Cogn Neurodyn (2010) 4:189–198
DOI 10.1007/s11571-010-9122-0

ORIGINAL RESEARCH

Visual saliency: a biologically plausible contourlet-like frequency domain approach

Peng Bian · Liming Zhang

Received: 28 February 2010 / Revised: 17 June 2010 / Accepted: 17 June 2010 / Published online: 27 June 2010

Abstract In this paper we propose a fast frequency domain saliency detection method that is also biologically plausible, referred to as frequency domain divisive normalization (FDN). We show that the initial feature extraction stage, common to all spatial domain approaches, can be simplified to a Fourier transform with a contourlet-like grouping of coefficients, and saliency detection can be achieved in frequency domain. Specifically, we show that divisive normalization, a model of cortical surround inhibition, can be conducted in frequency domain. Since Fourier coefficients are global in space, we extend to this model by conducting piecewise FDN (PFDN) using overlapping local patches to provide better biological plausibility. Not only do FDN and PFDN outperform current state-of-the-art methods in eye fixation prediction, they are also faster. Speed and simplicity are advantages of our frequency domain approach, and its biological plausibility is the main contribution of our paper.

Keywords Visual saliency · Attention selection · Saliency map · Divisive normalization

Introduction

Visual saliency refers to the perceptual quality that makes an object or location stand out or pop out relative to its neighbors, and thus attracting our attention. There has been great interest in computational models of saliency detection which are biologically plausible and consistent with physiological and psychophysical data. Just as attention modulation rapidly locates general areas of interest in our visual pathway, computational models can be used to detect perceptually salient locations given a complex natural scene, which is useful for object recognition (Rutishauser et al. 2004), adaptive image compression (Guo and Zhang 2010), image quality assessment (Lu et al. 2005), and many more. In this paper we consider only bottom-up saliency, which results from feed-forward connections in our visual pathway and is independent of memory and task-driven modulation.

One of the most influential computational models of bottom-up saliency detection was proposed by Itti et al. (1998), which is inspired by the neural architecture of our early visual system. However, the model suffers from over-parameterization and ad-hoc design choices. For a fair comparison, we use the saliency toolbox (STB), Walther's implementation of Itti's method (Walther and Koch 2006), which along with our method and other models used in this paper, are all coded in Matlab.

Following STB, some later models proposed are not explicitly biologically based, but adhere to the hypothesis that our sensory system develops in response to the statistical properties of the signals to which they are exposed (Barlow 1961). Bruce and Tsotsos (2006) proposed an attention model based on information maximization (AIM) which uses the self-information criteria to define saliency. Harel et al. (2007) proposed a graph based approach to visual saliency (GBVS) using dissimilarity as a measure of saliency. Gao et al. (2008) proposed a discriminant center-surround model (DISC) where saliency is computed by the mutual information between center and surround. While these methods show better consistency than STB with physiological and psychophysical data, they are more

P. Bian · L. Zhang (✉)
Department of Electronic Engineering, Fudan University,
Shanghai 200433, China
e-mail: lmzhang@fudan.edu.cn

P. Bian
e-mail: pengfish@gmail.com

computationally expensive and incapable of real-time application.

Frequency domain approaches have recently gained popularity due to their fast computational speed and good consistency with psychophysics, but they are not at all biologically based. These include spectral residual (SR), proposed by Hou and Zhang (2007), and phase spectrum of quaternion Fourier transform (PQFT), proposed by Guo et al. (2008). No justification is given as to why these frequency based methods can outperform biologically based spatial models.

In this paper we propose a biologically plausible frequency domain saliency detection method called frequency domain divisive normalization (FDN), which has the topology of a biologically based spatial domain model, but is conducted in frequency domain. This reduces computational complexity because unlike all spatial domain methods, we do not need to decompose the input image into numerous feature maps separated in orientation and scale, and then compute saliency at every spatial location of every feature map. Such an operation may be quick for the massively parallel connections of our visual pathway, but is slow for computer processors. In frequency domain, each feature map is represented as a sub-band in the spectrum and saliency is computed across all scales and orientations.

Saliency is always defined as some measure of difference between center and surround. While the spatial extent of this surround is limited in our visual cortex, Fourier coefficients are global in space, and so FDN is constrained by a global surround. In order to overcome this constraint, we propose a piecewise FDN (PFDN) using Laplacian pyramids and overlapping local patches. While PFDN is slower than FDN, it is more biologically plausible and performs better in eye fixation prediction.

Derivation of FDN

Topology of spatial domain approach

Examining the topology of current state-of-the-art spatial domain saliency detection models, we can see most approaches have steps which can be grouped into three common stages: feature extraction, center-surround rectification, and recombination.

The input image $\boldsymbol{I}$ is first decomposed into many feature maps separated in scale and orientation to mimic scale and orientation tuning of V1 simple cells. The tuning function can be modeled with oriented Gabor filters but ICA basis functions have also been used for this feature extraction stage. In this paper we consider the use of a contourlet transform (Do and Vetterli 2005) for the complete and invertible decomposition of the input image on many scales and orientations, which is given using simplified notation as

$$r_i[n] = C_i^n\{I\}, \tag{1}$$

where C_i denotes a contourlet transform of the i-th sub-band, and $r_i[n]$ is the n-th contourlet coefficient of the i-th sub-band. We will also refer to $\boldsymbol{r}_i$ as a feature map.

The feature maps are then nonlinearly rectified through some type of center-surround operation to emphasize salient locations. Li and Dayan (2006) hypothesized that the output firing rate of V1 neurons represents the salience of the input. Her spiking neuron model, which models lateral interactions between V1 simple cells for saliency detection, is capable of saliency detection. Inspired by her work, we use a simpler lateral surround inhibition model to conduct center-surround rectification, divisive normalization.

