This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

Unsupervised Salient Object Detection via Inferring from Imperfect Saliency Models

Rong Quan, Junwei Han, *Senior Member, IEEE*, Dingwen Zhang, Feiping Nie, Xueming Qian, and Xuelong Li, *Fellow, IEEE*

Abstract—Visual saliency detection has become an active research direction in recent years. A large number of saliency models which can automatically locate objects of interest in images have been developed. As these models take advantage of different kinds of prior assumptions, image features, and computational methodologies, they have their own strengths and weaknesses and may cope with only one or a few types of images well. Inspired by these facts, this paper proposes a novel salient object detection approach with the idea of inferring a superior model from a variety of previous imperfect saliency models via optimally leveraging the complementary information among them. The proposed approach mainly consists of three steps. First, a number of existing unsupervised saliency models are adopted to provide weak/imperfect saliency predictions for each region in the image. Then, a fusion strategy is used to fuse each image region's weak saliency predictions into a strong one by simultaneously considering the performance differences among various weak predictions and various characteristics of different image regions. Finally, a local spatial consistency constraint which ensures high similarity of the saliency labels for neighboring image regions with similar features is proposed to refine the results. Comprehensive experiments on five public benchmark datasets and comparisons with a number of state-of-the-art approaches can demonstrate the effectiveness of the proposed work.

Index Terms—Salient object detection, weak prediction, fusion strategy, local spatial consistency constraint.

I. Introduction

SALIENT object detection aims to automatically obtain objects of user interest from images by using bottom-up visual features. In recent years, as an cornerstone technique, it has been widely used in a variety of computer vision applications such as object recognition [1, 2], image/object segmentation [3, 4] , image compression [5], image cropping [6], content-based image retrieval [7], and so on.

After two decades of extensive study, a large amount of saliency detection models have been developed. By and large, existing saliency models can be categorized into three classes: local contrast based models, global contrast based models, and background prior based models. The local contrast based models [8-12] compute the center-surround difference at each location of an image and highlight the regions distinct within a local context. Normally, this class of methods can highlight the small salient objects precisely. However, for the salient objects with big size, these methods can only detect the objects' boundaries instead of the entire interior regions of them (See Fig. 1(c)). In contrast, the global contrast based models [13-20] calculate the saliency of an image location as the uniqueness in the entire image. These models can alleviate the problem of only detecting the boundaries of the salient objects to some extent. However they may fail to uniformly highlight the whole salient objects when the foreground regions are complex and with diverse appearance, or (See Fig. 1(d)). Besides, both the local and global contrast based models are likely to falsely consider the small-scale high-contrast background patterns as salient. The third class of models [21-25] rely on the background prior to assume that the image boundary regions are more likely to be the image background and then separate the salient foreground regions by calculating their contrast with the image boundary regions. Many previous works have demonstrated that this class of models can uniformly detect the salient objects and suppress the background in most cases. However, they still cannot achieve satisfactory performance in challenging scenarios especially when the images contain complex background and foreground regions or the salient objects significantly touch the image boundaries (See Fig. 1(e)).

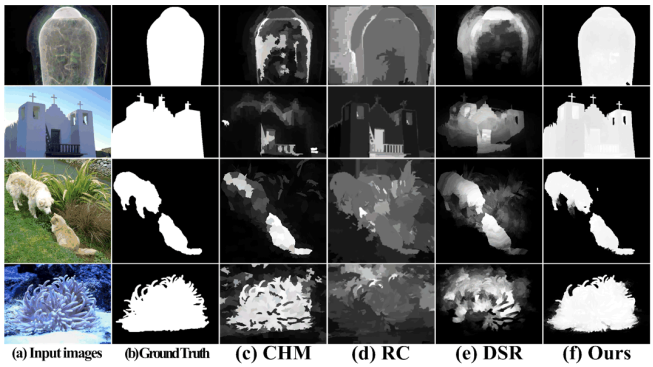


Fig. 1. Examples (including the local contrast based model, CHM [12], the global contrast based model, RC [15], and the background prior based model, DSR [22]) to show the differences among various saliency models and the advantages of our salient object detection method.

This work was supported in part by the National Science Foundation of China under Grant 61473231. (*Junwei Han is the corresponding author*).

R. Quan, J. Han, and D. Zhang are with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: rongquan0806@gmail.com,junweihan2010@gmail.com, zhangdingwen2006yyy@gmail.com).

F. Nie is with School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: feipingnie@gmail.com).

X. Qian is with School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: qianxm@mail.xjtu.edu.cn).

X. Li is with Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xuelong_li@ieee.org)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

As can be seen, existing saliency models are based on different prior assumptions, feature representations, and computational methodologies. Each of these methods may have their own strengths and weaknesses and normally only handle one or a few types of images well. It is extremely difficult, even impossible, for a single salient object detection model to work well under all various scenarios. This naturally motivates us to yield a superior saliency model by fusing the strength of each imperfect model, which on one hand is able to push forward salient object detection to conduct more robust prediction, and on the other hand could be a way to make the best use of the existing and forthcoming saliency models.

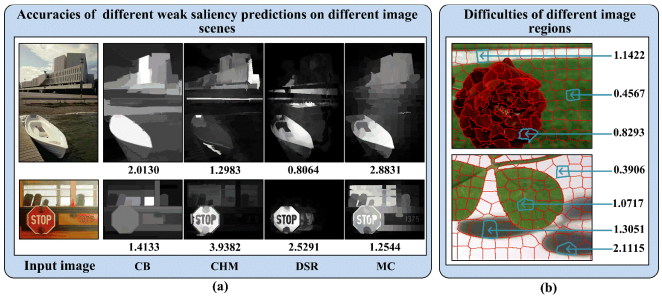


Fig. 2. Illustration of the image-dependent and region-dependent properties of the fusion process. (a) The value under each weak saliency prediction represents the accuracy of this prediction to certain image, which not only depends on the corresponding weak saliency models' saliency detection capacities (including CB [11], CHM [12], DSR [22], MC [21]), but also depends on the image scene itself. (b) The value corresponding to each example image region represents the difficulty of this image region to be labeled correctly. Even in the same image, different image regions have different difficulties to be labeled correctly.

To this end, in this paper, we propose to cast the salient object detection as a problem of fusing weak predictions from multiple existing imperfect saliency models. Given an image, we first apply a number of existing saliency models to yield the corresponding saliency maps, each of which provides a weak prediction of saliency at every location of the image. Then, we propose to fuse the weak predictions from various saliency models into a strong/superior saliency prediction. In the proposed work, we presume that, the fusion process should have two important properties: 1) Image-dependent: the accuracy of each weak saliency prediction should be considered with respect to the specific image scene because the prior assumptions or computational methodologies used in different saliency models may fit to the specific image scene in varying degrees. (See Fig. 2(a)); 2) Region-dependent: the difficulty of image regions should be considered with respect to the specific image regions because the agreement of the multiple weak saliency predictions is different in various image regions, which indicates that different image regions may have different difficulties to be labeled correctly (See Fig. 2(b)). Thus, we propose to model this fusion process in a GLAD (Generative model of Labels, Abilities, and Difficulties) [26] framework, where the accuracy of each weak saliency prediction and the difficulty of each image region can be inferred from the input image regions and their corresponding weak saliency labels automatically. Then, the region-level professionality of each weak saliency prediction, i.e. the confidence degree of each weak saliency prediction to each image region, can be further obtained to infer the strong saliency label for each image region in an unsupervised way. Finally, as the fusion strategy does not take into consideration the interaction between neighboring image regions, and it is obvious that the saliency of neighboring regions should have the mutual influence, we further refine the fusion results by utilizing a novel local spatial consistency constraint to enable the neighboring image regions with similar image features to have similar saliency labels.

