A Unified Approach to Salient Object Detection via Low Rank Matrix Recovery

Xiaohui Shen and Ying Wu
Northwestern University
2145 Sheridan Road, Evanston, IL 60208
{xsh835, yingwu}@eecs.northwestern.edu

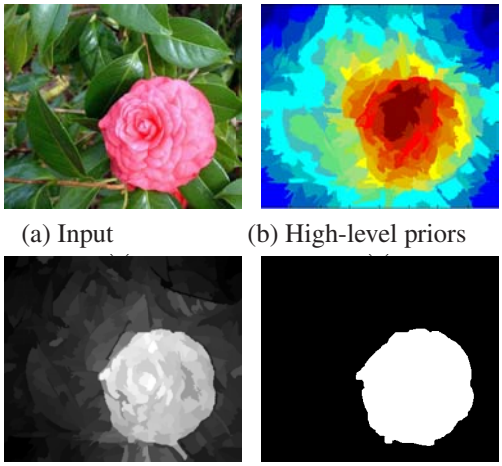
Abstract

Salient object detection is not a pure low-level, bottom-up process. Higher-level knowledge is important even for task-independent image saliency. We propose a unified model to incorporate traditional low-level features with higher-level guidance to detect salient objects. In our model, an image is represented as a low-rank matrix plus sparse noises in a certain feature space, where the non-salient regions (or background) can be explained by the low-rank matrix, and the salient regions are indicated by the sparse noises. To ensure the validity of this model, a linear transform for the feature space is introduced and needs to be learned. Given an image, its low-level saliency is then extracted by identifying those sparse noises when recovering the low-rank matrix. Furthermore, higher-level knowledge is fused to compose a prior map, and is treated as a prior term in the objective function to improve the performance. Extensive experiments show that our model can comfortably achieves comparable performance to the existing methods even without the help from high-level knowledge. The integration of top-down priors further improves the performance and achieves the state-of-the-art. Moreover, the proposed model can be considered as a prototype framework not only for general salient object detection, but also for potential task-dependent saliency applications.

1. Introduction

Image saliency is an important and fundamental research problem in neuroscience and psychology to investigate the mechanism of human visual systems in selecting regions of interest from complex scenes. Recently it has also been an active topic in computer vision, due to its applications to object detection[11, 20] and image editing techniques [19, 8, 13, 4].

Visual saliency can be viewed from different perspectives. Contrast-based and uniqueness-based methods are two typical categories. Local contrast on multiple low-level features can be used to detect low-level saliency [12],



(a) Input (b) High-level priors

(c) Saliency map (d) Detection

Figure 1. Illustration of our approach. By integrating the low-level visual features from the image in (a) and high-level priors from human perception in (b), we get the saliency map of the image as in (c). The salient object is then segmented based on the saliency map, which is shown in (d).

which has motivated various models and methods that combine local, regional and global contrast-based features [17, 23, 27, 18, 8, 4, 16]. In addition, uniqueness is another point of view for saliency, because salient regions can be regarded as those that cannot be well "explained" by its surroundings [2], i.e., being unique. To measure the uniqueness, different models such as self-information[3], graphic models [9], log-spectrum [10] and sparsity models [26] are studied. Uniqueness is in essence similar to high contrast, as the regions different from their surroundings usually have high responses on contrast-based features.

These methods may work well for low-level saliency (or saliency regions), but they are neither sufficient nor necessary, especially in the cases when the saliency is also related to the human perception or is task-dependent. While the salient regions are mostly unique, the inverse might not necessarily be true [14]. Not all unique regions are salient, and a small region with high local contrast might be con-

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

sidered as meaningless noise by the human. Thus, to differentiate real salient regions from other unique/high-contrast parts, priors from higher-level knowledge need to be integrated. Such priors can be based on the human perception, e.g., objects in red color are more pronounced (perhaps because 56% of the cones in our eyes are red-sensitive). They can also be objects with certain semantic meanings such as faces, which is either based on our daily experiences or is task-dependent. *Ad hoc* methods have been proposed to use some of these priors [13, 8], but the integration acts rather like post-processing at the saliency-map level, i.e., weighted averaging on the saliency maps generated from higher level cues and those from low level features. A more principled integration is more desirable.

In this paper, we propose a unified model to integrate bottom-up, lower-level features and top-down, higher-level priors for salient object detection. We represent an image as a low-rank matrix plus sparse noises in a learned feature space, where the low-rank matrix explains the non-salient regions (or background), and the sparse noises indicate the salient regions. This is because the background usually lies in a low-dimensional subspace, while the salient regions that are different from the rest (i.e., deviating from this subspace) can be considered as noises or errors. Therefore, salient regions can be identified by performing low rank matrix recovery using the Robust PCA technique [25]. Higher level knowledge is then converted to pixel-wise priors and incorporated to this model to achieve better performance. An illustration of our method is shown in Fig.1.

