Pattern Recognition Letters 107 (2018) 114–122

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Saliency fusion via sparse and double low rank decomposition

Junxia Li[a,1,*], Jian Yang[b], Chen Gong[b], Qingshan Liu[a]

[a] B-DAT, School of Information and Control, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[b] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China

ARTICLE INFO

Article history:
Available online 12 August 2017

Keywords:
Saliency detection
Saliency fusion
Low rank
Sparse noise

ABSTRACT

Video surveillance-oriented biometrics is a very challenging task and has tremendous significance to the security of public places. Saliency detection can support video surveillance systems by reducing redundant information and highlighting the critical regions, e.g., faces. Existing saliency detection models usually behave differently over an individual image, and meanwhile these methods often complement each other. This paper addresses the problem of fusing various saliency detection methods such that the fusion result outperforms each of the individual methods. A novel sparse and double low rank decomposition model (SDLRD) is proposed for such a purpose. Given an image described by multiple saliency maps, SDLRD uses a unified low rank assumption to characterize the object regions and background regions respectively. Furthermore, SDLRD depicts the noises covered on the whole image by a sparse matrix, based on the observation that the noises generally lie in a sparse subspace. After reducing the influence by noises, the correlations among object and background regions can be enhanced simultaneously. In this way, an image is represented as the combination of a sparse matrix plus two low rank matrices. As such, we cast the saliency fusion as a subspace decomposition problem and aim at inferring the low rank one that indicates the salient target. Experiments on five datasets demonstrate that our fusion method consistently outperforms each individual saliency method and other state-of-the-art saliency fusion approaches. Specifically, the proposed method is demonstrated to be effective on the applications of video-based biometrics such as face detection.

© 2017 Published by Elsevier B.V.

1. Introduction

Video surveillance-oriented biometrics has received intensive attentions in computer vision and machine learning for several decades. The main challenge is to develop and deploy reliable systems to detect, recognize and track moving objects, and further to interpret their activities and behaviors to meet the aim of increasing public security. With the rapid development of surveillance cameras, it is becoming more and more difficult for computers to handle the immense amount of video data. Particularly, the high-quality of video frames introduce a great deal of redundant spatial and temporal information which is time-consuming to handle, and there is no doubt that processing useless information deteriorates system performance.

Saliency detection, the task to detect objects attracted by the human visual system in an image or video, has attracted a lot of focused research in computer vision and has resulted in many applications, such as object detection, tracking and recognition, image/video retrieval, retargeting and compression, photo collage, video surveillance and so on. This paper aims to design an effective saliency fusion model to predict salient objects. Using saliency guides the video surveillance systems to reduce the search space for further processing and thus improve the computational efficiency of the whole system. As shown in Fig. 1, we can use the region covered by the red rectangular bounding box instead of the video frame for further object detection, recognition, tracking, *etc.*

With the goal both to achieve a comparable salience detection performance of human visual system and to facilitate various saliency-based applications, a rich number of saliency detection methods have been proposed in the past decade [2,6,16,18,19,27,28,37–39,44,46,51–53,57,60,64–66]. These approaches design a variety of models to simulate the visual attention mechanism or use data-driven methods to calculate a saliency map from an input image. Since different theories lead to different behaviors of saliency models, the saliency maps obtained by different approaches often vary remarkably from each other. Fig. 2 shows a few results produced by several representative saliency detection methods(*i.e.*, CA [27], HS [60], GC [15]). As

* Corresponding author.
E-mail address: junxiali99@163.com (J. Li).
[1] This work was done when the corresponding author was a Ph.D. student in the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.

http://dx.doi.org/10.1016/j.patrec.2017.08.014
0167-8655/© 2017 Published by Elsevier B.V.

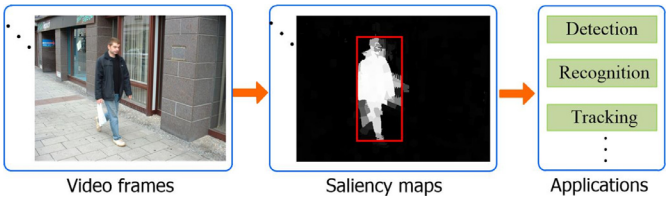

Fig. 1. Illustrating saliency's role of reducing the search space for further processing on the video surveillance systems.

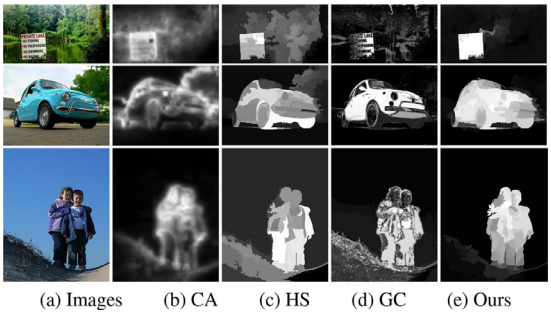

Fig. 2. Saliency fusion results. Individual saliency detection approaches often complement each other. Saliency fusion can effectively combine their results and perform better than each of them.

shown in Fig. 2(b) and (d), the object boundaries are well-defined, but some objects interiors are attenuated. Differently, the results shown in Fig. 2(c) highlight most of the object regions, but some background regions also stand out with the salient regions. Interestingly, these results often can complement each other. This motivates us to combine different saliency maps to achieve better results. Specifically, for a given image, we can first obtain various saliency maps by different saliency detection methods, and then try to find a way to utilize the advantages of these methods, aiming to effectively integrate these saliency maps.

By far, there are few methods attempting to fuse different saliency detection methods. Borji et al. [7] proposed a saliency fusion model using pre-defined combination functions. It treats each individual method equally in the fusion process. This simple strategy may not fully capture the advantages of each saliency detection approach. Mai et al. [48] use a conditional random field (CRF) to model the contribution from each saliency map. Although this method has been shown to be effective, the learnt CRF model parameters are somewhat biased toward the training dataset, due to which it suffers from limited adaptability.

The existing saliency fusion methods are often difficult to produce reliable results for images with diverse properties mainly due to the information contained across multiple saliency maps is not well utilized in the fusion process. To make use of such cross-saliency map information, in our previous work [40,41], we propose two low rank matrix recovery theory based saliency fusion methods, *i.e.*, the robust principle complement analysis (RPCA) model and the double low rank matrix recovery model (DLRMR). However, RPCA assumes that the image object has the sparsity property and hence it does not consider the correlation between object regions. Although DLRMR uses low rank constraint for the object and background regions respectively, it does not consider the noises covered the image in the saliency feature space, and thus leads to the poor robustness.