Actually, there are a few previous works close to our considerations in this paper. Specifically, Boji et al. [2] made an earlier attempt to directly apply several simple pre-defined combination functions, e.g., multiplication, averaging and et al, to fuse the existing saliency maps. Although these fusion functions are simple to implement, they cannot make full use of the complementary information in the saliency models under fusion and adapt them in various scenarios because they assign equal weights to different saliency models for every image. To solve this problem, Cao et al. [27] proposed to fuse the weak predictions of an input image with a self-adaptively weighting scheme. They calculated the self-adaptive weights of each weak saliency map based on consistency energy in a group of images with similar objects. Thus, it can adapt to the variation of the image scenes to some extent. As can be seen, both [2] and [27] fused the weak saliency predictions only in the image-level, which ignores the region-dependent property of the desirable fusion process. Lately, Mai et al. [28] presented a supervised saliency aggregation method which uses the ground-truth of some similar images to learn a fusion model to aggregate the weak saliency maps of the input image. They considered the image-dependent property and explored the interaction between neighboring pixels, which, however, still ignores the region-dependent property during the fusion process. In addition, by comparing the work in [28] with our work, we can find several other obvious differences:1) Our method works in an unsupervised manner and it is able to automatically fuse the given weak predictions without using the ground truth, thereby alleviating the time-consuming off-line training process or labeling positive samples manually. The purpose of this paper is to design effective inference mechanism to combine those simple unsupervised models to obtain stronger prediction result which is expected to be comparable with or even better than the state-of-the-art saliency methods. In our work, the candidate models to be combined are not necessary to be the best ones; 2) Our method infers the fusion model from the content of the specific input image rather than its nearest neighbor images, which could alleviate the ambiguity during the learning process; 3) When the pre-provided salient object detection results are changed, the work in [28] needs to spend much time to re-train the fusion models whereas our method could infer the fusion model immediately, which makes our method more flexible and efficient.

The contributions of the proposed work can be summarized as follows:

1) We make the earliest effort to consider the different characteristics of various image regions in saliency fusion, which is implemented by defining each image region using a labeling difficulty parameter to reflect its difficulty to be labeled correctly. Consequently, the proposed fusion strategy can autonomously choose to trust different weak saliency models on image regions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

with different labeling difficulties.

2) Our saliency fusion strategy can automatically infer the accuracies of each weak saliency model during the fusion process. The inferred accuracies are demonstrated to be consistent with the accuracies measured by comparing with the ground-truth, which, to some extent, guarantees the reasonability of the superior performance of the fusion results even without the ground-truth.
3) A novel smooth-optimization strategy is proposed based on the local spatial consistency constraint, to ensure high similarity of the saliency labels for the neighboring regions with similar image features, by considering the interactions between the saliency labels of neighboring regions.

II. Fusing The Weak Predictions

Given an input image I, we first obtain m weak saliency prediction maps $SM = \{sm_1, sm_2, ..., sm_m\}$ from m existing saliency detection models $M = \{M_1, M_2, ..., M_m\}$. The weak saliency predictions provide weak saliency labels (1/0 for salient/not salient) for each location/region in the image, and our goal is to fuse these weak saliency labels into the strong labels without any supervision. Specifically, we first define an accuracy parameter for each weak prediction to represent its contribution to the specific image scene, and a difficulty parameter for each image region to measure its difficulty to be detected correctly. By using the GLAD model [26], the accuracy of each weak saliency model and the difficulty of each image region can be determined automatically, which are further used to compute the professionality of each weak saliency label to each specific image region. Finally, the strong saliency label of each image region can be inferred according to the weak saliency labels and their corresponding professionalities.

A. Selection of the weak saliency models

For the set of candidate weak saliency models, our fusion strategy can always automatically analyze the professionality of each saliency model's weak prediction to each image region, and then fuse the weak predictions in the most appropriate way according to their professionalities. Consequently, the fusion results can always be much better than each individual weak prediction. Therefore, the candidate weak saliency models are not fixed and unchangeable. Any reasonable weak saliency models can be used as the candidate models in our work.

We basically use the following two criteria to select weak saliency models:

1. The selected weak saliency models are based on different prior assumptions, image features, and computational methods. Each model may have its own strengths and weaknesses and normally only handle one or a few types of images well.

2. The selected weak saliency models work in the manner of unsupervised learning and have the reasonable good performance. They do not need the training stage and have low computational complexity.

The number of combined weak models can be decided by concerning the tradeoff of the accuracy and computational complexity. Normally, combination of more models may lead to better accuracy but high computational complexity. If a task has high real-time requirement, it is better to select weak saliency models with low computation complexities and reasonably use less weak saliency models. However, if a task has high demand of accuracy, it is better to use a few more saliency models.

In our implementation, by considering both the efficiency and effectiveness of the fusion process, we adopt eight different unsupervised weak saliency models with low computation complexities. The fusion result is much better than each individual saliency detection result and also better than other state-of-the-art supervised saliency detection models as shown in the experiment section. The details of these weak modes are as follows:

(1) **CB [11]**: This model performs the saliency detection by integrating the context-based saliency and object-level shape prior into an iterative energy minimization framework. By considering the object-level prior and using the multi-scale technique, this model can preserve the boundary of the salient object well and obtain large homogeneous salient regions.

(2) **RC [15]**: This model measures each image region's saliency via its contrasts and space distances with all other image regions in the image. This model can generate high quality spatially coherent saliency maps, but some background regions of the images are still salient.

(3) **MC [21]**: This model executes saliency detection via an absorbing Markov chain on a graph model. It computes the saliency of each image region node as its absorbed time to the boundary absorbing nodes and further utilizes the equilibrium probability to suppress the saliency of the long-range smooth background regions.

(4) **DSR [22]**: This model first constructs dense and sparse appearance models from the boundary image regions, based on which it computes each image region's dense and sparse reconstruction errors. Multi-scale reconstruction errors are then propagated and integrated to produce pixel-wise saliency maps. This model can highlight salient regions uniformly and suppress the background saliency quite well.

(5) **RBD [24]**: This model first proposes a robust boundary connectivity measure to characterize each image region's spatial layout with the image boundaries, instead of simply treating all the image boundaries as the background. Then it exploits a principled optimization model to integrate low-level image cues (including the boundary connectivity measure) to produce uniform and clean saliency maps.

(6) **CHM [12]**: This model first uses a cost-sensitive SVM objective function to capture center-surround contrast information and a hypergraph model to capture more comprehensive contextual information, and then linearly combines the SVM and hypergraph saliency detection results to a final saliency map.

(7) **PISA [19]**: This model executes saliency detection by first computing each pixel's color and structure contrasts with spatial priors holistically, then aggregating these two complementary saliency cues in a pixel-wise adaptive manner. By using pixel-wise instead of homogeneous superpixel-based model and taking into consideration the spatial priors, this

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

model can produce spatially coherent yet detail-preserving saliency maps. However, the boundaries of the salient objects are always blurred.

(8) **BL[29]**: This model exploits both weak and strong models to detect salient object. It first constructs weak saliency maps using image priors to generate training samples, based on which it then trains a strong classifier to detect salient pixels.

The selected models exploit different kinds of prior assumptions, image features, and computational methodologies. For example, CB [11] and CHM [12] are based on local contrast. RC [11], PISA [19], and RBD [24] are based on global contrast. MC [21], DSR [22], RBD [24], and CHM [12] are based on background priors. In addition, BL[29] uses the center prior and dark channel prior. From the perspective of image features, CB [11] uses color and hue histograms, PISA [19] uses color and gradient, and BL[29] additionally exploits the Local Binary Pattern (LBP) features. As for the computational methodologies, MC [21] computes the image saliency via an absorbing Markov chain, CHM [12] uses a hypergraph model and a cost-sensitive SVM, and RBD [24] utilizes a principled optimization model.

B. The weak saliency labels for all the regions in the image

After getting the weak predictions $SM=\{sm_1,sm_2,...,sm_m\}$ of I, we generate weak saliency labels for all the image regions from these weak predictions. We first execute over-segmentation to image I by a graph-based image segmentation algorithm in [30], and decompose I into a set of superpixels $SP=\{sp_1,sp_2,...,sp_N\}$, where N is the number of all the superpixels in image I.

