Short Papers

A Model of Saliency-Based Visual Attention for Rapid Scene Analysis

Laurent Itti, Christof Koch, and Ernst Niebur

Abstract—A visual attention system, inspired by the behavior and the neuronal architecture of the early primate visual system, is presented. Multiscale image features are combined into a single topographical saliency map. A dynamical neural network then selects attended locations in order of decreasing saliency. The system breaks down the complex problem of scene understanding by rapidly selecting, in a computationally efficient manner, conspicuous locations to be analyzed in detail.

Index Terms—Visual attention, scene analysis, feature extraction, target detection, visual search.

1 Introduction

PRIMATES have a remarkable ability to interpret complex scenes in real time, despite the limited speed of the neuronal hardware available for such tasks. Intermediate and higher visual processes appear to select a subset of the available sensory information before further processing [1], most likely to reduce the complexity of scene analysis [2]. This selection appears to be implemented in the form of a spatially circumscribed region of the visual field, the so-called "focus of attention," which scans the scene both in a rapid, bottom-up, saliency-driven, and task-independent manner as well as in a slower, top-down, volition-controlled, and task-dependent manner [2].

Models of attention include "dynamic routing" models, in which information from only a small region of the visual field can progress through the cortical visual hierarchy. The attended region is selected through dynamic modifications of cortical connectivity or through the establishment of specific temporal patterns of activity, under both top-down (task-dependent) and bottom-up (scene-dependent) control [3], [2], [1].

The model used here (Fig. 1) builds on a second biologically-plausible architecture, proposed by Koch and Ullman [4] and at the basis of several models [5], [6]. It is related to the so-called "feature integration theory," explaining human visual search strategies [7]. Visual input is first decomposed into a set of topographic feature maps. Different spatial locations then compete for saliency within each map, such that only locations which locally stand out from their surround can persist. All feature maps feed, in a purely bottom-up manner, into a master "saliency map," which topographically codes for local conspicuity over the entire visual scene. In primates, such a map is believed to be located in the posterior parietal cortex [8] as well as in the various visual maps in the pulvinar nuclei of the thalamus [9]. The model's saliency map is endowed with internal dynamics which generate attentional shifts. This model consequently represents a complete account of bottom-up saliency and does not require any top-down guidance to shift attention. This framework provides a massively parallel method for the fast selection of a small number of interesting image locations to be analyzed by more complex and time-consuming object-recognition processes. Extending this approach in "guided-search," feedback from higher cortical areas (e.g., knowledge about targets to be found) was used to weight the importance of different features [10], such that only those with high weights could reach higher processing levels.

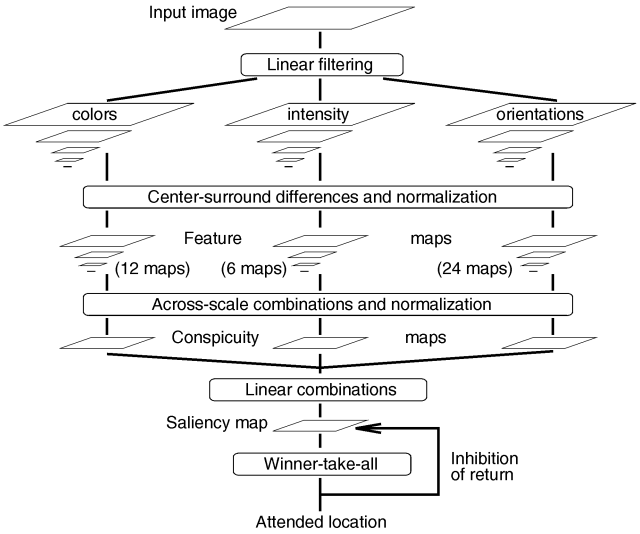

Fig. 1. General architecture of the model.

2 Model

Input is provided in the form of static color images, usually digitized at 640 × 480 resolution. Nine spatial scales are created using dyadic Gaussian pyramids [11], which progressively low-pass filter and subsample the input image, yielding horizontal and vertical image-reduction factors ranging from 1:1 (scale zero) to 1:256 (scale eight) in eight octaves.

