Image and Vision Computing 35 (2015) 1–13

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Multi-scale hybrid saliency analysis for region of interest detection in very high resolution remote sensing images☆

Libao Zhang *, Bingchang Qiu, Xianchuan Yu, Jindong Xu

College of Information Science and Technology, Beijing Normal University, No. 19 Xinjiekouwai Street, Haidian District, Beijing 100875, China

ARTICLE INFO

Article history:
Received 11 October 2013
Received in revised form 21 September 2014
Accepted 22 December 2014
Available online 29 December 2014

Keywords:
Computer vision
Remote sensing image processing
Region of interest
Visual attention
Saliency analysis

ABSTRACT

Researchers have recently been performing region of interest detection in such applications as object recognition, object segmentation, and adaptive coding. In this paper, a novel region of interest detection model that is based on visually salient regions is introduced by utilizing the frequency and space domain features in very high resolution remote sensing images. First, the frequency domain features that are based on a multi-scale spectrum residual algorithm are extracted to yield intensity features. Next, we extract the color and orientation features by generating space dynamic pyramids. Then, spectral features are obtained by analyzing spectral information content. In addition, a multi-scale feature fusion method is proposed to generate a saliency map. Finally, the detected visual saliency regions are described using adaptive threshold segmentation. Compared with existing models, our model eliminates the background information effectively and highlights the salient detected results with well-defined boundaries and shapes. Moreover, an experimental evaluation indicates promising results from our model with respect to the accuracy of detection results.

1. Introduction

In recent years, the spatial resolution of remote sensing images has increased greatly. There has been considerably more information in such remote sensing images compared with the previously used low spatial resolution remote sensing image; this change presents a great challenge to the analysis and processing of very high resolution (VHR) remote sensing images [1–4]. Compared with the traditional low spatial resolution remote sensing images, the VHR images contain complicated spatial information, clear details, and well-defined geography objects. Moreover, the structure, edge, and texture information in a VHR remote sensing image is abundant and clear. In summary, the background information in a VHR remote sensing image becomes much more complex. Thus, more efficient information processing technology for VHR remote sensing images is required.

Among various applications of remote sensing image processing technology, object recognition is one of the most popular. Usually, the first step toward object recognition is object detection, which aims at extracting an object from its background before recognition. However, before performing the feature analysis, how can a machine vision system extract the saliency regions from an unknown background? Traditional approaches are to convert this problem to the detection of specific categories of objects and to use statistical pattern recognition methods that are based on features such as the spectrum and texture. Most of these methods, called top-down approaches [5,6], require a prior knowledge library that is difficult to build and has a large influence on the detection result, and the expansibility becomes the bottleneck in generalized tasks. Moreover, an essential part of these methods is global searching, which is both time-consuming and storage expensive. Thus, the region of interest detection method based on the bottom-up visual attention model, which has the least reference on statistical knowledge of the object in remote sensing images, should be implemented.

Currently, the study of the human visual system (HSV) has become an important trend. Visual attention is one of the primary features of HVS to derive important and compact information from a scene [7–10]. Because the surrounding environment includes an excessive amount of information, the visual attention mechanism enables a reduction in the redundant data. The content that draws human beings' attention has a characteristic called visual saliency, which means to stand out from the surroundings. A point that draws a person's attention is called a focus of attention (FOA), and the region centered by the FOA is called a visually salient region. In recent years, many studies have attempted to build computational models to simulate visual attention, and many bottom-up visual attention models have been proposed [11–16].

One of the earliest bottom-up visual attention models was proposed by Itti et al. [6,11,12]. Itti [12] constructed this model by using a biologically plausible architecture, which was proposed by Koch and Ullman [7] and is the basis for several models [13–15]. Itti's model obtains the saliency map based on the intensity, color, and orientation conspicuity maps. These conspicuity maps are attained by the across-scale addition

☆ This paper has been recommended for acceptance by Nicu Sebe.
* Corresponding author. Tel.: +86 13366069027.
E-mail address: libaozhang@bnu.edu.cn (L. Zhang).

http://dx.doi.org/10.1016/j.imavis.2014.12.002
0262-8856/© 2014 Elsevier B.V. All rights reserved.

of feature maps, while the feature maps capture the center-surround differences between various Gaussian pyramids and oriented pyramid scales [12]. Because the saliency map is computed over coarser scales, local information loss is unavoidable in this algorithm. This model attempts to simulate the visual attention mechanism based on the HVS biological vision principles, has had a significant influence on the study of the visual attention mechanism, and has been improved since its proposal. Inspired by Itti's method, Frintrop et al. [16] present a method, use integral images to speed up the calculations, and compute center-surround differences with square filters.

After the introduction of Itti's model, other approaches, which are purely computational and are not based on biological vision principles, are proposed. Achanta et al. [17] attempt to build visual attention models by accounting for color contrast information. They first obtained the Gaussian-filtered image and then transformed the input images from RGB color space to CIE Lab color space. The CIE Lab color space is used for each image location to form a feature vector, and then, the absolute difference between the Gaussian-blurred image and the arithmetic mean vector is calculated to obtain the saliency map. This method generates full-resolution saliency maps, but it works well only on images that have large and homogeneous objects that have clear boundaries. Hence, because of its dependency on the object size and the uniformity, it has limitations with respect to the applications of remote sensing images.

