Contents lists available at ScienceDirect

Neurocomputing

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

Saliency detection based on global and local short-term sparse representation

Qiang Fan, Chun Qi*

School of Electronic and Information Engineering, Xi'an Jiaotong University, Xian 710049, China

A R T I C L E I N F O

Article history:
Received 9 October 2013
Received in revised form
24 September 2015
Accepted 11 October 2015
Communicated by Bijaya Ketan Panigrahi
Available online 24 October 2015

Keywords:
Saliency detection
Sparse representation
ICA algorithm
Low-Rank Representation
Global saliency
Local saliency

A B S T R A C T

Saliency detection has been considered to be an important issue in many computer vision tasks. In this paper, we propose a novel bottom-up saliency detection method based on sparse representation. Saliency detection includes two elements: image representation and saliency measurement. For an input image, first, the ICA algorithm is employed to learn a set of basis functions, then the image can be represented by this set of basis functions. Next, a global and local saliency framework is employed to measure the saliency. The global saliency is obtained through Low-Rank Representation (LRR), and the local saliency is obtained through a sparse coding scheme. The proposed method is compared with six state-of-the-art methods on two popular human eye fixation datasets, the experimental results indicate the accuracy of the proposed method to predict the human eye fixations.

*Corresponding author.

1. Introduction

A human vision system has to process an enormous amount of incoming information from retina every moment. The system selects and focuses on important regions instead of processing all this information at the same time. We call such important regions *Visual saliency*. Saliency detection is an easy task for human beings while it is hard for many computer vision systems. However, the detection of visual saliency is very important in many computer vision tasks [1–4]. Recently, modeling visual saliency has become a hot research topic, a lot of theories and applications have been raised.

In terms of application, the saliency detection algorithms mainly fall into three groups. Some algorithms focus on identifying the fixation points that a human viewer would focus on at the first glance [5–10]. This type of saliency helps us better understand human attention, which leads to the applications such as auto focusing and adaptive region-of-interest based image compression [11]. Others have concentrated on detecting a salient object as a whole, it has been used for object segmentation [2,3], object recognition [4] and content-aware image resizing [1]. In addition, some researchers pay attention to modeling a saliency map that the context of the salient object is as important as the salient object, examples include summarization of a photo collection [12] and image thumbnailing [13]. In this paper, we focus on predicting the human fixations.

According to whether the method requires human supervision, the saliency detection methods can be divided into two categories, bottom-up, data driven, task independent unsupervised saliency detection and top-down, goal driven, task dependent supervised saliency detection. The proposed method is under the first setting.

Two important issues of modeling a saliency detection problem are image representation and saliency measurement. Researchers have proposed several feature descriptors, such as color, intensity and orientation [5,14]. Different saliency measures have used one or several features to detect the saliency. However, using just one or several features will not represent the image comprehensively, and always leads to loss of image information. Some researchers have found an evidence that the receptive fields of simple cells in the primary visual cortex produce a sparse representation [15], which motivated lots of works to represent natural images by sparse features [6,10,16]. Most of these methods use Independent Component Analysis (ICA) to generate the basis functions which are learned from a large number of image patches sampled from different kinds of natural images. However, as shown in [17], these basis functions cannot provide a perfect representation for every input image without information loss. This is because some features of the input image cannot be captured by the predefined basis functions, and the information loss is mainly occurred in the salient regions, this influences the detection of saliency.

As for saliency measurement, we find that previous models are mainly fall into two groups: global models and local models.

* Corresponding author.
E-mail address: qichun@mail.xjtu.edu.cn (C. Qi).

http://dx.doi.org/10.1016/j.neucom.2015.10.030
0925-2312/© 2015 Elsevier B.V. All rights reserved.

Global models find salient regions by calculating rarity of features over the entire scene (e.g., AIM [6], ICM [10]). Local models find salient regions locally by calculating the rarity of features in a center-surround local region (e.g., Itti [5], GBVS [7]). Local high contrast will be found by the local measure, but these contrasts may not be salient globally. Global rarity features will be found by the global measure, but the local high contrast may be overlooked. However, the human visual system follows a center-surround approach in the early visual cortex and the human visual system is sensitive to the local high contrast [18]. Thus using either global or local measure will not be very reasonable.

The above problems encourage us to seek other better image representation and saliency measurement methods. In this paper, we use "short-term image representation" proposed by Sun et al. [17] recently as the image representation method to transform the image from color space into coefficient space. For each input image, a set of basis functions is learned from the image patches sampled from this image using the ICA algorithm. Then each image patch can be represented by the linear combination of this set of basis functions without information loss [17]. This method can overcome the information loss problem of the above-mentioned methods. We use a framework that combines global and local saliency measure as the saliency measure, which means the global and local saliency is first computed respectively, and then combined to get the final saliency. The global saliency is obtained through Low-Rank Representation (LRR). First, LLR decomposes the coefficient matrix into a low-rank part and a sparse part, which represents image information of regularity and singularity, respectively. Then the saliency can be accessed from the sparse part. A sparse coding scheme is used to measure the local saliency. First, it represents the center patch with the surrounding patches. The goal of sparse coding is to find the optimal balance between loss and sparseness, after solving a Lasso problem, we can get the representation coefficient for each patch. Then the local saliency is modeled as the product of sparse coding length and representation residual. The process of the proposed method can be seen in Fig. 1.

