Unsupervised Saliency Detection Based on 2D Gabor and Curvelets Transforms

Sheng-hua Zhong[1], Yan Liu[1], Ling Shao[2], Gangshan Wu[3]

[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, P. R. China

[2]Department of Electronic and Electrical Engineering, University of Sheffield, Sheffield, United Kingdom

[3]State Key Laboratory for Novel Software Technology, Nanjing University, P. R. China

ABSTRACT

Construction of saliency map in multimedia data is useful for applications in multimedia like object segmentation, quality assessment, and object recognition. In this paper, we propose a novel saliency map model called Gabor & Curvelets based Saliency Map (GCSMP) relying on 2D Gabor and Curvelet transforms. Compared with the traditional model based on DOG and wavelets, our model takes advantage of Garbor transforms's spatial localization and Curvelet transform's edge and directional information. We also discuss the influence of center bias and object detectors in our model. Empirical validations on standard dataset demonstrate the effectiveness of the proposed technique.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding –*Perceptual reasoning*

I.4.9 [**Computing Methodologies**]: Image Processing and Computer Vision—*Applications*

General Terms

Algorithm, Performance, Experimentations

Keywords

Saliency map, 2D Garbor, Curvelet transform.

1. INTRODUCTION

As a cognitive process of selectively concentrating on one aspect of the environment while ignoring other things, attention has been referred to as the allocation of processing resources [1]. In human vision system (HVS), to encode detailed visual information, eyes need to be moved so that this area is focused on the visual locations in which are interested [2]. As the most famous attention model, saliency map is proposed to measure of conspicuity and calculate the likelihood of a location to attract attention [3].

Owing to the models of image saliency provide predictions about which regions are likely to attract observers' attention [4], automatic detection of visually salient regions is useful in different multimedia applications. These applications include content aware resizing [5], quality assessment [6], segmentation [7], object detection and object recognition [8].

There are two different kinds of processing in attention, bottom-up and top-down processing. In existing works on visual saliency detection, most of them focus on the bottom-up processes of HVS. Typically, multiple low-level visual features such as intensity, color, orientation, texture and motion are extracted at multiple scales. After a feature map is computed, they are normalized and combined into a master saliency map that represents the saliency of each pixel [9]. Nearly all of existing bottom-up models are inspired by the theories from biology, psychology and neuropsychology [10]. Among them, the most famous one was proposed by Itti *et al.* [11]. They developed the center surround structure akin to on-type and off-type visual receptive field. In recent years, more proposed work simulated the multi-scale and multi-orientation function of primary visual cortex V1, Achanta *et al.* detected the saliency map with a Difference of Gaussians (DOG) model to describe the spatial properties of visual regions [12]. Gabor filters and Log-Gabor wavelets are utilized to explore the salient features such as spatial localization, spatial frequency characteristics in [13] and [10] respectively.

The DOG, wavelets and Gabor transforms are prevalent in saliency map construction in recent years; nevertheless, both of them have some inherent drawbacks. The DOG is a wavelet mother function of null total sum which approximates the Mexican Hat wavelet by subtracting a wide Gaussian from a narrow Gaussian. Compared with DOG, wavelets have the ability to capture the scale-space information in details. But wavelets are ill-suited for detecting or providing a compact representation of intermediate dimensional structures, for example, wavelets are very crude in representing directional features. The principal motivation to use Gabor transforms is biological relevance that the receptive field are oriented and have characteristic spatial frequencies. But due to the elimination of spectra overlap, 'holes' are created in the spectra plane of Garbor transforms, which causes loss of spectral information, especially the edge and fine directional information. As the latest multi-directional & multi-scale transform, Curvelet transforms have subtle capability to resolve directional feature than wavelet transform and improved ability to represent edges and other singularities along curves.

Due to the difficulty in refining the goal of attention in natural images [14], little work about saliency map construction simulating the top-down processing. In these work, the top-down processes often need supervised learning and lack the expansibility. In [15], Judd *et al.* collected eye tracking data and utilized the dataset to learn a model of saliency based on low, middle and high-level image features. In [6], Zhong *et al.* integrated visual features, center priority, and semantic meaning from tag information to learn a top-down & bottom-up saliency model based on the eye-tracking data. According to several

studies of subject's visual attention measure, human's attention is often biased toward the center of static image. Therefore, both of these techniques utilized the center bias as one important feature to simulate the top-down processing.

In this paper, we propose a novel unsupervised approach for visual saliency detection in natural images. The proposed method in this paper takes advantage of 2D Gabor's spatial localization and Curvelet transforms' edge and directional information. And in order to simulate the top-down processes in human visual system, we also consider the influence of center bias and object detection into our model. The rest of paper is organized as follows. Section 2 presents a novel saliency map construction method based on the Gabor and Curvelet transforms in detail. Experimental results are given in Section 3 and the paper is concluded in Section 4.

