

Saliency computation via whitened frequency band selection

Qi Lv^{1,2} · Bin Wang^{1,2} · Liming Zhang^{1,2}

Received: 8 April 2015 / Revised: 7 December 2015 / Accepted: 24 December 2015 / Published online: 6 January 2016
© Springer Science+Business Media Dordrecht 2016

Abstract Many saliency computational models have been proposed to simulate bottom-up visual attention mechanism of human visual system. However, most of them only deal with certain kinds of images or aim at specific applications. In fact, human beings have the ability to correctly select attentive focuses of objects with arbitrary sizes within any scenes. This paper proposes a new bottom-up computational model from the perspective of frequency domain based on the biological discovery of non-Classical Receptive Field (nCRF) in the retina. A saliency map can be obtained according to the idea of Extended Classical Receptive Field. The model is composed of three major steps: firstly decompose the input image into several feature maps representing different frequency bands that cover the whole frequency domain by utilizing Gabor wavelet. Secondly, whiten the feature maps to highlight the embedded saliency information. Thirdly, select some optimal maps, simulating the response of receptive field especially nCRF, to generate the saliency map. Experimental results show that the proposed algorithm is able to work with stable effect and outstanding performance in a variety of situations as human beings do and is adaptive to both psychological patterns and natural images. Beyond that, biological plausibility of nCRF and Gabor wavelet transform make this approach reliable.

Keywords Visual attention · Non-Classical Receptive Field (nCRF) · Extended Classical Receptive Field (ECRF) · Gabor wavelet · Whitening · 2D entropy

Introduction

The highly evolved human vision system enables us to rapidly attend to the conspicuous locations within a scene. It is attention mechanism that facilitates us to locate these salient regions. The visual system of human being receives an enormous amount of information from the outside world at each moment. But the information conveyed to the high level of brain is highly reduced through visual processing in the fovea and the ganglion cells in the retina, the lateral geniculation nucleus, the primary visual cortex V1 area and so on. This mechanism can be adopted in computer vision tasks like segmentation (Mishra et al. 2009), object recognition (Liu et al. 2011), visual tracking (Mahadevan and Vasconcelos 2009), image compression (Itti 2004), etc.

With regard to types of attention mechanism, top-down models which are task-driven, and bottom-up models which are stimuli-based, are two main branches. Virtually, these two types of mechanism interact with each other. Top-down attention refers to the process of biased visual perception based on specific tasks or intentions. For bottom-up models, the attended regions are in general sufficiently distinct with respect to surrounding areas, in terms of kinds of low-level features like intensity, color, orientation or motion. Many existing models fall into this category (Achanta et al. 2009; Guo et al. 2008; Itti et al. 1998). Among different kinds of bottom-up models, saliency map (Koch and Ullman 1987), a topological map containing global conspicuity information, is frequently assumed and utilized as it directly demonstrates the attended locations or

✉ Bin Wang
wangbin@fudan.edu.cn

¹ Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China

² Research Center of Smart Networks and Systems, School of Information Science and Technology, Fudan University, Shanghai 200433, China

regions. In this paper, the focus is on computational model in pure bottom-up manner.

The typical biologically inspired model named Neuro-morphic Vision Toolkit (NVT) is proposed by Itti et al. (1998) and it follows the Feature Integration Theory of Treisman (Treisman and Gelade 1980). This model mainly employs subtraction between filtered input with diverse scales to simulate center-surround difference [Difference of Gaussian (DoG) filter], which represents the on–off and off–on effects of visual receptive fields on ganglion cells. Besides, it adopts color opponents and Gabor filter of different orientations with multiple scales, simulating the process of simple cells on visual cortex, to extract visual features. After obtaining the across-scale difference, the model normalizes each feature map to emphasize the active location which is mainly inspired by cortical lateral inhibition mechanisms. By doing so, it calculates saliency with multiple channels and multiple scales on the mechanism of center-surround difference. This model has biological visual structure and basis in spatial domain, and the saliency map is mostly coincident with the fixation focuses of human being in both psychological patterns and natural images.

However, since it employs center-surround differences to simulate the ganglion cells' processing of the retina, some low-frequency information in scenes is largely discarded, which causes the failure of extracting saliency information of large regions that contain a lot of low-frequency components. The operation of normalization (inhibition mechanism) makes high-frequency components stand out as well. As a result, this model basically extracts salient points instead of consistent area. There are actually many other models derived from this baseline one (Harel et al. 2007; Le Meur et al. 2007; Walther and Koch 2006).

Another category of models (Guo et al. 2008; Hou and Zhang 2007; Bian and Zhang 2010; Yu et al. 2011; Li et al. 2013) are based on frequency domain analysis, which have fast computational speed. Guo et al. propose the Phase spectrum Quaternion Fourier Transform (PQFT) model which pops out the edge information of objects, since phase information is related to local properties (form and position) of the image (Oliva and Torralba 2001). Li et al. introduce Hypercomplex Fourier Transform (HFT) model, in which multiple Gaussian filters with diverse scales are used to filter the log-amplitude spectrum in order to highlight the salient information. No evidences show the occurrence of frequency operations in human brain, so such models are basically not consistent with the visual system of human being.

There are also some engineering application based models which are designed for specific applications such as large object segmentation, object recognition etc. From the viewpoint of frequency band, object with large size covers

more low frequency components. So Achanta et al. proposed a Frequency-Tuned Saliency (FTS) algorithm which retains most of the frequency components of images in order to realize better segmentation. It calculates saliency in spatial domain by simple subtraction between Gaussian filtered image and its global mean, i.e. most frequency components are preserved except the direct current (DC). This model is computationally efficient and has good performance. However, this kind of models (Cheng et al. 2011; Perazzi et al. 2012) is only effective for large object segmentation.

To our knowledge, it turns out that most existing models focus on certain kinds of objects or only adapt to fixed situations. For instances, the NVT model is merely able to extract salient points with much high-frequency information, while those engineering-based models are well applicable to large objects recognition. The idea of the paper is to build a computational model which has the ability to automatically make adjustments according to stimuli. So, this paper proposes a bottom-up model from the perspective of frequency domain though manipulated in spatial domain, based on the biological discovery of non-Classical Receptive Field (nCRF) of the ganglion cells in the retina. Thanks to the discovery of nCRF (Li et al. 1992), which complements and interacts with CRF according to the stimuli, the low-frequency loss caused by center-surround mechanism can be largely compensated by the tuning of nCRF areas outside the CRF. The conception of nCRF and CRF could be considered related to the processes of magnocellular cells of lateral geniculate nucleus (LGN) and those of parvocellular cells of LGN (Shi et al. 2011).

The phenomenal traits of the proposed model are: (1) improving the classical Itti's NVT model by employing Gabor wavelet transform, taking multi-orientation and multi-scale (Itti's) into account, as well as retaining the low-frequency components in order to better tune the frequency bandwidth adaptively; (2) based the discovery of nCRF, proposing one way to adaptively adjust the frequency band according to diverse stimuli and a method to select the optimal scale in Gabor wavelet domain.

