Pattern Recognition 48 (2015) 1337–1348

Contents lists available at ScienceDirect

Pattern Recognition

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

Nonlocal center–surround reconstruction-based bottom-up saliency estimation

Chen Xia, Fei Qi*, Guangming Shi, Pengjin Wang

State Key Laboratory of Integrated Services Networks, Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, School of Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, PR China

ARTICLE INFO

Article history:
Received 18 September 2013
Received in revised form
17 May 2014
Accepted 2 October 2014
Available online 23 October 2014

Keywords:
Saliency
Compressed sensing
Nonlocal means
Exemplar-based image processing

ABSTRACT

Many saliency models consider the feature extraction as the algorithmic core and the performance of their methods relies on the selection of the features to a great extent. However, there can hardly be a set of features effective to pop out the salient regions under various visual environments. Moreover, because saliency is not tuned to certain visual features, a location winning the spatial competition in any feature space can be defined as salient. Instead of seeking for or learning the features to highlight the difference between the salient areas and the background, we focus more on the sparsity and uniqueness carried by the original image itself, the source of all the features, to propose a nonlocal reconstruction-based saliency model. In the proposed approach, the saliency is measured by the sparse reconstruction residual of representing the central patch with a linear combination of its surrounding patches sampled in a nonlocal manner. In addition, this is generalized to model the global aspect saliency, which provides a complement to the nonlocal saliency and improves the performance further. As a generalization of Itti et al.'s classical center–surround comparison scheme, the proposed approach performs well on images where Itti et al.'s method fails, as well as on general natural images. Numerical experiments show the proposed approach produces better results compared with the state-of-the-art algorithms on three public databases.

1. Introduction

During the long-term evolutionary process, the human visual system (HVS) has been endowed with a remarkable ability to detect possible objects and seize the most significant contents in the complicated and changeable visual world with limited time [1–3]. Based on the capacity of the HVS to process mass data in real-time, the concept of saliency emerged and provided a way to solve the problems in a large range of fields with analogous demands, such as image and video compression [4,5], object detection [6,7], object recognition [8,9], image segmentation [10], image retrieval, and image retargeting [11]. In addition, the research of saliency also takes an important part in the development of multiple disciplines including neuroscience [12], cognitive psychology [13] and computer vision [14,15].

Among many models of saliency, feature selection has been extensively studied. Different features can highlight and characterize the salient regions from distinct aspects. To handle saliency estimation under more visual scenes, a common way is to increase the number of features. However, high-dimensional data will result in some problems from the computational standpoint, such as the curse of dimensionality. Moreover, according to the knowledge of information theory, the features cannot provide more data than the original image itself, the source of all the features. Therefore, how to utilize the sparsity and uniqueness carried by the original image to develop a saliency measure method without special requirements of features has become a problem to be solved.

The center–surround (C–S) hypothesis, which attests a region is salient if it is distinct from its surrounding region, is widely accepted in bottom-up saliency estimation [2,1,16]. Based on this hypothesis, Itti et al. [1,2] proposed to perform C–S difference operators and spatial competition on features like color, intensity, orientation, and motion. Saliency maps on various features are then combined to derive the overall saliency. However, Itti et al.'s method is based on local image processing techniques which are hard to handle texture regions well. As shown in Fig. 1, Itti et al.'s method fails on the images with texture structures. Furthermore, we do not know which feature space is complete when applying Itti et al.'s C–S operators.

Guided by the findings that textures imply the mutual relationship of pixels, Efros and Leung created a texture synthesis method by comparing context regions around the pixels rather than the pixels themselves in a large portion of the image [17]. Originating from this idea, the nonlocal means filter was proposed and

* Corresponding author. Tel.: +86 29 88202265; fax: +86 29 88204453.
E-mail addresses: cxia@stu.xidian.edu.cn (C. Xia), fred.qi@ieee.org (F. Qi), gmshi@xidian.edu.cn (G. Shi), wpjsolo@163.com (P. Wang).

http://dx.doi.org/10.1016/j.patcog.2014.10.007
0031-3203/

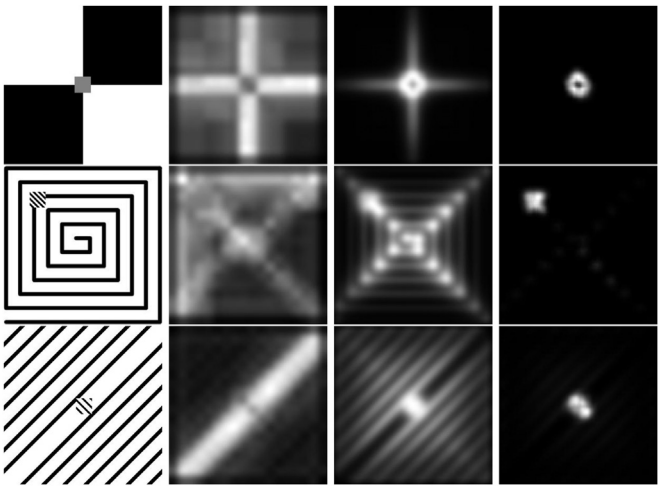

Fig. 1. A comparison with Itti et al.'s classical C–S saliency model. From left to right, columns are original images, results of Itti et al.'s [1], Wu et al.'s [19] and the proposed approaches, respectively.

achieved a better preservation of the image texture information during the denoising process [18].

