Saliency Detection Based on Frequency and Spatial Domain Analysis

Jian Li[12]
http://www.cim.mcgill.ca/~lijian
Martin D. Levine[2]
levine@cim.mcgill.ca
Xiangjing An[1]
anxiangjing@gmail.com
Hangen He[1]
hehangen@gmail.com

[1] Institute of Automation
National University of Defense Technology
Changsha, PR China

[2] Centre for Intelligent Machines
McGill University
Canada

Abstract

We propose a new saliency detection model by combining global information from frequency domain analysis and local information from spatial domain analysis. In the frequency domain analysis, instead of modeling salient regions, we model the nonsalient regions using global information; these so-called repeating patterns that are not distinctive in the scene are suppressed by using spectrum smoothing. In spatial domain analysis, we enhance those regions that are more informative by using a center-surround mechanism similar to that found in the visual cortex. Finally, the outputs from these two channels are combined to produce the saliency map. We demonstrate that the proposed model has the ability to highlight both small and large salient regions in cluttered scenes and to inhibit repeating objects. Experimental results also show that the proposed model outperforms existing algorithms in predicting objects regions where human pay more attention.

1 Introduction

As a component of low-level vision processing, saliency detection 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 has received extensive attention by both psychologists and computer vision researchers [3, 8, 10, 12]. Bottom-up saliency for selecting attentional regions is the focus of this paper. Many such computational models have appeared in the literature. There are several other models proposed which utilize the local information. Itti and Koch's saliency model [11, 12] is the milestone in saliency detection and is usually used for comparison. Gao *et al.* [5] proposed a bottom-up saliency model by using Kullback-CLeibler (KL) divergence to measure the difference between a location and its surrounding area. Recently, several models have been proposed to compute saliency by using global information. In [1], the authors first transform the input color image into the Lab space (an opponent color space),

BMVC 2011 http://dx.doi.org/10.5244/C.25.86

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.* [8] proposed a graph-based solution which uses local computation to obtain a saliency map which is everywhere dependent on global information. Hou and Zhang [9] proposed a Fourier Transform based saliency model, called the spectrum residual (SR). Successively, the phase spectrum of Fourier Transform (PFT) was presented, which achieved nearly the same performance as SR [9].

In this paper, we argue that a reasonable saliency detector should have the ability to: **(1) Detect both small and large saliency regions.** The size of salient regions vary greatly. As shown in row 2 of Fig.7, the yellow flower definitely attracts the most attention, and the fixation points should be distributed more or less uniformly throughout the whole salient region. However, because center-surround algorithms mainly use local information, they will respond heavily in boundary regions, where the texture, intensity or other features are locally different. **(2) Detect saliency in cluttered scenes.** Another drawback of local information based saliency models is that heavily textured regions are always highlighted. Cluttered scene are still a challenge for models based on local information and some based on global information[1, 7, 9]. **(3) Inhibit repeating patterns.** Objects in scenes viewed by the human visual system are thought to compete with each other to selectively focus attention on a subset [13]. These repeating patterns will suppress each other and then be inhibited.

In this paper, inspired by [9, 10, 14], we propose a new saliency model based on both frequency and spatial domain analysis, which utilizes both local and global information of the image. We show experimentally that the proposed model has the ability to highlight both small and large salient regions and to inhibit repeating patterns in cluttered scenes.

2 Related work

Recently, the simple and fast algorithm, *Spectrum Residual* (SR), was proposed [9]. This paper argued that the spectrum residual corresponded to image saliency. Thus given an image $f(x,y)$, it was first transformed into the frequency domain: $f(x,y) \xrightarrow{\mathscr{F}} \mathscr{F}(f)(u,v)$ and the amplitude, $\mathscr{A}(u,v) = |\mathscr{F}(f)|$, and phase, $\mathscr{P}(u,v) = angle(\mathscr{F}(f))$,spectra calculated, where $\mathscr{F}(f)$ is the Fourier Transformation (FT) of $f(x,y)$. The log amplitude spectrum is given by $\mathscr{L}(u,v) = log(\mathscr{A}(u,v))$. Given these definitions, the *spectrum residual* was defined as:

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

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

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

In order to obtain a better visual display, the final saliency map was actually given as:

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

where $\mathscr{F}$ and $\mathscr{F}^{-1}$ denote the Fourier and inverse Fourier Transforms, respectively; h_n and g are low-pass filters; i is the imaginary unit; $\mathscr{P}(u,v)$ denotes the phase spectrum of the image, which is assumed to be preserved during this process. (1-3) are from [9]. The authors argued that it is this residual, combined with the original phase spectrum, that corresponded to image saliency. However, in this paper, we will show that: 1) the spectrum residual is of little significance; 2) for natural images, SR and PFT [7]) are, to some extent, equivalent to a gradient operator; and 3) SR works in certain cases only.

