Pattern Recognition 60 (2016) 348–360

Contents lists available at ScienceDirect

Pattern Recognition

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

Bottom-up saliency detection with sparse representation of learnt texture atoms☆

Mai Xu [a], Lai Jiang [a], Zhaoting Ye [a], Zulin Wang [a,b,]*

[a] School of Electronic and Information Engineering, Beihang University, Beijing 100191, China
[b] Collaborative Innovation Center of Geospatial Technology, 129 Luoyu Road, Wuhan 430079, China

ARTICLE INFO

Article history:
Received 21 January 2016
Received in revised form
11 April 2016
Accepted 14 May 2016
Available online 21 May 2016

Keywords:
Visual attention
Saliency detection
Sparse representation
Dictionary learning

ABSTRACT

This paper proposes a saliency detection method by exploring a novel low level feature on sparse representation of learnt texture atoms (SR-LTA). The learnt texture atoms are encoded in salient and non-salient dictionaries. For salient dictionary, a formulation is proposed to learn salient texture atoms from image patches attracting extensive attention. Then, the online salient dictionary learning (OSDL) algorithm is presented to solve the proposed formulation. Similarly, the non-salient dictionary is learnt from image patches without any attention. Then, the pixel-wise SR-LTA feature is yielded based on the difference of sparse representation errors, regarding the learnt salient and non-salient dictionaries. Finally, image saliency can be predicted by linearly combining the proposed SR-LTA feature and conventional features, luminance and contrast. For the linear combination, the weights of different feature channels are determined by least square estimation on the training data. The experimental results show that our method outperforms 9 state-of-the-art methods for bottom-up saliency detection.

1. Introduction

Saliency detection refers to computing on image features to characterize the regions attracting different amounts of visual attention in a scene. Generally speaking, saliency detection is extensively studied in the context of the human visual system (HVS). Similar to the HVS, saliency detection enables machines to survive from processing a deluge of visual data. Thus, it has been widely applied in computer vision and image processing areas, such as object detection [1], object recognition [2], image retargeting [3], image quality assessment [4], and image/video compression [5].

For predicting visual attention, saliency detection can be traced back to feature integration theory [6] by Treisman and Gelade in 1980, which discussed on the possible visual features related to visual attention. To combine these features together, Koch and Ullman [7] in 1987 proposed to generate the saliency map for an input image, indicating which regions are conspicuous to attract attention in the HVS. Specifically, saliency map is a matrix with the same size as the input image, and the values of its elements range from 0 to 1. The large saliency value indicates high probability to attract human attention. Later, Itti and Koch [8] found out that the low level feature channels of intensity, color, and orientation are effective in generating the saliency map. In their method, these feature channels are decomposed for images at various scales subsampled by a Gaussian pyramid, and then conspicuity maps are constructed by center-surround responses to the decomposed feature channels. In each channel, conspicuity maps are aggregated across different scales. Finally, the saliency map can be obtained by the linear integration of conspicuity maps of all channels. Benefiting from the success of Itti's model [8], extensive saliency detection methods (e.g., [9–13]), using biological plausible features, have been proposed in the past decade.

Recently, several saliency detection methods (e.g., [14–18]) have been proposed to learn the parameters or even features from the ground-truth eye fixations[1] of training images, for saliency detection. From the perspective of parameters, Zhao and Koch [16] presented a method to learn weights associated with conspicuity maps for different feature channels, with least square fitting to fixations. This replaces the equal weight assignment in [8,19], thus improving the saliency detection accuracy. However,

☆A short version of this paper has been presented in ICCV Workshop 2015.
* Corresponding author. Tel.: +86 10 82317201.
E-mail addresses: MaiXu@buaa.edu.cn (M. Xu), wzulin@buaa.edu.cn (Z. Wang).

[1] Fixations are the points where people look during the eye tracking experiment. They are seen as the ground-truth of visual attention.

http://dx.doi.org/10.1016/j.patcog.2016.05.023
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

