2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)

M3S-NIR: Multi-Modal Multi-Scale Noise-Insensitive Ranking for RGB-T Saliency Detection

(Paper ID:10)

Zhengzheng Tu
Anhui University
School of Computer Science and Technology
Hefei,China
zhengzhengahu@163.com

Tian Xia
Anhui University
School of Computer Science and Technology
Hefei,China
tianxia.ahu@foxmail.com

Chenglong Li
Anhui University
School of Computer Science and Technology
Hefei,China
lcl1314@foxmail.com

Yijuan Lu
Texas State University
Department of Computer Science
San Marcos USA
lu@txstate.edu

Jin Tang
Anhui University
School of Computer Science
and Technology
Hefei,China
jtang99029@foxmail.com

***Abstract*—RGB-Thermal saliency detection is to use thermal infrared information to assist salient object detection with visible light information. Multi-Modal Multi-Scale Noise-Insensitive Ranking (M3S-NIR), is proposed for RGB-Thermal (RGB-T) saliency detection. Given spatially aligned RGB and thermal images, M3S-NIR first segments them together into a set of multi-scale superpixels. Second, it takes these superpixels as graph nodes and performs multi-modal multi-scale manifold ranking to achieve saliency calculation, in which the cross-modal and cross-scale collaborations are performed to integrate different kinds of information. Third, to handle noises and corruptions of ranking seeds (i.e., boundary superpixels) introduced by salient objects and RGB-T alignment, M3S-NIR introduces an intermediate variable to infer the optimal ranking seeds, and formulates it as a sparse learning problem. Finally, M3S-NIR uses a unified ADMM (Alternating Direction Method of Multipliers)-based optimization framework to solve the ranking model efficiently. Extensive experiments on the benchmark dataset demonstrate the effectiveness of the proposed approach over other state-of-the-art RGB-T saliency detection methods.**

***Keywords*-RGB-Thermal Saliency Detection; Multi-Modal Multi-Scale ; Noise-Insensitive Ranking; Manifold Ranking.**

I. Introduction

Image saliency detection aims to find out the most salient and important regions in an image, and has been rapidly developed in the field of computer vision in past decade. However, there is still an open problem due to existence of many challenging factors in detecting salient objects especially in various environmental conditions (e.g., low illumination, rain, haze and smog, etc.), which significantly limit the imaging quality of visible spectrum.

As a new branch in image saliency detection, RGB-Thermal (RGB-T) saliency detection supplements thermal infrared information to assist visible light information for salient object detection [1]. Thermal sensors convert invisible target surface temperature distribution into visible thermal images which are insensitive to lighting conditions and have a strong ability to penetrate haze and smog. Therefore, thermal images can overcome some drawbacks that traditional saliency detection models are hard to deal with, such as light changing, shadows or reflective light in the surroundings and so on.

Given potentials of thermal data, RGB-T saliency detection has attracted many attentions recently. Li *et al.* [1] proposed a multi-task manifold ranking algorithm for RGB-T saliency detection, and built up a unified RGB-T saliency detection benchmark dataset, based on which our work will be tested on. Despite of recent progresses, RGB-T saliency detection still faces the following challenging problems. i) There may exist different objects of different sizes in the image and there could be a big object without interior appearance consistency; ii) Proved in [1], there are some noises caused by image registration in the boundary of RGB-T images and some objects may be next to the boundary of an image, which is also very common in RGB-T saliency detection. To handle aforementioned issues, we employ the idea of graph-based manifold ranking [2], seeking for good saliency detection performance in terms of accuracy and speed. Moreover, we employ multi-modal information on the ranking function to integrate different modalities collaboratively. In order to integrate multi-scale appearance consistency of the salient object, we propose a unified computational framework to incorporate the multi-scale cues, which is also beneficial to integrate other priors or constraints. To handle noises and corruptions caused by registration errors or cross-boundary of salient objects, we introduce an intermediate variable to infer the optimal