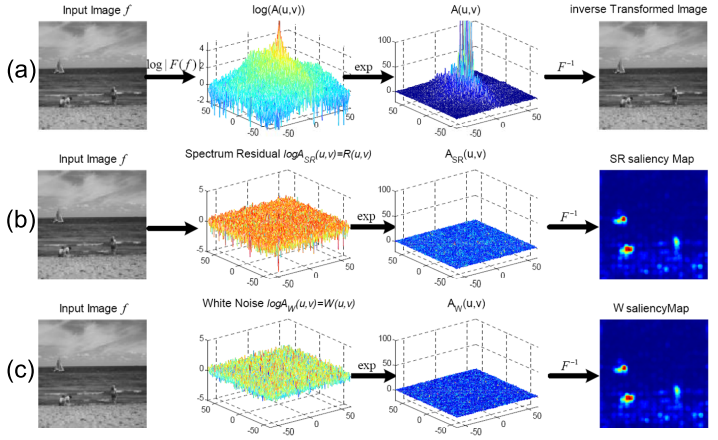

Figure 1: Analysis of Spectrum Residual. (a) Obviously, the original image is reproduced by performing the inverse FT using the original amplitude and phase spectra. (b) In SR, it is argued that saliency map can be obtained by replacing the $log(\mathscr{A}(u,v))$ by the Spectrum Residual $\mathscr{R}(u,v)$. (c) If we replace the log amplitude spectrum $log\mathscr{A}(u,v)$ by random white noise, we can obtain nearly the same saliency map.

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

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

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

Thus we can rewrite (2) as:

$$\mathscr{S}(x,y) = \mathscr{F}^{-1}[exp(\mathscr{R}(u,v).exp(i.\mathscr{P}(u,v))], \quad (6)$$

Define $exp(\mathscr{R}(u,v))$ as $\mathscr{A}_{SR}(u,v)$, so that (6) can be rewritten as:

$$\mathscr{S}(x,y) = \mathscr{F}^{-1}[\mathscr{A}_{SR}(u,v).exp(i.\mathscr{P}(u,v))]. \quad (7)$$

Then comparing (5) and (7), we observe that if we replace the amplitude spectrum $\mathscr{A}(u,v)$ by the exponential of the $\mathscr{R}(u,v)$, the saliency map is obtained[1](See Fig. 1(a,b)). In order to illustrate that the spectrum residual is of little significance, we generate a 2D white noise $\mathscr{W}(u,v)$, which has the same average value and maximum as the spectrum residual $\mathscr{R}(u,v)$. We then use $\mathscr{W}(u,v)$ to replace the spectrum residual and perform the inverse FT as follows:

$$\mathscr{S}(x,y) = \mathscr{F}^{-1}[exp(\mathscr{W}(u,v).exp(i.\mathscr{P}(u,v))], \quad (8)$$

Fig.1(c) illustrates this process. Defining $exp(\mathscr{W}(u,v))$ as $\mathscr{A}_W(u,v)$, (8) can be rewritten as:

$$\mathscr{S}(x,y) = \mathscr{F}^{-1}[\mathscr{A}_W(u,v).exp(i.\mathscr{P}(u,v))]. \quad (9)$$

Surprisingly, we can obtain nearly the same saliency map when we replace the spectrum residual by white noise. This result very clearly shows that the spectrum residual is of little

[1]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.

significance. Why is this the case? Examining both (7) and (9), we find that the amplitude spectra used to perform the inverse Fourier Transform are $A_{SR}(u,v)$ and $\mathscr{A}_W(u,v)$. As shown in the third columns of Fig.1(b,c), both $\mathscr{A}_{SR}(u,v)$ and $\mathscr{A}_W(u,v)$ are horizontal planes compared with $\mathscr{A}(u,v)$ shown in Fig.1(a). That is to say, in both (7) and (9), the amplitude information has been totally abandoned and only phase information has been used.