Then, a saliency thresholding is performed on $SM=\{sm_1,sm_2,...,sm_m\}$ to obtain m binary salient object segmentation maps $SM'=\{sm_1',sm_2',...,sm_m'\}$. This step aims at extracting the salient regions detected by the weak predictions. Specifically, for weak prediction sm_i, the binary map sm_i' is obtained by:

$$sm_i'(x_1,x_2)=\begin{cases}1, & sm_i(x_1,x_2)\ge T\\ 0, & otherwise\end{cases} \tag{1}$$

where $sm_i(x_1,x_2)$ is the saliency value of the pixel at location (x_1,x_2), and T is the adaptive threshold which is defined as:

$$T=\frac{\lambda}{W\times H}\sum_{x_1=1}^{W}\sum_{x_2=1}^{H}sm_i(x_1,x_2) \tag{2}$$

where W and H are the width and height of sm_i, and λ is the parameter that controls the extraction of candidate salient object regions from the saliency map. A small value of λ means more image regions will be extracted as candidate salient object regions from the corresponding weak prediction. Empirically, we set λ to 2 in our experiments.

Next, we define each superpixel m weak saliency labels (1/0 for salient/not salient) from $SM'=\{sm_1',sm_2',...,sm_m'\}$. For the j th superpixel sp_j, the weak saliency label provided by sm_i' can be described as follows:

$$l_{ij}=\begin{cases}1, & \frac{1}{N_j}\sum_{k=1}^{N_j}sm_i'(sp_j(k))\ge 0.5\\ 0, & otherwise\end{cases} \tag{3}$$

where N_j is the number of pixels in superpixel sp_j, and $sp_j(k)$ is the position index of the k th pixel of sp_j in sm_i'. As shown in (3), if more than half pixels of sp_j belong to the salient regions, then the weak saliency label l_{ij} is set to 1.

Finally, for each superpixel in the input image I, we obtain a set of m weak saliency labels, denoted as $\mathbf{l}_j=\{l_{1j},l_{2j},...,l_{mj}\}$. The whole input image can be described as N weak saliency label sets $\mathbf{l}=\{\mathbf{l}_1,\mathbf{l}_2,...,\mathbf{l}_N\}$.

C. Inferring the strong saliency label from weak saliency labels

1) Parameter definition

Accuracy of the weak saliency model: As different saliency models have different performances on the input image I, we define a labeling accuracy parameter $\alpha_i\in(-\infty,+\infty)$ for each saliency model M_i to represent the accuracy of its weak prediction to I. A large α_i means that the saliency model M_i can detect the salient regions of this image more accurately. The saliency model with an α_i greater than 0 can be considered as effective, while that with α_i less than 0 is considered as invalid and adversarial. When α_i equals to 0, it indicates that the corresponding saliency model cannot distinguish the salient object regions and the background. For this case, the labels given by this saliency model make no contribution to the inference of the real saliency labels of the superpixels. In the extreme case, when $\alpha_i=+\infty$, the corresponding saliency model M_i would be very skilled and it can label all the superpixels correctly, and vice versa.

Difficulty of each superpixel: As mentioned before, even in the same image, different image regions may have different difficulties to be labeled correctly. Thus we define a parameter $1/\beta_j\in[0,+\infty)$ to denote the labeling difficulty of superpixel sp_j .A large $1/\beta_j$ represents that the superpixel sp_j is difficult to be labeled correctly. When $1/\beta_j=+\infty$, it means that the superpixel sp_j is too difficult that even the most expertly saliency model can only have a chance of fifty percent to label it correctly. When $1/\beta_j=0$, it means that the superpixel sp_j is so simple that even the most obtuse saliency model can label it correctly.

With the parameters defined above, we can figure out that the weak saliency labels for one superpixel can be considered to depend on the following factors: (1) the weak saliency model's labeling accuracy; (2) the superpixel's labeling difficulty; (3) The real saliency label of the superpixel. Fig. 3 shows the causal relationship of these factors.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

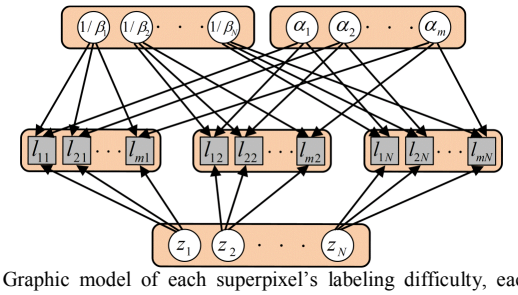


Fig. 3. Graphic model of each superpixel's labeling difficulty, each weak saliency label's labeling accuracy, the superpixel's weak saliency labels, and real saliency label. Only the shaded variables are available during the inference process.

Considering the relationships described above, the professionality of a weak saliency label l_{ij} can be defined as:

$$p\left(l_{ij}=z_j \mid \alpha_i,\beta_j\right)=\frac{1}{1+e^{-\alpha_i\beta_j}} \tag{4}$$

where $z_j\in\{0,1\}$ is the real saliency label of superpixel sp_j, and our ultimate goal is to infer the most likely value of z_j from the weak saliency labels and their professionalities. The professionality of the weak saliency label l_{ij} can also be interpreted as the probability that l_{ij} equals to the real saliency label z_j, i.e. the probability of sp_j being labeled correctly by M_i. As can be seen from the equation, more skilled saliency models (with higher α_i) can achieve a higher probability of labeling a simpler superpixel (with lower $1/\beta_j$) correctly. In addition, when the superpixel's labeling difficulty $1/\beta_j$ approaches to $+\infty$, or the saliency model's labeling accuracy α_i approaches to 0, the probability of sp_j being labeled correctly tends to be 0.5, which means that this weak saliency label l_{ij} makes no contribution to the final labeling result.

2) The posterior probability of z_j

After getting the professionality of each weak saliency label, we compute the posterior probability of $z_j\in\{0,1\}$ from the weak saliency labels as follows:

$$\begin{aligned} p\left(z_j \mid \mathbf{l},\boldsymbol{\alpha},\boldsymbol{\beta}\right) &= p\left(z_j \mid \mathbf{l}_j,\boldsymbol{\alpha},\beta_j\right) \\ &\propto p\left(z_j \mid \boldsymbol{\alpha},\beta_j\right) p\left(\mathbf{l}_j \mid z_j,\boldsymbol{\alpha},\beta_j\right) \\ &\propto p\left(z_j\right)\prod_{i=1}^{m} p\left(l_{ij} \mid z_j,\alpha_i,\beta_j\right) \end{aligned} \tag{5}$$

where $p\left(z_j\right)$ is the prior probability of z_j, and $p\left(z_j\right)=p\left(z_j \mid \boldsymbol{\alpha},\beta_j\right)$ is based on the conditional independence assumptions of the superpixel's strong saliency label, the superpixel's labeling difficulty, and the saliency models' labeling accuracies.

The physical significance of (5) can be interpreted as that the posterior probability of z_j is jointly determined by the prior probability of z_j, the weak saliency labels, and their corresponding professionalities. If we know the labeling difficulty and the prior probability of each superpixel, and the labeling accuracy of each weak saliency model, then we can compute the posterior probability of each superpixel being salient directly.

3) The optimal inference process

As a matter of fact, only the weak saliency labels for each superpixel are available, both the superpixel's labeling difficulty $1/\beta_j$ and the saliency model's labeling accuracy α_i are still unknown. Therefore, according to [26], we use an Expectation–Maximization algorithm (EM) to simultaneously infer the maximum likelihood estimation of the parameters and the posterior probability of z_j.

The EM algorithm can be described as follows:

E-step: According to a current estimation of $\boldsymbol{\alpha},\boldsymbol{\beta}$ from the last M-step and the observed weak saliency labels, we compute the posterior probabilities of all z_j according to (5).

M-step: To estimate the parameters, a standard auxiliary function Q:

$$\left(\boldsymbol{\alpha}',\boldsymbol{\beta}'\right)=\arg\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})} Q\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right) \tag{6}$$

where $\boldsymbol{\alpha}'$ and $\boldsymbol{\beta}'$ are the estimated parameters that locally maximize Q.