978-1-7281-1198-8/19/$31.00 ©2019 IEEE
DOI 10.1109/MIPR.2019.00032

ranking seeds, and formulate it as a sparse learning problem. Finally, we used ADMM [3](Alternating Direction Method of Multipliers)-based optimization framework to solve the ranking model efficiently. The main contributions of this work are summarized as follows:

- We propose an effective approach to adopt multiple cues in image saliency detection. Extensive experiments show that the proposed method outperforms the state-of-the-art methods on the benchmark dataset.
- In order to avoid noises and corruptions on the boundary, we infer the optimal seed nodes in graph-based manifold ranking by introducing an intermediate variable, and formulating it as a sparse learning problem.
- We integrate multi-scale appearance consistency of the salient object into a unified ranking model to detect salient objects accurately with different sizes and complex appearances.
- A unified ADMM-based optimization framework to solve each subproblem to get a closed-form solution with fast convergence speed and small computational complexity.

II. M3S-NIR Algorithm

In this section, we will introduce the proposed M3S-NIR model and the associated optimization algorithm. As M3S-NIR bases on graph-based multi-task manifold ranking, we first review it for clarity. Next, we detail M3S-NIR as follows.

A. Problem Formulation

Since different objects are with different size in the image or big objects are without interior appearance consistency, we adopt multi-grained superpixel segmentation to incorporate multi-scale cues to handle size variation of salient objects. Given a pair of RGB and thermal images, we segment them into n^k superpixels for the k-th scale using information of joint M modalities which can mitigate noise effects of some individual source, through the Simple Linear Iterative Clustering (SLIC) algorithm [4], where $k \in \{1, 2, ..., K\}$, K and M are the scale number and the modality number, respectively. In this work, RGB-T data is the special case of $M = 2$, and we discuss its general form for other potential applications.

For the k-th scale, we take superpixels as graph nodes to construct a k-regular graph $G^k = (V^k, E^k)$, where V^k is a node set and E^k is a set of undirected edges at the k-th scale. We denote $X^{m,k} = \{\mathbf{x}_1^{m,k}, ..., \mathbf{x}_{n^k}^{m,k}\} \in \mathbb{R}^{d \times n^k}$ as the feature descriptor of the m-th modal superpixels at the k-th scale, where d is the feature dimension. Some superpixels are labeled as queries and the rest need to be ranked according to their affinities to the queries. Let $\mathbf{s}^k : X^k \rightarrow \mathbb{R}^{n^k}$ denotes a ranking function that assigns a ranking value $\mathbf{s}_i^k$ to each superpixel $\mathbf{x}_i^k$, and $\mathbf{s}^k$ can be viewed as a vector $\mathbf{s}^k = [s_1^k, ..., s_{n^k}^k]^T$. Let $\mathbf{y}^k = [\mathbf{y}_1^k, ..., \mathbf{y}_{n^k}^k]^T$ denote an indication vector, where $\mathbf{y}_i^k = 1$ if $\mathbf{x}_i^k$ is a query, and $\mathbf{y}_i^k = 0$ otherwise. Through aggregating multi-scale graph information, the optimal ranking of queries are computed by solving the following optimization problem:

$$\min_{\{\mathbf{s}^k\},\mathbf{s}} \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j=1}^{n^k} \sum_{m=1}^{M} \mathbf{W}_{ij}^{m,k} || \frac{s_i^k}{\sqrt{\mathbf{D}_{ii}^{m,k}}} - \frac{s_j^k}{\sqrt{\mathbf{D}_{jj}^{m,k}}} ||^2 \tag{1}$$
$$+ \lambda ||\mathbf{s}^k - \mathbf{y}^k||^2 + \lambda_3 ||\mathbf{C}^k \mathbf{s}^k - \mathbf{s}||_F^2,$$

