Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Salient Object Detection via Low-Rank and Structured Sparse Matrix Decomposition

Houwen Peng[1], **Bing Li**[1], **Rongrong Ji**[2], **Weiming Hu**[1], **Weihua Xiong**[1], **Congyan Lang**[3]
[1] National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2] Department of Cognitive Science, School of Information Science, Xiamen University
[3] Department of Computer Science and Engineering, Beijing Jiaotong University
{houwen.peng, bli, wmhu}@nlpr.ia.ac.cn, {jirongrong, wallace.xiong}@gmail.com, cylang@bjtu.edu.cn

Abstract

Salient object detection provides an alternative solution to various image semantic understanding tasks such as object recognition, adaptive compression and image retrieval. Recently, low-rank matrix recovery (LR) theory has been introduced into saliency detection, and achieves impressed results. However, the existing LR-based models neglect the underlying structure of images, and inevitably degrade the associated performance. In this paper, we propose a Low-rank and Structured sparse Matrix Decomposition (LSMD) model for salient object detection. In the model, a tree-structured sparsity-inducing norm regularization is firstly introduced to provide a hierarchical description of the image structure to ensure the completeness of the extracted salient object. The similarity of saliency values within the salient object is then guaranteed by the ℓ_∞-norm. Finally, high-level priors are integrated to guide the matrix decomposition and enhance the saliency detection. Experimental results on the largest public benchmark database show that our model outperforms existing LR-based approaches and other state-of-the-art methods, which verifies the effectiveness and robustness of the structure cues in our model.

Introduction

Salient object detection is an emerging topic in computer vision as it provides an alternative solution to various image semantic understanding tasks, such as object detection (Rutishauser et al. 2004), region-based image retrieval (Itti, Koch, and Niebur 1998), and adaptive image compression (Cheng et al. 2011). A typical workflow of salient object detection involves detecting and extracting the most salient and attention-grabbing foreground object from the background. The output usually is a so-called "saliency map" where the intensity of each pixel represents the probability of that pixel belonging to the salient object (Borji, Sihite, and Itti 2012).

Many computational models have been proposed to calculate the saliency map of a given image. According to whether the prior knowledge is required or not, two classes of models are usually distinguished: bottom-up and top-down. Typical bottom-up models extract low-level features such as color, intensity, and orientation to construct a conspicuity map in each independent feature space. These conspicuity maps are then combined to form the final saliency map via a predefined fusion strategy (Itti, Koch, and Niebur 1998). Frequency domain analysis (Hou and Zhang 2007)(Achanta et al. 2009) and global contrast-based model (Cheng et al. 2011) are also introduced to compute low-level saliency. The main limitation of these approaches is that the detected salient regions may only contain parts of the target object, or be easily merged with background. On the other hand, the top-down models exploit prior knowledge, such as color (Khan, Weijer, and Vanrell 2009), location (Oliva et al. 2003) and context (Liu et al. 2007), to guide the subsequent saliency detection and estimation. However, the high variety of object types limits its generalization and scalability of these methods.

Recently, an emerging trend is to combine and take advantage of both models into a unified framework. One representative work comes from introducing the low-rank matrix recovery (LR) theory (Candès et al. 2011) into salient object detection. For instance, Shen et al. proposed a Unified method based on LR (ULR) to incorporate traditional low-level features with high-level prior knowledge. Lang et al. proposed a Multi-Task Sparsity Pursuit (MTSP) method to combine multiple types of features for detecting saliency collaboratively (Lang et al. 2012). The existing LR-based saliency detection models share a common assumption that an image can be represented as a highly redundant information part (e.g. background regions) plus a salient part (e.g. foreground object) including several homogeneous regions. The redundant information part usually lies in a low-dimensional feature subspace, which can be approximated as a low-rank feature matrix, while the salient part can be viewed as a sparse sensory matrix (Shen and Wu 2012). As a result, given the feature matrix $\mathbf{F}$ of an input image, it can be decomposed as a low-rank matrix $\mathbf{L}$ plus a sparse matrix $\mathbf{S}$ corresponding to the non-salient background and the salient object, respectively. It is formulated as low-rank matrix recovery problem (Candès et al. 2011):

$$\min_{\mathbf{L},\mathbf{S}} \|\mathbf{L}\|_*+\lambda\|\mathbf{S}\|_1 \qquad \text{s.t.} \qquad \mathbf{F}=\mathbf{L}+\mathbf{S}, \tag{1}$$

where the nuclear norm $\|\cdot\|_*$ (sum of the singular values of a matrix) is a convex relaxation of the matrix rank function (Candès et al. 2011), $\|\cdot\|_1$ indicates ℓ_1-norm which promotes

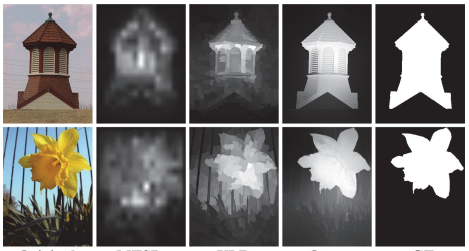

Figure 1: Saliency maps of two typical challenging cases computed by LR-based models and our proposed LSMD model. The outputs of MTSP and ULR are scattered and incomplete, while ours uniformly cover the whole salient objects which are close to ground truth (GT).

sparsity, and the parameter $\lambda > 0$ is a tradeoff between the two items.