It should be noted that it is not the first time that image saliency is explained as sparse noises over a low rank matrix. In [26], an image is decomposed to 8×8 patches, each of which is sparsely coded by over-complete bases. These sparse code vectors are then collected to form a matrix for low-rank matrix recovery. However, sparse encoding may not be a good feature transformation, as the sparse codes for the image patches have no relation to the sparsity of the noise (i.e., saliency) in the entire image representation. It neither guarantees that the background matrix has a low rank. Furthermore, when the images are divided into small patches, a large salient object will contribute many similar feature vectors. In this case, the noises expected to indicate saliency will no longer be sparse. This violates the underlying assumption of this model.

Different from this approach, we represent the image in another way. We decompose an image into small regions by image segmentation after multi-scale feature extraction. The mean of the feature vectors in a segment is treated as the feature of that segment. Stacking them forms the matrix representation of the image. By this means, the number of segments in a salient object is still small even when the object size is large, as salient objects usually have spatial- and appearance-wise coherence. Meanwhile, a linear feature transformation is further trained from labeled data to ensure that the matrix representing the background has a low rank in our learned feature space. As a result, the assumption of the model is valid in most cases, and our method yields good saliency detection results even without higher-level knowledge.

In addition, we propose to use this model to accommodate higher-level features and priors for unification. Different higher-level information is fused into a prior map, which is then incorporated to the objective function. By utilizing higher-level priors, the salient regions are further highlighted and the saliency detection performance is significantly improved. Moreover, our model provides solutions to potential task-dependent saliency extraction by incorporating different task-dependent and volition-driven priors.

The contributions of our approach are three-fold:

1. It proposes a new representation for images. Through segmentation and feature transformation learning, our model is based on the theory of low rank matrix recovery. This new model provides a new perspective for saliency detection, and achieves the state-of-the-art performance;
2. Our new model naturally integrates higher-level top-down information and lower-level bottom-up saliency in a unified way, which has rarely been done before;
3. The proposed approach can be further used in task-dependent saliency detection and subsequent applications as a prototype model.

2. Related Work

We briefly introduce the related work on image saliency detection in this section. Since saliency is explained as those parts standing out from the rest of the image, lots of efforts have been devoted to measure the differences of a region from others. Various contrast-based features have been proposed. In [12], center-surrounding operators are performed on feature maps to obtain the local maxima of activities. Gradient slope information and isophote symmetry can be used to detect saliency[23]. Regional contrast features such as center-surround histograms [17] and center-surround divergence of feature statistics [16] have also been introduced. Recent methods measure the contrast in a global scope with different types of features such as color or luminance differences [27, 18, 1, 4].

Other approaches focus on the mathematical models that can be utilized to model saliency. Entropy is used to select the best scale for saliency [14]. Site entropy rate is adopted [24], and the Shannon's self information is used to measure the local contrast [3]. In [9], an image is represent as an attribute graph, where its nodes represent the spatial locations and the weights on the edges are proportional to the dissimilarity between two locations. An SVM is trained

from the eye movement data to classify salient regions [15]. Saliency can also be represented as the residual of the image log-spectrum compared with an average log-spectrum from a set of natural images [10]. Sparse coding and low rank matrix recovery is employed in [26].

There are also some approaches attempting to integrate different features or priors to detect saliency. However, most of them simply combine the saliency responses from these features through weighted averaging. The weights are either pre-determined [8], or learned from examples by SVM [13] or Conditional Random Fields [17]. Top-down priors are used as the weights for saliency in object detection [20]. Mutual information between the features and the target is used to incorporate top-down information and bottom-up saliency[28]. Gao *at al.* propose[7] to use top-down discriminant principles for bottom-up saliency selection. Different from these approaches, we propose a unified model to integrate low-level and high-level knowledge.

3. Low-level Saliency Detection

In this section, we introduce our method of low-level saliency detection based on low rank matrix recovery.

3.1. Multi-scale visual feature extraction

Given an image, we extract different types of visual features around each pixel, including:

Color. Three RGB color values as well as the hue and the saturation components are extracted for each pixel, producing 5 color features. Each feature is normalized by subtracting its median value over the entire image;

Steerable pyramids[22]. Steerable pyramid filters with four directions on three different scales are performed on the image, yielding 12 filter responses at each location;

Gabor filters[6]. Gabor filter responses with 12 orientations and 3 scales are extracted. The bandwidth of the smallest filter is chosen to be 8, and the scaling factor is 2.

每一个像素点53个特征，颜色5个，金字塔12个，Gabor 36个

All those 53 features are then stacked vertically to form a feature vector, which captures color, edge and texture that are most common low-level visual features.

