Jianming Zhang · Filip Malmberg
Stan Sclaroff

Visual Saliency: From Pixel-Level to Object-Level Analysis

Springer

Visual Saliency: From Pixel-Level to Object-Level Analysis

Jianming Zhang • Filip Malmberg • Stan Sclaroff

Visual Saliency: From Pixel-Level to Object-Level Analysis

Springer

Jianming Zhang
Adobe Inc.
San Jose, CA, USA

Filip Malmberg
Centre for Image Analysis
Uppsala University
Uppsala, Uppsala Län, Sweden

Stan Sclaroff
Department of Computer Science
Boston University
Boston, MA, USA

ISBN 978-3-030-04830-3 ISBN 978-3-030-04831-0 (eBook)
https://doi.org/10.1007/978-3-030-04831-0

Library of Congress Control Number: 2018965600

This Springer imprint is published by the registered company Springer Nature Switzerland AG.
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Overview

Visual saliency computation is about detecting and understanding pertinent regions and elements in a visual scene. Given limited computational resources, the human visual system relies on saliency computation to quickly grasp important information from the excessive input from the visual world [189]. Modeling visual saliency computation can help computer vision systems to filter out irrelevant information and thus make them fast and smart. For example, saliency detection methods have been proposed to predict where people look [87], delineate between foreground regions and the background [1], and localize dominant objects in images [112]. These techniques have been used in many computer vision applications, e.g. image segmentation [69], object recognition [150], visual tracking [118], gaze estimation [169], action recognition [125], and so on.

In this book, we present methods for both traditional and emerging saliency computation tasks, ranging from classical low-level tasks such as pixel-level saliency detection to emerging object-level tasks such as subitizing and salient object detection. For low-level tasks, we focus on pixel-level image processing approaches based on efficient distance transform. For object-level tasks, we propose data-driven methods using deep convolutional neural networks. The book includes both empirical and theoretic studies, together with implementation details of the proposed methods. The rest of this chapter will introduce the background of those saliency computation tasks and the outline of this book.

1.1 Pixel-Level Saliency Detection

Early works in visual saliency computational modelling [79, 95] aim at computing a saliency/attention map that topographically represents humans' attentional priority when they view a scene. Computing such saliency maps is formulated as assigning importance levels to each pixel of a digital image. At the beginning of this line

J. Zhang et al., *Visual Saliency: From Pixel-Level to Object-Level Analysis*,
https://doi.org/10.1007/978-3-030-04831-0_1

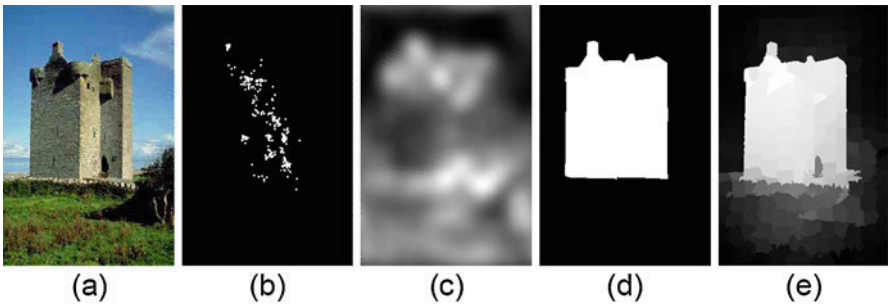

Fig. 1.1 (**a**) A sample image from the DUT-Omron dataset [194]. (**b**) The ground truth for eye fixation prediction. Each white spot on the map is an eye fixation position of some participant in the free-viewing experiment. (**c**) The saliency map by a state-of-the-art method for eye fixation prediction [61]. (**d**) The ground truth for salient region detection. (**e**) The saliency map by a state-of-the-art method for salient object detection [84]

of research, the saliency values are represented based on the likelihood of human eye gaze. Therefore, this type of saliency models are often named as eye fixation prediction models. Later, motivated by image segmentation, people started to compute saliency maps that delineate dominant objects from the background. This type of saliency detection task is named interchangeably as salient object detection, salient object segmentation, or salient region detection. In this book, we will refer to it as salient region detection, since most of the existing works of this topic are still based on very low-level image processing algorithms without semantic object-level understanding.

Figure 1.1 shows the difference between the two pixel-level saliency tasks mentioned above. Figure 1.1b, c is the ground truth eye fixation map and the predicted saliency map, respectively. The eye fixation data is collected by eye tracking devices from multiple subjects when they free-view the image, and the data is expected to have a significant amount of inter-subject variance. By aggregating the eye fixations of different subjects, the ground truth map shows that the human eye gaze is often focused on salient regions or objects, but there is still a lot of uncertainty about the exact positions of those eye fixations. As a result, the predicted saliency map tends to be very blurry. In contrast, the ground truth map for salient region detection is simply a binary foreground mask for the dominant object in the image, and thus has much less uncertainty (see Fig. 1.1c) when there is a well-defined dominant object. The predicted map needs to uniformly highlight the regions of dominant objects with precise object boundary details, as shown in Fig. 1.1d.