From the perspective of statistical signal processing, when using the ℓ_1-norm to promote the sparsity on the matrix $\mathbf{S}$ in Formula (1), it assumes that each element in $\mathbf{S}$ is independent (Zhao, Rocha, and Yu 2009)(Jia, Chan, and Ma 2012) regardless of the potential relationships and structures (such as spatially contiguity and pattern consistency) among them. This assumption inevitably brings two limitations for salient object detection: (i) The generated saliency map tends to be scattered salient pixels or patches instead of spatially contiguous regions. (ii) The existing LR-based methods can not uniformly highlight the whole salient object, which results in incompleteness of the detected object. Some typical examples about these two limitations are shown in Figure 1.

To circumvent these problems, we propose a Low-rank and Structured sparse Matrix Decomposition (LSMD) model that can capture the underlying structure of the salient object. To the end, (i) a tree-structured sparsity-inducing norm, which essentially is a group sparsity with a certain tree structure, is introduced to constrain the matrix in terms of multi-scale spatial connectivity and feature similarity. (ii) The ℓ_∞-norm is embedded into the tree-structured sparsity-inducing norm to replace the plain ℓ_1-norm so as to enforce pixels within the same object have similar saliency values. An effective optimization algorithm for LSMD model is also given out by extending the Augment Lagrange Multipliers (ALM) algorithm. Experimental results on the public benchmark database (Achanta et al. 2009)(Liu et al. 2007) show that our method outperforms the state-of-the-art approaches and can effectively extract the entire salient object.

Low-rank and Structured Sparse Matrix Decomposition

Problem Formulation: For efficiency, we partition the image into non-overlapping patches as the basic image elements in saliency estimation. Assume an input image is partitioned into N patches $\{P_i\}_{i=1}^N$. For each patch P_i, we extract the D-dimension feature and use vector $\mathbf{f}_i \in \mathbb{R}^D$ to represent it. The ensemble of feature vectors forms a matrix representation of the entire input image as $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N] \in \mathbb{R}^{D\times N}$. Then, the task of salient object detection is to design an effective algorithm to decompose the feature matrix $\mathbf{F}$ into a redundant information part $\mathbf{L}$ and a structured salient part $\mathbf{S}$ formulated as:

$$\min_{\mathbf{L},\mathbf{S}} \|\mathbf{L}\|_* + \lambda\Omega(\mathbf{S}) \qquad \text{s.t.} \qquad \mathbf{F} = \mathbf{L} + \mathbf{S}, \tag{2}$$

where $\Omega(\cdot)$ is a structured sparsity-inducing norm regularization to preserve relevant structure and latent relationship of patches in $\mathbf{S}$.

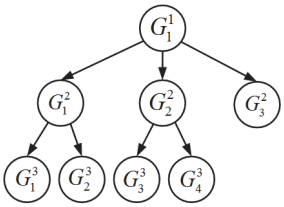

Figure 2: A sample index tree for illustration. Depth1 (Root): $G_1^1 = \{1,2,3,4,5,6,7,8\}$. Depth2: $G_1^2 = \{1,2,3,4\}$, $G_2^2 = \{5,6\}$, $G_3^2 = \{7,8\}$. Depth3: $G_1^3 = \{1,2\}$, $G_2^3 = \{3,4\}$, $G_3^3 = \{5\}$, $G_4^3 = \{6\}$.

LSMD model

Tree structure is widely existed and explored in natural image processing, e.g., tree-structured wavelet transforms, tree-based image segmentation (Felzenszwalb and Huttenlocher 2004), etc. Recent advances in the sparse representation research also exploit tree structure to pursuit the structured sparsity in terms of relationships between patterns (Jenatton et al. 2011). In this work, we consider a tree-structured sparsity-inducing norm, which essentially is a hierarchical group sparsity, to represent the underlying structure of images in feature space.

First, we give out the definition of so-called index tree (Liu and Ye 2010): For an index tree T with depth d, let G_j^i be the j-th node at the i-th level and $T_i = \{G_1^i, \ldots, G_j^i, \ldots, G_{n_i}^i\}$ contain all the nodes corresponding to depth i, where n_i is the number of nodes at the i-th level of T. Specially, for the root node, $n_1 = 1$, and $T_1 = \{G_1^1\} = \{1, 2, \ldots N\}$ (N as the patch number). Furthermore, the nodes in the tree satisfy the following conditions: (i) the nodes from the same depth level have non-overlapping indices, i.e. for any $1 \le j,k \le n_i, j \ne k$, we have $G_j^i \cap G_k^i = \emptyset$. (ii) Let G_{j0}^{i-1} be the parent node of a non-root node G_j^i, then $G_j^i \subseteq G_{j0}^{i-1}$ and $\bigcup_j G_j^i = G_{j0}^{i-1}$. Figure 2 shows a sample index tree with eight indexes ($N = 8$).

Assume we have a meaningful index tree that store underlying structure information of a natural image, and impose it on $\mathbf{S}$ as a structured constraint. Thus, a general tree-structured sparsity regularization can be written as:

$$\Omega(\mathbf{S}) = \sum_{i=1}^{d}\sum_{j=1}^{n_i} w_j^i \|\mathbf{S}_{G_j^i}\|_{p,q}, \tag{3}$$

where $w_j^i \ge 0$ is the weight for the node G_j^i, $\mathbf{S}_{G_j^i} \in \mathbb{R}^{D\times|G_j^i|}$ ($|\cdot|$ denotes the cardinality of a set) is the sub-

Algorithm 1 Solving LSMD via ALM algorithm.