To address this problem, in this paper we propose a sparse and double low rank decomposition (SDLRD) model for saliency fusion. Fig. 3 is an intuitive illustration on our motivation. Given an image, if we first segment the original image into many homogeneous super-pixels, both object and background contain multiple

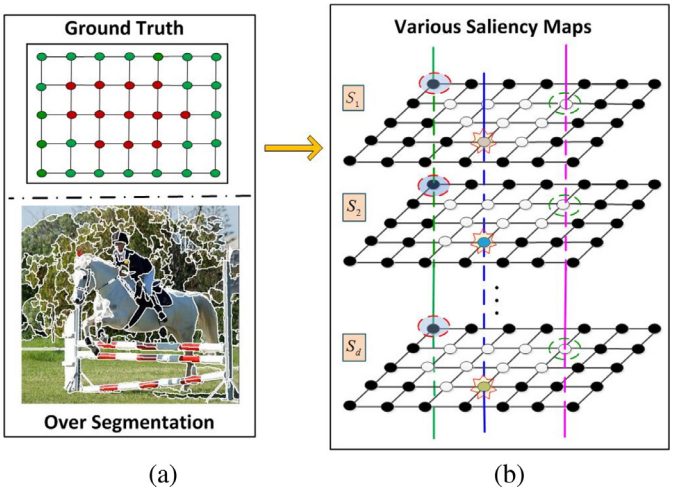

Fig. 3. An example to show the motivation of the proposed SDLRD model. (a) shows the over-segmentation of the original image and its simulated ground truth. In the ground truth image, super-pixels are represented by color nodes: red nodes denote object super-pixels and green ones represent background super-pixels. Clearly, both background and object contain multiple super-pixels. As shown in (b), in all the simulated saliency maps, white nodes denote the super-pixels that are with higher saliency values, while the black ones represent that the corresponding super-pixels are with lower saliency values. The nodes lying on the green (or the pink, blue) line show that they correspond to the same image super-pixel, and we drew a circle over the corresponding node for a visual discrimination. Moreover, there exist some super-pixels that are independent of background and object subspaces and can be considered as noises, *e.g.*, the super-pixel covered by the blue line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

super-pixels. For each super-pixel of object, the corresponding locations of the set of saliency maps are with high probability showing in brighter, indicating that they are with higher saliency values. With image regions being represented by the saliency values of multiple saliency maps, the object super-pixels are highly correlated and the corresponding feature vectors lie in a low-dimensional subspace. Meanwhile, most of background regions tend to have lower saliency values in various saliency maps. They are strongly correlated and lie in a low-dimensional subspace that is independent of the object subspace. Besides, in order to reduce the influence by noises and further to enhance the correlation among the object regions, we assume that the noises covered on the whole image lie in a sparse subspace and can be characterized by using a sparse matrix. Thus, an image can be represented as the combination of a sparse matrix plus two low rank matrices. SDLRD aims at inferring a unified low rank matrix that represents the salient objects. The inference process can be solved efficiently with the alternating direction method of multipliers (ADMM) [8]. Since the correlations within object regions as well as within background regions are well considered, SDLRD can produce more accurate and reliable results than previous saliency fusion models, and also can outperform the performance of each individual saliency detection method.

The contributions of our method mainly include:

1. Our method casts the saliency fusion as a subspace decomposition problem. It provides an interesting perspective for saliency fusion framework.
2. We propose a novel SDLRD model for saliency fusion. Theoretical analysis and experimental results demonstrate the feasibility and effectiveness of the presented method.
3. SDLRD well considers the cross-saliency map information. It performs better than the method which combines saliency maps through pre-defined combination functions.

2. Related work

2.1. Models for saliency detection

In our method, a number of saliency detection approaches are used to produce individual saliency maps. Recently, numerous models have been proposed for detecting salient objects based on variety of mathematical principles and techniques [11–14,21]. As 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 methods have been proposed [5,16,25,26,30,37,47,49,52]. Contrast-based methods have their difficulty in distinguishing among similar saliency cues (*e.g.*, color, pattern, or structure) in both background and foreground regions, and they often fail when the images are with large-scale objects. Besides the widely exploited contrast-based methods, there are lots of formulations for saliency detection based on other principles, such as graph theory [20,29,38,56,60,61], information theory [9,36], and spectral analysis [24,35]. These models may work well for objects within consistent scenes. However, they still lack robustness to detect objects in complex images with cluttered background or objects.

Recently, [50,55,59,66] exploit low rank matrix recovery to formulate saliency detection, in which an image is decomposed into a low rank matrix representing the background and a sparse noise matrix indicating the salient regions. To meet the low rank and sparse properties, [59] uses sparse coding as a representation of image features, and in [55], a learnt transform matrix is used to modulate the image features. Unfortunately, as pointed out in [55], the sparse coding cannot guarantee that the sparse codes of the background are of low rank and those of the salient regions are sparse, especially when the image object is not small. Besides, the learnt transform matrix in [55] is to some extent biased toward the training data set, therefore it suffers from limited adaptability.

Our approach is different from [50,55,59,66] in essence. First, the proposed method is under the saliency fusion scheme and uses the matrix combined by various saliency maps to conduct the matrix recovery. Second, we use the nuclear norm to depict the property of salient regions rather than consider the salient regions as sparse noises. Third, a novel double low rank plus sparse decomposition model is presented to infer the low rank matrix that indicates the salient target.

2.2. Saliency fusion models

Saliency fusion aims at combining various saliency detection methods such that the fusion result outperforms each of the combining ones. Borji et al. [7] use a predefined function (*e.g.*, averaging) to combine individual saliency maps. It treats each individual method equally in the fusion process. This simple strategy may not fully capture the advantages of each individual saliency detection approach. Mai et al. [48] employ a conditional random field (CRF) to model the contribution from individual saliency maps, which show very good results. Unfortunately, training is required and the learnt CRF model parameters are somewhat biased toward the training dataset, therefore it suffers from limited adaptability. Different from [7,48], in our method, we cast the saliency fusion as an object and background decomposition problem in the saliency feature space and propose a novel double low rank matrix recovery model.

3. SDLRD-based saliency fusion

As mentioned above, the existing saliency detection methods are still insufficient to effectively handle all the images, especially for the ones which are with heterogeneous objects, cluttered background, or low contrast between object and background. Fortunately, owing to based on different theories and principles, different saliency detection methods in general can complement each other. Therefore, to make full use of the advantages of existing models, we design a saliency fusion strategy which combines various saliency detection methods such that the fusion result outperforms each of them.

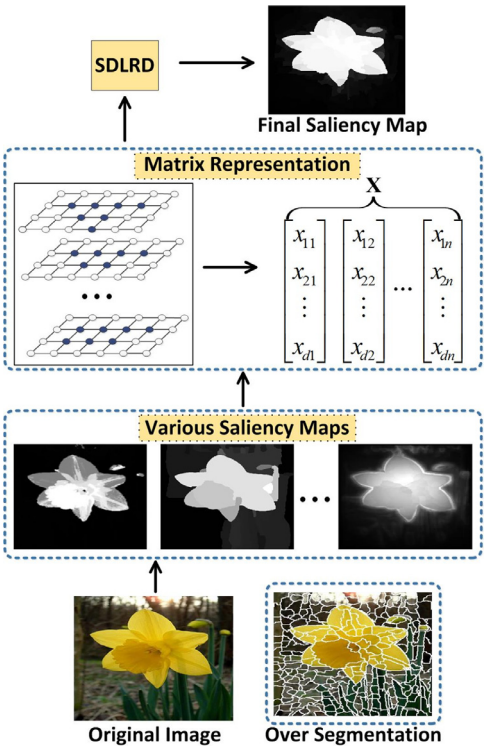

Fig. 4. Framework of the SDLRD model for saliency fusion.

3.1. Problem formulation for saliency fusion

Given an input image $\mathbf{I}$, we first conduct a set of d saliency detection methods and obtain d saliency maps $\{S_k | 1 \le k \le d\}$, one for each approach. Each element $S_k(p)$ in a saliency map denotes the saliency value at pixel p. In each saliency map, the saliency value is represented in gray and normalized to [0, 1]. Our task is to take these d saliency maps as original data and then obtain a final saliency map S.

