Salient Region Detection via Discriminative Dictionary Learning and Joint Bayesian Inference

Shigang Wang, Min Wang*, Shuyuan Yang, *Senior Member*, *IEEE*, Kai Zhang

***Abstract*—In the past decades, saliency detection has received increasing attention from computer vision communities, for its potential usage in many vision related tasks. However, findingrepresentative and discriminative features to accurately locate salient regions from complex scenes remains a challenging problem. Recent research on primary visual cortex(V1) shows that vision neurons are sparsely connected to form a compact representation of natural scenes and different visual stimuli are processed separately according to their semantic importance. Inspired by the above characteristics of visual perception, in this paper we advance a novel saliency detection method via Representative and Discriminative Dictionary Learning (RDDL). An assumption that salient and non-salient information are sparsely coded under two separate dictionaries is cast on the problem and we proposeto learn a compact background dictionary from the image itself for saliency estimation. Different from previous methods, our saliency cues are obtained via active learning strategies rather than artificially designed rules, thus is more adaptive. Followed by this, a probabilistic inference model is deduced to fully excavate multi-source information of the scenes for high quality saliency map generation. This joint inference scheme takes both spatial and color space information into consideration and is proved to be quite effective in practice. Finally, to investigate the performance of the proposed model, some experiments are conducted on two benchmark datasetsalong with other 20 state-of-the-art saliency detection approaches. Experimental results show that our method outperforms its counterparts and can correctly detect salient regions, even when other methods fail. Besides, the usability of the proposed method in real application based cases is verified by applying it to content based image resizing and promising results are obtained.**

***Index Terms*—Visual saliency modeling, discriminative dictionary learning, joint Bayesian inference, image retargeting.**

I. INTRODUCTION

VISUAL saliencyis a remarkable ability of primates by which one's attention can be attracted to certain regions in a scene. Massive visual big data enter our eyes every day,most of which are however useless for our cognition purposes. Handling the spatial-temporal high dimensional visual big data is quite a daunting and exhausting work for our limited brain resource. However, with the help of the selective attention mechanism, we can consciously or unconsciously focus on the regions of interest and allocate more computational resource to them for rapid scene understanding. Inspired by this, developing computationally plausible models to mimic this intelligent behavior has been a hot topic in computer vision for decades.

It is commonly accepted that visual saliency is driven both by a task-independent bottom-up manner and a task-dependent top-down manner [9]. Bottom-up saliency is considered as the pop-out in a scene during free look and our eyes are unconsciously attracted to the regions of interest, which generally contain meaningful semantic information for cognition. While, top-down saliency is dominated by specific search goals and only the regions that we are going to find will attract our attention. Due to the subjective (goal driven) character of top-down saliency, relatively small amount of work has been done to study this behavior from computational perspective. Instead, extensive works can be found in the literature to model bottom-up saliency [3], [5], [9], [23], [28], [40]for its invariance to personal preferences. Bottom-up saliency is itself a large category and basically contains two research directions, fixation point detection and salient object detection. In this paper, we mainly focus on extracting salient regions from complex scenes.

Since the underlying driven force of saliency is still not clear, in order to model it from computational perspective, it is required that reasonable assumptions should be made and also reliable cues be found. Existing salient region detection methods generally make strong assumptions on the problem and thus see a model mismatch under diversified situations. Also, the saliency cues they found are mostly intuitive and lack robustness in complex scenes. Along with the advancement of saliency research, these problems become even more severe. In order toobtain satisfactory detection results, more meaningful assumptions and reliable cues should be developed to build new computational models. On the other hand, findings on how biological visual information is processed are the source of inspirations that can motivate us to design effective computational models. Studies from biological and cognitive science demonstrate that visual information is sparsely represented under some learned basis functions or dictionaries [27] to form abstracted high level knowledge for scene understanding. Also, cells in V1 respond separately to different visual stimuli and visually important information are processed independently from the others to extract discriminative

The authors are with Xidian University, Xi'an, Shaanxi Province, 710071, China (e-mail: wangmin@xidian.edu.cn). This work was supported by the National Basic Research Program of China (973 Program) under Grant No. 2013CB329402, the Major Research Plan of the National Natural Science Foundation of China (No. 91438103, 91438201), the fundamental research funds for the Central Universities under Grant No.BDY021429, the National Natural Science Foundation of China (No. 61405150, 61072108, 61173090, 61501353), Natural Science Foundation of Ningbo under Grant No. 2015A610244..

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

features.Exploiting the representative and discriminative characteristics will contribute to the design of more advanced models.Motivated by the above observations, in this paper we propose a heuristic discriminative dictionary learning approach for saliency detection. Our basic assumption is that salient and non-salient information are sparsely coded under two independent dictionaries. We propose to learn a discriminative background dictionary whichfavors the sparse representation of the non-salient rather than salient information. In this way, the sparse reconstruction errors under the learned background dictionary can be used as reliable saliency measures.To achieve this, we first learn a representative global dictionary for the entire scene and then find atoms from it to form the discriminative background dictionary with guidance from the low-rank property of border regions. Compared with previous methods, our assumption on the problem can be more easily satisfied in practice and also the saliency cues in our model are automatically learned from the image rather than artificially designed. Therefore, it will be capable of providing more satisfactory salient region detection results from complex real world scenes.

Similar ideas can be found in the literature to model saliency using sparse and low rank structures, such as the low-rank matrix recovery model LRMR [10] and sparse and dense representation model DSR [42]. LRMR assumes that under a certain feature space, salient regions are sparse and background regions have low rank structure, and casts saliency detection as a low-rank matrix decomposition problem. Robust principal component analysis (RPCA) is used to solve the generalized low-rank matrix decomposition problem and the l_1-norm of each column in learned sparse matrix is taken as saliency measure. However, it is relatively difficult to find a good feature space to meet the above requirement, especially in complex scenes. In this paper, we assume that salient and non-salient information are processed independently and they are sparsely coded under two separate dictionaries. With the help of domain knowledge, we want to find a representative and discriminative background dictionarywith which to provide accurate saliency estimation. Though both adopt low-rank structure in their models, they have different motivations and moreover cast different assumptions on this problem. In LRMR, the low-rank constraint is cast on the feature space to recover the low-rank structure of background regions and meanwhile detect foreground regions as sparse outliers. Saliency is considered as the singularity part of the scene and measured by the decomposed sparse matrix. Different from LRMR, the low-rank constraint in our model is applied on the sparse coding coefficient space to facilitate the learning of a discriminative background subspace. The modeling bias caused by the complex scene layout can be compensated by the low-rank analysis so as to provide reliable saliency cues for detection. On the other hand, DSR assumes that there is a large difference between the reconstruction errors of foreground and background using the same sparse or dense bases. Templates from the image boundary are used to build the dense and sparse feature projection bases for saliency measure. However, simply depending on the image boundary information is not sufficient enough to learn a good basis to represent background ordiscriminate from foreground. Instead, our method makes full use of both foreground and background information to learn a representative and discriminative background dictionary. The learned background dictionary is only favorable for the projection of background regions even if the image boundary is partially corrupted by salient regions, making it robust in complex scenes.

Besides, based on the above error map, we propose a probabilistic inference model to make full use of multi-source information for high quality saliency map construction. In the previous works, probabilistic and statistical models are widely used in saliency detection and show their advantages in the optimization of saliency values.However, most of them adopt intuition inspired rules or depend on single source information for inference.It is believed that models that can simultaneously take multi-source information into account will be more helpful to this problem. Motivated by this, a joint Bayesian inference model is deduced for unified saliency prediction. Since the pop-out of saliency is highly related to spatial location and color appearance, we model the emergence of saliency as a stochastic event that depends on the above two attributes. Specifically, the spatial location and colorappearance are considered as two random variables and the conditional probability that saliency emerges given the two variables is used to quantify the saliency levels. Through deduction, we transform the conditional probability into a computationally plausible form, in which spatial and color information can be together guided by the above error map to produce more visually appealing saliency maps. As will be seen later, theconditional probability is finally determined by two color-related posterior probabilities and two position-related prior probabilities. This model formulation not only has good interpretability but also makes the elegant combination of multi-source information possible. Thus, it could be expected that this probabilistic inference model would boost the performance of our saliency detector.

The contribution of this paper is hence threefold. First, inspired by the basic visual perceptron characteristics, we propose a discriminative dictionary learning model for salient region detection from complex scenes. Foreground and background information are assumed to be sparsely coded under two separate dictionaries and we propose a heuristic approach to learn a discriminative background dictionary for accurate saliency prediction. Compared with existing methods, the proposed model can automatically learn robustsaliency cues for effective scene analysis. Secondly, a joint Bayesian inference model is developed to combine multi-source information for unified saliency map generation. We consider the emergence of saliency as a stochastic event that depends on spatial location and color appearance, and use a conditional probability to model this process. Through deduction, the conditional probability can be transformed to computationally plausible forms to favor the elegant fusion of multi-source information. Different from previous works, this model possesses good interpretability and meanwhile can make full

use of hybrid information for saliency modeling. Thirdly, the performance of the proposed method is compared with other 20 state-of-the-art saliency models on benchmark datasets andboth qualitative and quantitative results indicate its superiority. Motivated by its strongmodeling capability, we further extend our saliency model to seam carving and present a content-aware image resizing approach, which shows better performance than the traditional methods in real application based cases.