Input: Feature matrix $\mathbf{F}$, parameter λ, ρ and w_j^i (default as 1) for each G_j^i.

Initialize $\mathbf{L}^0$=$\mathbf{0}$, $\mathbf{S}^0$=$\mathbf{0}$, $\mathbf{Y}^0$=$\mathbf{0}$, μ^0=0.5, $\mu_{\max}$= 10^6, and ρ= 6.
While not converged **do**
  $\mathbf{L}^{k+1}$= $\arg\min_{\mathbf{L}} L(\mathbf{L},\mathbf{S}^k,\mathbf{Y}^k,\mu^k)$
  $\mathbf{S}^{k+1}$= $\arg\min_{\mathbf{S}} L(\mathbf{L}^{k+1},\mathbf{S},\mathbf{Y}^k,\mu^k)$
  $\mathbf{Y}^{k+1} = \mathbf{Y}^k + \mu^k(\mathbf{F} - \mathbf{L}^{k+1} - \mathbf{S}^{k+1})$
  μ^{k+1}= $\min(\rho\mu^k,\mu_{\max})$
  $k = k + 1$
End While

Output: $\mathbf{L}$ and $\mathbf{S}$.

matrix of $\mathbf{S}$ corresponding to the node G_j^i, and $\|\cdot\|_{p,q}$ is the mixed $\ell_{p,q}$-norm[1]. Consequently, $\Omega(\cdot)$ is essentially a weighted group sparsity with a certain tree structure, which can fuse similar image patches into identical groups, and meanwhile represent the relationships among groups.

The mixed $\ell_{p,q}$-norm on $\mathbf{S}_{G_j^i}$ in Formula (3) actually includes two components: (i) The ℓ_p-norm on each column of matrix $\mathbf{S}_{G_j^i}$ indicates saliency values calculation of corresponding patches. Inspired by the work of Lang et al., we use ℓ_2-norm ($p = 2$) to measure the saliency of each patch in this paper. (ii) The ℓ_q-norm on the resulting saliency values is to express the relationships among the corresponding patches within the same group. Since similar saliency values are expected to be induced for the patches within the same group, the ℓ_∞-norm ($q = \infty$) is used here. For the ℓ_∞-norm, it is the maximum saliency value of patches within a group that decides if the group is set to saliency or not, and it does encourage all the patches within the group take similar (hence close to the maximum) values (Jia, Chan, and Ma 2012).

After introducing the tree-structured sparsity regularization with the mixed $\ell_{2,\infty}$-norm, the LSMD model can be reformulated as

$$\min_{\mathbf{L},\mathbf{S}} \|\mathbf{L}\|_* + \lambda \sum_{i=1}^{d} \sum_{j=1}^{n_i} w_j^i \|\mathbf{S}_{G_j^i}\|_{2,\infty} \quad s.t. \quad \mathbf{F}=\mathbf{L}+\mathbf{S}. \quad (4)$$

It is noted that by setting the index tree to be a single layer one ($d = 1$, $w_j^i = 1$, and $q = 1$), Formula (4) will be degraded to Formula (1). Therefore, LSMD can be regarded as a generalization of the standard LR model (Candès et al. 2011).

Optimization via ALM

Considering both the nuclear norm and the tree-structured sparsity-inducing norm are convex, we can optimize them by extending the Augment Lagrange Multipliers (ALM)(Lin et al. 2009) algorithm. Correspondingly, Formula (4) is equivalently converted to the following augmented Lagrangian function form:

$$\begin{aligned}\mathcal{L}(\mathbf{L},\mathbf{S},\mathbf{Y},u) = \|\mathbf{L}\|_* + \lambda \sum_{i=1}^{d} \sum_{j=1}^{n_i} w_j^i \|\mathbf{S}_{G_j^i}\|_{2,\infty} \\ + \langle \mathbf{Y} \ , \mathbf{F} - \mathbf{L} - \mathbf{S} \rangle + \tfrac{\mu}{2} \|\mathbf{F} - \mathbf{L} - \mathbf{S}\|_F^2 ,\end{aligned} \quad (5)$$

where $\mathbf{Y}$ is the Lagrange multiplier, and $\mu > 0$ is a penalty parameter. To solve Formula (5), we search for the optimal $\mathbf{L}$, $\mathbf{S}$, and $\mathbf{Y}$ iteratively. The pseudo code of optimization procedure is outlined in Algorithm 1. We now discuss how to update these variables in each iteration.

