Saliency Detection Based on Adaptive DoG and Distance Transform

Hong-Yun Gao and Kin-Man Lam
Department of Electronic and Information Engineering
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong

***Abstract*—A novel computational model for detecting salient regions in color images is proposed, based on adaptive difference of Gaussian (DoG) filtering and distance transform. In our method, we first transform an image into the frequency domain, and perform adaptive DoG filtering, whose parameters are determined by the energy spectrum of the image. Then, the edge information is extracted from the DoG filtering output, and the distance transform is applied to the edge map. Finally, the Gaussian pyramids are used to enhance the distance transform performance. Our proposed method achieves spectral domain filtering as well as spatial domain edge extraction, thus exploiting the benefits from both the spatial domain and the spectral domain for saliency detection. We compare our proposed method with five existing saliency detection methods in terms of precision, recall, and F-measure. Experiments on the MSRA dataset show the outperformance of the proposed method over those saliency algorithms.**

Keywords—Saliency detection; difference of Gaussian; distance transform; Gaussian pyramid

I. INTRODUCTION

Visual saliency detection is one of the many extraordinary abilities of the human visual system (HVS), which directs eye attention to salient regions in the free-viewing process. Visual attention results from both fast, pre-attentive, bottom-up visual cues of the retinal input, as well as slow, top-down, task-dependent and memory-based processing [1]. Basically, the saliency value at a given pixel position is determined by the dissimilarity from the neighborhood pixels. This "dissimilarity" can be defined in different ways, such as center-surround contrast [2], self-information [3], or even Bayesian surprise [4]. A saliency map is useful in many applications, such as object detection, image compression, etc. The object detection task can be simplified by first calculating a saliency map, and then focusing on the identified salient regions. Image compression can also benefit from the saliency map, which can segment an image into salient and non-salient regions. For the non-salient regions, the compression ratio can be greatly increased, since people tend to neglect those regions.

One of the earliest computational models was proposed by Itti *et al*. [2]. The algorithm is based on center-surround contrast, which is computed as the difference between the fine and coarse scales in the intensity, color, and orientation maps, and is inspired by the behavioral and neuronal architecture proposed by Koch and Ullman [5]. Three conspicuity maps are generated by multiple-scales maps from the intensity, color, and orientation feature maps, respectively. The final saliency map is produced by combining the three conspicuity maps after normalization. However, this method suffers from severe computational complexity and over-parameterization. Ma and Zhang [6] simulated human perception and proposed a model based on local contrast analysis. Harel *et al*. [7] proposed a visual saliency model in a graph theory framework, namely Graph-Based Visual Saliency. Hou and Zhang [8] were the first to propose the spectral saliency model (spectral residual model), based on the difference between the perceived log-spectrum and the characteristic log-spectrum of natural images. Then, the final saliency map is obtained by transforming the spectral residual back to the spatial domain. Achanta *et al*. [9] proposed the frequency-tuned saliency detection method, which employs the difference of Gaussian (DoG) filter to eliminate redundant information, and outputs full-resolution saliency maps with well-defined boundaries of salient objects.

In this paper, we present a novel computational model for saliency detection based on adaptive DoG filtering and distance transform. Our algorithm performs spectral domain filtering as well as spatial edge extraction, thus exploiting the benefits from both the spatial domain and the spectral domain.

The rest of the paper is organized as follows. Section II describes the proposed framework for saliency detection. Experimental results are evaluated and compared with other existing algorithms in Section III. Finally, the conclusion is given in Section IV.

II. PROPOSED FRAMEWORK

A. Edge Detection and Distance Transform

Rosin [10] attempted to use edge density as a measure of saliency, because edge is easy to compute and requires simple or no parameters. The edge density is measured using distance transform [11]. The basic idea of distance transform is straightforward: the value of a particular pixel in the transform domain is the distance to the nearest edge. Hence, the larger distance a pixel is from its nearest edge, the higher its transform value is. Usually, the salient regions should have sharp intensity contrast or color transition, so the salient region generally presents high edge density. Since distance transform is a measure of edge intensity, this allows distance transform to be correlated with saliency detection. In order to generate saliency map by distance transform, the first step is to use edge operators, e.g. the Sobel or Canny operators, to produce the edge map. The edge is then subject to distance transform.

