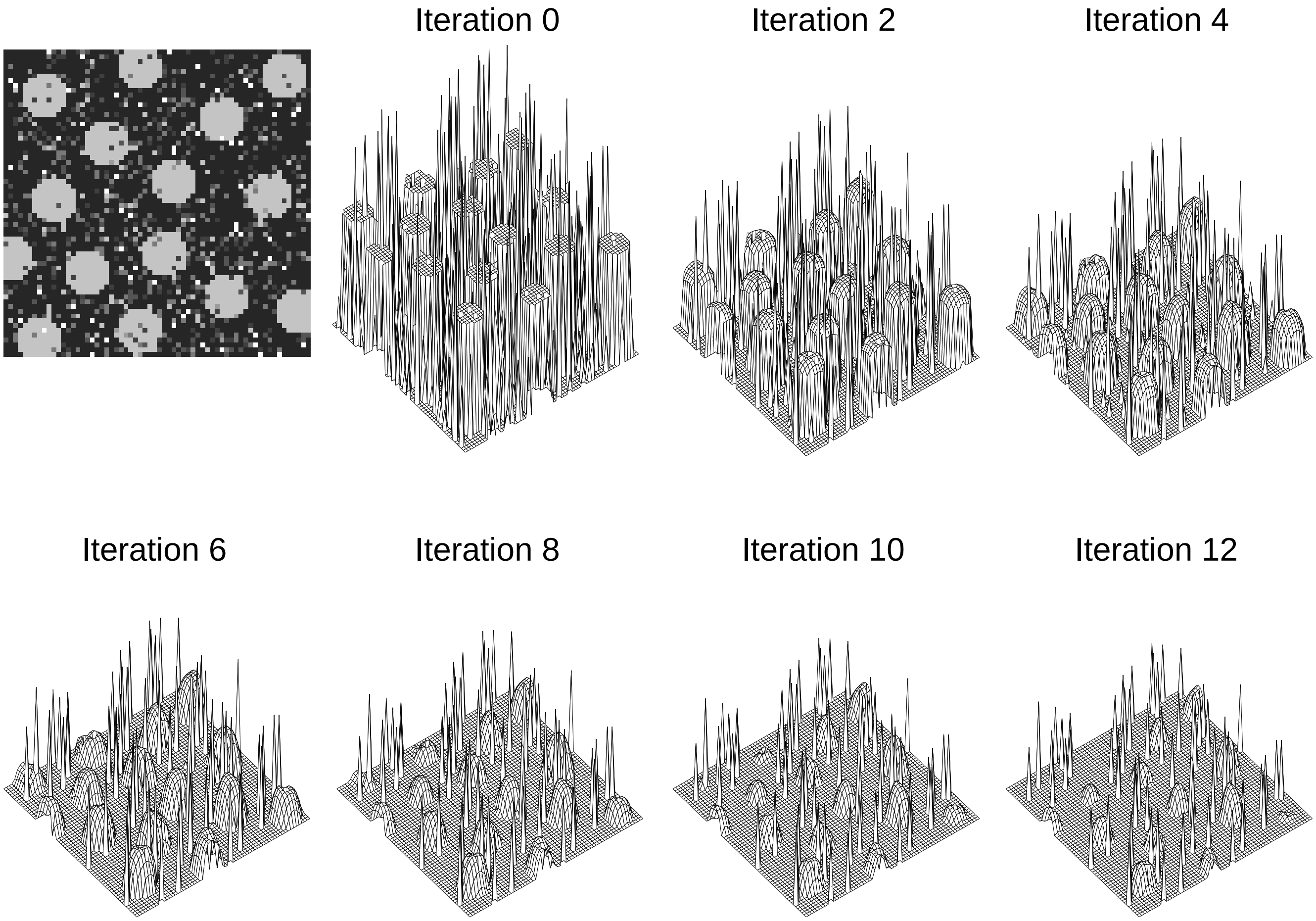

Figure 6. Iterative normalization of a feature map containing numerous strong activation peaks. This time, all peaks equally inhibit each other, resulting in the entire map being suppressed.

