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

A R T I C L E  I N F O

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

A B S T R A C T

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.

© 2014 Elsevier B.V. All rights reserved.

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 is 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
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 $X = [x_1, x_2, ..., x_n] \in \mathbb{R}^{d \times n}$ be the input data matrix (training dataset), where $x_i \in \mathbb{R}^{d \times 1}$ is a data instance. We use $\|A\|_F = (\sum_{j=1}^{n} \sum_{i=1}^{d} |A[i,j]|^2)^{1/2}$ to denote the Frobenius norm of matrix $A \in \mathbb{R}^{n \times n}$. For a vector $a \in \mathbb{R}^{n}$, we define its $\ell^2$ norm by $\|a\|_2 = (\sum_{i=1}^{n} |a[i]|^2)^{1/2}$, $\ell^1$ norm by $\|a\|_1 = \sum_{i=1}^{n} |a[i]|$, and $\ell^0$ norm by $\|a\|_0 = \#\{a[i] \neq 0\}$, which counts the number of non-zero entries in the vector $a$ and is a pseudo-norm in fact due to not satisfying the required axioms. When $\psi \subseteq \{1, 2, ..., n\}$ is a finite set of indices, $A_\psi \in \mathbb{R}^{n \times \#(\psi)}$ stands for the sub-matrix of $A$ without containing the columns of $A$ corresponding to the indices in $\psi$. Similarly, for $\varphi \subseteq \{1, 2, ..., n\}$, $A_\varphi \in \mathbb{R}^{n \times \#(\varphi)}$ denotes the sub-matrix of $A$ without containing the rows of $A$ corresponding to the indices in $\psi$.

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 $D = [d_1, d_2, ..., d_K] \in \mathbb{R}^{d \times K}$ with each column $d_i$ being an atom. Considering a data sample $x \in \mathbb{R}^d$, we say that it admits a sparse representation or approximation over dictionary $D$ 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 $D$ 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 $X = [x_1, x_2, ..., x_n]$, one can learn a dictionary $D \in \mathbb{R}^{d \times K}$ by solving the following optimization problem [22]:

$$\begin{align*}
(D, R) &= \arg \min_{D, R} \max_{x} \frac{\|X - XD - DR\|_F^2}{\text{reconstructionerror}} + \lambda \sum_{i=1}^{n} \|R[i,]\|_0 \\
\text{sparseness}
\end{align*}$$  \hspace{1cm} (1)

where $\lambda$ is a trading-off parameter to balance the reconstruction error term and sparseness penalty, and $R = [r_1, r_2, ..., r_n] \in \mathbb{R}^{K \times n}$ denotes the sparse coding matrix with each column $r_i$ being the sparse representation of data instance $x_i$. Here, for the purpose of promoting the sparseness penalty, it is an intuitive way to adopt the $\ell^0$ 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.
An alternative relaxation way is to replace the $\ell^0$ norm in Eq. (1) by using $\ell^1$ norm [23], yielding:

$$(D, R) = \arg \min_{D, R} \|X - DR\|_F^2 + \lambda \cdot \sum_{i=1}^n \|r_i\|_1$$  

(2)

The optimization problem Eq. (2) is in general not jointly convex with respect to variables $D$ and $R$. But if we fix one of them, either $D$ or $R$, the objective function with respect to the other variable becomes a convex function. Thus, to solve Eq. (2) will mainly consist of two steps: 1) Sparse coding with the available dictionary $D$; 2) Dictionary updating with the obtained sparse coding matrix $R$. The above two steps need to be implemented iteratively until a certain convergence condition is met. In fact, when we fix variable $D$ in Eq. (2), the sparse coding procedure can be transformed into a linear programming problem. For the dictionary updating step, the K-SVD algorithm [24] is applied due to its stability and efficiency.

With the leaned over-complete dictionary $D$, the sparse representation $r$ of the input data instance $x$ can be given by:

$$(r) = \arg \min_{r \in \mathbb{R}^{np}} \left(\|x - DR\|_2^2 + \lambda \cdot \|r\|_1\right)$$  

(3)

3. The proposed saliency detection framework based on ensemble learning

As a machine learning paradigm, ensemble methods try to construct a set of hypotheses from training data and combine them to provide a more reliable prediction on unknown data [25]. Ensemble learning has already been used in many applications, such as face recognition, text categorization, and computer-aided medical diagnosis, and demonstrated that a good ensemble model exhibited apparent advantage over a single one. In view of the powerful generalization ability of ensemble learning and in allusion to the limitations with SCSP, we proposed an ensemble dictionary learning (EDL) framework for salient region detection in this paper.