The rest of paper is organized as follows. In “[Biological background for the algorithm](#)” section, the biological background of the proposed model will be introduced. In the next part, “[The proposed algorithm](#)” section, the proposed algorithm will be described in detail including how to decompose frequency bands with Gabor wavelet and how to whiten and select them. Experimental results and comparisons are followed in “[Experimental results and discussions](#)” section with discussions on various models. Finally, conclusions are made in “[Conclusions and future work](#)” section.

Biological background for the algorithm

As mentioned before, some spatial and spectral models would fail in some scenarios especially when salient regions are relatively large due to the center-surround processing they employ or insufficient use of spectral information. On the contrary, some engineering-oriented models appear to be remarkable on larger objects while small objects or psychological patterns are beyond their reach. Therefore, to learn how human vision system really works, the receptive field models of the ganglion cells in the retina are examined here.

The Classical Receptive Field (CRF) is a center-surround antagonism structure of the retinal ganglion cells and Rodieck et al. propose the DoG function to depict it. In frequency domain, the DoG structure is typically represented as a ring band shown in Fig. 1a (left). Many bottom-up models, like NVT etc., adopt this structure in their saliency computations, but it is actually flawed since it merely contains certain high pass bands, while leaves out the low-frequency components at center region shown in Fig. 1a, even if several DoGs of different scales are adopted. It might result in high lighting only small salient regions or edges of large objects but failure to capture the whole object.

Physiologists have found out in 1960s, however, that center-surround CRF can be influenced by a larger region outside the CRF (Ikeda and Wright 1972). This area is regarded as non-Classical Receptive Field (nCRF) that can inhibit the antagonistic effect of center-surround and compensate the loss of low-frequency components. In order to explain the relation between CRF and nCRF, Ghosh et al. (2006) suggested the following equation named ECRF using three zero-mean Gaussians with different variances:

$$ECRF(\sigma_1, \sigma_2, \sigma_3) = A_{CRF} \left(\frac{1}{2\pi\sigma_1^2} e^{-\frac{x^2+y^2}{2\sigma_1^2}} - \frac{1}{2\pi\sigma_2^2} e^{-\frac{x^2+y^2}{2\sigma_2^2}} \right) + A_{nCRF} \frac{1}{2\pi\sigma_3^2} e^{-\frac{x^2+y^2}{2\sigma_3^2}} \tag{1}$$

where $ECRF(\cdot)$ represents the response function, σ_1, σ_2 and σ_3 ($\sigma_1 < \sigma_2 < \sigma_3$) are variances representing region size of the center, the antagonistic surround and the extended non-inhibitory surround respectively, A_{CRF} and A_{nCRF} represent the corresponding amplitudes of both structures. The first two terms refer to the classical center-surround structure (DoG filter). The last term is the compensating extended function which is a Gaussian with larger variance in spatial domain.

Figure 1a shows frequency band of classical DoG filter (the first two terms in Eq. (1)) with $\sigma_2 = 3\sigma_1$ in 1D (right) and 2D (left) cases. Figure 1b is the frequency band of ECRF structure with $\sigma_2 = 3\sigma_1$ and $\sigma_3 = 3\sigma_2$. Note that in Fig. 1b both structures have the same amplitudes that $A_{CRF} = A_{nCRF}$. In Fig. 1b, the central part of frequency domain, which contains much of low-frequency information, is well preserved. It should be noted that the DC component (original point in Fig. 1b) is not necessary which will be removed in subsequent processing. Although the structure in Fig. 1a does not include the DC component as well, it fails to contain the low-frequency information around the DC component.

It is illustrated in Fig. 1 that the structure of nCRF is to adjust the frequency bandwidth of DoG filter thus compensate the loss of low-frequency information. That is why human beings can easily pay attention to salient objects in a scene with arbitrary size. In contrast, those models using center-surround filters to calculate saliency actually do not include low-frequency components. And the model like PQFT, flattening the amplitude spectrum, is just to heighten high-frequency components and extrude edges of object. For the model like FTS, it always extracts the same range of frequency band no matter what the stimulus is, in fact however, the range of frequency bands should vary according to the stimuli.

Our purpose of saliency computation is mainly to extract certain frequency bands of input image according to the input stimuli, which might be consistent with the idea of grained-scale process and minute-scale process with M (magnocellular) and P (parvocellular) pathway respectively

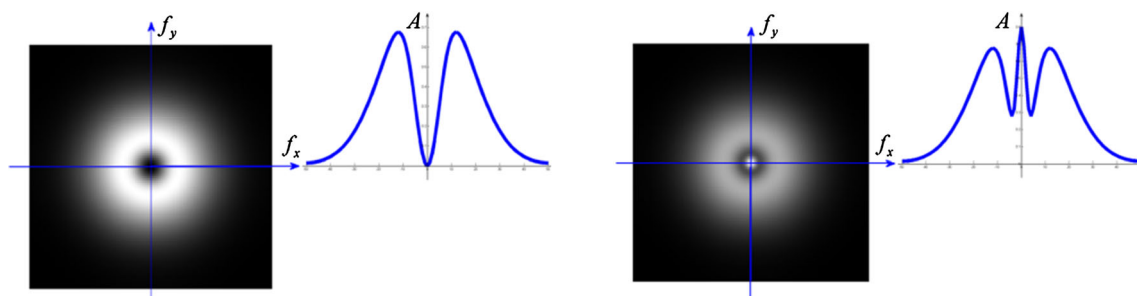


Fig. 1 Left frequency domain of DoG; right frequency domain of ECRF, the central part is the effective area of nCRF

(Shi et al. 2011). In order to achieve this, the whole frequency domain of input image is decomposed into several bands in a discrete way. Based on these discrete bands, some optimal bands containing meaningful saliency information are selected to build the final saliency map. By using decomposition and selection, it is convenient to take into account both the multi-scale subtraction of center-surround process employed by Itti's model and the retention of low-frequency information which is not considered in Itti's model.

With regard to the method of frequency domain decomposition, wavelet transform is utilized here to carry out bands division. Discrete wavelet transform, which takes both spatial and spectral information into consideration, performs a logarithmic division of frequency domain. This is more practical than FFT as low-frequency components are always with low spatial resolution and thus need detailed division in frequency domain. A fine division made on low-frequency components can achieve a better effect on saliency computation. In addition, wavelet transform is of multi scales, representing different bands in frequency domain, and it can categorize each frequency band into different orientations that like simple cells do in the primary visual cortex (except low-frequency part). This also facilitates the calculation of saliency. These orientations of wavelet transform in each frequency band correspond to different sub-bands. These sub-bands will be whitened across channels to highlight the saliency information, so do the low-frequency ones. After that, optimal bands among high-frequency or low-frequency ones could

be selected. These operations can partly simulate the mechanism of frequency bandwidth adjustment and achieves the same effect of frequency band selection.