2. PROPOSED METHOD

The framework of our technique is illustrated in Fig. 1. In this framework, since visual neurons are often excited by one color and inhibited by opponent color, we choose preattentive features as the red/green (RG), blue/yellow (BY) color and intensity. And we utilize the 2D Gabor filters and Curvelet transforms to build the feature maps. These feature maps are then computed into activation maps within Gabor and Curvelet channels. They are integrated with the center bias and object detection to construct the saliency map. In the rest of this section, the Gabor & Curvelet transforms are first introduced. Then, we describe the details of the saliency detection by applying the 2D Gabor & Curvelet transforms and integrating the top-down information priority.

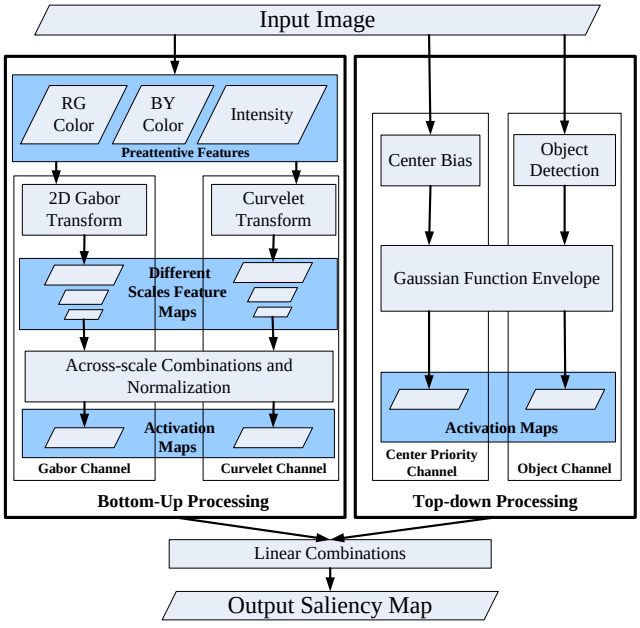


Figure 1. Framework of GCSMP construction.

2.1 Introduction to 2D Gabor Transforms

Gabor filter, named after Dennis Gabor, is a linear filter used in image processing. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave. The 2D Gabor function in space domain is defined as below:

$$g(x_1,x_2)=c_s(x_1,x_2)g_s(x_1,x_2) \quad (1)$$

where $c_s(x_1,x_2)$ is a complex sinusoid, known as the carrier, and $g_s(x_1,x_2)$ is a 2D Gaussian function, known as the envelope. $c_s(x_1,x_2)$ is denoted as Eq. (2).

$$c_s(x_1,x_2)=\exp[2\pi iF_g(x_1\cos\omega_g+x_2\sin\omega_g)] \quad (2)$$

where $F_g=\sqrt{u_g^2+v_g^2}$, $\omega_g=\tan^{-1}(v_g/u_g)$, i.e. $u_g=F_g\cos\omega_g$ and $v_g=F_g\sin\omega_g$. And (u_g,v_g) are the spatial frequencies of the sinusoid carrier in Cartesian coordinates.

The Gaussian envelope $g_s(x_1,x_2)$ is given as follows:

$$g_s(x_1,x_2)=\exp[-\pi a(x_1-x_{1g})_t^2-\pi b(x_2-x_{2g})_t^2] \quad (3)$$

where a,b are the scales of the two axis in the Gaussian envelope, (x_{1g},x_{2g}) is the location of the peak of the Gaussian envelope. The rotation and translation transformation is denoted as below, where θ_g is the rotation angle of the Gaussian envelope.

$$\begin{cases}(x_1-x_{1g})_t=\ (x_1-x_{1g})\cos\theta_g+(x_2-x_{2g})\sin\theta_g\\(x_2-x_{2g})_t=-(x_1-x_{1g})\sin\theta_g+(x_2-x_{2g})\cos\theta_g\end{cases} \quad (4)$$

2.2 Introduction to Curvelet Transforms

A special member of the emerging family of multiscale geometric transforms is the Curvelet transform. It was developed in an attempt to overcome inherent limitations of traditional multiscale representations such as wavelets [16]. The Curvelet transform is a multiscale pyramid with many directions and positions at each length scale, and needle-shaped elements at fine scales.