3.2. Saliency detection by low rank matrix recovery

After feature extraction, we perform image segmentation based on the extracted features by mean-shift clustering[5]. We select spatial and feature bandwidths to over-segment the image so that the background also contains multiple segments even if it is visually homogenous. See Fig.2(b) for an example. The image accordingly is decomposed into segments $\{B_i\}_{i=1,...,N}$, where N is the number of segments. For each B_i, we use the feature vector of its cluster center, i.e., the mean of the features in this segment, as its feature representation f_i. By stacking f_i into a matrix, we get the feature matrix representation of the entire image

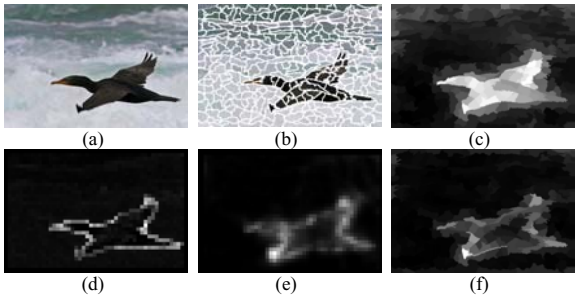


Figure 2. Illustration on low-level saliency detection by our model. (a) the original image, (b) over-segmentation result, (c) detected saliency with learned feature transformation, which is better than others, (d) detected saliency by SC[26], which only has high values on edges, (e) saliency by uniform sampling, which also cannot detect the entire object, (f) saliency by over-segmentation without feature learning, better than (d) and (e), but still has low saliency values inside the object.

$\mathbf{F} = [f_1, f_2, ..., f_N]$. $\mathbf{F} \in \mathbb{R}^{D \times N}$, where D is the dimension of the feature vector ($D = 53$ here).

Similar as in [26], we consider the image as a combination of a background residing in a low dimensional space with salient objects as sparse noises. Therefore, $\mathbf{F}$ can be decomposed into two parts $\mathbf{F} = \mathbf{L} + \mathbf{S}$, where $\mathbf{L}$ is the low-rank matrix corresponding to the background while $\mathbf{S}$ is a sparse matrix representing the salient regions. The low-rank matrix recovery problem can then be formulated as:

$$(\mathbf{L}^*, \mathbf{S}^*) = arg \min_{\mathbf{L},\mathbf{S}} (rank(\mathbf{L}) + \lambda \parallel \mathbf{S} \parallel_0) \quad s.t. \quad \mathbf{F} = \mathbf{L} + \mathbf{S} \tag{1}$$

Since the above problem is NP-hard and hard to approximate, one can alternatively solve the convex surrogate:

$$(\mathbf{L}^*, \mathbf{S}^*) = arg \min_{\mathbf{L},\mathbf{S}} (\parallel \mathbf{L} \parallel_* + \lambda \parallel \mathbf{S} \parallel_1) \quad s.t. \quad \mathbf{F} = \mathbf{L} + \mathbf{S} \tag{2}$$

where $\parallel \mathbf{L} \parallel_*$ is the nuclear norm of $\mathbf{L}$ and $\parallel \cdot \parallel_1$ indicates l_1-norm. Wright *et al.*[25] show that $\mathbf{L}$ and $\mathbf{S}$ can be perfectly recovered by Eqn.2 in most cases by Robust PCA.

After we obtain $\mathbf{S}$, the l_1-norm of each column S_i in $\mathbf{S}$ is used to measure the saliency of corresponding segments. If $\|S_i\|_1$ is larger, we assign a higher value to the image region of the i-th segment B_i. A saliency map is then accordingly generated and normalized to be a gray-scale image.

The rationale of using over-segmentation is that it is a better way of image decomposition than uniform sampling for our model. The uniformly sampled patches have larger feature variations as they are solely determined by the locations, while some sampled patches may contain both the background and the salient object. Therefore the assumption that the background matrix has a low rank may be invalid. Moreover, when the salient object is large, many patches will be sampled from this object. As a result, they

are not considered to be noises as they are no longer sparse. Over-segmentation can alleviate these two problems as it provides feature coherence within the segments and yields fewer feature vectors from the salient object. For example, Fig.2(e) shows the detected saliency on the image from Fig.2(a) using uniform sampling. We can see that the saliency mainly lies on the edges, and the regions inside the object are not detected, as they are not treated as sparse noises. Fig.2(f) gives a better result when using over-segmentation with the same feature representation.

Meanwhile, as mentioned in Section 1, finding a good feature space is also a key problem to ensure the validity of the model. In our original feature space, some salient regions without high texture feature responses may not be considered salient, as in Fig.2(f). Fig.2(d) shows the detected saliency by the method in [26], in which it also uses the model in Eqn.2 but chooses sparse coding as feature representation. We can see that they also tends to produce high saliency along the edges of the object, which indicates sparse coding may not be a good feature representation for saliency detection. Actually, sparse coding only ensures that the coded vector for each image patch is a sparse vector, which is not equivalent to the sparsity of the saliency over the entire image, or the low rank of the background matrix.

To address this problem, we instead learn a linear transformation on the original feature space $\mathbf{T}$ from a set of training images. After transformation, features can be represented as $g_i = \mathbf{T}f_i$ where $\mathbf{T} \in \mathbb{R}^{D\times D}$. Accordingly, $\mathbf{G} = [g_1, g_2, ..., g_N] = \mathbf{TF}$, and the formulation in Eqn. 2 is advanced to:

$$(\mathbf{L}^*, \mathbf{S}^*) = arg\min_{\mathbf{L},\mathbf{S}}(\| \mathbf{L} \|_* + \lambda \| \mathbf{S} \|_1) \quad s.t. \quad \mathbf{TF} = \mathbf{L} + \mathbf{S} \tag{3}$$

It is infeasible to specify $\mathbf{T}$, but it can be learned as shown in the next section.