Simoncelli and Schwartz (1999) used divisive normalization to model surround inhibition. The results were consistent with recordings from macaque V1 simple cells from Cavanaugh et al. (1997). Itti et al. (1999) used divisive normalization to model psychophysical data for orientation selectivity. In this paper the divisive normalization equation is given as

$$\hat{r}_i[n] = \frac{r_i[n]}{\sqrt{\sum_{n'} w[n'] |r_i[n']|^2 + \sigma^2}}, \tag{2}$$

where $\hat{r}_i[n]$ are the divisive normalized coefficients, $r_i[n']$ are coefficients in the spatial surround of $r_i[n]$, $w[n']$ are the spatial weights and σ is a constant. We can see that each contourlet coefficient is divided by a function of its surrounding coefficients in a way that models the tuning curves of surround inhibition. Note that in (Simoncelli and Schwartz 1999) the right side of Eq. 2 is squared, but in order to keep the coefficients invertible, we simply square the normalized coefficients after inverse contourlet transform.

The recombination stage is the spatial summation of all feature maps, which can be obtained by

$$S = \sum_i \left| C_i^{-1}\{\hat{r}_i\} \right|^2, \tag{3}$$

where C_i^{-1} denotes the inverse contourlet transform of the i-th sub-band and $\boldsymbol{S}$ is the corresponding saliency map. Since the feature maps only serve to be recombined after center-surround rectification, if we can conduct the rectification stage without having to decompose the image into so many feature maps, we can greatly reduce computational complexity.

Divisive normalization in the frequency domain

To see the connection in frequency domain, let us consider the example in Fig. 1. A simple input image $\boldsymbol{I}$ of a slanted-horizontal bar surrounded by slanted-vertical bars is given in Fig. 1a. Its Fourier transform is given as

$$R[k] = F_k\{I\}, \quad (4)$$

where F denotes a Fourier transform and $R[k]$ is the k-th Fourier coefficient. The amplitude spectrum of $\boldsymbol{R}$ is plotted in Fig. 1b. We also conduct a contourlet transform of the image using the frequency decomposition scheme given in Fig. 1e to produce feature maps $\boldsymbol{r}_i$ given in Fig. 1f. The decomposition scheme separates the image into feature maps in 4 scales with 16, 8, 4 and 0 orientations from the highest scale to the lowest. This is biologically plausible because cells with larger receptive fields (tuned to lower scales) are less orientation selective. Each feature map in Fig. 1f represents the corresponding frequency band in Fig. 1e.

Because Fourier coefficients are global in space, we first assume the surround size in Eq. 2 to be global with uniform weights. This means the denominator of Eq. 2 is constant for a given feature map. According to Parseval's theorem, the squared sum of all coefficients for a feature map in spatial domain should be proportional to the energy of its corresponding frequency band, which is given as

$$\sum\nolimits_n |r_i[n]|^2 = \frac{1}{N}\sum\nolimits_{k\in i} |R[k]|^2, \quad (5)$$

where N is the number of pixels. So we group the Fourier coefficients into frequency sub-bands equivalent to the contourlet transform decomposition scheme given in Fig. 1e, and rewrite the denominator of Eq. 2 in frequency domain as

$$E_i = \sqrt{w\sum\nolimits_{k\in i} |R[k]|^2 \Big/ N + \sigma^2}, \quad (6)$$

which we refer to as the normalization term. After calculating E_i for every sub-band, given in Fig. 1c, and the same divisive normalization can then be achieved in Frequency domain by

$$\hat{R}[k]|_{k\in i} = \frac{R[k]}{E_i}|_{k\in i}, \ \hat{R} = \bigcup\nolimits_i \left(\hat{R}[k]|_{k\in i}\right), \quad (7)$$

where $\hat{R}$ are the divisive normalized Fourier coefficients, given in Fig. 1d. We can see each Fourier coefficient is divided by a function of its sub-band energy. This suppresses frequency bands with high energy concentration, and through this suppressive effect we can achieve saliency detection.

After divisive normalization, the recombination stage is computed similar to Eq. 3, given as

$$S = W\left|F^{-1}\{\hat{R}\}\right|^2, \quad (8)$$

where F^{-1} denotes the inverse Fourier transform, $\boldsymbol{W}$ is a windowing function to remove edge effects, and $\boldsymbol{S}$ is the

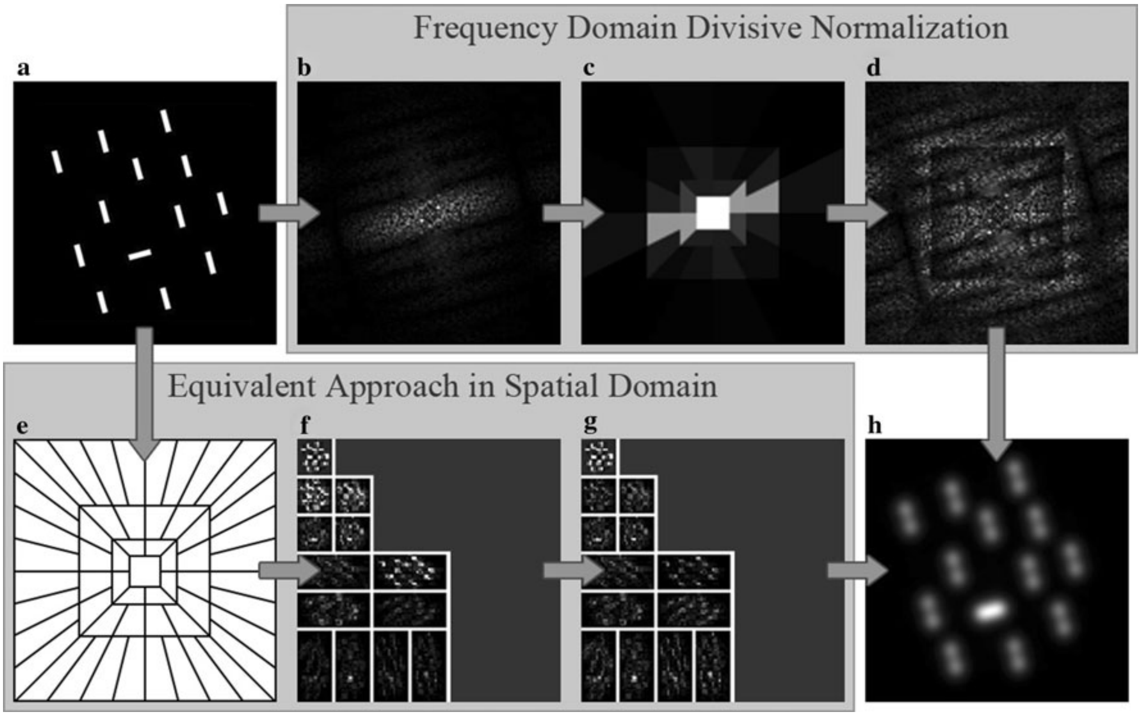
Frequency Domain Divisive Normalization
a
b
c
d
Equivalent Approach in Spatial Domain
e
f
g
h