These two pixel-level saliency tasks are useful in many applications. Predicting human eye fixation is useful in applications related to human–computer interaction and graphics, e.g. gaze estimation [169], eye tracker calibration [168], non-photorealistic rendering [47], stereoscopic disparity manipulations [92], image retargeting [146], photo quality assessment [131], etc. Other computer vision tasks

such as action recognition, tracking, and object detection can also benefit from this task by analyzing relevant regions indicated by human eye fixation [118, 125, 150]. Salient region detection methods can automatically generate saliency maps to extract dominant objects, and thus this task is useful for automatic image segmentation [32] and a lot of photo editing applications that need such segmentation [30, 37, 76].

1.2 Object-Level Saliency Computation

Object-level saliency computation is a relatively new topic. By object level, we mean that the analysis should be performed based on the understanding of object instances. One of the basic problems, for example, is to predict the existence of salient object(s) in an image. Existence prediction leads to the differentiation between object-centric images and scene-centric images, which is a very general and fundamental attribute for image understanding.

In this book, we propose a new problem of visual saliency computation, called salient object subitizing (SOS), which is to predict not only the existence but also the number of salient objects in an image using holistic cues, without the need to localize them. This task is inspired by humans' subitizing ability to quickly and accurately tell the number of items within the subitizing range (1–4) [90]. Because the appearance and size of salient objects can vary dramatically from category to category, and from image to image, the SOS problem poses very different challenges than traditional object counting problems [3, 129].

Knowing the existence and the number of salient objects without the expensive detection process can enable a machine vision system to select different processing pipelines at an early stage, making the vision system more intelligent and reducing computational cost. Furthermore, differentiating between scenes with zero, a single and multiple salient objects can also facilitate applications such as image retrieval, iconic image detection [11], image thumbnailing [38], robot vision [152], egocentric video summarization [104], snap point prediction [191], etc.

Besides the subitizing task, detecting generic salient object instances in unconstrained images, which may contain multiple salient objects or no salient object, is also a fundamental problem (see examples in Fig. 1.2). We refer to this task as unconstrained salient object detection. Solving this problem entails generating a compact set of detection windows that matches the number and the locations of salient objects. Detecting each salient object (or reporting that no salient object is present) can be very helpful in the weakly supervised or unsupervised learning scenario [31, 89, 207], where object appearance models are to be learned with no instance level annotation.

The unconstrained salient object detection task arguably solves the problem of salient object subitizing, as the existence and the number of salient objects can be derived from the detection result. However, salient object subitizing can be solved by much faster and lighter-weight models, since it does not require accurate

Fig. 1.2 Examples of unconstrained salient object detection. Note that for the input image in the right column, there is no dominant object

localization of salient objects. Moreover, for training a subtizing model, the labor cost for annotation is also much less than the detection model. Therefore, one can expect that the subitizing model can do a better job in detecting the existence and counting the number for salient objects given the same amount of labor and computation resources. This will also be empirically verified in this book.

1.3 Book Outline

The book will cover each of the saliency computation tasks mentioned above and present computational approaches as well as empirical and theoretic studies. Here we provide a summary of the main content.

1.3.1 Boolean Map Saliency for Eye Fixation Prediction

A majority of existing eye fixation prediction models are based on the contrast and the rarity properties of local image patches, e.g. [13, 19, 79]. However, these local image properties have limited ability to model some global perceptual phenomena [96] known to be relevant to the deployment of visual attention. One such global perception mechanism is figure-ground segregation. Several factors are likely to influence figure-ground segregation, e.g. size, surroundedness, convexity, and symmetry [133]. As Gestalt psychological studies suggest, figures are more likely to be attended to than background elements [126, 145] and the figure-ground assignment can occur without focal attention [94]. Neuroscience findings also show that certain responses in monkey and human brains involved in shape perception are critically dependent on figure-ground assignment [10, 98], indicating that this process may start early in the visual system.

In the first part of this book, we explore the usefulness of the surroundedness cue for eye fixation detection [201, 202]. We propose a simple, training-free, and computationally efficient Boolean map saliency model (BMS). Our model uses basic image processing operations to find surrounded regions in binary maps which are generated by randomly thresholding the color channels of an image. The saliency map is computed based on the probability that a pixel belongs to a surrounded region in a sampled binary map. Despite its simplicity, BMS consistently achieves state-of-the-art performance across all the testing datasets. Regarding the efficiency, BMS can be configured to run at about 100 FPS with only a little drop in performance, which makes it quite suitable for many time-critical applications.

1.3.2 A Distance Transform Perspective

We provide an explanation of the BMS algorithm in a perspective of image distance transform [120, 202]. First, we propose a novel distance function, the Boolean map distance (BMD), that defines the distance between two elements in an image based on the probability that they belong to different components after thresholding the image by a randomly selected threshold value. We show that the BMS algorithm is a straightforward implementation to compute the Boolean map distance of each pixel to the image border.

