Global salient object detection based on multiple visual features

Zhengyi Li, Heng Li*, Xinyu Chai *
Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
lizhenyi963.cn@qq.com
*Corresponding author: xychai@sjtu.edu.cn and liheng@sjtu.edu.cn

Abstract. Salient object detection has become an important tool in the fields of computer vision and image processing. In this paper, we propose a novel global salient object detection model based on multiple visual features. Firstly, following the mechanism of parallel processing of visual information, we obtain three saliency maps based on colour, intensity difference, and spatial distribution. Secondly, introducing the concept of spatial gaze point of an image, we calculate three Gaussian weighted maps based on three saliency maps to constrain their background noise and get three weighted saliency maps. Finally, the final saliency map is produced by fusing three weighted saliency maps. We compare our method with 10 state-of-the-art saliency methods on a public dataset, and the results show that our method outperforms the other 10 methods.

1. Introduction

The human visual system is always able to quickly locate features or objects that are of interest to them when observing a natural scene. This is a bottom-up and data-driven attention mechanism [1]. The human visual system is also able to quickly search for the information needed in a scene based on commands sent by its own brain. This is a top-down and task-driven attention mechanism [2]. Realistic, top-down visual saliency models involve more complex knowledge in the fields of neuroscience, biology, and computer vision. Moreover, as the cognition to cognitive mechanisms of human brain is limited, and the principle of this attention mechanism cannot be profoundly revealed, studies in this area are limited by now. On the contrary, currently, the bottom-up visual saliency model has good applications in many fields, like target recognition [3], image segmentation [4], image and video resizing [5], image retrieval [6], and artificial vision [7].

At present, the bottom-up saliency model can be roughly divided into local methods and global methods. Local saliency models achieve a saliency map by calculating the difference between the centre and the periphery based on multi-scale visual features (like brightness, colour, edges, texture, motion) in parallel, and the greater the difference is, the more salient the centre will be [1]. The saliency map calculated by local methods generally is scattered or fuzzy. For example, Itti et al. [1] propose a classic local saliency model to produce a blur saliency map by calculating the central-peripheral differences based on multi-scale visual features. On this basis, Harel et al. [8] propose a graph-based visual saliency model that further highlights the salient regions. Ma and Zhang [9] construct a model by highlighting the difference of quantized colour distance between the perception centre and its neighbourhood. The major drawback of above methods is in overemphasizing the boundaries of salient region, but ignoring the entire salient region. Conversely, the global saliency
models consider the visual feature difference between a unit (pixel or block) and the entire picture and can produce the entire salient region. For example, there are the saliency methods globally considering pixel brightness (LC) [10], average colour (FT) [11], three channel RGB (HC) [12], amplitude and phase in frequency domain (SR) [13]. However, the saliency models only considering a visual feature are imperfect, like LC and FT, which loss other potential salient cues. As an improvement, HC method fuses three channel cues and obtains superior performance than LC and FT. On this basis, Cheng et al. further take count spatial relationship inside the image and construct the RC model [12]. However, this method needs to segment the image as a pre-processing, which adds an extra time cost. Li et al. [7] introduce the concept of intensity difference as another cue to remove the effect of light reflection from salient object. In addition, there are some saliency models combine the local and global methods to improve the performance of salient detection, like SF method [14], which effectively combines the local and global contrast considered the colour and spatial relationship to produce a finer saliency map.

In this paper, following the mechanism of parallel processing of visual features, we construct an advanced saliency model based on three visual features: colour, intensity difference, and spatial distribution. We compare our method with other 10 state-of-the-art methods by using widely admitted benchmarks on a public dataset. The results show that the saliency map obtained by our method is superior over other 10 methods.

2. Method

In this work, we firstly calculate three saliency maps based on three visual features (see Figure 1-b/c/d). Then similar to [7, 15], we introduce the concept of gaze points of image, and respectively calculate the spatial gaze points of the three saliency maps to obtain three Gaussian weighted maps. Finally, we multiply three saliency maps with its corresponding Gaussian weighted maps to get three weighted saliency maps. After normalizing, the final saliency map is obtained by fusing three weighted saliency maps. Figure 2 shows the image processing flow of our method.

