Pattern Recognition Letters 84 (2016) 85–90

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Salient object detection via point-to-set metric learning☆

Jia You, Lihe Zhang*, Jinqing Qi, Huchuan Lu

School of Information and Communication Engineering, Dalian University of Technology, Dalian 116023, China

ARTICLE INFO

Article history:
Received 9 February 2016
Available online 31 August 2016

Keywords:
Salient object detection
Metric learning
Point-to-set classification

ABSTRACT

Distance metric is an essential step of salient object detection, in which the pairwise distances are often used to distinguish salient image elements (pixels and regions) from background elements. Instead of using the point-to-point distance metrics which possibly implicitly take into account the context information around data points, we learn the point-to-set metric to explicitly compute the distances of single points to sets of correlated points and cast saliency estimation as the problem of point-to-set classification. First, we generate a series of bounding box proposals and region proposals for an input image (i.e., some pre-detected regions which possibly include object instances), and exploit them to compute a recall-preference saliency map and a precision-preference one, based on which the background and foreground seed regions are respectively determined. Next, we collect positive and negative samples (include point samples and set samples) to learn the point-to-set distance metric, and employ it to classify the image elements into foreground and background classes. Last, we update the training samples and refine the classification result. The proposed approach is evaluated on three large publicly available datasets with pixel accurate annotations. Extensive experiments clearly demonstrate the superiority of the proposed approach over the state-of-the-art approaches.

1. Introduction

Saliency detection aims to simulate human visual attention mechanisms and find out the focus of visual attention in images and videos. Although much progress has been made in recent years, it remains a challenging problem. Essentially, saliency detection is to explore the properties of salient stimuli, and characterize each element (is also called point) in images and further distinguish salient points from non-salient ones, where the computation of the distance or similarity among points is necessary. Some purely computational saliency methods directly compute the point-to-point affinities using the Euclidean distance [1,12,14,30,32,38]. They separately deal with each point and ignore the mutual dependence among points. Therefore, the direct distance-calculation may not be a good measure of the affinities between samples. The sparse reconstruction based methods map each point to the subspace spanned by the definite foreground or definite background points, and then estimate its label according to the reconstruction residual [28,34,40]. These methods implicitly measure the distance of single point to a set of seed points (i.e., the atoms in the dictionary) using the residual, whereas each observed point is also handled independently.

Numerous graph-based approaches integrate local and global spatial connections between points on data manifold to learn the point-to-point affinities for saliency detection [9,15,21,31,35,42]. These methods actually imply the notion of metric learning, that is, learn a valid distance metric, measured by which the samples with the same class label or similar samples could be as close as possible, while the samples with the different class labels or dissimilar samples could be as far as possible.

In saliency detection, whether a superpixel is salient depends on the context it lies in. Therefore, we learn a point-to-set distance metric to explicitly compute the affinities between single superpixel (i.e., a point) and a set of correlated ones (i.e., a set), thereby obtaining the context-aware pairwise affinities. We adopt the supervised metric learning with two class labels. Different from the point-to-point metric learning, the training samples used in this work include point samples and set samples. The pipeline of the proposed algorithm is shown in Fig. 1. We first exploit bounding box and region proposals to compute a recall-preference map and a precision-preference map, based on which the pseudo training samples are generated. Second, we learn the point-to-set distance metric and employ it to classify all superpixels. Third, we iteratively learn the distance metric using the updated training samples and re-classify superpixels. The main contributions of this work include:

☆ This paper has been recommended for acceptance by Dr. D. Coeurjolly.

* Corresponding author.
E-mail address: zhanglihe@dlut.edu.cn (L. Zhang).

http://dx.doi.org/10.1016/j.patrec.2016.08.018
0167-8655/

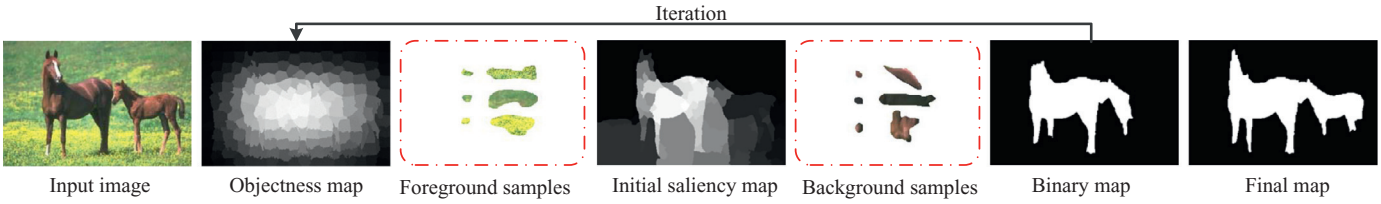

Fig. 1. Pipeline of the proposed algorithm. We first compute an objectness map with high recall, based on which the background training samples are selected. Second, we compute an initial saliency map with high precision, from which the foreground training samples are selected. Third, we learn the point-to-set distance metric and classify all superpixels, thereby obtaining a binary map. Finally, we update the training samples and re-learn the distance metric and re-classify the superpixels.

- We learn the point-to-set distance metric to capture the context cues around each superpixel, and formulate saliency estimation as the problem of point-to-set classification.
- We propose a simple yet effective rough object localization method by comprehensively utilizing bounding box and region object proposals.
- Experimental results on several large benchmark datasets show that the proposed algorithm performs favorably against the state-of-the-art saliency methods.

2. Related work

Numerous saliency models have been proposed in the most recent decade. A thorough review on this topic can be found in Borji and Itti [7] and we discuss the most related methods in this section.