For efficiency, we segment the input image into super-pixels as the basic image elements in saliency estimation. Let $\mathbf{P} = \{P_i\}_{i=1,\ldots,n}$ be a set of n super-pixels of image $\mathbf{I}$. Combining the obtained d saliency maps, super-pixel P_i can be represented by a vector $\mathbf{X}_i = [x_{1i}, x_{2i}, \ldots, x_{di}]^T \in R^{d\times 1}$, where x_{ki} corresponds to the mean saliency values of P_i in saliency map S_k. By arranging $\mathbf{X}_i$ into a matrix, we get the combinational matrix representation of the whole image $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n]$. $\mathbf{X} \in R^{d\times n}$, where n denotes the number of super-pixels. Then, our goal is to find an assignment function $S(P_i) \in [0, 1]$. Function $S(P_i)$ is referred to as the final saliency map, where the higher value indicates higher salient location. Fig. 4 gives an illustration for the easy understanding of our problem formulation procedure.

3.2. SDLRD model

The task described by the above formulation is to build a criterion for measuring the final saliency. Since each saliency detection

method can be regarded as a nonlinear transformation from the original image to the saliency map, matrix $\mathbf{X}$ can be treated as a feature matrix representation of the image $\mathbf{I}$ in the saliency feature space. In the saliency feature space, we assume that an image is composed by three part: foreground part, background part and the noises. Naturally, feature matrix $\mathbf{X}$ can be decomposed as:

$$\mathbf{X} = \mathbf{F} + \mathbf{B} + \mathbf{L}, \tag{1}$$

where $\mathbf{F}$, $\mathbf{B}$ and $\mathbf{L}$ denote matrices corresponding to foreground, background and noises, respectively.

Problem (1) is actually a subspace decomposition problem. To recover the matrix $\mathbf{F}$ that corresponds to the foreground regions, some criteria are needed for characterizing the matrices $\mathbf{F}$, $\mathbf{B}$ and $\mathbf{L}$. We here consider three basic principles to formulate the inference process. As shown in the over-segmentation map in Fig. 4, the image object and background both contain multiple super-pixels even if they are visually homogeneous. For each super-pixel of image object, the corresponding coordinates of different saliency maps are often with higher saliency values shown in brighter. With these super-pixels being represented by the saliency values of a series of saliency maps, the feature vectors corresponding to the image object have strong correlations and lie in a low-dimensional subspace. Thus, the matrix $\mathbf{F}$ should be encouraged to be low rank. Meanwhile, the background super-pixels generally show similar appearance as they tend to have lower saliency values shown in black. The strong correlations among the background super-pixels suggest that matrix $\mathbf{B}$ may have the property of low rankness. Besides, there exist some noises covering both on the object regions and the background components. In order to deviate them from the whole image and further to enhance the correlations among object regions and background regions simultaneously, we assume that the noises lie in a sparse subspace, *i.e.*, matrix $\mathbf{L}$ is sparse. By considering these three sides, the matrix $\mathbf{F}$ can be inferred by solving the following problem:

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{B},\mathbf{L}} rank(\mathbf{F}) + \lambda(rank(\mathbf{B})) + \gamma\|\mathbf{L}\|_0 \\ &s.t. \quad \mathbf{X} = \mathbf{F} + \mathbf{B} + \mathbf{L}, \end{aligned} \tag{2}$$

where $\|\cdot\|_0$ is the ℓ_0-norm, and parameters λ, $\gamma > 0$ balance the effects between three matrices.

Problem (2) is NP-hard and hard to approximate as the matrix rank and ℓ_0-norm are not convex, no efficient solution is known in both theory and practice [4]. A popular heuristic is to replace the rank with the nuclear norm, and the ℓ_0-norm with the ℓ_1-norm. It has been shown that nuclear norm based models can obtain the optimal low rank solution in a variety of scenarios [23]. Thus we turn to relax minimization problem (2) and obtain a tractable optimization problem, yielding the following convex surrogate:

$$\begin{aligned} &\min_{\mathbf{F},\mathbf{B},\mathbf{L}} \|\mathbf{F}\|_* + \lambda\|\mathbf{B}\|_* + \gamma\|\mathbf{L}\|_1 \\ &s.t. \quad \mathbf{X} = \mathbf{F} + \mathbf{B} + \mathbf{L}, \end{aligned} \tag{3}$$

where $\|\cdot\|_*$ denotes the matrix nuclear norm (sum of the singular values of a matrix), and $\|\cdot\|_1$ is the ℓ_1-norm.

We call the model (3) *sparse and double low rank decomposition* (SDLRD). This minimization problem is convex, and can be efficiently solved via a variety of methods [43]. We will discuss how to solve it in the following subsection.

Fig. 5 gives an example to visually show the ability of SDLRD for subspaces decomposition in the saliency fusion problem. Note that each row of matrix $\mathbf{X}$ corresponds to an individual saliency map, and the rows in different matrices $\mathbf{X}$, $\mathbf{F}$, $\mathbf{B}$ and $\mathbf{L}$ with the same index correspond to the same saliency map. From the second column of Fig. 5, we can clearly see that SDLRD can well extract salient objects from the original saliency maps.

Saliency Assignment. Let $\mathbf{F}^*$ be the optimal solution (with respect to $\mathbf{F}$) to problem (3). To obtain a saliency value for each super-pixel P_i, we define a simple assignment function on the low rank matrix $\mathbf{F}^*$:

$$S(P_i) = \frac{\sum_{j=1}^{d} |\mathbf{F}^*(j,i)|}{d}. \tag{4}$$

A larger response of $S(P_i)$ means a higher saliency rendered on the corresponding super-pixel P_i. The resulting saliency map is obtained though merging all super-pixels together. After normalizing, we can get the final saliency map S. S actually is the 'average map' of all the recovered $\mathbf{F}_k$ shown in the second column of Fig. 5. Algorithm 1 summarizes the whole procedure of our SDLRD based saliency fusion.

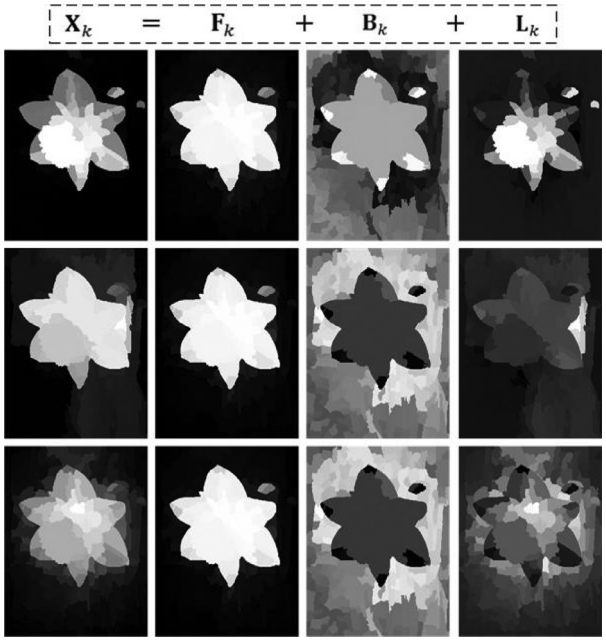

Fig. 5. Illustrating SDLRD's mechanism of decomposing the data. Given matrix $\mathbf{X}$ composed by 11 saliency maps, SDLRD decomposes it into a low rank part $\mathbf{F}$ that represents the object regions, a low rank part $\mathbf{B}$ that links to background regions and a sparse part $\mathbf{L}$ that fits noise. $\mathbf{X}_k$, $\mathbf{F}_k$, $\mathbf{B}_k$ and $\mathbf{L}_k$ correspond to the k^{th} row of $\mathbf{X}$, $\mathbf{F}$, $\mathbf{B}$ and $\mathbf{L}$, respectively. The pixel values in $\mathbf{X}_k$, $\mathbf{F}_k$, $\mathbf{B}_k$ and $\mathbf{L}_k$ are normalized to [0,1]. Here we just give the results corresponding to three original saliency maps.