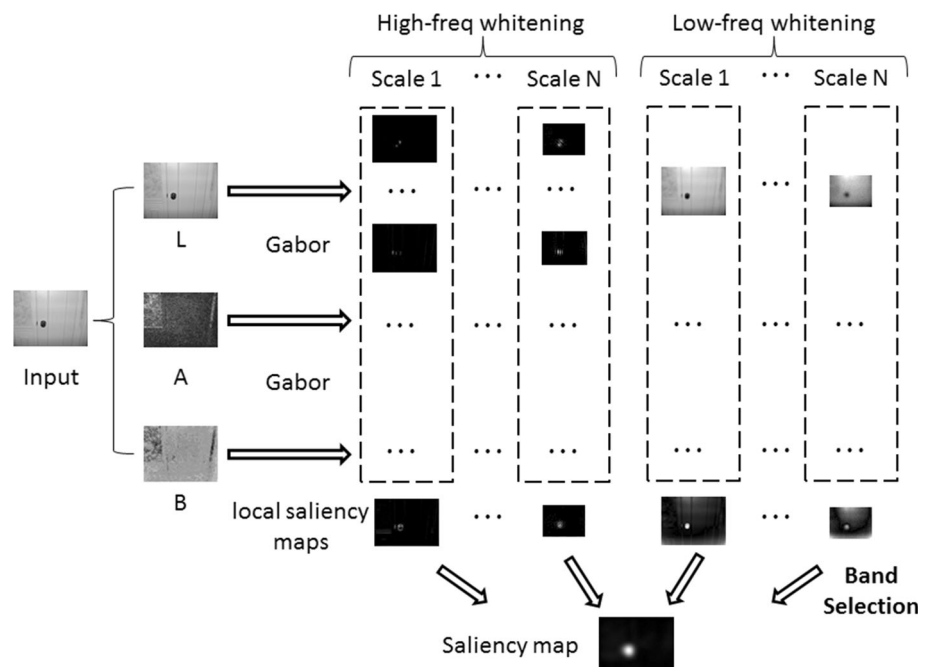
The proposed algorithm

The basic steps of proposed algorithm are band division, whitening and band selection. The diagram is shown in Fig. 2. Prior to the processing, the original color image is converted to Lab color space to form three channel images. For each channel image, Gabor Wavelet is employed to decompose the channel into different feature maps corresponding to different frequency bands. After whitening and fusion, one or two local saliency maps corresponding to certain frequency bands are selected to generate the final saliency map. Each step in diagram of the algorithm is shown in Fig. 2.

Gabor decomposition

With regard to different categories of wavelet functions, Gabor function is adopted here to carry out decomposition because it is similar to the process, which is also employed by Itti's NVT model to analyze the orientation information, of simple cells in the primary visual cortex. Additionally, the low-frequency components of Gabor wavelet domain are maintained. The illustration and formula of Gabor filter are shown in Fig. 3 and Eq. (2), respectively.

Fig. 2 Diagram of the algorithm



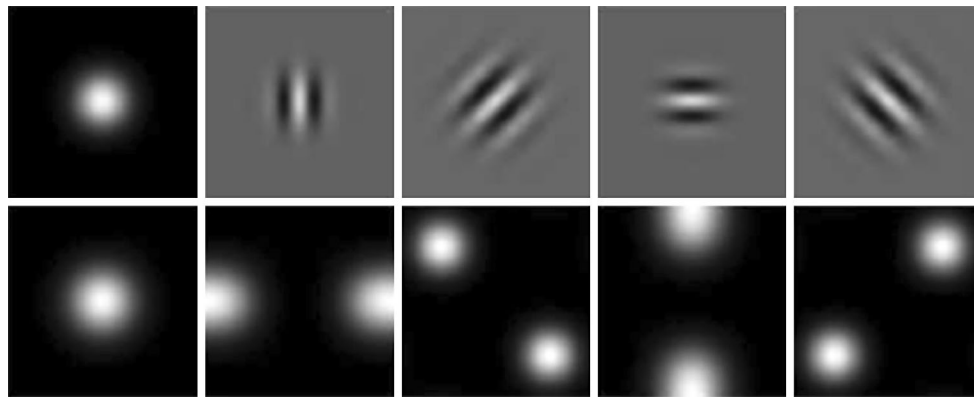
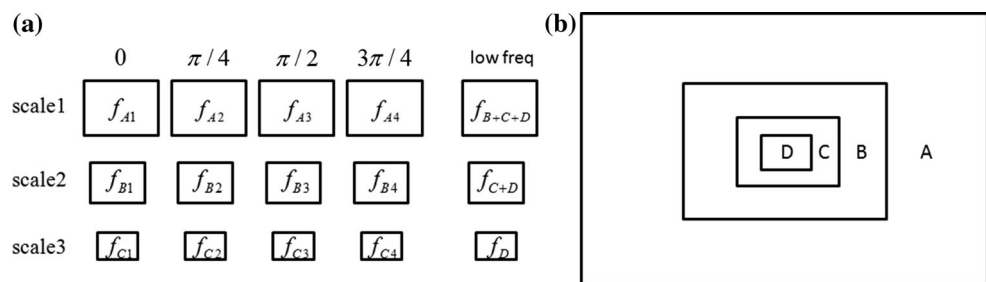


Fig. 3 Five Gabor 2D filters (*top row*) with their corresponding amplitude spectrum (*bottom row*). From *left to right* low-frequency part, high-frequency part with 0°, 45°, 90°, 135° orientations

Fig. 4 Illustration of Gabor decomposition and relationships between feature maps and bands. **a** High-frequency sub feature maps and low-frequency feature maps. **b** Corresponding frequency domain



Therefore, 2D Gabor filter rather than Fast Wavelet Decomposition is employed to accomplish wavelet decomposition in order to obtain more information on orientations. The 2D Gabor function is:

$$g(x, y; \lambda, \theta, \sigma) = \exp\left(-\frac{x'^2 + y'^2}{2\sigma^2}\right) \cos\left(2\pi\frac{x'}{\lambda}\right) \quad (2)$$

where $x' = x \cos \theta + y \sin \theta$, $y' = -x \sin \theta + y \cos \theta$, and $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$. And λ is the wavelength, σ^2 is the variance of the Gaussian envelope. Four band pass filters and one low pass filter (when λ approaches to infinity), together amount to five 2D Gabor filters. These five Gabor filters can almost cover the whole frequency domain at each scale. They are shown in Fig. 3.

The diagram of Gabor decomposition with three scales in one channel is shown in Fig. 4 and the relationships between spatial feature maps (Fig. 4a) and spectral bands (Fig. 4b) are also illustrated.

Please note that the actual sizes of feature maps f_{Bx} are half of those of f_{Ax} and f_{Cx} are also half of those of f_{Bx} , for $x = 1, 2, 3, 4$ and sizes of feature maps at the same scale are equal. Since these feature maps correspond to various frequency components, selections made on these maps are equivalent to those made on frequency components. Therefore, this approach calculates saliency based on the feature maps in spatial domain. In the following sections, feature maps are used to represent frequency bands.