The overall framework is shown in Fig. 1. For an input image, it is first partitioned into abundant of overlapping patches $X = [x_1, x_2, \ldots, x_N]$, where $N$ is the total number of overlapping patches. In our case, the size of each patch is $n_p \times n_p$ and we set $n_p = 8$ throughout the paper.

Instead of directly learning a universal over-complete dictionary as in SCSP [21], $N_d$ over-complete dictionaries $D_i$’s, $i = 1, \ldots, N_d$, are independently learned with a small portion of randomly selected patches from $X$. Furthermore, to boost the distinctness of salient patch from the one of background, the reconstruction residual based dictionary atom reduction over the learned dictionary is proposed. Based on the multiple saliency responses for each of the patches, the combination of them is finally carried out in a probabilistic way to generate the saliency map for achieving better predictive performance on saliency region.

4. Ensemble dictionary learning

4.1. Dictionary learning based on random sampling

To improve the generalization performance of a model to be trained, random sampling based ensemble method, also known as ‘bagging’ [15], has shown promising potentiality by training individual sub-model independently on a random redistribution of the training set. Each sub-model’s training set is generated by randomly selecting, with replacement, a portion of original training samples. Many of the original examples may appear repeatedly in the resulting training set while others may be left out.

Following the idea behind bagging ensemble learning, we propose an ensemble dictionary learning approach with its individual over-complete dictionary to be learned on randomly selected samples. Specifically, as illustrated in Fig. 1, let $X_{trn}$ be a randomly generated training subset of the set of all overlapping patches with $i = 1, 2, \ldots, N_d$. Thus, a set of over-complete dictionaries $D_i$’s and the corresponding sparse representation $R_i$’s for all overlapping patches can be learned according to Eq. (2) and Eq. (3), respectively, by repeating such procedure $N_d$ times. In fact, the proposed ensemble dictionary learning has a good probabilistic interpretation.

4.2. Probabilistic interpretation of ensemble dictionary learning

As previous works on saliency detection, a saliency map, usually in a probabilistic way, will be generated to indicate where the salient region locates. Consider the following joint probability formula:

$$P(x) = \sum_{i=1}^{N_d} P(H_i) \cdot P(x|H_i)$$  

(4)

Fig. 1. The framework of the ensemble dictionary learning based saliency detection.
where \( H_i \)'s are some instances in the hypothesis space \( H \). When we assume that the parameter hypothesis space \( H \) to be dictionary space \( D \), the right hand of Eq. (4) can be seen as a probabilistic ensemble with \( P(x|H_i) \) being a conditional probability given the learned dictionary \( H_i \). Hereafter, to make symbol expression consistent throughout the paper, we will use \( D \) to substitute for \( H_i \). The detail about the derivation of \( P(x|D') \) will be given in the following part.

### 4.3. Dictionary atom reduction

Given the learned dictionary \( D \), we use the following Gaussian distribution to model the conditional probability \( P(x|D') \) of the input patch \( x \) to be dealt with:

\[
P(x|D') = C \cdot \exp \left( -\frac{\|x-D' \cdot r\|_2^2 + \lambda \|r\|_1}{2\sigma^2} \right)
\]  

where \( \sigma \) is a scale parameter and \( C \) is a constant. In fact, the above Gaussian distribution has been popularly adopted and shown to take favorable discrimination ability in sparse representation based classification.

Assume that the salient region generally occupies a fraction of area compared with the background region in a scene. Thus, much more samples (patches) will come from background region than the former. Based on this assumption, saliency detection can be treated as a novelty detection problem with those patches from salient region to be outliers. From this viewpoint, to make the overall reconstruction residual as in Eq. (2) reach its minimum, there will exist some of the atoms of the learned dictionary, called 'representative atoms', that tend to be highly shared by background samples but not those samples from salient region. As opposed to those representative atoms, the remaining 'non-representative atoms' will partially play an insignificant role in minimizing the reconstruction residual of background samples but may be significant for samples from a salient region. In other words, if these 'non-representative atoms' are taken away, bigger reconstruction residual will occur with high probability on samples from a salient region. It also means that the distinctness of a salient region from background can be boosted in this case. Aiming at achieving this goal, we propose a dictionary atom reduction method over an over-complete dictionary.