Algorithm 1 Saliency fusion by SDLRD.

Input: An image $\mathbf{I}$.

1: Conduct individual saliency detection methods and obtain d saliency maps;

2: Conduct image segmentation and compute the matrix representation $\mathbf{X}$ by Section 3.1;

3: Obtain the low rank matrix $\mathbf{F}$ by solving problem (3);

4: Compute the saliency map S by (4);

Output: A map that encodes the saliency value of each super-pixel.

3.3. Optimization via ADMM

Alternating direction method of multipliers (ADMM) is a popular method to solve convex optimization problems, especially in large-scale cases arising in statistics, machine learning and related areas [8]. Problem (3) is convex and can be solved with ADMM. To solve problem (3) by ADMM, let us form the augmented Lagrange

function:

$$L_\mu(\mathbf{F},\mathbf{B},\mathbf{L},\mathbf{Y}) = \|\mathbf{F}\|_* + \lambda\|\mathbf{B}\|_* + \gamma\|\mathbf{L}\|_1 + Tr(\mathbf{Y}^T(\mathbf{X}-\mathbf{F}-\mathbf{B}-\mathbf{L})) + \frac{\mu}{2}\|\mathbf{X}-\mathbf{F}-\mathbf{B}-\mathbf{L}\|_F^2, \quad (5)$$

where $\mathbf{Y}$ is the Lagrange multiplier, $\mu > 0$ is the penalty parameter, and $Tr(\cdot)$ is the trace operator. The standard augmented Lagrange multiple method minimizes L_μ with respect to variables $\mathbf{F}$, $\mathbf{B}$ and $\mathbf{L}$ simultaneously. However, to exploit the property that the variables $\mathbf{F}$, $\mathbf{B}$ and $\mathbf{L}$ in objective function are separable, ADMM decomposes the minimization of L_μ into three sub-problems which minimizes $\mathbf{F}$, $\mathbf{B}$ and $\mathbf{L}$, respectively. The detailed ADMM algorithm for SDLRD is summarized in Algorithm 2. Steps 3 and 4 are solved via the singular value thresholding operator [10], while step 5[2] is solved via a soft-thresholding (shrinkage) operator.

Algorithm 2 Solving SDLRD via ADMM.

Input: Data matrix $\mathbf{X}$, parameters $\lambda > 0$, $\gamma > 0$, ε^{abs}, ε^{rel}.

1: Initializing: $\mathbf{Y}^0 = 0$, $\mathbf{B}^0 = 0$, $\mathbf{L}^0 = 0$, $\mu = 1$, $\mu_{max} = 10^6$, $\rho = 1.2$, $k = 0$.
2: **while** not converged **do**
3: Update $\mathbf{F}$: $\mathbf{F}^{k+1} = D_{\frac{1}{\mu}}(\mathbf{X} - \mathbf{B}^k - \mathbf{L}^k + \frac{1}{\mu}\mathbf{Y}^k)$;
4: Update $\mathbf{B}$:$\mathbf{B}^{k+1} = D_{\frac{\lambda}{\mu}}(\mathbf{X} - \mathbf{F}^{k+1} - \mathbf{L}^k + \frac{1}{\mu}\mathbf{Y}^k)$;
5: Update $\mathbf{L}$:$\mathbf{L}^{k+1} = sgn(\mathbf{X} - \mathbf{F}^{k+1} - \mathbf{B}^{k+1} + \frac{1}{\mu}\mathbf{Y}^k) \circ \max\{|\mathbf{X} - \mathbf{F}^{k+1} - \mathbf{B}^{k+1} + \frac{1}{\mu}\mathbf{Y}^k| - \frac{\gamma}{\mu}, 0\}$;
6: Update $\mathbf{Y}$:$\mathbf{Y}^{k+1} = \mathbf{Y}^k + \mu(\mathbf{X} - \mathbf{F}^{k+1} - \mathbf{B}^{k+1} - \mathbf{L}^{k+1})$;
7: Update μ:$\mu = \min(\rho\mu, \mu_{max})$;
8: If Eq. (6) is not satisfied go to Step 3.
9: **end while**

Output: The optimal solution $\mathbf{F}^*$, $\mathbf{B}^*$ and $\mathbf{L}^*$.

3.3.1. Stopping criterion

Boyd et al. [8] give the optimality conditions and stopping criteria of the ADMM algorithm. Based on the results in [8], we use the following termination criterion: the primal and dual residuals must be small, *i.e.*,

$$\|r^k\|_2 \le \varepsilon^{pri}, \quad \|s^k\|_2 \le \varepsilon^{dual}, \quad \|t^k\|_2 \le \varepsilon^{dual}, \quad (6)$$

where r^k, s^k, t^k, ε^{pri} and ε^{dual} are defined as follows

$$\begin{aligned} r^k &= \mathbf{X} - \mathbf{F}^k - \mathbf{B}^k - \mathbf{L}^k, \\ s^k &= \mu(\mathbf{B}^k - \mathbf{B}^{k-1}), \\ t^k &= \mu(\mathbf{L}^k - \mathbf{L}^{k-1}), \end{aligned} \quad (7)$$

$$\begin{aligned} \varepsilon^{pri} &= \sqrt{dn}\varepsilon^{abs} + \varepsilon^{rel}\max(\|\mathbf{X}\|_F, \|\mathbf{F}^k\|_F, \|\mathbf{B}^k\|_F, \|\mathbf{L}^k\|_F), \\ \varepsilon^{dual} &= \sqrt{dn}\varepsilon^{abs} + \varepsilon^{rel}\|\mu\mathbf{B}^k\|_F + \varepsilon^{rel}\|\mu\mathbf{L}^k\|_F, \end{aligned} \quad (8)$$

where ε^{abs} and ε^{rel} are absolute tolerance and relative tolerance, respectively. The factor $\sqrt{dn}$ accounts for the fact that the ℓ_2 norm is in $R^{d\times n}$.

[2] $sgn(\cdot)$ is the symbolic function, and the absolute value $|\cdot|$ act on each element of the matrix $\mathbf{X} - \mathbf{F}^{k+1} - \mathbf{B}^{k+1} + \frac{1}{\mu}\mathbf{Y}^k$, and $\circ$ is the Hadamard product.

3.3.2. Computational complexity and convergence analysis

The main computational costs of the proposed model are due to the SVD steps. In the steps of updating the matrices $\mathbf{F}$ and $\mathbf{B}$, we need to perform SVD in a matrix of $d \times n$. The computational complexity is $O(nd^2)$, assuming that $n > d$. This is quite efficient because the number of the used saliency detection methods d is usually smaller than n in our experiments, where n denotes the number of super-pixels. Therefore, our algorithm has the computational cost of $O(nd^2)$.

There have been many studies focusing on the convergence of ADMM. Especially, utilizing the properties of the saddle points, Boyd et al. [8] analyzed convergence of ADMM with two variables. He et al. [31,32] presented some significant convergence results by virtue of variational inequalities. What's more, He et al. [33] illuminated that the ADMM owns a convergence rate of $O(1/k)$, where k is the iteration number. Recently, Hong et al. [34] solved the convergence of the ADMM when the number of blocks is more than two. Considering the above results, it is enough that we use (6) as a stopping criterion.

3.4. Connections to the existing saliency fusion models

It is interesting to compare the proposed SDLRD with the other two models, *i.e.*, RPCA [40] and DLRMR [41]. RPCA assumes that in the saliency feature space, an image can be represented as a low rank matrix corresponding to the background, plus a sparse matrix that relates to foreground objects. To well consider the correlation between object regions, DLRMR uses the nuclear norm to constraint the object matrix and enhances the fusion performance to some extent. Actually, SDLRD is an enhanced version of RPCA and DLRMR. SDLRD uses a unified low rank assumption to characterize the object and background regions, respectively. Furthermore, it depicts the noises covered on the image by a sparse matrix.

