Visual Saliency Based on Scale-Space Analysis in the Frequency Domain

Jian Li, *Student Member*, *IEEE*, Martin D. Levine, *Fellow*, *IEEE*, Xiangjing An, *Member*, *IEEE*, Xin Xu, *Senior Member*, *IEEE*, and Hangen He

Abstract—We address the issue of visual saliency from three perspectives. First, we consider saliency detection as a frequency domain analysis problem. Second, we achieve this by employing the concept of *nonsaliency*. Third, we simultaneously consider the detection of salient regions of different size. The paper proposes a new bottom-up paradigm for detecting visual saliency, characterized by a scale-space analysis of the amplitude spectrum of natural images. We show that the convolution of the *image amplitude spectrum* with a low-pass Gaussian kernel of an appropriate scale is equivalent to an image saliency detector. The saliency map is obtained by reconstructing the 2D signal using the original phase and the amplitude spectrum, filtered at a scale selected by minimizing saliency map entropy. A Hypercomplex Fourier Transform performs the analysis in the frequency domain. Using available databases, we demonstrate experimentally that the proposed model can predict human fixation data. We also introduce a new image database and use it to show that the saliency detector can highlight both small and large salient regions, as well as inhibit repeated distractors in cluttered images. In addition, we show that it is able to predict salient regions on which people focus their attention.

Index Terms—Visual attention, saliency, hypercomplex Fourier transform, eye tracking, scale space analysis

1 Introduction

Visual attention facilitates our ability to rapidly locate the most important information in a scene [1], [2]. Such image regions are said to be salient since it is assumed that they attract greater attention by the visual system than other parts of the image. These salient regions are expected to possess distinctive features when compared with others in the image. The study of saliency detection may reveal the attentional mechanisms of biological visual systems, as well as model their fixation selection behavior. On the other hand, as a component of low-level artificial vision processing, it facilitates subsequent processing such as object detection or recognition by reducing computational cost, which is a key consideration in real-time applications. For object detection, this would always be more efficient than dense sampling, provided one could ensure the accuracy of the attentional mechanism.

Visual saliency detection has received extensive attention by both psychologists and computer vision researchers [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], and many models have been proposed based on different assumptions. Generally speaking, there are two different processes that influence visual saliency: One is top-down and depends on the task at hand and the other is bottom-up, which is driven by the input image. The focus of the paper is bottom-up saliency for selecting attentional regions.

Many bottom-up computational models that simulate primate perceptual abilities have appeared in the literature. For example, in [3], [31], [16] a center-surround mechanism is used to define saliency across scales, which is inspired by the putative neural mechanism. It has also been hypothesized that some visual inputs are intrinsically salient in certain background contexts and that these are actually task independent [3], [31]. This model has established itself as the exemplar for saliency detection and is consistently used for comparison in the literature. Similarly, there are also several proposed models which use other types of local information in different ways. In [17], saliency is defined as the local complexity. Gao et al. [29], [32], [33] proposed a bottom-up saliency model by using Kullback-Leibler (KL) divergence to measure the difference between a location and its surrounding area. In [5], a model of overt attention with selection based on self-information is proposed where the patches of an image are decomposed into a set of prelearned bases and kernel density estimation is used to approximate the self-information.

Several models have been suggested to compute saliency using global information. In [18], the authors first transform the input color image into the *Lab* color space and then define the saliency at each location as the difference between the *Lab* pixel value and the mean *Lab* value of the entire image. Harel et al. [7] proposed a graph-based solution that uses local computation to obtain a saliency map, which is everywhere dependent on global information. In [19], a saliency model called "extended saliency" was proposed in which the "global exceptions" concept is

- *J. Li, X. An, X. Xu, and H. He are with the Institute of Automation, National University of Defense Technology, Changsha 410073, Hunan Province, P.R. China. E-mail: lijian@nudt.edu.cn, {anxiangjing, hehangen}@gmail.com, xuxin_mail@263.net.*
- *M.D. Levine is with the Department of Electrical and Computer Engineering and the Centre for Intelligent Machines (CIM), McGill University, 3480 University Street, Montreal, QC H3A 2A7, Canada. E-mail: m.levine@ieee.org.*

Manuscript received 18 Sept. 2011; revised 1 Apr. 2012; accepted 25 June 2012; published online 11 July 2012.
Recommended for acceptance by R. Manmatha.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-09-0660.
Digital Object Identifier no. 10.1109/TPAMI.2012.147.

used to replace the traditional preference for local contrast. Recently, a simple and fast algorithm, called the spectrum residual (SR), was proposed based on the Fourier Transform [20]. The paper argues that the *spectrum residual* corresponds to image saliency. Following this, the Phase spectrum of the Fourier Transform (PFT) was introduced, which achieved nearly the same performance as the SR [30]. Based on PFT, PQFT [30] was also proposed by combining more features and using the quaternion Fourier Transform.

In this paper, we address the issue from three perspectives. Inspired by [20], we first consider saliency detection as a frequency domain problem. Unlike recent approaches which model saliency as a local phenomenon, we propose a new frequency domain paradigm which permits the full use of global information. Second, instead of modeling saliency in an image, we define the concept of *nonsaliency* using global information. Although in this paper we are solely concerned with determining saliency computationally, it is also interesting to consider the biological point of view. Research has suggested that objects viewed by the human visual system are thought to compete with each other to selectively focus our attention on a subset [34], [21]. Objects that appear in the visual field will influence how they are viewed by suppressing each other. Consequently, many are inhibited, while those that are not will ultimately predominate in the visual cortex to provide a focus of attention. In this paper, we model these inhibited regions as nonsaliency. Compared with salient regions, which are very distinctive in the image, nonsaliency can usually be modeled by common or uniform regions. These are then suppressed, thereby permitting salient objects to literally pop out. In this paper, nonsaliency is modeled in the frequency domain. Third, we also address another issue, that of detecting salient regions of different sizes. To date, there is no consistent definition of saliency in the literature. The models of saliency are diverse. In several models, saliency detection mimics the fixation selection mechanism and tends to find small distinct regions or points, for example, SR [20], PFT [30], PQFT [22], and AIM [5]. However, these may fail when detecting large saliency regions. Other papers tend to find large salient regions [18], [23], [24], [25], [35], [26]. Recently, scale-aware saliency [36] has been introduced to alleviate the problem of fixed scale in the spatial domain. We consider both small salient points as well as salient regions. For convenience, we will refer to both of these as salient regions, but of different size. We will show that the size of saliency regions is related to a scale parameter in the frequency domain.

We propose a new framework for saliency detection which ostensibly, at first sight, seems to be similar to the *Convolution Theorem* but in fact we will show that it is not. We will demonstrate that the convolution of the amplitude spectrum with a Gaussian kernel of an appropriate scale is equivalent to a saliency detector. The proposed framework has the ability to both highlight small and large salient regions and to inhibit repeated distractors in cluttered images.

The contribution of this paper is threefold: 1) A frequency domain paradigm for saliency detection is proposed, 2) the detection of both small and large salient regions is treated as a whole in the proposed model, 3) we show that SR, PFT, and the frequency-tuned model [18] are, to some extent, special cases of the proposed model.

The paper is organized as follows: Section 2 is the description and review of related work. In Section 3, we present the theoretical background of the proposed framework, called Spectrum Scale Space analysis (SSS). We present the saliency model based on the Hypercomplex Fourier Transform (HFT) in Section 4. In Section 5, we discuss experimental results. Concluding remarks and possible extensions are discussed in Section 6.

2 Related Work

Recently, a simple and fast algorithm called the *Spectrum Residual* was proposed in [20]. This paper argues that the spectrum residual corresponds to image saliency. Thus, given an image $f(x, y)$, it was first transformed into the frequency domain: $f(x, y) \xrightarrow{\mathcal{F}} \mathcal{F}(f)(u, v)$. The amplitude $\mathcal{A}(u, v) = |\mathcal{F}(f)|$ and phase $\mathcal{P}(u, v) = angle(\mathcal{F}(f))$ spectra are calculated, and then the log amplitude spectrum is obtained: $\mathcal{L}(u, v) = log(\mathcal{A}(u, v))$. Given these definitions, the spectrum residual was defined as

$$\mathcal{R}(u, v) = \mathcal{L}(u, v) - h_n \star \mathcal{L}(u, v), \tag{1}$$

and the saliency map $\mathcal{S}(x, y)$ of the original image as

$$\mathcal{S}(x, y) = \mathcal{F}^{-1}[exp(\mathcal{R}(u, v) + i \cdot \mathcal{P}(u, v))]. \tag{2}$$

In order to obtain a better visual display, the final saliency map was actually presented as[1]

$$\mathcal{S}(x, y) = g \star |\mathcal{F}^{-1}[exp(\mathcal{R}(u, v) + i \cdot \mathcal{P}(u, v))]|^2, \tag{3}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier and inverse Fourier Transforms, respectively; h_n and g are low-pass filters; i is the imaginary function; $\mathcal{P}(u, v)$ denotes the phase spectrum of the image, *which is assumed to be preserved* when transforming back to the spatial domain. Equations (1)-(3) are from [20]. The spectrum residual is the key idea of the SR, and the authors argued that it is this residual, combined with the original phase spectrum, that corresponds to the saliency in the image. However, in this paper, we will: 1) show that the spectrum residual is of little significance; 2) show that, for natural images, SR (or other similar models such as PFT [30]) are, to some extent, equivalent to a gradient operator; and 3) provide an explanation of why SR works in certain cases.

For convenience, we rewrite the standard *inverse Fourier Transform* as follows:

$$f(x, y) = \mathcal{F}^{-1}[exp(log\mathcal{A}(u, v) + i \cdot \mathcal{P}(u, v))], \tag{4}$$

$$\Leftrightarrow f(x, y) = \mathcal{F}^{-1}[\mathcal{A}(u, v) \cdot exp(i \cdot \mathcal{P}(u, v))], \tag{5}$$

$$\Leftrightarrow f(x, y) = \mathcal{F}^{-1}[\mathcal{F}(f)(u, v)]. \tag{6}$$

Thus, we can rewrite (2) as follows:

$$\mathcal{S}(x, y) = \mathcal{F}^{-1}[exp(\mathcal{R}(u, v) \cdot exp(i \cdot \mathcal{P}(u, v))], \tag{7}$$

1. In this paper, $|\cdot|^2$ indicates computing the square of each element in the matrix.

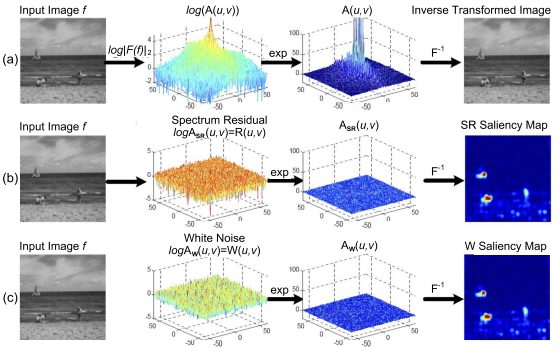

Fig. 1. Spectrum residual given by SR contains little information corresponding to image saliency. (a) Obviously, the original image is reproduced by performing the inverse FT using the original amplitude and phase spectrum. (b) In SR, it is argued that saliency map can be obtained by replacing the $log(\mathcal{A}(u,v))$ by the Spectrum Residual $\mathcal{R}(u,v)$. (c) If we replace the log amplitude spectrum $log\mathcal{A}(u,v)$ by random white noise(sic), we can obtain nearly the same saliency map.