where λ and λ_3 are balanced parameters, and $\mathbf{s}$ is the target ranking vector which will be used for saliency detection in the next section. $\mathbf{D}^{m,k} = diag\{\mathbf{D}_{11}^{m,k}, ..., \mathbf{D}_{nn}^{m,k}\}$ is the degree matrix of the m-th modal graph at the k-th scale, and $\mathbf{D}_{ii}^{m,k} = \sum_j \mathbf{W}_{ij}^{m,k}$. $\mathbf{W}_{ij}^{m,k}$ is the affinity matrix, which is defined as: if nodes V_i^k and V_j^k is connected, we assign it with an edge, and the edge weight is as follows:

$$\mathbf{W}_{ij}^{m,k} = e^{-\gamma ||\mathbf{c}_i^{m,k} - \mathbf{c}_j^{m,k}||}, \tag{2}$$

where $\mathbf{c}_i^{m,k}$ denotes the mean CIELab colors of the i-th superpixel in the m-th modality at the k-th scale, and γ is a scaling parameter. The multi-scale ranking vectors are collaboratively integrated to compute the target ranking vector by introducing the cross-scale consistent matrix $\mathbf{C}^k$, The $\mathbf{C}^k = [C_{ij}^k]_{\pi \times \pi_k}$ is defined as follows:

$$C_{ij}^k = \begin{cases} \frac{N_j{}^k}{N_i}, & overlap \\ \\ 0, & otherwise \end{cases} \tag{3}$$

N_i denotes the total pixel numbers of the i-th superpixel in the original scale. The number of superpixels in the k-th scale are π_k and the number of superpixels in the original scale are π. $N_j{}^k$ denotes the total pixel numbers in the j-th superpixel of k-th scale, overlapping with the i-th superpixel in the original scale. In (1), we can see that the multi-scale ranking vectors (i.e., $\{\mathbf{s}^k\}$) are codetermined by multi-modal information (i.e., $\{\mathbf{W}_{ij}^{m,k}\}$), and the final ranking vector (i.e., $\mathbf{s}$) is optimized by utilizing multi-scale ranking vectors (i.e., $\{\mathbf{s}^k\}$) collaboratively. As discussed above, there are some flaws in image boundaries of the dataset we use, in order to solve this problem, we introduce two constraints for inferring the optimal indication vectors $\hat{y}^k$, i.e., *visual similarity constraint* and *noise sparsity constraint*, as follows:

$$\min_{\hat{\mathbf{y}}^k} \lambda_1 \sum_{i,j=1}^{n^k} \sum_{m=1}^{M} \mathbf{W}_{ij}^{m,k} (\hat{\mathbf{y}}_i^k - \hat{\mathbf{y}}_j^k)^2 + \lambda_2 ||\hat{\mathbf{y}}^k - \mathbf{y}^k||_1, \tag{4}$$

where λ_1 and λ_2 are balanced parameters. We call it as the noise-insensitive model in this paper, as it can alleviate indication vector noises (i.e., boundary seed noises) in ranking process. Note that the first constraint assumes that visually similar superpixels should have same labels and saliency values, and vice versa. Therefore, we add a smoothness term $\sum_{i,j=1}^{n^k}\sum_{m=1}^{M}\mathbf{W}_{ij}^{m,k}(\hat{\mathbf{y}}_i^k - \hat{\mathbf{y}}_j^k)^2$ that can make visual similarity become a graph smoothness constraint. The second constraint aiming to compel noise sparsity in $\hat{\mathbf{y}}^k$ is enlightened by the common use of l_1-norm sparsity regularization term in data noise, which has been proven to be effective even when the data noise is not sparse [5]. Therefore, we formulate it as $||\hat{\mathbf{y}}^k - \mathbf{y}^k||_1$, where l_1-norm is used to promote sparsity on the difference of initial seeds and inferred seeds (because most of the initial seeds should be correct and the remaining ones are noises). We integrate the noise-insensitive model into (1), and obtain the final Multi-Modal Multi-Scale Noise-Insensitive Ranking (M3S-NIR) model as follows:

