Computer Vision and Image Understanding 118 (2014) 40–49

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Tag-Saliency: Combining bottom-up and top-down information for saliency detection

Guokang Zhu [a,b], Qi Wang [a], Yuan Yuan [a,*]

[a] Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, PR China
[b] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, PR China

ARTICLE INFO

Article history:
Received 14 September 2012
Accepted 4 July 2013
Available online 30 August 2013

Keywords:
Computer vision
Saliency detection
Visual attention
Image tagging
Visual media
Semantic

ABSTRACT

In the real world, people often have a habit tending to pay more attention to some things usually noteworthy, while ignore others. This phenomenon is associated with the top-down attention. Modeling this kind of attention has recently raised many interests in computer vision due to a wide range of practical applications. Majority of the existing models are based on eye-tracking or object detection. However, these methods may not apply to practical situations, because the eye movement data cannot be always recorded or there may be inscrutable objects to be handled in large-scale data sets. This paper proposes a Tag-Saliency model based on hierarchical image over-segmentation and auto-tagging, which can efficiently extract semantic information from large scale visual media data. Experimental results on a very challenging data set show that, the proposed Tag-Saliency model has the ability to locate the truly salient regions in a greater probability than other competitors.

1. Introduction

Visual attention is an important mechanism of the human visual system. It helps access the enormous amount of complex visual information from the world effectively through rapidly selecting the most prominent or highly relevant subjects. In computer vision, this mechanism is modeled as saliency detection, which can provide the computational identification of scene regions that are more attractive to human observers than their surroundings. Based on the detected results, a higher and more complex processing can focus only on the salient regions when there is large-scale visual media data to be handled.

It is believed that visual attention is driven by two independent factors: (1) a bottom-up component, which is a task-independent component purely based on the low-level information, and (2) a top-down component, which is based on high-level information and guides attention through the volitionally controlled mechanisms. In recent years, many bottom-up saliency detection methods have been designed because they can provide a lot of useful information without prior knowledge about the scene. This kind of methods has already achieved a laudable performance in practical multimedia applications. For example, they have been successfully used for object detection and recognition [1,2], image quality assessment [3,4], video summarization [5], image/video compression and resizing [6,7], adaptive content delivery [8], and image segmentation [9,10]. In the meantime, in contrast to the focused interest in modeling the bottom-up guided attention, few studies have attempted to explore top-down factors.

Evidence from visual cognition researches indicates that the low-level factors dominate the early visual stage. But late on, it is mainly the high-level factors that direct eye gazing and changing [11,12]. For instance, when looking at images in the first row of Fig. 1, people are usually attracted by some specific regions standing out from the rest of the scene with respect to color, intensity, or orientation during the first few hundreds of milliseconds. Then immediately they are able to allocate attention spotlight to the cars, texts, pedestrians, or other objects with special concepts or meanings, while are easy to ignore the things usually treated as background. Thus, top-down factors should also be taken into account separately in visual saliency detection, and it is reasonable to believe that efficient and effective utilization of high-level information can help improve current saliency detection performance.

Most studies on top-down attention are still at the descriptive and qualitative level. Few completely implemented computational models are available, and most of them are based on eye tracker or a series of specific object detectors. However, in practical situations, there may be no eye movement data or too many objects to be handled in large-scale data sets. These methods therefore will be greatly limited by the harsh conditions of use and the high computational complexities.

In fact, there are many other techniques in computer vision besides eye-tracking and object detection, which can help automatic extraction of high-level information. Inspired by the works of

* Corresponding author. Tel.: +86 29 88889302.
E-mail address: yuany@opt.ac.cn (Y. Yuan).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.07.011