Defining $exp(\mathcal{R}(u,v))$ as $\mathcal{A}_{SR}(u,v)$, (7) is rewritten as

$$\mathcal{S}(x,y) = \mathcal{F}^{-1}[\mathcal{A}_{SR}(u,v) \cdot exp(i \cdot \mathcal{P}(u,v))]. \tag{8}$$

Comparing (5) and (8), we observe that if we replace the amplitude spectrum $\mathcal{A}(u,v)$ by the exponential of $\mathcal{R}(u,v)$, the saliency map is obtained.[2] (See the comparison in Figs. 1a and 1b). This is the key idea of SR.

In order to illustrate that the spectrum residual is of little significance, we generate a 2D white noise signal $\mathcal{W}(u,v)$ which has the same average value and maximum as the spectrum residual $\mathcal{R}(u,v)$. We then use $\mathcal{W}(u,v)$ to replace the spectrum residual and perform the inverse Fourier Transform as follows:

$$\mathcal{S}(x,y) = \mathcal{F}^{-1}[exp(\mathcal{W}(u,v) \cdot exp(i \cdot \mathcal{P}(u,v))]. \tag{9}$$

Fig. 1c shows this process. If we define $exp(\mathcal{W}(u,v))$ as $\mathcal{A}_W(u,v)$, (9) can be rewritten as follows:

$$\mathcal{S}(x,y) = \mathcal{F}^{-1}[\mathcal{A}_W(u,v) \cdot exp(i \cdot \mathcal{P}(u,v))]. \tag{10}$$

Surprisingly, we can obtain nearly the same saliency map when we use white noise to replace the spectrum residual. This result clearly shows that the spectrum residual in [20] contains little information corresponding to saliency. Why is this the case? Comparing (8) and (10), we find that the amplitude spectra used to perform the inverse Fourier Transform are $A_{SR}(u,v)$ and $\mathcal{A}_W(u,v)$. As shown in the third columns of Figs. 1b, 1c, both $\mathcal{A}_{SR}(u,v)$ and $\mathcal{A}_W(u,v)$ are nearly horizontal planes compared (at the same scale) with $\mathcal{A}(u,v)$ shown in Fig. 1a. That is to say, in both (8) and (10), the amplitude information is totally abandoned and only phase information plays a role.

Two questions arise: 1) Why does SR yield a saliency map using only phase information? 2) More important, is there any information corresponding to image saliency contained in the amplitude spectrum? For the first question, our answer is that it only works for certain cases (detecting small salient regions in uncluttered scenes).

2. The phase spectra will no longer be plotted in the remaining figures in this paper, although obviously they exist and are required for computing the transforms.

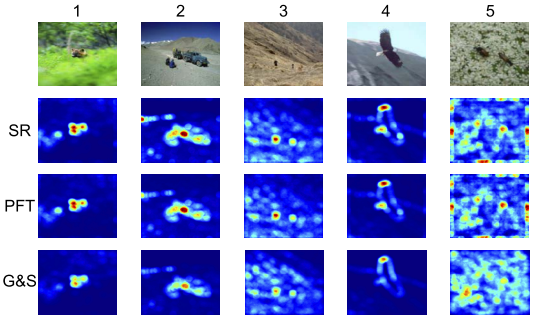

Fig. 2. G&S achieves nearly the same performance as SR and PFT.

Also, consider [30], [22] where the authors propose a new saliency model called the Phase Fourier Transform. The saliency is computed using only phase information as follows:

$$\mathcal{S}(x,y) = \mathcal{F}^{-1}[exp(i \cdot \mathcal{P}(u,v))], \tag{11}$$

$$\Leftrightarrow \mathcal{S}(x,y) = \mathcal{F}^{-1}[1(u,v) \cdot exp(i \cdot \mathcal{P}(u,v))]. \tag{12}$$

We observe that in PFT, the amplitude spectrum is (implicitly) also replaced by a horizontal plane. Therefore, we can deduce that, for natural images, both SR and PFT will produce nearly the same saliency map.

So what does using the inverse Fourier Transform solely with phase information imply? We argue that for natural images, both SR and PFT are, to a certain degree, equivalent to a gradient operator combined with Gaussian postprocessing (like the g in (3)). This is because the amplitude spectrum of natural images always has higher values at low than at high frequencies [37], [38]. Thus, if the amplitude spectrum is replaced by a horizontal plane, all of the frequencies are being treated equally. That is to say, the lower frequencies are suppressed and the higher frequencies are enhanced. As is well known, this implies a gradient enhancement operation. Based on the above discussion, we conclude that both SR and PFT will enhance the object boundaries and textured parts in an image. This indicates that they could work well only in detecting small salient regions where the center-surround contrast is very strong (see columns 1 and 3 of Fig. 2). However, they will have difficulty detecting large salient regions (column 4) and those in a cluttered background (column 5). To illustrate this point, we use a simple gradient operation combined with Gaussian postfiltering (as given by Algorithm 1 below) and obtain nearly the same performance as the other two methods, as shown in Fig. 2.

Why is the performance of these models inadequate? The reason is that the information contained in the amplitude spectrum has been abandoned.

Algorithm 1. Procedure for computing the **gradient and smoothing(G&S)**

Input:

The resized image $\mathcal{I}$ with resolution 128×128

Output:

Saliency map $\mathcal{S}$ of $\mathcal{I}$.

1: Convolve the input image with a Laplacian kernel, $L = [0\ -1\ 0;\ -1\ 4\ -1;\ 0\ -1\ 0]$. Obtain the gradient magnitude map Gra;

Fig. 3. Regular (repeated) and anomalous patterns. Top: Four images; bottom: collection of fragments from the last image above.

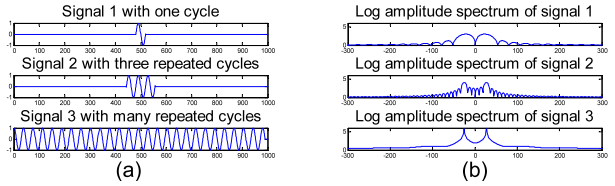

Fig. 4. Repeated patterns lead to sharp spikes. (a) Signals with different number of repeated cycles; (b) corresponding amplitude spectra.

2: Convolve Gra with a low-pass Gaussian filter kernel g, giving $\mathcal{S} = g \star |Gra|^2$;
3: **return** $\mathcal{S}$.

In the next section, we will discuss the question regarding whether the amplitude spectrum does contain any useful information about saliency. We will illustrate that the amplitude spectrum contains very important information and will develop a new framework for saliency detection in which we make full use of both the amplitude and phase.

3 Convolution of the Amplitude Spectrum with a Low-Pass Gaussian Kernel Equals a Saliency Detector

Many researchers have proposed models of saliency, which invariably then require the detection of salient *regions*. These regions are described as *distinctive* or *irregular* patterns which possess a distinct feature distribution when compared with the rest of the image. In this paper, instead of searching for these irregular patterns, we model regular or the so-called common patterns that do not attract much attention by our visual system. We refer to these patterns as being *nonsalient*.

3.1 Suppressing Repeated Patterns for Saliency Pop-Out

In the proposed model, we assume that a natural image consists of several salient and many so-called regular regions. All of these entities (whether distinct or not) may be considered as visual stimuli that compete for attention in the visual cortex. In this regard, it has been shown that nearby neurons constituting receptive fields in the visual cortex mutually inhibit each other and interact competitively [39]. As an example, in Fig. 3, if we divide the image into many patches (at a particular scale), we find that some are distinctive, while others are quite similar to each other. The bottom part of Fig. 3 shows the collection of patches from the last natural image above. We observe that several patterns appear many times (e.g., blue sky and grassy patches). We refer to these regular patches as *repeated patterns*, which correspond to *nonsaliency*.

Clearly, the primate visual system is more sensitive to distinctive rather than repeated patterns in an image. Furthermore, the latter are very diverse. For example, consider the top row of Fig. 3. These exhibit several different examples of repeated patterns at different scales (including at the "scale" of 0 frequency for the uniform areas): grassy and sky patches (image 4), similar objects (image 1), road patches of the same color and texture (image 2), the "L"s (image 3), and so on. We model these repeated patterns and then suppress them, thereby producing the pop-out of the salient objects.

3.2 Spikes in the Amplitude Spectrum Correspond to Repeated Patterns

In this paper, we will illustrate that the amplitude spectrum contains important information corresponding to saliency and nonsaliency. To be more precise, the spikes in the amplitude spectrum turn out to correspond to repeated patterns which should be suppressed for saliency detection.

For convenience, we take a 1D periodic signal $f(t)$ as an example. Suppose $f(t)$ can be represented by $f(t) = \sum_{n=-\infty}^{\infty} F(n) e^{jn\omega_1 t}$, where $F_n = \frac{1}{T}\int_{-T/2}^{T/2} f(t) e^{-jn\omega_1 t} dt$. Then the Fourier transform is given by

$$\mathcal{F}(w) = 2\pi \sum_{n=-\infty}^{\infty} F(n)\delta(\omega - n\omega_1). \tag{13}$$

From (13), we can conclude that the spectrum of a periodic signal (repeated cycles) is a set of impulse functions (spikes). We note that this is based on the assumption that the signal is infinite. Therefore, given a more realistic finite length periodic signal, the shape of the spectrum will obviously be different but not degraded greatly.

Fig. 4 provides an illustration of this point. Fig. 4a shows three signals with a different number of repeated patterns (cycles), while Fig. 4b shows their corresponding amplitude spectra. We observe that the larger the number of repeated cycles, the sharper the spikes in the spectrum. In order to quantify this notion, we define the *sharpness* of a spectrum X. We note that if we smooth the spikes by convolving the spectrum with a low-pass filter, the sharper the original spike, the more its peak height will be reduced. Therefore, the *sharpness* of X can be defined as $\gamma(X) = \| X - X \star h_m \|_\infty$, where h_m is a Gaussian kernel with fixed scale. The sharpness values of these three spectra in Fig. 4 are 0.2320, 0.6091, and 1.3227, respectively. Besides the sinusoid shown in the figure, other repeated signals also have this characteristic.