Fig. 1 Method flow for FDN and its spatial equivalent. **a** The input image. **b** Amplitude spectrum of Fourier transform of Input. **c** Normalization terms for each Fourier coefficient in frequency domain. **d** Divisive normalized amplitude spectrum. **e** Contourlet decomposition scheme in frequency domain. **f** Contourlet coefficients. **g** Divisive normalized contourlet coefficients. **h** Resulting saliency map. Note that in **b**, **c**, **d** and **e** the frequency origin is located in the center of the image, and in **f** and **g** the feature maps have been down-sampled and some less significant feature maps have been left out to save space

corresponding spatial domain saliency map, given in Fig. 1h. Normally $\boldsymbol{S}$ is post-processed by convolution with a Gaussian function $\boldsymbol{G}$ for smoothing. The slanted-horizontal bar is salient which is perceptually consistent. This method is referred to as frequency domain divisive normalization (FDN). If we divide the spatial domain feature maps in Fig. 1f by their corresponding E_i to obtain Fig. 1g, the corresponding recombination from Eq. 3 can give us the same saliency map, but the computational cost is much higher. By conducting divisive normalization in frequency domain, we eliminate the need to decompose the image into feature maps.

To summarize, feature extraction essentially separates the image into many frequency sub-bands and center-surround rectification suppresses sub-bands with high energy concentration, which we compute using divisive normalization. The inverse Fourier transform is then able to give us the spatial domain saliency map.

The FDN algorithm

In the human visual pathway, the color space of natural images is decomposed into well decorrelated channels. The RGB color space is highly correlated, but an LAB color space transformation results in well decorrelated color channels for natural color images. In addition, the transformation is perceptually uniform, and it produces three biologically plausible channels: a luminance channel, a red-green opponent channel and a blue-yellow opponent channel.

The complete FDN algorithm from input image to final saliency map is given as follows.

1. Perform an LAB color space transformation
2. Resize the image to a suitable scale
3. Perform a Fourier transform for each color channel using Eq. 4
4. Group the Fourier coefficients using the scheme given in Fig. 1e
5. Calculate normalization terms using Eq. 6, with constants w and σ from Eq. 6 set to 1 for simplicity
6. Following Eq. 7, obtain the divisive normalized Fourier coefficients
7. Obtain the saliency maps of each color channel using Eq. 8
8. Take the spatial maximum across all color channels to obtain the final saliency map
9. Smooth saliency map with Gaussian filter $\boldsymbol{G}$

Psychophysical consistency

In this section we show the consistency of FDN with well known properties of psychophysics: feature pop-out, search asymmetry and conjunction search. For each psychophysical pattern, we calculate the saliency maps for FDN, PQFT, and STB. In all examples, the target is located in the center of the pattern. First we give a simple example in Fig. 2 of an orientation pop-out and color pop-out.

The pop-out locations for all three methods are consistent with perception, but for STB the disparity between saliency values of target and distractors are not as clear.

During visual search, there is asymmetry in search difficulty when switching the target and distractor which differ in the inclusion or exclusion of a feature. In general, targets with an added feature are easier to detect than targets with a missing feature. Figure 3 gives an example of such search asymmetry.

The target in the top row is a plus sign located at the center of the pattern, which is surrounded by vertical bar distractors. In this case the target is salient. In the bottom row the target and distractors become switched, and this time the target does not pop out. The pop-out locations for FDN and STB are consistent with perception. For PQFT, even when there is no pop-out, there are large disparities between saliency values amongst distractors.

A pop-out occurs when the target contains a feature unique from its distractors, but when the target contains no single unique feature but rather a unique conjunction of two or more features, the target does not pop out, increasing the difficulty of visual search. Figure 4 gives an example of such conjunction search.

The target in the top row consists of a horizontal bar, a feature which is not present in the surrounding distractors. Accordingly, the target should be salient. In the bottom row, the horizontal bar is no longer unique to the target. Rather, the target is only unique in that the conjunction of the two oriented bars that form the target is unique, and the target should not pop out. The pop-out locations for FDN and STB are consistent with perception. For PQFT, even when there is no pop-out, there are large disparities between saliency values amongst distractors.

Analyzing the saliency maps produced by these three methods, it can be seen that for the most part all methods are consistent with these psychophysical patterns. FDN and STB show a closer resemblance while PQFT can sometimes find salient locations in patterns where there is no pop-out.