The function Q is defined as the expectation of the joint log-likelihood of the observed and hidden variables $(\mathbf{l},\mathbf{z})$ given the parameters $(\boldsymbol{\alpha},\boldsymbol{\beta})$, w.r.t. the posterior probabilities of all the z_j computed in the last E-step:

$$\begin{aligned} Q\left(\boldsymbol{\alpha},\boldsymbol{\beta}\right) &= E\left[\ln p\left(\mathbf{l},\mathbf{z}\middle|\,\boldsymbol{\alpha},\boldsymbol{\beta}\right)\right] \\ &= E\left[\ln\prod_{j=1}^{N}\left(p\left(z_j\right)\prod_{i=1}^{m} p\left(l_{ij}\middle|\,z_j,\alpha_i,\beta_j\right)\right)\right] \\ &\quad \left(\text{since } l_{ij} \text{ are cond. indep. given } \mathbf{z},\boldsymbol{\alpha},\boldsymbol{\beta}\right) \\ &= \sum_{j=1}^{N} E\left[\ln p\left(z_j\right)\right]+\sum_{j=1}^{N}\sum_{i=1}^{m} E\left[\ln p\left(l_{ij}\middle|\,z_j,\alpha_i,\beta_j\right)\right] \end{aligned} \tag{7}$$

where the expectation is taken over $\mathbf{z}$ given the previous parameter values $\boldsymbol{\alpha}^{old}$, $\boldsymbol{\beta}^{old}$ as estimated in the last E-step.

The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are updated in iterations. We set their initial values to be 0.5 since we have no idea of the saliency models' labeling accuracies or the superpixels' labeling difficulties initially. In addition, the prior probability of z_j plays a vital role to the final result, and we have to carefully define it. According to our observation, if all the weak saliency labels of a superpixel are 1, the strong saliency label should have no chance to be 0 after the fusion process. Thus, we define the prior probability of each superpixel as follows: we define the superpixels with all their weak saliency labels as 1 to be the positive superpixels and set their prior probabilities to 1. Conversely, for the superpixels with all their weak saliency labels as 0, we define them as the negative superpixels and set their prior probabilities to 0. For the superpixels whose weak saliency labels contain both 1 and 0, we first find their most similar positive superpixel sp_{pos} and most similar negative

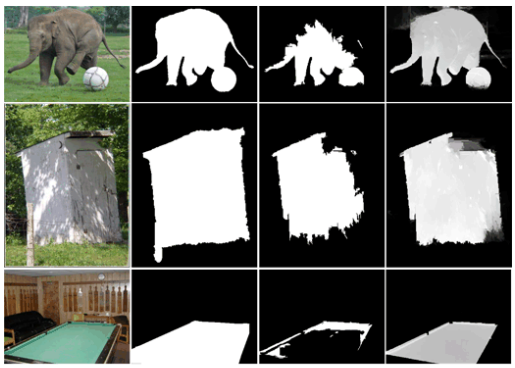
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

superpixel sp_{neg} in the input image via KD-tree search, and then define their prior probabilities based on their similarities with respect to both sp_{pos} and sp_{neg}, which is defined as:

$$p(z_j) \propto \exp(D_{neg} - D_{pos}) \tag{8}$$

where D_{neg} is the Euclidean distance in the feature space from sp_j to sp_{neg}, and D_{pos} is the Euclidean distance from sp_j to sp_{pos}. Note that all the values of $p(z_j)$ are normalized to (0,1).

Finally, we use gradient ascent to estimate the values of $\boldsymbol{\alpha}'$ and $\boldsymbol{\beta}'$ by locally maximizing Q. The final saliency label of each superpixel is obtained by thresholding at 0.5 the superpixel's posterior probability of being salient.



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

Fig. 4. An illustration of the significance of the local spatial consistency constraint. (a) Input images. (b) Ground truth. (c)-(d) Salient object detection results before and after introducing the local spatial consistency constraint.

III. The Local Spatial Consistency Constraint

During the fusion process, all the superpixels in the GLAD model are considered to be independent from each other, which ignores the interaction between neighboring superpixels. This may lead to the poor continuity between the saliency labels of neighboring image regions, i.e. the saliency labels of neighboring superpixels with similar image features may be quite different from each other (as shown in Fig. 4(c)). Therefore, we propose a local spatial consistency constraint to refine the fusion results, which will lead to high similarity for the saliency labels of neighboring superpixels with similar image features.

The local spatial consistency constraint is implemented through a similarity matrix $\mathbf{W}$. Specifically, we construct a graph $G(V,E)$ on the input image, where each node $v_i \in V$ corresponds to a superpixel and each edge $e_{ij} \in E$ connects certain pair of neighboring superpixels. The edges E are weighted by the similarity matrix $\mathbf{W} = [w_{ij}]_{N\times N}$, where the weight w_{ij} of the edge e_{ij} connecting nodes v_i and v_j, is defined as:

$$w_{ij} = \begin{cases} \exp\left(-\dfrac{\|f_i - f_j\|_2^2}{2\sigma^2}\right) & \text{if } e_{ij} \in E, \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where f_i and f_j are the mean image feature vectors of the node v_i and node v_j, respectively, and σ is a constant that controls the strength of the weight. For the image features, we use two kinds of features in this work, including color and texture. In our experiments, we set σ=0.07. Then, the degree matrix of the graph can be computed as $\mathbf{D} = diag\{d_{11},...,d_{nn}\}$, where $d_{ii} = \sum_{j=1} w_{ij}$, and the Laplacian matrix is computed as $\mathbf{S} = \mathbf{D} - \mathbf{W}$.

With the similarity matrix $\mathbf{W}$ and the Laplacian matrix $\mathbf{S}$, we introduce the local spatial consistency constraints to all the superpixels' saliency labels through solving the following optimization problem:

$$\min_{y} \gamma \sum_{i=1}^{N}\sum_{j=1}^{N_n} (y_i - y_j)^2 w_{ij} + \sum_{i=1}^{N} \|y_i - z_i\|_2^2 \tag{10}$$

where $y_i \in [0,1]$ is the final saliency label of node v_i after this optimization process. v_i and v_j are adjacent nodes, and N_n is the number of the nodes neighboring to node v_i. This equation can also be written as:

$$\min_{\mathbf{y}} \gamma(\mathbf{y}^{\mathrm{T}}\mathbf{S}\mathbf{y}) + Tr(\mathbf{y}-\mathbf{z})^{\mathrm{T}}(\mathbf{y}-\mathbf{z}) \tag{11}$$

where $\mathbf{y} = [y_1, y_2, ..., y_N]^{\mathrm{T}}$ and $\mathbf{z} = [z_1, z_2, ..., z_N]^{\mathrm{T}}$.

As can be seen, the first term in (10) and (11) illustrates the local spatial consistency constraint for the saliency labels, which indicates that if two neighboring nodes v_i and v_j have similar image features (w_{ij} is large), they should have similar saliency labels in the final result. The second term is used to ensure that the usage of the proposed constraint does not change the original saliency labels too much. The parameter γ is used to balance these two constraints. Empirically, we set γ=20. The result of (10) is:

$$\mathbf{y} = (\gamma\mathbf{S} + \mathbf{I}_N)^{-1}\mathbf{z} \tag{12}$$

where $\mathbf{I}_N \in \mathbb{R}^{N\times N}$ is an identity matrix. As shown in Fig. 4(d), the saliency labels of neighboring image regions with similar image features in the final salient object detection result are more similar to each other. Thus, the final salient object detection results can have larger continuity between the saliency labels of neighboring superpixels.