Next, suppose there is one salient part that is embedded in a finite length periodic signal (row 1 of Fig. 5). We will illustrate that this salient interval will not largely influence the spikes in the spectrum. That is to say, 1) the spikes will remain even though a salient part is embedded in the periodic signal, 2) the embedded salient part will not lead to very sharp spikes in the spectrum. The signal to be analyzed is defined as follows:

$$f(t) = g(t) + g_\tau(t) + s(t), \tag{14}$$

where $g(t)$ is a periodic signal of finite length L and equals $p(t)$ inside the interval $(0, L)$ and 0 elsewhere; $g_\tau(t) = -p(t) \cdot r(t)$; $s(t) = p_s(t) \cdot r(t)$, where $s(t)$ is the salient part of

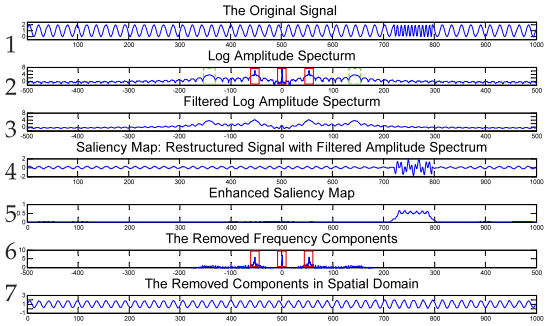

Fig. 5. Suppression of repeated patterns by using spectrum filtering. It is clear that the larger the repeated background, the sharper the spikes, leading to the suppression of the amplitude spectrum via filtering.

$f(t)$, which for convenience is also defined as a portion of yet another periodic function $p_s(t)$; $p(t)$ and $p_s(t)$ are periodic functions with frequencies ν and ν_s, respectively; $r(t)$ is a rectangular window function that equals 1 inside the interval $(t_0, t_0+\tau)$ and 0 elsewhere. We also suppose that $(t_0, t_0+\tau) \in (0, L)$ and $\tau \ll L$ (see row 1 of Fig. 5). Thus, the Fourier Transform of $f(t)$ can be represented as follows:

$$\begin{aligned}\mathcal{F}(f)(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt = \int_0^L g(t)e^{-j\omega t}dt \\ + \int_{t_0}^{t_0+\tau} g_\tau(t)e^{-j\omega t}dt + \int_{t_0}^{t_0+\tau} s(t)e^{-j\omega t}dt.\end{aligned} \tag{15}$$

From (15), the spectrum of $f(t)$ consists of three terms. We assume that $\tau \ll L$. This implies that the first term has very sharp spikes in the amplitude spectrum as it contains many repeated patterns, while this is not true of the second and third terms. Consider $g_\tau(t)$ as an example. $g_\tau(t)$ is the pointwise product of signal $-p(t)$ and $r(t)$. According to the convolution theorem, $\mathcal{F}(g_\tau)(\omega)$ equals the convolution of $-\mathcal{F}(p)(\omega)$ with $\mathcal{F}(r)(\omega)$. Since $\mathcal{F}(r)(\omega) = \frac{2sin(\tau/2)}{\omega} e^{j\omega(t_0+\tau/2)}$ is a low-pass filter, the spikes in the amplitude spectrum of $-\mathcal{F}(p)(\omega)$ will be greatly suppressed. That is to say, there are no sharp spikes in the second term. This also occurs for the third term. As discussed above, the sharpness of $\mathcal{F}(f)(\omega)$ is mainly determined by $g(t)$, while the latter two terms in (15) do not make a significant contribution to the spikes in the spectrum. In other words, since the first term corresponds to repeated patterns (nonsalient) which lead to spikes, they can be suppressed by smoothing (SM) the spikes in the amplitude spectrum of $\mathcal{F}(f)(\omega)$.

3.3 Suppressing Repeated Patterns Using Spectral Filtering

A Gaussian kernel h can be employed to suppress spikes in the amplitude spectrum $|\mathcal{F}\{f\}|$ of an image as follows[3]:

$$\mathcal{A}_\mathcal{S}(u,v) = |\mathcal{F}\{f(x,y)\}| \star h. \tag{16}$$

The resulting smoothed amplitude spectrum $\mathcal{A}_S$ and the *original* phase spectrum are combined to compute the inverse transform, which in turn, yields the saliency map:

$$\mathcal{S} = \mathcal{F}^{-1}\{\mathcal{A}_\mathcal{S}(u,v)e^{i\cdot\mathcal{P}(u,v)}\}. \tag{17}$$

In order to improve the visual display of saliency, we define it hereafter as

$$\mathcal{S} = g \star \left|\mathcal{F}^{-1}\{\mathcal{A}_\mathcal{S}(u,v)e^{i\cdot\mathcal{P}(u,v)}\}\right|^2. \tag{18}$$

Consider the very simple example shown in Fig. 5. The input signal (row 1) is periodic, but there is a short segment for which a different frequency signal is apparent. The short segment is quite distinct from the background for human vision, so a saliency detector should be able to highlight it. Row 2 shows the amplitude spectrum: There are three very sharp spikes (labeled by solid boxes), one of which corresponds to the constant background (uniform part) at zero frequency and the other two correspond to the periodic background. In addition, there are two rounded maxima (labeled by dashed boxes) corresponding to the salient parts. The amplitude spectrum is then smoothed by a Gaussian kernel (row 3), and the signal is reconstructed using the smoothed amplitude and original phase spectrum (row 4). It is clear that both the periodic and the uniform background are largely suppressed, while the salient segment is well preserved. Row 5 shows the saliency map after enhancing the signal shown in row 4 using postprocessing. We can further analyze this in the frequency domain, as shown in row 6, which illustrates the components actually removed by the previous operations. Here, the eliminated frequency components are mainly the low frequencies near zero frequency, as well as the periodic background. Row 7 presents these removed components in the spatial domain. We find that nonsalient parts (including uniform parts) are well suppressed using amplitude filtering. This process suggests that convolution in the frequency domain of the amplitude spectrum with a Gaussian kernel is equivalent to an image saliency detector.[4]

3.4 Spectrum Scale-Space Analysis

Repeated patterns (including uniform patterns) can be suppressed by smoothing the amplitude spectrum at an appropriate scale. However, which scale is the best in (16)? As shown in Fig. 6, if the filter scale is too small, the repeated patterns cannot be suppressed sufficiently (row 2), while if the filter scale is too large, only the boundaries of the salient region are highlighted (row 4 and 5). Therefore, it is important to select a proper scale for the Gaussian kernel. In fact, we will illustrate that different filter scales are required for different types of saliency. For example, a small-scale kernel is needed to detect large salient regions, while a large-scale kernel could be used to detect texture-rich or small salient regions (e.g., distant objects in the scene).

In this paper, we propose a Spectrum Scale-Space for handling amplitude spectra at different scales, yielding a one-parameter family of smoothed spectra which is parameterized by the scale of the Gaussian kernel. Given an amplitude spectrum, $\mathcal{A}(u,v)$, of an image, the SSS is a family of derived signals $\Lambda(u,v;k)$ defined by the convolution of $\mathcal{A}$ with the series of Gaussian kernels:

3. In the implementation of this equation, we found that suppressing spikes in the log amplitude spectrum rather than the amplitude spectrum yielded better results.

4. One might mistakenly be confused to think that this convolution in the frequency domain is equivalent to multiplication in the spatial domain as in the convolution theory. Yet, this is not the case as we convolve only the amplitude and do not change the phase.

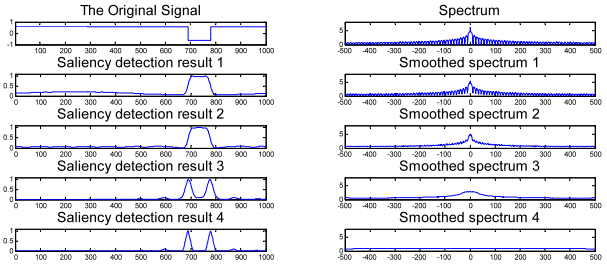

Fig. 6. The original 1D signal is shown in the first row of column 1 with the 1D saliency maps below. The spectrum of the original signal is shown in the first row of column 2, followed by the smoothed spectra associated with the paired saliency map in column 1.

$$g(u,v;k) = \frac{1}{\sqrt{2\pi}2^{k-1}t_0} e^{-\left(u^2+v^2\right)/\left(2^{2k-1}t_0^2\right)}, \tag{19}$$

where k is the scale parameter, $k = 1, \ldots, K$. K is determined by the image size: $K = \lceil log_2 min\{H, W\} \rceil + 1$, where H, W indicate the height and width of the image; $t_0 = 0.5$. Thus, the scale space is defined as follows:

$$\Lambda(u,v;k) = (g(.,.;k) \star \mathcal{A})(u,v). \tag{20}$$

As an example, assume a 1D signal. We first compute a series of filtered spectra according to the SSS model, then the saliency map is computed for each scale, as shown in Fig. 6. The significance of the scale for saliency detection can easily be observed. In this example, smoothed spectrum 2 gives the best result. As the kernel scale goes to infinity, the spectrum tends to be a constant (horizontal plane in 2D), as shown in the last row of Fig. 6. This is exactly the case proposed in [20], [30], [22].

Fig. 7 shows 2D results obtained using different kernel scales, increasing from left to right. The best saliency map is labeled by a red square. We observe that broad regions pop out when smaller scale kernels are used, while distant objects or those with rich texture pop out when larger scale kernels are used. Thus, given a natural image, a set of saliency maps is obtained from which one must be selected as the final saliency map. The criterion for achieving this will be discussed in Section 4.

Here, we suggest that the frequency-tuned model [18] is, to some extent, a special case of the proposed model. In [18], the saliency map is defined as: $\mathcal{S}(x,y) = \|I_\mu - I_{\omega hc}(x,y)\|$, where I_μ is the average Lab vector of the entire image and $I_{\omega hc}(x,y)$ is a specific Lab pixel vector from the Gaussian-filtered version of the original image. The authors compute a saliency map by removing the frequencies around the DC frequency (the "mean"). Previously, we have illustrated that there is always a very sharp spike around zero frequency, which corresponds to this "mean." Hence, if we use a very small scale Gaussian kernel to smooth the spectrum, those components corresponding to the "mean" will be suppressed significantly.

4 Saliency Using the Hypercomplex Fourier Transform

In Section 3, we discussed the saliency computation using only one feature map (that of intensity). However, in order to obtain better performance, more features are required, for example, color and motion information. Inspired by [30], [40], [41], we use the so-called hypercomplex matrix to combine multiple feature maps. Consequently, the *Hypercomplex Fourier Transform* is employed to replace the *Fourier Transform* used in Section 3 for saliency computing.

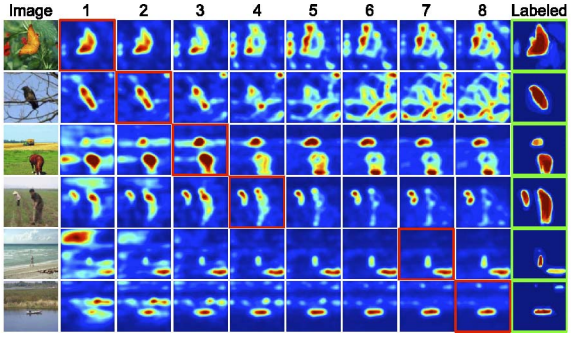

Fig. 7. Five 2D examples are shown. The first column shows the original 2D signals (images). The remaining images in each row present the set of saliency maps computed by smoothing the original image amplitude spectrum using different scales for the Gaussian kernels.