Each feature is computed by a set of linear "center-surround" operations akin to visual receptive fields (Fig. 1): Typical visual neurons are most sensitive in a small region of the visual space (the center), while stimuli presented in a broader, weaker antagonistic region concentric with the center (the surround) inhibit the neuronal response. Such an architecture, sensitive to local spatial discontinuities, is particularly well-suited to detecting locations which stand out from their surround and is a general computational principle in the retina, lateral geniculate nucleus, and primary visual cortex [12]. Center-surround is implemented in the model as the difference between fine and coarse scales: The center is a pixel at scale $c \in \{2, 3, 4\}$, and the surround is the corresponding pixel at scale $s = c + \delta$, with $\delta \in \{3, 4\}$. The across-scale difference between two maps, denoted "$\ominus$" below, is obtained by interpolation to the finer scale and point-by-point subtraction. Using several scales not only for c but also for $\delta = s - c$ yields truly multiscale feature extraction, by including different size ratios between the center and surround regions (contrary to previously used fixed ratios [5]).

2.1 Extraction of Early Visual Features

With r, g, and b being the red, green, and blue channels of the in-

• *L. Itti and C. Koch are with the Computation and Neural Systems Program, California Institute of Technology—139-74, Pasadena, CA 91125. E-mail: {itti, koch}@klab.caltech.edu.*
• *E. Niebur is with the Johns Hopkins University, Krieger Mind/Brain Institute, Baltimore, MD 21218. E-mail: niebur@jhu.edu.*

Manuscript received 5 Feb. 1997; revised 10 Aug. 1998. Recommended for acceptance by D. Geiger.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 107349.

used to create a Gaussian pyramid $I(\sigma)$, where $\sigma \in [0..8]$ is the scale. The r, g, and b channels are normalized by I in order to decouple hue from intensity. However, because hue variations are not perceivable at very low luminance (and hence are not salient), normalization is only applied at the locations where I is larger than 1/10 of its maximum over the entire image (other locations yield zero r, g, and b). 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 $R(\sigma)$, $G(\sigma)$, $B(\sigma)$, and $Y(\sigma)$ are created from these color channels.

Center-surround differences ($\ominus$ defined previously) between a "center" fine scale c and a "surround" coarser scale s yield the feature maps. The first set of feature maps is concerned with intensity contrast, which, in mammals, is detected by neurons sensitive either to dark centers on bright surrounds or to bright centers on dark surrounds [12]. Here, both types of sensitivities are simultaneously computed (using a rectification) in a set of six maps $\mathcal{I}(c, s)$, with $c \in \{2, 3, 4\}$ and $s = c + \delta$, $\delta \in \{3, 4\}$:

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)|. \tag{1}$$

A second set of maps is similarly constructed for the color channels, which, in cortex, are represented using a so-called "color double-opponent" system: In the center of their receptive fields, neurons are excited by one color (e.g., red) and inhibited by another (e.g., green), while the converse is true in the surround. Such spatial and chromatic opponency exists for the red/green, green/red, blue/yellow, and yellow/blue color pairs in human primary visual cortex [13]. Accordingly, maps $\mathcal{RG}(c, s)$ are created in the model to simultaneously account for red/green and green/red double opponency (2) and $\mathcal{BY}(c, s)$ for blue/yellow and yellow/blue double opponency (3):

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \tag{2}$$

$$\mathcal{BY}(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))|. \tag{3}$$

Local orientation information is obtained from I using oriented Gabor pyramids $O(\sigma, \theta)$, where $\sigma \in [0..8]$ represents the scale and $\theta \in \{0°, 45°, 90°, 135°\}$ is the preferred orientation [11]. (Gabor filters, which are the product of a cosine grating and a 2D Gaussian envelope, approximate the receptive field sensitivity profile (impulse response) of orientation-selective neurons in primary visual cortex [12].) Orientation feature maps, $\mathcal{O}(c, s, \theta)$, encode, as a group, local orientation contrast between the center and surround scales:

$$\mathcal{O}(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)|. \tag{4}$$

In total, 42 feature maps are computed: six for intensity, 12 for color, and 24 for orientation.