IV. Experiments

A. Experimental setup

Datasets: **Datasets**: We evaluated our proposed method on five standard benchmark datasets: SOD [31], MSRA-B [10], HKU-IS [32], DUT-OMRON [23], and THUR-15K [33]. All these datasets contain different kinds of images and somehow

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

challenging. The SOD dataset consists of 300 images from the Berkeley segmentation dataset. This dataset is challenging since most images in this dataset contain salient objects either with low contrast or overlapping with the image boundary. Pixel-wise annotations of these images can be obtained for this dataset. The MSRA-B dataset contains 5000 images with pixel-wise annotations. Most of the images in this dataset contain only one object and have complex background. The HKU-IS dataset contains 4447 images, and most of these images have multiple salient objects which often have low contrast to the image backgrounds. The DUT-OMRON dataset contains 5168 images. This dataset is also challenging since the images have unknown number of salient objects and complex backgrounds. The THUR15K dataset consists of 15000 images downloaded from Flickr with five keywords: butterfly, coffee mug, dog jump, giraffe, and plane. Not every image in this dataset contains a salient object, besides, only the images with salient objects have pixel-wise annotations.

Evaluation criterion: We evaluated the salient object detection performance using the standard precision-recall (PR) curves, F-measure and the intersection-over-union (IOU) score [34]. To obtain the PR curve, the saliency map is first converted to a binary mask using a threshold, then the corresponding precision and curve values are obtained by comparing the binary mask with the ground truth. The PR curve is then obtained by changing the threshold from 0 to 1 and averaged on each dataset.

The F-measure value is the joint performance of the precision and recall:

$$F_\rho = \frac{\left(1+\rho^2\right)\cdot precision \cdot recall}{\rho^2 \cdot precision + recall} \tag{13}$$

We set $\rho^2 = 0.3$ here to emphasize precision and obtain the binary salient object segmentation map by thresholding the saliency map at twice its mean saliency value. The intersection-over-union (IOU) score of one dataset is defined as:

$$\frac{1}{|\tau|}\sum_{t\in\tau}\frac{R_t \cap GT_t}{R_t \cup GT_t} \tag{14}$$

where R_t is the binary salient object segmentation map of image t obtained by thresholding the saliency map at twice its mean saliency. GT is the ground truth, τ represents the image dataset, and $|\tau|$ means the number of images in the dataset.

B. Examination of design options

Parameter analysis. Our fusion strategy defines each weak saliency model a labeling accuracy parameter to measure its contribution to certain image during the fusion process. To examine this parameter, we compare the performance ranks of the weak saliency models based on their obtained labeling accuracy parameters with those based on their mean F-measure and IOU scores on the SOD, MSRA-B, HKU-IS, and DUT-OMRON dataset, respectively. As shown in Fig. 5, the performance ranks of the weak saliency models based on their mean F-measure and IOU scores are consistent on all the four datasets. Meantime, the performance ranks based on the labeling accuracy parameter values of the weak saliency models are mostly consistent with those based on the mean F-measure and IOU scores.

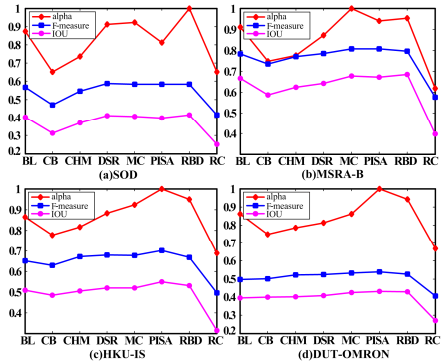


Fig. 5. Comparison between the weak saliency models' performance ranks based on their labeling accuracy parameter values and those based on their mean F-measure and IOU scores on the SOD, MSRA-B, HKU-IS, and DUT-OMRON datasets.

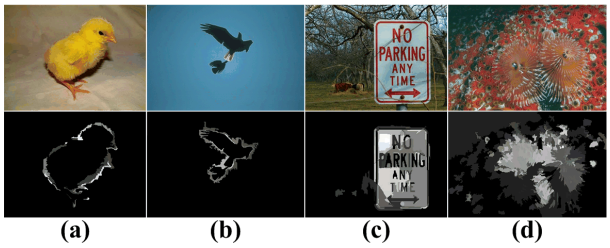


Fig. 6. The visualization of labeling difficulty parameter values of the superpixels in some example input images.

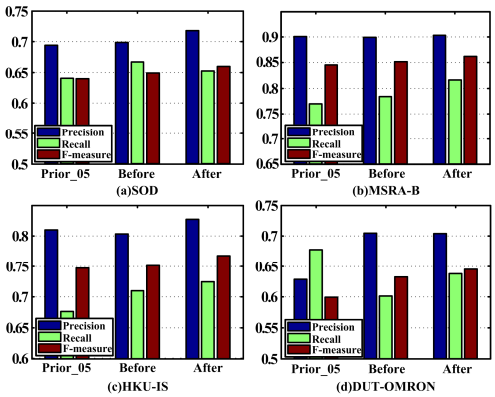


Fig. 7. Evaluation of the prior probability determination scheme and the local spatial consistency constraint in terms of precision, recall and F-measure on the SOD, MSRA-B, HKU-IS, and DUT-OMRON dataset. 'Prior_05': fusion results before introducing the local spatial consistency constraint, where each superpixel's prior probability is defined as 0.5. 'Before': fusion results before introducing the local spatial consistency constraint, where each superpixel's prior probability is defined by our prior probability determination scheme. 'After': fusion results after introducing the local spatial consistency constraint, where each superpixel's prior probability is defined by our prior probability determination scheme.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

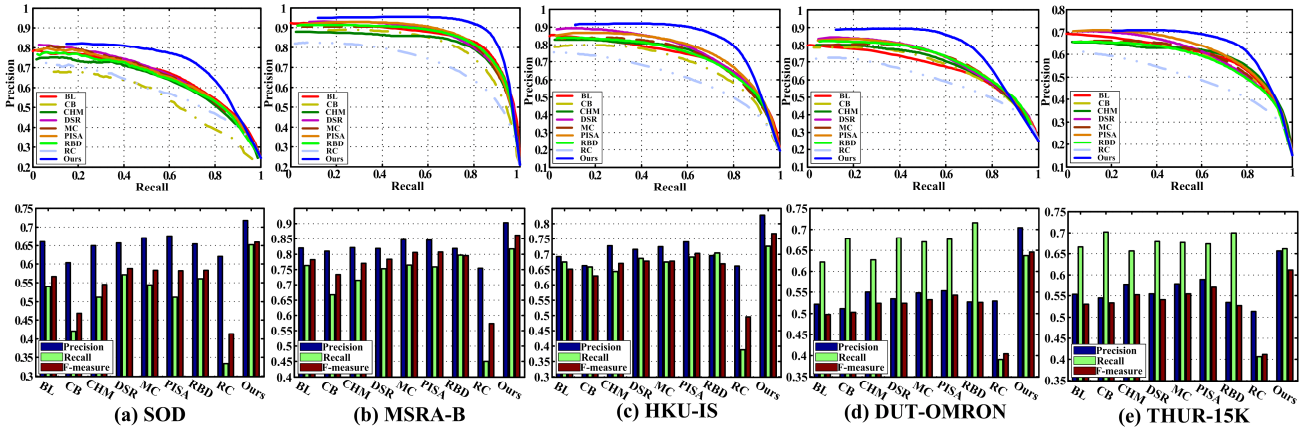


Fig. 8. Performance comparison of the proposed approach and the used weak saliency models in terms of PR-curves, Precision, Recall and F-measure on the SOD, MSRA-B, HKU-IS, DUT-OMRON, and THUR-15K datasets.

In addition to the labeling accuracy parameter, our fusion strategy also defines each superpixel a labeling difficulty parameter to differentiate them during the fusion process. Here we visualize the labeling difficulty parameter values of the superpixels in some example images in Fig. 6. As we can see, for the input image with large homogeneous salient object and clean background regions which have strong contrast with the salient object, only the superpixels on the edges of the salient objects have large labeling difficulties (see Fig. 6 (a)-(b)). However, if the salient objects have complex patterns (see Fig. 6 (c)), the superpixels inside the salient objects are also difficult to label correctly. Even worse, when the background regions have similar image features with the salient object (see Fig. 6 (d)), all the superpixels in the input image are difficult to label during the fusion process.