Updating $\mathbf{L}$: To update $\mathbf{L}^{k+1}$ at the $(k + 1)$-th iteration in Algorithm 1, we fix $\mathbf{L}$ and $\mathbf{S}$, and solve the following problem accordingly:

$$\begin{aligned}&\mathbf{L}^{k+1}= \arg\min_{\mathbf{L}} \mathcal{L}(\mathbf{L},\mathbf{S}^k,\mathbf{Y}^k,\mu^k) \\ &= \arg\min_{\mathbf{L}} \|\mathbf{L}\|_* + \left\langle \mathbf{Y}^k, \ \mathbf{F} - \mathbf{L} - \mathbf{S}^k \right\rangle + \tfrac{\mu^k}{2} \left\|\mathbf{F}-\mathbf{L} -\mathbf{S}^k\right\|_F^2 \\ &= \arg\min_{\mathbf{L}} \tau\|\mathbf{L}\|_* + \tfrac{1}{2} \|\mathbf{L} - \mathbf{M}_\mathbf{L}\|_F^2 ,\end{aligned} \quad (6)$$

where τ=$\frac{1}{\mu^k}$ and $\mathbf{M}_\mathbf{L}$=$\mathbf{F}-\mathbf{S}^k+\frac{1}{\mu^k}\mathbf{Y}^k$. The solution to Formula (6) can be solved as

$$\mathbf{L}^{k+1}=\mathbf{U}T_\tau[\mathbf{\Sigma}]\mathbf{V}^T, where(\mathbf{U},\mathbf{\Sigma},\mathbf{V}^T) =\mathcal{SVD}(\mathbf{M}_\mathbf{L}). \quad (7)$$

Note that $\mathbf{\Sigma}$ is the singular value matrix of $\mathbf{M}_\mathbf{L}$. The operator $T_\tau[\cdot]$ in Formula (7) is a Singular Value Thresholding (SVT) operator (Chen, Wei, and Wang 2012), which is defined by element-wise τ thresholding of $\mathbf{\Sigma}$, i.e., $diag(T_\tau[\mathbf{\Sigma}]) = [t_\tau[\sigma_1], t_\tau[\sigma_2], \ldots, t_\tau[\sigma_r]]$ for $\mathrm{rank}(\mathbf{\Sigma}) = r$, where each $t_\tau[\sigma]$ is determined as

$$t_\tau[\sigma] = \begin{cases} \sigma - \tau, & if \sigma > \tau, \\ \sigma + \tau, & if \sigma < -\tau, \\ 0, & otherwise. \end{cases} \quad (8)$$

Updating $\mathbf{S}$: To update $\mathbf{S}^{k+1}$, we derive Formula (5) with fixed $\mathbf{L}$ and $\mathbf{Y}$, and obtain the following form:

$$\begin{aligned}&\mathbf{S}^{k+1}= \arg\min_{\mathbf{S}} \mathcal{L}(\mathbf{L}^{k+1},\mathbf{S},\mathbf{Y}^k,\mu^k) \\ &= \arg\min_{\mathbf{S}} \lambda \sum_{i=1}^{d} \sum_{j=1}^{n_i} w_j^i \|\mathbf{S}_{G_j^i}\|_{2,\infty} \\ &+ \left\langle \mathbf{Y}^k, \ \mathbf{F} - \mathbf{L}^{k+1} - \mathbf{S} \right\rangle + \tfrac{\mu^k}{2} \left\|\mathbf{F} - \mathbf{L}^{k+1} - \mathbf{S}\right\|_F^2 \\ &= \arg\min_{\mathbf{S}} \varepsilon' \sum_{i=1}^{d} \sum_{j=1}^{n_i} w_j^i \|\mathbf{S}_{G_j^i}\|_{2,\infty} + \tfrac{1}{2} \|\mathbf{S}-\mathbf{M}_\mathbf{S}\|_F^2 ,\end{aligned} \quad (9)$$

where $\varepsilon' = \frac{\lambda}{\mu^k}$ and $\mathbf{M}_\mathbf{S}$=$\mathbf{F} - \mathbf{L}^{k+1}+\frac{1}{\mu^k}\mathbf{Y}^k$. The above tree-structured sparsity optimization problem can be solved by a hierarchical group thresholding operator (Jenatton et al. 2011), which uses the orthogonal projection onto the ℓ_1-ball as a relaxation for the ℓ_∞-norm.

LSMD-based Salient Object Detection

This section elaborates on the salient object detection using the proposed LSMD model. This detection is basically

[1]The mixed $\ell_{p,q}$-norm of a matrix $\mathbf{X} \in \mathbb{R}^{\mathbf{m}\times\mathbf{n}}$ is defined as : $\|\mathbf{X}\|_{p,q}$= $(\sum_{i=1}^{n} \|\mathbf{x}_i\|_p^q)^{1/q}$ = $\left\|(\|\mathbf{x}_1\|_p, \ldots, \|\mathbf{x}_n\|_p)\right\|_q$, where $\mathbf{x}_i$ is the i-th column of the matrix.