Setting $\lambda = 0$ in (3), we have

$$\min_{\mathbf{F},\mathbf{L}} \|\mathbf{F}\|_* + \gamma\|\mathbf{L}\|_1, \quad s.t. \quad \mathbf{X} = \mathbf{F} + \mathbf{L}. \quad (9)$$

Clearly, it is the same as the RPCA model for saliency fusion in [40]. Let $\gamma = 0$ in (3), then (3) amounts to

$$\min_{\mathbf{F},\mathbf{B}} \|\mathbf{F}\|_* + \lambda\|\mathbf{B}\|_*, \quad s.t. \quad \mathbf{X} = \mathbf{F} + \mathbf{B}, \quad (10)$$

which is actually the double low rank matrix recovery model (DLRMR) presented in [41].

As a result, SDLRD generalizes (9) and (10) with different parameter settings, *i.e.*, different assumptions for object and background regions. SDLRD can exhibit a better performance than RPCA and DLRMR models. This can be further verified by the experimental results in Section 4.2.

4. Experiments

4.1. Experimental setup

Datasets. Experiments are performed on five publicly-available datasets, including ASD [2], SED1 [3], SED2 [3], SOD [58] and PASCAL-1500 [66]. ASD is a subset of MSRA [45]. It is the most commonly used dataset for saliency detection performance evaluation, and the images in this dataset are relatively simpler than the other four datasets. SED1 and SED2 contain objects of largely different sizes and locations. SOD contains many images with different natural scenes making it challenging for saliency detection. PASCAL-1500 is with 1500 real-world images from PASCAL VOC 2012 segmentation challenging [22]. Many images in this dataset contain multiple objects with various locations and scales, and highly cluttered background.

Evaluation Metrics. We use standard precision-recall curve and F-measure to evaluate the performance of saliency methods.

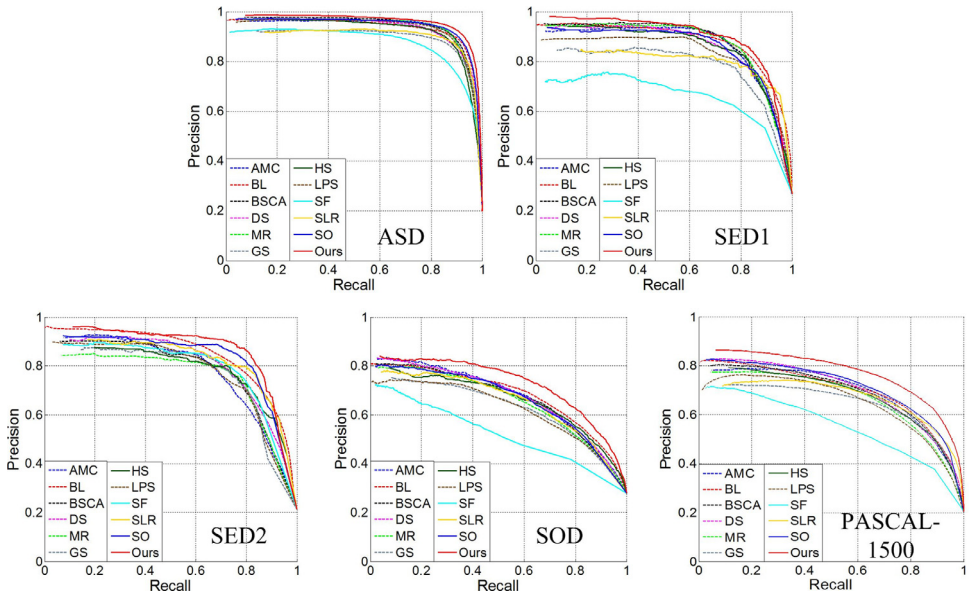

Fig. 6. Precision recall curves of all the twelve methods on the five datasets. Clearly, our method achieves a better PR performance than the other individual methods.

Specifically, the precision-recall curve is obtained by binarizing the saliency map using a number of thresholds ranging from 0 to 255, following [2,16,53]. As described in [2], F-measure is computed as $F-measure = \frac{(1+\beta^2)P\times R}{\beta^2 P+R}$ (P =precision, R=recall), where precision and recall rates are obtained by binarizing the saliency map using an adaptive threshold that is twice the overall mean saliency value. We set $\beta^2 = 0.3$ which is the same as in [2,16,53].

In addition, we measure the quality of the saliency maps using the precision rates at equal error rates (EER) where precision is equal to recall. As complementary to precision and recall rates, we also report the VOC score to evaluate the performance of our proposed method. The VOC Overlap score [54] is defined as $VOC = \frac{|S\cap G|}{|S\cup G|}$, where S is the object segmentation result obtained by binarizing the saliency map using the same adaptive threshold as in the computation of F-measure, and G is the ground-truth.

Parameters. We perform mean shift algorithm [17] to over-segment the original image, where the minimum segment area is set to 200 pixels. Besides, there are two tradeoff parameters λ and γ in our model (3). For fair comparison, we use images from MSRA dataset that has no intersection with the ASD dataset to find the optimal parameters λ and γ, and set $\lambda = 0.7$ and $\gamma = 0.06$ empirically.

4.2. Experimental results

Quantitative Evaluation. Our fusion framework requires a set of saliency detection results from existing saliency detection methods. For each image in the above mentioned five datasets, we produce the saliency maps using eleven saliency detection methods, including: AMC [38], BL [57], BSCA [53], DS [42], MR [61], GS [62], HS [60], LPS [39], SF [52], SLR [66][3] and SO [65]. In order to examine the saliency fusion performance of our proposed method, we compare our fusion result with that of eleven used individual saliency detection methods.

Fig. 6 shows the quantitative results of the presented method against the eleven methods in the aspect of PR curves on the five datasets. It can be seen from Fig. 6 that the proposed method obtains the highest precision rate when the recall rate is fixed. This demonstrates that the fusion method consistently outperforms each individual saliency detection method. Table 1 summarizes the corresponding F-measures, precision rates at EER and VOC overlap scores of all the twelve methods. We can see from Table 1 that our method achieves the highest F-measures, precision rates at EER and VOC overlap scores over the five datasets. This demonstrates that our approach can appropriately consider the performance gaps among individual methods and performs better than them, including the state-of-the-arts.

Comparison with Saliency Fusion Methods. To further illustrate the effectiveness of the proposed method, we first compare our method with other saliency fusion models in literatures [7,48]. Borji et al. [7] uses a pre-defined combination function and takes each individual approach all equal in the fusion process. We denote this method as LA for convenience. Mai et al. [48] adopt a conditional random field framework to saliency aggregation (abbreviated by SA). Fig. 7 shows the evaluation results of our method against LA in the PR curves on the five datasets. The other metrics scores, *i.e.*, F-measure, EER and VOC, for LA are also reported in Table 1. For a fair comparison, like SA, we conduct our saliency fusion model using ten saliency detection methods, *i.e.*, IT [37], MZ [47], LC [63], GBVS [30], SR [35], AC [1], FT [2], HC [16], RC [16], and CA [27] on the ASD dataset. The comparison result is reported in Fig. 8. From Figs. 7, 8 and Table 1, we observe that our method achieves superior saliency fusion performance with respect to previous saliency fusion models for all of the five datasets.

