REGION BASED MULTI-SPECTRAL SALIENCY DETECTION

G. P. Zhao, Y. M. Bo, M. F. Yin

Academy of Automation, Nanjing University of Science and Technology
Xiaolingwei st. 200, 210094, Nanjing, China
Phone: +86-025-84439725
E-mail: zhaogaopeng@sina.com, byming@njust.edu.cn, qiaodancumt@163.com

Visual saliency detection is useful for applications like image segmentation, image retrieval et al. According to the thought that the near-infrared image can provide complementary information to the corresponding visible image, a new region based multi-spectral saliency detection method is proposed. The source images are decomposed into homogeneous regions by superpixel algorithm in 6-dimentional vector space which is composed of the LAB colour values of the visible image, the grey value of the near-infrared image and the two-dimensional pixel coordinate. Then the final saliency map is obtained by measuring the region’s contrast which is computed by using multi-spectral colour information and spatial distance. The publicly RGB-NIR dataset is adopted for experiment evaluation; Compared with five other state-of-the-art methods, the proposed method produces superior performance. Experimental results show that the proposed method is effective.

Keywords: saliency detection, multi-spectral, superpixel, region contrast

1. Introduction

The visual attention mechanism in humans can rapidly focus on the context relevant stimuli while suppressing the unimportant ones. The mechanism has been studied in computer vision for a long time; Visual saliency detection [1, 2] is useful for applications like image retrieval, image compression and object segmentation et al, by assuming there are one or several salient objects in the image.

Most of the existing saliency detection methods can be categorized as top-down computational model and bottom-up computational model. Generally the top-down model is in conjunction with task-driven or visual search, and the bottom-up model is stimulus-driven. Many researches [3, 4, 5, 6, 7] have focused on the bottom-up model by computing the feature contrast including colour variation of individual pixels, edge and gradient, texture, histogram, spatial frequency, or combinations thereof. Goferman et al. [3] propose the context-aware saliency method, which detects the important parts of the scene rather than the traditional salient object, however, the method uses multi-scale computation and it tends to acquire a high salience value on the edge, not highlight the entire salient region. Li et al. [4] combine the global information form frequency domain analysis and local information from spatial domain analysis to generate the saliency map. Cheng et al. [5] employ the colour histogram feature in Lab colour space to calculate the image contrast but it is difficult to handle cluttered and textured background. Achanta et al. [6] determine the saliency by using the maximum symmetric surrounds and Vikram et al. [7] compute the saliency map by random regions. The above three methods [5, 6, 7] do not downscale the input image to low resolution and generate full resolution saliency map, however, they calculate the saliency value of each pixel independently like many other methods, so the correlation between pixels is not considered properly. Cheng et al. [5] propose another contrast analysis method, region contrast, by combine region segmentation and histogram representation. It can generate spatially coherent saliency map at the cost of the reduced calculation performance. The above methods can show significant performance on the specific testing dataset, Borji et al. [1] review a large body of recent start-of-the-art visual attention models and provide a critical comparison of their capabilities and shortcomings.

Despite so many methods have been proposed and contributed to the specific task, the start-of-the-art performance is not satisfying. The same object should have an identical saliency value in an ideal case; however most of the existing methods ignore the spatial correlation between pixels. Also, almost methods are studied based on the visible image. Due to the differences in the characteristics of image sensor, the infrared image can provide differentially information which is complementary to the visible image [8, 9].

Motivated by this, our focus in this paper is proposing a new region based multi-spectral saliency method based on a pair of visible and infrared images. Our main contribution is that the source images are decomposed into homogeneous regions by superpixel algorithm in 6-dimentional vector space; we modify the SLIC (simple linear iterative clustering) superpixel algorithm [10] to execute the superpixel
segmentation in the 6-dimentional vector space which is constituted by the Lab colour component in visible image, the grey component in infrared image and the position of the pixel in image coordinates; and then the saliency map can be computed by the region’s contrast, which is more emphasis on the object rather than single pixel. The proposed method is tested on the public benchmark [9] and it can obtain superior performance than the other comparable state-of-the-art methods.

2. Proposed Method

The framework of the proposed method is illustrated in Figure 1.