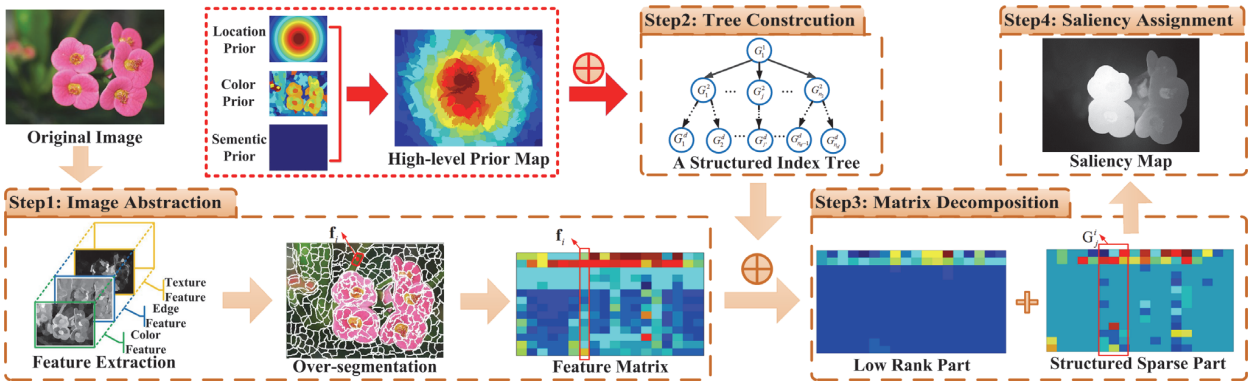

Figure 3: Framework of the LSMD model for salient object detection.

deployed over the low-level features as detailed in the first part of this section, then high-level prior knowledge can be further introduced into the proposed LSMD model, as discussed in the second part. Figure 3 shows the LSMD-based saliency object detection framework.

Low-level Salient Object Detection

Our framework for low-level salient object detection based on the LSMD model consists of four steps as below.

Image Abstraction. In the first step, we aim to partition an input image into compact and perceptually homogeneous elements. Following (Shen and Wu 2012), we first extract the low-level features including RGB color, steerable pyramids (Simoncelli and Freeman 1995), and Gabor filter (Feichtinger and Strohmer 1998) to construct a 53-dimension feature space. Then, we perform the mean-shift clustering (Comanicu and Meer 2002) in the feature space to over-segment the image into N basic patches $\{P_i\}_{i=1}^{N}$. Each patch is represented by $\mathbf{f}_i$, the ensemble of which forms the feature matrix as $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N] \in \mathbb{R}^{D \times N}$ (here $D = 53$).

Tree Construction. The second step is to construct an index tree to represent the image structure via divisive hierarchical k-means clustering. During the tree construction, for each image patch P_i, we get its position coordinate $\mathbf{p}_i$ and feature representation $\mathbf{f}_i$. All patches from an image composite a set of N data points $\mathbf{W} = \{\mathbf{w}_i\}_{i=1}^{N} = \{[\mathbf{p}_i, \mathbf{f}_i]\}_{i=1}^{N}$. Then the divisive hierarchical clustering starts with all the points in a single cluster, and then recursively divides each cluster into k child clusters using k-means algorithm. The recursion terminates when all the clusters contain less than k data points. In this paper, we exploit a quad-tree structure ($k = 4$). Figure 4 shows a visualized example of hierarchical clustering, resulting in a 6-layer index tree structure.

Matrix Decomposition. After obtaining the feature matrix representation and the corresponding structured index tree of the input image, the third step is to use the proposed LSMD model defined in Formula (4) to decompose $\mathbf{F}$ into a low-rank component $\mathbf{L}$ and a structured sparse component $\mathbf{S}$. By introducing the tree-structured sparsity regularization into the LSMD model, we can group perceptually homogeneous patches of the foreground object, while discarding the non-salient background.

Figure 4: Illustration of index tree construction based on the divisive hierarchical k-means clustering. From left to right, each image indicates one layer in the index tree, and each patch represents one node.

Saliency Assignment. The last step is to transform image representation from the feature domain to the spatial domain. To the end, we define a simple assignment function on the structured matrix $\mathbf{S}$ to give a saliency value for each patch P_i:

$$Sal(P_i) = \|\mathbf{s}_i\|_2, \tag{10}$$

where $\mathbf{s}_i$ is the i-th column of matrix $\mathbf{S}$. A larger response of $Sal(P_i)$ means a higher saliency rendered on the corresponding image patch. The resulting saliency map is obtained though merging all patches together. After normalizing and high-dimensional Gaussian filtering (Adams, Baek, and Davis 2010)(Perazzi et al. 2012) on each pixel (x, y), we can get the final pixel-level saliency map $Map(x, y) = Sal(P_i)$ where $(x, y) \in P_i$.

Generalized to Integrate High-level Priors

We further extend the proposed LSMD-based saliency detection to integrate high-level priors. Inspired by the work of Shen et al., we employ the Gaussian distribution to fit the location, semantic and color priors, and fuse them to generate a high-level prior map (see Figure 3). Then the map is incorporated into the proposed LSMD model through setting the weight parameter w_j^i defined in Formula (4).

For each patch P_i, we can use its corresponding average prior value, i.e. $\pi_i \in [0, 1]$, to represent its high-level information. Thus, the prior map is formulated as a vec-

tor $\Pi = [\pi_1, \pi_2, \ldots, \pi_N]$ and then embedded into the tree-structured sparsity-inducing norm as the weight via:

$$w_j^i = 1 - \max(\Pi_{G_j^i}), \tag{11}$$

where $\Pi_{G_j^i} \in \mathbb{R}^{|G_j^i|}$ is the sub-vector of Π corresponding to the node G_j^i. Formula (11) indicates that the node with high prior probability value tends to be salient, and therefore has a small penalty w_j^i. As a result, the high-level prior knowledge is seamlessly integrated with the proposed LSMD model to guide the matrix decomposition to enhance the saliency detection. It is worth noting that if we fix $w_j^i = 1$ for each node G_j^i, the proposed model is degraded to be a pure low-level saliency detection model.