Next, we proceed to compare the proposed SDLRD model with other low-rank theory based saliency fusion methods, *i.e.*, RPCA [40] and DLRMR [41]. Table 2 shows the F-measure, precision rates at EER and VOC overlap scores of all the three models on the five datasets. From this table, we can see that SDLRD consistently outperforms RPCA and DLRMR models. This comparison results verify that SDLRD is more robust to noise and can lead to better saliency fusion result, indicating that adding sparse constraint and thus reducing the influence by noises is a reasonable strategy for saliency fusion.

Subjective Evaluation. Some saliency maps generated by the proposed model, the eleven state-of-the-art saliency models and three saliency fusion methods are shown in Fig. 9 for a subjective comparison. We can observe that most saliency detection methods can handle well the images with relatively simple background and homogenous objects, such as the examples shown in row 1 and

[3] SLR is the extension of LR [55], so here we no longer report the results produced by other low-rank matrix recovery based saliency detection methods.

Table 1
Quantitative performance of our proposed method, fusion method LA and all the eleven individual methods in F-measure, precision rates at EER and VOC overlap scores on the five datasets. The best results are shown in bold.

Dataset	Metric	Ours	LA	AMC	BL	BSCA	DS	MR	GS	HS	LPS	SF	SLR	SO
ASD	F-measure	**0.9161**	0.9016	0.8944	0.8703	0.8745	0.8568	0.8943	0.8260	0.8526	0.8871	0.8157	0.8443	0.8825
	EER	**0.9261**	0.9127	0.8922	0.8974	0.8934	0.8746	0.8918	0.8670	0.8774	0.8864	0.8290	0.8791	0.9029
	VOC	**0.8535**	0.8448	0.8140	0.7987	0.8033	0.7547	0.8100	0.7528	0.7651	0.7941	0.6680	0.7785	0.8206
SED1	F-measure	**0.8415**	0.8257	0.8218	0.7700	0.8056	0.7857	0.8267	0.7259	0.7157	0.7609	0.5737	0.7337	0.7854
	EER	**0.8426**	0.8308	0.8112	0.8278	0.8171	0.8008	0.8225	0.7763	0.8158	0.7959	0.6646	0.7901	0.8002
	VOC	**0.6579**	0.6296	0.6325	0.6046	0.6275	0.5865	0.6408	0.5630	0.5523	0.5778	0.3674	0.5993	0.6145
SED2	F-measure	**0.7870**	0.7730	0.7230	0.7095	0.6983	0.7123	0.7274	0.6830	0.6797	0.7173	0.7203	0.7274	0.7721
	EER	**0.8271**	0.8085	0.7273	0.7830	0.7586	0.7554	0.7585	0.7523	0.7580	0.7333	0.7663	0.7981	0.8062
	VOC	**0.6405**	0.6240	0.5606	0.5739	0.5580	0.5573	0.5690	0.5582	0.5375	0.5420	0.5177	0.5690	0.6306
SOD	F-measure	**0.6278**	0.6107	0.5906	0.5820	0.5855	0.5983	0.5722	0.5685	0.5140	0.5222	0.4258	0.5736	0.6006
	EER	**0.7024**	0.6777	0.6519	0.6628	0.6543	0.6505	0.6337	0.6234	0.6478	0.6161	0.5088	0.6440	0.6424
	VOC	**0.4319**	0.4126	0.3941	0.3984	0.3992	0.4006	0.3753	0.3931	0.3267	0.3256	0.2288	0.4005	0.4091
PASCAL-1500	F-measure	**0.6693**	0.6426	0.6269	0.5983	0.6102	0.6078	0.6107	0.5819	0.5797	0.5870	0.4932	0.5976	0.6347
	EER	**0.7297**	0.7185	0.6833	0.6834	0.6760	0.6734	0.6631	0.6520	0.6684	0.6493	0.5490	0.6738	0.6911
	VOC	**0.5221**	0.5018	0.4683	0.4589	0.4627	0.4466	0.4521	0.4430	0.4164	0.4186	0.3112	0.4567	0.4940

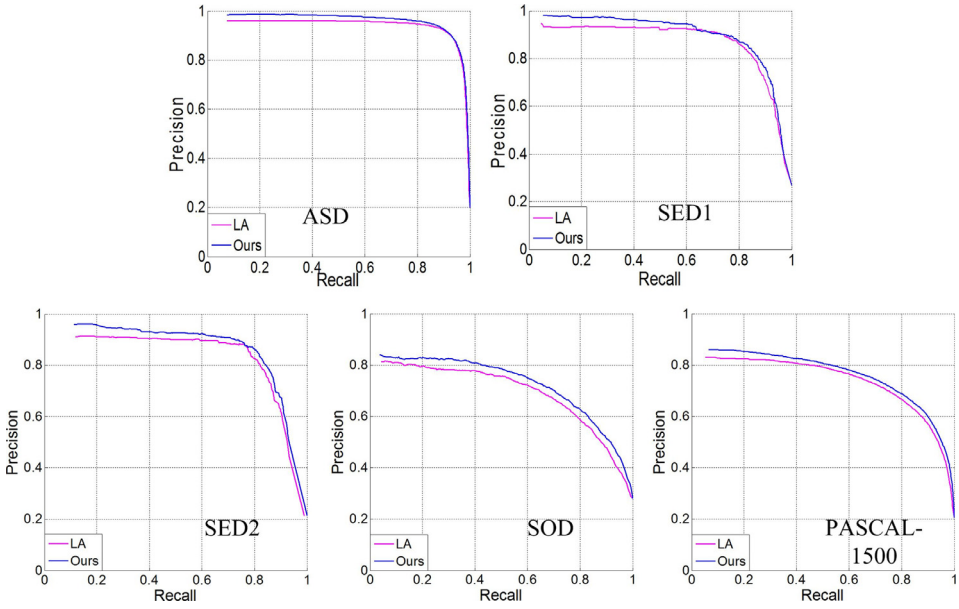

Fig. 7. Precision recall curves of our method with the saliency fusion method proposed in [7] on five datasets.

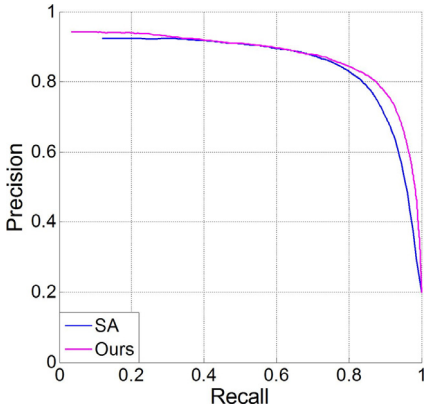

Fig. 8. Comparison with the method proposed in [48] on the ASD dataset.

Table 2
Comparison of F-measure, precision rates at EER and VOC overlap scores between SDLRD, DLRMR [41] and RPCA [40] on the five datasets. The best results are shown in bold.