2.1 Saliency based on color

As discussed in [12], the region that has large colour difference compared to the entire image should have higher salient values. Similar to [12], the saliency of a pixel \( i \) based on colour is defined as:

\[
S_c = \sum_{j \neq i} f_i |C_j - C_i|^2
\]

where \( f_i \) represents the colour frequency corresponding to this pixel appears in the image, and \( C_j \) represents Lab values (that are converted by RGB values) of the pixel \( i \). The complexity of Eq. (1) is related to the number of pixel and the color count in the image. In order to improve computational efficiency, we employ color quantization to quantize R-G-B color channels by using an adapted weight term \( WT = [1, k, k^2] \) [7, 12]. The quantized \( R' \)-\( G' \)-\( B' \) color channels are defined as:

\[
R' = \text{floor}(R \times WT(2) \times WT(1))
\]

\[
G' = \text{floor}(G \times WT(2) \times WT(2))
\]

\[
B' = \text{floor}(B \times WT(2) \times WT(3))
\]

where the R-G-B value has been normalized, and the k is set as 10 in our work. Then we simply define the colour pallet value of pixel \( i \) in RGB space as:

\[
C'_i = R' + G' + B'
\]
By Eq. (2) and Eq. (3), we can quantize $256^3$ colours to $(k^3+k^2+k)$. Then, we can count all the colour pallet frequencies in the image, and construct a mapping relationship between each pixel and colour pallet values. Finally, we define the new R-G-B value of this colour: the average of the sum of the original R-G-B values of all the pixels corresponding to this colour.

2.2 Saliency based on intensity difference
Taking into consideration the inhomogeneous saliency results due to ignoring the colour difference caused by the object itself and light reflections, similar to [7], we define the intensity difference value of a pixel as:

$$I_i = \frac{\max(L_R, L_G, L_B) - \min(L_R, L_G, L_B)}{\max(L_R, L_G, L_B)}$$

where the $L_R$, $L_G$, $L_B$ represents the R-G-B values of this pixel. Then, the saliency based on the colour intensity difference is defined as:

$$S_{i} = \sum_{j \in c(i)} f_i ||l_j - l_i||^2$$

where $f_i$ represents the frequency of the colour corresponding to pixel $i$.

2.3 Saliency based on spatial distribution
Spatial information is an important saliency cue, because salient object always has compact distribution. To this end, we segment the raw image into $N$ perceptually uniform patches by using SLIC method proposed by [16]. We set $N=50$ in this work. It is easy to understand that the patches with compact distribution compared to the whole image are more likely a salient region, and should have higher saliency. Therefore, we define the saliency based on spatial distribution as:

$$S_{Di} = 1 - \frac{\sum_{j=1}^{N} ||P_j - \mu_i||^2 W_{ij}}{\text{normalization}}$$

Where $P_j$ represents the position of segment $j$, $c_i$ represents the mean colour of segment $i$ in Lab colour space, $\mu_i = \sum_{j=1}^{N} W_{ij}$ represents the mean position of $c_i$ relative to whole image, and $W_{ij}$ represents the similarity of colour $c_i$ and $c_j$. Similar to [14], we define $W_{ij}$ as:
\[ W_{ij} = \frac{1}{Z_i} e^{-A}, \quad A = -\frac{1}{2\sigma_c^2} ||c_i - c_j||^2 \]  

(7)

where \( Z_i \) ensure the \( \sum_{j=1}^{N} W_{ij} = 1 \), \( \sigma_c \) represents the colour sensitivity of spatial distribution, and we set \( \sigma_c = 20 \) in our work.