However, the standard distance transform is defined on binary images rather than on gray-level images or color images, it is thus required to threshold images to form the binary image counterparts. We use a threshold value starting from an intermediate value to eliminate some false alarms. In the final step, the saliency map is obtained by simply complementing the saliency map or reversing the sign in the thresholding, because a smaller value indicates more saliency, while a larger value manifests less saliency. As shown in the second row of Fig. 1, the saliency in those images with a homogeneous background or a simple texture can be detected appropriately, e.g. the car and the tower.

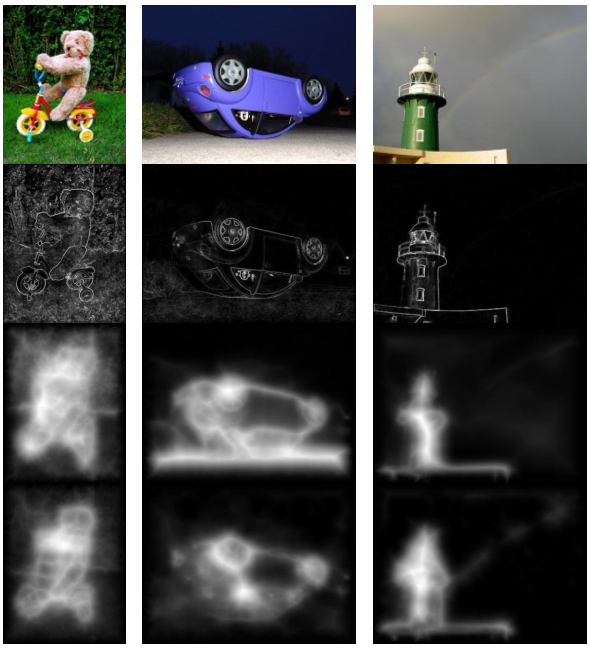


Fig. 1. The first row shows images from the MSRA dataset [12]. The second row shows the results after the distance transform. The third row shows the saliency maps enhanced by using Gaussian pyramids. The fourth row shows the saliency maps obtained by DoG filtering before the distance transform.

B. Gaussian Pyramid

The simple distance transform works well for images with homogeneous background. However, for images in complex scenes or with complex textures, the performance becomes unsatisfactory, since regions with high edge density are not always salient, e.g. the "teddy bear" image in Fig. 1. It is highly probable that a texture or a background has a high edge density, but observers usually do not focus on these areas. As can be seen in Fig. 2, the salient objects in the three images are the dandelion, the horses, and the human face, respectively. However, the background leaves, grass texture, and bark texture obscure the salient objects.

This problem can be solved by using the Gaussian pyramid. Inspired by the combination of image pyramids [13], we propose the use of Gaussian pyramids to enhance saliency detection. First, images of several spatial scales are computed by using the Gaussian pyramids, which progressively low-pass and down-sample an image into its half-sized counterparts. Then, the images of the different scales are processed individually, i.e. subject to distance transform, to construct the saliency maps of different scales. Finally, each of the saliency maps is rescaled to its original size using bilinear interpolation, and summed together to form the final saliency map. The reason for using Gaussian pyramids to enhance saliency detection is straightforward. Gaussian pyramids accentuate important edges by diffusing the large values to the neighbors; this can better represent the true salient regions. However, weak edges are attenuated by diffusing their small values to their neighbors; this makes the false salient regions less salient, and they can be removed by thresholding. We follow [13] in that the number of Gaussian pyramid levels is set at $n = \log_2(\min(w, h)/10)$, where w and h are the width and height of the image, respectively. As shown in the third row of Fig. 1, the salient regions become more conspicuous by using Gaussian pyramids. However, some non-salient regions are also enhanced and focused, e.g. the border between the bright and the black region in the car image.

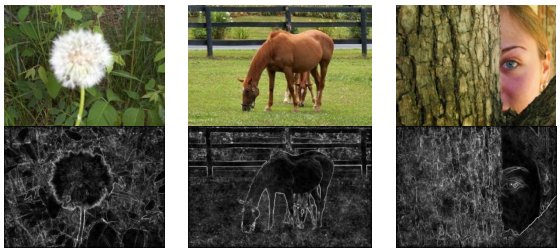


Fig. 2. The first row shows the original images. The second row shows the saliency maps using distance transform.