II. Related Work

Till now, a great number of works have been proposed to address this problem from different perspectives [2], [17], [30]-[32], [34]-[36], [38], [39]. Visual attention has found its applications in a broad range of scientific and engineering fields, such as targetdetection, visual surveillance, contrast enhancement, image/videosegmentation and retargeting, object discovery and so on [23]-[26], [29],[43]. Saliency modeling is a profound challenge because it is in nature a highly ill-posed problem. Most of the time, we have to depend on the image itself to predict which region attracts our attention the most. However, findings on how visual saliency is formed in cognitive science can provide cues for computational modeling, such as the center-surround contrast [9], feature or structure rarity [39], neuron state transition [5] and so on. Also, images generally follow certain rules in capturing natural scenes, such as the center prior [3], border effects [15], closure property [31] and so on, which can contribute to the modeling of the ill-posed problem. Generally speaking, the above cues formed the foundation of many saliency detection models. Starting from some of the basicprinciples, computationally plausible models can be built to locate regions of interest. Hence, it is believed that a good saliency model should incorporate meaningful knowledge into its computation process.

In [9], Itti developed a "central-surround" operation to highlight local pop-out with respect to illumination, orientation and color opponency. Harel et al. proposed a graph model to mimic the saliency transition process among neurons and recorded the equilibrium state as the final saliency measure [5]. Inspired by the idea that saliency originates from its rarity or singularity nature, Borji and Itti adopted local and global patch rarities as the saliency descriptors [39]. Since salient regions barely exist in image borders for some databases, in Bruce and Rahtu's saliency maps, border areas are left unprocessed and assumed to be non-salient [1], [15]. Instead of simply ignoring the borders while searching for salient regions, advanced models make use of the border information to guide their search process. Among them [41] and [42] are the most successful ones. In [41], the authors took border nodes as absorbing nodes and the absorbed time from each node to the absorbing nodes is used to measure saliency. Similarly, Li et al. constructed dense and sparse dictionaries from borders to approximate the subspaces of background [42]. Corresponding to border prior, center prior and object bias are also introduced to give a further refinement of the resulting saliency maps. In [3], Goferman et al. designed a 2D Gaussian filter to highlight the locations near the image center. Since center prior can work only in certain situations,object bias isproposed by Li et al. [42] to enhance the regions that are with high beliefs to be salient. However, it may also weaken the saliency of small salient regions in face of multiple objects. Recently, there is also a rising interest in taking the closure property of salient objects into consideration when modeling saliency. Geodesic saliency connected the length of the shortest path from each node to the virtual node with saliency. Since salient objects are commonly closed areas surrounded by background, further distance is needed for salient nodes to reach the virtual node [17]. Inspired by the recent progress in cognitive psychology, Zhang et al. [31] advanced an interesting theory called "Boolean map", which adopts a dynamic process to model the shifts in color and figure based saliency.

Along with the progress made in saliency modeling, more complex datasets are published, which provide new challenges for the researchers in this area. Traditional saliency cues may not be sufficient enough to cover these newly emerged characteristics. Consequently, a deeper understanding of the problem will be of great significance for developing robust models. In this paper, we try to develop a learning based saliency model that is capable of detecting salient regions in complex natural scenes. The rest of this paper is organized as follows. Section 3 gives a detailed description of the proposed saliency detection method. Experimental results and comparative analysis are shown in Section 4. Finally, we conclude this paper and point out possible future works in Section 5.

III. Dictionary Learning based Saliency Detection Model

Saliency detection can be viewed as a binary classification problem which usually lacks reliable external information for supervision. Learning models, such as independent component analysis (ICA), conditional random field (CRF) and support vector machine (SVM) borrow information from natural images and saliency database to build saliency detectors [1], [16], [33], [37]. However, bottom-up saliency is only concerned with the scene under observation. Thus the most reliable information for saliency modeling is perhaps the image itself. How to find representative and discriminative internal information for detection is the key towards a successful saliency model. In this paper, we try to design a "learn from the image itself" strategy to find concrete cues for saliency detection. Fig.1 shows the general framework of the proposed method.

For a given image, we assume that the salient foreground and non-salient background information in it are sparsely coded under two separate dictionaries. That is to say, under the space spanned by saliency features, the foreground and background regions lie in two different subspaces. Our goal is to learn a compact background dictionary which can not only represent background regions but also well discriminate from foreground regions. To achieve this, we first build a global dictionary learning model based on our sparse coding assumption. The learned global dictionary can sparsely code both salient and non-salient information. Therefore, it can be considered as the union of the foreground and background subspaces. This global

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

dictionary learning procedure corresponds to the "representative part" of our model and it is used to mimic the sparse and representative characteristics of visual perception. Followed by this, we try to find atoms from the global dictionary to compose a subspace, with which to approximate the true background subspace. This is a non-trivial problem if no extra knowledge is available. Luckily, there are still some priors we can develop to settle this problem. According to the basic photographic composition rule, salient objects will not be cropped along the view frame by most photographers. Therefore, for natural images, most of, if not all, the border regions will come from background, which is also verified on several salient object databases [17]. According to our assumption, most of them will choose atoms corresponding to background dictionary to get sparse representation. Thus, the matrix composed by their sparse coefficient vectors will have low-rank structure (plus random noise), and the non-zero rows in the decomposed low-rank matrix can indicate the positions of background atoms. In this way, a compact background dictionary can be formed by arranging all the selected atoms into a matrix. This background dictionary learning process corresponds to the "discriminative part" of our model and it is used to learn selective knowledge for attention. Based on this learned background dictionary, we can measure the saliency of each input using its sparse reconstruction error under the background dictionary.

The obtained error map is a prototype of the final saliency map with some undesired defects. To further improve the quality of error maps, we design a probabilistic inference model. This model is based on the heuristics that the emergence of saliency is closely related to the spatial location and color appearance of visual stimuli. We use a conditional probability to model the above relationship and by deduction we obtain a computationally plausible form for joint inference of saliency scores. This probabilistic inference process is based on the aforementioned error map and makes full use of the statistical properties of the image itself. Therefore, the defects caused by coarse segmentations or learning bias can be largely overcome and high quality saliency maps would be generated. In the following, we will give a detailed description of the proposed dictionary learning model and probabilistic inference scheme.

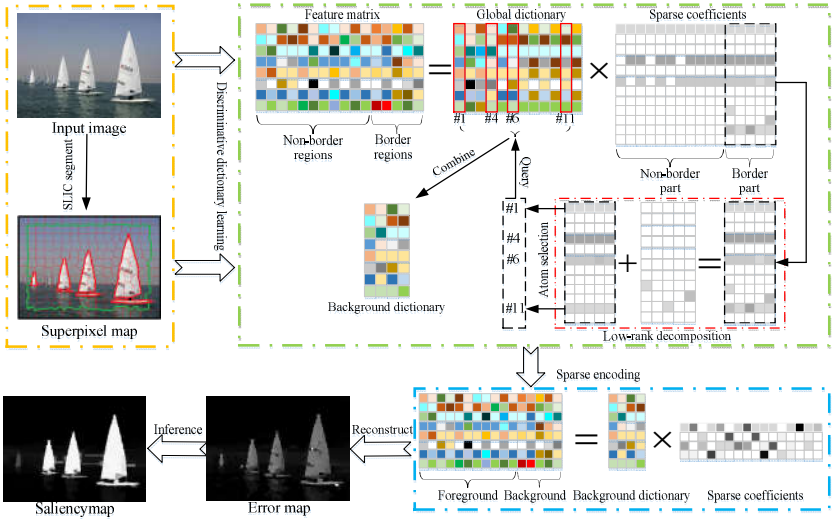


Fig. 1. General framework of the proposed visual saliency model.

A. Compact Background Dictionary Learning

Using the simple linear iterative clustering (SLIC) algorithm [19], we first segment the input image into several uniform superpixel regions, each of which is considered as a basic element. The LAB color space, as is verified by the experiments, is more suitable as saliency feature than the RGB color space, and the union of them can give better detection results than each of the single space. In this paper, the union of CIE Lab and RGB color spaces is used as the feature space, and each visual stimuli can be represented as a point in this feature space. For each superpixel, the feature values of the pixels inside it are averaged to describe it. Thus, the feature matrix of the image can be represented as $Y=[y_1,y_2,\cdots,y_n]\in\Re^{m\times n}$, where m is the dimension of the feature space and n is the number of superpixel regions after SLIC segmentation. Each column $y_i\in\Re^m, i=1,2,\cdots,n$ corresponds to a superpixel region and either belongs to salient or non-salient part of the image.

