Image and Vision Computing 32 (2014) 180–188

Contents lists available at ScienceDirect

Image and Vision Computing

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

Ensemble dictionary learning for saliency detection☆

Zhenfeng Zhu *, Qian Chen, Yao Zhao

Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China

ARTICLE INFO

Article history:
Received 19 April 2013
Received in revised form 22 October 2013
Accepted 30 December 2013

Keywords:
Saliency detection
Dictionary learning
Sparse representation
Ensemble learning

ABSTRACT

The human visual system (HSV) is quite adept at swiftly detecting objects of interest in complex visual scene. Simulating human visual system to detect visually salient regions of an image has been one of the active topics in computer vision. Inspired by random sampling based bagging ensemble learning method, an ensemble dictionary learning (*EDL*) framework for saliency detection is proposed in this paper. Instead of learning a universal dictionary requiring a large number of training samples to be collected from natural images, multiple over-complete dictionaries are independently learned with a small portion of randomly selected samples from the input image itself, resulting in more flexible multiple sparse representations for each of the image patches. To boost the distinctness of salient patch from background region, we present a reconstruction residual based method for dictionary atom reduction. Meanwhile, with the obtained multiple probabilistic saliency responses for each of the patches, the combination of them is finally carried out from the probabilistic perspective to achieve better predictive performance on saliency region. Experimental results on several open test datasets and some natural images demonstrate that the proposed *EDL* for saliency detection is much more competitive compared with some existing state-of-the-art algorithms.

1. Introduction

The human visual system (HSV) has a remarkable ability to quickly detect the salient regions in complex static or dynamic scenes and can easily understand scenes based on this selective functionality. In recent years, simulating human visual system to detect visually salient regions of an image has been arousing great research interests in computer vision. A wide range of potential applications of the saliency detection technology encompass image/video compression [1], image segmentation [2] and retrieval [3–5], video analysis [6], object recognition, detection and tracking [7,8], and so on.

In light of the massive studies in the past years in neuropsychology, the deployment of visual attention has long been believed that there are two different approaches in visual processing mechanism: bottom-up approach and top-down approach.

Bottom-up approach, which is also known as data-driven processing and task independent, means that the sensory information is analyzed in one direction: from simple analysis of raw sensory data to ever increasing complexity of analysis through the visual system. Lots of studies have attempted to explain on this area by observing the correlation between fixations of observers and basic features such as edge and local contrast [9,10]. The most classical saliency model was proposed by Itti et al. [11]. This model was based on the Feature Integration Theory (FIT) of Treisman and Gelade [12] and used a Difference of Gaussians (DoG) approach combining three kinds of low level features, i.e., intensity, color, and orientation, to determine center-surround contrast. Based on Itti's model, Harel et al. [13] proposed a graph-based visual saliency model to highlight conspicuous parts and permit combination with other maps. Ma and Zhang [14] proposed an approach that used color feature contrast analysis, and developed a fuzzy growing algorithm to extract conspicuous regions from the saliency map.

On the contrary, top-down approach is related to the recognition process according to the prior knowledge such as tasks having been performed and the feature distribution of the object. The basic idea of this model is that various basic features are extracted from the scene and subsequently integrated into the representation of saliency map. Inspired by the theory of visual routines, Sprague and Ballard [15] proposed a top-down attention model based on RL (reinforcement learning) to make eye movements of an operator in virtual environments clear. Kanan et al. [16] proposed saliency detection method SUN using natural statistics to estimate the probability of a target at every location.

1.1. Related works

As in other computer vision tasks, the visual representation problem in saliency detection still keeps to be one of the significant issues. In

☆ This paper has been recommended for acceptance by Nicu Sebe.
* Corresponding author.
E-mail address: zhfzhu@bjtu.edu.cn (Z. Zhu).

0262-8856/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2013.12.015

general, pixel-based and patch-based visual representations are two popularly adopted ways. In patch-based saliency detection algorithms, a patch is utilized as the basic representation unit instead of a pixel. For each of the patches, some low level visual features such as color and texture are usually extracted to form the patch-based visual representation [17–19].

Inspired by the recent development of sparse coding in the field of machine learning, some saliency detection algorithms based on sparse representations of patches have been proposed. In order to integrate multiple types of features for detecting saliency collaboratively, Lang et al. [20] have posed saliency detection as a problem of multi-task learning and proposed a saliency detection algorithm *MTSP*, i.e., multi-task sparsity pursuit. Although good performance has been reported, it didn't give the real running time for detecting saliency region of a given image. In the work of Han et al. [5], they denoted the weighted residual using sparse coding length as saliency. For each patch, its sparse coding can be obtained by taking its surrounding patches as dictionary.

Assuming that an image is composed of redundancy (background) part and saliency (foreground) part, the 'sparse + low-rank' matrix decomposition technique has been applied for saliency detection by Yan et al. [21]. For their proposed *SCSP* algorithm, it mainly consists of two steps: using the learned over-complete sparse bases to represent image patches and detecting saliency by 'sparse and low-rank' matrix decomposition. However, since the learned over-complete dictionary needs to be pre-trained by using a large number of randomly collected natural image patches as training samples, it lacks of considering the inherent discrimination of salient region from background of an input image to be dealt with, which will make its scalability to the different input images insufficient. In addition, if the resolution of the input image is relatively high, the 'sparse and low-rank' decomposition of a large scale matrix will be involved. Thus, the high computational complexity with it won't be avoided. Meanwhile, the stability or convergence of the above matrix decomposition might not be guaranteed.