4.1 Hypercomplex Fourier Transform

The input to the traditional Discrete Fourier Transform is a real matrix. Each image pixel is an element of the input matrix and is a real number. However, if we combine more than one feature into a hypercomplex matrix, each element is a vector and this hypercomplex matrix is a *vector field*. Thus, the traditional Fourier Transform becomes unsuitable for computational purposes.

The Hypercomplex Fourier Transform was proposed in [40], in which the hypercomplex input was specified to be a quaternion.[5] Given a hypercomplex matrix

$$f(n,m) = a + bi + cj + dk, \tag{21}$$

the discrete version of the HFT of (21) is given by

$$\mathcal{F}_\mathcal{H}[u,v] = \frac{1}{\sqrt{MN}} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-\mu 2\pi((\frac{mv}{M})+(\frac{nu}{N}))} f(n,m), \tag{22}$$

where μ is a unit pure quaternion and $\mu^2 = -1$. Note that $\mathcal{F}_\mathcal{H}[u,v]$ is also a hypercomplex matrix. The inverse Hypercomplex Fourier Transform is given as

$$f(n,m) = \frac{1}{\sqrt{MN}} \sum_{v=0}^{M-1} \sum_{u=0}^{N-1} e^{\mu 2\pi((\frac{mv}{M})+(\frac{nu}{N}))} \mathcal{F}_\mathcal{H}[u,v]. \tag{23}$$

4.2 Hypercomplex Representation of Multiple Feature Maps

The Hypercomplex representation can be employed to combine multiple features (e.g., in [22] the authors combine color, intensity, and motion as the features). We define the input hypercomplex matrix as follows:

$$f(n,m) = w_1 f_1 + w_2 f_2 i + w_3 f_3 j + w_4 f_4 k, \tag{24}$$

5. The quaternion is represented as $q = a + bi + cj + dk$, where a, b, c, and d are real numbers and i, j, k satisfy $i^2 = j^2 = k^2 = ijk = -1$. A quaternion can also be represented as $q = S(q) + V(q)$, where $S(q) = a$ is the scalar and $V(q) = bi + cj + dk$ is the vector part. q is called a pure quaternion if $S(q) = 0$.

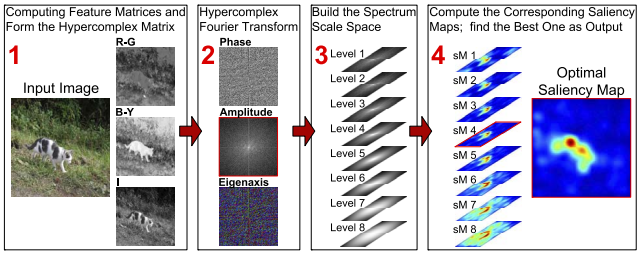

Fig. 8. Procedure for computing saliency using the Hypercomplex Fourier Transform.

where $w_1 - w_4$ are weights and $f_1 - f_4$ are the feature maps (matrices). Similarly to [3], we use three features to compute the saliency for the static input case:

$$f_2 = \mathcal{I}_s = (r + g + b)/3, \tag{25}$$

$$f_3 = \mathcal{RG} = R - G, \tag{26}$$

$$f_4 = \mathcal{BY} = B - Y, \tag{27}$$

where r, g, b are the red, green, and blue channels of an input color image and $R = r - (g + b)/2$, $G = g - (r + b)/2$, $B = b - (r + g)/2$, $Y = (r + g)/2 - |r - g|/2 - b$. These three feature maps comprise the opponent color space representation of the input image (see part 1 of Fig. 8). Based on the work in [22], our approach has also been experimentally confirmed using videos by defining a motion feature $\mathcal{M}$ and setting $f_1 = \mathcal{M}$ in (24). In this paper, we consider only the static image case by employing just intensity and color information. We select the weights so that $w_1 = 0$, $w_2 = 0.5$, $w_3 = w_4 = 0.25$.

4.3 Computing the Saliency Map

Given an image, the input is defined according to Section 4.2. The Hypercomplex Fourier Transform, $\mathcal{F}_{\mathcal{H}}[u, v]$, can be rewritten in polar form as follows:

$$\mathcal{F}_{\mathcal{H}}[u, v] = \|\mathcal{F}_{\mathcal{H}}[u, v]\| e^{\mu \Phi(u,v)}, \tag{28}$$

where $\|\cdot\|$ indicates the modulus for each element of a hypercomplex matrix; $\mathcal{F}_{\mathcal{H}}[u, v]$ can be considered as the frequency domain representation of $f(m, n)$. Its amplitude spectrum $\mathcal{A}(u, v)$, phase spectrum $\mathcal{P}(u, v)$, and the so-called eigenaxis spectrum $\mathcal{X}(u, v)$ are defined as

$$\mathcal{A}(u, v) = \|\mathcal{F}_{\mathcal{H}}(u, v)\|,$$

$$\mathcal{P}(u, v) = \Phi(u, v) = tan^{-1} \frac{\|\mathcal{V}(\mathcal{F}(u, v))\|}{\mathcal{S}(\mathcal{F}(u, v))},$$

$$\mathcal{X}(u, v) = \mu(u, v) = \frac{\mathcal{V}(\mathcal{F}(u, v))}{\|\mathcal{V}(\mathcal{F}(u, v))\|},$$

where $\mathcal{X}(u, v)$ is a pure quaternion matrix. These three spectra are shown in part 2 of Fig. 8.[6]

As discussed in Section 3, the amplitude spectrum contains important information about the scene. Similar to the discussion in Section 3.4, we create the *Spectrum Scale Space* $\Lambda = \{\Lambda_k\}$ by smoothing $\mathcal{A}(u, v)$ with a series of Gaussian kernels according to (20) (see Fig. 8 part 3) while retaining unchanged the phase spectrum $\mathcal{P}(u, v)$ and eigenaxis spectrum $\mathcal{X}(u, v)$.

Observing the images in part 3 of Fig. 8 reveals that when the scale k is very small, the information contained in the amplitude plots is retained quite well, while when it becomes very large, the pertinent information basically is lost. Actually, the PQFT model is a special case of the proposed framework when the scale goes to infinity.

Thus, given a single (smoothed) amplitude spectrum Λ_k (one layer in Λ) and the original phase and eigenaxis spectra, we can perform the inverse transform (23) to give the saliency map at each scale:

$$\mathcal{S}_k = g \star \|\mathcal{F}_{\mathcal{H}}^{-1}\{\Lambda_k(u, v) e^{\mathcal{X}\mathcal{P}(u,v)}\}\|^2, \tag{29}$$

where g is a Gaussian kernel at a fixed scale.[7] Thus, we again obtain a series of saliency maps $\{\mathcal{S}_k\}$, shown in part 4 of Fig. 8. In the approach proposed in this paper, the final saliency map $\mathcal{S}$ is chosen from $\{\mathcal{S}_k\}$ by selecting the best scale k_p according to criteria discussed in Section 4.4. The saliency model based on the Hypercomplex Fourier Transform is referred to as HFT in this paper.

The HFT model is summarized in Algorithm 2.[8]

Algorithm 2. HFT saliency model

Input:

The resized color image $\mathcal{C}$ with resolution $m \times n$

Output:

Saliency map $\mathcal{S}$ of $\mathcal{C}$

1: Compute the feature maps $\{\mathcal{I}, \mathcal{RG}, \mathcal{BY}\}$ of $\mathcal{C}$ according to (25)-(27);
2: Form the hypercomplex matrix $f(n, m)$ by combining these feature maps according to (24);
3: Perform the Hypercomplex Fourier Transform on $f(n, m)$ and compute the amplitude spectrum $\mathcal{A}$, phase spectrum, $\mathcal{P}$ and eigenaxis spectrum $\mathcal{X}$;
4: Smooth the amplitude spectrum with Gaussian kernels according to (19), thereby obtaining a spectrum scale space $\{\Lambda_k\}$;
5: Obtain a saliency map $\mathcal{S}_k$ according to (29) for each Λ_k, thereby producing a sequence of saliency maps $\{\mathcal{S}_k\}$;
6: Find the best saliency map $\mathcal{S}$ from the set $\{\mathcal{S}_k\}$ and use it as the final saliency map according to the criterion introduced in (31) in section 4.4;
7: **return** $\mathcal{S}$.

4.4 Finding the Proper Scale

In Section 3, we assumed that the best saliency map would appear at a specific scale in the sequence $\{\mathcal{S}_k\}$. Unlike the use of *entropy* in [17], we employ it as the criterion for determining the optimal scale:

$$k_p = \arg\min_k \{\mathcal{H}(\mathcal{S}_k)\}, \tag{30}$$

6. Here, we use a monochrome image to represent the phase spectrum $\mathcal{P}(u, v)$ as it is a real matrix. This is different from [41].

7. For convenience, the scale parameter has been set to $0.05 \cdot W$. Though this has been done to improve the visual display, it will nevertheless influence the ROC score when predicting human fixation [27]. We discuss this issue in detail in Section 5.

8. The input image is resized to 128×128 in the experiments.

Fig. 9. Binary images with the same histogram, but with different spatial structures.

where $\mathcal{H}(x) = -\sum_{i=1}^{n} p_i \log p_i$ is the definition of entropy of x. The reason for using entropy is as follows: The saliency map can be considered as a probability map. In a desirable saliency map, the regions of interest would be assigned higher values and the rest of the map would be largely suppressed. Thus, it is expected that the values in the saliency map histogram would cluster around certain values. The entropy of the saliency map would then be very small according to the definition of entropy.

Conventional entropy is based on the distribution of a variable x; if the histogram is given, the entropy of x is determined. However, the spatial geometric information is ignored. As shown in Fig. 9, images may possess the same histograms and therefore have the same entropy values, even though the spatial structure becomes more and more chaotic. Obviously, in saliency detection, we wish to avoid selecting a map with a high level of chaos. Spatial geometric information needs to be considered in 2D signal analysis, and work related to this issue has been reported, such as the so-called 2D Entropy [42], [43]. Here, we present a simple improved definition of entropy in order to make use of the spatial geometric information. We consider each pixel individually and require it to also depend on the values of its neighbors. We achieve this objective by employing a Gaussian kernel to filter the 2D signal, and then computing the conventional entropy on the smoothed 2D signal. Consequently, the new entropy is defined as: $\mathcal{H}_{2D}(x) = \mathcal{H}\{g_n \star x\}$, where g_n is a low-pass Gaussian kernel with a scale of ς.