C. Difference of Gaussian (DoG)

To cope with non-salient edges (e.g. the border edges in the car image), we propose applying band-pass filtering to eliminate these kinds of non-salient edges before performing distance transform. In our experience, these non-salient edges usually appear with low-frequency components in images, such as the edge between the sea and a beach, the boundary between the sky and land, etc. Using high-pass filtering can remove these clear boundaries while keeping all other frequency components untouched. Also, as shown in Fig. 2, cluttered backgrounds and textures are always the obstacles in generating accurate salient regions, since the variation of the pixel values is so high that these regions are sometimes mistaken as salient regions. We can perform low-pass filtering at the same time, since cluttered backgrounds and complex textures always indicate high-frequency components. Taking the low-pass and high-pass filtering together, a band-pass filter is derived. The Difference of Gaussian (DoG) is one of the most widely-used band-pass filters, as shown below.

$$DoG(x,y) = \frac{1}{2\pi}\left[\frac{1}{\sigma_1^2}e^{-\frac{x^2+y^2}{2\sigma_1^2}} - \frac{1}{\sigma_2^2}e^{-\frac{x^2+y^2}{2\sigma_2^2}}\right] = g(x,y,\sigma_1) - g(x,y,\sigma_2), \quad (1)$$

where σ_1 and σ_2 ($\sigma_1 > \sigma_2$) are the standard deviations of the Gaussian function. Furthermore, many saliency models have

also exploited DoG, but the objectives of using DoG vary from one model to another. Itti *et al*. [2] used DoG to calculate the center-surround contrast, while Achanta [9] employed DoG to preserve more frequency components in images than other saliency algorithms. According to Achanta [9], some of the state-of-the-art saliency models, such as IT [2], MZ [6], GB [7], SR [8], primarily operate using extremely low-frequency content in images. Consequently, the boundaries of those salient objects are blurred, as can be seen in Fig. 3.

In the fourth row of Fig. 1, we calculate the saliency map using DoG with σ_1=2 and σ_2=1; this is the simple DoG (SDOG) method. This method removes the textures, such as the grass and trees in the background of the teddy bear image, which are high-frequency components. Moreover, the SDOG eliminates entirely the border between the bright and the black region in the car image, which demonstrates the effectiveness of DoG in removing low-frequency non-salient boundaries. Assume that the pass band is within [ω_l ω_h], where ω_l and ω_h are determined by σ_1 and σ_2, respectively. The pass-band width is controlled by the ratio of σ_1:σ_2. Marr [14] proposed that DoG is the most suitable operator for detecting intensity changes when the standard deviations of the Gaussians have the ratio of 1.6:1. It is possible to combine several narrow band-pass DoG filters by fixing the standard deviation ratio [9] as follows:

$$\sum_{n=0}^{N-1}[g(x,y,\rho^{n+1}\sigma)-g(x,y,\rho^{n}\sigma)]=g(x,y,\sigma\rho^{N})-g(x,y,\sigma), \quad (2)$$

D. Adaptive DoG

As can be seen in the fourth row of Fig. 1, the rainbow in the "green tower" image and the trees in the "teddy bear" image are not sufficiently salient in the respective saliency maps, but they are somehow emphasized by our algorithm. The reason for this is that the algorithm is image-independent, i.e. a fixed pass-band of the DoG cannot accurately filter out the non-salient low-frequency and high-frequency components of different images. To improve the performance, images should be subject to image-dependent DoG filtering. We propose an adaptive DoG for our algorithm such that images with a lot of high-frequency background or texture should have more high-frequency components eliminated, while images with a simple configuration should preserve more of the low-frequency components. This filtering process should therefore take the image frequency components into account. According to Ruderman [15], natural images are not random; rather, they obey the distribution in (3), which is called scale invariance or the *1/f* law.

$$E\{A(f)\}\propto 1/f, \quad (3)$$

where *A(f)* is the amplitude of the Fourier transform.

Inspired by this fast-decaying energy distribution of (3), we propose to use energy to define the adaptive standard deviations to be used in the DoG filter. The two standard deviations are defined as follows:

$$\sigma_1=\arg En\{G(u,v,\sigma_1)I(u,v)\}/En\{I(u,v)\}=T_1, \text{ and} \quad (4)$$

$$\sigma_2=\arg En\{G(u,v,\sigma_2)I(u,v)\}/En\{I(u,v)\}=T_2, \quad (5)$$

where (u, v) represents the frequency indices. $G(u,v,\sigma)$ and $I(u,v)$ are the Gaussian low-pass filter and the image in the frequency domain, respectively. *En* is the energy operator in the frequency domain. T_2 and T_1, where $T_2 > T_1$, are the ratio parameters of the preserved frequency components compared with the original image frequency components.