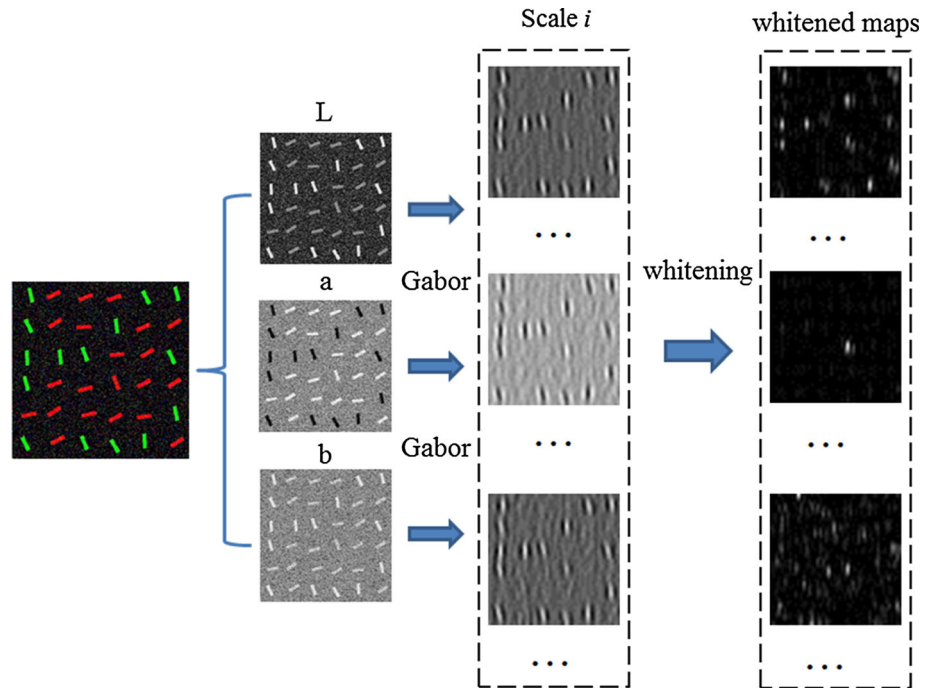
For Eq. (2), $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ and the scale is $\sigma = 7/5$. The low-pass 2D Gabor filter sets λ a large number like $\lambda_{low} = 2.5^{10}$, and four other high-pass ones set $\lambda_{high} = 2.5$. The sizes of these filters are 15×15 pixels (shown in Fig. 3). Experimental results indicate that saliency map computation is insensitive to the parameters of Gabor filters, as long as the 5 Gabor filters could cover the whole frequency domain. Many more orientations can be included as well.

Concretely, the input channel image is filtered to generate feature maps repeatedly with Gabor wavelet filters and the low-frequency feature map is keeping down-sampled until the height of decomposed map is less than 32 pixels. Besides, if the height of input image is greater than 256 pixels, the feature maps at first scale are discarded as they contain less significant information even most can be considered as noise. After filtering, four high-frequency feature maps and one low-frequency feature map at each scale (scale 1 to scale N , the typical value of N is 3–5 according to the original size of input image) shown in Fig. 2 are obtained.

Whitening and computation of local saliency maps

The processing of feature maps includes whitening which aims to extract saliency information and fusion which sums the whitened feature maps to generate the local saliency maps.

Fig. 5 Illustration of whitening



At each scale for all channels, the high-frequency feature maps and low-frequency ones are whitened separately using Zero-Phase Component Analysis (ZCA) whitening (<http://ufldl.stanford.edu/wiki/index.php/Whitening>), a method similar to Principal Component Analysis (PCA) whitening. Whitening is the process of decorrelation and orthogonalization between feature maps. After whitening, DC component is removed and the variance is normalized for each map. As a result, the unique part of data is underlined while the redundancy is suppressed. The idea of whitening is also employed in several works (Garcia-Diaz et al. 2012; Bian and Zhang 2010). Then the square of whitened feature maps is summed to get local saliency map. There are two local saliency maps at each scale, corresponding to low- and high-frequency bands respectively.

Whitening

ZCA-based signal whitening is operated on feature maps at the same scale across channels for high- or low-frequency bands separately, as shown in the dashed boxes of Fig. 2, so that the resulting feature maps become orthogonal and uncorrelated with each other. Let \mathbf{f}_i , $i = 1 \dots n$ be the vectorized feature map for a given scale. $\mathbf{W} \in \mathbf{R}^{n \times n}$ is a ZCA whitening matrix and the whitened result \mathbf{F}' is:

$$\mathbf{F}' = \mathbf{W}\mathbf{F}, \mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2 \dots \mathbf{f}_n]^T, \mathbf{f}_i \in \mathbf{R}^{M \times 1} \quad (3)$$

where n is the number of feature maps at the same scale, $n = 4 \times 3$ for high-frequency feature maps, and $n = 3$ for low-frequency feature maps. M is pixel number of a feature

map. It should be noted that for high- or low-frequency feature maps at different scales, the whitening matrix is quite differed. After whitening, each whitened feature map has unitary variance and zero mean. Whitening can not only remove the DC component which is not necessary for further processing, as mentioned above, but can also highlight the saliency information.

Figure 5 gives a target search example of psychological pattern in conjunctional conditions, i.e. the unique red 90° bar is inserted in red 0° bars and green 90° bars with orientation disturbance. Most visual attention models would fail in this case but ours would not. By whitening the feature maps of high frequency at scale i , the unique bar is popped out in a channel while the others are suppressed.

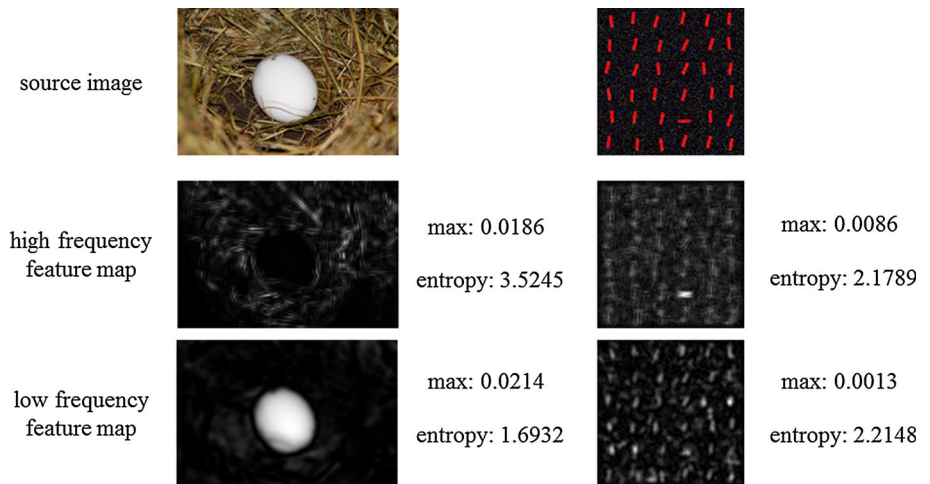
Computation of local saliency maps

For high-frequency whitened feature maps, the local saliency map is the simple quadratic sum of all whitened maps at each scale. The formula is:

$$S_{hi} = \sum_{j=1}^{3 \times K} f_{hij}'^2 \quad (4)$$

where S_{hi} is the high-frequency local saliency map of i th scale, K is number of orientations and is set as 4 here, f_{hij}' is the j th whitened high-frequency feature map. After computation according to Eq. (4), N local saliency maps for

Fig. 6 Illustration of selection principle



N scales in high-frequency bands are obtained shown in Fig. 2, which correspond to N bands.