3.3. Learn a good feature transformation

In a good feature space, most image background should lie in a low dimensional sub-space so that they can be represented as a low rank matrix. To learn this feature space, images with labeled salient objects from the MSRA dataset[17] are used, in which the salient objects have been manually labeled and highlighted by rectangles.

Given an training image, we perform feature extraction and image segmentation as well, and the image is accordingly represented by $\mathbf{F} = [f_1, f_2, ..., f_N]$ as in Section 3.2. Given the labeled rectangle, we use q_i to indicate whether or not f_i belongs to the salient regions ($q_i = 0$ when the corresponding region is salient and 1 otherwise). Such information can be represented in a diagonal matrix $\mathbf{Q} = \mathrm{diag}(q_1, q_2, ...q_N)$. By multiplying $\mathbf{Q}$ to our transformed feature matrix $\mathbf{TF}$, we get $\mathbf{TFQ} \in \mathbb{R}^{D\times N}$. Apparently the difference of $\mathbf{TFQ}$ from $\mathbf{TF}$ is that the feature vectors in $\mathbf{TFQ}$ corresponding to the salient regions become zero-vectors as they are multiplied by $p_i = 0$. Therefore $\mathbf{TFQ}$ only has the information of the background and should have a low rank given a good transformation $\mathbf{T}$. Therefore, the problem of learning $\mathbf{T}$ can be formulated as:

$$\mathbf{T}^* = arg\min_{\mathbf{T}} \mathbf{O}(\mathbf{T}) \triangleq \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}} \| \mathbf{TF}_k\mathbf{Q}_k \|_* - \gamma\|\mathbf{T}\|_* \quad s.t. \quad \|\mathbf{T}\|_2 = c \tag{4}$$

where $\mathbf{F}_k$ and $\mathbf{Q}_k$ are the feature representation and saliency indicator of the k-th training image respectively. K is the number of training images. The regularization term $-\gamma\|\mathbf{T}\|_*$ is to avoid the trivial solution where the rank of $\mathbf{T}$ becomes arbitrarily small so that the rank of $\mathbf{TFQ}$ is largely dominated by $\mathbf{T}$, and γ is a weighting parameter. The constraint $\|\mathbf{T}\|_2 = c$ is to prevent $\mathbf{T}$ from becoming arbitrarily large or small, where c is a constant. We use gradient-descent approach to obtain the local-optimal solution. The partial derivative of $\mathbf{O}(\mathbf{T})$ with respect to $\mathbf{T}$ can be written as

$$\frac{\partial \mathbf{O}(\mathbf{T})}{\partial \mathbf{T}} = \frac{1}{\mathrm{K}}\sum_k \frac{\partial \| \mathbf{TF}_k\mathbf{Q}_k \|_*}{\partial \mathbf{T}} - \gamma\frac{\partial\|\mathbf{T}\|_*}{\partial \mathbf{T}} = \frac{1}{\mathrm{K}}\sum_k \frac{\partial \| \mathbf{TF}_k\mathbf{Q}_k \|_*}{\partial \mathbf{TF}_k\mathbf{Q}_k}(\mathbf{F}_k\mathbf{Q}_k)^T - \gamma\frac{\partial\|\mathbf{T}\|_*}{\partial \mathbf{T}} \tag{5}$$

The sub-differential of a matrix's nuclear norm can be easily obtained by singular value decomposition (SVD). Let $X = U\Sigma V^T$ be a singular value decomposition of X, the sub-differential of the nuclear norm at X is given by [21]:

$$\partial\|X\|_* = UV^T + W \tag{6}$$

where W is a matrix such that $U^TW = 0$, $WV = 0$ and $\|W\| \leq 1$.

After obtaining the gradient, we perform gradient descent $\mathbf{T}_{t+1} = \mathbf{T}_t - \alpha\frac{\partial \mathbf{O}(\mathbf{T})}{\partial \mathbf{T}}$, where α is the step. We then normalize $\|\mathbf{T}_{t+1}\|_2 = c$ at each iteration. The algorithm stops when it converges to local optima. Although Eqn.4 is not convex, we found that $\mathbf{T} = \mathbf{I}$ (i.e., an identity matrix) is a reasonable initialization value. Fig.3 shows the change of the objective function values in Eqn.4 when we choose different steps α. We can see the algorithm converges within 100 iterations with α in a certain range.

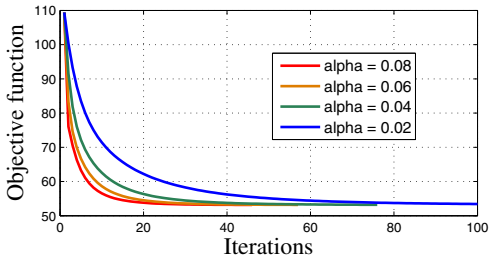


Figure 3. The change of objective function values with different step lengths in feature transformation learning. When α varies from 0.02 to 0.08, the algorithm converges within 100 iterations.