Visual information is found to be sparsely coded before being processed further and we use the following sparse model to mimic this information processing mechanism

$$(D^*,A^*)=\min_{D,A}\frac{1}{2}\|Y-DA\|_F^2+\lambda_1\|A\|_1 \quad (1)$$

where, $Y\in\Re^{m\times n}$ is the raw data to be processed, $D\in\Re^{m\times k}$ is the global encoder, $A\in\Re^{k\times n}$ is the sparse encoding result and λ_1 is the regularization parameter. Under our assumption, the global encode D can be divided into two irrelevant parts $D=[D_f,D_b]$, where $D_f\in\Re^{m\times k_1}$ is the sub-encoder for salient foreground and $D_b\in\Re^{m\times k_2}$ the sub-encoder for non-salient background. Correspondingly, if the feature matrices of foreground and background of the image are $F\in\Re^{m\times n_1}$ and $B\in\Re^{m\times n_2}$ ($n_1+n_2=n$), the model in Eq.1 can be rewritten as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

$$\min_{D_f, D_b, A_{ff}, A_{bf}, A_{fb}, A_{bb}} \frac{1}{2}\left\| \begin{bmatrix} F & B \end{bmatrix} - \begin{bmatrix} D_f & D_b \end{bmatrix} \begin{bmatrix} A_{ff} & A_{bf} \\ A_{fb} & A_{bb} \end{bmatrix} \right\|_F^2 + \lambda_1 \left\| \begin{bmatrix} A_{ff} & A_{bf} \\ A_{fb} & A_{bb} \end{bmatrix} \right\|_1 \tag{2}$$

where, A is divided into four blocks $A_{ff} \in \Re^{k_1 \times n_1}$, $A_{bf} \in \Re^{k_1 \times n_2}$, $A_{fb} \in \Re^{k_2 \times n_1}$ and $A_{bb} \in \Re^{k_2 \times n_2}$ which are cross-section coefficients matrices. By constraining A_{bf} and A_{fb} to be both zero matrices, the above problem can be divided into two independent sub-problems and from which the two sub-encoders D_f and D_b can be learned. However, this is intractable because in practice both F and B are not known due to the ill-posed nature of this problem. In fact, the above two sub-problems are complementary to each other and addressing either one of them is enough for saliency detection. In this paper, we turn to find a compact background sub-encoder D_b^* with the help of image border prior so as to approximate the true background sub-encoder. To do this, we first learn the global encoder D^* from Eq.1 and then adopt a heuristic strategy to select atoms from it to get the sub-encoder D_b^*.

The optimization problem in Eq.1 can be intuitively interpreted as decomposing the feature matrix of a given image into the multiplication of a dictionary and its corresponding sparse coefficients matrix. It can be solved efficiently using the online dictionary learning method in sparse modeling software (SPAMS) [20]. For a natural image, its feature vectors scatter around the original feature space and usually lie in a low dimensional subspace. The learned encoder D^* can be seen as an estimation of this low dimensional subspace. Followed by this, we use a heuristic approach to select atoms from the encoder D^* to form a compact background sub-encoder D_b^*. According to our assumption, background regions will choose atoms only from D_b^* to get sparse representation. Thus, their sparse coding coefficient matrix $A_b^* \in \Re^{k \times n_2}$ should have low-rank structure plus random noise, and the non-zero rows in the decomposed low-rank matrix indicate the positions of atoms in D_b^*. In practice, we do not know where the background lies in and instead use the image borders as surrogates because they cover a wide range of background regions. Denote the sparse coefficient matrix of the image borders as A_d^* and we perform low-rank decomposition to it using augmented Lagrange multiplier (ALM) method [21].

$$\min_{A_{dl}^*, A_{ds}^*} \left\| A_{dl}^* \right\|_* + \lambda_2 \left\| A_{ds}^* \right\|_1 \quad s.t.\ A_d^* = A_{dl}^* + A_{ds}^* \tag{3}$$

where, A_{dl}^* and A_{ds}^* are the low-rank and sparse parts of A_d^* and λ_2 is the weighting parameter. Ideally, if the image borders purely contain non-salient regions, the atoms in D^* that correspond to the non-zero rows of A_{dl}^* can be used to build a good sub-encoder D_b^*. However, in real-world situations, regions of interest are sometimes cropped at the image borders. In these cases, there will be a few non-zero rows in A_{dl}^* with small l_2 norms and their corresponding atoms in D^* do not belong to D_b^*. If they are included in the learned sub-encoder D_b^*, the independent condition will be severely damaged. Since these illegal atoms are selected for quite a limited number of times, the l_2 norms of their corresponding rows in A_{dl}^* are relatively small values. However, this is just the opposite for the legal atoms. Based on this, we define the following rule to select legal atoms for the sub-encoder D_b^*. Denote the set of column atoms in D^* as $\{d_i^* \in \Re^m \mid i = 1, 2, \cdots, k\}$ and their corresponding row vectors in A_{dl}^* as $\{a_i^* \in \Re^n \mid i = 1, 2, \cdots, k\}$. We want to select some atoms from D^* to form a subset D_b^* such that the following conditions should be satisfied.

$$\begin{aligned} &1: \forall d_i^* \in D_b^* \ and\ d_j^* \in D^* - D_b^*, \exists\ \left\| a_i^* \right\|_2 \geq \left\| a_j^* \right\|_2; \\ &2: \sum_{i=1, d_i^* \in D_b^*}^{k} \left\| a_i^* \right\|_2 \geq \eta \sum_{j=1}^{k} \left\| a_j^* \right\|_2, \ \eta \in (0,1]; \end{aligned} \tag{4}$$

With the above defined rule, we can select atoms from D^* to form a background sub-encoder $D_b^* \in \Re^{m \times p} (p \leq k)$. The decay factor η is set empirically and the choice of its value is universal for all the images. Thelearned D_b^* is an approximation to the true background sub-encoder and it contains more intrinsic information for saliency detection. Eq.4 will support the selection of atoms that are frequently used to represent the image border regions and meanwhile those atoms that are used for quite a limited number of times will be discarded according to the selection rule. This is especially important for compensating the bias caused by the image border prior. In practice, the existence of salient foreground in the image border regions will corrupt the intrinsic structure of the low rank matrix A_{dl}^*. Simply depending on the non-zero rows in A_{dl}^* for atom selection will bring outliers in the learned background sub-encoder. With help from the atom selection rule in Eq.4, the less frequently used illegal atoms will be excluded. In this way, the bias caused by the image border prior will be fixed and more reliable background sub-encoder can be learned for saliency detection. In the experimental section, we will verify the effectiveness of the atom selection rule in Eq.4 by comparing it with other pure low-rank methods in a quantitative manner. The general framework of the proposed background dictionary learning model is listed in Algorithm 1. In what follows, we will use this learned background sub-encoder (dictionary) to generate saliency maps.

In this dictionary learning model, a global encoder is first learned and then decomposed into two sub-encoders that correspond to the foreground and background regions respectively. By exploiting the low-rank property of the image border regions, we find an effective way to learn a discriminative background dictionary for saliency detection. The global encoder can be considered as the joint space that foreground and background lie in and we want to find from it a discriminative background subspace, which only favors the sparse representation of background regions. Since the image border regions cover a wide range of background areas, most of them will choose atoms from background subspace to get

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

representation according to the independent assumption. Thus, their sparse coding coefficients under the global encoder should have low-rank structure and the non-zero rows in the decomposed low-rank matrix indicate the possible positions of the background atoms. By selecting these atoms from the global encoder and rearranging them into a matrix, we can get a compact background sub-encoder. This learned background sub-encoder can approximate the true background subspace and is complementary to the foreground subspace. Thus, it can well represent the background regions and meanwhile discriminate from the foreground regions. Generally speaking, the global dictionary ensures that a representative space be provided for the image and the border regions can provide discriminative cues for the finding of the two subspaces. By making use of the representative and discriminative information, we can learn the two dictionaries that correspond to the salient foreground and non-salient background regions.

Algorithm1 Framework of the ProposedBackground Dictionary Learning Model

Input: The feature matrix $Y \in \Re^{m\times n}$, encoder size k, regularization parameter λ_1, weighting parameter λ_2 and decay factor η.

Step 1. Use SPAMS to learn the global encoder $D^* \in \Re^{m\times k}$ and sparse encoding result $A^* \in \Re^{k\times n}$ for the entire scene according to $(D^*, A^*) = \min_{D,A} \frac{1}{2}\|Y - DA\|_F^2 + \lambda_1 \|A\|_1$.

Step 2. Get the sparse coefficient matrix of the image border regions A_d^* and perform low-rank decomposition to it using ALM $\min_{A_{dl}^*, A_{ds}^*} \|A_{dl}^*\|_* + \lambda_2 \|A_{ds}^*\|_1 \ \ s.t.\ A_d^* = A_{dl}^* + A_{ds}^*$.

Step 3. Based on the A_{dl}^* obtained in step 2, select atoms from D^* to form a compact sub-encoder D_b^* according to the atom selection rule in Eq.4.

Output: The learned background sub-encoder (dictionary) $D_b^* \in \Re^{m\times p}$.

B. Saliency Emergence with Probabilistic Inference

Since our dictionary learning model exploits both the representative and discriminative information of the scene, reliable saliency cues can be found to address this highly ill-posed problem. According to our previous assumption, the learned sub-encoder D_b^* is only responsible for the sparse coding of non-salient background. Thus, the sparse reconstruction errors of non-salient regions under D_b^* will be much smaller than that of the salient regions. In this paper, we will use the sparse reconstruction errors to quantify the saliency scores. Given the learned sub-encoder D_b^*, each of the feature vector in Y is sparsely coded with it in the following way

$$\min_{\alpha_i} \|\alpha_i\|_0 \ \ s.t.\ y_i = D_b^* \alpha_i, \quad i = 1, 2, \cdots n \quad (5)$$

where, $\alpha_i \in \Re^p$ is the sparse encoding result of y_i. Relaxing the l_0 norm with convex l_1 norm and applying the Lagrange multiplier method, the above optimization problem can be rewritten as

$$\min_{\alpha_i} \frac{1}{2}\|y_i - D_b^* \alpha_i\|_2^2 + \lambda_3 \|\alpha_i\|_1, \quad i = 1, 2, \cdots n \quad (6)$$

where, λ_3 is the regularization parameter and the optimal solution to the above problem $\alpha_i^* \in \Re^p (i = 1, 2, \cdots, n)$ can be determined using Lasso [22], [45]. Till now, we can use the following reconstruction error to describe how suitable D_b^* is to encode each feature vector

$$error_i = \|y_i - D_b^* \alpha_i^*\|_2^2, \quad i = 1, 2, \cdots, n \quad (7)$$

where, $error_i$ is the reconstruction error of the i-th element. A large error means D_b^* is not suitable as an encoder for this element, implying that it is more likely to come from foreground regions. Therefore, the errors are organized into an error map (same size with the original image) and taken as a prototype of our saliency map. Some example error maps are shown in the second row of Fig.2.