For low-frequency feature maps, the fusion strategy is somewhat diverse. After whitening, the whitened feature maps are combined with certain weights. The weight function is monotonically decreasing with the increase of 2D entropy:

$$S_{li} = \sum_{j=\{L,a,b\}} \exp(-0.5 \times (entropy_{2D}(f'_{lij})^2) \times f'_{lij} \quad (5)$$

where S_{li} is the low-frequency local saliency map of i th scale, f'_{lij} is the j th whitened low-frequency feature map, and $entropy_{2D}(\cdot)$ is the 2D entropy value calculated by Eq. (7) (which will be detailed in the following section) with quantization level setting to eight for compromise between accuracy and computational cost.

The reason for taking weight into account is that large objects are with low responses after whitening due to unitary variance. If the sum is made with same weights like Eq. (4), the larger salient objects will be suppressed. It will be shown in Fig. 6 that a clear structure of large object has low 2D entropy. By adding this weight, the importance of feature maps containing large salient objects will not be diminished. Thus, the saliency for large objects is popped out by whitening and preserved by weighted summation. As mentioned above, there are N local saliency maps for low-frequency band.

Selection via importance measure

Dozens of local saliency maps covering different frequency bands are obtained. Among them one or two maps (bands) with most significant saliency information will be picked out. Therefore, an importance measure criterion is set up to complete the selection task. It incorporates two metrics: one is the maximum response of feature map and the other

measures the clutter degree. The 2D entropy (Abutaleb 1989; Yang et al. 1996) is used in this paper to estimate the clutter degree of a map.

To calculate 2D entropy, a 2D gray-level histogram taking spatial relations into account is formed in advance by comparing the original image $f(x, y)$ and the averaging filtered version $g(x, y) = m * f(x, y)$, where m is a 2D mean filter with 3×3 pixels. The 2D histogram is a $L \times L$ square matrix, where L represents number of gray levels. A pixel located at (x, y) in a map which refers to i th gray level in $f(x, y)$ and j th gray level in $g(x, y)$ contributes one counting unit on r_{ij} , where r_{ij} denotes the number of pixels which are at i th gray level of $f(x, y)$ and at j th level of $g(x, y)$. After scanning all pixels, the element of 2D histogram p_{ij} is calculated as follows:

$$p_{ij} = \frac{r_{ij}}{PNUM} \quad (6)$$

where $PNUM$ represents total pixel number of a map. Then the 2D entropy of a map can be calculated based on the generated 2D histogram:

$$entropy_{2D} = - \sum_{i=1}^L \sum_{j=1}^L p_{ij} \log p_{ij} \quad (7)$$

where p_{ij} is calculated according to Eq. (6).

According to the definition, the 2D histogram mainly takes edge change into account since uniform regions scarcely alter their grey level after averaging filtering. If a map is topologically compact, which means less edge information, the averaged map may still contain relatively less edge information. On the other hand, when a scene is cluttered, smooth filtering may lead to excessive gray level change which accordingly generates relatively greater value of 2D entropy. So, the smaller the 2D entropy value is, the more significant a map is.

By conducting experiments, we find out that 2D entropy value and maximum response of a map can be employed together to be the criterion to select bands. Actually, 2D entropy is defined to measure the clutter degree of an image. The low-frequency part of a scene containing large objects with compact structures usually has lower 2D entropy value. As a result, we tend to choose feature maps of low-frequency in this case. But an image with a single small object appears to have higher response on high-frequency side. So we are inclined to select high-frequency components in this scenario. Taking both factors into account, we set a criterion which favors maps with higher responses and smaller entropy values to make selection. Two examples are shown in Fig. 6, where the low-frequency feature map in the left column has smaller 2D entropy value and the high-frequency map in the right column has higher response.

In other words, optimal local saliency maps with more intensive response and smaller 2D entropy value are picked out, which indicates less clutter or chaos, simultaneously. These operations are meant to adjust the bandwidth information so as to achieve better effect of saliency information extraction. The selection of bands is based on the importance measure of each local saliency map shown as follows:

$$IM = \max(\text{map}) \times \exp(-0.5 \times (\text{entropy}_{2D}(\text{map}))^2) \quad (8)$$

where IM is the importance of each map, $\max(\cdot)$ is the function used to measure the maximum response of a map, $\text{entropy}_{2D}(\cdot)$ is the 2D entropy value. And map represents the local saliency map indicated in Fig. 2.

Suppose the original image is decomposed into N scales. There are $2 \times N$ candidate local saliency maps to be selected from, where N maps are related to low-frequency part while the other N are related to high-frequency bands. One or two local saliency maps with no overlapped bands are picked out as optimal maps, either high-frequency or low-frequency, or even both. The selection of optimal local saliency maps depends on their importance values which are calculated by Eq. (8).

To begin with, the local saliency map with the largest importance value IM_{\max} is picked out. If there are no other maps with their importance values larger than $0.5 \times IM_{\max}$, then the local saliency map is just the final saliency map. But if there exist other maps with their importance values larger than $0.5 \times IM_{\max}$, another one with the second largest importance value is selected. In this case, the final saliency map is the result of equal combination of the two maps if their bands are not overlapped. It is worth noting that if the map with the second largest importance is overlapped with the largest one, the third or fourth largest one will be considered.

Experimental results and discussions

To make a comprehensive evaluation on the proposed model, the testing databases include both natural images and psychological patterns/images. Natural images contain not only small sized salient objects but also large ones.

In order to better illustrate the superiority of the proposed model, the comparisons are performed between our model and several state-of-the-art models, including the typical model in space domain, NVT (Itti et al. 1998), the representative spectral model using FFT, PQFT (Guo et al. 2008), the large object segmentation oriented model on engineering, FTS (Achanta et al. 2009) and the model adaptive to various kinds of stimuli, HFT (Li et al. 2013).

Among all of these models, PQFT resizes input to the resolution 64×64 and HFT resizes input to 128×128 as optimal defaults while others do not carry out resolution adjustment as well as our approach. It is proved that resize of input may make the computation fast but probably leads to irreversible information loss. This will be discussed later.

Quantitative and subjective evaluation

For psychological patterns, the saliency results of several common cases are listed to make a subjective evaluation on each model. For natural data, the outputs of each model are compared with the ground truth in a quantitative manner.

The ground truth data are based on human visual behaviors and mainly include two types: fixation maps and labeled maps. A fixation map is record of human fixation within one image by eye tracking apparatus. Data of this kind are binary maps, with logical 1 (fixation points) dotted over the whole image. Ground truth maps of the other kind are also binary maps, but with consistent areas indicating logical 1 which are labeled by a number of subjects. For the fixation dataset, sAUC (shuffled Area under ROC Curve, the larger is better) is used to measure the performance as it eliminates the influence of center-bias (Tatler et al. 2005; Zhang et al. 2008) while all other metrics are all susceptible to center-bias effect. For the other dataset, segmentation dataset, Precision/Recall is adopted to be a metric. In the calculation of precision and recall, the saliency maps are transformed to a binary map with varying threshold from 0 to 1. Comparing the binary map with ground truth, precision rate is calculated as the number of true positive (intersection of predicted foreground and true foreground) to the number of predicted foreground while the definition of recall rate is the number of true positive to the number of true foreground. A better model has larger area covered by P/R curve.

For all of these models, the final saliency maps are blurred to get optimal effects. We use Gaussian filters with different sigma parameters to blur the saliency maps and pick out the optimal one for each model. The blurring factor is chosen from $0.01 \times \text{width}$ to $0.1 \times \text{width}$ with 0.01 as interval.