The learned **T** is then used for saliency detection in Eqn.3 for all the images. In the learning process we observe that the image background regions are relatively more coherent in the color spaces than in the gradient or texture feature space. As a result, our learned model is more sensitive to color differences in the saliency detection, which is consistent with human perception. Fig.2(c) gives us the saliency detection result using **T**, which is much better than the result obtained in the original feature space (Fig.2(f)).

4. Higher-level Prior Integration

In this section, we mainly discuss how to incorporate higher-level priors into our model for saliency detection.

4.1. Explored higher-level priors

The higher-levels are generally based on human perception. Currently the following higher-level priors are generated and integrated to our model:

Location prior. Objects near the image center are more attractive to people[13]. Therefore we generate a prior map using a Gaussian distribution based on the distances of the pixels to the image center, in which $p_l(x) = \exp(-d(x,c)/\sigma_1^2)$.

Semantic prior. People pay more attention to certain semantic objects such as faces even without specific purposes. Therefore, similar as in [8, 13], we perform face detection on the images. The regions near the detected faces are assigned higher priors $p_f(x) = \exp(-d(x,f_c)/\sigma_2^2)$, where f_c is the center of the face.

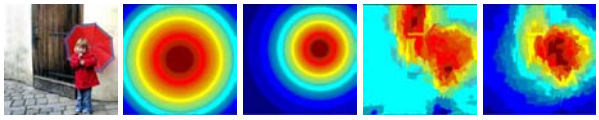
Color prior. From our daily experience, we find that warm colors such as red and yellow are more pronounced. To use such information, we obtain a 2-D histogram distribution $\mathrm{H}(S)$ in nR-nG color space ($nR = \frac{R}{R+G+B}$, $nG = \frac{G}{R+G+B}$) from the labeled salient objects in the MSRA dataset. Similarly a histogram indicating the color distribution of the background $\mathrm{H}(B)$ is generated as well. For each quantized color, we get the values from the two histograms, denoting by h_S and h_B. We then set the color prior for saliency to be $p_c(x) = \exp((h_S(c_x) - h_B(c_x))/\sigma_3^2)$, where c_x indicates the color at location x.

These prior maps are then multiplied together to produce a final prior map. Fig.4 gives an example of those individual prior maps and the fused high-level prior map.

4.2. Integration to the model

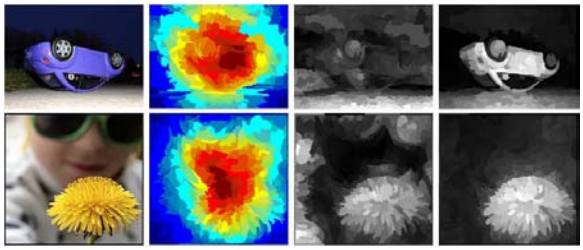
According to the prior map, we know the probability of being salient for each segment based on the location of the segment center, which is denoted by p_i. Such prior probability can also be represented by a diagonal matrix $\mathbf{P} = \mathrm{diag}(p_1, p_2, ..., p_N)$. **P** can be naturally incorporated in our formulation:

$$(\mathbf{L}^*, \mathbf{S}^*) = arg\min_{\mathbf{L},\mathbf{S}} (\| \mathbf{L} \|_* + \lambda \| \mathbf{S} \|_1) \quad s.t. \quad \mathbf{TFP} = \mathbf{L} + \mathbf{S} \tag{7}$$



(a) (b) (c) (d) (e)

Figure 4. Example of high-level prior maps. (a) original image, (b) location prior map, (c) prior maps generated by face detection, (d) color prior map, (e) final fused prior map.



(a) Input (b) Prior map (c) w/o prior(d) w/ prior

Figure 5. The integration of higher-level priors further improves the performance.

By integrating the high-level priors to the formulation, the effects to the saliency detection are two-fold:

1. In **P**, most of p_i are relatively small. Therefore feature vectors multiplied by a larger p_i will be considered as outliers in the low rank matrix **L** and more likely to be included in the sparse noise matrix **S**.