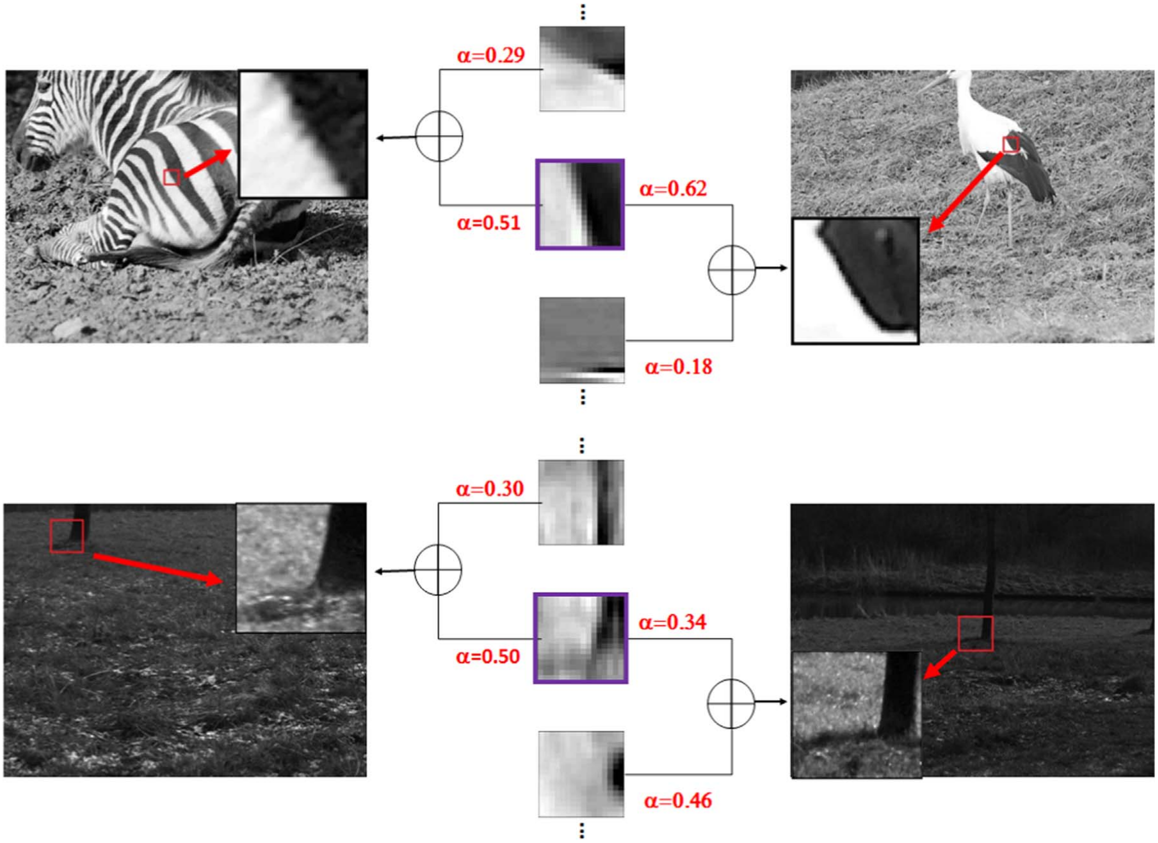

Fig. 1. An example of salient patches with similar texture patterns. The regions inside the red squares (enlarged in the corners) are salient patches, in the images of the eye tracking Kienzle database (the first row) and Doves database (the second row). Some atoms of the dictionaries, learnt from the salient regions of other training images, are shown in the middle of two images. In addition, the sparse representation coefficients α of the salient patterns regarding the learnt dictionaries are also provided. It can be seen that the salient patches across the different images may share some similar basic patterns, and these basic patterns may be learnt from the training data. Note that the patch sizes are 96 × 96 for DOVES and 41 × 41 for Kienzle et al., to ensure that the corresponding fovea degrees are around 1.5° in each database.

only few parameters can be learnt in these methods, such that the performance of these methods depends on the features of the conventional methods.

From the perspective of features, Kienzle et al. [17,18] proposed to directly learn patch patterns of salient and non-salient regions from the ground-truth eye tracking data. These patterns can be seen as low level features attracting different amount of visual attention. Specifically, two center-surround texture patches are learnt as the most relevant patterns for drawing visual attention, and two other patches are learnt as the least possible patterns for receiving eye fixations. Then, the saliency of an image patch can be detected, on the basis of the distance to the learnt texture patterns. However, the learnt four patch patterns have limited expression, since only two positive and two negative patterns are available for saliency detection.

Fig. 1 shows the possibility of learning hundreds of salient patterns (by applying the dictionary learning algorithm) for saliency detection. Accordingly, this paper proposes to learn extensive positive and negative patterns from the eye tracking data of training images, for bottom-up saliency detection. Specifically, this paper first proposes a formulation with a novel center-surround term, for learning two discriminative dictionaries. These two dictionaries contain the atoms for basic texture patterns of salient and non-salient regions, respectively. In light of online dictionary learning [20], we develop an online salient dictionary learning (OSDL) algorithm to solve the proposed formulation, and then the salient and non-salient dictionaries can be learnt from the eye tracking data of training images. Given the learnt dictionaries, a novel feature based on sparse representation of learnt texture atoms (SR-LTA) is worked out in our method. Such a feature is generally based on the errors of sparse representation regarding salient and non-salient dictionaries. Next, the saliency of an image can be predicted, via combining the SR-LTA feature with conventional luminance and contrast features. For the linear combination, the weights corresponding to each feature channel are estimated via least square fitting on the training data. Similar to other bottom-up methods [21,17,18], this paper only works on gray images with natural scenes.

In summary, the main contributions of this paper are two-folds:

- We address a novel dictionary learning formulation solved by the proposed OSDL algorithm, for generalizing salient and non-salient dictionaries from training eye tracking data.
- We propose the SR-LTA feature in light of the learnt dictionaries, together with other two conventional features (luminance and contrast), for bottom-up saliency detection of gray images.

This paper is the extended version of our conference paper [22], with some advanced work. The advances are summarized as follows. First, the related work of saliency detection is extensively reviewed, from biologically inspired and learning based aspects. Second, this paper provides technical details about the derivation of our method, e.g., the derivation of dictionary updating in our ODSL algorithm. Third, we analyze the computational time of our method, by comparing to other methods. At last, we provide more comprehensive comparison and analysis in this paper. For example, we compare our method with the latest work of [23], and show that our method still outperforms [23] in bottom-up saliency