### 2.4 Gaze weighted term

Similar to [15], an attention centre point is assumed in each image. Therefore, after calculating the saliency map based on three visual features, we introduce a Gaussian weighted term to constrain the potential background noise. Gaussian weighted term is defined as:

\[ W^S = e^{-A}, \quad A = \frac{1}{2\sigma_1^2}(i_0 - i)^2 + \frac{1}{2\sigma_2^2}(j_0 - j)^2 \]  

(8)

where \( \sigma_1 = 0.5M \cdot \gamma \) and \( \sigma_2 = 0.5N \cdot \gamma \) represents the standard deviation of the Gauss weight map at two directions, the \( M \) and \( N \) represents the height and width of the map, and \( \gamma \) is set as 1/2.355 in our work. The \( i_0 \) and \( j_0 \) represents the centre point of the Gauss weight map, and the region near the centre point will give more weight, conversely the region around will be suppressed. According to the law of Gestalt, The \( i_0 \) and \( j_0 \) can be obtained by Eq.(9):

\[ i_0 = \frac{U_{00}}{U_{01}}, \quad j_0 = \frac{U_{10}}{U_{01}} \]  

(9)

where the \( U_{00} \) represents the zero-order moment of saliency map, and \( (U_{10}, U_{01}) \) represents the first-order moment. We specifically define as follows:

\[ U_{00} = \sum_{i=1}^{M} \sum_{j=1}^{N} S(i,j) ; \quad U_{10} = \sum_{i=1}^{M} \sum_{j=1}^{N} i \cdot S(i,j) ; \quad U_{01} = \sum_{i=1}^{M} \sum_{j=1}^{N} j \cdot S(i,j) \]  

(10)

After we get three Gaussian weight terms based on three saliency maps, we define the final saliency map as:

\[ S = (S_c \cdot W^S)_{\text{normalization}} + (S_t \cdot W^S)_{\text{normalization}} + (S_d \cdot W^S)_{\text{normalization}} \]  

(11)

where \( S \) is normalized to the range \([0 \cdots 1]\).

### 3. Test Results and Discussions

![Figure 3: A visual comparison of 10 saliency methods to ours and ground truth (GT). (a) Source, (b) IT, (c) SR, (d) AC, (e) CA, (f) GB, (g) LC, (h) FT, (i) HC, (j) SF, (k) GCSR, (l) Ours and (m) GT.](image-url)

To evaluate the superiority of our method, we compare our method with 10 state-of-the-art saliency methods (IT [1], SR [13], AC [17], CA [18], GB [19], LC [10], FT [11], HC [12], SF [14], GCSR [7])
on the MSRA-1k set [14] of 1000 images with binary ground truth. Figure 3 shows the results of these saliency detection methods and our method. From the visual comparison, the saliency map calculated by our method is closer to the ground truth.

### 3.1. Precision and recall

Similar to [13, 14], we use the precision-recall (PR) curve to evaluate the superiority of our method. Given a saliency map normalized to [0, 255], 256 binary saliency maps are generated by varying the threshold in the range [0,255]. Precision is also called positive predictive value, which represents the ratio of correctly assigned salient pixels to all salient region pixels. Recall is also called sensitivity, which represents the ratio of correctly assigned salient pixels to all ground truth pixels. Figure 4 A and B shows performance comparison of PR curve with varied thresholds, in which the PR curve of our method is closer to the top right of the chart, representing a higher precision rate with the same recall rate.

Figure 4: Performance comparison of our method with other 10 methods: A and B are PR curve; C and D are F-measure curve; E and F are ROC curve.

### 3.2. F-measure

As discussed in [14], both the precision and recall only consider the true positive counts, they ignore the true negative counts. Therefore, PR curve cannot evaluate the efficiency of detecting the true negative counts. Meanwhile, high precision can be achieved at the expense of reducing the recall and vice-versa. So it is necessary to evaluate the efficiency of all methods by considering the precision and recall simultaneously. Similar to [14], we compute a metric called F-measure, which is defined as:

\[ F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \]  

where, \( \beta^2 = 0.3 \). Figure 4 C and D shows performance comparison of F-measure curve with varied thresholds. As shown in the chart, our method has a higher F-measure value than other methods, except SF method within the range [0,55] of threshold. This is because SF method has a cleaner background than ours, but has inhomogeneous saliency results.