Two questions arise: (1) Why does SR yield a saliency map using only phase information? (2) 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 detecting small salient regions on an uncluttered background. Consider [4, 7] where the authors propose a new saliency model called Phase Fourier Transform (PFT), where The saliency is computed using only phase information. What does using the inverse Fourier Transform solely with phase information imply? In fact, it implies a gradient operation. We argue that for natural images, both SR and PFT are, to some extent, equivalent to a gradient operator combined with Gaussian post-processing (like the g in (3)). This is because the amplitude spectrum of natural images always have higher values at low rather than at high frequencies. Thus, if the amplitude spectrum is replaced by a horizontal plane, we are treating all of the frequencies equally. That is to say, the lower frequencies are suppressed and the higher ones are enhanced. Obviously, 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. So why is the performance of these models not good enough? Because the information contained in the amplitude spectrum has been totally abandoned!

Next, we will illustrate in section.3 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 amplitude and phase information.

3 The methodology

The human visual systems pays different attention to different regions in a scene. For example, a region that contain a unique and well defined target may be allocated more attention, while numerous and similar regions could be given less attention. A saliency map produced by a detection algorithm should assign a saliency probability for each location in a similar manner. The authors in [14] use a Bayesian framework to formalize the visual saliency in this way. Thus the saliency value at a location z in an image is defined as the probability that this point belongs to a salient region ($C = 1$), given the location L, the feature F at this point, local information I_l and global information I_g. The saliency is defined as $\log s_z$, and we have:

$$\log s_z = \underbrace{-\log p(F,I_l,I_g)}_{\text{Bottom up saliency}} + \underbrace{\log p(F,I_l,I_g|C=1)}_{\text{Top down knowledge}} + \underbrace{\log p(C=1|L)}_{\text{Center bias}}. \tag{10}$$

There are three items on the right side of (10). The last item is the location prior, which is also called the center bias in saliency detection. This concept was first discussed theoretically in [12]. In this paper, we will take it into account in the evaluation of experiments in section 4. The second item constitutes the top-down knowledge. However, in this paper, we will focus on task-independent saliency (free viewing), so this will be omitted. The first item corresponds to the bottom-up saliency. From this item, we know that, given the features at a position, bottom-up saliency is determined by two factors: local and global information. In this section, we will discuss a model of bottom-up saliency that combines global information from the frequency domain analysis and local information from spatial domain analysis.

Figure 2: Repeating and anomalous patterns. Left: A natural image; right: Collection of fragments from the image.

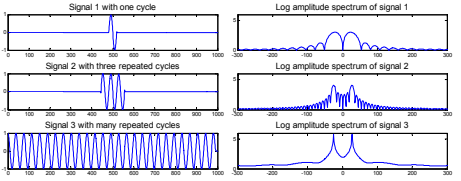

Figure 3: Repeating patterns lead to sharp spikes. Left: Signals with repeating cycles; right: Corresponding amplitude.

3.1 Frequency domain analysis

Frequency analysis presents an opportunity to deal with the global information in an image. In the proposed model, we investigate the relationship between the amplitude spectrum and non-salient regions in the image. In the existing models in the literature, salient regions are usually assumed to be distinctive or irregular patterns, which possess a distinct feature distribution compared with the rest of the image. In this paper, instead of searching for these so-called distinctive patterns, we model regular patterns that would not attract much attention by our visual system. We refer to these as being non-salient. The analysis is based solely on the Fourier Transform.

Suppressing repeating patterns for saliency pop-out. In our model, it is assumed that a natural image consists of several salient and many common (non-salient) regions. All of these (whether distinct or not) may be considered as visual stimuli that compete for attention in the visual cortex. As shown in left part of Fig.2, if we divide the image into patches (at any scale), we find that some are distinctive, while many are quite similar to each other (blue sky and ground). The right part of Fig.2 shows the complete collection of patches. We observe that several patterns appear many times in the image and refer to them as repeating patterns, which is consistent with the human visual system, which treats these as being non-salient. In the next section, we model the repeating patterns and then suppress them to achieve pop-out, thereby yielding the salient regions.

