Digital Signal Processing 63 (2017) 44–55

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Supervised training and contextually guided salient object detection

Mengnan Du [a], Xingming Wu [a,*], Weihai Chen [a], Zhengguo Li [b]

[a] School of Automation Science and Electrical Engineering, Beihang University, Beijing, 100191, China
[b] Robotics Department, Institute for Infocomm Research, 138632, Singapore

ARTICLE INFO

Article history:
Available online 6 January 2017

Keywords:
Salient object detection
Supervised training
Contextual optimization
CRF
Hierarchical feature

ABSTRACT

Existing salient object detection research generally focused on designing diverse saliency features and integrating them heuristically. In this paper, a novel salient object detection method is proposed by employing supervised training and contextual modeling. Gradient boosting decision trees are explored to aggregate features on segmented regions using a supervised training manner. Feature representation of hierarchically segmented regions is exploited to capture salient objects at different levels so as to extract discriminative features better. A region-level pairwise conditional random field (CRF) method is constructed to further boost the accuracy of saliency estimation as well as to improve the perceptual consistency of saliency maps. Experimental results demonstrate that the proposed method could achieve state-of-the-art performance over all public datasets. The F-measure is improved by 3.9%, 13.0%, 4.3% on the MSRA-B, DUT-OMRON and HKU-IS dataset respectively, and the mean absolute error (MAE) is reduced by 31.6%, 26.4% and 21.2% respectively on these three datasets.

1. Introduction

Visual saliency, which aims to find the objects that could effectively represent a visual scene, has received much attention in recent years since the first computational visual saliency model presented by Itti et al. [1]. It has a wide variety of applications in the fields of computer vision and robotics, including object detection and segmentation [2], image thumbnailing [3], image retrieval [4], video coding [5], tone mapping of high dynamic range images [6] and so on. Visual saliency models can be categorized into two major classes [7,8], i.e., fixation prediction [1] and salient object detection. Fixation prediction models aim at predicting where human look in an image. In contrast, salient object detection models are designed to detect the most attention-grabbing objects in a scene, which are also the focus of this paper.

Most existing salient object detection methods focused on designing diverse saliency features and integrating them heuristically using simple approaches, such as multiplication [9,10] or weighted average [11], *etc*. The majority of features deals with the global color contrast calculation at the pixel wise [12] or the region-level [13]. This is mainly due to its conceptual simplicity and it can simulate the mechanism of human visual perception. Region compactness or color spatial distribution [11,10], which assumes that regions with low color spatial variance are more likely to belong to salient objects, is another frequently explored feature that also exploits the contrast principle. These color contrast based methods perform well on images with simple backgrounds or high color contrasts among the foreground objects and the background regions. However, it is proved that merely utilizing the color contrast is insufficient to capture salient objects in more cluttered scenes.

In order to compensate for the weakness of contrast principle, a variety of saliency priors were introduced recently, such as center prior [9,14], backgroundness prior [15] and focusness prior [16]. Nevertheless, the performance of these priors heavily relies on the object arrangement of the test dataset. Algorithms utilizing these priors may overfit to a specific dataset, *e.g.*, the widely used ASD dataset [12], and fail to work when tested on an alternative dataset. As such, the aforementioned heuristic methods are limited in their ability to characterize complex scenarios. Clearly, saliency maps produced by these heuristic methods impose restriction on their application to many fields, such as image segmentation and object recognition. Although significant progress has been made [8], it is still desired to further improve saliency estimation accuracy.

A novel saliency estimation method is introduced using a supervised training manner in this paper. Inspired by the works in [17] and [18], the salient region detection task is formulated as a regression problem. Specifically, a gradient boosting frame is adopted to learn a regression function that maps features extracted from a region to a saliency score. Compared to the heuristic methods, our supervised training method can learn more complex non-linear

* Corresponding author.
E-mail addresses: dumengnan@gmail.com (M. Du), wxmbuaa@163.com (X. Wu), whchen@buaa.edu.cn (W. Chen), ezgli@i2r.a-star.edu.sg (Z. Li).

http://dx.doi.org/10.1016/j.dsp.2017.01.004
1051-2004/

integration scheme of features. In addition, our training based method can automatically find the discriminative features from high-dimension feature vectors, resulting in a richer representation of salient objects on the segmented regions with relatively large sizes. The image is segmented at multi-scale using the gPb/UCM method in [19], which allows us to extract features at hierarchically segmented regions, leading to a better saliency representation than extracting feature just at a single segmentation scale. The gPb/UCM segmentation approach segments an image into regions where the boundaries are captured well. Besides, the saliency map generated by taking the integration of saliency maps from multiple segmentations is often more consistent with its ground truth mask than saliency map from a specific segmentation scale. It is well known that a single image region only contains limited information, which could lead to misleading saliency score. To reduce the ambiguity, it is necessary to take all the regions in an image into consideration. As such, a region-level pairwise CRF method is proposed to model the spatially context dependencies among all the segmented regions. By penalizing inconsistent saliency scores derived from similar color features, it enhances the consistency of adjacent regions and significantly boosts the accuracy of the overall salient object detection system. In summary, three major contributions of this paper are:

1. An effective gradient boosting regressor to map saliency features extracted from segmented regions to saliency scores.
2. A hierarchical feature representation on regions generated by multi-scale segmentation to capture salient objects at different scales.
3. A region-level pairwise CRF method to boost accuracy as well as to enhance perceptual consistency of saliency maps.