Extending to piecewise FDN

Overcoming the global surround constraint

There is much evidence suggesting that the spatial extent of our surround inhibition mechanism is limited (Nothdurft 2000), which means the surround size is limited. So while the assumption of a global surround may be adequate for

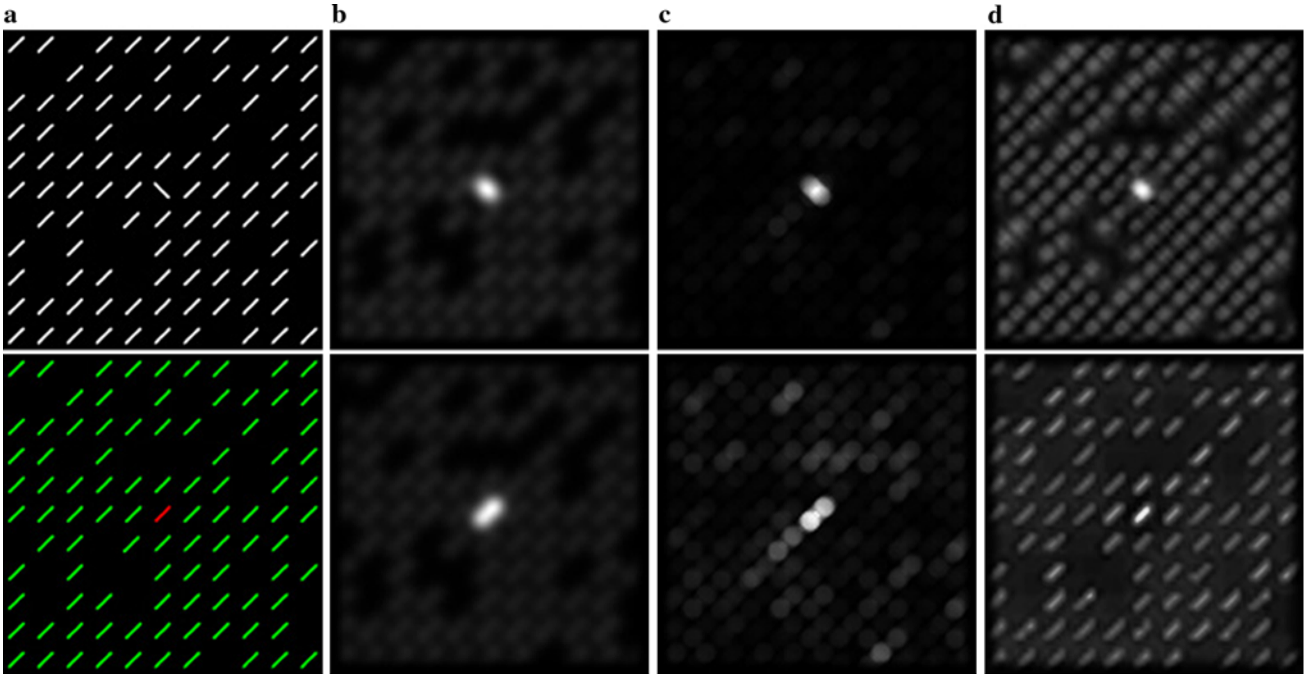

Fig. 2 **a** Psychophysical pattern. **b** FDN saliency map. **c** PQFT saliency map. **d** STB saliency map

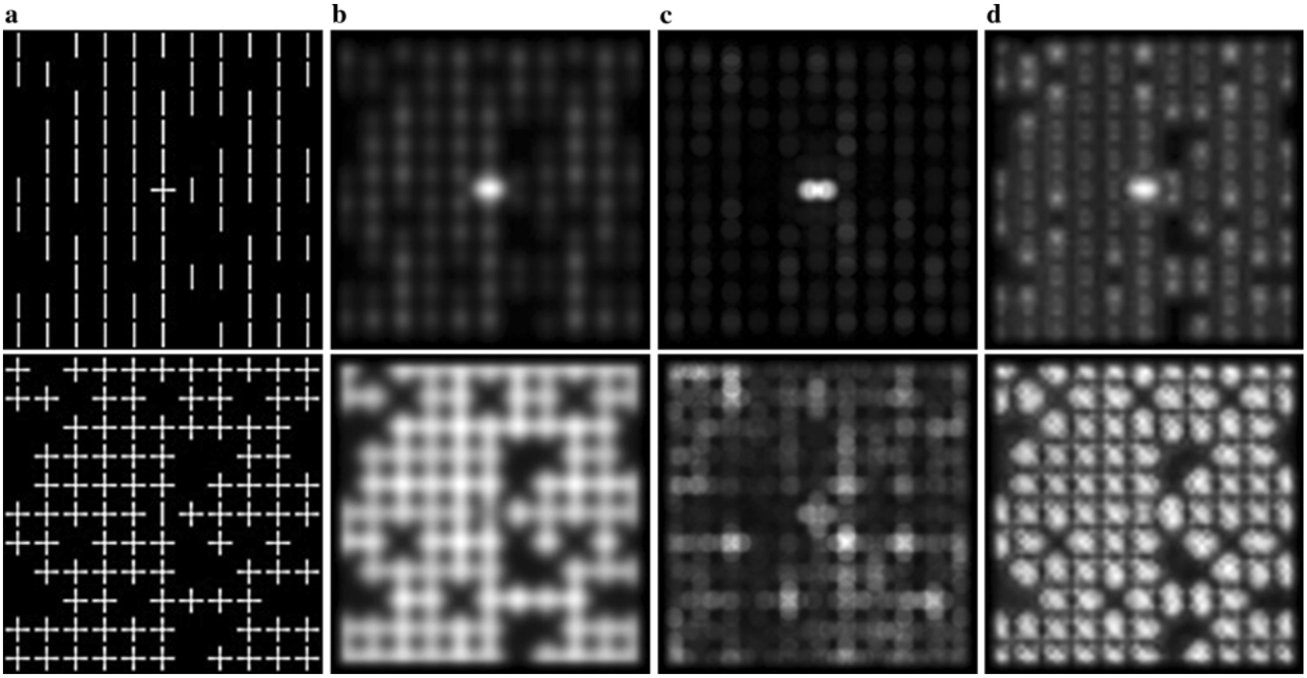

Fig. 3 **a** Psychophysical patterns. **b** FDN saliency maps. **c** PQFT saliency maps. **d** STB saliency maps

simple psychophysical patterns, we need to factor in the limited extent of this surround size for complex visual stimuli found in natural color images. From the example given in Fig. 5, we can see the limitation of assuming a global surround. Sometimes a feature map has high energy globally, but has low energy relative to a local area. On the bottom right of Fig. 5a, there is a large empty area comprised of asphalt with a small rubber object in the middle. The rubber object is perceptually salient because the area surrounding the rubber object has very low energy across all feature maps. Using FDN, we are not able to locate the salient rubber object, illustrated by Fig. 5b.