3. RESULTS AND DISCUSSION

We used three databases of natural color images to evaluate the different feature combination strategies proposed above. 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[11]). 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.[11] 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

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[1]	1.84 ± 2.13	0.49 ± 1.06	0.52 ± 1.05	0.24 ± 0.77
Traffic[2]	3.26 ± 2.80	1.27 ± 2.12	0.70 ± 1.18	0.77 ± 1.93

[1] before *first* sign found. [2] before *all* signs found.

images, 39 contained one traffic sign, 35 contained two, 12 contained three, 2 contained four, and 1 contained five traffic signs.

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 5 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 non-iterative "$\mathcal{N}(.)$" normalization;
3. Model with 12 iterations of the "Iterative" normalization;
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 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, 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 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 pre-attentively detect learned conjunctive features remains controversial.[10] Nevertheless, it was remarkable that the trained model performed best of the three alternative 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 non-iterative approximation $\mathcal{N}(.)$, yielded reliable yet non-specific 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 non-linearity (threshold) which suppresses negative values; as we can see in **Figures 5, 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 inhibits all other peaks.

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 non-linear "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.[16,17] Also, we have modeled interactions between neighboring regions of the visual field by simple self-excitation and subtractive neighbor-induced inhibition, while much more complicated patterns of interactions within the "non-classical receptive field" of visual neurons have been reported.[18,19]

In conclusion, we compared three 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, non-specific normalization which coarsely replicates biological within-feature spatial competition for saliency. Both the additive learning and the non-linear (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).

REFERENCES

1. C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum Neurobiol* **4**, pp. 219–27, 916 1985.
2. A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cogn Psychol* **12**(1), pp. 97–136, 1980.
3. J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg, "The representation of visual salience in monkey parietal cortex," *Nature* **391**, pp. 481–4, Jan 1998.
4. 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 Intell* **78**(1-2), pp. 507–45, 1995.
5. R. Rarasuraman, *The Attentive Brain*, MIT Press, Cambridge, MA, 1998.
6. 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**, pp. 4700–19, Nov 1993.
7. J. M. Wolfe, "Guided search 2.0: a revised model of visual search," *Psychon Bull Rev* **1**, pp. 202–38, 1994.

8. R. Milanese, S. Gil, and T. Pun, "Attentive mechanisms for dynamic and static scene analysis," *Opt Eng* **34**(8), pp. 2428–34, 1995.
9. S. Baluja and D. A. Pomerleau, "Expectation-based selective attention for visual monitoring and control of a robot vehicle," *Robotics Auton Sys* **22**(3-4), pp. 329–44, 1997.
10. E. Niebur and C. Koch, "Computational architectures for attention," in *The Attentive Brain*, R. Rarasuraman, ed., MIT Press, Cambridge, MA, 1998.
11. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans Patt Anal Mach Intell)* **20**(11), pp. 1254–9, 1998.
12. M. W. Cannon and S. C. Fullenkamp, "A model for inhibitory lateral interaction effects in perceived contrast," *Vision Res* **36**, pp. 1115–25, Apr 1996.
13. 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**, pp. 615–22, Jan 1996.
14. C. D. Gilbert and T. N. Wiesel, "Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex," *J Neurosci* **9**, pp. 2432–42, Jul 1989.
15. 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, ed., pp. 706–12, MIT Press, Cambridge, MA, 1997.
16. H. R. Wilson, "A transducer function for threshold and suprathreshold human vision," *Biol Cybern* **38**, pp. 171–8, Cyb 1980.
17. D. J. Heeger, "Normalization of cell responses in cat striate cortex," *Vis Neurosci* **9**, pp. 181–97, Aug 1992.
18. A. M. Sillito and H. E. Jones, "Context-dependent interactions and visual processing in v1," *J Physiol Paris* **90**, pp. 205–9, sio 1996.
19. A. M. Sillito, K. L. Grieve, H. E. Jones, J. Cudeiro, and J. Davis, "Visual cortical mechanisms detecting focal orientation discontinuities," *Nature* **378**, pp. 492–6, Nov 1995.