![Figure 1. The framework of the proposed method](image)

(a) (b) (c) (d) (e) (f)

(a) and (b) are the visible source image and the corresponding near-infrared image; (c) and (d) are the homogeneous regions using the superpixel segmentation method; (e) is the saliency map by the proposed method and (f) is the ground truth.

2.1. Color Space Illustration

Almost all saliency methods utilize a colour space to describe the visible image; some have used RGB while others have used the Lab. The Lab space is preferred to RGB space because the \( L \) component matches human perception of lightness, while the \( a \) component and \( b \) component approximate the human chromatic opponent system. We also adopt the Lab colour space to represent the visible image.

Near-infrared image is introduced into the saliency. It can be taken to interpret the different characteristics of the scene because its wavelength is discriminative to the visible wavelength. So the grey value of the near-infrared image represented by variable \( g \) is used as a complementary component to the Lab component of the visible image.

The visible image and the corresponding near-infrared image have the same resolution. We consider the resolution is \( rc \), where \( r \) and \( c \) are the number of rows and columns respectively, thus the \( \text{Lab}_{xyg} \) components have the same resolution \( rc \).

2.2. Generating Homogeneous Regions

We aim to decompose the input image into a series of perceptually homogeneous regions by clustering pixels with similar properties. Superpixel algorithm is able to group pixels into perceptually meaningful atomic regions. The SLIC algorithm [10] is faster, more memory efficient than the comparable superpixel methods and exhibits start-of-the-art boundary adherence. In this paper, the input images include a visible image which is a colour image and a near-infrared image which is a grey image. We extend the original SLIC algorithm to the \( \text{Lab}_{xyg} \) 6-dimentional space, where \( \text{Lab}_{g} \) is defined in the 2.1 section, \([xy]\) is the pixel position.

Assuming \( K \) is the desired number of superpixel, \( N \) is the pixel number of the input number. We initialize the cluster centres \( C_k = [l_k, a_k, b_k, g_k, x_k, y_k]^T \) with \( k = [1, K] \) at regular grid interval \( S \). The grid interval \( S \) is set to \( S = \sqrt{N/K} \) for approximately equally sized superpixels. To avoid initializing the cluster centre on the edge or a noisy pixel, the centres are moved to seed locations corresponding to the lowest gradient position in a \( 3 \times 3 \) neighbourhood.

After initialization, each pixel \( i \) is associated with the nearest cluster centre according to the distance measure. Assuming the pixels that have similar properties lie in a region \( 2S \times 2S \) around the superpixel centre, the search is done in the \( 2S \times 2S \) region on the \( xy \) plane. The distance measure \( D \) is defined as follows:

\[
d_{\text{labg}} = \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2 + (g_k - g_i)^2},
\]

(1)
Mathematical and Computer Modelling

\[ d_{xy} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2}, \]  
\[ D = \sqrt{d_{labg}^2 + \left(\frac{m}{S}d_{xy}^2\right)}, \]  

where \( D \) represents the distance between a pixel \( i \) with the cluster centre \( C_k \). \( d_{labg} \) represents the colour distance and \( d_{xy} \) represents the spatial distance. The variable \( m \) is used to weigh the relative importance between colour similarity and spatial proximity. The greater the value of \( m \), the more spatial proximity is emphasized and the more compact the region. The range of \( m \) is \([1, 40]\) and \( m \) is set to 10 for all experiments.

Next, after each pixel has been associated to the nearest cluster centre, the new cluster centre is computed by the mean \( \frac{1}{S} \) vector of all the pixels belonging to the cluster region. And the residual error \( E \) measured by \( L_2 \) norm is computed between the new centre and the previous centre.

We iteratively repeat the process of associating each pixel with the cluster centre and recomputing the new cluster centre until the residual error converges. We set the number of iterations is 10 which is appropriate for most images.

After the clustering step, a post-processing step is adopted by relabeling disjoint region with the labels of the largest neighbouring cluster using a connected components algorithm as depicted in [10]. We illustrate the effect of generating homogeneous regions in Figure 2.

![Figure 2](image.png)

Figure 2. (a) and (b) are the visible image and the nir image; (c) and (d) are the images segmented using SLIC method; (e) and (f) are the images segmented using the proposed method.