Saliency prediction for natural images

The fixation datasets include Bruce's (Bruce and Tsotsos 2005), Kootstra's (Kootstra et al. 2008) and Judd's (Judd et al. 2009). They contain 120, 100 and 1003 natural images, respectively. The segmentation datasets consist of Achanta's (Achanta et al. 2009), Li's (Li et al. 2013) and Zou's (Zou et al. 2013). They have 1000, 235 and 1500 images, respectively. Zou's dataset is derived from PASCAL VOC 2012 segmentation challenging.

Some results for natural images are shown in Fig. 7. The quantitative comparisons for fixation datasets and segmentation datasets are shown in Table 1 and Fig. 8 respectively.

Table 1 shows the quantitative comparison between models in terms of sAUC. The proposed model is proved to be effective over all of these fixation datasets. Meanwhile, PQFT also shows relatively good performance over such type of datasets.

Table 1 sAUC comparison on fixation datasets

Datasets	Ours	HFT	NVT	PQFT	FTS
Bruce's (120)	0.7180	0.6852	0.6052	0.7108	0.5822
Kootstra's (100)	0.6088	0.5945	0.5788	0.6031	0.5665
Judd's (1003)	0.6849	0.6646	0.6132	0.6838	0.5862

Bold values indicate the highest performance

The first column of Fig. 7 consists of five original images with different sizes. Their resolutions are 681×511 , 400×300 , 763×512 , 400×300 and 333×500 from the first to the fifth row respectively. The sizes of objects in these original images are also different. The images of the first and third row include small objects (a man stands by a tree, two people near to a snow mountain), while the large sized objects are arranged on second and fourth row. The image in the last row contains multiple objects. Some models, except NVT, FTS and the proposed, resize the input images (i.e. HFT resize to 256×256 , PQFT resize to 64×64) for better performance. Though NVT and FTS do not perform resizing of input images, they are not effective for both small targets and large ones simultaneously, i.e. NVT is effective only for small ones and FTS only for large ones. It is worth noting that the proposed model is not subject to the original