Spikes in the amplitude spectrum correspond to repeating patterns. It is argued in [9] that the spectrum residual corresponds to the saliency in an image, while contradictorily in [7], the amplitude information was totally abandoned. However, in this paper, we will illustrate that the amplitude spectrum also contains important information corresponding to image saliency. To be more exact, spikes in the amplitude spectrum correspond to repeating patterns, which should be suppressed for saliency detection.

For natural images, repeating patterns always lead to spikes in the amplitude spectrum. Taking a 1-D periodic signal $f(t)$ as an example, suppose that it 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 its Fourier transform is given by: $F(w) = 2\pi \sum_{n=-\infty}^{\infty} F(n)\delta(\omega - n\omega_1)$. From the latter, we can conclude that the spectrum of a periodic signal (repeating cycles) is a set of impulse functions (spikes). We note that this is based on the assumption that the periodic signal is infinite in extent. Therefore, given a more realistic finite length periodic signal, the shape of the spectrum will obviously be different but not degraded greatly. From Fig.3(b) we note that, the larger the number of repeating cycles, the sharper the spectrum. Thus one can conclude that repeating patterns lead to sharp spikes in the amplitude spectrum. Besides the sinusoid, other repeating signals also have this characteristic.

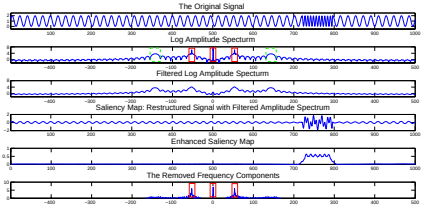

Figure 4: Suppression of repeated patterns using spectrum filtering.

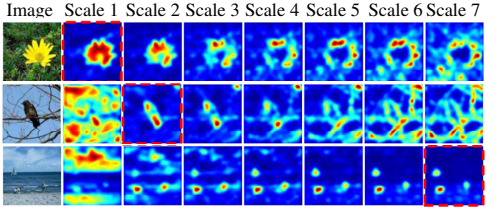

Figure 5: Saliency maps computed by smoothing the image amplitude spectrum at different scales. The best result is indicated by a dashed box.

As a simple illustrative example, suppose there is one salient part that is embedded in a finite length periodic signal (see the original signals in Fig.4). We will illustrate that this salient part 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 signal; 2) The embedded salient part will not lead to very sharp spikes in the amplitude spectrum. The signal to be analyzed is defined as follows: $f(t) = g(t) + g_\sigma(t) + s(t)$, where $g(t) = p(t)$if $t \in (0,L)$, else $g(t) = 0$ $g_\sigma(t) = -p(t) \cdot W(t)$, $s(t) = -p_s(t) \cdot W(t)$; $s(t)$ is the salient part of $f(t)$, 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 f and f_s, respectively; $W(t)$ is a rectangular window function that equals 1 inside the interval $(t_0, t_0+\sigma)$ and 0 elsewhere; we also assume that $(t_0, t_0+\sigma) \in (0,L)$ and $\sigma \ll L$. Thus the Fourier Transform of $f(t)$ can be represented as follows:

$$\mathscr{F}(f)(\omega) = \int_0^L g(t)e^{-j\omega t}dt + \int_{t_0}^{t_0+\sigma} g_\sigma(t)e^{-j\omega t}dt + \int_{t_0}^{t_0+\sigma} s(t)e^{-j\omega t}dt. \quad (11)$$

From (11), the spectrum of $f(t)$ consists of three terms. Since $\sigma \ll L$, the first term will exhibit very sharp spikes in the amplitude spectrum, while this is not true of the second and third terms. In order to illustrate this, we define a notion of "sharpness" of an amplitude spectrum X. Suppose that we smooth an amplitude spectrum, containing several spikes, using a low-pass filter. Then we will find that the sharper the original spike, the more its peak height will be reduced. Therefore, we describe the "sharpness" of X as follows: $S(X) = \| X - X * h \|_\infty$, where h is a Gaussian kernel at scale σ. Taking $g_\sigma(t)$ as an example, we compute the point-wise product of a periodic signal $-p(t)$ and a rectangular window function $W(t)$. According to the convolution theorem, $\mathscr{F}(g_\sigma)(\omega)$ equals the convolution of $-\mathscr{F}(p)(\omega)$ with $\mathscr{F}(W)(\omega)$. Since $\mathscr{F}(W)(\omega) = \frac{2sin(\sigma/2)}{\omega} e^{j\omega(t_0+\sigma/2)}$ is a low-pass filter, the spikes in the amplitude of $-\mathscr{F}(p)(\omega)$ will be greatly suppressed. This also occurs for the third term.