Experiments and Comparisons

We evaluate our LSMD-based saliency computation model on the 1000-image publicly available dataset provided by Achanta et al., which is a subset of MSRA dataset (Liu et al. 2007). This 1000-image dataset is the largest of its kind (Cheng et al. 2011) with accurate manual labels as binary ground truth. We also provide an extensive comparison of our method to the existing LR-based methods and 10 prevailing algorithms, as detailed later. In our experiments, we try to answer the following two key questions on saliency detection:

Q1: Is the tree-structured sparsity regularization beneficial for salient object detection. And how does the proposed LSMD method compare to the existing LR-based methods?

Q2: Compared with other state-of-the-art algorithms, does the generalized LSMD model with high-level prior knowledge have its superiority?

Evaluation Measures and Parameter Selection

Following the standard evaluation protocols in (Achanta et al. 2009)(Cheng et al. 2011), two evaluation measures are exploited. In the first one, we segment saliency maps using every fixed threshold in the range [0, 255]. The segmented binary masks are then compared with the ground truth to compute the precision and recall at each value of the threshold, resulting in the precision versus recall curve. In the second evaluation, we use the image dependent adaptive threshold, defined as twice the mean saliency value of the entire image (Achanta et al. 2009):

$$T_a = \frac{2}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} Map(x, y), \tag{12}$$

where W and H are the width and height of the saliency map in pixels respectively. In this test, in addition to precision and recall, we also compute their weighted harmonic mean measure or F-measure, which is defined as:

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision \times Recall}. \tag{13}$$

The same as previous works (Achanta et al. 2009)(Cheng et al. 2011), β^2 is set to be 0.3.

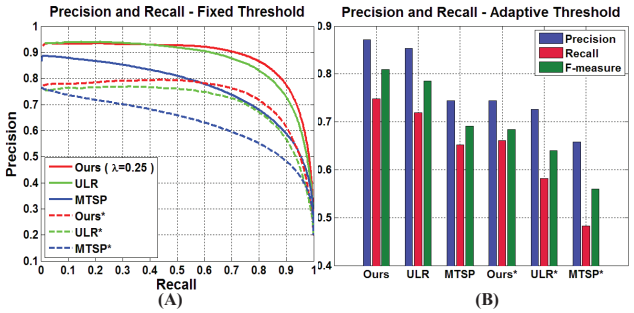

Figure 5: Comparisons with the existing LR-based methods. The superscript "*" in the figure indicates the method without priors.

In the proposed LSMD-base saliency detection models, the tradeoff parameter λ in Formula (4) has a notable influence on the detection performance. To avoid the dependency between datasets for parameter selection and performance evaluation, we use images from MSRA dataset that has no intersection with the 1000-image test dataset to find the optimal parameter λ. We find experimentally that the best choice is λ=0.25.

Comparisons with LR-based methods

To answer the question **Q1**, we compare our method (LSMD) with existing LR-based methods (ULR and MTSP) under two conditions: without priors and with priors.

In the case of pure low-level saliency detection, as shown in Figure 5, our method outperforms other LR-based methods under the two evaluation criteria. It demonstrates that the tree-structured sparsity regularization is much more effective than the plain ℓ_1-norm regularization for salient object detection, because the former one improves the completeness of the extracted salient object based on multi-scale representation of image structure.

If we consider the high-level priors, the performances of these algorithms are all further improved as validated in Figure 5. The proposed LSMD method still achieves the best performance due to the tree-structured regularization. It indicates that both the structured regularization and high-level priors are beneficial for salient object detection. Interestingly, our method without priors (Ours*) even has comparable performance to MTSP with priors as shown in Figure 5(B). This phenomenon further implies that the structure cue is another important and recommendable factor for visual saliency detection.

Comparisons with state-of-the-art methods

In this experiment, we aim to answer the question **Q2** by comparing our method with the other 10 prevailing approaches, including classical works: bio-inspired saliency (IT)(Itti, Koch, and Niebur 1998), fuzzy growing (MZ)(Ma and Zhang 2003), graph-based saliency (GB)(Harel, Koch, and Perona 2006), spatiotemporal cues (LC)(Zhai and Shah. 2006), spectral residual saliency (SR)(Hou and Zhang 2007); and recent leading methods: salient region detection