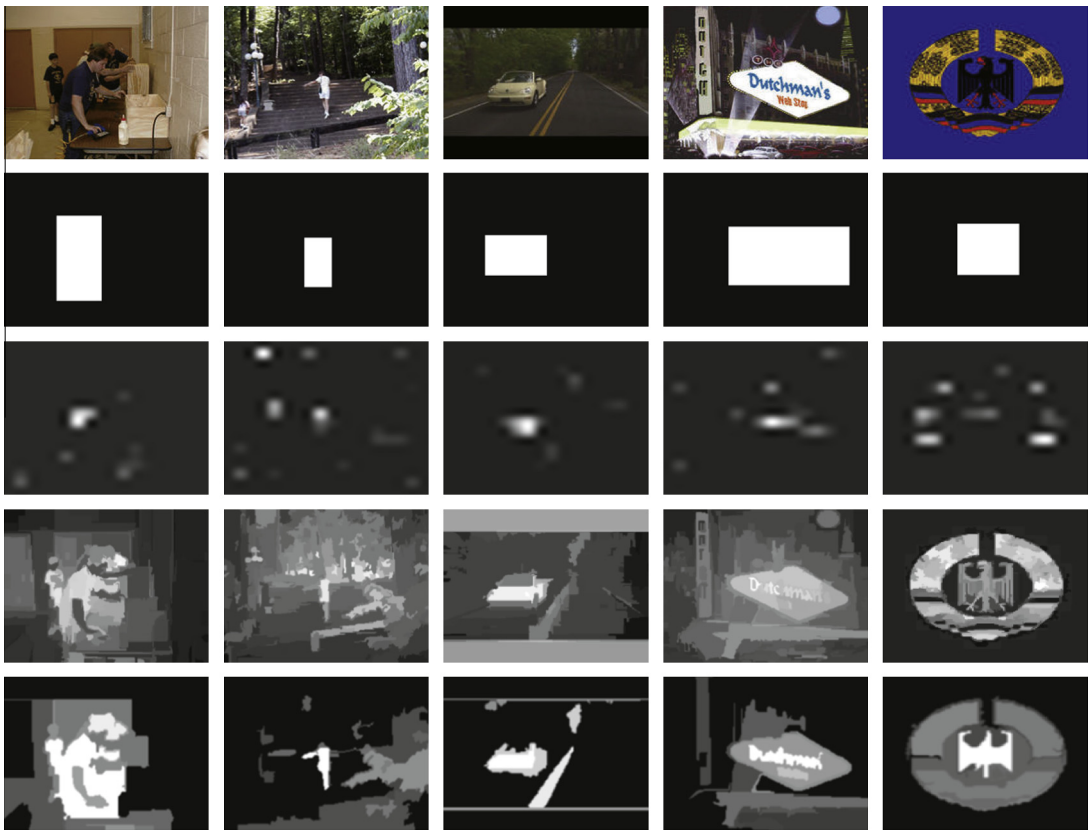

Fig. 1. Saliency detection. From top to bottom, each row respectively represents the original images, the ground truths, the saliency maps calculated by IT [13], RC [14], and the proposed model.

image auto-tagging [15–17] and tag completion [18], which can predict for images the relevant keywords from a vocabulary, this paper proposes a Tag-Saliency model capable of tackling high-level information and convenient to use. The proposed model takes advantage of the following two aspects:

First, the image auto-tagging technique is introduced into saliency detection task. Compared with the previous top-down saliency methods, the proposed method can extract high-level semantic information from large scale visual media data through only one unified image tagging model. This image tagging model can be obtained from the suitable training sets, without the need for a large number of specific object detectors.

Second, a hierarchical over-segmentation and global contrast based paradigm is proposed, which can integrate the high-level and low-level information for effective saliency estimation. In this paradigm, each image will be pyramid decomposed into a sequence of hierarchical regions, where regions at one segmentation level may be partitioned into several more finer spatial subregions at the next level. Based on the hierarchical over-segmented regions, both low-level and high-level information are extracted for global contrast analyzing. The main advantage of hierarchical over-segmentation is the ability to provide multiple information for regional tagging, which can yield the excellent tagging accuracy. The global contrast analysis is employed here. This is mainly because the experimental experiences which indicate that global contrast based models are often connected with outstanding performances in practice [14,19,20].

The rest of this paper is organized as follows. Section 2 reviews some relevant mainstream works. Section 3 details the proposed model. Section 4 presents the extensive experiments to verify the effectiveness of the proposed model. Section 5 gives a quantitative discussion to analyze all the factors that influence the performance of our model, and the conclusion follows in Section 6.

2. Related work

Classical bottom-up saliency detection methods choose to utilize low-level information to calculate contrasts of image regions with respect to their surroundings. According to the range of comparative reference regions they used, these methods can be roughly classified into local contrast based and global contrast based.

Local contrast based methods determine saliency of each image region by calculating the low-level contrast between the examined region and its local neighborhoods. Many early studies are related with the biologically inspired visual model introduced by Koch and Ullman [21]. The ground-breaking implementation of this model is the work of Itti et al. [13]. They employ a Difference of Gaussian (DoG) approach to extract multi-scale low-level information from images, and utilize this information to define saliency by calculating center-surround differences. Later on, in the work of Walther et al. [22], the method of [13] is modified with a hierarchical recognition system to recognize salient objects. Similarly, Han et al. [9] modify the method of [13] with Markov Random Field (MRF) based region growing to produce salient regions.

Besides the aforementioned approaches, there are many other methods which are not clearly related with a biologically inspired visual model, but strongly related with pure local contrast analysis. For instance, Gao et al. [23] detect saliency by estimating the Kullback–Leibler (KL) divergence of a series of DoG and Gabor filter responses to calculate the local contrast between the examined location and its surrounding local region. Harel et al. [24] design a graph based method. They firstly form activation maps by combining some excellent feature maps, and then use graph algorithms and a measure of contrast to achieve conspicuous parts. Hou and Zhang [25] introduce a model in frequency domain, which defines saliency of a location based on the difference between the