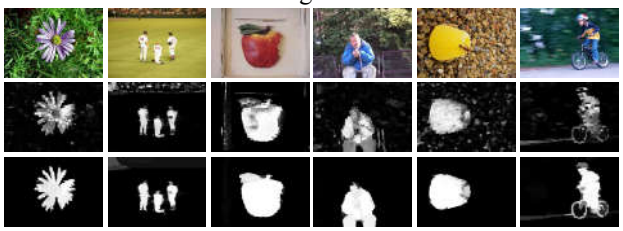


Fig.2. Some examples of the resulting error maps and saliency maps. The first row is the input images and second and third rows are the corresponding error maps and saliency maps.

As can be seen from the results, our error maps can generally give good estimations of where the regions of interest lie in. Besides, it is quite robust to the cases when there are multiple objects, cropped objects, cluttered background and motion blur in the scenes. This verifies that the proposed dictionary learning model can capture the intrinsic cues for saliency detection in diversified scenes. However, due to the following reasons, further processing is still needed to get high quality saliency maps. First of all, the error maps only show saliency in a coarse scale and also are not precise due to some incorrect segmentations. If the colors in the superpixel are not consistent, they will still be assigned with the same error value, which will lead to biased estimations in the error maps. Furthermore, the mismatch of model with complex real world situations should be compensated by extra operations. Finally, high belief saliency maps (foreground to background contrast) are needed to favor both qualitative and quantitative evaluations. Motivated by this, we propose a probabilistic inference model, which can combine multi-source information to enhance the above detection results.Fig.3 outlines the proposed probabilistic inference model with a simple example.

Since saliency is the competitive result of various physical attributes in a scene, it can be modeled as a stochastic event. For a visual stimuli with spatial position l and color value x (we use Lab color because it is more consistent with human perception habits), the probability that it can pop-out to attract our attention given its spatial position l and color value x can be denoted as $p(f \mid l, x)$. According to Bayesian formula, $p(f \mid l, x)$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

can be written in the following form

$$p(f|l,x)=\frac{p(l,x|f)p(f)}{p(l,x)} (8)$$

Since l and x are two uncorrelated attributes for saliency detection, we assume that they are independent variables in the above formula and thus we have

$$p(f|l,x)=\frac{p(l|f)p(x|f)p(f)}{p(l)p(x)} (9)$$

where, $p(l|f)$ is the probability that a salient stimuli comes from the spatial position l. Again, with the Bayesian formula, $p(l|f)$ can be written as

$$p(l|f)=\frac{p(f|l)p(l)}{p(f)} (10)$$

By substituting Eq.10 into Eq.9, we have

$$p(f|l,x)=\frac{p(f|l)p(x|f)}{p(x)} (11)$$

Since l and x are two independent variables, according to probability theory the following formula holds true

$$\underbrace{p(x)=p(x|l)}_{x \text{ and } l \text{ are independent}}=\underbrace{p(x|f)p(f|l)+p(x|b)p(b|l)}_{p(f)+p(b)=1} (12)$$

After substituting Eq.12 into Eq.11, we can get the final probabilistic inference model as follows

$$p(f|l,x)=\frac{p(x|f)p(f|l)}{p(x|f)p(f|l)+p(x|b)p(b|l)} (13)$$

where, $p(x|f)$ and $p(x|b)$ represent the probabilities that a salient and non-salient stimuli has color value x, and $p(f|l)$ and $p(b|l)$ represent the probabilities that a visual stimuli pops out or is inhibited given its spatial position l.

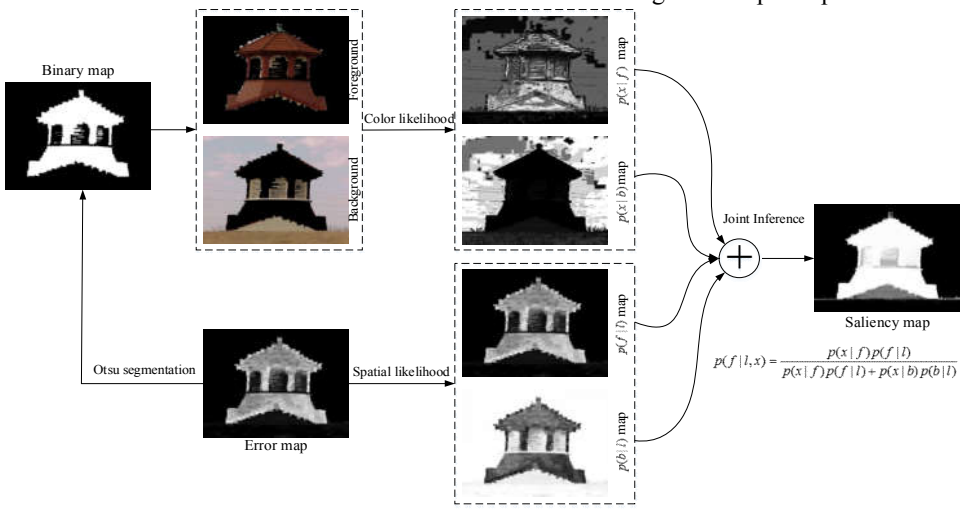


Fig.3. General framework of the proposed probabilistic inference model.

In order to obtain $p(f|l,x)$, the four likelihood probabilities $p(x|f)$, $p(x|b)$, $p(f|l)$ and $p(b|l)$ should be determined in advance. Since the above obtained error map indicates how likely a salient stimuli occurs in a spatial location, it is reasonable to derive the two probabilities $p(f|l)$ and $p(b|l)$ from the error map. To achieve this, we first normalize the values of the error map into [0, 1] and for each spatial location, its $p(f|l)$ is set to the corresponding value in this normalized error map. Also, $p(b|l)$ is determined according to $p(b|l)=1-p(f|l)$ based on the fact that a spatial location either pops out or is inhibited. The problem left is how to calculate the probabilities that a pop-out or inhibited visual stimuli has color value x, i.e., $p(x|f)$ and $p(x|b)$. We first segment the error map into two parts using Otsu algorithm [46] and the locations with values above the optimal threshold are considered as foreground and otherwise background. Each color channel of Lab space is divided into 12 bins and thus there are totally 12^3 color bins in the cubic space. In this way, each visual stimuli will belong to one of the 12^3 color bins based on its color value. For a color value x, its $p(x|f)$ and $p(x|b)$ is defined as

$$p(x|f)=\frac{n_{fx}}{n_f} \quad and \quad p(x|b)=\frac{n_{bx}}{n_b} (14)$$

where, n_f and n_b are the total number of foreground and background stimuli, and n_{fx} and n_{bx} are the number of stimuli in the foreground and background that fall into the same bin with x. Fig.4 shows the scatter plot of an example image in the Lab color space. As we can see from the figure, there is a tendency for the points from foreground regions to gather around and keep away from the points from background regions (also true for the points from background regions). By jointly learning from the nearby points in the Lab color space for inference, more comprehensive results can be obtained.

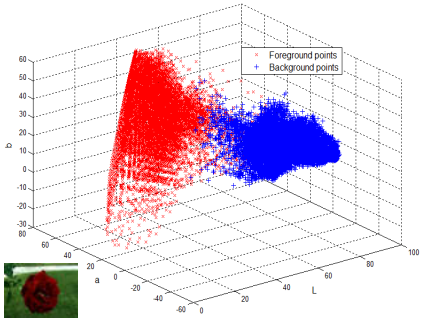


Fig.4. Scatter plot of an example image in the Lab color space. The original

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

image is put in the bottom-left corner for display convenience. Points from the foreground and background are represented with red cross and blue plus respectively.

The above probabilistic inference model is proved to be quite effective in practice. Let us consider the extreme condition when a non-salient pixel is incorrectly given high belief to be salient in the error map. Its probability of being salient after inference will be relatively small since its neighbors in the color space come mostly from background regions. As a result of this, this pixel will be given low belief to be salient by the above inference. Meanwhile, superpixel segmentation will inevitably bring misleading results especially along the object contours. The proposed strategy can achieve pixel level accuracy and also has well-kept object boundaries. Finally, to gain better visual effect, we adopt morphological operations called opening-by-reconstruction and closing-by-reconstruction to post-process the results [46]. Some of the saliency maps after probabilistic inference and morphological operations are shown in the third row of Fig.2. We can see that the defects in the error maps can be remedied with probabilistic inference and high quality saliency maps are consistently produced. Besides, in practice our "learn from the image itself" model is fast in speed. The whole process takes a few seconds per image, which can meet real-time requirements for many related tasks.