As discussed above, the "sharpness" of $\mathscr{F}(f)(\omega)$ is mainly determined by $g(t)$, while the latter two terms in (11) do not contribute much to the spikes in the spectrum. In other words, since the first term corresponds to repeated patterns that lead to spikes, they can be suppressed by smoothing the spikes in the amplitude spectrum of $\mathscr{F}(f)(\omega)$.

Suppressing repeated patterns using spectrum smoothing. A Gaussian kernel h can be employed to suppress spikes in the amplitude spectrum[2] as follows:

$$\mathscr{A}_\mathscr{S}(u,v) = |\mathscr{F}\{f(x,y)\}| \star h, \quad (12)$$

[2]In the computer implementation of this, we found that suppressing spikes in the log amplitude spectrum rather than the amplitude spectrum yielded better results.

where h is a Gassian kernel with a scale σ and $|\mathscr{F}\{f\}|$ is the amplitude spectrum of $f(x,y)$. The resulting smoothed amplitude spectrum $\mathscr{A}_{\mathscr{S}}(u,v)$ and the *original* phase spectrum are combined to produce the inverse Transform, which in turn, yields the saliency map:

$$\mathscr{S} = \mathscr{F}^{-1}\{\mathscr{A}_{\mathscr{S}}(u,v)e^{i\cdot\mathscr{P}(u,v)}\}. \tag{13}$$

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

$$\mathscr{S} = g \star |\mathscr{F}^{-1}\{\mathscr{A}_{\mathscr{S}}(u,v)e^{i\cdot\mathscr{P}(u,v)}\}|^2. \tag{14}$$

Again consider the very simple illustrative example shown in Fig.4. 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) which correspond to the constant at zero frequency plus two which correspond to the periodic background. In addition, there are two rounded maxima (labeled by a dashed box) corresponding to the salient parts. The complete amplitude spectrum is then smoothed by a Gaussian kernel (row 3), and the signal is reconstructed in the spatial domain using the smoothed amplitude and original phase spectrum (row 4). It is clear that both the periodic background and the near zero-frequency components are largely suppressed while the salient segment is well preserved. Row 5 shows the (spatial domain) saliency map after enhancing the signal shown in row 4 using post-processing. Row 6 illustrates the components actually removed by the previous operations. We find that the non-salient and uniform parts are properly suppressed using amplitude filtering. This process suggests that convolution in the frequency domain of the amplitude spectrum with a Gaussian kernel produces the saliency pop-out in an image using only global information.

Spectrum scale-space analysis: choose the best scale for the Gaussian kernel. Repeating patterns can be suppressed by smoothing the amplitude spectrum using a Gaussian kernel. However, which scale is the best? We propose a Spectrum Scale-Space (SSS) for handling amplitude spectra at different scales, yielding a one-parameter family of smoothed spectra parameterized by the scale of the Gaussian kernel. Given an amplitude spectrum, $A(u,v)$, of an image, the SSS is a family of derived signals $L(u,v;k)$ defined by the convolution of $A(u,v)$ with the Gaussian kernel $g(u,v;k) = \frac{1}{2^k t_0}e^{-(u^2+v^2)/(2^{k+1}t_0)}$, where k is the scale parameter, $k = 1,...,K$. K is determined by the image size: $K = \lceil log_2 min\{X,Y\}\rceil$, where X and Y indicate the width and height, of the image. Thus scale-space is defined as:

$$\mathscr{L}(u,v;k) = (g(.,.;k) * \mathscr{A})(u,v). \tag{15}$$

Fig.5 shows saliency results obtained using different kernel scales, increasing from left to right.. The best saliency map is labeled by a dashed red box. As shown in Fig.5, if the kernel scale is too small, the repeating patterns cannot be suppressed sufficiently, while if the kernel scale is too large, only the boundaries of the salient region are highlighted (see rows 1 and 2 in the figure). Therefore it is important to select a proper scale. Entropy is used to determine the best scale. the appropriate scale k_p is defined as:

$$k_p = \text{argmin}\,(\text{entropy}(saliencymap(k)), \tag{16}$$

where entropy is given by $H(x) = -\sum_{i=1}^{n} p_i log p_i$. The explanation for using entropy to select the best scale is as follows. If a saliency map is good enough, the salient region will pop out from the image, while the common regions will be greatly suppressed. Thus the histogram of the saliency map must cluster around certain values, yielding a very small entropy for the signal. Thus we can find the best scale by finding the map with the smallest entropy value.