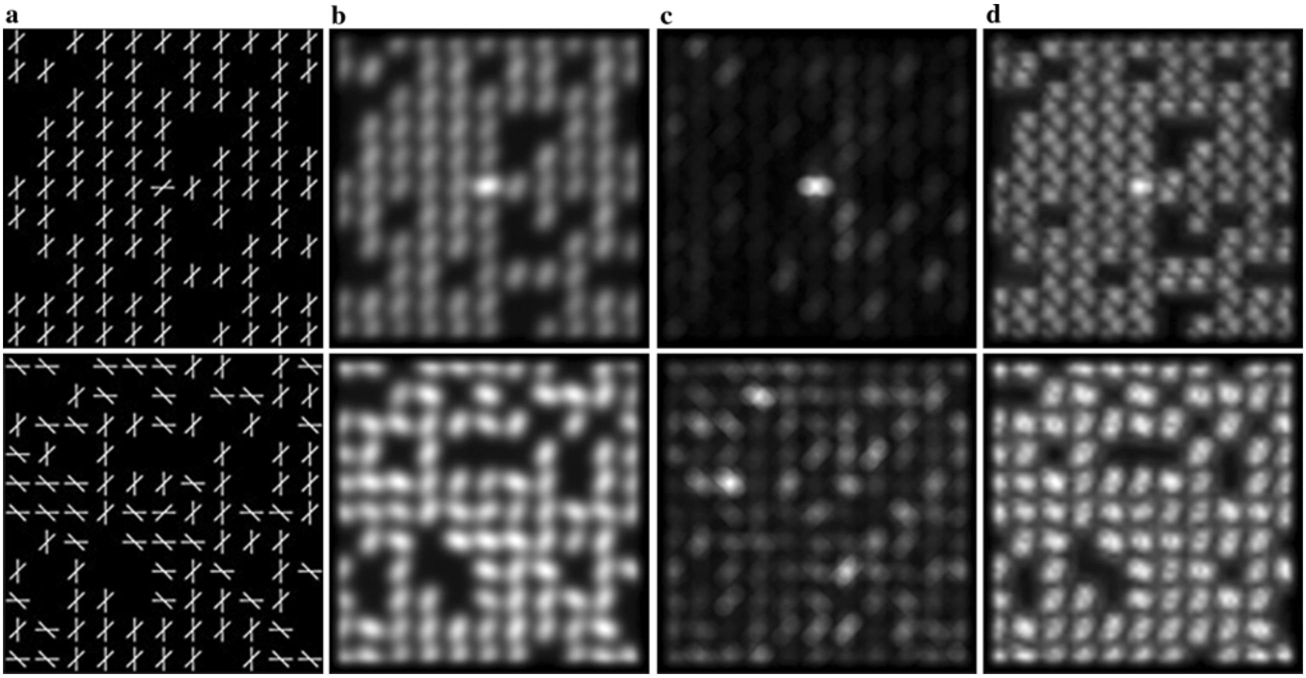

Fig. 4 **a** Psychophysical patterns. **b** FDN saliency maps. **c** PQFT saliency maps. **d** STB saliency maps

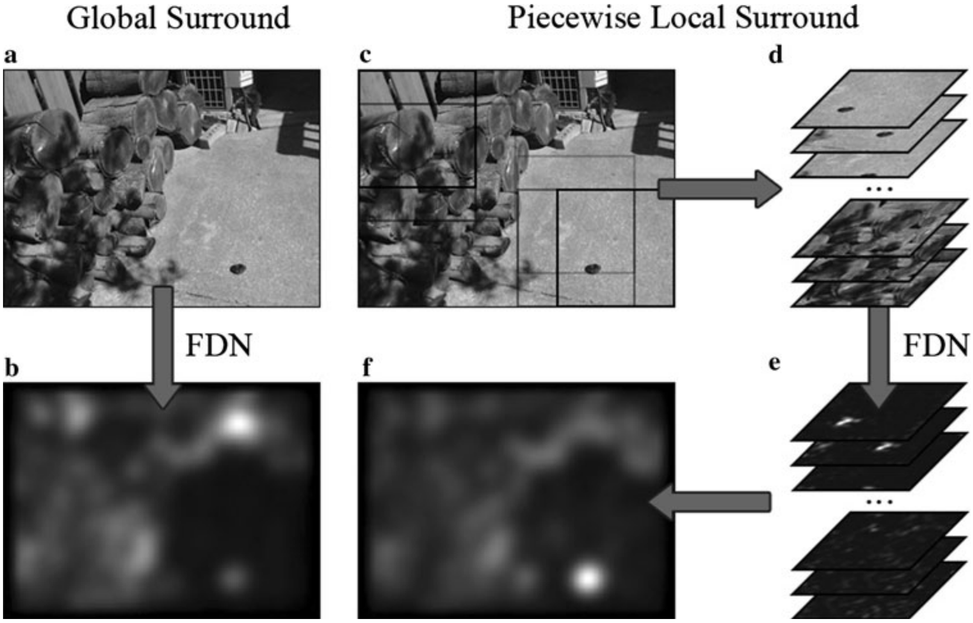

Fig. 5 Difference between FDN and PFDN: **a** original image and **b** saliency map generated by FDN, **c** original image, **d** separation into local patches, **e** normalized patches through FDN, and **f** recombined saliency map generated by PFDN

In order to relax the global surround constraint, we separate the input image into overlapping local patches, conduct FDN on every patch, and recombine the divisive normalized patches to obtain the final saliency map. For recombination, we take the maximum value, as argued by Li and Dayan (2006), at each pixel location of the corresponding normalized patches instead of a spatial summation used by most models. From Fig. 5f, we can see that by using this piecewise approach, we are able to locate the salient rubber object.