TABLE I
EVALUATION OF THE PRIOR PROBABILITY DETERMINATION SCHEME AND THE LOCAL SPATIAL CONSISTENCY CONSTRAINT IN TERMS OF IOU SCORES ON THE SOD, MSRA-B, HKU-IS, AND DUT-OMRON DATASETS. 'PRIOR_05': FUSION RESULTS BEFORE INTRODUCING THE LOCAL SPATIAL CONSISTENCY CONSTRAINT, WHERE EACH SUPERPIXEL'S PRIOR PROBABILITY IS DEFINED AS 0.5. 'BEFORE': FUSION RESULTS BEFORE INTRODUCING THE LOCAL SPATIAL CONSISTENCY CONSTRAINT, WHERE EACH SUPERPIXEL'S PRIOR PROBABILITY IS DEFINED BY OUR PRIOR PROBABILITY DETERMINATION SCHEME. 'AFTER': FUSION RESULTS AFTER INTRODUCING THE LOCAL SPATIAL CONSISTENCY CONSTRAINT, WHERE EACH SUPERPIXEL'S PRIOR PROBABILITY IS DEFINED BY OUR PRIOR PROBABILITY DETERMINATION SCHEME.

	SOD	MSRA-B	HKU-IS	DUT-OMRON
Prior_05	0.484	0.713	0.584	0.482
Before	0.495	0.734	0.607	0.500
After	**0.516**	**0.758**	**0.629**	**0.520**

Effectiveness of the prior probability determination scheme. As described in Section II-C, each superpixel's prior probability of being salient plays a vital role to the final fusion results. Therefore we use a novel scheme based on KD-tree search to carefully define them. To show the effectiveness of the prior probability determination scheme, we performed comparison experiments on the SOD, MSRA-B, HKU-IS, and DUT-OMRON dataset, where the prior probability of each superpixel is set to 0.5. The comparison results are shown in Fig. 7 and Table. 1. As can be seen, the superpixels' prior probabilities of being salient affect a lot on the final fusion results and our prior probability determination scheme is very effective.

Effectiveness of the local spatial consistency constraint. After the fusion process, we further introduce a local spatial consistency constraint to ensure high similarity for the saliency labels of neighboring image regions with similar features. Here we compared the salient object detection results of our method before and after using the constraint in Fig. 7 and Table 1. The comparison results show that the usage of this constraint can improve the salient object detection results.

TABLE 2
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH AND THE USED WEAK SALIENCY MODELS IN TERMS OF IOU SCORES ON THE SOD, MSRA-B, HKU-IS, DUT-OMRON, AND THUR-15K DATASETS.

	SOD	MSRA-B	HKU-IS	DUT-OMRON	THUR-15K
BL	0.403	0.666	0.509	0.397	0.432
CB	0.313	0.587	0.485	0.400	0.432
CHM	0.370	0.624	0.505	0.402	0.438
DSR	0.411	0.642	0.519	0.408	0.426
MC	0.407	0.677	0.520	0.425	0.444
PISA	0.401	0.673	0.545	0.433	0.454
RBD	0.415	0.684	0.531	0.430	0.431
RC	0.254	0.399	0.315	0.271	0.287
Ours	**0.516**	**0.758**	**0.629**	**0.520**	**0.503**

C. Performance comparison of the proposed approach with the weak saliency models

We compared our salient object detection results with the weak saliency predictions to be combined, i.e. $SM = \{sm_1, sm_2, ..., sm_m\}$ described in Section II-B. The comparison results are shown in Fig. 8 and Table 2. As can be seen, the saliency detection performance improves significantly compared with the weak saliency predictions used to fuse, especially on the DUT-OMRON dataset, where both the F-measure and IOU scores improve about 10% compared with the best weak saliency prediction, which directly demonstrates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

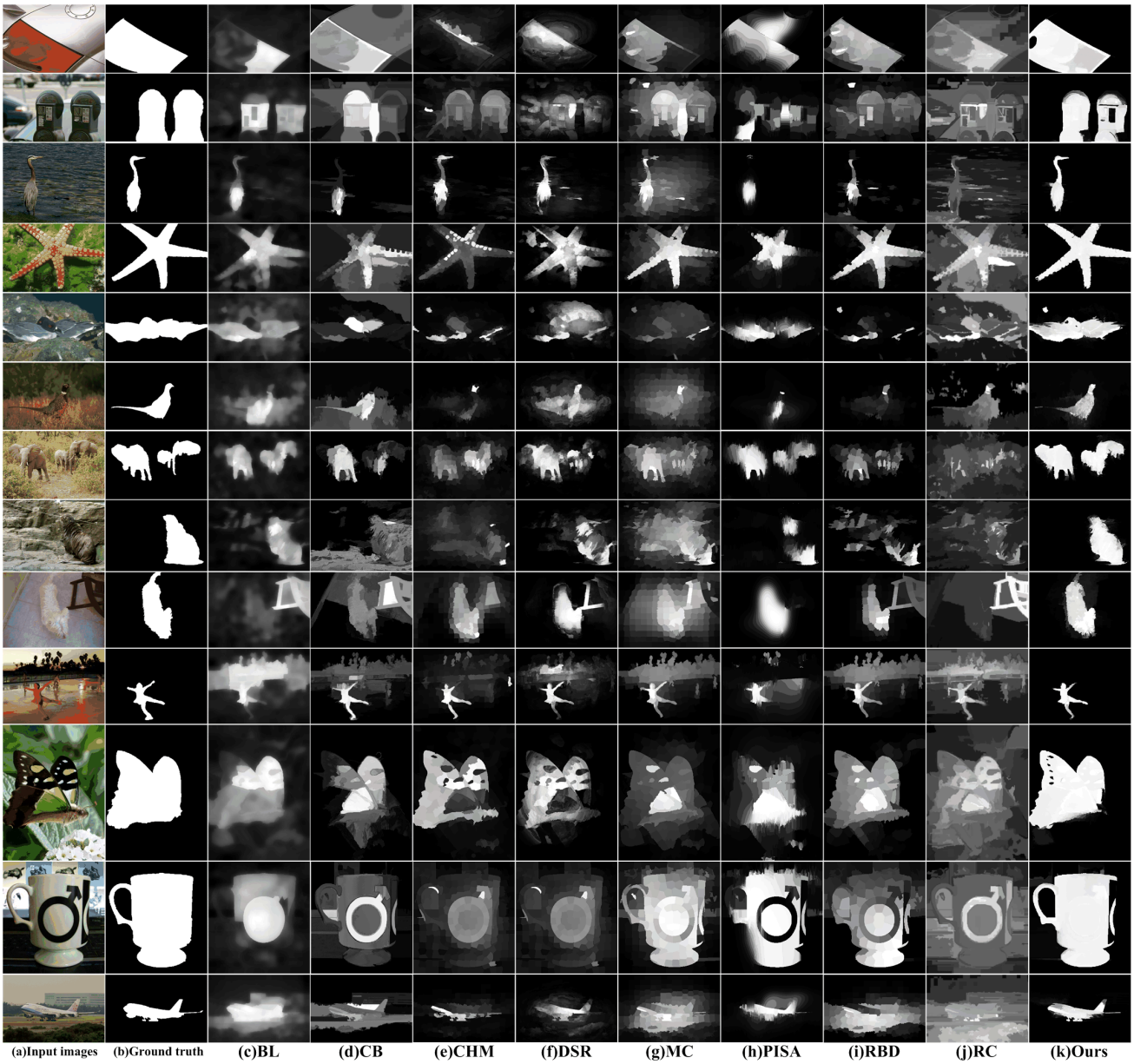


Fig. 9. Visual sample comparisons of the proposed approach and the used salient object detection approaches.

the effectiveness of the proposed framework for saliency fusion.

Fig. 9 shows some visual comparison results of our salient object detection results and the weak predictions of the weak saliency models. We can see from the comparison results that our approach can consistently perform better than the used weak saliency models.