3.2 Spatial domain analysis

In this section, we model salient pixels and regions locally. A center-surround template is usually employed to evaluate the distinctiveness of a local area by measuring the local contrast [3, 6].

Use the independent components of natural scenes as the center-surround filters. Difference of Gaussian (DOG) and Gabor filters are commonly used to measure the local contrast. However, recently, researchers have also used Independent Component Analysis (ICA) bases as the filters [6, 10, 14]. It has been shown that by training on tens of thousands of natural image patches, the resulting filters turn out to be quite similar to receptive fields found in the visual cortex [2]. In this paper, we use the 192 color features from [10] as the filters and obtain192 response maps.

Use entropy to assign a weight to each response map. Given the 192 response maps, we calculate a weighted sum to obtain a single saliency map. Thus the saliency is defined as:

$$S(x,y) = \sum w_i (f(x,y) * h_i), \tag{17}$$

where the h_i are the local filters obtained by ICA and w_i is given by the following: $w_i = \text{entropy}\,(f(x,y) * h_i)^{-1}$.

Unlike frequency domain analysis which highlights saliency using global information, spatial domain analysis will enhance only those salient regions that exhibit strong local contrast. Such a "center-surround" model has been adopted in previous work [10, 14].

3.3 The final saliency map

As we have two processing channels (frequency and spatial), we obtain two saliency maps. For convenience, we denote them as S_g, S_l respectively here. S_g is obtained as follows: (1) we first decompose the input color image into a opponent color space: $I = \max\{r,g,b\}$, $RG = r - g$ and $BY = b - \frac{r+g}{2} - \frac{min(r,c)}{2}$. Then we can compute saliency map in each channel as introduced in section 3.1. The entropy values of these three best saliency maps are also used as weights for combining them into the S_g. In the spatial analysis channel, the S_l is computed according to 17. With S_g and S_l, the final saliency map S_f is given by: $S_f = S_g + k.\frac{\text{entropy}(S_g)}{\text{entropy}(S_l)} S_l$. Here, k is a free parameter.

4 Experimental results

Psychological patterns such as those shown in Fig.6 are widely used to evaluate saliency detectors. In row 1, the red bar has a salient color, and our algorithm detects this region properly; row 2 shows a salient region in which there is a bar with a different orientation from others; rows 3-5 show typical patterns used in Gestalt studies and both our algorithm and SR can detect these regions properly; row 6 shows an example of asymmetry and row 7 a salient region where the item is missing. Our algorithm works well in both these examples.

In section 1 we indicated that a good saliency detector should be capable of detecting both small and large salient regions confounded by a cluttered background as well as suppressing repeating objects. Fig.7 presents some results and compares them to important methods in the literature. Row 1 show examples in which the salient regions are quite small. Most existing algorithms work well in this case. However, large salient regions will present a challenge for most algorithms, (see row 2). Our algorithm highlights the flower nearly uniformly.