In addition, data from (Cavanaugh et al. 2002) suggest that the surround size depends on the receptive field size of V1 simple cells. In order to be consistent with these findings, we first decompose the image into many scales using the Laplacian pyramid. The patch sizes remain constant for all pyramid scales and are also chosen to be consistent with biological data.

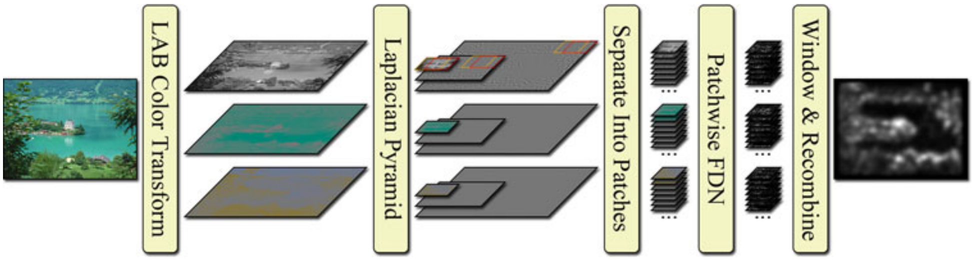

Fig. 6 The PFDN algorithm from original image (*left*) to saliency map (*right*)

The PFDN algorithm

The complete PFDN algorithm flow from input image to final saliency map is illustrated by Fig. 6 and given as follows.

1. Perform an LAB color space transformation
2. Decompose the image into a number of scales using a Laplacian pyramid.
3. For each scale and every color channel, separate into overlapping local patches with a shift between patches
4. Perform a Fourier transform for each patch using Eq. 4
5. Group the Fourier coefficients using the scheme given in Fig. 1e.
6. Calculate normalization terms using Eq. 6, with constants w and σ from Eq. 6 set to 1 for simplicity
7. Following Eq. 7, obtain the divisive normalized Fourier coefficients
8. Obtain the saliency maps of each patch using Eq. 8
9. For each scale and color channel, recombine the saliency maps of all patches by taking the maximum value at each pixel location
10. Resize all scales to be equal in size and take the spatial maximum across all scales and color channels to obtain the final saliency map
11. Smooth saliency map with Gaussian filter $\boldsymbol{G}$

Experimental validation for natural images

Implementation details

For this experiment we use the data set of 120 color images from an urban environment and corresponding eye fixations from 20 subjects provided by (Bruce and Tsotsos 2006). For FDN we resize the image to a width of 64 px. The aspect ratio is kept, so for this experiment the pixel dimension of all resized images is 64 × 48 px. The scales are chosen following the heuristics set by previous frequency domain papers including PQFT, which uses the same scale for this data set. For PFDN the Laplacian pyramid consists of two scales: 32 × 24 px, 64 × 48 px. The patch size is 24 × 24 px with an 8 px shift size between patches. The patch size is chosen from a careful consideration of viewing environment and biological data.

Eye fixation prediction metric

In this section we quantify the consistency of FDN and PFDN with fixation locations for human subjects during free viewing. Tatler et al. (2005) proposed the use of a receiver operating characteristic (ROC) curve by considering the saliency map as a binary classifier for fixation versus non-fixation points. We compute the ROC curve for FDN, PFDN and 5 other state-of-the-art models: PQFT, AIM, DISC, GBVS and STB. We then plot the ROC curve for the 7 methods using fixation data as ground truth and calculate the area under the curve (AUC) for each method. The results are given in Table 1. Note higher AUC denotes better consistency with eye fixation data.

AIM and DISC parameters for this image set are optimized by their respective authors to produce the highest AUC, and in order to maintain objectivity, we tune GBVS and STB parameters to obtain the highest AUC possible. Consequently, the results for GBVS and STB are higher than given from other papers, which use default parameters. Results show PFDN is the best performer, and is well above all other methods. FDN is the second with better performance than PQFT, although all frequency domain methods outperform spatial domain methods. The spatial domain methods have very similar performance, and the AUC difference between the best and worst performing spatial domain methods is less than the AUC difference between PFDN and FDN.

We also give a qualitative comparison of performance for 8 select images in Fig. 7. A fixation density map, generated for each image by convolution of the fixation map for all subjects with a Gaussian filter (Bruce and Tsotsos 2006), serves as ground truth.

Analyzing the qualitative results, we can see that PFDN and FDN show more resemblance to the ground truth. In

Table 1 Eye fixation predication AUC for state-of-the-art saliency detection methods