In Curvelet transform, the work is throughout in two dimensions, i.e., $\mathbb{R}^2$, with spatial variable $x=(x_1,x_2)\in\mathbb{R}^2$, with the frequency domain variable ω, and with r and θ polar coordinates in the frequency-domain. The basic pair of windows includes the "radial window" $W(r)$ with $r\in(1/2,2)$ and "angular window" $V(t)$ with $t\in[-1,1]$ .Then, the frequency window U_j is defined in the Fourier domain as follows:

$$U_j(r,\theta)=2^{-3j/4}W(2^{-j}r)V(\frac{2\lfloor j/2\rfloor\theta}{2\pi}) \quad (5)$$

where $j=0,1,...$ is a scale parameter, $\lfloor j/2\rfloor$ is the integer part of $j/2$. The support of U_j is a polar "wedge" defined by W and V which is applied with scale-dependent window widths in each direction.

Define the waveform $\varphi_j(x)$ by means of its Fourier transform $\hat{\varphi}_j(\omega)=U_j(\omega)$, $\omega=(\omega_1,\omega_2)\in\mathbb{R}^2$ is slightly abuse by letting $U_j(\omega_1,\omega_2)$ be the window defined in the polar coordinate system. The equispaced sequence of rotation angle is denoted as $\theta_l=2\pi\cdot2^{-\lfloor j/2\rfloor}\cdot l$, with the orientation parameter $l=0,1,...$ such that $0\le\theta_l\le2\pi$. And the sequence of translation parameter $k=(k_1,k_2)\in\mathbb{Z}^2$. With these notations, the Curvelets are defined as function of $x=(x_1,x_2)\in\mathbb{R}^2$ at scale 2^{-j}, orientation θ_l and position $x_{j,l,k}=R_{\theta_l}^{-1}(k_1\cdot2^{-j},k_2\cdot2^{-j/2})$ by Eq. (6).

$$\varphi_{j,l,k}(x) = \varphi_j(R_{\theta_l}(x - x_{j,l,k})) \quad (6)$$

Where R_θ is the rotation by θ radians as follows:

$$R_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \quad (7)$$

So the Curvelet coefficient map $c_{j,l,k}(x_1, x_2)$ is then simply the inner product between an element $f(x_1, x_2) \in L^2(\mathbb{R}^2)$ of image and a Curvelet $\varphi_{j,l,k}$.

$$c_{j,l,k}(x_1, x_2) = < f(x_1, x_2), \varphi_{j,l,k} > \quad (8)$$

In digital Curvelet Transforms, similar with the continuous-Time Curvelet transform, U_j smoothly extracts frequencies near the dyadic corona $\{2^j \le r \le 2^{j+1}\}$ and near the angle $\{-\pi \cdot 2^{-j/2} \le \theta \le \pi \cdot 2^{-j/2}\}$. But due to the coronae and rotation are not especially adapted to Cartesian arrays, in digital Curvelet Transform, the "Cartesian coronae" based on the concentric squares and shears are utilized.

2.3 Saliency Detection using 2D Gabor and Curvelets Transform

Firstly, the red/green (*RG*) , blue/yellow (*BY*) color and intensity (*I*) are selected as preattentive features. Then, we build the 2D Gabor and Curvelet transform functions $g(x_1, x_2)$ and $\varphi(x_1, x_2)$ as we described before. These transforms are utilized to build the feature maps of 2D Gabor $F_G(x_1, x_2)$ and the feature maps $F_C(x_1, x_2)$ of Curvelet as follows:

$$F_G(x_1, x_2) = < f(x_1, x_2), g > \quad (9)$$

$$F_C(x_1, x_2) = < f(x_1, x_2), \varphi > \quad (10)$$

The feature maps $F_G(x_1, x_2)$ and $F_C(x_1, x_2)$ are then across-scale combined and normalized into activation maps within Gabor and Curvelet channels based on the commonly used combination and normalization method in [11][13]. The activation maps are the components to simulate the bottom-up processing of attention.

2.4 Integrating Top-down Features Channel

To discuss whether the top-down information is useful in an unsupervised learning fashion, we add the top-down features channel in our model. Typically, viewers may reorient to the center of a scene at a greater frequency than to other locations. In order to exactly simulate human visual information processing, we should consider the influences of center bias. In this paper, the distance to the center for each pixel is calculated as the influence of center bias as follows:

$$Dis(x_1, x_2) = \frac{\sqrt{(x_1 - x_{1c})^2 + (x_2 - x_{2c})^2}}{4\sqrt{WT^2 + HT^2}} \quad (11)$$

where *WT* and *HT* are the width and height of the image. And (x_{1c}, x_{2c}) is the center of the image. In our model, $Dis(x_1, x_2)$ is the activation map of center priority channel constructed for image $f(x_1, x_2)$.