C. Implementation Details

The main parameters of our model are superpixelnumber n, encoder size k, regularization parameters λ_1 and λ_3, weightingparameter λ_2 and decay factor η. In the experiments, the above parameters are set to be $n=1200$, $k=16$, $\lambda_1=\lambda_3=0.2$, $\lambda_2=1/\sqrt{m}$ and $\eta=0.99$. We use the default settings for the maximum number of iterations in SPAMS and ALM. The radius $r$in the morphological operations is fixed to be 10. Also, before feature extraction, we rescale all the color channels into the same range [0,255]. All thefollowing experiments are conducted on a HP workstation with dual-core CPU of 2.40GHz frequency and 16GB memory under Windows 7 operation system and our codes are written in MATLAB and C.

We use the precision-recall curve (PR curve), average precision, recall and F-measures, and F-measure curve as quantitative indexes to evaluate the performance of saliency detection methods. For all the saliency methods, their resulting saliency maps are resized to the same size of the original images and with values into interval [0,255]. Using a threshold ranging from 0 to 255, we segment the saliency maps and get a series of binary maps. Each binary map is compared with the corresponding ground truth map to get a pair of precision and recall values. All the 256 precision-recall pairs in each database are averaged and plotted to get the PR curve. Besides, for each binary image, we use the following formula to determine its F-measure

$$F_\beta = \frac{(1+\beta^2)\times Precision \times Recall}{\beta^2 \times Precision + Recall} \quad (15)$$

where, we set $\beta^2=0.3$ to put more emphasis on precision than recall as suggested in [4].The average precision, recall and F-measures in each dataset are plotted in a bar graph to give an intuitive comparison. Also, F-measure curve is used to verify the effectiveness of each component in RDDL for saliency detection. For each threshold, an F-score can be computed and all the corresponding F-scores are averaged in each database and plotted to get the F-measure curve. Besides, statistical indexes, including mean absolute error (MAE) and area under curve (AUC), are also used to check the real relevance and impact of our proposal for saliency detection. AUC is the area under the receiver operator characteristic (ROC) curve and MAE is the average deviation of saliency maps from their ground truth maps, which is defined as follows

$$MAE = \frac{1}{N}\times\sum_{n=1}^{N}\frac{1}{p_n \times q_n}\sum_{i=1}^{p_n}\sum_{j=1}^{q_n}\left|S_n(i,j)-G_n(i,j)\right| \quad (16)$$

where, N is the total number of test images in the dataset, D_n and G_n are the n-th saliency map and its corresponding ground-truth map, all with size p_n and q_n and being normalized into [0,1].

The contrast of saliency maps is proved to be quite important for quantitative evaluations like PR curve [18]. If two maps contain exactly the same prediction results, the one with higher belief will be more favored in terms of PR measure. Therefore, a good saliency map should not only contain accurate predictions for saliency but also be with high beliefs. Fig.5 (a) and (d) shows an example image and its ground truth map. Our error map and saliency map are shown in Fig.5 (b) and (c) respectively. Since the probabilistic inference strategy combines multi-source information, it can enhance the beliefs of the previous prediction results. This improvement, as can be seen from the PR curves in Fig.6, leads to better evaluation result. In the followingexperiments, we will take this factor into consideration when evaluating the performances of saliency models.

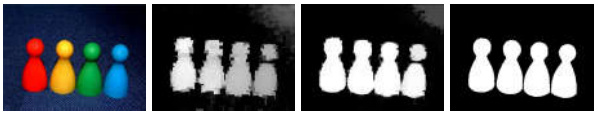


Fig. 5. Effect of contrast on PR measure. (a) original image, (b) our error map, (c) our saliency map and (d) the ground truth map.

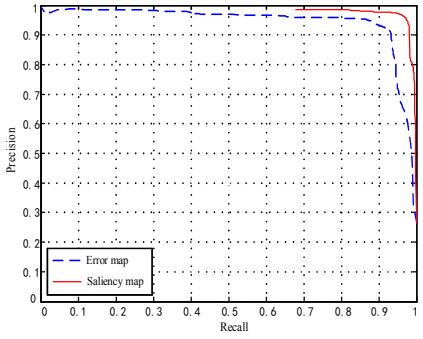


Fig. 6. PR curves of the error map and saliency map corresponding to Fig.5.

IV. Experimental Results

In this section, we evaluate our proposed model RDDL along with other 20 state-of-the-art saliency detection models on two publicly available datasets, MSRA-1000 [4], [37], [49] and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

THUS-10000 [32].For thecomparison methods, we denote them as AIM [1], CSP [2], CW [3], DSR [42], FT [4], GBVS [5], GRSD [6], HFT [7], ICL [8], IT [9], LG [39], LRMR [10], MSSS [11], SDSR [12], SR [13], SRDS [14], SRIV [15], SUN [16], GD and SP [17] respectively. These methods try to address this problem from different perspectives and have shown competitive detection performance. The experiments below are based on the codes and exe files provided by the authors and all the parameters are fixed and set according to the suggestions of the authors. We implemented GD and SP since their source codes are not publicly available yet and our results are quite close to the ones reported in [17]. In the following, both qualitative and quantitative results will be used to evaluate the performance of various saliency detection models.

A. Performance Evaluation on MSRA-1000

MSRA-1000 (also called as ASD) dataset is made publicly available by Achanta et al. [4]. It contains 1,000 images and their corresponding carefully labeled ground truth maps, and is one of the most widely used benchmark dataset forperformance evaluation. We first test the performance of various methods on it and some of the resulting saliency maps

by the top performing methods and our method are shown in Fig.7. Note that the ground truth mapsare not displayed because of space limitations and readers may refer to the labeled results provided by the corresponding authors. Since the background dictionary of DSR is built on the boundary regions, salient objects near the image boundary are not fully highlighted in their maps as can be seen from the third image in Fig.7. Also, salient objects are not uniformly highlighted in the saliency maps of LRMR (refer to the second and fourth images) and low beliefs are assigned to part of the foreground regions. Compared with other methods, our method can locate the salient objects more accurately and meanwhile gives higher contrast saliency maps. Besides, it is quite robust to the cases where there are multiple objects, cluttered background and cropped objects in the scenes. This verifies the effectiveness of the proposed dictionary learning and probabilistic inference models in modeling saliency.

Besides, in Fig.8 we plot the PR curves of our method along withthe other 20methods on MSRA-1000 for quantitative evaluations. For ease of observation, the PR curves of the 20 methods are shown in two subplots and our curve is shown in both. We can see from Fig.8 that the curve of RDDL keeps lying in the upper right most among all, indicating its better precision and recall performance. This numerical result is also quite in accordance with our qualitative assessment of the saliency maps. Besides, we divide each saliency map into a binary map using Otsu algorithm and the precision, recall and F-measure values are determined respectively. Fig.9 shows the average precision, recall and F-measures of various methods on MSRA-1000. As the bar graph shows, the average precision of RDDL is comparable to that of the DSR and GRSD, and RDDL achieves the highest recall and F-measure scores among all the 20 methods. The good performance of RDDL benefits from both the coarse scale dictionary learning and fine scale probabilistic inference. The error map provides a general estimation of where the regions of interest lie in and the saliency map after probabilistic inference accuratelydescribes to what degree the locations can be considered as salient.

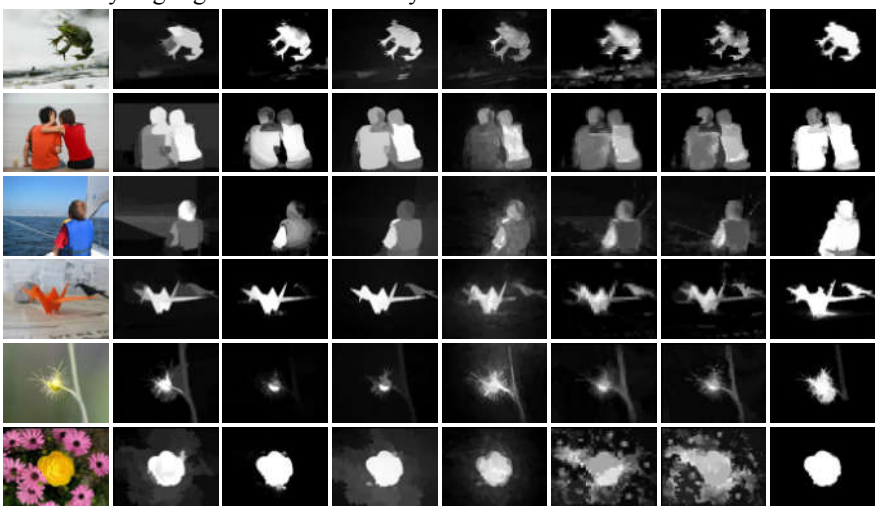


Fig. 7. Some saliency maps produced by the top performing methods on MSRA-1000. Each row corresponds to a test image. From left to right are the original image and saliency maps produced by CSP, DSR, GRSD, LRMR, GD, SP and RDDL.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

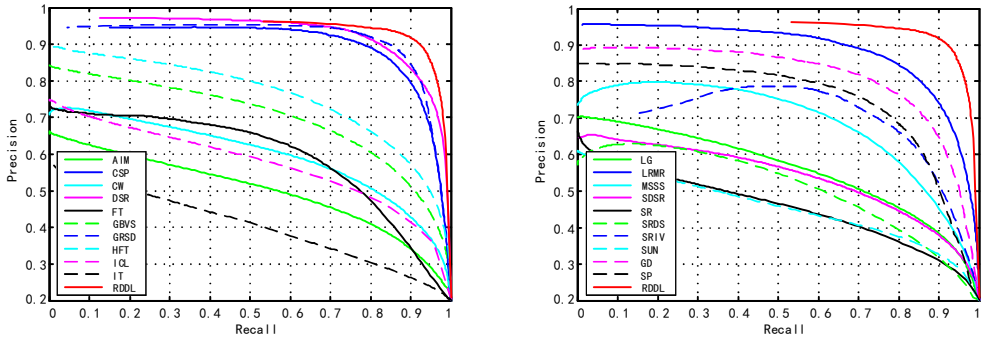


Fig. 8. Precision-Recall curves on MSRA-1000.