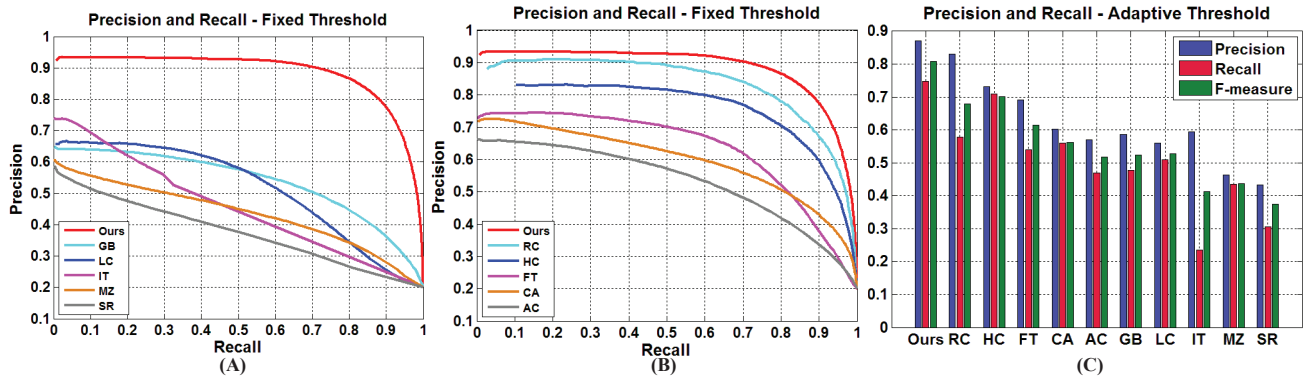

Figure 6: A and B: precision-recall curves for fixed thresholding of saliency maps. C: precision, recall, and F-measure for adaptive thresholding. In all experiments, our method consistently outperforms all the state-of-the-art approaches.

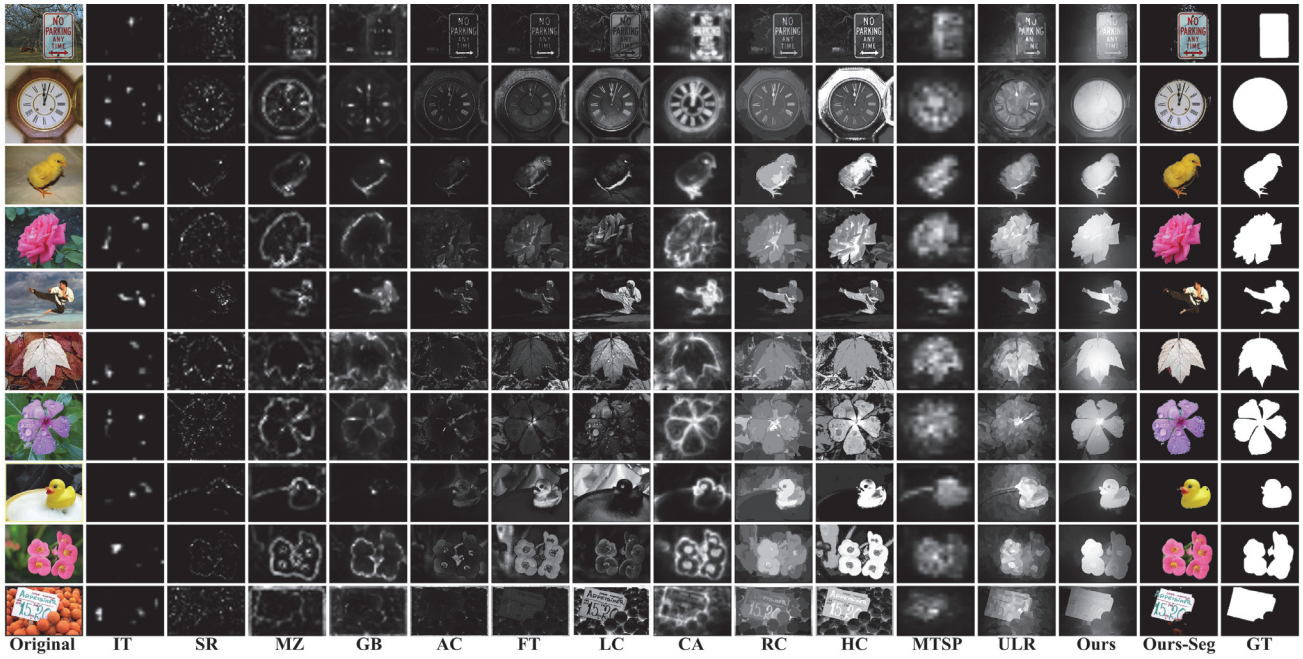

Figure 7: Visual comparison of saliency maps. We compare our method (Ours) to existing LR-based methods and the other 10 prevailing methods. Our segmentation results (Ours-Seg), which are based on saliency maps (Ours) using simple adaptive threshold, are close to ground truth.

(AC)(Achanta et al. 2008), frequency-tuned saliency (FT) (Achanta et al. 2009), context-aware saliency (CA)(Goferman, Manor, and Tal 2010), global-contrast saliency (HC and RC)(Cheng et al. 2011). We use authors' implementation or the resulting saliency maps provided in (Cheng et al. 2011)(Achanta et al. 2009) for evaluation and give out comparison results in Figure 6. The precision versus recall curves in Figure 6(A) and (B) show that the saliency maps generated by our LSMD method with fixed thresholding are much more accurate than those given by the other 10 prevailing algorithms, and even close to the ground truth. Meanwhile, as shown in Figure 6(C), the performance of our method using adaptive threshold is also superior to the other algorithms. For instance, the F-measure of our method is better than that of the best (HC) among the 10 prevailing algorithms by more than 10%.

Figure 7 shows the visual comparison[2] on several challenging images that most existing methods failed. We can clearly see that, compared with other methods, the proposed LSMD-based method can not only completely extract the entire salient object from each image without many scattered patches, but also produce nearly equal saliency values of the pixels within the salient object. This phenomenon further confirms the effect of the structural constraint.