Other top-down features include the object detection results. In [15], some top-down features are learnt to model the saliency map by SVM, such as the face detection. In our model, the object activation map $Obj(x_1, x_2)$ is modeled by the Gaussian envelope whose location of the peak (x_{1o}, x_{2o}) is the object center determined by object detector.

$$Obj(x_1, x_2) = \exp[-2\sqrt{(x_1 - x_{1o})^2 + (x_2 - x_{2o})^2} / (WT \times HT)] \quad (12)$$

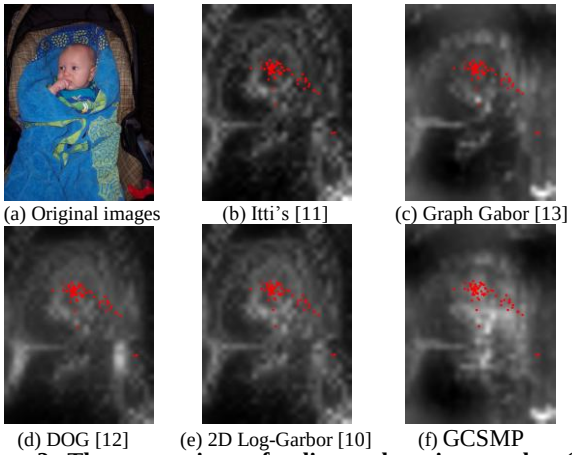
Finally, the activation map $Dis(x_1, x_2)$ of center priority and $Obj(x_1, x_2)$ of the object detection are equally weighted and linearly combined with the bottom-up activation maps just as Fig. 1. The output is proposed saliency map GCSMP.

3. EXPERIMENTAL RESULTS

In this section, we evaluate the performance on a public image dataset with collected eye tracking data on 1003 images [15]. This dataset is the largest one with eye tracking data and popularly utilized in saliency map construction. We compare proposed GCSMP saliency map with four other unsupervised saliency map models, including basic Itti's model [11], Graph model based on Gabor filter [13], DOG model [12] and 2D Log-Garbor model [10]. We will firstly demonstrate the experimental results based on bottom-up processing. Then, the influence of the top-down information will be discussed. To evaluate the performance of various saliency models, we choose the ROC curves and ROC areas. The ROC curves can be plotted as the False Positive Rate vs. Hit Rate. The ROC area can be then calculated as the area under the ROC curve to demonstrate the overall performance of a saliency model. Perfect prediction corresponds to the ROC area of 1, while random prediction generates an ROC area of 0.5.

3.1 Bottom-up Processing Results

In this section, we demonstrate the experimental results based on bottom-up processing, any features extracted based on the top-down processing are not considered into our model. Firstly, the comparison of the saliency maps constructed by our technique with other techniques is shown in Fig. 2. Although all techniques focus on salient changes to capture attention, it is obvious that the saliency regions detected by our model is more concentrate to fixation points. Therefore, our method can predict where human look more efficiently and more accurately.



(a) Original images (b) Itti's [11] (c) Graph Gabor [13]

(d) DOG [12] (e) 2D Log-Garbor [10] (f) GCSMP

Figure 2. The comparison of saliency detection results. (a) Original images. From (b) to (f) is the saliency maps with human fixations marked as red dots. (b) Itti's saliency map [11], (c) Graph Gabor saliency map [13], (d) DOG saliency map [12], (e) 2D Log-Garbor saliency map [10] (f) GCSMP saliency map.

Furthermore, the ROC areas of these techniques are shown in Table 1. We could easily observe from Table 1 that our model has the largest ROC area and achieves the best overall performance.

Table 1. ROC area comparison based on bottom-up model

Model	Basic Itti [11]	Graph Gabor [13]	DOG [12]	Log-Gabor [10]	Proposed model
ROC Area	0.6736	0.6820	0.6816	0.6874	**0.6990**

3.2 Experiment Results Integrated with Top-down Information

We have compared proposed technique based on 2D Gabor and Curvelet transforms with other techniques. As we described before, the top-down processing is one of the important components in human's attention. So in this section, we will discuss the influence from features of top-down processing.

The center bias is a common phenomenon when humans look in natural scenes, which has been considered into top-down attention model [6] [15]. Therefore, we firstly integrate the center priority into our model. After integrating the center bias for proposed model and all other compared models, we could obtain the ROC areas over all users and all images in Table 2. From Table 2, it can be seen that center bias could improve the performance greatly. Based on this observation, we could get the conclusion that human fixation is near the center of the image.