2.2 The Saliency Map

The purpose of the saliency map is to represent the conspicuity—or "saliency"—at every location in the visual field by a scalar quantity and to guide the selection of attended locations, based on the spatial distribution of saliency. A combination of the feature maps provides bottom-up input to the saliency map, modeled as a dynamical neural network.

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 all 42 feature maps are combined, salient objects appearing strongly in only a few maps may be masked by noise or by less-salient objects present in a larger number of maps.

In the absence of top-down supervision, we propose a map normalization operator, $\mathcal{N}(.)$, which globally promotes maps in which a small number of strong peaks of activity (conspicuous locations) is present, while globally suppressing maps which contain

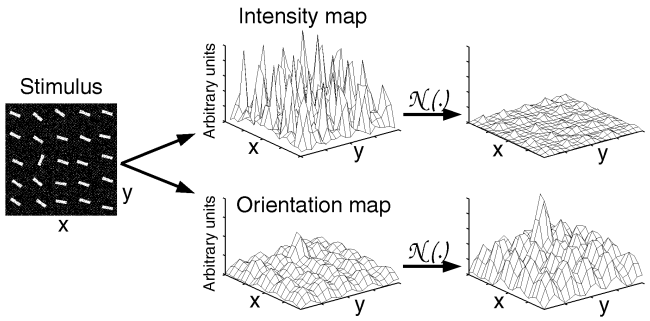

Fig. 2. The normalization operator $\mathcal{N}(.)$.

1) normalizing the values in the map to a fixed range $[0..M]$, in order to eliminate modality-dependent amplitude differences;
2) finding the location of the map's global maximum M and computing the average $\overline{m}$ of all its other local maxima; and
3) globally multiplying the map by $(M - \overline{m})^2$.

Only local maxima of activity are considered, such that $\mathcal{N}(.)$ compares responses associated with meaningful "activitation spots" in the map and ignores homogeneous areas. Comparing the maximum activity in the entire map to the average overall activation measures how different the most active location is from the average. When this difference is large, the most active location stands out, and the map is strongly promoted. When the difference is small, the map contains nothing unique and is suppressed. The biological motivation behind the design of $\mathcal{N}(.)$ is that it coarsely replicates cortical lateral inhibition mechanisms, in which neighboring similar features inhibit each other via specific, anatomically defined connections [15].

Feature maps are combined into three "conspicuity maps," $\overline{\mathcal{I}}$ for intensity (5), $\overline{\mathcal{C}}$ for color (6), and $\overline{\mathcal{O}}$ for orientation (7), at the scale ($\sigma = 4$) of the saliency map. They are obtained through across-scale addition, "$\oplus$," which consists of reduction of each map to scale four and point-by-point addition:

$$\overline{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} \mathcal{N}(\mathcal{I}(c, s)) \tag{5}$$

$$\overline{\mathcal{C}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \left[\mathcal{N}(\mathcal{RG}(c, s)) + \mathcal{N}(\mathcal{BY}(c, s))\right]. \tag{6}$$

For orientation, four intermediary maps are first created by combination of the six feature maps for a given θ and are then combined into a single orientation conspicuity map:

$$\overline{\mathcal{O}} = \sum_{\theta \in \{0°, 45°, 90°, 135°\}} \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c, s, \theta))\right). \tag{7}$$

The motivation for the creation of three separate channels, $\overline{\mathcal{I}}$, $\overline{\mathcal{C}}$, and $\overline{\mathcal{O}}$, and their individual normalization is the hypothesis that similar features compete strongly for saliency, while different modalities contribute independently to the saliency map. The three conspicuity maps are normalized and summed into the final input S to the saliency map:

$$S = \frac{1}{3}\left(\mathcal{N}(\overline{\mathcal{I}}) + \mathcal{N}(\overline{\mathcal{C}}) + \mathcal{N}(\overline{\mathcal{O}})\right). \tag{8}$$

At any given time, the maximum of the saliency map (SM) defines the most salient image location, to which the focus of attention (FOA) should be directed. We could now simply select the most active location as defining the point where the model should next attend. However, in a neuronally plausible implementation, we model the SM as a 2D layer of leaky *integrate-and-fire* neurons at scale four. These model neurons consist of a single capacitance