For a given dictionary \( D = [d_1, d_2, ..., d_K] \in \mathbb{R}^{d \times K} \) and the corresponding sparse coding matrix \( R \in \mathbb{R}^{K \times N} \), let \( f(j) \) denote the support frequency counter of the \( j \)th atom \( d_j \) that is defined by:

\[
f(j) = \frac{\|R_{j\cdot}\|_0}{N}
\]  

where \( R_{j\cdot} \) denotes the \( j \)th row of \( R \). In addition, we define the reconstruction residual evaluator \( E(j) \) over training set \( X \) without involvement of the \( j \)th atom \( d_j \) by:

\[
E(j) = \|X-D \cdot R_{j\cdot}\|_F^2
\]  

where the \( X \) stands for \( X - D \cdot R \). In fact, by some mathematical manipulation, Eq. (7) can be simplified into:

\[
E(j) = Tr([X' - X] + 2Tr[d_j' \cdot X \cdot r_j'] + d_j' \cdot d_j \cdot r_j') \cdot r_j
\]  

which will be with much lower computational complexity.

Obviously, both the support frequency counter \( f(j) \) and the reconstruction residual evaluator \( E(j) \) defined above can be good indicators to reflect the confidence of the \( j \)th atom \( d_j \) being one of the representative atoms. Thus, it is straightforward to combine them by:

\[
S(j) = \alpha \cdot E(j) + (1-\alpha) \cdot f(j)
\]  

where \( \alpha \in [0,1] \) denotes a weight. Fig. 2 shows an example of an over-complete dictionary and the corresponding sparse coding coefficients of training samples. According to Eq. (9), the \( j \)th atom \( d_j \) with pink color will be selected as one of 'representative atoms' since it is highly shared by training samples, especially by those from background. The detail implementation for obtaining the reduced set of representative atoms \( D_{\varphi} \) from \( D \) is given in Algorithm 1.

---

**Algorithm 1: Dictionary Atom Reduction (DAR)**

**Input**: The over-complete dictionary \( D = [d_1, ..., d_K] \in \mathbb{R}^{d\times K} \) and the corresponding sparse coding matrix \( R = [r_1, ..., r_N] \in \mathbb{R}^{K\times N} \).

1. For \( j = 1:K \),
   1. Compute the support frequency counter \( f(j) \) according to Eq. (6).
   2. Compute the reconstruction residual evaluator \( E(j) \) according to Eq. (7).
   3. Obtain the overall confidence \( S(j) \) of the \( j \)th atom being representative one according to Eq. (9).
2. End for
3. Sort \( S(j) \), \( j = 1, ..., K \), in ascending order.
4. Choose the leading 15% atoms to be "non-representative atoms" and let \( \varphi = \{j = 1, 2, ..., K\} \) be the index set of them in the original atom set.

**Output**: The dictionary \( D_\varphi \) and corresponding sparse coding matrix \( R_\varphi \).
With the obtained reduced set $D'_{\phi}$ of dictionary atoms, we can rewrite Eq. (5) by:

$$P_r(x|D'_{\phi}) = C \cdot \exp \left( -\frac{\|x - D'_{\phi} \cdot r^\phi \|^2 + \lambda \|r^\phi\|_1}{2\sigma^2} \right)$$

(10)

For convenience of illustration, we name the conditional probability $P_r(x|D'_{\phi})$ by probabilistic saliency response.

4.4. Saliency map generalization

Before generalizing the final saliency map based on the ensemble of individual probabilistic saliency response $P_r(x|D'_{\phi})$, a slight modification on $P_r(x|D'_{\phi})$ is first made to further highlight the salient region by:

$$\tilde{P}_r(x|D'_{\phi}) = \begin{cases} T & \text{if } P_r(x|D'_{\phi}) \geq T \\ P_r(x|D'_{\phi}) & \text{otherwise} \end{cases}$$

(11)

where $T$ is a threshold. Fig. 3 illustrates an example to adaptively determine the threshold $T$ by finding the percentile $p$ of $P_r(x|D'_{\phi})$.

Based on the modified $\tilde{P}_r(x|D'_{\phi})$, the final saliency map following Eq. (4) can be given by:

$$\text{Map}(x) = \sum_{i=1}^{N_d} P(D^i) \cdot \tilde{P}_r(x|D'_{\phi})$$

(12)

4.5. Summarization of EDL model for saliency detection

In light of the presentations given above, the pseudocodes of the proposed EDL model for saliency detection is summarized in Algorithm 2.