Hou and Zhang [18] attempt to obtain a saliency map for images in the transform domain. They first obtained a down-sampled image with a height or width equal to 64 pixels, and then, they performed Fourier Transform (FT) of the down-sampled image to obtain the phase and amplitude spectrums. By analyzing the log-spectrum of the images, they extracted the spectral residuals of the down-sampled image in the spectral domain. The saliency map is derived by applying an inverse FT on an exponential function that combines the spectral residual and phase spectrum information. This method can obtain fast saliency detection, but the saliency map has very low resolution, and much detailed information is unavoidably lost.

Perazzi et al. [19] proposed a saliency filter method, which obtained perfect results in natural images. It decomposes a given image into compact, perceptually homogeneous elements that abstract unnecessary detail. Based on this abstraction it computes two measures of contrast that rate the uniqueness and the spatial distribution of these elements. From the element contrast it then derives a saliency measure that produces a pixel-accurate saliency map.

Another bottom-up visual saliency model, Graph-Based Visual Saliency (GBVS), was proposed by Harel et al. [20]. This method uses a novel application of ideas from graph theory to concentrate mass on activation maps and to form activation maps from raw features. The saliency map yielded by GBVS also has low resolution, and some spatial information is lost. As an improvement of Harel's model, Sun et al. [21] combined edge-based and Graph-Based Visual Saliency computation methods, and obtain better results in VHR remote sensing images. In addition, to such models, significant advancements have also been made in automatic salient object detection [22–27], and the visual saliency model is also applied to image and video compression [28,29].

In these methods, the detection results of the visually salient region have low resolution, poorly defined borders, or are sensitive to the object size and uniformity. The VHR remote sensing images contain complicated spatial information and clear details, and the structure, edge, and texture information is abundant and clear. Although the good performance of these methods has been proved in natural images, their limitations will be magnified in VHR images. Even some approaches have been trying to be applied in remote sensing images, like Sun's model [21]. However, current methods of saliency detection fail to address the complex background information that is found in VHR images.

The focus of this paper is on the automatically detected and described region of interest for VHR remote sensing images based on visually salient analysis. Thus, the salient regions in VHR remote sensing images should be detected effectively and described accurately. We introduce a new model that offers two advantages over existing methods: eliminate the background information effectively and highlighting the salient detected results with well-defined boundaries and shapes. In our model, a ROI extraction model for VHR remote sensing images is presented, which allows the input image to be processed along four feature channels and involves saliency analysis that combines spatial and frequency analysis.

In this model, the intensity, color, orientation, and spectral features are extracted to compute the saliency map. We first extract the intensity and color information from the input remote sensing images. Next, the frequency domain features based on a multi-scale spectrum residual algorithm are extracted to yield intensity features. We subsequently extract the color and orientation features by generating space dynamic pyramids. Then, spectral features are obtained by analyzing spectral information content. In addition, a multi-scale feature fusion method based on a weighted across-scale combination strategy is proposed to generate the saliency map. Finally, the detected visual saliency regions are described using adaptive threshold segmentation. Experimental evaluation depicts the promising results from our model in the accuracy of the detection results.

2. Background of the visual attention models

2.1. Itti's model

In this model, the early visual features are extracted from the input image, including the intensity, color and orientation. For the color, red/green, green/red, blue/yellow, and yellow/blue are color pairs that exist in the human visual cortex. Orientation features are obtained by using Gabor pyramids. A linear "center-surround difference" operation, including interpolating the coarser aspects to a finer scale and point-by-point subtraction, is used between different levels of the pyramid to compute the multi-scale feature maps.

The feature maps are globally prompted using the normalization operation $N(\bullet)$ and are added by using the across-scale combination operation "$\oplus$". This operation includes reducing each map to scale four and point-by-point addition to generate the conspicuity maps, $\bar{I}$ for intensity, $\overline{C}$ for color and $\bar{O}$ for orientation:

$$\bar{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c,s)) \tag{1}$$

$$\overline{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} (N(RG(c,s)) + N(BY(c,s))) \tag{2}$$

$$\overline{O} = \sum_{\theta \in \{0^0, 45^0, 90^0, 135^0\}} N\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c,s,\theta)) \right). \tag{3}$$

The final saliency map is calculated by normalizing and adding the three conspicuity maps, as follows:

$$S = \frac{1}{3}\left(N(\bar{I}) + N\left(\overline{C}\right) + N\left(\overline{O}\right)\right). \tag{4}$$

2.2. Achanta's model

In 2009, Achanta et al. presented a frequency-tuned approach of computing saliency in images by using the low level features of color and luminance; this method is easy to implement, is fast, and provides full resolution saliency maps. In the Achanta et al. model, the original image is first transformed to the CIE Lab color space, and each pixel