Dataset	Metric	RPCA	DLRMR	SDLRD
ASD	F-measure	0.9046	0.9089	**0.9161**
	EER	0.9137	0.9205	**0.9261**
	VOC	0.8478	0.8498	**0.8535**
SED1	F-measure	0.8295	0.8394	**0.8415**
	EER	0.8331	0.8401	**0.8426**
	VOC	0.6307	0.6470	**0.6579**
SED2	F-measure	0.7806	0.7839	**0.7870**
	EER	0.8152	0.8209	**0.8271**
	VOC	0.6238	0.6339	**0.6405**
SOD	F-measure	0.6042	0.6193	**0.6278**
	EER	0.6726	0.6862	**0.7024**
	VOC	0.3964	0.4272	**0.4319**
PASCAL-1500	F-measure	0.6431	0.6564	**0.6693**
	EER	0.7051	0.7243	**0.7297**
	VOC	0.4893	0.5134	**0.5221**

2 of Fig. 9, and generate high-quality saliency maps. It is natural that our model can obtain good results for these simple images. However, for some complicated images containing heterogeneous objects (*e.g.*, the person in the fourth row), having a cluttered background (*e.g.*, row 3 in Fig. 9), and showing a low contrast between objects and background (*e.g.*, the bus in the fifth row), most of existing saliency methods cannot effectively highlight the salient objects for all these images. It can be seen that our model in general can suppress background regions and highlight the complete salient object regions with well-defined boundaries more effec-

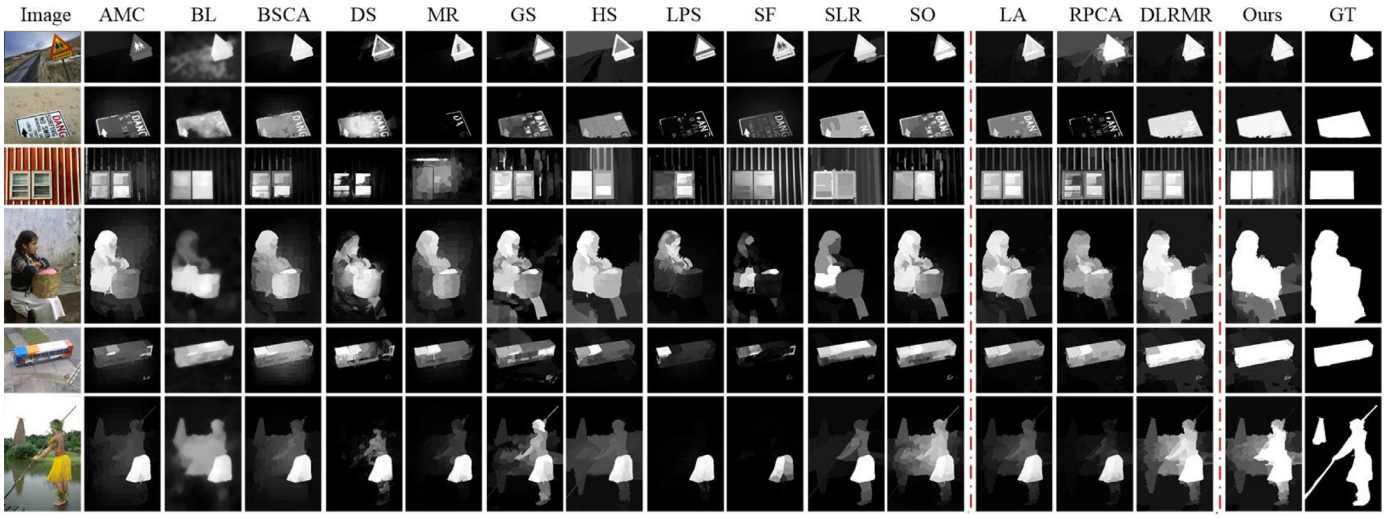

Fig. 9. Examples of saliency detection results. The last row shows a failure case where our method detects more salient regions or powerlessly segments the salient object from the complex background.

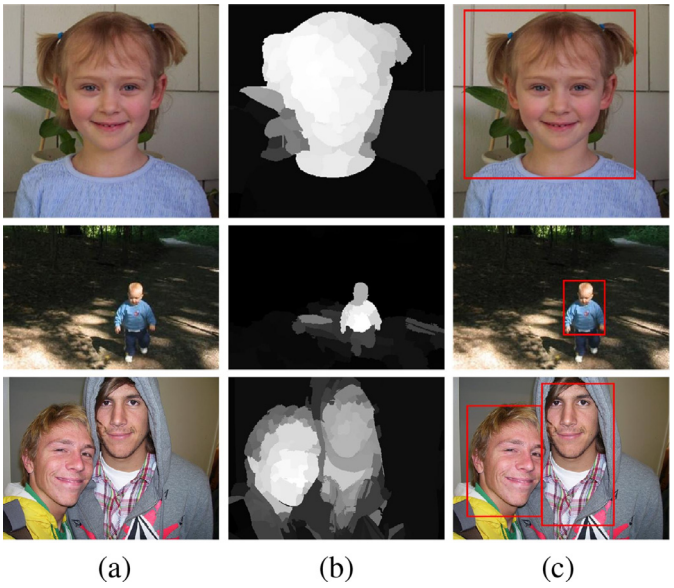

Fig. 10. Face localization using our saliency fusion result. (a) Input image, (b) saliency fusion result of our proposed method, and (c) face localization using our saliency map.

tively than other methods. From Fig. 9, we can clearly see that our approach consistently outperforms every individual saliency detection method. This confirms that our model can effectively integrate the results of these methods.

4.3. Application

Saliency and object localization mainly differ in their granularity of representation. Object localization produces a tight rectangular bounding box around all instances of objects belonging to user-defined categories. In this subsection, we detail the use of our saliency fusion for object localization, especially for face localization.

For an input image, we first binarize the saliency map by applying a threshold at 0.5. The smallest rectangular box enclosing each disconnected region is the localization box for object. Fig. 10 shows some qualitative results obtained using the proposed method in localizing face. From the results, we can see that with this simple strategy, we achieve a satisfactory performance. Using the regions covered by the red bounding box instead of the whole image in the video surveillance systems can reduce the search space for further processing, and there is no doubt that this preprocessing improves system performance.

4.4. Discussions

We have evaluated the effectiveness of our method which can consistently improve the performance of each individual saliency method. However, there is a limit of improvement, since the proposed saliency fusion is based solely on the saliency maps produced by individual methods. When all the used saliency detection methods fail to identify a salient region in an image, our model will usually fail too. The image in the last row of Fig. 9 shows a failure case where the proposed method as well as all the individual methods are unable to detect the salient object in some scenarios.

5. Conclusions and future work

This paper presented a saliency fusion framework to combine saliency maps such that the fusion result outperforms each individual one. Specifically, we cast the saliency fusion as a subspace decomposition problem and proposed a novel sparse and double low rank decomposition model. It provides a robust way to combine individual saliency detection methods into a more powerful one. Experimental results prove that the presented approach performs better than the individual saliency detection methods and outperforms other state-of-the-art saliency fusion approaches.

References

[1] R. Achanta, F. Estrada, P. Wils, S. Susstrunk, Salient region detection and segmentation, in: ICVS, 2008, pp. 66–75.
[2] R. Achanta, S. Hemami, F.J. Estrada, S. Susstrunk, Frequency-tuned salient region detection, CVPR, 2009.
[3] S. Alpert, M. Galun, R. Basri, A. Brandt, Image segmentation by probabilistic bottom-up aggregation and cue integration, CVPR, 2007.
[4] E. Amaldi, V. Kann, On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems, Theor. Comput. Sci. (1998) 237–260.
[5] A. Borji, L. Itti, Exploiting local and global patch rarities for saliency detection, CVPR, 2012.
[6] A. Borji, M. Cheng, H. Jiang, J. Li, Salient object detection: a benchmark, IEEE TIP (2015) 5706–5722.
[7] A. Borji, D.N. Sihite, L. Itti, Salient object detection: a benchmark, ECCV, 2012.
[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. (2011) 1–122.
[9] N. Bruce, J. Tsotsos, Saliency based on information maximization, NIPS, 2005.