In summary, the contribution of this paper mainly lies in that, we propose a novel framework that is the first to combine the better image representation measure and the better saliency measure. More specifically, the "short-term image representation" and the global and local saliency measure, which are more accurate for saliency detection problem.

The remainder of the paper is organized as follows. Related works are reviewed in Section 2. The proposed scheme is introduced in Section 3. Some experimental results and comparisons are presented in Section 4. In Section 5, conclusions and some future works are demonstrated.

2. Related work

In recent years, a lot of saliency detection methods have been proposed, these methods mainly fall into three groups: biologically structure motivated, mathematical motivated, and those motivated by both aspects.

Itti et al. [5] propose a framework based on Koch and Ullman's biologically plausible architecture [19]. They compute saliency in three channels: color, intensity, and orientation. For each channel, the saliency is obtained by center-surround contrast using a Difference of Gaussians (DoG) approach across different scales. Then these three channels' saliency maps are combined to get the final saliency map. Le Meur et al. [20] adapted the Koch–Ullmans model to include features of contrast sensitivity, perceptual decomposition, visual masking, and center-surround interactions. Some models have added features such as symmetry [21], texture contrast [22], curvedness [23], or motion [24] to the basic structure.

The other type is purely mathematically motivated, Ma et al. [25] propose a local contrast-based method, the saliency map is obtained by summing up the differences of image pixels with their respective surrounding pixels in a small neighborhood under just one scale. Hou and Zhang [8] compute saliency by extracting spectral residual in the amplitude spectrum of Fourier transform. Not long after Hou's work, Guo et al. [9] argue that not the amplitude, but the phase spectrum of an image is the key to obtain the salient regions, and propose the method PFT. They also extend PFT from a two-dimensional Fourier to a Quaternion Fourier Transform, and propose the method QFT. Zhai et al. [26] define the saliency of a pixel due to its contrast to all the other pixels in the image, however, they use only the luminance channel, therefore other channels' information has been ignored. However, many of these methods suffer from the problems that the objects' borders are often assigned with higher saliency than the salient regions and the objects' borders cannot be well protected in the saliency map. In order to overcome such problems, many image segmentation based methods like [27] have been proposed. However, these methods heavily rely on the performance of image segmentation, which itself is a challenging problem.

Some other methods are motivated by both biological structure and mathematical computation. Bruce et al. [6] base their model on sparse coding representation and the principle of information maximization. They use self-information of the sparse coefficient as the saliency. Itti and Baldi [28] defined surprising stimuli as those which significantly change beliefs of an observer, measured as the Kullback–Leibler (KL) distance between posterior and prior beliefs. Hou and Zhang [10] introduce the Incremental Coding Length (ICL) to measure the perspective energy gain of each feature, the objective of their model is to maximize energy consumption, the features they used are learned by the ICA based method from natural image patches, which are biologically

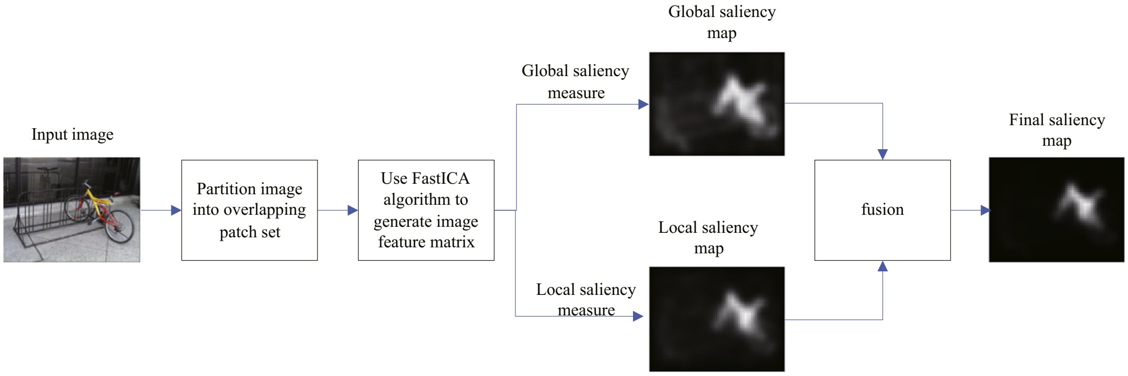

Fig. 1. The framework of the proposed method.