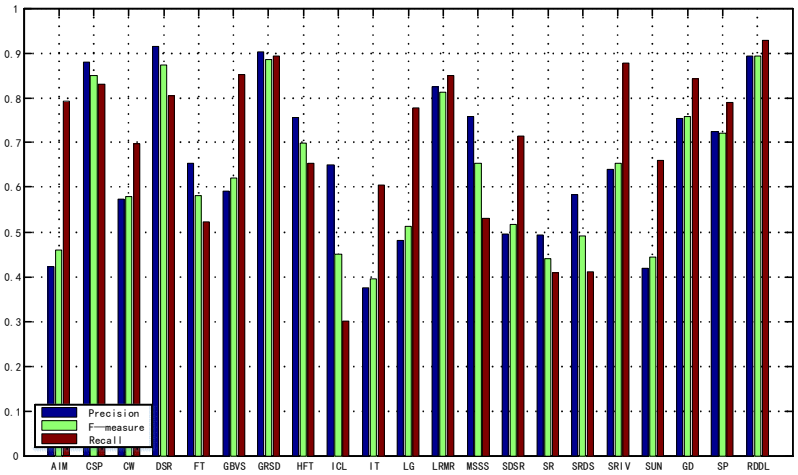


Fig. 9. Average precision, recall and F-measures on MSRA-1000.

B. Performance Evaluation on THUS-10000

THUS-10000 is made publicly available by Cheng et al. [32], which contains 10,000 images with pixel-wise labeled ground truth maps. It is a more challenging dataset with complex scenes and salient objects from it appear in quite different forms. For methods with high computational complexity, quite a long time is needed to process the 10,000 images. While, the time needed for RDDL is about 3 seconds per image under our experimental platform. Shown in Fig.10 are the saliency maps produced by the top performing methods and the proposed RDDL. We can see from the results that even under these difficult cases, RDDL is still capable of detecting regions of interest and giving more visually pleasing saliency maps.

Like in MSRA-1000, we also use PR curves as well as the average precision, recall and F-measures to evaluate the performance of various methods on THUS-10000. Fig.11 shows the PR curves of various methods on THUS-10000 and the average precision, recall and F-measure bars are plotted in Fig.12. We can see from the results that nearly all the methods see adrop in performance on this datasetdue to the complexscenes in it. Methods like AIM, GBVS, IT, LG, SDSR, SRIV and SUN have better recall than precision, i.e., they try to locate salient regions at the expense of introducing more false alarms. On the contrary, methods like DSR, FT, HFT, ICL, MSSS, SR and SRDS have better precision than recall, i.e., their detection results are mostly correct but only part of the salient object is detected. For other methods (including RDDL), their precision and recall values keep at nearly the same level, which means they can make a good compromise between precision and recall. In fact, this analysis is in accordance with the intuitive observation of the saliency maps. Similarly, the average F-measure of RDDL is the highest among all the comparative methods, which further confirms its potential in complex scene modeling.

Besides, we use the average MAE and AUC scores to give statistical evaluations of the various methods on the two datasets and the numerical results are listed in Table I. Smaller MAE and larger AUC mean better performance of saliency models. As we can see from the results, RDDL has better MAE and AUC scores than the other comparative methods in most cases. This implies that our method can produce higher quality saliency maps that are much more close to the ground truth maps.

C. Comparative Study and Model Verification

Since the dictionary learning and Bayesian inference models are two basic components in RDDL, in this subsection we will check their usefulness for saliency detection by quantitative evaluations on MSRA-1000. We denote our model when no Bayesian inference is involved as “without Bayesian” and when only border regions are used to learn the background dictionary as “Border”, and the corresponding PR and F-measure curves are shown in Fig.13.Besides, by keeping the other components unchanged, the usefulness of our atom selection rule is compared with these pure low-rank methods, including the exact and inexact Augmented Lagrange Multiplier (ALM) [21] (refer to “exact-ALM” and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

"inexact-ALM"), Accelerated Proximal Gradient (APG) [47], Dual Method (DM) [47] and Singular Value Thresholding(SVT) [48]. The PR and F-measure curves of the various low-rank based variants of RDDL are also shown in Fig.13.Finally, we denote our method when only the LAB color space is used as "lab-RDDL", which is also included for comparison.

As we can see from the results, the union of the LAB and RGB color spaces (RDDL) can provide better detection results than the single LAB color space (lab-RDDL).Despite the differences in optimization strategies, the five low-rank versions of RDDL (inexact-ALM, exact-ALM, APG, DM and SVT) provide nearly equivalent detection performance (Their curves are quite close to each other). As has been discussed previously, pure low-rank model is not sufficient enough to address this problem. Both the PR and F-measure curves in Fig.13 show that our atom selection rule provides more favorable detection results. Meanwhile, the background dictionary learned from the border regions only keeps part of the information for saliency detection and thus is less effective. And the probabilistic inference model, as is expected, has boosted the performance of our model with low computational cost.

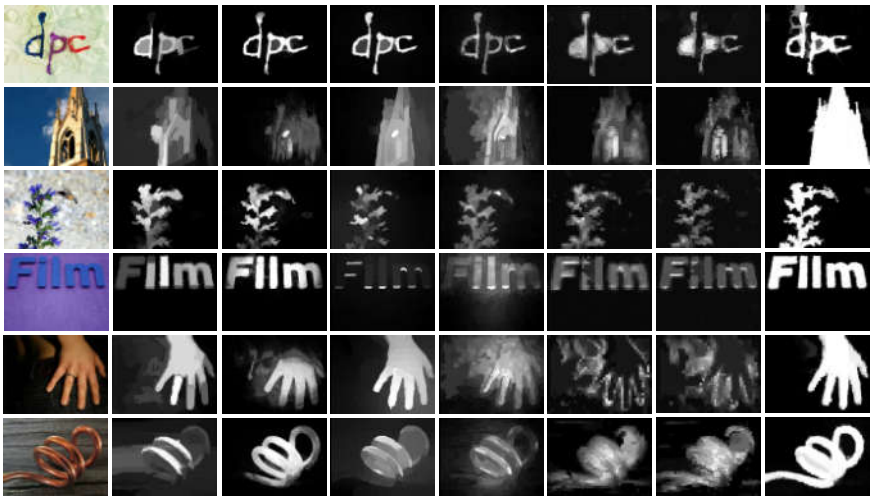


Fig.10. Some saliency maps produced by the top performing methods on THUS-10000. Each row corresponds to a test image. From left to right are the original image and saliency maps produced by CSP, DSR, GRSD, LRMR, GD, SP and RDDL.

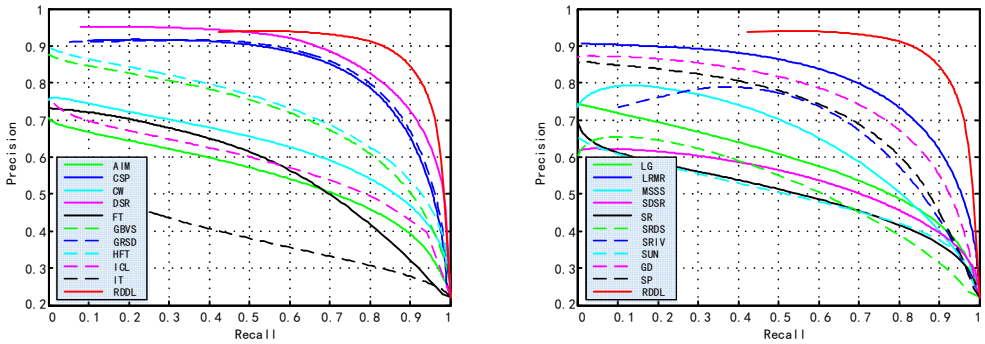


Fig.11. Precision-Recall curves on THUS-10000.