The rest of this paper is organized as follows. Section 2 summarizes related salient object detection works. Section 3 presents the proposed supervised training method. Section 4 introduces our proposed region-level pairwise CRF method. Section 5 includes experimental results to verify effectiveness of the proposed method. Conclusional remarks and future works are presented in Section 6.

2. Related work

Salient object detection is a widely studied topic in computer vision. Generally, saliency object methods can be classified according to two major factors: saliency feature selection (what features are the algorithms used) and feature aggregation (how these features are aggregated and mapped into the final saliency score). As such, we classify the salient object detection literature into two types, i.e., heuristic salient object detection and supervised learning based salient object detection (see Table 1), and review the relevant salient object detection methods in the following.

2.1. Heuristic salient object detection

Most early stage research on salient object detection focused on designing features that can discriminate an object from the background. Perhaps the most frequently utilized feature is the global color contrast, partly because it can simulate the mechanism of human visual perception. Achanta et al. [12] formulated the saliency computation as the difference between the mean image feature and the blurred image feature. The work in [13] extended the global contrast from pixel-level to region-level by introducing a spatially weighted contrast between a region and its surrounding regions. Margolin et al. [9] calculated the color distinctness of an image region using the sum of the color distances between that region with all other regions in that image. Yan et al. [20] formulated color contrast in a hierarchical manner so as to coping with images

Table 1
Comparison of related works from two aspects: feature selection and feature aggregation. Among the features, CC = Color contrast, SD = Color spatial distribution, GP = Generic properties, BP = Background prior, CP = Center prior, SP = Shape prior, CNN = Convolutional neural network.

	Model	Feature	Aggregation
Heuristic	FT [12]	CC	–
	RC [13]	CC	–
	PCA [9]	CC + CP	Non-linear
	GS [15]	BP	–
	HS [20]	CC + CP	Hierarchical integration
	SF [10]	CC + SD	Non-linear
	CB [21]	CC + CP + SP	Linear
	CA [14]	CC	Non-linear
	MC [22]	BP	Non-linear
	MR [23]	BP	–
	RWR [24]	BP	–
	RBD [25]	CC + BP	–
Supervised	HDCT [18]	CC + SD + GP	Boosted decision tree
	BL [26]	GP	Support vector machine
	DRFI [17]	CC + BP + GP	Random forest
	MDL [27]	CNN	–
	MDF [28]	CNN	Neural network

with complex structures. Color spatial distribution [10,11] is another widely used saliency feature. It assumes that a region with compact spatial distribution is more likely to represent a meaningful object. Perazzi et al. [10] formulated regional contrast and spatial distribution in the form of high-dimensional Gaussian filters for efficient implementation. These features are limited in the sense that they may fail to capture saliency, *e.g.*, in complicated scenes with cluttered texture or lighting variations.

To compensate for the weakness of contrast calculation, various priors were developed in recent works. Center prior, the hypothesis is that salient object is likely to be put at the center of an image, was investigated in [9,14], and [20]. Background prior and boundary connectivity prior view the saliency detection problem from the perspective of background rather than the object [15]. Using segmented regions at the four image boundaries as queries, Yang et al. [23] employed graph based ranking for saliency detection. Rather than simply regarding the four image boundaries as background [23], Li et al. [24] first selected from the four boundaries with the largest possibility belong to the background, and produced foreground saliency estimation. They then constructed random walks ranking by extracting seeds from the foreground saliency estimation and finally generates pixel-wise saliency map. A background measure that can capture the spatial layout of image patches with respect to image boundaries was introduced in [25]. Based on the background prior, Jiang et al. [22] computed the visual saliency through absorbing Markov chain on an image graph model. Jiang et al. [21] presented shape prior, which meant that salient object should have a well defined closed boundary, for automatic saliency estimation. Similarly, Singh et al. [29] assumed that salient object should have a closed shape, which can thus be covered by a convex hull. They first extracted key points from the input image and then constructed a convex hull around these key points, generating the rough salient region. Later, they utilized Gaussian mixture model to refine the rough saliency map and obtained the final saliency map. The majority of these methods built on heuristic image priors lack sufficient theoretical support. As a result, these priors usually fail to handle complex scenes where objects touch the image boundaries or several salient objects exist.

2.2. Supervised salient object detection

Instead of integrating the saliency features heuristically, there were a few methods addressing the salient object detection problem by utilizing supervised training manners, learning the map-