2.3. Measuring Region Based Saliency

In this work, we define the saliency measure on the region level and then combine the saliency value with centre prior to get a pixel-accurate saliency map. Human are prone to pay more attention to meaningful regions instead of single pixels and they often concentrate the objects of interest near the centre of the image. So the saliency method based on region and centre prior is consistent with the characteristics of the human vision and it is robust to image noises.

When the labelling of the homogeneous regions is obtained, for each region, we calculate its saliency by measuring its colour contrast and spatial distance to other regions. The spatial weighting term is incorporated to emphasize the contrast of the region and its closer surrounding regions, the father regions in image plane are less considered.

The saliency of a region is defined as follows:

\[ S(r_k) = \sum_{i \in k} w(r_i, r_k) d_c(r_i, r_k), \]  
\[ d_c(r_i, r_k) = \max \left(||a_i - a_k||, ||b_i - b_k||, ||g_i - g_k||\right), \]  
\[ w(r_i, r_k) = e^{-\frac{d_c(r_i, r_k)}{\sigma^2}}, \]

where \( S(r_k) \) is the saliency value of the region \( r_k \). \( d_c(r_i, r_k) \) represents the colour contrast which is calculated by the formula (5). We believe that there exists an optimal colour component representation.
which would benefit measuring the difference of the two regions. We normalize the \( l, a, b, g \) components to \([0, 1]\) and compute mean colour vector \( \left[ l, a, b, g \right]^T \) is computed for each region, then select the most discriminate colour component by formula (5). \( w(r_i, r_k) \) represents the spatial weighting term, which is calculated by the spatial distance in the \( xy \) plane. \( d_s\left( r_i, r_k \right) \) represents the spatial distance between the region centre position vector \( \left[ x, y \right]^T \) of the region \( r_i \) and \( r_k \), which is calculated by the formula (2). And \( d_s\left( r_i, r_k \right) \) is normalized to \([0, 1]\). \( \sigma^2 \) is used to control the strength of spatial weighting. The greater the value of \( \sigma^2 \), the more the effect of the farther regions is emphasized. \( \sigma^2 \) is set to 0.4 for all experiments.

We measure \( S\left( r_i \right) \) for each region, when \( S\left( r_i \right) \) is obtained, the value is assigned to each pixel of the region \( r_i \) and a pixel-accurate saliency map \( S(r) \) is obtained.

As a consequence we further incorporate a centre prior to estimate our final saliency map. A two dimensional anisotropic Gaussian functions is employed to describe the centre prior knowledge which is defined as follows.

\[
S_G = \exp\left\{-\left[ \frac{(x_c - x_0)^2}{2\sigma_x^2} + \frac{(y_c - y_0)^2}{2\sigma_y^2} \right] \right\},
\]

where \( (x_c, y_c) \) is the centre of the image, \( (x_0, y_0) \) is the pixel coordinate in image plane, \( \sigma_x \) and \( \sigma_y \) are variants along the two directions respectively, and \( \sigma_x \) is set to 0.5\( W \) and \( \sigma_y \) is set to 0.5\( H \) where \( W \) and \( H \) are the width and height of the image respectively.

Therefore, the final saliency map can be obtained as

\[
S = S_G S\left( r \right).
\]

We compare the performance with and without the spatial weighting term and the centre prior as illustrated in Figure 3.

![Figure 3](image)

(a) and (b) are the homogeneous regions representation in Figure 2; (c) and (d) are without and with the spatial weighting term separately; (e) is the centre prior; (f) is the final saliency map

3. Experiments

To validate the proposed method in the salient region detection task, we select the publicly RGB-NIR dataset provided by Brown et al [9]. We choose 18 pairs of images from the dataset. Each pair contains a visible image in RGB format and a near-infrared image in grey format. And we manually label the ground truth image.