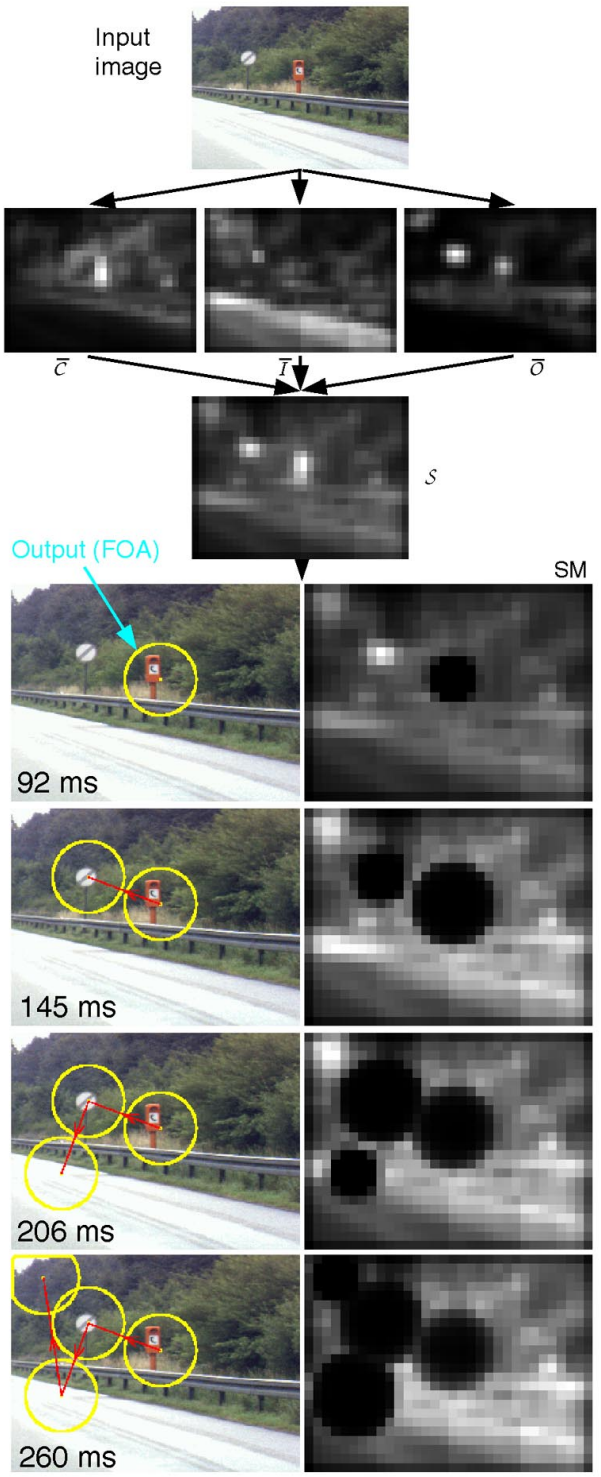

Fig. 3. Example of operation of the model with a natural image. Parallel feature extraction yields the three conspicuity maps for color contrasts ($\bar{C}$), intensity contrasts ($\bar{I}$), and orientation contrasts ($\bar{O}$). These are combined to form input S to the saliency map (SM). The most salient location is the orange telephone box, which appeared very strongly in $\bar{C}$; it becomes the first attended location (92 ms simulated time). After the inhibition-of-return feedback inhibits this location in the saliency map, the next most salient locations are successively selected.

reached, a prototypical spike is generated, and the capacitive scale $\sigma = 4$, in which synaptic interactions among units ensure that only the most active location remains, while all other locations are suppressed.

The neurons in the SM receive excitatory inputs from S and are all independent. The potential of SM neurons at more salient locations hence increases faster (these neurons are used as pure integrators and do not fire). Each SM neuron excites its corresponding WTA neuron. All WTA neurons also evolve independently of each other, until one (the "winner") first reaches threshold and fires. This triggers three simultaneous mechanisms (Fig. 3):

1) The FOA is shifted to the location of the winner neuron;
2) the global inhibition of the WTA is triggered and completely inhibits (resets) all WTA neurons;
3) local inhibition is transiently activated in the SM, in an area with the size and new location of the FOA; this not only yields dynamical shifts of the FOA, by allowing the next most salient location to subsequently become the winner, but it also prevents the FOA from immediately returning to a previously-attended location.