Method	PFDN	FDN	PQFT	AIM	DISC	GBVS	STB
AUC	0.8217	0.7981	0.7822	0.7787	0.7683	0.7602	0.7575

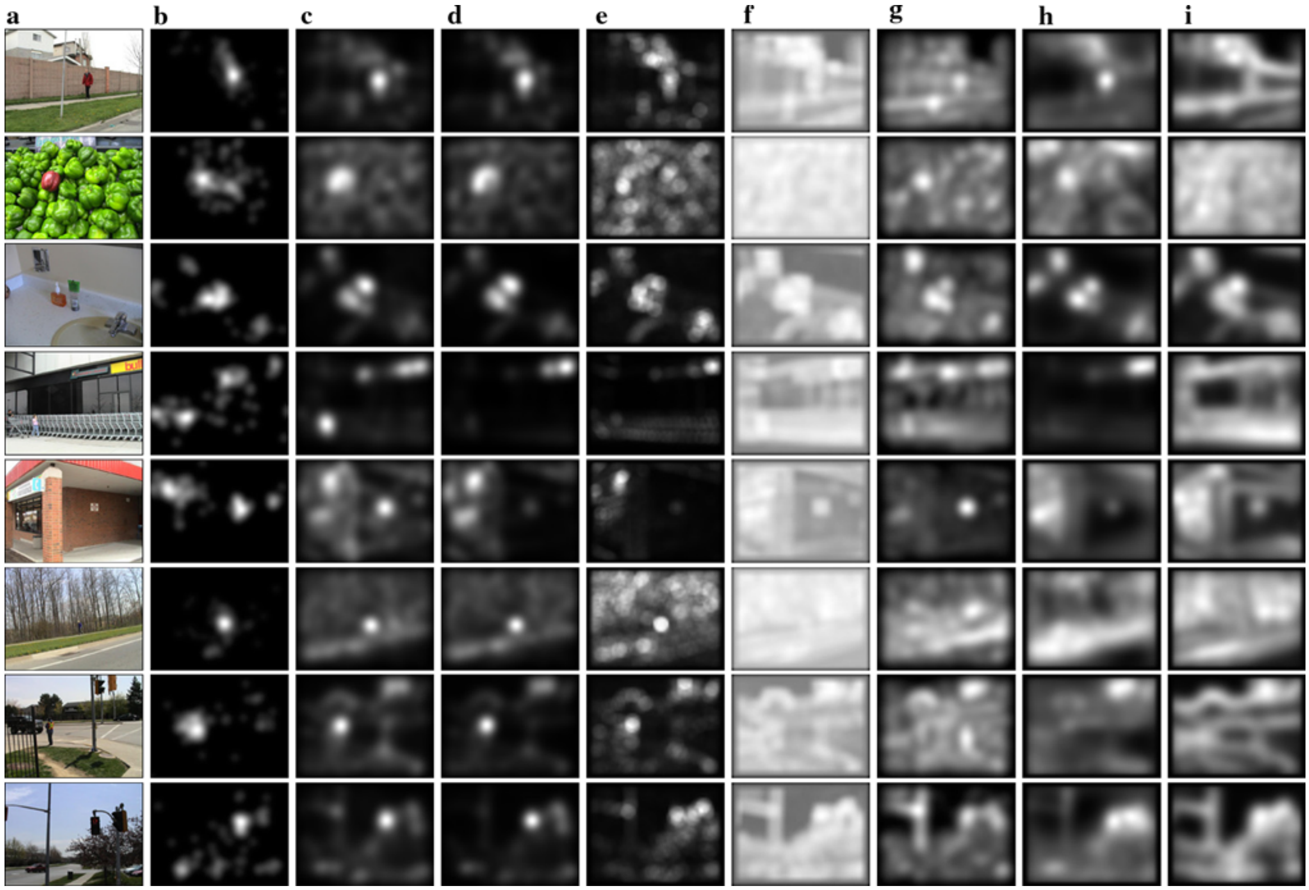

Fig. 7 **a** Natural images from (Bruce and Tsotsos 2006). **b** Corresponding fixation density maps. **c** PFDN saliency maps. **d** FDN saliency maps. **e** PQFT saliency maps. **f** AIM saliency maps. **g** DISC saliency maps. **h** GBVS saliency maps. **i** STB saliency maps

addition, high contrast straight edges are suppressed to a much greater extent using frequency domain approaches (rows 1 and 8 of Fig. 7). In our approach, we assume many more directional filters than most spatial domain approaches. This provides greater suppression for straight lines, which are concentrated around one orientation. Good performance with respect to color salience is also observed with FDN compared to the other models. In addition, PFDN is able to locate some local surround pop-outs FDN fails to detect (rows 4 and 5 of Fig. 7).

Computational cost

We calculate the time cost per image for each method averaged over the 120 images in the eye fixation prediction experiment. All methods are computed using Matlab 2009a using an Intel Core 2 Duo T7200 clocked at 2.00 GHz with 2 GB of memory. The results are given in Table 2.

FDN and PQFT are extremely fast and capable of real-time saliency detection at a very high frame rate. PFDN is slower than the other two frequency domain methods due to its inclusion of scales and overlapping patches. It is, however, still faster much than STB and the other spatial domain methods. Note that for Matlab, the algorithm for PFDN contains many nested for loops, which has a very inefficient runtime in Matlab. A comprehensive analysis of conducting FDN and PFDN at different scales, including the tradeoff between performance and cost, are discussed in the next section.

Table 2 Average time–cost per image for state-of-the-art saliency detection methods

Method	PFDN	FDN	PQFT	AIM	DISC	GBVS	STB
Time (s)	0.0634	0.0093	0.0152	9.6483	3.6186	4.9333	0.6624

Discussion

The effect of scale

Natural image statistics are scale invariant due to its characteristic 1/*f* amplitude spectrum (Ruderman 1994),

Table 3 AUC and time cost for PFDN and FDN using different scales for the eye fixation prediction experiment

Method	PFDN			FDN		
Scale	128 × 96	64 × 48	32 × 24	128 × 96	64 × 48	32 × 24
AUC	0.8214	0.8217	0.8109	0.7741	0.7981	0.7969
Time (s)	0.1698	0.0634	0.0382	0.0339	0.0093	0.0041

which is to say statistics of natural images do not tend to change with respect to rescaling. For modeling attention selection in the frequency domain, however, rescaling the image and processing it at different resolutions can lead to a large change in the resulting saliency map. It has been shown that in the human visual pathway, attention selection operates from coarse-to-fine (Navon 1997). Since bottom-up processing occurs quickly from the onset of visual stimuli, a coarse scale for modeling bottom-up attention is plausible. Table 3 shows the results of conducting FDN and PFDN at different scales for the data set used in the previous experiment.