Table 2. The comparison of ROC area with center bias

Model	Basic Itti [11]	Graph Gabor [13]	DOG [12]	Log-Gabor [10]	Proposed model
ROC Area	0.7763	0.8169	0.8095	0.8176	**0.8200**

The average ROC curves of these different models are shown in Fig. 3. From Fig. 3, it can also be seen that although all models benefits the center bias, our proposed model reaches the highest Hit Rate when False Positive Rate is low. These results indicate that our model achieves the best performance.

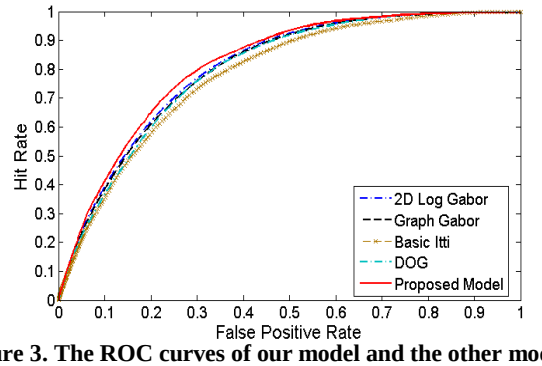


Figure 3. The ROC curves of our model and the other models.

Based on the research of neuroscience, neurons in visual association cortex for example the inferior temporal cortex (IT), respond selectively to a particular object, especially to human faces. And the feedback originating in some higher level areas such as V4, V5, and IT can modify the V1 responses and influence the human's attention in a top-down manner. Therefore, we also discuss the effectiveness of adding other higher level information: the objects detection features.

The face detection result is utilized as high-level feature to construct the saliency map learnt by SVM in [6] [15]. In our model, we add the face detection channel into our GCSMP model to obtain the saliency map. The ROC area increases from 0.8180 to 0.8281 in the images with human. But to the images without human, the false alarm of the detector will lead to the performance have a 1.08% reduction, from 0.8086 to 0.7999. Therefore, this result proves that the reliable tag information is useful to determine which object is in the image and build a better saliency map even not in a supervised learning fashion.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we propose an effective multi-scale and multi-orientation technique for saliency detection by using Garbor transforms based on its spatial localization ability and Curvelet transforms based on its better directional and edges representation ability. In experimental results, our model has the best performance and the highest ROC area in comparison with other three state-of-the art bottom-up techniques. Moreover, we also discuss the influence of center bias and object detection feature map in saliency map. Even after adding these priorities to other models, our model still has the best performance. In this work, we just consider the still feature and test our model in static images. For future work, we will consider the motion feature in temporal domain and then test our improved model in video.

5. ACKNOWLEDGEMENTS

This research was supported by Video Retargeting using Spatio-temporal Optimization (KFKT2011A09).

6. REFERENCES

[1] Anderson, John R., "Cognitive psychology and its implications", 6th Edition, Worth Publishers, 2004.

[2] E.A. Styles, "Attention, Perception, and Memory: An Integrated Introduction", First edition, Psychology Press, 2005.

[3] Koch, C. & Ullman, S., "Shifts in selective visual attention: Towards the Underlying Neural Circuitry," In Human Neurobiology, 1985.

[4] Parkhurst, D., Law, K., & Niebur, E., "Modeling the role of salience in the allocation of overt visual attention," In Vision Res, 2002.

[5] S. Avidan and A. Shamir. "Seam carving for content-aware image resizing," In ACM Transactions on Graphics, 2007.

[6] Zhong, S., Liu, Y., Liu, Y., and Chung, F.-L., "A semantic no-reference image sharpness metric based on top-down and bottom-up saliency map modeling", In ICIP, 2010.

[7] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-Based Video Segmentation with Graph Cuts and Sequentially Updated Priors", In ICME, 2009.

[8] Yu, H., Li, J., Tian, Y., Huang, T., "Automatic interesting object extraction from images using complementary saliency maps", In ACMMM, 2010.

[9] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," In Vision Res., 2000.

[10] Wang, M., Li, J., Huang, T., Tian, Y. Duan, L., and Jia, G.,"Saliency detection based on 2D log-gabor wavelets and center bias", In ACMMM, 2010.

[11] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", In TPAMI, 1998.

[12] R. Achanta, S. Hemami, F. Estrada and S. Süsstrunk, "Frequency-tuned salient region detection", In CVPR, 2009.

[13] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency", In NIPS, 2006.

[14] R. Fergus, P. Perona and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning", In CVPR, 2003.

[15] Tilke Judd, Krista Ehinger, Fr´edo Durand and Antonio Torralba, "Learning to predict where humans look," In ICCV, 2009.

[16] E. J. Cand`es and D. L. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise-C^2 singularities", Comm. On Pure and Appl. Math., 2004.