Based on cognitive studies of visual search, Itti et al. [17] calculate the center-surround differences on the image pyramid to characterize pixel saliency with respect to local context. Cheng et al. [12] compute the global contrast by the linear accumulation of appearance differences weighted by spatial distances in the entire image. While Goferman et al. [14] combine the point-to-point distance metric at multiple scales to comprehensively evaluate the local and global context of the dominant objects. Given a set of basis vectors, the encoding residual can be used to denote the overall distance of an input vector to these bases. Inspired by it, some sparse reconstruction based saliency methods are proposed [28,40].

Supervised learning is also often applied in saliency detection, which learns to distinguish salient regions from the background. Borji [5] trains several linear and non-linear classifiers from bottom-up and top-down features to fixations. Kienzle et al. [23] and Judd et al. [22] utilize support vector machines to learn saliency from human eye tracking data. Jiang et al. [20] integrate the regional contrast, regional property and regional backgroundness features together and learn a regressor to directly map the regional feature to a saliency score. While Zhao and Koch [41] use least-squares regression to learn the weights associated with a set of feature maps from subjects freely fixating natural scenes drawn from eye tracking datasets. These methods require a large number of annotated images in order to train the classifiers and regressors. There also exist some approaches to train saliency models using the pseudo labels which are assigned according to various prior knowledge. Zhang and Yuan [39] respectively exploit the contrast prior and the boundary prior to label positive samples and negative samples.

The graph-based affinity measure well infers relationships between data points, thereby describing the underlying manifold structure lied in the data space. Harel et al. [15] use dissimilarity to define edge weights on graphs which are interpreted as Markov chains, and treat the equilibrium distribution over the activation map as saliency values. Wang et al. [35] introduce the entropy rate and incorporate the equilibrium distribution to measure the average information transmitted from a node to the others at one step, which is used to predict visual attention. Recently, Li et al. [25] propose the regularized random walks ranking and define a new fitting constraint to consider local image data and prior estimation. Li and Yu [26] refine the initial saliency map on the graph model to enhance the spatial coherence of the saliency results. Similarly, there are many methods that formulate the pairwise similarities as a labeling problem on the vertices of a graph [21,36,39]. While some other approaches exploit the graphical model to integrate multiple saliency cues [9,31,42].

The approach most related to ours is Li et al. [27]. They first learn a generic distance metric from the training image set for all images. Based on the optimal distances, they learn another distance metric again for each to-be-processed image using the predetermined pseudo training samples. Next, the two metrics are fused together to compute the pairwise affinities between any pair of superpixels. Last, the distances of each superpixel to foreground and background seeds are integrated to formulate the saliency. This method learns the point-to-point distance metrics, and accumulate the pairwise distances to compute the distance of a point to a seed set. And, the annotated images are required in the phase of generic metric learning. Different from [27], we learn the point-to-set distance metric to directly and explicitly calculate the distance of a point to a set, and cast saliency detection into the problem of point-to-set classification. Moreover, the human annotated images (i.e., ground truth) do not need to be provided in the whole procedure of saliency detection in this work.

3. Point-to-set metric learning

Sets (of images) are typically modeled as a subspace lied on specific Riemannian manifold in computer vision. Huang et al. [16] propose the framework of Euclidean-to-Riemannian metric learning for point-to-set classification. Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\} \subset \mathbb{R}^D$ denote a set of Euclidean points and $\mathrm{Y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_k\} \subset \mathcal{M}$ denote a collection of Riemannian points, where $\mathrm{y}_i \in \mathbb{R}^{D \times D'}$ is a set of Euclidean points. Employing the kernel trick, the point-to-set distance $d(\boldsymbol{x}, \mathrm{y})$ can be written as follows:

$$d(\boldsymbol{x}, \mathrm{y}) = \left\| \sum_i \boldsymbol{W}_{\boldsymbol{x}}^{\top} \boldsymbol{K}_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{x}_i) - \sum_j \boldsymbol{W}_{\mathrm{y}}^{\top} \boldsymbol{K}_{\mathrm{y}}(\mathrm{y}, \mathrm{y}_j) \right\|_2, \tag{1}$$

where $\boldsymbol{K}_{\boldsymbol{x}}(\boldsymbol{x}, \boldsymbol{x}_i)$ denotes the Euclidean metric based kernel, and $\boldsymbol{K}_{\mathrm{y}}(\mathrm{y}, \mathrm{y}_j)$ denotes the Riemannian metric based kernel.

The projection matrixes $\boldsymbol{W}_{\boldsymbol{x}}$ and $\boldsymbol{W}_{\mathrm{y}}$ can be obtained by solving the following objective function:

$$\min_{\boldsymbol{W}_{\boldsymbol{x}}, \boldsymbol{W}_{\mathrm{y}}} \left\{ D(\boldsymbol{W}_{\boldsymbol{x}}, \boldsymbol{W}_{\mathrm{y}}) + \lambda_1 G(\boldsymbol{W}_{\boldsymbol{x}}, \boldsymbol{W}_{\mathrm{y}}) + \lambda_2 T(\boldsymbol{W}_{\boldsymbol{x}}, \boldsymbol{W}_{\mathrm{y}}) \right\} \tag{2}$$

where $D(\cdot, \cdot)$, $G(\cdot, \cdot)$ and $T(\cdot, \cdot)$ are the distance, discriminant geometry and transformation constraints respectively, $\lambda_1 > 0$, $\lambda_2 > 0$ are the balancing parameters. For more details on their definitions, please refer to [16].

Image pixels are often described in terms of feature vectors, each of which can be taken as a point lying in the Euclidean space.