**Algorithm 2: Ensemble Dictionary Learning for Saliency Detection**

**Input**: The image $I$ for saliency detection, the number $N_d$ of dictionaries to be ensemble.

1. Partition the input image $I$ into $N$ overlapping patches: $X = [x_1, ..., x_N]$.
2. For $i = 1 : N_d$:
   1. Randomly select $N_{trn}$ patches from $X$ to build the training dataset $X_{trn}^i = [x_{trn}^i, ..., x_{N_{trn}}^i]$.
   2. Learn the $i^{th}$ dictionary $D^i$ according to Eq. (2).
   3. Compute the sparse coding matrix $R^i = [r_{trn}^i, ..., r_N^i]$ according to Eq. (3).
   4. $[D'_{\phi}, R_{\phi}^i] = DAR(D^i, R^i)$
5. For each $x_j \in X$, $j = 1, ..., N$, compute the probabilistic saliency response $\tilde{P}(x_j|D'_{\phi})$ according to Eq. (10).
6. End for

**Output**: The saliency map $\text{Map}(x) = \sum_{i=1}^{N_d} P(D^i) \cdot \tilde{P}_r(x|D'_{\phi})$ according to Eq. (12)

5. Experimental results and analyses

5.1. Experiment setup

5.1.1. Datasets

To evaluate the effectiveness of our model, we carry out the experiments on three public image datasets including Bruce, Achanta and MIT datasets, which have been popularly served as the benchmark for evaluating visual saliency detection performance. Bruce dataset consists of a variety of 120 images ranging from indoor and outdoor scenes. To provide the eye fixations, the images were presented to 20 human subjects in random order and the eye movements made within the first 4 s were recorded. Achanta dataset, which was originally published by Achanta et al. [20], is a sub-dataset of MSRA dataset that contains 1000 images.
and the corresponding binary images labeled by human. The last MIT dataset contains 1003 images and the corresponding eye tracking data were collected when the 1003 images were shown to 15 different users.

5.1.2. Baselines
Particularly, six state-of-the-art saliency detection algorithms, i.e., AC [27], FT [26], GBVS [13], ICL [28], IT [11] and SCSP [21] were compared with our proposed EDL model.

5.1.3. Feature extraction
For each of the partitioned overlapping patches of an input image, a 240-dimensional feature vector was extracted to form its low-level representation which includes 192-dimensional color feature vector \( (8 \times 8 \times 3) \) from R, G and B channels and 48-dimensional texture feature vector from all the channels that contains contrast, correlation, energy and homogeneity texture feature. The number \( K \) of atoms of a dictionary is set \( K = 500 \). In addition, one third of all the overlapping patches were randomly selected with replacement to serve as the training dataset for learning multiple dictionaries independently, i.e., \( N_{cm} = N/3 \) (See step 3 of Algorithm 2). The percentile \( p \) for adaptively determining the threshold \( T \) in Eq. (11) is uniformly set to be 0.85.

5.1.4. Evaluation metrics
In order to make quantitative evaluation, the area under the Receiver Operating Characteristic curve (AUC) [10] was used as the criterion [29,13,30,31], which indicates how well the saliency map predicts actual human eye fixations. Perfect prediction corresponds to an AUC of 1, while chance performance occurs at an area of 0.5. In addition, two kinds of Correlation Coefficients (CC) between the ground truth map \( M_p \) and the predicted saliency map \( M_{\hat{p}} \), i.e., micro Correlation Coefficient \( CC_{micro} \) and macro Correlation Coefficient \( CC_{macro} \), were utilized for evaluation. Here, the micro Correlation Coefficient refers to the conventional correlation coefficient [32]:

\[
CC_{micro} = \frac{M_{\hat{p}} \odot M_p}{\left\| M_{\hat{p}} \right\|_F \cdot \left\| M_p \right\|_F}
\]

where \( \odot \) denotes the Hadamard product. Different from the micro Correlation Coefficient, the macro Correlation Coefficient reflects the correlation in macroscopic view and is defined as:

\[
CC_{macro} = \frac{A_0 \cap A_{\hat{p}}}{A_0 \cup A_{\hat{p}}}
\]

where \( A_0 \) and \( A_{\hat{p}} \) stand for the area of the binarized ground truth map and the area of the binarized predicted saliency map, respectively.