As shown in Fig. 10, if ς were too small, especially when $\varsigma = 0$, Gaussian filtering would have a minor effect. If $\varsigma = 1.2$, the entropy value would increase as the image became more and more chaotic in the spatial domain. This is quite reasonable. However, if ς were too large, the entropy value would decrease. This is because the small structures in the 2D signal would be heavily destroyed by the Gaussian filter. Thus, on the one hand, we desire that ς should be as large as possible because with larger ς the influence of a pixel could spread farther. On the other hand, we do not wish to destroy the small spatial structures. Therefore, ς should be related to the size of smallest region we expect to detect. Experiments indicate that $\varsigma = 0.01 \sim 0.03 \cdot W$ yields acceptable results.

Besides entropy, there is another issue to consider when choosing the proper scale k. In HFT, given $\{\mathcal{S}_k\}$, we avoid choosing saliency maps with a strong response at the border region by using a border avoidance strategy. Thus, a parameter λ is defined for each candidate saliency map: $\lambda_k = \sum\sum \mathcal{K}(n,m) \cdot \mathcal{N}(\mathcal{S}_k(n,m))$, where $\mathcal{K}$ is a 2D centered Gaussian mask of the same size as $\mathcal{S}$, $\sigma_w = W/4$, $\sigma_h = H/4$, and $\sum\sum \mathcal{K}(n,m) = 1$. $\mathcal{N}(\cdot)$ is used to normalize $\mathcal{S}$ so that the summation of all the pixel values is 1. Note that λ is not the same as the *center-bias* (CB) or *border cut* (BC) described in [28, since it is used only to *choose* a proper scale, but not to *modify*

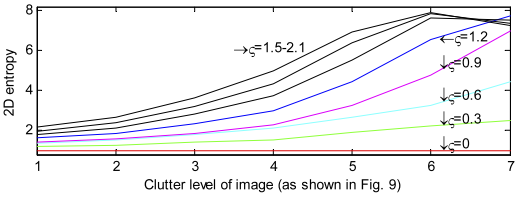

Fig. 10. Computing 2D entropy of the 2D images in Fig. 9 using Gaussian kernels of different size.

the saliency map as done in [7], [44]. *Center-bias* and *border cut* will be discussed in Section 5. Thus, with this definition of 2D entropy and λ, k_p in (30) is revised as follows:

$$k_p = \arg\min_k \left\{\lambda_k^{-1} \mathcal{H}_{2D}(\mathcal{S}_k)\right\}. \tag{31}$$

The model that uses the criterion in the above equation is HFT. In addition, we use entropy without the border-avoidance strategy as the criterion, and the results are labeled as HFT(e). Of course, it is possible that the performance of the proposed model might be improved with an even better criterion. For example, as shown in row 2 of Fig. 16F, HFT failed to highlight the two salient objects uniformly. However, we note that this was caused by the selection of an improper scale, notwithstanding the fact that what the optimal scale did was present in the existing set, as shown in the third row of Fig. 7. In order to illustrate the potential power of the proposed model, we have also determined the optimal scale for each image visually by examining the ROC scores of the saliency maps. The results are reported in this paper and are labeled as HFT*.

5 Experimental Results

Three experiments to evaluate the performance of the proposed HFT are discussed in this section: 1) response to psychological patterns, 2) predicting human fixations, and 3) predicting the object regions to which humans pay attention. Eight state-of-the-art methods were employed to perform the comparisons: Itti's model [3],[9] DVA [14],[10] GBVS [7],[11] SR [20],[12] PFT [30],[13] PQFT [30], [22],[14] AIM [5],[15] and SUN [28].[16]

We evaluate the performance of saliency detection algorithms both qualitatively and by comparison to human observers. For the former, we essentially compare the saliency map to the original image by using a simple algorithm to determine an object map based on the saliency map. For the latter, we require ground-truth data. We use two kinds of ground truth in this paper, fixation data and salient regions labeled by human observers. In Section 5.3, we have used freely available human fixation data [5] as ground truth to evaluate the algorithms listed earlier. The

9. This implementation comes in the GBVS package, see http://www.klab.caltech.edu/~harel/share/gbvs.php. The saliency toolbox STB (based on Itti's model) was used in the first experiment.
10. The code is available at http://www.its.caltech.edu/~xhou/.
11. See http://www.klab.caltech.edu/~harel/share/gbvs.php.
12. The code is available at http://www.its.caltech.edu/~xhou/.
13. Phase Fourier Transform is an improved version of SR. Our implementation was done according to [30].
14. The code was provided by the first author in [22].
15. See http://www-sop.inria.fr/members/Neil.Bruce/AIM.zip.
16. The code is available at http://cseweb.ucsd.edu/~l6zhang/.

TABLE 1
Three Subsets of Algorithms Employed for Comparison

Subset	Model	Image size	post-processing effects		
			SM	BC	CB
1	HFT	$128 \times 128^{\ddagger}$	explicit	no	no
1	SR/PFT	$64 \times 64^{\dagger}$	explicit	no	no
1	SUN	$\frac{1}{8}$full size†	implicit	no	no
2	AIM	$\frac{1}{2}$full size†	implicit	yes	no
2	DVA	$80 \times 120^{\ddagger}$	explicit	yes	no
2	PQFT	$64 \times 64^{\dagger}$	explicit	yes	no
2	Itti	full size‡	explicit	yes	no
3	GBVS	full size‡	explicit	no	yes

†*The optimal image size for this model.* ‡*The default image size.*

ROC score (area under the ROC curve, AUC) is adopted to measure their performance. In Section 5.4, we have also evaluated the algorithms using object regions labeled by humans (some examples of "labeled" results are shown in each second column of Fig. 16) as ground truth. In fact, the *available* eye tracking data only contain *positional* information [20]. However, saliency detection algorithms in computer vision are assumed and expected to have the ability to detect salient object *regions* in a scene [45]. For example, given a region such as a flower (see row 4 of Fig. 16A as an example), an algorithm should respond more or less uniformly within the whole region and not just along the boundary of the flower or at several points on the flower. Therefore, we use *salient region maps* labeled by humans as ground truth. In this experiment, besides ROC, we also use the Dice Similarity Coefficient (DSC) as a measure to evaluate the overlap between the thresholded saliency map and the ground truth. The peak value of the DSC curve (PoDSC) is an important index of performance as it corresponds to the optimal threshold and the best possible algorithm performance [46].

5.1 How to Make Fair Quantitative Comparisons?

There are two aspects which should be considered when making quantitative comparisons between two saliency models: scale and postprocessing.

Certain models permit the usage of different image scales (input image size). Therefore, in these cases we find the optimal scale by maximizing their performance, but for the other models it is necessary to use the default settings, as shown in Table 1.

With regard to postprocessing, most previous work has used the ROC directly without investigating any of the postprocessing factors affecting the fairness of this approach. However, it is important to note that three factors dramatically influence the ROC score and PoDSC: **1) border cut** [28], **2) center-bias setting** [7], and **3) smoothing** [27], [7]. In this paper, in order to make a fair comparison, the postprocessing is calibrated. We first consider BC and CB by dividing the saliency models into three *subsets*: 1) models without any BC and CB, 2) models with BC, and 3) models with CB, as shown in Table 1. In addition, the optimal smoothing parameter for each model is learned in order to eliminate the influence of SM. We compare HFT class models (HFT, HFT(e), and HFT*) with each of the three subsets:

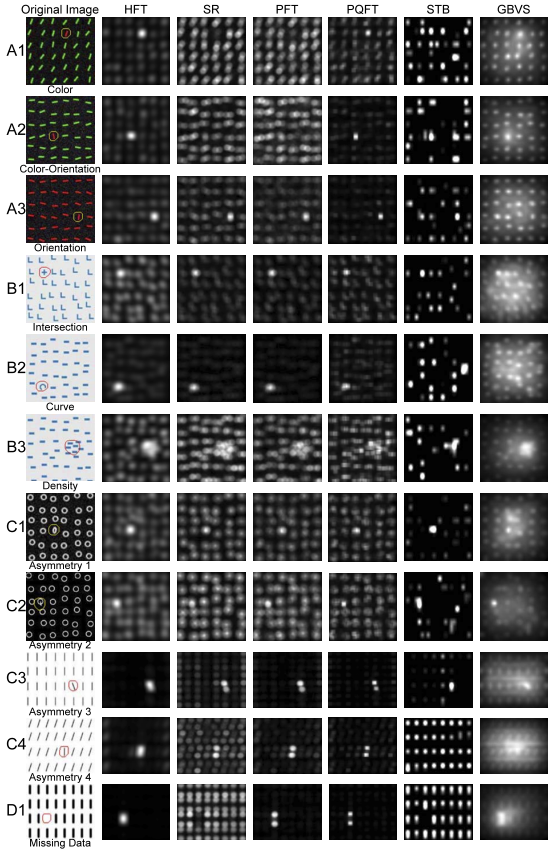

Fig. 11. Responses to the so-called psychological patterns. The first column shows the boundary of the primary object computed by HFT, superimposed on the original image. For comparison, the remaining columns present the results obtained by the methods mentioned earlier.

1. When comparing HFT class models with models in subset 1, we compute the ROC and/or PoDSC directly;
2. When comparing HFT class models with subset 2, we set the border cuts for all of these models to be of equal size.[17]
3. When comparing HFT class models with models in subset 3, we apply an *optimal* center-bias for each model individually, thereby ensuring that the ROC score for each model is maximized.

5.2 Response to Psychological Patterns

Psychological patterns are employed to evaluate three aspects of the algorithms: 1) First, we use them to evaluate basic detection ability; 2) then, we evaluate their ability to detect salient regions of different sizes; and 3) we evaluate their tolerance to random noise.

Four kinds of psychological patterns are employed: salient orientation and colored patterns (part A in Fig. 11), salient shape patterns (part B), asymmetric patterns (part C), and patterns with missing items (part D). The first column in Fig. 11 shows the original images and the second shows the saliency maps produced by HFT. The proto-objects given by HFT are superimposed on the original images in

17. In our experiments, we considered only the interior of the frame and the corresponding region in the ground truth when computing the ROC curve.

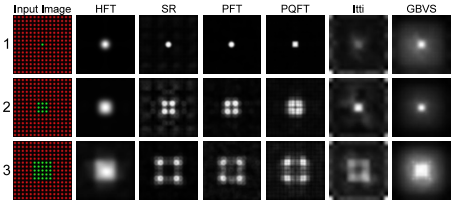

Fig. 12. Responses to psychological patterns with salient regions of different size.

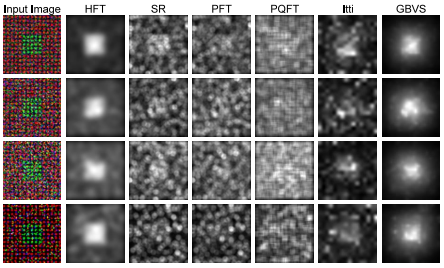

Fig. 13. Responses to psychological patterns with noise.

the first column. Our results are compared with SR, PFT, PQFT, STB, and GBVS.