Such an "inhibition of return" has been demonstrated in human visual psychophysics [16]. In order to slightly bias the model to subsequently jump to salient locations spatially close to the currently-attended location, a small excitation is transiently activated in the SM, in a near surround of the FOA ("proximity preference" rule of Koch and Ullman [4]).

Since we do not model any top-down attentional component, the FOA is a simple disk whose radius is fixed to one-sixth of the smaller of the input image width or height. The time constants, conductances, and firing thresholds of the simulated neurons were chosen (see [17] for details) so that the FOA jumps from one salient location to the next in approximately 30–70 ms (simulated time), and that an attended area is inhibited for approximately 500–900 ms (Fig. 3), as has been observed psychophysically [16]. The difference in the relative magnitude of these delays proved sufficient to ensure thorough scanning of the image and prevented cycling through only a limited number of locations. All parameters are fixed in our implementation [17], and the system proved stable over time for all images studied.

2.3 Comparison With Spatial Frequency Content Models

Reinagel and Zador [18] recently used an eye-tracking device to analyze the local spatial frequency distributions along eye scan paths generated by humans while free-viewing gray-scale images. They found the spatial frequency content at the fixated locations to be significantly higher than, on average, at random locations. Although eye trajectories can differ from attentional trajectories under volitional control [1], visual attention is often thought as a pre-occulomotor mechanism, strongly influencing free-viewing. It was, hence, interesting to investigate whether our model would reproduce the findings of Reinagel and Zador.

We constructed a simple measure of spatial frequency content (SFC): At a given image location, a 16×16 image patch is extracted from each $I(2)$, $R(2)$, $G(2)$, $B(2)$, and $Y(2)$ map, and 2D Fast Fourier Transforms (FFTs) are applied to the patches. For each patch, a threshold is applied to compute the number of nonnegligible FFT coefficients; the threshold corresponds to the FFT amplitude of a just-perceivable grating (1 percent contrast). The SFC measure is the average of the numbers of nonnegligible coefficients in the five corresponding patches. The size and scale of the patches were chosen such that the SFC measure is sensitive to approximately the same frequency and resolution ranges as our model; also, our SFC measure is computed in the RGB channels as well as in intensity,

