Journal of Visual Communication and Image Representation 54 (2018) 1–9

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Salient object detection via a local and global method based on deep residual network☆

Dandan Zhu[a], Ye Luo[a,*], Lei Dai[b], Xuan Shao[a], Qiangqiang Zhou[c], Laurent Itti[d], Jianwei Lu[a,e,*]

[a] School of Software Engineering, Tongji University, China
[b] School of Automotive and Traffic Engineering, Jiangsu University, China
[c] School of Electronics and Information, Tongji University, China
[d] Dept. of Computer Science and Neuroscience Program, University of Southern California, United States
[e] Institute of Translational Medicine, Tongji University, China

ARTICLE INFO

Keywords:
Salient object detection
Deep residual network
Local and global features

ABSTRACT

Salient object detection is a fundamental problem in both pattern recognition and image processing tasks. Previous salient object detection algorithms usually involve various features based on priors/assumptions about the properties of the objects. Inspired by the effectiveness of recently developed deep feature learning, we propose a novel Salient Object Detection via a Local and Global method based on Deep Residual Network model (SOD-LGDRN) for saliency computation. In particular, we train a deep residual network (ResNet-G) to measure the prominence of the salient object globally and extract multiple level local features via another deep residual network (ResNet-L) to capture the local property of the salient object. The final saliency map is obtained by combining the local-level and global-level saliency via Bayesian fusion. Quantitative and qualitative experiments on six benchmark datasets demonstrate that our SOD-LGDRN method outperforms eight state-of-the-art methods in the salient object detection.

1. Introduction

Saliency detection attempts to identify the most important and conspicuous object regions in an image by the human visual and cognitive system. It is a fundamental problem in neural science, psychology and computer vision. Many computer vision researchers propose computational models to simulate the process of human visual attention or identify salient objects. Recently, salient object detection had drawn a large amount of attention in variety of computer vision tasks, such as object detection [1], person re-identification [2], object retargeting [3], image retrieval [4,5], video summarization [6] and image deblurring [7], etc.

Visual saliency can be viewed into different perspectives and contrast is one of them. Based on the observation that salient object is always distinguishing itself from its surroundings, contrast as a prior has been widely used to detect salient object. According to the range of the context that the contrast is computed to, it can be further categorized into local contrast and global contrast methods. The local contrast based methods usually compute center-surround difference to obtain the object-of-interest region standing out from their surroundings [8]. Due to the lack of the global information, methods of this category tend to highlight the boundaries of salient objects and neglect the interior content of the object. Meanwhile, the global contrast based methods take the entire image into consideration to estimate the saliency of every pixel or every image segment, thus the whole salient object is detected but the details of the object structure are always missing. There are also some methods proposed to improve the performance of salient object detection via integrating both local and global-level cues [9]. The aforementioned methods may work well for low-level saliency, but they are neither sufficient nor necessary, especially in the cases when the saliency is also related to the human perception or is task-dependent.

In order to obtain more accurate semantic features for salient object detection, deep neural networks have recently been widely used. These methods include Multisacle Deep Feature (MDF) [10], Deep Image Saliency Computing via Progressive Representation Learning (DISC) [11], Local Estimation and Global Search (LEGS) [12], and Encoded Low level Distance map (ELD) [13]. They utilized high-level features from the deep convolution neural network (CNN) and demonstrate superior results over previous works that utilizes only low-level features. CNN based methods did show their superiority on deep feature extraction, but compared to the methods using deep residual network

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding authors at: School of Software Engineering, Tongji University, China (Y. Luo).
E-mail addresses: yeluo@tongji.edu.cn (Y. Luo), jwlu33@tongji.edu.cn (J. Lu).

https://doi.org/10.1016/j.jvcir.2018.03.017
Received 27 August 2017; Received in revised form 23 January 2018; Accepted 28 March 2018
Available online 24 April 2018
1047-3203/

(ResNet), there are still some limitations. Specifically for MDF, the author treats each region as an independent unit in feature extraction without any shared computation. In addition, this method uses not only high-level features but also low-level features for complement and enhancement when doing salient detection. Due to the fact that CNNs fail to extract distinguishing features when the textures of salient object regions are similar to the background. Moreover, since CNN is generally performed on image patches, only features at local level are extracted, thus it fails to capture the global relationship of image regions and can't maintain the label consistency in a relative large region.

In order to solve these problems, in this paper, we propose a novel image salient object detection model named Salient Object Detection via a Local and Global Method Based on Deep Residual Network (i.e. SOD-LGDRN). Instead of CNN, we apply the deep residual network (ResNet) to salient object detection, from which more distinctive features of the salient objects can be obtained. Different from previous methods, high-level semantic features are extracted from both local and global levels via two ResNets, respectively. Features extracted from an entire image via a global ResNet (ResNet-G) roughly identify the global concepts of the salient object(e.g. the location, the scale and the size of the salient object). However, the residual network at the global level considers the entire image but pays less attention to the local context information, and this may misinterpret the background as salient regions or lack the subtle structure of a salient object. Hence, local features of the image are extracted via local ResNets (ResNet-L) as complements to the global information. By considering deep features at the local and the global levels simultaneously, we can expect to obtain a salient object with homogeneously highlighted region and accurate object boundary.