5.2. Results and analyses

5.2.1. Performance analysis on ensemble dictionary learning
In the aspect of ensemble learning, random sampling based ‘bagging’ method has demonstrated much more promising performance than each of its individual component for some applications like image retrieval and object classification, etc. For the saliency detection task under consideration, we first made performance comparisons of the proposed ensemble dictionary learning method and its individuals.

As illustrated in Fig. 4, the first column shows some original input images. The second to the fourth columns (b)–(d) are the saliency maps generated by \( \hat{P}_i(\mathbf{x}|D^{(i)}) \), \( i = 1,..,3 \), respectively, and the fifth column (e) corresponds to the map by the ensemble of them based on Eq. (12). Taking the hand-drawn binary map as benchmark (the last column (g) in Fig. 4), it is not hard to find that the ensemble of multiple \( \hat{P}_i(\mathbf{x}|D^{(i)}) \)'s obviously outperforms its individuals. In addition, the saliency map (the sixth column (f)) generated by using all the overlapping patches to train the dictionary shows no advantage over each of \( \hat{P}_i(\mathbf{x}|D^{(i)}) \)'s. These results are also consistent with the above clarification for the merit of ‘bagging’ ensemble learning method to some extent.

Table 2

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Datasets</th>
<th>AC</th>
<th>FT</th>
<th>GBVS</th>
<th>ICL</th>
<th>IT</th>
<th>SCSP</th>
<th>EDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>Bruce</td>
<td>0.5608</td>
<td>0.5358</td>
<td>0.7952</td>
<td>0.6576</td>
<td>0.7289</td>
<td>0.6840</td>
<td>0.7601</td>
</tr>
<tr>
<td></td>
<td>Achanta</td>
<td>0.7895</td>
<td>0.8021</td>
<td>0.8219</td>
<td>0.7716</td>
<td>0.6334</td>
<td>0.7444</td>
<td>0.8010</td>
</tr>
<tr>
<td></td>
<td>MIT</td>
<td>0.6677</td>
<td>0.5674</td>
<td>0.7234</td>
<td>0.7790</td>
<td>0.7050</td>
<td>0.7667</td>
<td>0.7907</td>
</tr>
</tbody>
</table>

Table 1

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Datasets</th>
<th>AC</th>
<th>FT</th>
<th>GBVS</th>
<th>ICL</th>
<th>IT</th>
<th>SCSP</th>
<th>EDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>Bruce</td>
<td>0.5608</td>
<td>0.5358</td>
<td>0.7952</td>
<td>0.6576</td>
<td>0.7289</td>
<td>0.6840</td>
<td>0.7601</td>
</tr>
<tr>
<td></td>
<td>Achanta</td>
<td>0.7895</td>
<td>0.8021</td>
<td>0.8219</td>
<td>0.7716</td>
<td>0.6334</td>
<td>0.7444</td>
<td>0.8010</td>
</tr>
<tr>
<td></td>
<td>MIT</td>
<td>0.6677</td>
<td>0.5674</td>
<td>0.7234</td>
<td>0.7790</td>
<td>0.7050</td>
<td>0.7667</td>
<td>0.7907</td>
</tr>
</tbody>
</table>
Unless otherwise specified, the number \( N_d \) of dictionaries to be ensembled is uniformly set to be 3 in the following experiments.

5.3. Comparison to state-of-the-art algorithms

Table 1 reports the average AUC values of all the compared algorithms. It can be seen that the proposed EDL performs well to find the salient objects compared with other six state-of-the-art saliency detection algorithms. Particularly, EDL achieves the best performances on Bruce and MIT datasets while GBVS shows its more competitive advantage over others on Achanta dataset. Meanwhile, we also evaluate the correlation between the ground truth and the predicted saliency maps. For both macro Correlation Coefficient and micro Correlation Coefficient, we can observe from Table 2 that the results are consistent with those based on the criterion of AUC. Note that 100 thousands of training samples for SCSP have

Fig. 5. Performance comparisons on some natural images. From the left to the right are: (a) Original image and saliency maps produced by (b) AC, (c) FT, (d) GBVS, (e) ICL, (f) IT, (g) SCSP and (h) EDL.
been randomly collected from natural images to learn a universal over-complete dictionary off-line.

5.4. Visual evaluation on some natural scenes

In addition to the quantitative evidence to show the effectiveness of the proposed ensemble dictionary learning for saliency detection, we also conduct visual evaluation on some natural scenes. Fig. 5 shows performance comparisons on some natural images. Visually, the proposed EDL model achieves better saliency detection results. Since each of the multiple independent dictionaries was trained based on random sampling, we also evaluated the stability of EDL on a sequence of video showing a running wolf in a clutter background with weeds. As we can see from Fig. 6, the EDL model takes good stability of continuously capturing object even in a complex scene.