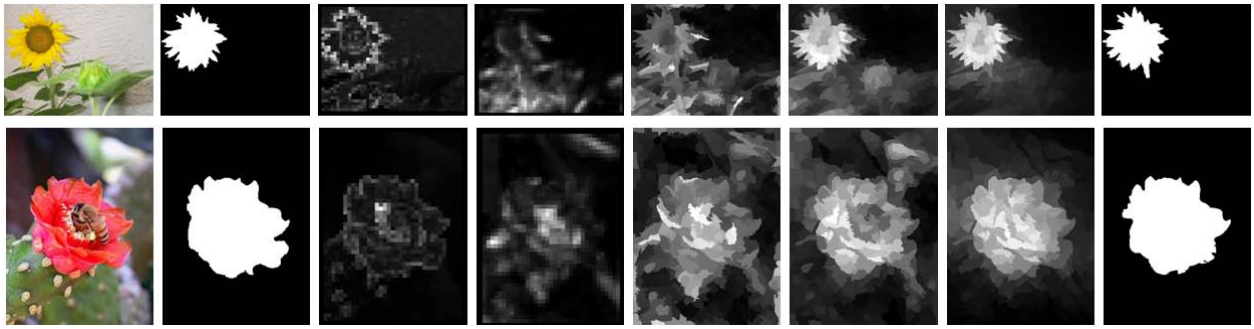
2. The error terms in **S** is also magnified with a larger p_i. Since our saliency map is generated according to the errors in **S** as introduced in Section 3.2, regions with larger p_i tend to produce higher saliency.

Therefore regions with larger priors will be highlighted in the final saliency map by solving Eqn.7. Fig.5(d) shows some results after integrating the high-level prior, which are better than the results without such information in Fig.5(c).

Moreover, by unifying low-level information in F and higher-level priors in P, our model is robust to incorrect top-down higher-level guidance to some extent. Consider the case when some parts of a homogenous background in an image are falsely assigned with large priors due to incorrect guidance from higher-level (see Fig.5 for example, some background regions are also marked with high priors), corresponding feature vectors from these areas are still highly correlated to other background regions. As a result, they are not considered to be noises and will not be labeled as salient regions, as we can observe in Fig.5(d).

5. Experiments

We evaluate our method quantitatively on the 1000-image public dataset[1], which is a subset of the MSRA



(a) Image (b) GT (c) SC (d) US (e) Seg (f) Learn (g) HP (h) Cut

Figure 6. Examples of extracted saliency by our method with different components. GT: ground truth, SC: sparse coding[26], US: uniform sampling, Seg: over-segmentation, Learn: using learned transformation, HP: results after high-level prior integration, Cut: the object is cut out based on the results from (g) using simple segmentation with adaptive threshold. The extracted object is very close to the ground truth.

dataset. Instead of using a rectangle to bound the salient object, accurate human-marked labels are provided as ground truth in this 1000-image dataset. These images are excluded when we learn the feature transformation **T**, and build the color prior histogram. Our model is trained on other images from the MSRA dataset and tested on these 1000 test images.

We follow Achanta *et al.*'s two methodologies[1] to evaluate the accuracy of the detected saliency. In the first evaluation, the image is segmented according to the saliency values with a fixed threshold. Given a threshold $\mathrm{T} \in [0, 255]$, the regions whose saliency values are higher than T are marked as foreground. The segmented image is then compared with the ground truth mask to obtain the precision and recall. When T varies from 0 to 255, different precision-recall pairs are obtained, and a precision-recall curve can be drawn. The average precision-recall curve is generated by combing the results from all the 1000 test images.

In the second evaluation, the test image is segmented by an adaptive threshold method. The image is first over-segmented by mean-shift. An average saliency is then calculated for each segment, and an overall mean saliency value over the entire image is obtained as well. If the saliency in this segment is larger than twice of the overall mean saliency value, the segment is marked as foreground. Precision and recall values are then calculated, and F-measure is also obtained for evaluation where $F = ((\beta^2+1)P*R)/(\beta^2 P+R)$ (P=Precision, R=Recall). We set $\beta^2 = 0.3$ which is the same as in [1, 4].

Comparisons with different components. We first compare the performance of our method with different components. The approach in [26] (denoted by SC) using sparse coding for low rank matrix recovery is also included for comparison. We implemented their algorithm using the same optimization package as in their paper. The precision-recall curves with fixed thresholds are shown in Fig.7, and some examples are presented in Fig.6. We can see from

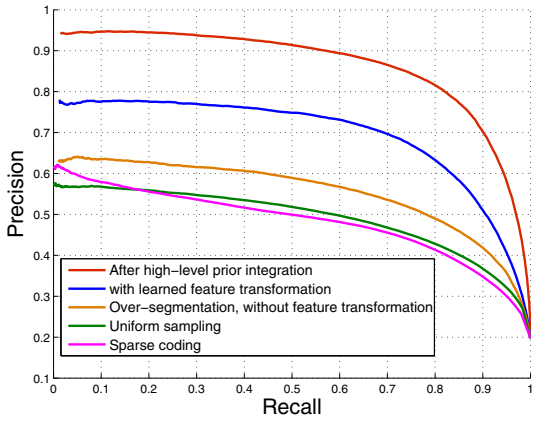


Figure 7. Precision-recall curves of our method with different components on the 1000-image dataset. The performance of sparse coding[26] is also included.

Fig.7 that when we use uniform sampling scheme in the original space, the performance is very close to the results by SC[26]. Over-segmentation noticeably improves the performance, which indicates it is a better image decomposition approach. By detecting the saliency in the learned feature space after feature projection **T**, the performance is significantly improved. This shows the importance of finding a good feature space for saliency detection based on low rank matrix recovery. This performance is already better than most of the existing saliency detection methods, whose results are shown in Fig.8. The accuracy is further improved when incorporating the higher-level priors. From Fig.6, we can observe that using uniform sampling and a sub-optimal feature space tends to produce high response on the edges and texture regions, while our learned model with higher-level priors can cover the entire regions of the object.