III. EXPERIMENTAL RESULTS

In this section, we evaluate the proposed saliency detection algorithm on the MSRA dataset [12]. As mentioned in Section 1, we have selected the methods IT [2], MZ [6], GB [7], SR [8], and FTS [9] for comparison.

In the experiment, we empirically set T_1 and T_2 at 25% and 50%, respectively, to perform the band-pass filtering. Fig. 3 shows the results for the different methods and for our proposed method, denoted as SDOG and ADOG, respectively. SDOG can detect salient objects well, but some non-salient information is also highlighted, such as the background of the "puppy" image and the "soccer" image. However, ADOG can achieve a better performance in eliminating false alarms to a certain extent, by taking individual image's characteristics into account, as shown in the (h) column of Fig. 3.

To compare and evaluate the performances of the different methods, we use the Precision, Recall, and F-measures, computed based on the binary ground-truth images and the thresholded saliency maps. Precision measures the percentage of true salient pixels detected with respect to the total number of salient pixels in the ground-truths. Recall is used to measure the percentage of correct salient pixels detected with respect to the total number of true salient pixels in the ground-truths. F-measure is an overall performance indicator, which is defined as the weighted harmonic mean of Precision and Recall with a nonnegative parameter α, to tune the relative importance of Precision and Recall. In the evaluation, we set α =0.5.These three performance measures are defined as follows:

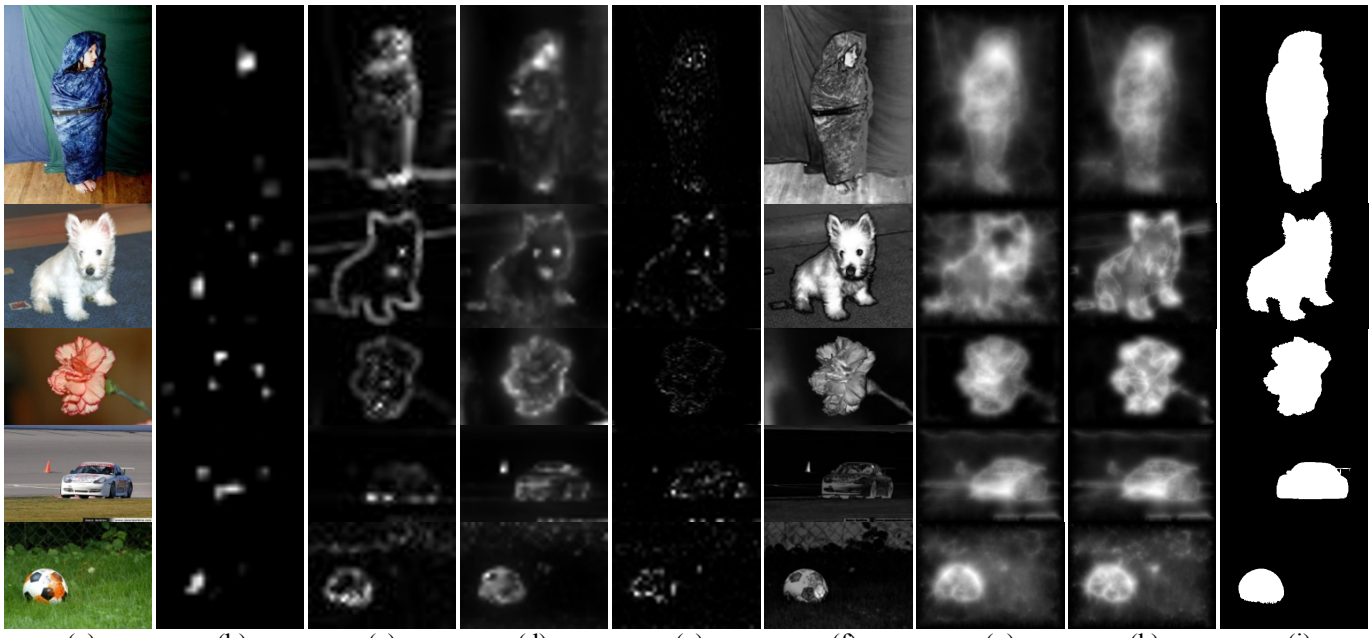
$$\text{Precision}=\frac{\sum_{x,y}s(x,y)g(x,y)}{\sum_{x,y}s(x,y)}, \quad (6)$$

$$\text{Recall}=\frac{\sum_{x,y}s(x,y)g(x,y)}{\sum_{x,y}g(x,y)}, \text{ and} \quad (7)$$

$$F=\frac{(1+\alpha)\times\text{Precision}\times\text{Recall}}{\alpha\times\text{Precision}+\text{Recall}}, \quad (8)$$

where $s(x, y)$ and $g(x, y)$ are the values of the saliency mask and the ground-truth image, respectively, at (x, y). The saliency mask is obtained by using the adaptive thresholding scheme [9], where the threshold is twice the mean of the saliency map.