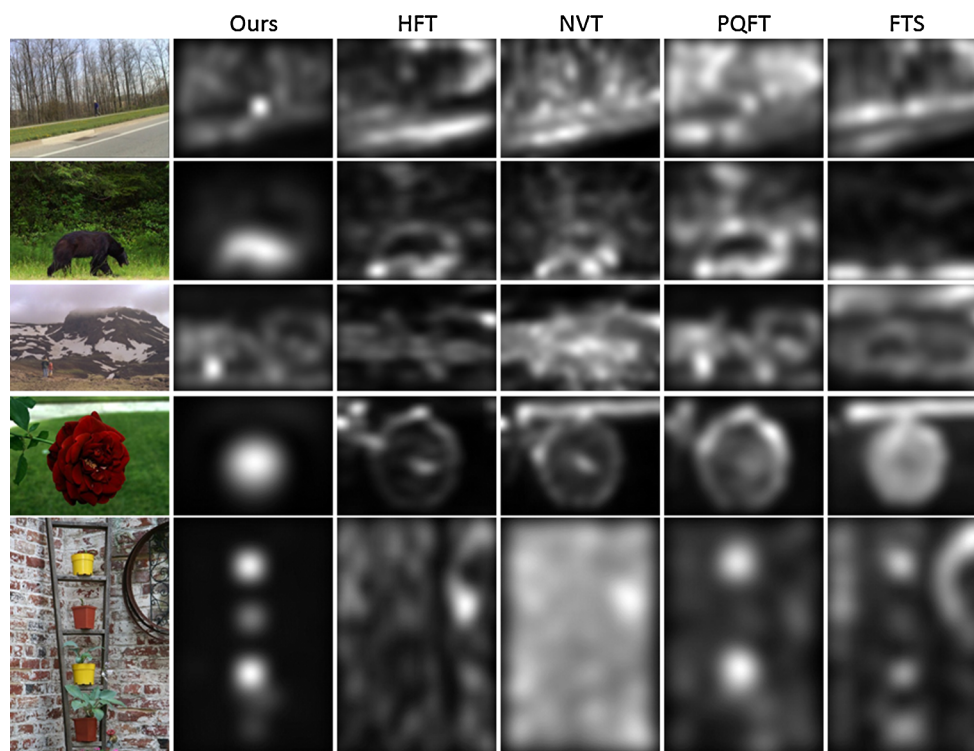


Fig. 7 Some saliency maps of different models

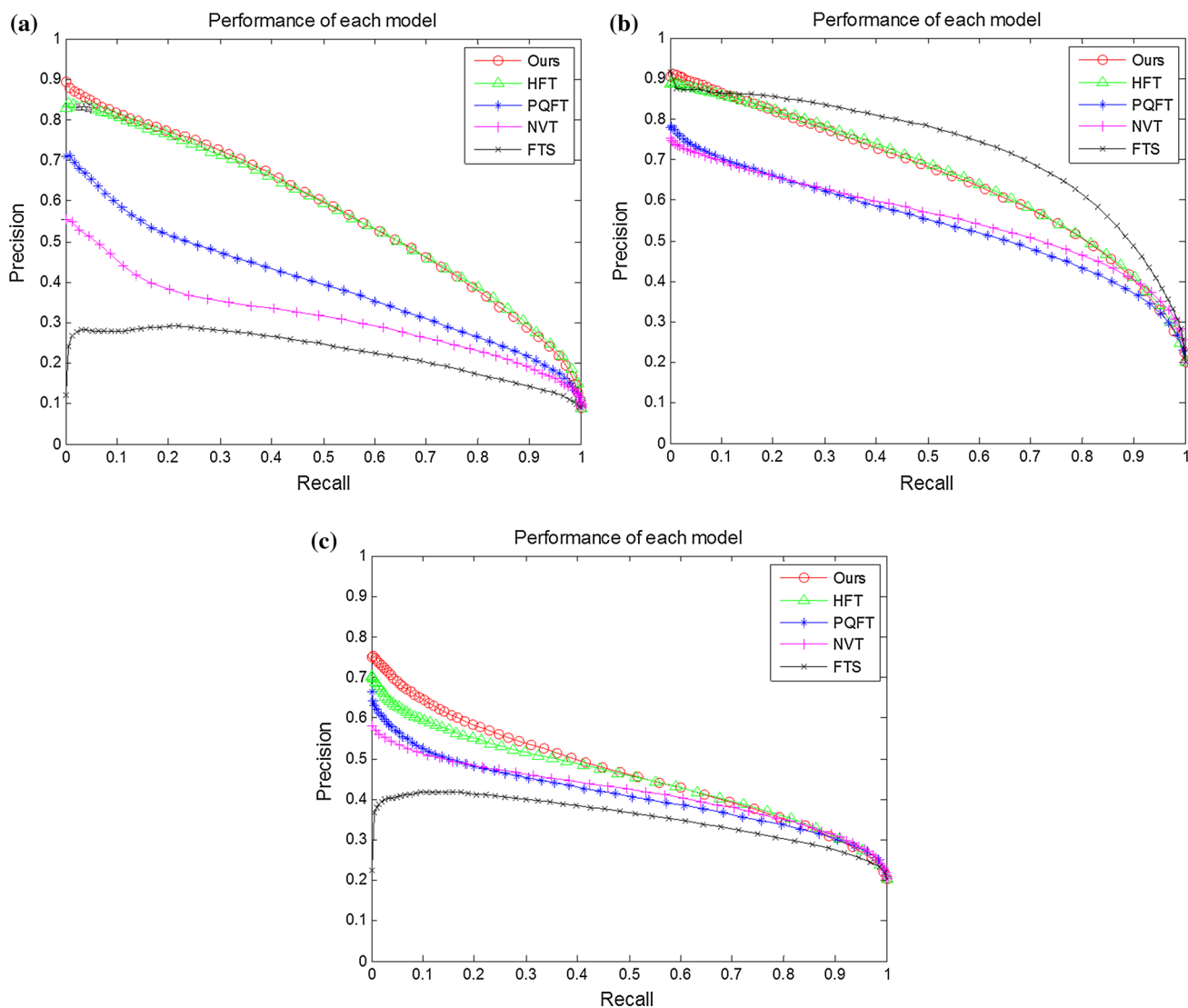


Fig. 8 Quantitative comparison on segmentation datasets. **a** Li's dataset (235). **b** Achanta's dataset (1000). **c** Zou's dataset (1500)

size of input image (do not need to resize image for subsequent process) and is able to pop out objects in diverse sizes. The original height-width ratio is maintained during the whole calculation process.

Figure 7 illustrates that our model is proved to be effective on small salient regions, large ones and images with multiple targets while PQFT and NVT only highlight small objects or edges. FTS always fails when salient objects are relatively small. The results of HFT are not very satisfying. Besides, Fig. 8 indicates that the proposed model is also able to deal with large salient objects. It is worth noting that FTS only has good performance on its own dataset. For Zou's dataset (Zou et al. 2013), it contains large amount of images with multi segments or objects of different scales, which clearly shows that our model is more robust than others (Fig. 8c).

Table 2 Average computational time per image

Models	Ours	HFT	FTS	PQFT	NVT
Time (s)	0.3396	0.3605	0.0885	0.095	1.3091

In Table 2, the average time consumption of each model by calculating 120 images of Bruce's database is listed, where images are uniformly of size 681×511 . All codes are written in Matlab and the computer works on Windows 7 platform with an Intel i7-2600 CPU.

For unbiased comparison, the input images are resized to 256×256 for all models. PQFT and FTS are the fastest as their processes are very simple. HFT is relatively slower because it employs 8 scale spaces to analyze the frequency domain. Time consumption of our model consists of

decomposition, whitening and map selection. The NVT model is the most computationally expensive as it produces too many features maps and uses iterative normalization.

In order to show the importance of band selection with both 2D entropy and maximum response, we have conducted experiments with different strategies. Several cases are compared: the proposed model, bands simply combined without selection, bands selected only using 2D entropy and selected only with maximum response. Experiments are conducted over the Bruce dataset (Bruce and Tsotsos 2005) and the comparison is shown in Table 3.

The comparison in Table 3 indicates that taking both 2D entropy and maximum response of maps is the optimal strategy to generate saliency map. And combining all the frequency bands has the least satisfying effect since much unnecessary information is included.

Saliency prediction for psychological patterns

Different types of psychological patterns construct another test bench which is important criterion to measure the performance of attention models. Figure 9 shows that the proposed model can deal with all cases of psychological patterns. All models fail case 1 and 3 except ours. The reason is that the whitening process makes the unique color component salient, thus our model is able to predict the saliency of them. These patterns prove the biological plausibility of the proposed model cogently.

PQFT fails some cases especially when distinct pattern is relatively large. It turns out that HFT shows good potential on these patterns as well as our model, except that it fails the first and third rows. FTS focuses on image segmentation and it fails most of these patterns naturally. NVT is also less effective for these patterns.

Discussions

The proposed model is built on the basis of nCRF and it takes the low-frequency information, which is mostly ignored by existing models, into account by considering range of frequency bands. This is, to some extent, consistent with human visual system. The experimental results also prove the feasibility of the proposed model.

A few parameters are involved in the proposed model. The modification of parameters for Gabor filter does not make much difference. And the number of scales decomposed is relatively fixed at 3–5. Besides, the process of whitening is almost parameter free.

With regard to other models, PQFT totally discards the amplitude information (by flattening the amplitude spectrum) and only phase information is utilized for saliency map construction, which leads to only edges being popped out. Besides, top-down instructions are difficult to be contained in this model because it employs quaternion and Fourier transform.

For FTS, as it is only effective on its own database (most images with large salient areas) but fails others, it indicates that retaining most of frequency components is effective for large objects (low-frequency components are crucial for large objects and are contained in FTS). The key defect is that it extracts fixed band width of information for all images. Despite this strategy is effective for large objects, but for small objects, specific bands are required even of certain orientations. Retaining too much frequency information is not necessary and may obstruct saliency prediction in some cases.

The NVT model suffers the problem that center-surround operations exclude much of the low-frequency information, and this information actually contributes a lot when the salient region is relatively large. However, this model is a cognitive model based on biological plausibility and top-down manner is easily manipulated in this kind of model (Zhang et al. 2008).

About HFT, it utilizes Gaussian filters of different scales to filter the log amplitude spectrum of input image to calculate saliency, which lacks enough biological support and the meaning remains unclear in spatial domain.

Conclusions and future work

The paper proposed a saliency model from the perspective of frequency domain by selecting certain bands though implemented in spatial domain. Three main steps are: dividing input image to different feature maps (frequency bands), whitening the feature maps to extract saliency information and picking out optimal maps containing significant saliency information according to the mechanism of receptive field. Our approach turns out to be superior compared with others on various kinds of stimuli, including psychological patterns and natural ones with large or small salient areas. Beyond that, top-down manners or prior knowledge can be easily included. As images are divided into many channels, scales and orientations, diverse weights can be assigned to feature maps when specific tasks are involved.

Table 3 Comparison of different strategies

Method	Proposed	Simply combine	Entropy only	Max response only
sAUC	0.7180	0.6573	0.6958	0.7012