$$\begin{aligned}
&\min_{\{\mathbf{s}^k\},\mathbf{s},\hat{\mathbf{y}}^k} \frac{1}{2}\sum_{k=1}^{K}\sum_{i,j=1}^{n^k}\sum_{m=1}^{M}\mathbf{W}_{ij}^{m,k}||\frac{{s_i}^k}{\sqrt{\mathbf{D}_{ii}^{m,k}}} - \frac{{s_j}^k}{\sqrt{\mathbf{D}_{jj}^{m,k}}}||^2 + \\
&+ \lambda_1\sum_{i,j=1}^{n^k}\sum_{m=1}^{M}\mathbf{W}_{ij}^{m,k}(\hat{\mathbf{y}}_i^k - \hat{\mathbf{y}}_j^k)^2 + \lambda_2||\hat{\mathbf{y}}^k - \mathbf{y}^k||_1 \\
&+ \lambda||\mathbf{s}^k - \hat{\mathbf{y}}^k||_F^2 + \lambda_3||\mathbf{C}^k\mathbf{s}^k - \mathbf{s}||_F^2.
\end{aligned} \tag{5}$$

With simple algebra, the problem (5) can be converted into:

$$\begin{aligned}
&\min_{\{\mathbf{s}^k\},\mathbf{s},\hat{\mathbf{y}}^k} \sum_{k=1}^{K}\sum_{m=1}^{M}(\mathbf{s}^k)^T\mathbf{L}^{m,k}\mathbf{s}^k + \lambda||\mathbf{s}^k - \hat{\mathbf{y}}^k||_F^2 \\
&+ \lambda_2||\hat{\mathbf{y}}^k - \mathbf{y}^k||_1 + \lambda_3||\mathbf{C}^k\mathbf{s}^k - \mathbf{s}||_F^2 \\
&+ 2\lambda_1\sum_{m=1}^{M}(\hat{\mathbf{y}}^k)^T(\mathbf{D}^{m,k} - \mathbf{W}^{m,k})\hat{\mathbf{y}}^k,
\end{aligned} \tag{6}$$

where $\mathbf{L}^{m,k} = \mathbf{I}^k - (\mathbf{D}^{m,k})^{-\frac{1}{2}}\mathbf{W}^{m,k}(\mathbf{D}^{m,k})^{-\frac{1}{2}}$, $\mathbf{I}^k$ is the identity matrix with the same size of $\mathbf{L}^{m,k}$. Although (6) seems complex, as demonstrated in the experiments, its parameters are easy to adjust, and the detection performance is insensitive to parameter variations.

B. Model Optimization

Although the variables of (6) are not joint convex, the subproblem of each variable with others fixed is convex and has a closed-form solution. We apply ADMM to solve the problems like (6), we introduce auxiliary variables $\hat{\mathbf{y}}^k = \mathbf{f}^k, k = 1, 2, .., K$ to make (6) separable:

$$\begin{aligned}
&\min_{\{\mathbf{s}^k\},\mathbf{s},\hat{\mathbf{y}}^k} \sum_{k=1}^{K}\sum_{m=1}^{M}(\mathbf{s}^k)^T\mathbf{L}^{m,k}\mathbf{s}^k + \lambda||\mathbf{s}^k - \hat{\mathbf{y}}^k||_F^2 + \lambda_3||\mathbf{C}^k\mathbf{s}^k - \mathbf{s}||_F^2 \\
&+ 2\lambda_1\sum_{m=1}^{M}(\mathbf{f}^k)^T(\mathbf{D}^{m,k} - \mathbf{W}^{m,k})\mathbf{f}^k + \lambda_2||\hat{\mathbf{y}}^k - \mathbf{y}^k||_1, \\
&s.t. \quad \hat{\mathbf{y}}^k = \mathbf{f}^k, k = 1, 2, .., K.
\end{aligned} \tag{7}$$