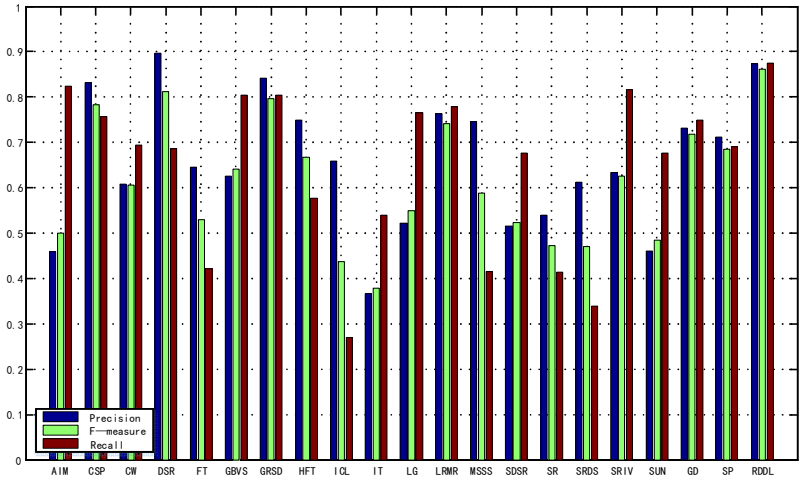
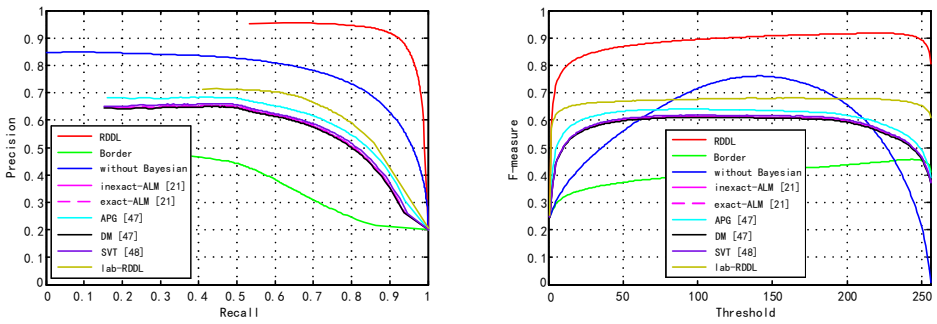


Fig.12. Average precision, recall and F-measures on THUS-10000.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

TABLE I
AVERAGE MAE AND AUCSCORES OF VARIOUS METHODS ON MSRA-1000 AND THUS-10000DATASETS

MSRA-1000 (MAE/AUC)	AIM	CSP	CW	DSR	FT	GBVS	GRSD
	0.2874/0.8101	0.1394/0.9625	0.2341/0.8705	0.0756/0.9784	0.2166/0.8200	0.2140/0.9218	0.1561/0.9737
	HFT	ICL	IT	LG	LRMR	MSSS	SDSR
	0.1769/0.9360	0.1955/0.8424	0.3443/0.6989	0.2914/0.8384	0.1857/0.9612	0.1830/0.8931	0.2332/0.8304
	SR	SRDS	SRIV	SUN	GD	SP	RDDL
	0.2308/0.7623	0.2143/0.7983	0.3051/0.9229	0.3112/0.7508	0.1812/0.9367	0.1960/0.9014	0.0747/0.9791
THUS-10000 (MAE/AUC)	AIM	CSP	CW	DSR	FT	GBVS	GRSD
	0.2863/0.8251	0.1776/0.9361	0.2371/0.8765	0.1207/0.9537	0.2348/0.7947	0.2223/0.9125	0.1976/0.9543
	HFT	ICL	IT	LG	LRMR	MSSS	SDSR
	0.2000/0.9189	0.2153/0.8329	0.3558/0.6566	0.2898/0.8409	0.2241/0.9277	0.2026/0.8780	0.2433/0.8235
	SR	SRDS	SRIV	SUN	GD	SP	RDDL
	0.2318/0.7426	0.2274/0.7638	0.3146/0.8888	0.3060/0.7694	0.2086/0.9057	0.2228/0.8686	0.1287/0.9614



(a)(b)

Fig.13. Comparative study and model verification results on MSRA-1000. (a) PR curves of RDDL, Border, without Bayesian, inexact-ALM, exact-ALM, APG, DM, SVT and lab-RDDL. (b) F-measure curves of RDDL, Border, without Bayesian, inexact-ALM, exact-ALM, APG, DM, SVT and lab-RDDL.

D. Model Applications and Future Work

With the increasing number of display devices, there is a demand for smarter ways to adjust the aspect ratios of images so as to better deliver visually important information under the given display dimension. Traditional methods simply resize the image or add black bars to fill the empty space, which will lower the visual perception quality and even destroy the regions of interest in the images. To address the above problem, Avidan and Shamir proposed an operator called seam carving for content-aware image resizing [44]. They pointed out that visual saliency can be a good measure for the energy in an image and thus may benefit the reduction or expansion process. Inspired by this, we use our saliency maps as energy measure in seam carving to guide the resizing process more user-friendly. We also compare our retargeting results with that of the resize operation and gradient based seam carving.

Fig.14 shows some retargeting results on images from TCD [38] and ECSSD [36] datasets when forty percent of the columns or rows are removed or added. From the results we can see that our method can keep the regions of interest intact during seam carving and moreover the retargeted images look more natural than that of the others. Conventional resize operation and gradient based method may easily distort the object regions, which will lower the visual perception quality. Since visual perception quality is closely connected with the regions we are interested in, our saliency model can be a useful tool for content-aware image retargeting.

Despite the satisfactory performance obtained, our method also has limitations in dealing with some specific situations. Since visual saliency is a complex problem, depending simply on knowledge from certain aspects will not grasp the entire nature of it. In the future, we will try to incorporate more meaningful cues into our modelto further enhance its performance and also extend this model to deal with saliency in video sequences.

V. CONCLUSION

Inspired by the representative and discriminative characteristics of human visual perception, in this paper a novel background dictionary learning based saliency detection method is proposed. Our assumption is that salient and non-salient information in a scene are sparsely coded under two separate dictionaries and we propose to learn a compact background dictionary that can well represent the background regions and at the same time discriminate from the foreground regions. To achieve this, we first learn a global encoder for the entire scene and then select atoms from it to build the sub-encoder for background with the help of the image border prior. The learned background dictionary takes information from the immediate context and can provide intrinsic cues for saliency detection. Besides, we proposed a probabilistic inference model that can combine multi-source information for more reliable and fine-grained saliency map generation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2016.2642341, IEEE Transactions on Circuits and Systems for Video Technology

Specifically, we use a conditional probability to describe the impact of physical attributes on the pop-out of visual stimuli. By deduction, computationally plausible forms can be obtained to favor the joint inference of saliency scores. Experimental results on two benchmark datasets verified the effectiveness of the proposed method and it constantly outperforms other 20 state-of-the-art saliency methods with respect to both qualitative and quantitative evaluations. The "learning from the image itself" strategy is quite efficient in practice and the whole process takes only a few seconds per image, which can meet real time requirements for many tasks. Besides, the effectiveness of each component in our model is investigated by the designed comparative experiments. Finally, we apply our saliency model to the image retargeting task and favorable results are obtained, implying its potential usage in attention related real world applications. Also, by analyzing the cases where our method fails to get satisfactory results, we pointed out possible future works along this research direction.

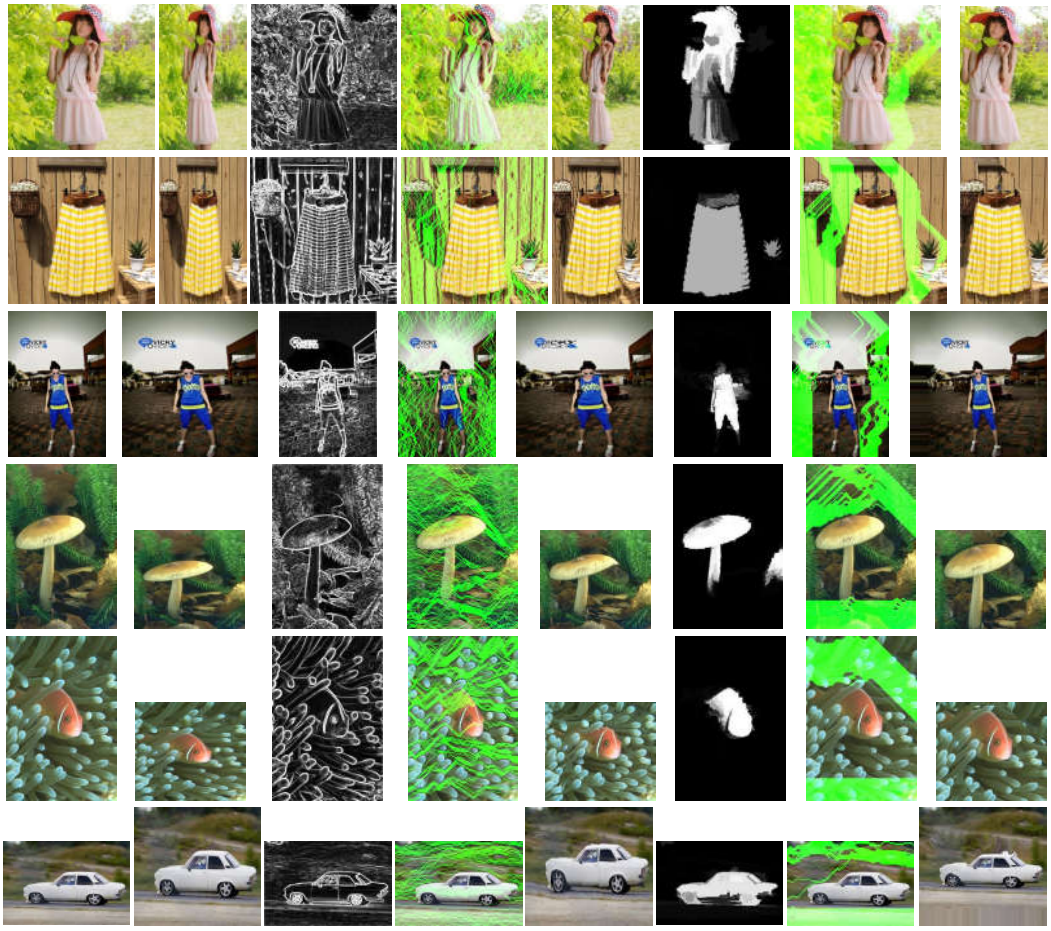


Fig.14. Retargeting results on TCD and ECSSD datasets. The images in each row correspond to the original image, result by resize operation, gradient map, seams from gradient map, result by gradient based seam carving, our saliency map, seams from our saliency map and result of our saliency map based seam carving. The first and second images are shown in 60% of the original width and the third image is shown in 140% of the original width. The fourth and fifth images are shown in 60% of the original height and the last image is shown in 140% of the original height.