As shown in part A of Fig. 11, the first pattern is salient due to the distinguishing color. Both HFT and PQFT obtain acceptable results, while SR, PFT, and STB are unable to highlight the salient bar. The second image contains a salient bar having both different color and orientation. HFT, GBVS, and PQFT succeed in highlighting the red bar, while the other methods fail. The salient bar in the third image has a distinguishing orientation, and only GBVS failed to locate it.

In part B of Fig. 11, HFT, SR, PFT, and PQFT function well. However, in B3, although both SR and PFT are able to highlight the salient region, the so-called common regions are not suppressed correctly. In B2, STB highlights the wrong area and in B3, both STB and PQFT cannot detect the salient region.

All of the algorithms are able to find the asymmetric salient regions in C1-C2. However, the common regions are not suppressed sufficiently by the SR, PFT, GBVS, and PQFT. All of the algorithms perform well for C3, although HFT achieves the best result. Finding the salient bar in C4 is apparently a more difficult task for humans and this also seems to be the case for SR, PFT, and PQFT. In general, the results are not as good as those for C3, and STB and GBVS has even failed completely.

Sometimes a salient region is simply an empty area, as shown in D1 of Fig. 11. A good salient detector should be able to locate such a region as well. We find that HFT, PFT, GBVS, and PQFT can detect the missing item successfully, although HFT does a better job.

Overall, HFT performs the best for all the cases shown in Fig. 11. We also note that SR and PFT obtain more or less the same results in cases A1-C2. However, they produce different results in cases C3-D1.[18] Since PQFT is an advanced version of PFT, its performance is an improvement over the latter, especially in the case of colored tokens. However, in the rest of the cases, PQFT achieves nearly the same performance as PFT and SR.

We would expect that an image saliency detector would highlight the different sized salient regions that people would normally pay attention to [47], [18]. In order to examine this issue, we created three patterns in which the size of the tokens increased progressively, as shown in Fig. 12. All of the algorithms responded well to the small regions (see row 1). However, as the size increased, the performance of PQFT, Itti, and SR/PFT decreased. We observe that both SR/PFT and PQFT only respond to the boundaries of the regions when the salient region is large, while both HFT and GBVS highlight the salient region uniformly.

Finally, in order to evaluate the noise tolerance of each model, we added different amounts of Gaussian (rows 1-3 of Fig. 13) and salt and pepper (row 4) noise to the patterns. As shown in Fig. 13, the proposed HFT obtained the best overall performance, while GBVS also performed quite well. GBVS is an improved version of Itti's saliency model, and its antinoise properties have also improved. We observe that SR/PFT and PQFT are quite sensitive to both Gaussian and salt and pepper noise.

5.3 Predicting Human Attention Using Fixations

We have evaluated HFT and compared it with state-of-the-art methods using human fixation data. Bruce's database[5] was employed for this purpose. It includes 120 natural images as well as corresponding eye-tracking data. The *quantitative* results are shown in Fig. 14.

We first compare HFT class models with models in subset 1. There is no border cut and center-bias in these models, so we need only find the optimal smoothing scale to compare the models. Fig. 14A shows the ROC scores for each model with different smoothing scales; we observe that they achieve their maximum ROC scores at different smoothing levels. We use the peak ROC score to establish the performance of each model and compensate for the influence of smoothing. In

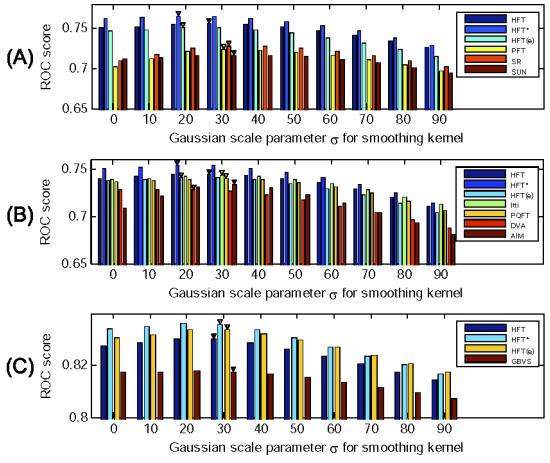

Fig. 14. Performance (peak ROC score) comparison between the HFT class models and those in the three subsets. (A) Comparing HFT with models in subset 1. (B) Comparing HFT with models in subset 2. (B) Comparing HFT with models in subset 3.

18. In Section 2, we draw the conclusion that both SR and PFT will yield nearly the same results based on the assumption of a natural image input. However, for "unnatural" images, they will sometimes produce different results, as is also discussed in [22].

Figs. 14A, 14B, and 14C, the peak performance of each algorithm is labeled by a triangle. As shown in Fig. 14A, it is obvious that HFT obtains the best performance, while SR, PFT, and SUN perform about the same.

When comparing HFT with models in subset 2, the border cut is set for both HFT class models and the models in the subset. In fact, none of the models employ the same border cut, and therefore are not immediately comparable. Without calibration, the models in subset 2 will have a mendacious ROC score. As shown in Fig. 14B, both HFT class algorithms and Itti's model are the highest performing models, while AIM has a higher peak performance than DVA.

In the more recent literature, GBVS always yields a very high ROC score and outperforms other models. However, this is most likely because GBVS incorporates a global center-bias [44]. When comparing HFT class models with GBVS, we selected the optimal center-bias for both. We note from Fig. 14C that HFT class algorithms achieve quite a high performance level, all of them outperforming GBVS.

5.4 Predicting Salient Regions that Humans Attend

Besides using fixation data, we also used object maps labeled by humans to evaluate the algorithms (Refer to http://www.cim.mcgill.ca/~lijian for details of how this database was obtained). Obviously, different images present different levels of difficulty for any saliency detector. However, the existing saliency benchmarks in the literature are collections of images, with no attempt to categorize the difficulty of analysis required. In this paper, a database containing 235 images was collected using Google as well as the recent literature. The images in this database were divided into six categories:

1. 50 images with large salient regions,
2. 80 images with intermediate salient regions,
3. 60 images with small salient regions,
4. 15 images with cluttered backgrounds,
5. 15 images with repeating distractors,
6. 15 images with both large and small salient regions.

In this section, we report both the overall performance of each model as well as the performance of each model for each category.

In this experiment, we report only the performance at the optimal smoothing level. Both the ROC score (AUC) and the peak value of the DSC curve (PoDSC) for each model were calculated as shown in Tables 3, 4, and 5. Fig. 16 shows some examples which permit a qualitative comparison for each category in the data set. However, due to space limitations, we are unable to show all of the *qualitative* results. However, these are available at http://www.cim.mcgill.ca/~lijian.

Fig. 16A shows natural images with large salient regions, a situation that is challenging for many models. It is clear that HFT achieves the best performance. The AUC and PoDSC criteria also support this conclusion. We note that GBVS achieves reasonable results, but SR, PQFT, and AIM only enhance the boundaries instead of highlighting the whole salient region uniformly.

In Fig. 16B, there are five images with intermediate salient regions. For example, in the second one, there are five salient flowers in the scene. HFT and GBVS have detected these object regions correctly. However, all of the other models failed to highlight them uniformly.

TABLE 2
Performance of the Revised One-Resolution Models

Model	AUC (improvement)	PoDSC (improvement)
SR	0.8733 (↑ 0.0210)	0.4316 (↑ 0.0387)
PFT	0.8769 (↑ 0.0243)	0.4420 (↑ 0.0456)
PQFT	0.8951 (↑ 0.0197)	0.4963 (↑ 0.0234)
SUN	0.8470 (↑ 0.0067)	0.4139 (↑ 0.0298)
AIM	0.8858 (↑ 0.0027)	0.4992 (↑ 0.0050)

The images in Fig. 16C contain distant objects and distractors (e.g., the skyline in row 1). We observe that most of the algorithms work well in detecting the small salient regions. However, sometimes Itti's method and GBVS fail to suppress distractors (see row 1). HFT only slightly outperforms the others for this category.

The backgrounds of the images in Fig. 16D are quite cluttered. This case is also difficult because many algorithms are quite sensitive to background noise. For example, consider the image in row 2. For SR, PQFT, Itti, and AIM, the nonsalient grassy surface of the ground is enhanced as much as the two insects. However, HFT and GBVS detect these two regions correctly. HFT achieves excellent performance in this category, which is also supported by the quantitative results.

Compared to salient objects, repeating distractors in the scene should not attract much attention from humans [34]. Fig. 16E shows images with salient objects among repeating objects. In both rows 2 and 3, HFT and GBVS suppress the repeating objects and enhance the salient object correctly. In row 4, there is a salient playing card among the five, and HFT, PQFT, and SR highlight the salient one and suppress the other four. HFT achieved the best performance for this category as well.

If an image contains both large and small salient regions (see Fig. 16F), a detector should be able to detect both simultaneously. For example, in row 1, there are two flowers of different size, but SR, PQFT, Itti, and AIM only respond strongly on their boundaries. However, HFT and GBVS respond correctly. Nevertheless, as discussed earlier, HFT selects just one optimal scale to determine the final output. Hence, objects of different size are not all detected or enhanced uniformly in the saliency map (see row 5 of Fig. 16F).

Overall, the experimental results shown in Tables 3, 4, and 5 indicate that the HFT model achieves the best performance for all six categories. Moreover, HFT exhibits superior performance when detecting large salient regions and saliency in cluttered scenes.