Then we draw a connection between the Boolean map distance and the minimum barrier distance (MBD) [166]. We prove that the Boolean map distance gives a lower bound approximation of the minimum barrier distance. As such it shares many of the favorable properties of the MBD discovered in [40, 166], while offering some additional advantages such as more efficient distance transform computation and straightforward extension to multi-channel images. These analyses provide insight into why and how BMS can capture the surroundedness cue via Boolean maps.

Finally, we discuss efficient algorithms for computing the Boolean map distance and the minimum barrier distance. In the next chapter, we propose a fast raster-scanning algorithm to approximate BMD and MBD, and show how to use that for real-time salient region detection.

1.3.3 Efficient Distance Transform for Salient Region Detection

Due to the emerging applications on mobile devices and large-scale datasets, a desirable salient region detection method should not only output high quality saliency maps, but should also be highly computationally efficient. In this chapter, we address both the quality and speed challenges for salient region detection using an efficient distance transform algorithm.

The surroundedness prior, also known as the *image boundary connectivity prior*, assumes that background regions are usually connected to the image borders. This prior is shown to be effective for salient region detection [187, 194, 201, 208]. To leverage this prior, previous methods, geodesic-distance-based [187, 208] or diffusion-based [84, 194], rely on a region abstraction step to extract superpixels. The superpixel representation helps remove irrelevant images details, and/or makes these models computationally feasible. However, this region abstraction step also becomes a speed bottleneck for this type of methods.

To boost the speed, we propose a method to exploit the image boundary connectivity prior without region abstraction [203]. Inspired by the connection between our BMS eye fixation prediction method and the *minimum barrier distance* (MBD) [40, 166], we use the MBD [40, 166] to measure a pixel's connectivity to the image boundary. Compared with the widely used geodesic distance, the MBD is much more robust to pixel value fluctuation. Since the exact algorithm for the MBD transform is not very efficient, we present FastMBD, a fast raster-scanning algorithm for the MBD transform, which provides a good approximation of the MBD transform in milliseconds, being two orders of magnitude faster than the exact algorithm [40]. The proposed salient region detection method runs at about 80 FPS using a single thread, and achieves comparable or better performance than the leading methods on four benchmark datasets. Compared with methods with similar speed, our method gives significantly better performance.

1.3.4 Salient Object Subitizing

We introduce a new computer vision task, salient object subitizing (SOS), to estimate the existence and the number of salient objects in a scene [199, 200]. To study this problem, we present a salient object subitizing image dataset of about 14K everyday images. The number of salient objects in each image was annotated by Amazon Mechanical Turk (AMT) workers. The resulting annotations from the AMT workers were analyzed in a more controlled offline setting; this analysis showed a high inter-subject consistency in subitizing salient objects in the collected images.

We formulate the SOS problem as an image classification task, and aim to develop a method to quickly and accurately predict the existence and the number of generic salient objects in everyday images. We propose to use an end-to-end trained deep convolutional neural network (CNN) model for our task, and show that an implementation of our method achieves very promising performance. In particular, the CNN-based subitizing model can approach human performance in identifying images with no salient object and with a single salient object. To further improve the training of the CNN SOS model, we propose a method to leverage synthetic images. Moreover, we demonstrate the application of our CNN-based SOS method in salient object detection and image retrieval.

1.3.5 Unconstrained Salient Object Detection

We introduce another new task, called unconstrained salient object detection. Many previous so-called "salient object detection" methods [1, 12, 34, 85, 113, 158] only solve the task of salient region detection, i.e. generating a dense foreground mask (saliency map). These methods do not individuate each object and assume the existence of salient objects. In contrast, we present a salient object detection system that directly outputs a compact set of instance detection windows for an unconstrained image, which may or may not contain salient objects. Our system leverages the high expressiveness of a convolutional neural network (CNN) model to generate a set of scored salient object proposals for an image. Inspired by the attention-based mechanisms of [8, 102, 127], we propose an adaptive region sampling method to make our CNN model "look closer" at promising images regions, which substantially increases the detection rate. The obtained proposals are then filtered to produce a compact detection set.

A key difference between salient object detection and object class detection is that saliency greatly depends on the surrounding context. Therefore, the salient object proposal scores estimated on local image regions can be inconsistent with the ones estimated on the global scale. This intrinsic property of saliency detection makes our proposal filtering process very challenging. Using the common greedy non-maximum suppression (NMS) method often leads to suboptimal results for our proposals. To attack this problem, we propose a subset optimization formulation based on the *maximum a posteriori* (MAP) principle, which jointly optimizes the number and the locations of detection windows. The effectiveness of our optimization formulation is validated on three benchmark datasets, where our formulation attains about 15% relative improvement in average precision (AP) over the NMS approach. Moreover, our method also attains about 15–35% relative improvement in AP over previous methods on these datasets.