Comparisons to the state-of-the-art. We also compare our method with other state-of-the-art approaches, in-

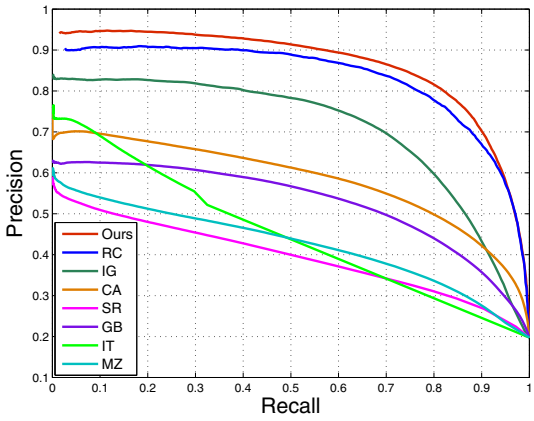


Figure 8. Precision-recall curves on the 1000-image dataset. Compared with IT[12], MZ[18], IG[1], RC[4], GB[9], SR[10] and CA[8], our method achieves the best performance.

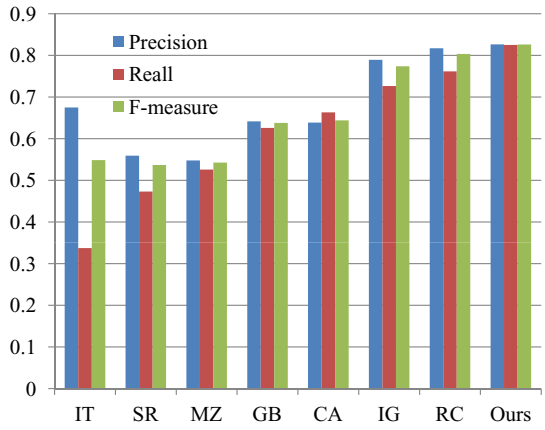


Figure 9. Average precision, recall and F-measure on the 1000-image dataset with adaptive-thresholding segmentation. Our method achieves the best precision, recall and F-measure.

cluding contrast-based approaches(IT[12], MZ[18], IG[1], RC[4]), graph-based(GB[9]), spectrum-based(SR[10]), and the one with high-level priors(CA[8]). Most of them were proposed recently. We use authors' implementation or results for evaluation. Among these baseline methods, the global-contrast based method RC[4] achieves the best performance in the 1000-image dataset. When varying the threshold in segmentation from 0 to 255, the precision-recall curves of these approaches on the 1000 test images along with ours are presented in Fig.8, and the average precision, recall and F-measure using adaptive threshold in segmentation is shown in Fig.9. We can see that the precision-recall curve of our approach is better than RC[4] and achieves the state-of-the-art. When segmenting the salient object using an adaptive threshold, we also achieves the best precision, recall and F-measure. The improvement of recall over other methods is more significant, which means our method are likely to detect more salient regions, while keeping a high accuracy. Some examples of salient object detection and segmentation are shown in Fig.10.

6. Conclusions

In this paper, we propose a novel and unified model based on low rank matrix recovery to integrate the low-level visual features and the higher-level priors for saliency detection. In this model, an image is decomposed into a low rank matrix representing the background, and a sparse noise matrix indicating the salient regions. To ensure the model to be valid for visual saliency, a linear transformation of the feature space is introduced and learned. Higher-level priors can be naturally integrated into this model. Saliency is then jointly determined by low-level and high-level cues in a unified way. Our approach achieves the state-of-the-art on the public salient object benchmark dataset. Furthermore, it can be used as a prototype model in task-dependent saliency applications by integrating different high-level guidance, which merits further study.

Acknowledgements

This work was supported in part by National Science Foundation grant IIS-0347877, IIS-0916607, US Army Research Laboratory and the US Army Research Office under grant ARO W911NF-08-1-0504, and DARPA Award FA 8650-11-1-7149.

References

[1] R. Achanta, S. Hemami, F. Estrada, , and S. Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[2] O. Boiman and M. Irani. Detecting irregularities in images and in video. *IJCV*, 2007.

[3] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, 2006.

[4] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, 2011.

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 2002.

[6] H. G. Feichtinger and T. Strohmer. *Gabor analysis and algorithms: theory and applications*. 1998.