[2]Please access to our project webpage for more comparisons and details. http://sites.google.com/site/saliencydetection

Conclusions

In this paper, we present a generic low-rank and structured sparse matrix decomposition model for visual saliency computation. In the proposed model, a hierarchical tree-structured sparsity-inducing norm is introduced to represent the underlying structure of image patches in feature space. The ℓ_∞-norm is embedded into the tree-structured sparsity to enforce patches within the same object have similar saliency values. Moreover, high-level prior knowledge is seamlessly integrated into our model to enhance saliency detection. Experiments indicate that the proposed model consistently achieves the superior performance on the public benchmark dataset. For future work, we believe that the proposed model can be extended from matrix decomposition to tensor decomposition, which will explicitly find most representative features for salient object detection.

Acknowledgments

This work is partly supported by the National Nature Science Foundation of China (Grant No. 60935002, 61005030, 61272352), the National 863 High-Tech R&D Program of China (Grant No. 2012AA012504, 2012AA012503), the Natural Science Foundation of Beijing (Grant No. 4121003), Guangdong Natural Science Foundation (Grant No. S2012020011081) and Xiamen University 985 Project.

References

Achanta, R.; Estrada, F.; Wils, P.; and Süsstrunk, S. 2008. Salient region detection and segmentation. In *Proc. of ICVS*, 66–75.

Achanta, R.; Hemami, S.; Estrada, F.; and Süsstrunk, S. 2009. Frequency-tuned salient region detection. In *Proc. of CVPR*, 1597–1604.

Adams, A.; Baek, J.; and Davis, M. A. 2010. Fast high-dimensional filtering using the permutohedral lattice. *Comput. Graph. Forum* 29(2):753–762.

Borji, A.; Sihite, D. N.; and Itti, L. 2012. Salient object detection: A benchmark. In *Proc. of ECCV*, 414–429.

Candès, E.; Li, X.; Ma, Y.; and Wright, J. 2011. Robust principal component analysis? *J. ACM* 58(3):1–39.

Chen, C.-F.; Wei, C.-P.; and Wang, Y.-C. 2012. Low-rank matrix recovery with structural incoherence for robust face recognition. In *Proc. of CVPR*, 2618–2625.

Cheng, M.; Zhang, G.; Mitra, N. J.; Huang, X.; and Hu, S. 2011. Global contrast based salient region detection. In *Proc. of CVPR*, 409–416.

Comanicu, D., and Meer, P. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE PAMI* 24(5):603–619.

Feichtinger, H. G., and Strohmer, T. 1998. *Gabor analysis and algorithms:theory and applications.* Springer.

Felzenszwalb, P., and Huttenlocher, D. 2004. Efficient graph-based image segmentation. *International Journal of Computer Vision* 59(2):167–181.

Goferman, S.; Manor, L. Z.; and Tal, A. 2010. Context-aware saliency detection. In *Proc. of CVPR*, 1915–1926.

Harel, J.; Koch, C.; and Perona, P. 2006. Graph-based visual saliency. In *Proc. of NIPS*, 545–552.

Hou, X., and Zhang, L. 2007. Saliency detection: A spectral residual approach. In *Proc. of CVPR*, 1–8.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI* 20(11):1254–1259.

Jenatton, R.; Mairal, J.; Obozinski, G.; and Bach, F. 2011. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research* 12:2297–2334.

Jia, K.; Chan, T.; and Ma, Y. 2012. Robust and practical face recognition via structured sparsity. In *Proc. of ECCV*, 331–344.

Khan, F. S.; Weijer, J. V. D.; and Vanrell, M. 2009. Top-down color attention for object recognition. In *Proc. of ICCV*, 979–986.

Lang, C.; Liu, G.; Yu, J.; and Yan, S. 2012. Saliency detection by multitask sparsity pursuit. *IEEE TIP* 21(3):1327–1338.

Lin, Z.; Chen, M.; Wu, L.; and Ma, Y. 2009. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Technical Report* 09(2215):1–18.

Liu, J., and Ye, J. 2010. Moreau-yosida regularization for grouped tree structure learning. In *Proc. of NIPS*, 1459–1467.

Liu, T.; Sun, J.; Zheng, N.; and Tang, X. 2007. Learning to detect a salient object. In *Proc. of CVPR*, 1–8.

Ma, Y.-F., and Zhang, H.-J. 2003. Contrast-based image attention analysis by using fuzzy growing. In *Proc. of ACM MM*, 374–381.

Oliva, A.; Torralba, A.; Casthelano, M.; and Henderson, J. 2003. Top-down control of visual attention in object detection. In *Proc. of ICIP*, 253–256.

Perazzi, F.; Krähenbühl, P.; Pritch, Y.; and Hornung, A. 2012. Saliency filters: Contrast based filtering for salient region detection. In *Proc. of CVPR*, 733–740.

Rutishauser, U.; Walther, D.; Koch, C.; and Perona, P. 2004. Is bottom-up attention useful for object recognition? In *Proc. of CVPR*, 37–44.

Shen, X., and Wu, Y. 2012. A unified approach to salient object detection via low rank matrix recovery. In *Proc. of CVPR*, 2296–2303.

Simoncelli, E. P., and Freeman, W. T. 1995. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proc. of ICIP*, 444–447.

Zhai, Y., and Shah., M. 2006. Visual attention detection in video sequences using spatiotemporal cues. In *Proc. of ACM MM*, 815–824.

Zhao, P.; Rocha, G.; and Yu, B. 2009. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics* 37(6):3468–3497.