REFERENCES

[1] N. Bruce and J. Tsotsos, "Saliency based on information maximization,"in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 155-162.

[2] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li,"Automatic salient object segmentation based on context and shape prior,"In*Proc.Brit. Mach. Vis. Conf.,* 2011, p. 7.

[3] S.Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection,"*IEEE Trans. Pattern Anal. Mach. Intell.,*vol. 34, no. 10, pp. 1915-1926, Oct. 2012.

[4] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection,"in *Proc.IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597-1604.

[5] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency,"in *Proc. Adv. Neural Inf. Process. Syst.,* 2007, pp. 545-552.

[6] C.Yang, L.Zhang, and H.Lu, "Graph-regularized saliency detection with convex-hull-based center prior,"*IEEE Signal Process. Lett.*, vol. 20, no. 7, pp. 637-640, 2013.

[7] J. Li, M. D. Levine, X. J. An, X. Xu, and H. G. He, "Visual saliency based on scale-space analysis in the frequency domain,"*IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 35, no. 4, pp. 996-1010, 2013.

[8] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments,"in *Proc. Adv. Neural Inf. Process. Syst.,* 2008, pp. 681-688.

[9] L. Itti, C. Koch, and E. Niebur,"A model of saliency-based visual attention for rapid scene analysis,"*IEEE Trans. Pattern Anal. Mach. Intell.,*vol. 20, no. 11, pp. 1254-1259, Nov. 1998.

[10] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery,"in*Proc. IEEE Conf. Comput. Vis. PatternRecognit.,* Jun. 2012, pp. 853-860.

[11] R. Achanta and S. Susstrunk, "Saliency detection using maximum symmetric surround," in*Proc. IEEE Int. Conf. Image Process.,* 2010, pp. 2653-2656.

[12] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 1-27, Nov. 2009.
[13] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach,"in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* 2007, pp. 1-8.
[14] R. Achanta, F. Estrada, P. Wils, and S. Susstrunk, "Salient region detection and segmentation,"in*Proc. ICVS*,May. 2008, pp. 66-75.
[15] E. Rahtu, J. Kannala, M. Salo, and J. Heikkila, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis.,* 2010, pp. 366-379.
[16] L. Zhang, M. Tong, T.Marks, H. Shan, and G. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics,"*J. Vis.,* vol. 8, no. 7, pp. 1-20, 2008.
[17] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors,"in *Proc. Eur. Conf. Comput. Vis.,* 2012, pp. 29-42.
[18] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proc. Int.Conf. Mach. Learning*, 2006, pp. 233-240.
[19] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 34, no. 11, pp. 2274-2282, 2012.
[20] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Machine Learning Research,* vol. 11, pp. 19-60, 2010.
[21] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Math. Program.*, 2009.
[22] J.Liu, S. W. Ji, and J. P.Ye, "SLEP: Sparse learning with efficient projections,"Arizona State University, 2009.http://www.public.asu.edu/~jye02/Software/SLEP, date of the last visit: Dec. 26, 2015.
[23] Y. Yu, B. Wang, and L. M. Zhang, "Hebbian-based neural networks for bottom-up visual attention and its applications to ship detection in SAR images," *Neurocomputing*, vol. 74, no. 11, pp. 2008-2017, 2011.
[24] M. T. Lopez, A. F. Caballero, M. A. Fernandez, J. Mira, and A. E. Delgado, "Visual surveillance by dynamic visual attention method," *Pattern Recognit.,* vol. 39, no. 11, pp. 2194-2211, 2006.
[25] K. Gu, G. T. Zhai, X.K. Yang, W.J. Zhang, and C. W. Chen, "Automatic contrast enhancement technology with saliency preservation,"*IEEE Trans. Circuits Syst. Video Technol.,* vol. 25, no. 9, pp. 1480-1494, 2015.
[26] C.C. Qin, G.P.Zhang, Y.C. Zhou, W.B. Tao, and Z. G. Cao, "Integration of the saliency-based seed extraction and random walks for image segmentation," *Neurocomputing*, vol. 129, pp. 378-391, 2014.
[27] B.A.Olshausen and D.J.Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607-609, 1996.
[28] W.L. Hou, X.B. Gao, D.C. Tao, and X.L. Li, "Visual saliency detection using information divergence," *Pattern Recognit.,* vol. 46, no. 10, pp. 2658-2669, 2013.
[29] W. Kim and C. Kim, "Spatiotemporal saliency detection using textural contrast and its applications,"*IEEE Trans. Circuits Syst. Video Technol.,* vol. 24, no. 4, pp. 646-659, 2014.
[30] J. W. Han, D.W. Zhang, X.T. Hu, L. Guo, J.C. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual,"*IEEE Trans. Circuits Syst. Video Technol.,*vol. 25, no. 8, pp. 1309-1321, 2015.
[31] J. M. Zhang and S. Sclaroff,"Exploiting surroundedness for saliency detection: A Boolean map approach," *IEEE Trans. Pattern Anal. Mach. Intell.,*vol. 38, no. 5, pp. 889-902, 2015.
[32] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. L. Huang, and S. M. Hu, "Global contrast based salient region detection,"*IEEE Trans. Pattern Anal. Mach. Intell.,*vol. 37, no. 3, pp. 569-582, 2015.
[33] J.W. Han, D.W. Zhang, S.F. Wen, L. Guo, T.M. Liu, and X.L. Li, "Two-stage learning to predict human eye fixations via SDAEs,"*IEEE Trans. Cybern.,* vol. 46, no. 2, pp. 487-498, Feb. 2016.
[34] J. Liu and S.J. Wang, "Salient region detection via simple local and globalcontrast representation,"*Neurocomputing*, vol. 147, pp. 435-443, 2015.
[35] J. Yan, M.Y. Zhu, H.X. Liu, and Y.C. Liu, "Visual saliency detection via sparsity pursuit," *IEEE Signal Process. Lett.,* vol. 17, no. 8, pp. 739-742, 2010.
[36] J.P. Shi, Q. Yan, L. Xu, and J.Y. Jia, "Hierarchical image saliency detectionon extended CSSD,"*IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 38, no. 4, pp. 717-729, 2015.
[37] T. Liu, Z.J. Yuan, J. Sun, J.D. Wang, N. N. Zheng, X. O. Tang, and H. Y. Shum, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 33, no. 2, pp. 353-367,2011.
[38] K.Z. Wang, L. Lin, J.B. Lu, C.L. Li, and K. Y. Shi, "PISA: Pixelwiseimage saliency by aggregating complementary appearance contrast measureswith edge-preserving coherence,"*IEEE Trans. Image Process.,*vol. 24, no. 10, pp. 3019-3033, 2015.
[39] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* 2012, pp. 478-485.
[40] Y.M. Fang, W.S. Lin, Z.Z. Chen, C.M. Tsai, and C.W. Lin, "A video saliency detection model in compressed domain,"*IEEE Trans. Circuits Syst. Video Technol.,*vol. 24, no. 1, pp. 27-38, 2014.
[41] B. W. Jiang, L. H. Zhang, H. C. Lu, C. Yang, and M. H. Yang, "Saliency detection via absorbing Markov chain," in *Proc. IEEE Int. Conf. Comput. Vis.,* 2013, pp. 1665-1672.
[42] X. H. Li, H. C. Lu, L. H. Zhang, X. Ruan, and M. H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.,* 2013, pp. 2976-2983.
[43] Y. Luo, J.S. Yuan, P. Xue, and Q. Tian, "Saliency density maximization for efficient visual objects discovery,"*IEEE Trans. Circuits Syst. Video Technol.,*vol. 21, no. 12, pp. 1822-1834, 2011.
[44] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing,"*ACM Trans. Graphics,* vol. 26, no. 3, p. 10, 2007.
[45] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statistical Soc. B,* vol. 58, no. 1, pp. 267-288, 1996.
[46] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, Digital Image Processing Using MATLAB. Tennessee: Gatesmark Publishing, 2009.
[47] Z.C. Lin, A. Ganesh, J. Wright, L.Q. Wu, M.M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix,"in *Proc. CAMSAP*, 2009, pp. 1-8.
[48] J.F. Cai, E.J. Candes, and Z.W. Shen, "A singular value thresholdingalgorithm for matrix completion,"*SIAM J. Optimization,*vol. 20, no. 4, pp. 1956-1982, 2010.
[49] T. Liu, J. Sun, N. N. Zheng, X. O. Tang, and H. Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.,* 2007, pp. 17-22.