Bold value indicates the highest performance

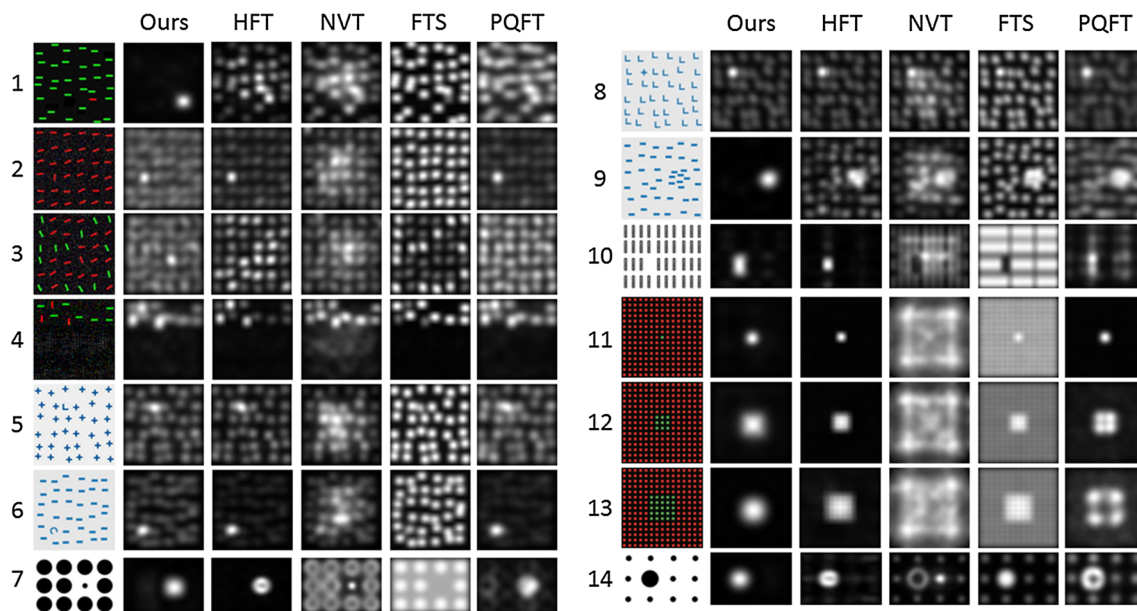


Fig. 9 Saliency maps of psychological images

However, our model still suffers a couple of drawbacks. For one thing, this algorithm requires a bit more computational cost compared to spectral methods. For another, some points are not entirely consistent with specific biological mechanism, for example, 2D entropy as a measure to select feature maps lacks of biological support.

The future work will make more emphasis on how to better match with biological mechanism and how to simplify the calculation of decomposing and selecting bands since these processes are both spatially and temporally complex. Moreover, top-down mechanism appears to be more important as we have interest in target detection in remote sensing images with attention models. We will attempt to combine top-down mechanism with this bottom-up model to deal with target detection tasks in the future.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (Grant No. 61572133) and the Project from State Key Laboratory of Earth Surface Processes and Resource Ecology under Grant 2015-KF-01.

References

- Abutaleb AS (1989) Automatic thresholding of gray-level pictures using two-dimensional entropy. *Comput Vis Graph Image Process* 47(1):22–32
- Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, 2009. IEEE, pp 1597–1604
- Bian P, Zhang L (2010) Visual saliency: a biologically plausible contourlet-like frequency domain approach. *Cogn Neurodyn* 4(3):189–198
- Bruce N, Tsotsos J (2005) Saliency based on information maximization. In: *The proceedings of the neural information processing systems conference (NIPS 2005)*, Vancouver, British Columbia, Canada, pp 155–162
- Cheng MM, Zhang GX, Mitra NJ, Huang X, Hu SM (2011) Global contrast based salient region detection. In: *2011 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, pp 409–416
- Garcia-Diaz A, Fdez-Vidal XR, Pardo XM, Doslil R (2012) Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image Vis Comput* 30(1):51–64
- Ghosh K, Sarkar S, Bhaumik K (2006) A possible explanation of the low-level brightness–contrast illusions in the light of an extended classical receptive field model of retinal ganglion cells. *Biol Cybern* 94(2):89–96
- Gu Y, Liljenström H (2007) A neural network model of attention-modulated neurodynamics. *Cogn Neurodyn* 1(4):275–285
- Guo C, Ma Q, Zhang L (2008) Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, 2008. IEEE, pp 1–8
- Harel J, Koch C, Perona P (2007) Graph-based visual saliency. In: *Advances in neural information processing systems*, vol 19. MIT Press, Cambridge, pp 545–552
- Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. In: *IEEE conference on computer vision and pattern recognition (CVPR)*, 2007. IEEE, pp 1–8
- Itti L (2004) Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Trans Image Process* 13(10):1304–1318
- Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Pattern Anal Mach Intell* 20(11):1254–1259
- Judd T, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. In: *2009 IEEE 12th international conference on computer vision*. IEEE, pp 2106–2113
- Koch C, Ullman S (1987) Shifts in selective visual attention: towards the underlying neural circuitry. In: *Vaina LM (ed) Matters of intelligence*. Springer, Netherlands, pp 115–141

- Kootstra G, Nederveen A, De Boer B (2008) Paying attention to symmetry. In: Proceedings of the British machine vision conference (BMVC2008). The British Machine Vision Association and Society for Pattern Recognition, pp 1115–1125
- Le Meur O, Le Callet P, Barba D (2007) Predicting visual fixations on video based on low-level visual features. *Vis Res* 47(19):2483–2498
- Li C, Zhou Y, Pei X, Qiu F, Tang C, Xu X (1992) Extensive disinhibitory region beyond the classical receptive field of cat retinal ganglion cells. *Vis Res* 32(2):219–228
- Li J, Levine MD, An X, Xu X, He H (2013) Visual saliency based on scale-space analysis in the frequency domain. *IEEE Trans Pattern Anal Mach Intell* 35(4):996–1010
- Liu T, Yuan Z, Sun J, Wang J, Zheng N, Tang X, Shum HY (2011) Learning to detect a salient object. *IEEE Trans Pattern Anal Mach Intell* 33(2):353–367
- Mahadevan V, Vasconcelos N (2009) Saliency-based discriminant tracking. In: IEEE conference on computer vision and pattern recognition, 2009. CVPR 2009. IEEE, pp 1007–1013
- Mishra A, Aloimonos Y, Fah CL (2009) Active segmentation with fixation. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp. 468–475
- Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int J Comput Vis* 42(3):145–175
- Perazzi F, Krahenbuhl P, Pritch Y, Hornung A (2012) Saliency filters: contrast based filtering for salient region detection. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, pp 733–740
- Rodieck RW, Stone J (1965) Analysis of receptive fields of cat retinal ganglion cells. *J Neurophysiol* 28(5):833–849
- Shi X, Bruce ND, Tsotsos JK (2011) Fast, recurrent, attentional modulation improves saliency representation and scene recognition. In: 2011 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 1–8
- Tatler BW, Baddeley RJ, Gilchrist ID (2005) Visual correlates of fixation selection: effects of scale and time. *Vis Res* 45(5):643–659
- Treisman AM, Gelade G (1980) A feature-integration theory of attention. *Cogn Psychol* 12(1):97–136
- Walther D, Koch C (2006) Modeling attention to salient proto-objects. *Neural Networks* 19(9):1395–1407
- Wright MJ (1972) Functional organization of the periphery effect in retinal ganglion cells. *Vis Res* 12(11):1857-IN8
- Yang CW, Chung PC, Chang CI (1996) Hierarchical fast two-dimensional entropic thresholding algorithm using a histogram pyramid. *Opt Eng* 35(11):3227–3241
- Yu Y, Wang B, Zhang L (2011) Bottom-up attention: pulsed PCA transform and pulsed cosine transform. *Cogn Neurodyn* 5(4):321–332
- Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) SUN: a Bayesian framework for saliency using natural statistics. *J Vis* 8(7):32
- Zou W, Kpalma K, Liu Z, Ronsin J (2013) Segmentation driven low-rank matrix recovery for saliency detection. In: 24th British machine vision conference (BMVC), pp 1–13