[10] J.-F. Cai, E. Cands, Z. Shen, A singular value thresholding algorithm for matrix completion, SIAM J. Optim. (2010) 1956–1982.
[11] X. Chang, F. Nie, S. Wang, Y. Yang, X. Zhou, Compound rank-k projections for bilinear analysis, IEEE TNNLS (2016) 1502–1513.
[12] X. Chang, Y. Yang, Semi-supervised feature analysis by mining correlations among multiple tasks, IEEE TNNLS (2017), doi:10.1109/TNNLS.2016.2582746.
[13] X. Chang, Y. Yang, E. Xing, Y. Yu, Complex event detection using semantic saliency and nearly-isotonic svm, ICML, 2015.
[14] X. Chang, Y. Yu, Y. Yang, E. Xing, Semantic pooling for complex event analysis in untrimmed videos, IEEE TPAMI (2017), doi:10.1109/TPAMI.2016.2608901.
[15] M. Cheng, J. Warrell, W. Lin, S. Zheng, V. Vineet, N. Crook, Efficient salient region detection with soft image abstraction, ICCV, 2013.
[16] M. Cheng, G. Zhang, N.J. Mitra, X. Huang, S. Hu, Global contrast based salient region detection, CVPR, 2011.
[17] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE PAMI (2002) 603–619.
[18] C. Ding, J. Choi, D. Tao, L. Davis, Multi-directional multi-level dual-cross patterns for robust face recognition, IEEE TPAMI (2016) 518–531.
[19] C. Ding, D. Tao, Robust face recognition via multimodal deep face representation, IEEE TMM (2015) 2049–2058.
[20] C. Ding, D. Tao, Pose-invariant face recognition with homography-based normalization, PR (2017) 144–152.
[21] C. Ding, C. Xu, D. Tao, Multi-task pose-invariant face recognition, IEEE TIP (2015) 980–993.
[22] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, in: IJCV, 2010, pp. 303–338.
[23] M. Fazel, Matrix Rank Minimization with Applications, 2002 Ph.D. dissertation.
[24] C. Gao, L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, IEEE TIP (2010) 185–198.
[25] D. Gao, V. Mahadevan, N. Vasconcelos, The discriminant center-surround hypothesis for bottom-up saliency, NIPS, 2007.
[26] D. Gao, N. Vasconcelos, Bottom-up saliency is a discriminant process, ICCV, 2007.
[27] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, CVPR, 2010.
[28] C. Gong, D. Tao, W. Liu, S. Maybank, M. Fang, K. Fu, J. Yang, Saliency propagation from simple to difficult, CVPR, 2015.
[29] V. Gopalakrishnan, Y. Hu, D. Rajan, Random walks on graphs for salient object detection in images, IEEE TIP (2010) 3232–3242.
[30] J. Harel, C. Koch, P. Perona, Graph-based visual saliency, NIPS, 2006.
[31] B. He, M. Xu, X. Yuan, Solving large-scale least squares covariance matrix problems by alternating direction methods, SIAM J. Matrix Anal. Appl. (2011) 136–152.
[32] B. He, H. Yang, Some convergence properties of a method of multipliers for linearly constrained monotone variational inequalities, Oper. Res. Lett. (1998) 151–161.
[33] B. He, X. Yuan, On the o(1/n) convergence rate of the douglas-rachford alternationg direction method, SIAM J. Numer. Anal. (2012) 700–709.
[34] M. Hong, Z. Luo, On the linear convergence of the alternating direction method of multipliers, Math. Program. (2013) 1–35.
[35] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, CVPR, 2007.
[36] X. Hou, L. Zhang, Dynamic visual attention: searching for coding length increments, NIPS, 2008.
[37] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE TPAMI (1998) 1254–1259.
[38] B. Jiang, L. Zhang, H. Lu, M. Yang, Saliency detection via absorbing markov chain, ICCV, 2013.
[39] H. Li, H. Lu, Z. Lin, X. Shen, B. Price, Inner and inter label propagation salient object detection in the wild, IEEE TIP (2015) 1–11.
[40] J. Li, J. Ding, J. Yang, Visual salience learning via low rank matrix recovery, ACCV, 2014.
[41] J. Li, L. Luo, F. Zhang, J. Yang, D. Rajan, Double low rank matrix recovery for saliency fusion, IEEE TIP (2016) 4421–4432.
[42] X. Li, H. Lu, L. Zhang, X. Ruan, M. Yang, Saliency detection via dense and sparse reconstruction, ICCV, 2013.
[43] Z. Lin, M. Chen, L. Wu, Y. Ma, The augmented lagrange multiplier method for extract recovery of corrupted low rank matrices, UIUC Technical Report, 2009. UILU–ENG–09–2215
[44] N. Liu, J. Han, Dhsnet: deep hierarchical saliency network for salient object detection, CVPR, 2016.
[45] T. Liu, J. Sun, N. Zheng, X. Tang, Learning to detect a salient object, CVPR, 2007.
[46] Z. Liu, W. Zou, O.L. Meur, Saliency tree: a novel saliency detection framework, IEEE TIP (2014) 1937–1952.
[47] Y. Ma, H. Zhang, Contrast-based image attention analysis by using fuzzy growing, ACM Multimedia (2003).
[48] L. Mai, Y. Niu, F. Liu, Saliency aggregation: a data-driven approach, CVPR, 2013.
[49] R. Margolin, A. Tal, L. Manor, What makes a patch distinct? CVPR, 2013.
[50] H. Peng, B. Li, R. Ji, W. Hu, W. Xiong, C. Yan, Salient object detection via low-rank and structured sparse matrix decomposition, AAAI, 2013.
[51] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, S. Maybank, Salient object detection via structured matrix decomposition, IEEE TPAMI (2017) 818–832.
[52] F. Perazzi, P. Krahenbuhl, Y. Pritch, A. Hornung, Saliency filters: contrast based filtering for salient object detection, CVPR, 2012.
[53] Y. Qin, H. Lu, Y. Xu, H. Wang, Saliency detection via cellular automata, CVPR, 2015.
[54] A. Rosenfeld, D. Weinshall, Extracting foreground masks towards object recognition, ICCV, 2011.
[55] X. Shen, Y. Wu, A unified approach to salient object detection via low rank matrix recovery, CVPR, 2012.
[56] J. Sun, H. Lu, X. Liu, Saliency region detection based on markov absorption probabilities, IEEE TIP (2010) 1639–1649.
[57] N. Tong, H. Lu, X. Ruan, M. Yang, Salient object detection via bootstrap learning, CVPR, 2015.
[58] V. Movahedi, J. Elder, Design and perceptual validation of performance measures for salient object segmentation, POCV, 2010.
[59] J. Yan, M. Zhu, H. Liu, Y. Liu, Visual saliency detection via sparsity pursuit, in: SPL, 2010, pp. 739–742.
[60] Q. Yan, L. Xu, J. Shi, J. Jia, Hierarchical saliency detection, CVPR, 2013.
[61] C. Yang, L. Zhang, H. Lu, X. Ruan, M. Yang, Saliency detection via graph-based manifold ranking, CVPR, 2013.
[62] Y. Wei, F. Wen, W. Zhu, J. Sun, Geodesic saliency using background priors, ECCV, 2012.
[63] Y. Zhai, M. Shah, Visual attention detection in video sequences using spatiotemporal cues, ACM Multimedia (2006).
[64] X. Zhou, Z. Liu, G. Sun, X. Wang, Adaptive saliency fusion based on quality assessment, Multimed. Tools Appl. (2016), doi:10.1007/s11042-016-4093-8.
[65] W. Zhu, S. Liang, Y. Wei, J. Sun, Saliency optimization from robust background detection, CVPR, 2014.
[66] W. Zou, K. Kpalma, Z. Liu, J. Ronsin, Segmentation driven low-rank matrix recovery for saliency detection, BMVC, 2013.