### 3.3. ROC

In addition, we also compute the true positive (TP) ratio and false positive (FP) ratio, and draw the ROC curve. Figure 4 E and F shows the ROC curve results of all methods. It is obvious that the curve of our method is closer to the top left, which means the higher TP value under the same FP value.
3.4. Performance
We calculate the average running time of our method based on MSRA-1k. It takes 0.606s at Matlab2018a to process an image on average under the Intel Core i5-7500 3.4GHz CPU and 16GB RAM.

4. Conclusion
In this paper, we propose a global salient object detection algorithm based on multiple visual features. Furthermore, we introduce the concept of spatial gaze point of an image to constrain the background noise in saliency map. Experiment results show that our method outperforms various state-of-the-art methods.

Acknowledgments
This work is financially supported by Basic Research Program of China (973 Program), The National Natural Science Foundation of China (61472247,61773256).

References
[1] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis[M]. IEEE Computer Society, 1998.
[2] Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry.[J]. Hum Neurobiol, 1987, 4(4):219-227.
[3] Li Z, Itti L. Saliency and gist features for target detection in satellite images,[J]. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society, 2011, 20(7):2017-29.
[4] H. Li, X. Su, J. Wang, H. Kan, T. Han, Y. Zeng, X. Chai, Image processing strategies based on saliency segmentation for object recognition under simulated prosthetic vision, Artificial Intelligence in Medicine, 84 (2018) 64-78.
[5] L. Heng, Z. Yajie, L. Zhuofan, C. Xiaofei, S. Xiaofan, S. Xiaohong, W. Jing, C. Xinyu, An optimized content-aware image retargeting method: toward expanding the perceived visual field of the high-density retinal prosthesis recipients, Journal of neural engineering, 15 (2018) 026025.
[6] Li L, Jiang S, Zha Z J, et al. Partial-Duplicate Image Retrieval via Saliency-Guided Visual Matching[J]. IEEE Multimedia, 2013, 20(3):13-23.
[7] Li H, Han T, Wang J, et al. A real-time image optimization strategy based on global saliency detection for artificial retinal prostheses[J]. Information Sciences, 2017, s 415–416(2017):1-18.
[8] Harel J , Koch C , Perona P . Graph-Based Visual Saliency.[J]. Proc of Neural Information Processing Systems, 2006, 19:545-552.
[9] Ma Y F , Zhang H J . Contrast-based image attention analysis by using fuzzy growing[C]// Eleventh Acm International Conference on Multimedia. ACM, 2003.
[10] Zhai Y , Shah M . Visual attention detection in video sequences using spatiotemporal cues[C]// Acm International Conference on Multimedia. ACM, 2006.
[11] Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2009.
[12] Cheng M M , Mitra N J , Huang X , et al. Global Contrast Based Salient Region Detection.[C]// Computer Vision and Pattern Recognition. IEEE, 2011.
[13] Hou X , Zhang L . Saliency Detection: A Spectral Residual Approach[C]// 2007 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2007.
[14] Perazzi F, Krähenbühl P, Pritch Y, et al. Saliency filters: Contrast based filtering for salient region detection[C]// IEEE Conference on Computer Vision & Pattern Recognition. 2012.
[15] Liu J, Wang S. Salient region detection via simple local and global contrast representation[J]. Neurocomputing, 2015, 147(1):435-443.
[16] Achanta R, Shaji A, Smith K, et al. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2012, 34(11):2274-2282.

[17] Achanta R, Estrada F, Wils P, et al. Salient Region Detection and Segmentation[M]// Computer Vision Systems. Springer Berlin Heidelberg, 2008.

[18] Goferman S, Zelnikmanor L, Tal A. Context-Aware Saliency Detection[J]. IEEE Trans Pattern Anal Mach Intell, 2012, 34(10):1915-1926

[19] Schölkopf B, Platt J, Hofmann T. Graph-Based Visual Saliency[J]. Proc of Neural Information Processing Systems, 2006, 19:545-552