Algorithm 1 Optimization Procedure to (8)

```
Input: Multi-scale affinity matrices $\{\mathbf{W}^k\}$, indication vectors $\{\mathbf{y}^k\}$, and the parameters $\lambda$, $\lambda_1$, $\lambda_2$ and $\lambda_3$;
    Set $\mu = 10^{-3}$, $\rho = 1.5$, $max_\mu = 10^{10}$, $\varepsilon = 10^{-4}$ and $maxIter = 50$.
Output: $\mathbf{s}$, $\{\mathbf{s}^k\}$ and $\{\hat{\mathbf{y}}^k\}$.
for $t = 1 : maxIter$ do
    for $k = 1 : K$ do
        Update $\mathbf{f}^k$ by (9);
        Update $\hat{\mathbf{y}}^k$ by (10);
        Update $\mathbf{s}^k$ by (11);
    end for
    Update $\mathbf{s}$ by (12);
    Update Lagrangian multiplier by $\mathbf{p}^k = \mathbf{p}^k + \mu(\hat{\mathbf{y}}^k - \mathbf{f}^k)$ for all $k$;
    Update Lagrangian parameter by $\mu = \min\{\rho * \mu, max_\mu\}$;
    if Check the convergence condition: the maximum element changes of all variables are lower than $\varepsilon$ or the iteration number reaches $maxIter$ then
        Terminate the loop.
    end if
end for
```

The augmented Lagrange function of (5) is:

$$\begin{aligned}
&\sum_{k=1}^{K}(\mathbf{s}^k)^T(\sum_{m=1}^{M}\mathbf{L}^{m,k})\mathbf{s}^k + \lambda||\mathbf{s}^k - \hat{\mathbf{y}}^k||_F^2 + \lambda_3||\mathbf{C}^k\mathbf{s}^k - \mathbf{s}||_F^2 \\
&+ 2\lambda_1\sum_{m=1}^{M}(\mathbf{f}^k)^T(\mathbf{D}^{m,k} - \mathbf{W}^{m,k})\mathbf{f}^k + \lambda_2||\hat{\mathbf{y}}^k - \mathbf{y}^k||_1 \\
&+ \frac{\mu}{2}||\hat{\mathbf{y}}^k - \mathbf{f}^k + \frac{\mathbf{p}^k}{\mu}||_F^2 - \frac{1}{2\mu}||\mathbf{p}^k||_F^2,
\end{aligned} \tag{8}$$

where $\mathbf{p}^k$ is the Lagrangian multiplier, and μ is the Lagrangian parameter. We then alternatively update one variable by minimizing (8) with fixing other variables. The solutions to these sub-problems are shown below.

Solving $\mathbf{f}^k$: When fixing other variables, we reformulate (8) as follows:

$$\begin{aligned}
&\min_{\mathbf{f}^k} 2\lambda_1(\mathbf{f}^k)^T\sum_{m=1}^{M}(\mathbf{D}^{m,k} - \mathbf{W}^{m,k})\mathbf{f}^k + \frac{\mu}{2}||\hat{\mathbf{y}}^k - \mathbf{f}^k + \frac{\mathbf{p}^k}{\mu}||_F^2, \\
&\Rightarrow \mathbf{f}^k = (4\lambda_1\sum_{m=1}^{M}(\mathbf{D}^{m,k} - \mathbf{W}^{m,k}) + \mu\mathbf{I})^{-1}(\mu\hat{\mathbf{y}}^k + \mathbf{p}^k),
\end{aligned} \tag{9}$$

Solving $\hat{\mathbf{y}}^k$: Similar to $\mathbf{f}^k$, $\hat{\mathbf{y}}^k$ can be solved by the soft thresholding method with closed-form solution:

$$\begin{aligned}
&\min_{\hat{\mathbf{y}}^k} \lambda||\mathbf{s}^k - \hat{\mathbf{y}}^k||_F^2 + \lambda_2||\hat{\mathbf{y}}^k - \mathbf{y}^k||_1 + \frac{\mu}{2}||\hat{\mathbf{y}}^k - \mathbf{f}^k + \frac{\mathbf{p}^k}{\mu}||_F^2, \\
&\Rightarrow \hat{\mathbf{y}}^k = soft\_thr(\mathbf{s}^k, \mathbf{f}^k - \frac{\mathbf{p}^k}{\mu}, \hat{\mathbf{y}}^k, \lambda, \frac{\mu}{2}, \lambda_2),
\end{aligned} \tag{10}$$

where $soft\_thr$ indicates the soft threshold operation.