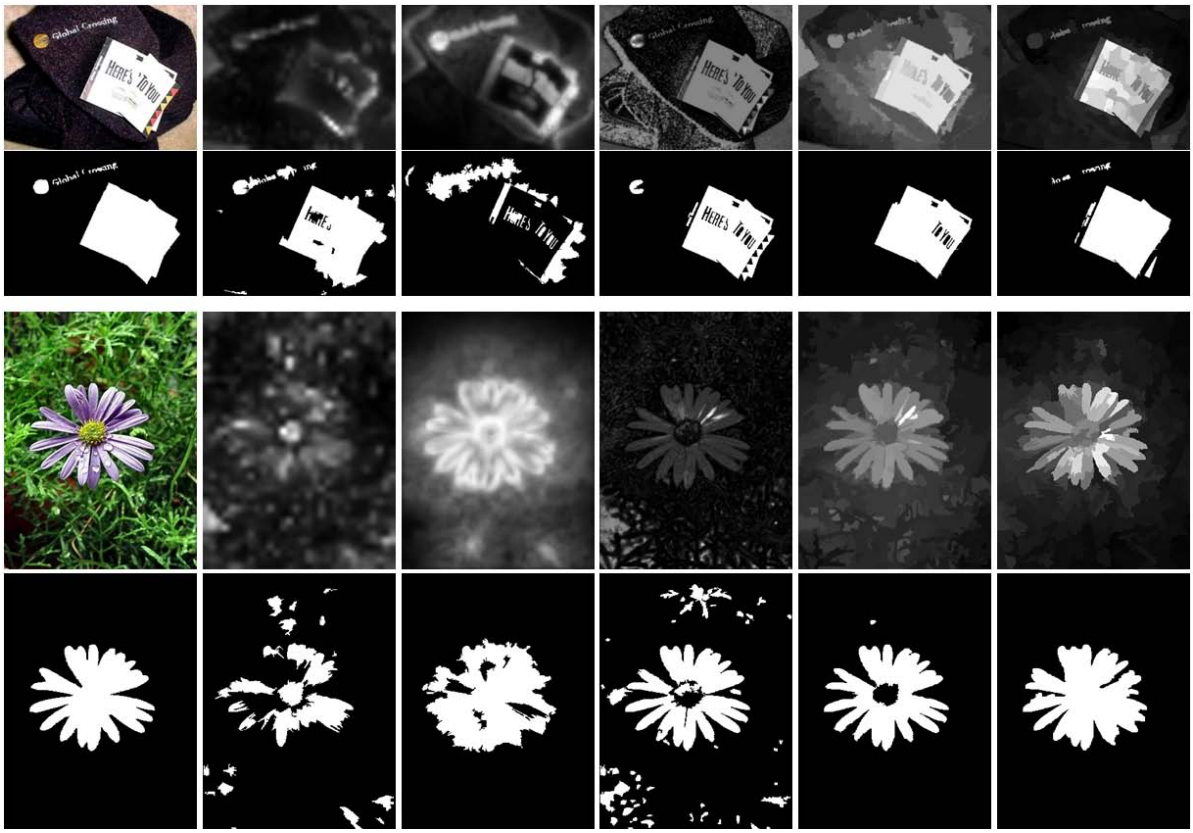
[7] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. In *ICCV*, 2007.

[8] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, 2010.

[9] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, 2007.

[10] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[11] L. Itti, C. Gold, and C. Koch. Visual attention and target detection in cluttered natural scenes. *Optical Engineering*, 2001.



(a) Image/GT (b) GB[9] (c) CA[8] (d) IG[1] (e) RC[4] (f) Ours

Figure 10. Examples of extracted salient objects by different methods. Due to space limit, only the results from four other methods that give good precision-recall curves in Fig.8 are presented. For each example, the first row are the saliency maps, and the second row are the segmented objects. Our approaches gives accurate object segmentation even using a simple thresholding method. The texts in the first image and the center of the flower in the second image are detected by our approach while they are mostly missing in the other methods.

[12] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 1998.

[13] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.

[14] T. Kadir and M. Brady. Saliency, scale and image description. *IJCV*, 2001.

[15] W. Kienzle, F. Wichmann, B. Scholkopf, and M. Franz. A nonparametric approach to bottom-up visual saliency. In *NIPS*, 2007.

[16] D. A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. In *ICCV*, 2011.

[17] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. *PAMI*, 2010.

[18] Y.-F. Ma and H.-J. Zhang. Contrast-based image attention analysis by using fuzzy growing. In *ACM MM*, 2003.

[19] L. Marchesotti, C. Cifarelli, and G. Csurka. A framework for visual saliency detection with applications to image thumbnailing. In *ICCV*, 2009.

[20] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimal object detection. In *CVPR*, 2006.

[21] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum rank solutions to linear matrix equations via nuclear norm minimization. *SIAM Review*, 2010.

[22] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *ICIP*, 1995.

[23] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *ICCV*, 2009.

[24] W. Wang, Y. Wang, Q. Huang, and W. Gao. Measuring visual saliency by site entropy rate. In *CVPR*, 2010.

[25] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*, 2009.

[26] J. Yan, M. Zhu, H. Liu, and Y. Liu. Visual saliency detection via sparsity pursuit. *IEEE Signal Processing Letters*, 2010.

[27] Y. Zhai and M. Shah. Visual attention detection in video sequences using spatiotemporal cues. In *MM*, 2006.

[28] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 2008.