D. Comparison with other salient object detection approaches based on fusing existing saliency models

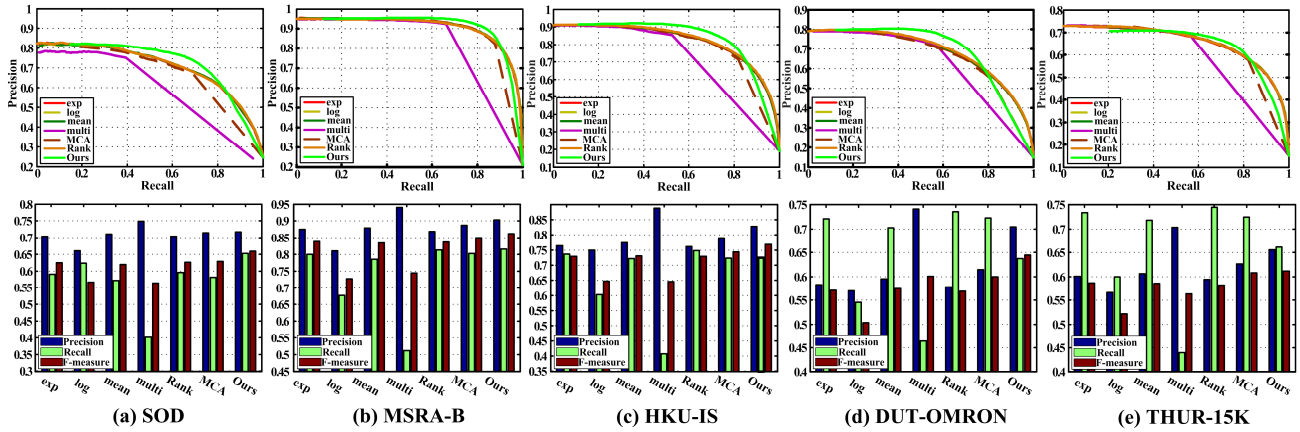
As mentioned before, Boji et al. [2] presented a salient object detection approach which also fuses multiple existing saliency models. Here we compared the proposed approach with the approaches in [2], which fuses the weak predictions via several simple pre-defined standard combination functions. As all of these simple combination functions are executed on pixel level in [2], we run an improved version of these models by executing them on superpixel level. The fusion results of these simple combination functions are denoted as mean, exp, log, multi, respectively, in the following parts.

In addition, we also compared our proposed approach with the approaches in [27, 35]. We adopted the fusion method proposed in [27, 35] to combine the same set of weak saliency models in our work and the obtained results are denoted as Rank and MCA, respectively.

TABLE 3
COMPARISON OF THE MEAN RUNNING TIME BETWEEN OUR METHOD AND [2, 27, 35].

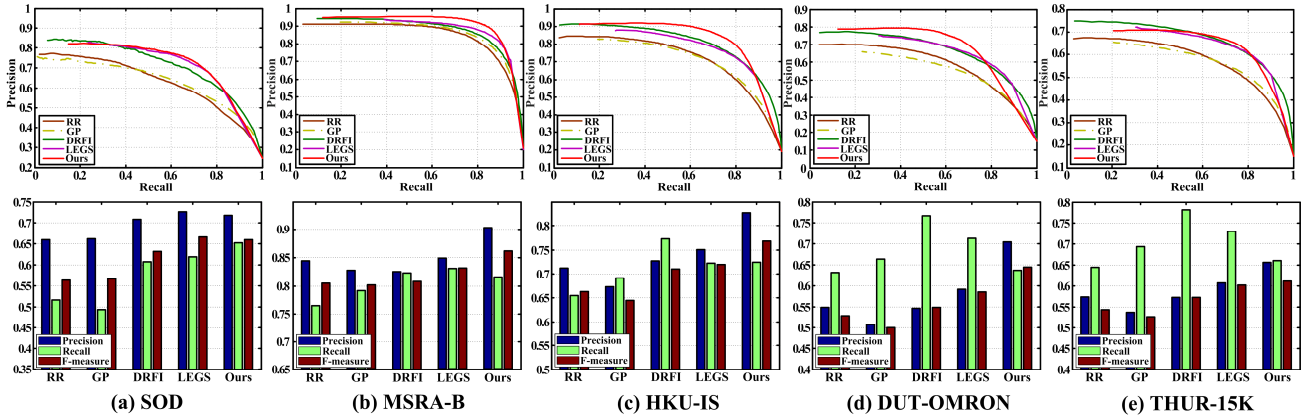
	[2]	[27]	[35]	Ours
Time(s)	0.31	3.65	0.35	0.63

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia



(a) SOD (b) MSRA-B (c) HKU-IS (d) DUT-OMRON (e) THUR-15K

Fig. 10. Performance comparisons of the proposed approach and other fusion based saliency detection approaches [2], [27], and [35] in terms of PR-curves, Precision, Recall and F-measure on the SOD, MSRA-B, HKU-IS, DUT-OMRON, and THUR-15K datasets.



(a) SOD (b) MSRA-B (c) HKU-IS (d) DUT-OMRON (e) THUR-15K

Fig. 11. Performance comparison of the proposed approach and four state-of-the-art salient object detection approaches in terms of PR-curves, Precision, Recall and F-measure on the SOD, MSRA-B, HKU-IS, DUT-OMRON, and THUR-15K datasets.

We first compared the mean running time of our method for one input image after obtaining the weak saliency predictions with that of [2, 27, 35] in Table 3. As can be seen, [2] spent the least time for fusion since it only exploited some simple pre-defined standard combination functions for fusion. In addition, the mean running time of our method is slightly longer than [35]. This is because we execute the fusion process on superpixel level and have to perform superpixel segmentation and define each superpixel a set of weak saliency labels first. On the other hand, although those additional steps have to be executed first, the mean running time of our method is still lower than [27] to a large extent.

The performance comparison results are shown in Fig. 10 and Table 4. As can be seen, the proposed approach outperforms [2, 27, 35] on all five benchmark datasets, which demonstrates the superiority of the proposed saliency fusion strategy over other saliency fusion approaches.

E. Comparison with state-of-the-art salient object detection approaches

Next, to further evaluate our saliency detection performance, we compared the proposed approach with some state-of-the-art saliency detection models such as RR [36], GP [37], DRFI [38], LEGS [39], where DRFI [38] and LEGS [39] work in the supervised manner. The comparison results are shown in Fig. 11 and Table 5. We can see that our method significantly outperforms the unsupervised state-of-the-art saliency detection methods on all tested datasets. More encouragingly, our saliency detection method has also shown to perform even better than the supervised saliency detection methods, including the recent deep learning-based LEGS method, which demonstrates the core insight of this paper, i.e., fusing imperfect unsupervised saliency models may yield superior saliency prediction that is better than the state-of-the-art saliency models, even the supervised ones trained on massive of labeled data.

V. Conclusion

In this paper, we have proposed to tackle the salient object detection as a model fusion problem, where only the weak predictions from the existing imperfect saliency models are offered and no ground truth information is required. We proposed to fuse these weak predictions to obtain the strong saliency predictions by fully making use of each saliency model's strength. During the fusion process, we defined a labeling accuracy parameter for each saliency model to measure its contribution to the input image and a labeling difficulty parameter for each superpixel to differentiate it from all other superpixels. Then we adopted the GLAD [26] model to simultaneously infer each saliency model's labeling accuracy, each superpixel's labeling difficulty, and each superpixel's strong saliency label, respectively. Furthermore, we also

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

TABLE 4
PERFORMANCE COMPARISONS OF THE PROPOSED APPROACH AND OTHER FUSION BASED SALIENCY DETECT APPROACHES [2], [27], AND [35] IN TERMS OF IOU SCORES ON THE SOD, MSRA-B, HKU-IS, DUT-OMRON, AND THUR-15K DATASETS.

	SOD	MSRA-B	HKU-IS	DUT-OMRON	THUR-15K
exp	0.444	0.716	0.583	0.462	0.484
log	0.329	0.608	0.500	0.395	0.427
mean	0.429	0.709	0.582	0.463	0.482
multi	0.278	0.495	0.381	0.399	0.374
Rank	0.459	0.722	0.589	0.465	0.482
MCA	0.456	0.731	0.590	0.490	0.497
Ours	**0.516**	**0.758**	**0.629**	**0.520**	**0.503**

TABLE 5
PERFORMANCE COMPARISON OF THE PROPOSED APPROACH AND FOUR STATE-OF-THE-ART SALIENCY OBJECT DETECTION APPROACHES IN TERMS OF IOU SCORES ON THE SOD, MSRA-B, HKU-IS, DUT-OMRON, AND THUR-15K DATASETS.