We compare the proposed region based multi-spectral saliency method (RMSS) with five start-of-the-art methods which are simply depicted in the section 2. Hereby they are respectively referred to as CA (Gorferman et al [3]), FS (Li et al [4]), RC (Cheng et al [5]), MSSS (Achanta et al [6]), RCSS (Vikram et al [7]). These methods are implemented with the author’s original codes and our method is implemented in Matlab.
3.1. Visual Comparison of Saliency Maps

We select five typical images to illustrate the visual effects. Figure 4 shows the visual comparison of saliency maps of the various methods. It can be observed in Figure 4 that the performance of each method is different. CA method highlights the edges of the salient region. FS method can detect the salient region effectively, but it cannot uniformly highlight the total salient region. MSSS method is less effective than the others. RC method, RCSS method and the proposed RMSS method can highlight the whole salient region. Almost all of the effects are poorer than the others images due to the cluttered background as displayed in the column 5.

![Figure 4: Visual comparison of saliency maps for state-of-the-art methods. Row 1: the original visible image; row 2: the ground truth; row 3–8: results on CA, FS, MSSS, RCSS and RMSS](image)

3.2. Quantitative Evaluation

To quantitative comparison, we further adopt the metrics of the recall, precision, F-measure performance evaluation, and the experiment results are depicted in Figure 5.

The performance evaluation [11] of $P$ (Precision), $R$ (Recall), $F$ (F-measure) are computed as in formula (9) – formula (11).
\[ P = \frac{\sum(B \times A)}{\sum(B)} , \quad (9) \]

\[ R = \frac{\sum(B \times A)}{\sum(A)} , \quad (10) \]

\[ F = \left(1 + \beta^2\right) \frac{P \cdot R}{\beta^2 \cdot P + R} , \quad (11) \]

where \( B \) represents the segment result of the saliency map by threshold, \( A \) is the ground truth by manually annotation, \( \beta^2 \) is set to 0.3 which indicates the F-measure value has more emphasis on accuracy.

We perform two different experiments. In both cases we generate a binary saliency map based on some saliency threshold. In the first experiment we compare binary masks for every threshold in the range \([0, 255]\). The step of the threshold is set to 10. The resulting Precision-Recall curves in Figure 5 show that our algorithm (RMSS) consistently produces results closer to ground truth at every threshold and for any given recall rate. In the second experiment we use the adaptive threshold which is computed as formula (11).

The results are observed in Figure 5. The larger the \( P \) value and \( F \) value, the better the effect of the method.

\[ T = \frac{2}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} S(x, y) , \quad (12) \]

where \( W \) and \( H \) are the width and the height of the saliency map \( S \).

From the quantitative evaluation in Figure 5, we can see that the RMSS method can obtain better results than the other comparable start-of-the-art methods.

### 3.3. Discussion

The parameter \( K \) is the only parameter of the proposed method. So we adapt one typical image to illustrate the effect of the number of superpixels. Figure 6 shows the visual effect which is computed by different \( K \) from 100 to 1000. And Figure 7 shows the precision, recall, and F-measure values which is computed by adaptive threshold.

![Figure 5: Left: Precision-Recall curves, Right: Precision, Recall, F-measure – adaptive threshold](image)

![Figure 6: Visual comparison of saliency maps for the different number of superpixels](image)
From Figure 6 and Figure 7, we can see that the proposed method performs robustly over a wide range of the number of the super-pixels. We evaluate the runtime of the proposed method by computing the average processing time with reference to CA method, FS method and the matlab version of the RC method. The CA method is time-consuming due to the multi-scale computation. Our running time is similar to RC method because both of them adapt the image clustering processing.

As depicted in the column 5 of the Figure 4, our method also cannot obtain satisfied saliency maps because the difference between the object and the cluttered background is small. The concept of the generic object [12] may be able to solve this problem, so in the future we plan to consider the generic object detection strategy to improve the effect.

4. Conclusions

We present a new region based multi-spectral saliency detection method. The algorithm incorporates the near-infrared and the visible information into the procedure of saliency detection. The superpixel segmentation algorithm is employed to generate the homogeneous regions, and the saliency map is measured by calculating the region’s contrast. We then compare the proposed algorithm with other five state-of-the-art methods and experimental results demonstrate that our algorithm performs better. Its shortcomings is that the performance get worse in the images which have the cluttered background, so in the future we plan to consider the generic object concept which can be combined with the saliency detection method to improve the saliency detection effect.

Acknowledgements

This Project is supported by the National Natural Science Foundation of China (Grant No. 61203266).

References


Received on the 21st of October 2013