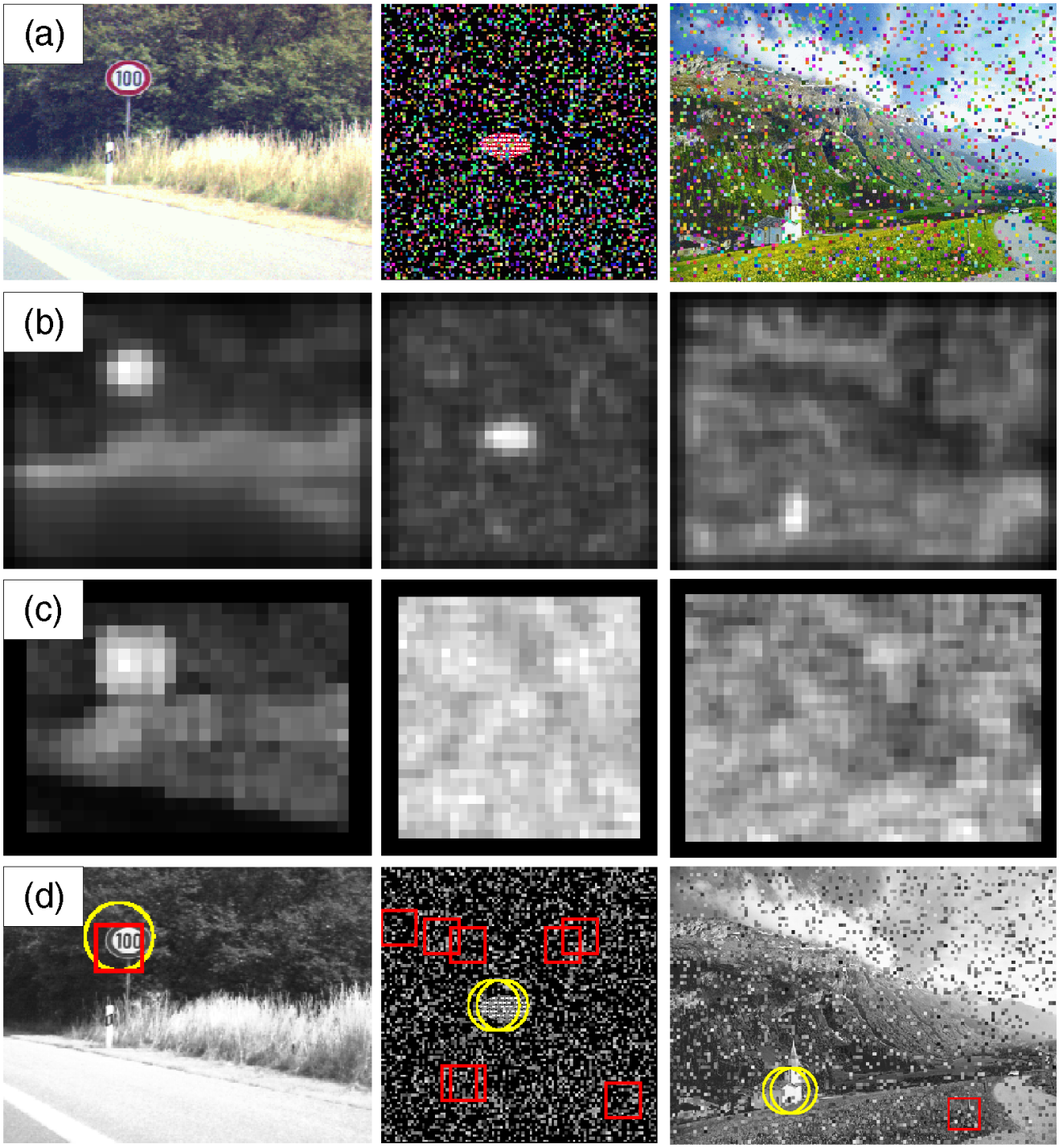

Fig. 4. (a) Examples of color images. (b) The corresponding saliency map inputs. (c) Spatial frequency content (SFC) maps. (d) Locations at which input to the saliency map was higher than 98 percent of its maximum (yellow circles) and image patches for which the SFC was higher than 98 percent of its maximum (red squares). The saliency maps are very robust to noise, while SFC is not.

3 Results and Discussion

Although the concept of a saliency map has been widely used in FOA models [1], [3], [7], little detail is usually provided about its construction and dynamics. Here, we examine how the feed-forward feature-extraction stages, the map combination strategy, and the temporal properties of the saliency map all contribute to

3.1 General Performance

The model was extensively tested with artificial images to ensure proper functioning. For example, several objects of the same shape but varying contrast with the background were attended to in the order of decreasing contrast. The model proved very robust to the addition of noise to such images (Fig. 5), particularly if the prop-