To handle texture regions in saliency computation, Wu et al. proposed a nonlocal redundancy reduction approach [19]. This approach is independent of certain features and performs well on several databases. However, because of their formulation of the redundancy coefficient as the simple sum of similarities between the central patch and all others, all the patches in the nonlocal region will have influence on the central patch without a unified manner. Hence, as shown in Fig. 1, in their method, some edges of the background will also be highlighted as salient parts.

Inspired by the development of the nonlocal image processing technique [18,19] and the successful application of the compressed sensing [20] in face recognition [21,22], we propose a novel nonlocal reconstruction-based C–S comparison model. Our idea is to represent the central patch with a linear combination of its surrounding patches in the nonlocal neighborhood. The reconstruction residual indicates the degree how the central patch can be represented by its surrounding patches in the nonlocal area. Hence, in the proposed method, the smaller the residual is, the smaller/less the difference/saliency is.

In our approach, the most salient region corresponds to the region with the largest residual, where the center and the surround share the fewest attributes of the same class and can be distinguished into distinct classes with the largest discriminant power. Thus, the proposed model is intrinsically relevant to a recognition problem [21] and essentially consistent with Gao and Vasconcelos' opinion that "bottom-up saliency is a discriminant process" [16].

The main differences of the proposed approach compared against existing approaches are twofold. On one hand, the image textures are processed based on patches, which are in a high dimensional space where the sparsity must be exploited. On the other hand, the saliency is estimated according to the reconstruction residual instead of selecting, combining or learning features. It should be noted that this paper is an extension of our previous work [23]. The main improvements include two aspects. One is the integration of global saliency and the other is that more experiments are conducted to evaluate the model comprehensively.

The rest of this paper is organized as follows. Section 2 reviews the related works. Section 3 introduces the proposed reconstruction-based saliency. In Section 4, we compare the performance of the proposed with the existing state-of-the-art saliency estimation methods. Discussions and conclusions are given in Section 5 and 6, respectively.

2. Related works

One of the main expectations and targets in saliency research is to understand where we look or pay attention in the given scenes. To this end, models have been proposed from various perspectives and can be broadly classified into two classes according to their computational measures of saliency. On one hand, biologically inspired methods usually take inspiration from the neurobiological or psychophysical findings in the HVS to model saliency. For example, Itti et al. evaluated the saliency with the center–surround difference of distinct features, which is based on the architecture of typical visual neurons [1]. Ma and Zhang used a fuzzy growing method, an efficient way to imitate human mental behaviors, to extract attended regions from the saliency maps derived by a local contrast method [24]. Although almost all the saliency estimation models are inspired by the biological concepts more or less, biologically inspired methods have more direct and closer correlation with the mechanisms of the HVS.

On the other hand, computationally oriented methods concentrate more on the common attributes of the salient regions and utilize a computational and mathematical framework to formulate them. Hou and Zhang found the similar trends of the log-spectra shared by different images and detected the unpredictable regions in the images by spectral residual (SR) [25]. Different from the method of SR, Guo et al. reconstructed the image with the phase spectrum alone to pop out the regions with less periodicity and homogeneity [26]. Harel et al. represented the image by a graph model and treated the saliency estimation problem as getting the equilibrium distribution of a Markov chain to assign high saliency values to the pixels with high dissimilarity to the surroundings [27]. During the past decade, information theory also participated in the development of saliency detection. Bruce and Tsotsos proposed an information maximization model to calculate the saliency with Shannon's self-information [28]. Itti and Baldi measured the saliency as the Kullback–Leibler Divergence (KLD) between the posterior and prior distributions [3]. Based on their work, Hou et al. employed KLD to compute information divergence on the coefficient maps obtained by independent component analysis (ICA) [29]. Klein and Frintrop also used KLD to determine the difference between the distributions of center and surround regions [30]. Gao and Vasconcelos presented a decision-theoretic approach based on mutual information between features and class labels [16]. Wu et al. proposed a saliency model based on redundancy reduction to compensate the Shannon entropy [19]. Rather than modeling the space-based saliency, Hou and Zhang presented a feature-based method according to the rarity of features which was computed by Incremental Coding Length [31]. Besides information theory, machine learning algorithms such as SVM and AdaBoost were also used to model the saliency by learning from recorded human fixations [15,32]. However, expansibility to generalized situations is still a bottleneck for training-based saliency models.

Recently, Ren et al. proposed to detect spatio-temporal saliency for videos based on reconstruction principle [33]. Though both their work and the proposed method are based on reconstruction, there are two main technical differences. On one hand, the focusing scopes are different: they focus on videos while we focus on static images. On the other hand, the actual reconstruction models are different: their model is an overdetermined optimization while ours is an underdetermined one where sparsity is the key to solving it. Due to the differences of models, as the experimental results reveal, Ren et al.'s method shows limited performance to predict human fixations on static images compared with the proposed method.

In recent years, a new trend called salient object detection has emerged and attracted much interest. Works belonging to this direction are aimed at detecting or segmenting the most salient objects in images rather than creating saliency maps to predict eye