As can be seen in Fig. 4, our proposed SDOG method achieves a similar performance to FTS [9], with a larger value in Recall, but a slightly smaller value in Precision. The overall performance measure, the F-measure, of SDOG is lower than that of FTS [9]. However, with the adaptive mechanism, the performance of ADOG is greatly improved, with an increase in the Precision, Recall, and F-measure.



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 3. Saliency maps generated from images in the MSRA dataset [12]: (a) Original images, (b) IT [2], (c) MZ [6], (d) GB [7], (e) SR [8], (f) FTS [9], (g) SDOG, (h) ADOG, and (i) ground-truth images.

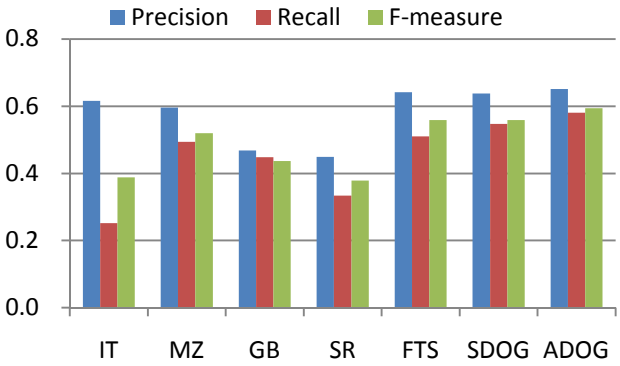


Fig. 4. Performance comparison between the proposed model and other models.

IV. CONCLUSION

In this paper, we propose a novel method to perform saliency detection based on the adaptive Difference of Gaussian (DoG) and distance transform. Our method performs spectral-domain filtering as well as spatial-edge extraction, thus exploiting the benefits from both the spatial domain and the spectral domain. In addition, we propose using an adaptive DoG scheme, which determines the two scales according to an image's content. This makes our algorithm detect saliency much more accurately. The experiment results on the MSRA [12] dataset demonstrate the superior performance of our proposed method when compared to five existing saliency detection methods.

ACKNOWLEDGEMENT

This work is supported by an internal grant from the Hong Kong Polytechnic University (Grant No. G-YN21).

REFERENCES

[1] E. Niebur and C. Koch, "Computational architectures for attention," *MIT Press*, pp.163-186, October 1995.

[2] L. Itti, C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254-1259, 1998.

[3] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, 2005.

[4] L. Itti and P. Baldi, "Bayesian surprise attracts human visual attention, " in *Advances in Neural Information Processing Systems*, 2006.

[5] C. Koch and S. Ullman, "Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219-227, 1985.

[6] Y. Ma and H. Zhang, "Contrast-based Image Attention Analysis by Using Fuzzy Growing," in *ACM Int. conference on Multimedia*, 2003.

[7] J. Harel, C. Koch and P. Perona, "Graph-Based Visual Saliency," in *Advances in Neural Information Processing Systems*, pp. 545-552, 2006.

[8] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[9] R. Achanta, S. Hemami, F. Estrada and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[10] P. Rosin, "A simple method for detecting salient regions," in *Pattern Recogniton*, vol. 42, no. 11, pp. 2363-2371, 2009.

[11] P. Rosin and G. West, "Salience distance transform," *Graphical Models and Image Processing*, vol. 57, no. 6, pp. 483-521, 1995.

[12] T. Liu, J. Sun, N-N. Zheng, X Tang and H Shum, "Learning to Detect a Salient Object," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[13] F. Liu and M. Gleicher, "Region Enhanced Scale-invariant Saliency Detection," in *IEEE Int. Conference on Mutimedia and Expo*, 2006.

[14] D. Marr, "Vision: a computational investigation into the human representation and processing of visual information," *MIT Press*, 2010.

[15] D. Ruderman, "The Statistics of Natural Images," *Network: Computation in Neural Systems*, vol. 5, no. 4, pp. 517–548, 1994.