	SOD	MSRA-B	HKU-IS	DUT-OMRON	THUR-15K
RR	0.389	0.679	0.506	0.422	0.429
GP	0.381	0.682	0.499	0.401	0.425
DRFI	0.458	0.695	0.580	0.451	0.479
LEGS	0.506	0.725	0.575	0.488	0.506
Ours	**0.516**	**0.758**	**0.629**	**0.520**	**0.503**

introduced a local spatial consistency constraint to ensure high similarity of the saliency labels for neighboring image regions with similar features. The experimental results on five public benchmark datasets have demonstrated that the proposed approach is superior compared with a number of state-of-the-art salient object detection approaches. In future work, we tend to explore some other factors that may also influence the fusion performance and utilize more forthcoming saliency detection models to further refine our results. We will also apply the proposed approach for the tasks of event saliency detection [40] and co-saliency detection [41-42].

REFERENCES

[1] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?," in CVPR, pp. 1137-1144, 2004.

[2] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in ECCV, pp. 414-429, 2012.

[3] B. C. Ko, and J.-Y. Nam, "Object-of-interest image segmentation based on human attention and semantic region clustering," J. Optical Soc. of Am. A, vol. 23, no. 10, pp. 2462-2470, 2006.

[4] W. Wang, J. Shen, X. Li, and F. Porikli, "Robust video object co-segmentation," IEEE Trans. Image Process., vol. 24, no. 10, pp. 3137-3148, 2015.

[5] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," IEEE Trans. Image Process., vol. 13, no. 10, pp. 1304-1318, 2004.

[6] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in ICCV, pp. 2232-2239, 2009.

[7] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2Photo: Internet image montage," ACM Trans. Graph., vol. 28, no. pp. 1-10, 2009.

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

[9] Y.-F. Ma, and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in MM, pp. 374-381, 2003.

[10] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 2, pp. 353-367, 2011.

[11] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in BMVC, pp. 1-12, 2011.

[12] X. Li, Y. Li, C. Shen, A. Dick, and A. V. D. Hengel, "Contextual hypergraph modeling for salient object detection," in ICCV, pp. 3328-3335, 2013.

[13] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in ICCVS, pp. 66-75, 2008.

[14] R. Achantay, S. Hemamiz, F. Estraday, and S. Susstrunky, "Frequency-tuned salient region detection," in CVPR, pp. 1597-1604, 2009.

[15] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in CVPR, pp. 409-416, 2011.

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

[17] X. Hou, and L. Zhang, "Saliency detection: A spectral residual approach," in CVPR, pp. 1 - 8, 2007.

[18] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?," in CVPR, pp. 1139-1146, 2013.

[19] K. Shi, K. Wang, J. Lu, and L. Lin, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors," in CVPR, pp. 2115-2122, 2013.

[20] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in CVPR, pp. 1155-1162, 2013.

[21] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing Markov Chain," in ICCV, pp. 1665-1672, 2013.

[22] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in ICCV, pp. 2976-2983, 2013.

[23] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in CVPR, pp. 3166-3173, 2013.

[24] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in CVPR, pp. 2814-2821, 2014.

[25] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," IEEE Trans. Circuits Syst. Video Technol., vol. 25, no. 8, pp. 1309 - 1321, 2015.

[26] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in NIPS, pp. 2035-2043, 2009.

[27] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-Adaptively Weighted Co-Saliency Detection via Rank Constraint," IEEE Trans. Image Process., vol. 23, no. 9, pp. 4175-4186, 2014.

[28] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in CVPR, pp. 1131-1138, 2013.

[29] N. Tong, H. Lu, R. Xiang, and M. H. Yang, "Salient object detection via bootstrap learning," in CVPR, pp. 1884-1892, 2015.

[30] P. F. Felzenszwalb, and D. P. Huttenlocher, "Efficient graph-based image segmentation," IJCV, vol. 59, no. 2, pp. 167-181, 2004.

[31] V. Movahedi, and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," IEEE Comput. Socie. Workshop. on Percept. Organ. in Comput. Vision, pp. 49-56, 2010.

[32] G. Li, and Y. Yu, "Visual saliency based on multiscale deep features," in CVPR, pp. 5455-5463, 2015.

[33] M. M. Cheng, N. J. Mitra, X. Huang, and S. M. Hu, "SalientShape: group saliency in image collections," The Visual Computer, vol. 30, no. 4, pp. 443-453, 2014.

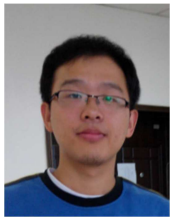
[34] Y. Li, J. Liu, Z. Li, Y. Liu, and H. Lu, "Object co-segmentation via discriminative low rank matrix recovery," in MM, pp. 749-752, 2013.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2017.2763780, IEEE Transactions on Multimedia

[35] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via Cellular Automata," in CVPR, pp. 110-119, 2015.
[36] C. Li, Y. Yuan, W. Cai, and Y. Xia, "Robust saliency detection via regularized random walks ranking," in CVPR, pp. 2710-2717, 2015.
[37] P. Jiang, N. Vasconcelos, and J. Peng, "Generic Promotion of Diffusion-Based Salient Object Detection," in ICCV, pp. 217-225, 2015.
[38] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in CVPR, pp. 2083-2090, 2013.
[39] L. Wang, H. Lu, X. Ruan, and M. H. Yang, "Deep networks for saliency detection via local estimation and global search," in CVPR, pp. 3183-3192, 2015.
[40] D. Zhang, J. Han, L. Jiang, S. Ye, X. Chang, "Revealing Event Saliency in Unconstrained Video Collection," IEEE Trans. Image Process., vol. 26, no. 4, pp. 1746-1758, 2017.
[41] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 5, pp. 865-878, 2017.
[42] X. Yao, J. Han, D. Zhang, and F. Nie, " Revisiting Co-Saliency Detection: A Novel Approach Based on Two-Stage Multi-View Spectral Rotation Co-clustering," IEEE Trans. Image Process., vol. 26, no. 7, pp. 3196-3209, 2017.

Rong Quan received the B.S. degree in information engineering from Northwestern Polytechnical University, in 2013. She is currently working toward the Ph.D. degree in pattern recognition and intelligent system, School of Automation, Northwestern Polytechnical University. Her research interests include computer vision, machine learning, visual saliency detection, and object co-segmentation.

Junwei Han (M'12–SM'15) is a currently a Full Professor with Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, multimedia processing, and brain imaging analysis. He is an Associate Editor of IEEE Trans. on Human-Machine Systems, Neurocomputing, and Multidimensional Systems and Signal Processing.



Dingwen Zhang received his B.E. degree from the Northwestern Polytechnical University, Xi'an, China, in 2012. He is currently pursuing the Ph.D. degree at Northwestern Polytechnical University. His research interests include computer vision and multimedia processing, especially on saliency detection, co-saliency detection, and weakly supervised learning.

Feiping Nie received the Ph.D. degree in Computer Science from Tsinghua University, China in 2009. His research interests are machine learning and its applications, such as pattern recognition, data mining, computer vision, and information retrieval. He has published more than 100 papers in the following top journals and conferences: TPAMI, IJCV, TIP, ICML, NIPS, KDD, IJCAI, AAAI. He is now serving as Associate Editor or PC member for several prestigious journals and conferences in the related fields.

Xueming Qian (M'10) received the B.S. and M.S. degrees in Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the Ph.D. degree in the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2008, after that he was an assistant professor. He was an associate professor from Nov. 2011 to March 2014, and now he was a full professor. His research interests include social media big data mining and search.

Xuelong Li (M'02–SM'07–F'12) is a Full Professor with the Center for OPTical IMagery Analysis and Learning, State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.