As mentioned earlier, images in this database contain salient regions of different sizes. Interestingly, in [20], it is suggested that in order to find objects at different scales it should be possible to use different resolutions of the input image. In order to investigate this issue, we created different resolutions of the input image and then fed them into SR and the other one-resolution models (see Table 2). We used the criterion described in (31) to find the optimal saliency map as the final output. We note that the

TABLE 3
Comparison between HFT Class Models and Models in Subset 1 (Optimal Smoothing Parameters for Each Algorithm)

	Category 1		Category 2		Category 3		Category 4		Category 5		Category 6		Overall	
Model	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC
HFT	**0.9424**	**0.7252**	**0.9146**	**0.5481**	**0.9351**	**0.4563**	**0.9448**	**0.5856**	**0.9193**	**0.5778**	**0.9535**	**0.6976**	**0.9281**	**0.5417**
HFT(e)	0.9101	0.6592	0.9050	0.5112	0.9348	0.4502	0.9463	0.6306	0.8907	0.5123	0.9418	0.6489	0.9159	0.5029
HFT*	0.9543	0.7438	0.9425	0.6184	0.9572	0.5217	0.9686	0.6846	0.9413	0.6148	0.9709	0.7568	0.9497	0.5963
SR	0.8148	0.5104	0.8495	0.4321	0.9091	0.3281	0.7595	0.2796	0.7929	0.3266	0.8924	0.5404	0.8523	0.3929
PFT	0.8064	0.5029	0.8426	0.4292	0.9269	0.3780	0.7294	0.2931	0.7724	0.3377	0.8967	0.5574	0.8526	0.3964
SUN	0.8218	0.5393	0.8457	0.4522	0.8838	0.3026	0.6994	0.2452	0.8067	0.3773	0.8778	0.5555	0.8403	0.4018

TABLE 4
Comparison between HFT Class Models and Models in Subset 2
(Optimal Smoothing Parameters and the Same Border Cut for Each Model)

	Category 1		Category 2		Category 3		Category 4		Category 5		Category 6		Overall	
Model	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC
HFT	**0.9338**	**0.7395**	**0.9064**	**0.5697**	**0.9328**	**0.4871**	**0.9378**	**0.6074**	**0.9137**	**0.5893**	**0.9441**	**0.7114**	**0.9217**	**0.5627**
HFT(e)	0.9010	0.6867	0.9004	0.5402	0.9312	0.4621	0.9471	0.6568	0.8660	0.5286	0.9340	0.6743	0.9102	0.5289
HFT*	0.9478	0.7584	0.9387	0.6434	0.9578	0.5503	0.9660	0.7012	0.9374	0.6330	0.9644	0.7658	0.9470	0.6226
AIM	0.8511	0.6011	0.8761	0.5226	0.9359	0.4506	0.8370	0.3969	0.8668	0.4987	0.9124	0.6489	0.8831	0.4942
PQFT	0.8571	0.6201	0.8771	0.5350	0.9096	0.3901	0.8205	0.3819	0.8421	0.4304	0.9105	0.6398	0.8754	0.4729
DVA	0.8075	0.5736	0.8565	0.5095	0.9038	0.3957	0.7618	0.3639	0.8250	0.4553	0.9048	0.6262	0.8510	0.4642
Itti	0.8768	0.6533	0.8886	0.5317	0.9239	0.3843	0.8107	0.3687	0.8983	0.5194	0.9191	0.6530	0.8910	0.4949

TABLE 5
Comparison between HFT Class Models and Model in Subset 3 (Optimal Smoothing Parameters and Center-Bias for Each Model)

	Category 1		Category 2		Category 3		Category 4		Category 5		Category 6		Overall	
Model	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC	AUC	PoDSC
HFT	**0.9565**	**0.7548**	**0.9296**	**0.5688**	**0.9504**	**0.4755**	**0.9381**	**0.5799**	**0.9523**	**0.6318**	**0.9578**	**0.7020**	**0.9414**	**0.5665**
HFT(e)	0.9409	0.7213	0.9287	0.5697	0.9614	0.5024	0.9460	0.6315	0.9361	0.6038	0.9579	0.7089	0.9403	0.5609
HFT*	0.9665	0.7771	0.9533	0.6396	0.9724	0.5640	0.9709	0.7028	0.9644	0.6841	0.9743	0.7586	0.9609	0.6250
GBVS	0.9363	0.6990	0.9135	0.5304	0.9173	0.3678	0.9223	0.5644	0.9453	0.6145	0.9249	0.6329	0.9211	0.5154

performance of these revised models has improved (as shown in Table 2), although it is still lower than the HFT class models (See Tables 3, 4, and 5.)

Although HFT has performed well in the experiments described in Sections 5.2 to 5.4, it does fail in certain cases. HFT could not satisfactorily predict the correct human fixations for several of the "hard" images collected in [44]. Although HFT did predict the human fixations correctly in Figs. 15a and 15b, some incorrect responses did occur. For example, in (c) it incorrectly highlighted some parts of the clothes and failed to highlight the eyes, while in (g), some parts of the boundary of the face were wrongly highlighted. In both (e) and (f), people tended to pay attention to the text, but HFT locates regions with salient low-level features (e.g., the red flag and the clock). In (d), HFT totally failed to locate the salient heads. Clearly prior knowledge and task information is not employed for bottom-up models. Therefore, these approaches focus on regions possessing distinct low-level features (color, intensity, etc.) and sometimes may fail to highlight the regions that are known to interest people (e.g., humans, animals, and other common objects). One way to solve this problem is to employ more complex features or invoke top-down cues.

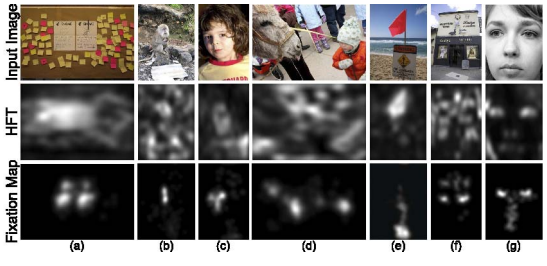

Fig. 15. Hard image cases of HFT in predicting human fixations.

Most of the bottom-up saliency models, such as Itti, Gao's model, AIM, and so on, use local contrast or a center-surround paradigm. Similarly, models like SR can also be considered as pixel-level local contrast models (gradient operation). These work well for detecting small salient regions, but do not perform well in predicting large salient regions. There are two ways to alleviate this problem; one is to adopt a multiscale strategy (as used in Itti's model), the other is to decrease the resolution of the input image and employ a large amount of blurring of the saliency maps (as used in SR, PFT). Finally, perhaps it is unfair to characterize SR class models as being only pixel-level local contrast detectors. As discussed earlier, SR and PFT are special cases of the proposed HFT model when the scale goes to infinity in the frequency domain. Hence, they have the ability to globally inhibit and suppress repeated patterns. However, other models based on local contrast will perform poorly in this case. Nevertheless, if there are no repeated patterns in

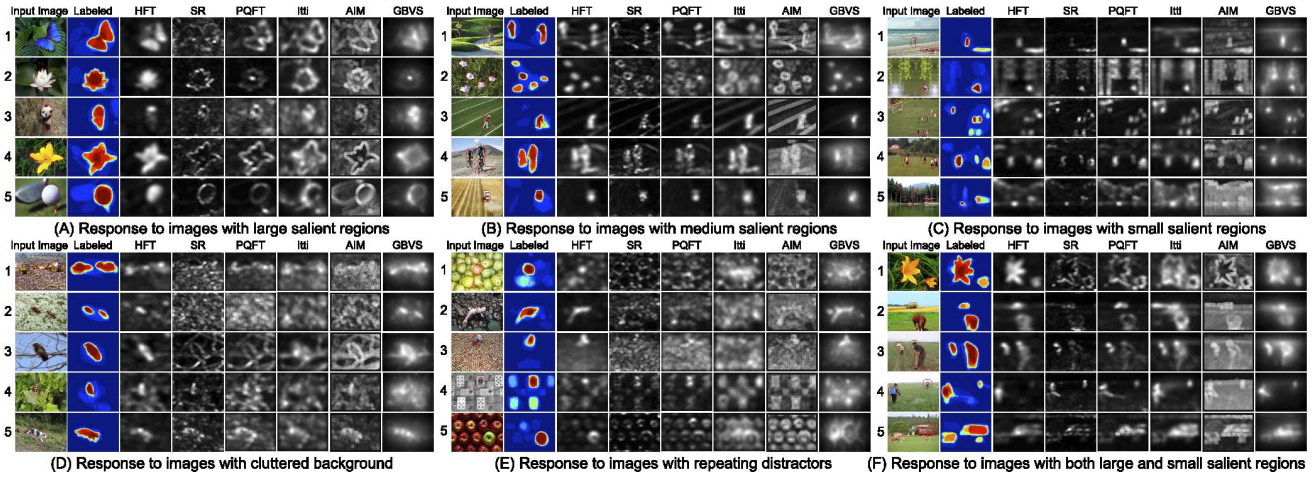

Fig. 16. Procedure for computing saliency using the Hypercomplex Fourier Transform.

the scene, the SR model will function as a gradient detector, and only enhance boundaries of objects.

6 Conclusions

This paper proposes a new saliency detection framework for images based on analyzing the spectrum scale-space. We show that the convolution of the image amplitude spectrum with a low-pass Gaussian kernel of an appropriate scale is equivalent to such an image saliency detector. The proposed approach is able to highlight both small and large salient regions and inhibit repeated patterns. We also illustrate that both SR and PFT are special cases of the proposed model when the scale parameter goes to infinity. In order to fuse multidimensional feature maps, we employ the Hypercomplex Fourier Transform to replace the standard Fourier Transform for spectrum scale-space analysis.

To validate the proposed approach, we have performed saliency computations on both commonly used synthetic data as well as natural images, and then compared the results with state-of-the-art algorithms. In order to make a fair comparison when using the ROC and PoDSC as measures of performance, we have proposed an improved comparison procedure by considering the border cut, center-bias, and smoothing effects. Experimental results indicate that the proposed model can predict human fixation data as well as the object regions labeled by humans. We also show that sometimes HFT may fail to predict human fixations. This is most likely because only low-level features are employed; clearly, top-down or task-orientated cues are necessary for improving the performance of current saliency models in predicting human attention.

With regard to future work, first, it would be interesting to investigate other criteria for optimal scale selection. Entropy was employed in this paper to select the optimal scale automatically. However, we have observed that much better performance can be obtained by manually selecting the optimal scale. Second, in the proposed model, only one saliency map, corresponding to the optimal scale, is selected as the final one. However, we have noted that certain of the abandoned maps also contain meaningful saliency information. How to incorporate these in the determination of the saliency is left to future investigation. Third, we intend to include top-down information to improve performance. The ultimate goal of our research is to develop a system for on-board pedestrian and vehicle detection, for which a considerable amount of top-down temporal data exists.

Acknowledgments

The authors thank all of the reviewers for their insights and suggestions, which were very helpful in improving the manuscript. They also thank J. Harel and X. Hou for useful discussions. This work is partially supported by the National Natural Science Foundation of China under Grants 61075072, 90820302, and the New-Century Excellent Talent Plan of Chinese Education Ministry (No. NCET-10-0901).

References

[1] A. Yarbus, *Eye Movements and Vision.* Plenum Press, 1967.
[2] U. Neisser, *Cognitive Psychology.* Appleton-Century-Crofts, 1967.
[3] L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1254-1259, Nov. 1998.
[4] J. Tsotsos, "What Roles Can Attention Play in Recognition?" *Proc. Seventh IEEE Int'l Conf. Development and Learning,* pp. 55-60, 2008.
[5] N. Bruce and J. Tsotsos, "Saliency Based on Information Maximization," *Proc. Advances in Neural Information Processing Systems,* 2006.
[6] S. Chikkerur, T. Serre, C. Tan, and T. Poggio, "What and Where: A Bayesian Inference Theory of Attention," *Vision Research,* vol. 50, pp. 2233-2247, 2010.
[7] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," *Proc. Advances in Neural Information Processing Systems,* 2007.
[8] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A Coherent Computational Approach to Model Bottom-Up Visual Attention," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 5, pp. 802-817, May 2006.
[9] W. Kienzle, F. Wichmann, B. Schölkopf, and M. Franz, "A Nonparametric Approach to Bottom-Up Visual Saliency," *Proc. Advances in Neural Information Processing Systems,* vol. 19, pp. 689-696, 2007.
[10] V. Mahadevan and N. Vasconcelos, "Spatiotemporal Saliency in Dynamic Scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 1, pp. 171-177, Jan. 2010.
[11] D. Gao, V. Mahadevan, and N. Vasconcelos, "The Discriminant Center-Surround Hypothesis for Bottom-Up Saliency," *Proc. Advances in Neural Information Processing Systems,* 2008.
[12] M. Cerf, J. Harel, W. Einhaeuser, and C. Koch, "Predicting Human Gaze Using Low-Level Saliency Combined with Face Detection," *Proc. Advances in Neural Information Processing Systems,* 2008.
[13] J. Tsotsos and A. Rothenstein, "Computational Models of Visual Attention," *Scholarpedia,* vol. 6, no. 1, p. 6201, 2011.