5.5. Computational complexity analysis

As illustrated in Section 4, the proposed EDL model consists mainly of four procedures, random dictionary learning based on randomly selected samples, sparse coding, dictionary atom reduction, and saliency map generalization. Essentially, all of these procedures should be implemented online. In the case of computational complexity, random dictionary learning and sparse coding will take most of the computation cost. Different from EDL model, to obtain the sparse representation for each of the patches, the dictionary learning in SCSP model can be carried out off-line. What need to be done online for SCSP are sparse coding and ‘sparse + low-rank’ decomposition.

Taking an image of size 296 x 400 as an example, Table 3 shows the main computation costs of EDL and SCSP for the saliency detection. Here, we use $T(D)$, $T(R)$ and $T(L + S)$ to denote the computation costs of dictionary learning, sparse coding and ‘sparse + low-rank’ decomposition, respectively. As stated above, we set here the number of the ensembled dictionaries $N_d = 3$. In addition, the test platform is based on Matlab2007.

In order to boost the scalability of the over-complete dictionary for forming a sparse representation in SCSP model, much more training samples collected from natural scenes are required. With the increase of the number of the training samples, a higher computational cost for dictionary learning will be unavoidable. But for EDL model, the computational cost for learning an over-complete dictionary can be satisfactorily reduced since only a small portion of paths from the image to be dealt with ($N_{trn} = N/3$ in our work as mentioned in Section 5.1(C))

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Off-line</th>
<th>On-line</th>
<th>Total time on-line</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCSP</td>
<td>437.26 s</td>
<td>27.98 s</td>
<td>105.18 s</td>
</tr>
<tr>
<td>EDL</td>
<td>7.68 s</td>
<td>25.91 s</td>
<td>34.59 s</td>
</tr>
</tbody>
</table>

Fig. 6. A sequence of video containing a running wolf in a clutter background with weeds. From the left to the right are: (a) Original image and saliency maps produced by (b) AC, (c) FT, (d) GBVS, (e) ICL, (f) IT, (g) SCSP and (h) EDL.
are used to serve as training samples. Meanwhile, during dictionary learning, these training samples can simultaneously obtain their corresponding sparse representation. Thus, the computational cost for following sparse coding procedure can also be reduced correspondingly. Compared with SCSP, the proposed EDL can nearly save half of the time as shown in Table 3.

It should be noted that the efficiencies of those methods with on-line learning involved, like EDL, SCSP [21], and MTSP [20], are not nearly so satisfactory as those without requiring online learning. For example, for some state-of-the-art algorithms such as FT [26], GBVS [13], ICL [28], and IT [11], they will generally take only 1–2 s to fulfill the task of saliency detection. One way to improve the efficiency of EDL can be to implement EDL on a reduced resolution of the input image like other state-of-the-art algorithms. In addition, a more efficient sparse coding method will also be helpful for efficiency improvement since it takes the vast majority of the total time as shown in Table 3.

6. Conclusions

Inspired by ensemble learning, we developed an ensemble dictionary learning (EDL) framework for saliency detection. Within this framework, randomly selected samples are used to learn multiple self-contained over-complete dictionaries. By treating saliency detection as a problem of novelty detection, we proposed a reconstruction residual based model for atom reduction over the learned dictionary to further boost the distinctness of salient patch from the one of background. Finally the multiple independent dictionary-driven saliency response models were assembled in a probabilistic way to obtain the saliency map. We experimentally evaluate our proposed model on several open test datasets and natural images, and also compare with some previous models. As indicated in the experimental results, the proposed EDL method preferably outperforms over some existing state-of-the-art algorithms.

Acknowledgments

This work was supported in part by the National Basic Research Program of China (no. 2012CB316400), the National Natural Science Foundation of China (no. 61025013, no. 61172129, and no. 61210006), the Program for Changjiang Scholars and Innovative Research Team in University (no. IRT201206), the Program for New Century Excellent Talents in University (no. 13-0661), the Fundamental Research Funds for the Central Universities (no. 2012JB2012), and the Natural Science Foundation of Beijing, China (no. 4112043). The authors would like to thank the anonymous reviewers for their constructive and valuable comments.

References