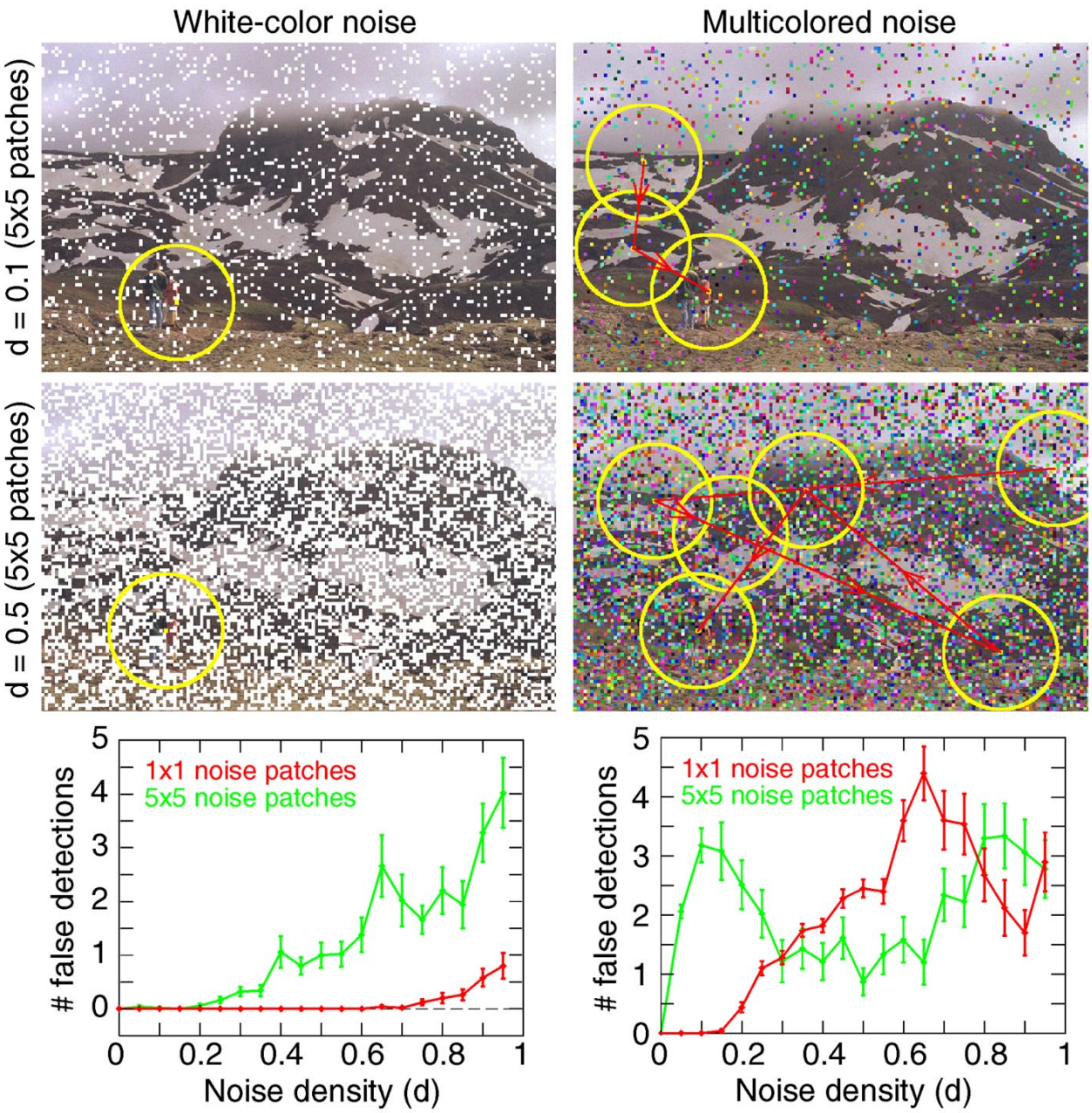

Fig. 5. Influence of noise on detection performance, illustrated with a 768×512 scene in which a target (two people) is salient by its unique color contrast. The mean ± S.E. of false detections before target found is shown as a function of noise density for 50 instantiations of the noise. The system is very robust to noise which does not directly interfere with the main feature of the target (left; intensity noise and color target). When the noise has similar properties to the target, it impairs the target's saliency and the system first attends to objects salient for other features (here, coarse-scale variations of intensity).

The model was able to reproduce human performance for a number of pop-out tasks [7], using images of the type shown in Fig. 2. When a target differed from an array of surrounding distractors by its unique orientation (as in Fig. 2), color, intensity, or size, it was always the first attended location, irrespective of the number of distractors. Contrarily, when the target differed from the distractors only by a conjunction of features (e.g., it was the only red horizontal bar in a mixed array of red vertical and green horizontal bars), the search time necessary to find the target increased linearly with the number of distractors. Both results have been widely observed in humans [7] and are discussed in Section 3.2.

Model predictions were compared to the measure of local SFC, in an experiment similar to that of Reinagel and Zador [18], using natural scenes with salient traffic signs (90 images), a red soda can (104 images), or a vehicle's emergency triangle symbol (64 images). Similar to Reinagel and Zador's findings, the SFC at attended lo-

the feature maps (Fig. 3 and [17]). With many such images, it is difficult to objectively evaluate the model, because no objective reference is available for comparison, and observers may disagree on which locations are the most salient. However, in all images studied, most of the attended locations were objects of interest, such as faces, flags, persons, buildings, or vehicles.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.