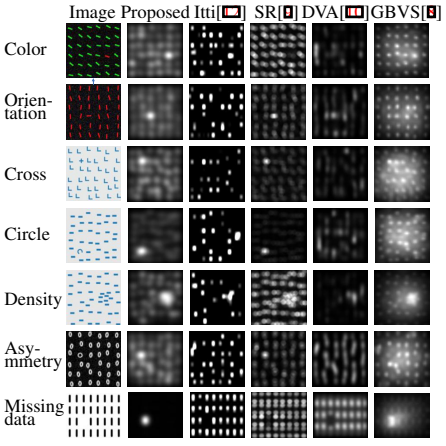

Figure 6: Responses to psycho. patterns

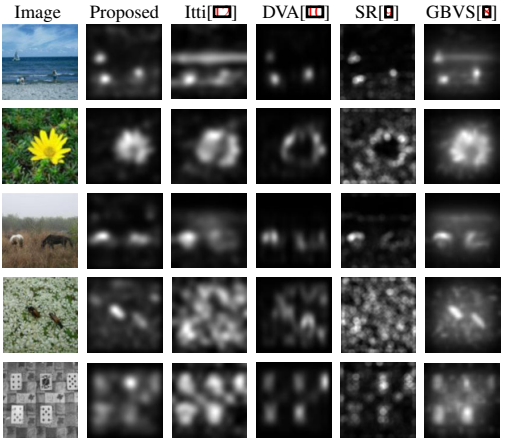

Figure 7: Responses to natural images

Method	ROC (large)	DSC (large)	ROC (small)	DSC (small)	ROC (overall)	DSC (overall)
Our model S_g	0.9293	0.6980	0.9072	0.3421	0.9172	0.5039
Our model S_f	0.9266	0.7103	0.9124	0.3680	0.9189	0.5236
Itti's	0.9020	0.6171	0.9071	0.3344	0.9048	0.4629
SR	0.8152	0.5087	0.9245	0.3637	0.8748	0.4296

Table 1: Comparison between the proposed model with SR[9] and Itti's model [11].

However, the other algorithms only enhance the boundary regions. Row 3-4 show images with cluttered backgrounds. In this case, only the proposed algorithm and GBVS work well, while the other approaches always highlight highly textured backgrounds or other distractors. In row 5, there are five cards in the image. Among these, there is one card that is more distinctive. All of the algorithms can detect these five cards. However, only the proposed method can highlight the most salient card.

Quantitative evaluation was also performed. There are two kinds of categories in our database, one containing 50 images with larger salient regions (labeled as large in the table) and the other 60 images with smaller salient regions (labeled as small in the table). Groundtruth images were labeled by 19 subjects using a method similar to [9]. To evaluate the different models, we used both the Receiver Operator Curve score (area under the curve) and the DSC value (peak value of the Dice Similarity Coefficient curve). DSC curve is defined as $DSC = 2TP/(TP+FP+T)$ and is obtained by sliding the threshold across the whole range, where TP is true positive, FP is false positive and T is the positive in the ground truth. From Table.1, we find that our global model, (S_g), is improved when local and global information are combined (see S_f) . Comparing our model to IttiâĂŹs and SR, we note that the proposed model produces superior performance.

5 Conclusions

In this paper, we argue that, besides predicting human fixation, a reasonable saliency detector should possess the ability to detect both small and large salient regions in cluttered

backgrounds and inhibit repeating objects. Based on these considerations, we propose a new bottom-up saliency model that combines global information from frequency domain analysis and local information from spatial domain analysis. In the frequency domain analysis, instead of modelling the salient regions, we model the common regions (non-salient regions) using global information. Then these so-called repeating patterns that are not distinctive in the scene are suppressed by using spectrum smoothing. In the spatial domain analysis, we enhance those points or regions that are more informative by using a centre-surround mechanism. We demonstrate experimentally that the proposed model has the ability to highlight both small and large salient regions and to inhibit repeating patterns in images. In addition, the model also properly detects saliency in cluttered background images.

Acknowledgments

This work is supported by NSFC under grant 90820302.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned Salient Region Detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. URL http://www.cvpr2009.org/.

[2] A.J. Bell and T.J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision research*, 37(23):3327, 1997.

[3] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *Advances in Neural Information Processing Systems 18*.

[4] Guo C and Zhang L. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Trans Image Process*, 19(1):185–198, 2010.

[5] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–6. IEEE, 2007.

[6] Dashan Gao, Vijay Mahadevan, and Nuno Vasconcelos. The discriminant center-surround hypothesis for bottom-up saliency. In *Advances in Neural Information Processing Systems 20*.

[7] C. Guo, Q. Ma, and L. Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pages 1–8, 2008.

[8] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19*.

[9] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8, 2007.

[10] X Hou and Liqing Zhang. Dynamic visual attention: searching for coding length increments. In *Advances in Neural Information Processing Systems 21*.

[11] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.

[12] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, Nov 1998.

[13] S. Yantis. How visual salience wins the battle for awareness. *Nature neuroscience*, 8(8):975–977, 2005.

[14] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 2008. ISSN 1534-7362.