We briefly describe the implementation of our approach as shown in Fig. 1. At first, the entire image as input is sent to ResNet-G to extract global level features thus to generate a saliency map in a global context. Secondly, a multilevel image segmentation method is employed to segment the target image into multiple segments, and the segments at each level are warped and then fed to a ResNet-L to extract local features. The obtained local features at multiple level are further used to estimate the local saliency map. In order to deal with the noise caused by the image segmentation, we also introduce a spatial coherence refinement method to enhance the smoothness of the local saliency map. At last, the global and the local saliency maps are fused to obtain the final saliency map via a Bayesian integration method. In summary, our major contribution is threefold:

(1) A novel salient object detection model is proposed based on the global and the local semantic features extracted from two deep residual networks: ResNet-G and ResNet-L.
(2) Features learned from deep residual network are at the first time extracted and applied to salient object detection.
(3) We proposed the model only uses high-level features for saliency detection, and the performance can be significantly improved without using low-level features for complement or refinement.

The rest of the paper is organized as follows. In Section 2, we first review and evaluate the related work. Then we introduce our proposed SOD-LGDRN model in Section 3. In Section 4, we conduct experiments on six public datasets: MSRA-B, PASCAL-S and ECSSD, ASD, SED1 and THUR15K and then make comparisons with eight state-of-the-art methods. In the last Section, we present the conclusion and discussion of the future work.

2. Related work

In this section, we discuss the related work on salient object detection. In addition, we also briefly review deep neural network that are closely related to this work.

To estimate visual saliency, various methods are proposed and most of the saliency detection approaches can be generally categorized as global and local schemes. Local methods measure saliency by computing local contrast and rarity. Itti et al. [14] propose the center-surround scheme to extract low-level features such as color, intensity, orientation and texture, and use a linear and non-linear combination of multi-scale saliency map. Hereafter, Ma and Zhang [15] utilize color contrast in a local neighborhood as a measure of saliency. Goferman et al. [9] provide a context-aware (CA) method that used three principles including local low-level cues, global consideration and visual representation rules to highlight salient objects along with their contexts. However, these methods still have some problems to be solved, for example, they tend to be more sensitive to the boundaries of salient objects, but it can't highlight the object interior uniformly. On the other hand, global methods generally detect saliency by using holistic contrast and color statistics of the entire image. Achanta et al. [16] propose a frequency tuned method to calculate the image pixel saliency by subtracting the average color of the image. Cheng et al. [17] compute image saliency on the basis of color histogram contrast and region contrast. Xu et al. [18] utilize histograms based global contrast and spatial coherence to detect saliency. Most of the global contrast methods depend on color uniqueness when doing global statistics. However, these methods can be insufficient to capture semantic features in a natural image. Besides, they usually ignore the spatial relation (i.e. spatial weight and distance) between different parts in the images. Liu et al. [19] propose a set of features from both global and local perspectives, which are integrated by conditional random field (CRF) to generate final saliency map. Fareed et al. [20] proposed a salient region detection algorithm based on sparse representation and graph ranking, which combined the Gaussian and Bayesian procedures to produce smooth and precise saliency map. While the combination of global and

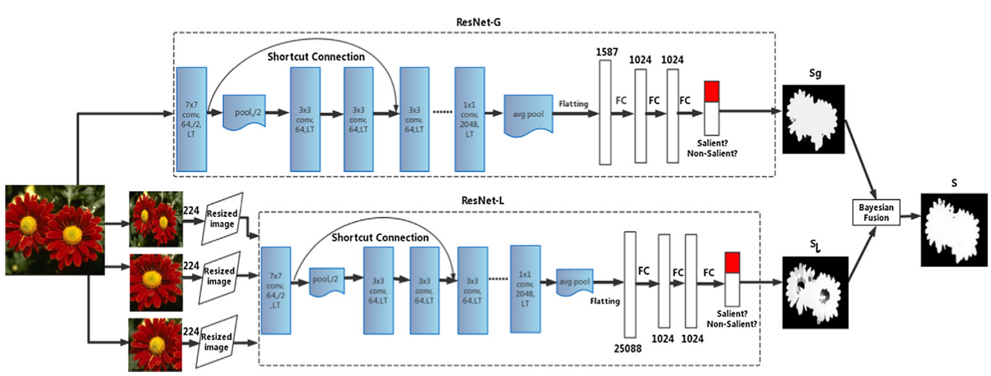

Fig. 1. Architecture overview of our proposed deep salient object detection model. The ResNet-G takes the entire image as input and generates global saliency map S_g. The ResNet-L takes image segments at multiple levels as input and produces the local-level saliency map S_l. The aggregated saliency map S is obtained by fusing S_g and S_l via a Bayesian integration method. LT denotes to the linear transformation operation.