1.2. Main contributions

To address the aforementioned limitations with dictionary learning based saliency detection algorithms, the following points highlight several contributions of the paper:

- Instead of training a universal dictionary, multiple over-complete dictionaries with good scalability to the input image itself are learned independently, generating more flexible multiple sparse representations for each of image patches. With the learned multiple over-complete dictionaries, a novel ensemble dictionary learning framework for saliency detection is proposed.
- Within the proposed ensemble dictionary learning framework, the task of saliency detection is posed as a novelty detection problem. At the heart of this framework lies a good probabilistic interpretation for combining multiple dictionary-driven saliency response.
- To obtain more 'representative' atoms that can well characterize samples from background exclusively, we present a reconstruction residual based method for dictionary atom reduction. Thus, the distinctness of salient patch from background region can be further boosted.

1.3. Organization

The remainder of the paper is organized as follows. In Section 2, some preliminaries for notation definitions and dictionary learning are presented. The overview of the proposed ensemble dictionary learning (*EDL*) framework for saliency detection is illustrated in Section 3. Section 4 gives some detail discussions of the proposed *EDL* framework for saliency detection. Some experimental results and analyses on publicly available test datasets and natural scenes can be found in Section 5. Finally, we give the concluding remarks in Section 6.

2. Preliminaries

2.1. Notations

Let's begin with introducing some useful notations. Throughout the paper, we use bold uppercase letter to denote matrix and bold lowercase letter to denote vector. Let $\mathbf{X} = [\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n]\in\mathbb{R}^{d\times n}$ be the input data matrix (training dataset), where $\mathbf{x}_i\in\mathbb{R}^{d\times 1}$ is a data instance. We use $||\mathbf{A}||_F = (\sum_{i=1}^{m}\sum_{j=1}^{n}A[i,j]^2)^{1/2}$ to denote the Frobenius norm of matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$. For a vector $\mathbf{a}\in\mathbb{R}^m$, we define its ℓ^2 norm by $||\mathbf{a}||_2 = (\sum_{i=1}^{m}a[i]^2)^{1/2}$, ℓ^1 norm by $||\mathbf{a}||_1 = \sum_{i=1}^{m}|a[i]|$, and ℓ^0 norm by $\|\mathbf{a}\|_0 = \#\{j,a[j]\neq 0\}$, which counts the number of nonzero entries in the vector $\mathbf{a}$ and is a pseudo-norm in fact due to not satisfying the required axioms. When $\psi\subseteq\{1,2,\ldots,n\}$ is a finite set of indices, $\mathbf{A}_{:\psi}\in\mathbb{R}^{m\times(n\text{-}|\psi|)}$ stands for the sub-matrix of $\mathbf{A}$ without containing the columns of $\mathbf{A}$ corresponding to the indices in ψ. Similarly, for $\psi\subseteq\{1,2,\ldots,m\}$, $\mathbf{A}_{..\psi}\in\mathbb{R}^{m\times(n\text{-}|\psi|)}$ denotes the sub-matrix of $\mathbf{A}$ without containing the rows of $\mathbf{A}$ corresponding to the indices in ψ.

2.2. Dictionary learning for sparse representation

A common way to represent real-valued data is with a linear combination of a collection of basis functions, which are generally referred to as atoms of a dictionary $\mathbf{D} = [\mathbf{d}_1,\mathbf{d}_2,\ldots,\mathbf{d}_K]\in\mathbb{R}^{d\times K}$ with each column $\mathbf{d}_i$ being an atom. Considering a data sample $\mathbf{x}\in\mathbb{R}^d$, we say that it admits a sparse representation or approximation over *dictionary* $\mathbf{D}$ when only a few of the selected atoms of dictionary are involved in the linear combination. Particularly, if the number of atoms is much larger than the dimensionality of data, i.e., $K\gg d$, the dictionary $\mathbf{D}$ can be called over-complete dictionary.

An over-complete dictionary that leads to sparse representations can be chosen as a predefined set of transform basis functions. Such is with the cases like the DFT, DCT, orthogonal wavelet, and some other transforms. Although powerful, the limitation with them is also apparent since they may not be favorable of characterizing the intrinsic structure of signals under consideration. An alternative approach, termed '*dictionary learning*', has received considerable investigations by inferring the dictionary directly from a set of existing data samples. Thus the learned dictionary can be well adapted to the purpose of sparse representation. Recent research progress has shown that the learning of data-driven dictionary quite outperforms those using a predefined one.

More precisely, given the training data-set $\mathbf{X} = [\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n]$, one can learn a dictionary $\mathbf{D}\in\mathbb{R}^{d\times K}$ by solving the following optimization problem [22]:

$$\langle D,R\rangle = arg\,\min_{\mathbf{D},\mathbf{R}}\underbrace{\|\mathbf{X}-\mathbf{D}\cdot\mathbf{R}\|_F^2}_{reconstructionerror}+\lambda\cdot\underbrace{\sum_{i=1}^{n}\|\mathbf{r}_i\|_0}_{sparseness} \tag{1}$$

where λ is a trading-off parameter to balance the reconstruction error term and sparseness penalty, and $\mathbf{R} = [\mathbf{r}_1,\mathbf{r}_2,\ldots\mathbf{r}_n]\in\mathbb{R}^{K\times n}$ denotes the sparse coding matrix with each column $\mathbf{r}_i$ being the sparse representation of data instance $\mathbf{x}_i$. Here, for the purpose of promoting the sparseness penalty, it is an intuitive way to adopt the ℓ^0 norm according to its definition. However, Eq. (1) is hard to solve due to its non-convex and non-smooth quality and has indeed been shown to be an NP-hard problem.