[14] X. Hou and L. Zhang, "Dynamic Visual Attention: Searching for Coding Length Increments," *Proc. Advances in Neural Information Processing Systems,* 2009.
[15] L. Itti and P. Baldi, "Bayesian Surprise Attracts Human Attention," *Vision Research,* vol. 49, pp. 1295-1306, 2009.
[16] L. Itti and C. Koch, "A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention," *Vision Research,* vol. 40, nos. 10-12, pp. 1489-1506, 2000.
[17] T. Kadir and M. Brady, "Saliency, Scale and Image Description," *Int'l J. Computer Vision,* vol. 45, no. 2, pp. 83-105, 2001.
[18] R. Achanta, S. Hemami, F. Estrada, and S. Ssstrunk, "Frequency-Tuned Salient Region Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2009.
[19] T. Avraham and M. Lindenbaum, "Esaliency (Extended Saliency): Meaningful Attention Using Stochastic Image Modeling," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 4, pp. 693-708, Apr. 2010.
[20] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2007.
[21] S. Yantis, "How Visual Salience Wins the Battle for Awareness," *Nature Neuroscience,* vol. 8, no. 8, pp. 975-977, 2005.
[22] C. Guo and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression," *IEEE Trans. Image Processing,* vol. 19, no. 1, pp. 185-198, Jan. 2010.
[23] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum, "Learning to Detect a Salient Object," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 33, no. 2, pp. 353-367, Feb. 2011.
[24] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-Aware Saliency Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2010.
[25] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global Contrast Based Salient Region Detection," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2011.
[26] S. Khan, J. van de Weijer, and M. Vanrell, "Top-Down Color Attention for Object Recognition," *Proc. IEEE Int'l Conf. Computer Vision,* 2009.
[27] X. Hou, J. Harel, and C. Koch, "Image Signature: Highlighting Sparse Salient Regions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 34, no. 1, pp. 194-201, Jan. 2012.
[28] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "SUN: A Bayesian Framework for Saliency Using Natural Statistics," *J. Vision,* vol. 8, no. 7, article 32, 2008.
[29] D. Gao, S. Han, and N. Vasconcelos, "Discriminant Saliency, the Detection of Suspicious Coincidences, and Applications to Visual Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 31, no. 6, pp. 989-1005, June 2009.
[30] C. Guo, Q. Ma, and L. Zhang, "Spatio-Temporal Saliency Detection Using Phase Spectrum of Quaternion Fourier Transform," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2008.
[31] L. Itti and C. Koch, "Computational Modelling of Visual Attention," *Nature Rev. Neuroscience,* vol. 2, no. 3, pp. 194-203, Mar. 2001.
[32] D. Gao and N. Vasconcelos, "Bottom-Up Saliency Is a Discriminant Process," *Proc. IEEE Int'l Conf. Computer Vision,* 2007.
[33] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the Plausibility of the Discriminant Center-Surround Hypothesis for Visual Saliency," *J. Vision,* vol. 8, no. 7, pp. 1-18, 2008.
[34] D. Beck and S. Kastner, "Stimulus Context Modulates Competition in Human Extrastriate Cortex," *Nature Neuroscience,* vol. 8, no. 8, pp. 1110-1116, 2005.
[35] Z. Yu and H. Wong, "A Rule Based Technique for Extraction of Visual Attention Regions Based on Real-Time Clustering," *IEEE Trans. Multimedia,* vol. 9, no. 4, pp. 766-784, June 2007.
[36] N. Jacobson, T.Q. Nguyen, and A. Tal, "Video Processing with Scale-Aware Saliency: Application to Frame Rate Up-Conversion," *Proc. IEEE Conf. Acoustics, Speech and Signal Processing,* 2011.
[37] D. Ruderman, "The Statistics of Natural Images," *Network: Computation in Neural Systems,* vol. 5, no. 4, pp. 517-548, 1994.
[38] A. Srivastava, A. Lee, E. Simoncelli, and S. Zhu, "On Advances in Statistical Modeling of Natural Images," *J. Math. Imaging and Vision,* vol. 18, no. 1, pp. 17-33, 2003.
[39] J. Duncan and G. Humphreys, "Visual Search and Stimulus Similarity," *Psychological Rev.,* vol. 96, no. 3, pp. 433-458, 1989.
[40] T. Ell, "Quaternion-Fourier Transforms for Analysis of Two-Dimensional Linear Time-Invariant Partial Differential Systems," *Proc. IEEE Conf. Decision and Control,* 2002.
[41] T. Ell and S. Sangwine, "Hypercomplex Fourier Transforms of Color Images," *IEEE Trans. Image Processing,* vol. 16, no. 1, pp. 22-35, Jan. 2007.
[42] A. Abutaleb, "Automatic Thresholding of Gray-Level Pictures Using Two-Dimensional Entropy," *Computer Vision, Graphics, and Image Processing,* vol. 47, no. 1, pp. 22-32, 1989.
[43] W. Chen, C. Wen, and C. Yang, "A Fast Two-Dimensional Entropic Thresholding Algorithm," *Pattern Recognition,* vol. 27, no. 7, pp. 885-893, 1994.
[44] T. Judd, F. Durand, and A. Torralba, "A Benchmark of Computational Models of Saliency to Predict Human Fixations," *IEEE Trans. Pattern Analysis and Machine Intelligence,* in review.
[45] L. Elazary and L. Itti, "Interesting Objects Are Visually Salient," *J. Vision,* vol. 8, no. 3, pp. 1-15, 2008.
[46] T. Veit, J. Tarel, P. Nicolle, and P. Charbonnier, "Evaluation of Road Marking Feature Extraction," *Proc. IEEE Conf. Intelligent Transportation Systems,* 2008.
[47] W. Einhauser, M. Spain, and P. Perona, "Objects Predict Fixations Better than Early Saliency," *J. Vision,* vol. 8, no. 14, pp. 1-26, 2008.

Jian Li received the BE and ME degrees in control science and engineering from the National University of Defense Technology (NUDT), Changsha, Hunan, P.R. China, where he is currently working toward the PhD degree. He is also a visiting PhD student at the Center for Intelligent Machines (CIM) at McGill University, Canada. His research interests include computer vision, pattern recognition, image processing, and machine learning. He is a student member of the IEEE.

Martin D. Levine received the BEng and MEng degrees in electrical and computer engineering from McGill University, Montreal, Canada, in 1960 and 1963, respectively, and the PhD degree in electrical engineering from the Imperial College of Science and Technology, University of London, United Kingsom, in 1965. He is currently a professor in the Department of Electrical and Computer Engineering, McGill University, and served as the founding director of the McGill Center for Intelligent Machines (CIM) from 1986 to 1998. During 1972-1973, he was a member of the technical staff at the Image Processing Laboratory of the Jet Propulsion Laboratory, Pasadena, California. During the 1979-1980 academic year, he was a visiting professor in the Department of Computer Science, Hebrew University, Jerusalem, Israel. His research interests include computer vision, image processing, and artificial intelligence, and he has numerous publications to his credit on these topics. As well, he has consulted for various government agencies and industrial organizations in these areas. He was a founding partner of AutoVu Technologies, Inc., and VisionSphere Technologies, Inc. He authored the book *Vision in Man and Machine* and coauthored *Computer Assisted Analyses of Cell Locomotion and Chemotaxis*. He is an area editor for face detection and recognition and on the editorial board of the journal *Computer Vision and Image Understanding*, and has also served on the editorial boards of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and *Pattern Recognition*. He was the editor of the Plenum Book Series on advances in computer vision and machine intelligence. He was the general chairman of the Seventh International Conference on Pattern Recognition held in Montreal during the summer of 1984 and served as the president of the International Association of Pattern Recognition during 1988-1990. He was also the founding president of the Canadian Image Processing and Pattern Recognition Society. He was elected as a fellow of the Canadian Institute for Advanced Research in 1984. During the period 1990-1996, he served as a CIAR/PRECARN associate. He was presented with the 1997 Canadian Image Processing and Pattern Recognition Society Service Award for his outstanding contributions to research and education in computer vision. He is a fellow of the IEEE, the Canadian Academy of Engineering, and the International Association for Pattern Recognition.

Xiangjing An received the BS degree in automatic control from the Department of Automatic Control, National University of Defense Technology (NUDT), Changsha, P.R. China, in 1995 and the PhD degree in control science and engineering from the College of Mechatronics and Automation (CMA), NUDT in 2001. He was a visiting scholar for cooperation research at Boston University during 2009-2010. Currently, he is an associate professor at the Institute of Automation, CMA, NUDT. His research interests include image processing, computer vision, machine learning, and biologically inspired feature extraction. He is a member of the IEEE.

Xin Xu received the BS degree in electrical engineering from the Department of Automatic Control, National University of Defense Technology (NUDT), Changsha, P.R. China, in 1996 and the PhD degree in control science and engineering from the College of Mechatronics and Automation (CMA), NUDT. He has been a visiting scientist for cooperation research at the Hong Kong Polytechnic University, University of Alberta, University of Guelph, and University of Strathclyde, respectively. Currently, he is a full professor at the Institute of Automation, CMA, NUDT. He has coauthored four books and published more than 70 papers in international journals and conferences. His research interests include reinforcement learning, learning control, robotics, data mining, autonomic computing, and computer security. He is one of the recipients who received the first-class Natural Science Award from Hunan Province, P.R. China, in 2009 and the Fork Ying Tong Youth Teacher Fund of China in 2008. He is a committee member of the IEEE Technical Committee on Approximate Dynamic Programming and Reinforcement Learning (ADPRL) and the IEEE Technical Committee on Robot Learning. He has served as a program committee member or session chair for many international conferences. He is a senior member of the IEEE.

Hangen He received the BSc degree in nuclear physics from Harbin Engineering Institute, China, in 1968. He was a visiting professor at the University of the German Federal Armed Forces in 1996 and 1999, respectively. He is currently a professor in the College of Mechatronics and Automation (CMA), National University of Defense Technology (NUDT), Changsha, Hunan, China. His research interests include artificial intelligence, reinforcement learning, learning control, and robotics. He has served as a member of editorial boards of several journals and has cochaired many professional conferences. He is a joint recipient of more than a dozen academic awards in China.