For both FDN and PFDN, the performance is optimal at 64 × 48. So even though saliency map calculation at higher scales is more costly, there is no improvement in performance. This is consistent with the findings from (Hou and Zhang 2007) and (Guo et al. 2008).

Relation to whitening

There has been a large amount of literature published regarding the topic of the goal of early sensory coding. Decorrelation (Atick and Redlich 1992) and whitening (Field 1987) are some of the goals hypothesized for early sensory coding. Li and Dayan (2006) summarized many of the previous papers with the proposed efficient coding principle, where retinal cells and V1 simple cells decompose the visual stimulus into decorrelated channels, and gain control distributes the amount of power given to each channel in a way which maximizes the amount of information gained. In a noiseless environment, optimal gain control is equal to whitening, and energy is distributed equally amongst all channels.

This efficient coding principle can be observed from our equation for divisive normalization, which acts as gain control for the decorrelated feature maps. Revisiting the equation for divisive normalization, we take the squared sum of both sides of Eq. 2 with the assumption of global surround and uniform weights as mentioned previously

$$|\hat{r}_i[n]|^2 = \frac{|r_i[n]|^2}{w\sum_{n'}|r_i[n']|^2 + \sigma^2}. \tag{9}$$

If $w = 0$ and $\sigma = 1$, then no normalization occurs. If $w = 1$ and $\sigma = 0$, then maximal normalization occurs, and sub-band energy for all normalized feature maps $\hat{r}_i$ is equalized. If we conduct FDN using the constants in the latter case and consider a decomposition scheme where all frequency sub-bands are 1 px in size, the result will be a flattened amplitude spectrum with the phase spectrum unchanged, producing results equivalent to PQFT. Such decomposition, however, is not biologically plausible. This explains why FDN is able to achieve better results than PQFT.

Conclusion

In this paper we achieved two goals. First and foremost, we proposed a biologically plausible frequency domain saliency detection method. The motivation is to combine the speed of frequency domain methods with the biologically inspired topology of spatial domain models. We accomplished this by showing the frequency domain equivalent of each biologically based spatial domain step. This allowed us to achieve our second goal, which is to provide a link between spatial and frequency domain models and a way to consider saliency detection in the frequency domain. Experimental results show the proposed frequency domain methods FDN and PFDN are more effective at predicting human eye fixation.

The parameters for this paper are mostly chosen out of heuristics and simplicity, and more research into biologically sound parameters is needed to further biological plausibility. For the patch size, this means taking into account the size of the monitor screen as well as the distance between the screen and the viewer. There is potential to further improve the performance if we choose parameters consistent with biological data.

Acknowledgments This work is supported by the National Science Foundation of China (NSF60571052) and by Key Subject Foundation of Shanghai (B112).

References

Atick JJ, Redlich AN (1992) What does the retina know about natural scenes? Neural Comput 4:196–210

Barlow HB (1961) The coding of sensory messages. In: Thorpe WH, Zangwill OL (eds) Current problems in animal behavior. Cambridge University Press, Cambridge, pp 331–360

Bruce ND, Tsotsos JK (2006) Saliency based on information maximization. Adv NIPS 18:155–162

Cavanaugh JR, Bair W, Movshon JA (1997) Orientation-selective setting of contrast gain by the surrounds of macaque striate cortex neurons. Neurosci Abstr 23:227.2

Cavanaugh JR, Bair W, Movshon JA (2002) Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. J Neurosci 88:2547–2556

Do MN, Vetterli M (2005) The contourlet transform: an efficient directional multiresolution image representation. Off J Eur Union Inf Not 49:2091–2106

Field DJ (1987) Relations between the statistics of natural images and the response profiles of cortical cells. J Opt Soc Am 4:2379–2394

Gao D, Mahadevan V, Vasconcelos N (2008) The discriminant center-surround hypothesis for bottom-up saliency. Adv NIPS 20:497–504

Guo CL, Zhang LM (2010) A novel multiresolution spatio temporal saliency detection model and its applications in image and video compression. IEEE Trans IP 19(1):185–198

Guo CL, Ma Q, Zhang LM. (2008). Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In: Proceedings of CVPR 2008

Harel J, Koch C, Perona P (2007) Graph-based visual saliency. Adv NIPS 19:545–552

Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. In: Proceedings of CVPR 2007

Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans PAMI 20(11): 1254–1259

Itti L, Braun J, Lee DK, Koch C (1999) Attentional modulation of human pattern discrimination psychophysics reproduced by a quantitative model. Neural Netw 19:143–1439

Li Z, Dayan P (2006) Pre-attentive visual selection. Neural Netw 19:143–1439

Lu ZK, Lin WS, Yang XK, Ong E, Yao S (2005) Modeling visual attention's modulatory after effects on visual sensitivity and quality evaluation. IEEE Trans IP 14(11):1928–1942

Navon D (1997) Forest before trees—precedence of global features in visual-perception. Cogn Psychol 9:353–383

Nothdurft HC (2000) Salience from feature contrast: variations with texture density. Vision Res 40(23):3181–3200

Ruderman D (1994) The statistics of natural images. Netw Comput Neural Syst 5(4):517–548

Rutishauser U, Walther D, Koch C, Perona P (2004) Is bottom-up attention useful for object recognition? In: Proceedings of CVPR 2004

Simoncelli EP, Schwartz O (1999) Modeling surround suppression in V1 neurons with a statistically-derived normalization model. Adv NIPS 11:153–159

Tatler BW, Baddeley RJ, Gilchrist ID (2005) Visual correlates of fixation selection: effects of scale and time. Vision Res 45:643–659

Walther D, Koch C (2006) Modeling attention to salient proto-objects. Neural Netw 19:1395–1407