Solving $\mathbf{s}^k$: By fixing others, the $\mathbf{s}^k$ subproblem can be reformulated as follow:

$$\begin{aligned}&\min_{\mathbf{s}^k}(\mathbf{s}^k)^T(\sum_{m=1}^{M}\mathbf{L}^{m,k})\mathbf{s}^k+\lambda||\mathbf{s}^k-\hat{\mathbf{y}}^k||_F^2+\lambda_3||\mathbf{C}^k\mathbf{s}^k-\mathbf{s}||_F^2,\\&\Rightarrow\mathbf{s}^k=(\frac{\mathbf{L}^k}{\lambda}+\mathbf{I}+\frac{\lambda_3}{\lambda}(\mathbf{C}^k)^T\mathbf{C}^k)^{-1}(\hat{\mathbf{y}}^k+\frac{\lambda_3}{\lambda}(\mathbf{C}^k)^T\mathbf{s}),\end{aligned}\tag{11}$$

where $\mathbf{L}^k$=$\sum_{m=1}^{M}\mathbf{L}^{m,k}$.

Solving $\mathbf{s}$: Once obtained $\mathbf{s}^k$ and $\hat{\mathbf{y}}^k$, updating $\mathbf{s}$ turns to be:

$$\begin{aligned}&\min_{\mathbf{s}}\sum_{k=1}^{K}\lambda_3||\mathbf{C}^k\mathbf{s}^k-\mathbf{s}||_F^2,\\&\Rightarrow\mathbf{s}=\frac{1}{K}\sum_{k=1}^{K}\mathbf{C}^k\mathbf{s}^k.\end{aligned}\tag{12}$$

We summarize whole optimization procedure in Algorithm 1.

III. Two-stage RGB-T Saliency Detection

In this section, we detail the two-stage ranking scheme for bottom-up RGB-T saliency detection based on background and foreground queries.

Ranking with Background Queries. Based on some early works [2], [6], we also utilize the boundary superpixels as background seeds initially, then choose high confident superpixels (low ranking scores in all modalities) belonging to the foreground as the foreground seeds. To be specific, we construct four saliency maps via boundary priors, then combine them for the final saliency map, which is also called the separation/combination (SC)approach [2].

For example, we use the left boundary superpixels as the background queries and the other nodes as the unlabeled superpixles. We run the M3S-NIR algorithm to obtain the ranking vector $\mathbf{s}$ and then normalize it to the range between 0 and 1 denoting as $\hat{\mathbf{s}}$. The saliency map $\mathbf{s}_l$ using the left boundary prior and can be written as:

$$\mathbf{s}_l=1-\hat{\mathbf{s}}$$

Similarly we can obtain other ranking vectors with right, bottom, top boundary superpixels, denoting as $\mathbf{s}_r$, $\mathbf{s}_b$, $\mathbf{s}_t$, respectively, and the final saliency map $\mathbf{s}_{bq}$ with background queries is computed as follows:

$$\mathbf{s}_{bq}=\mathbf{s}_l\circ\mathbf{s}_r\circ\mathbf{s}_b\circ\mathbf{s}_t,$$

where $\circ$ denotes the element-wise product operation.

Ranking with Foreground Queries. After obtaining $\mathbf{s}_{bq}$, we first set an adaptive threshold to produce foreground seeds, and then run the M3S-NIR algorithm by ranking with foreground queries. At last, the final saliency map can be gained by normalizing the saliency map based on foreground seeds into the range of 0 to 1.

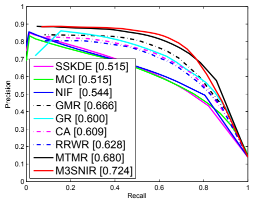

Figure 1. PR curves of our M3S-NIR against other methods with both RGB and thermal inputs.

IV. Experiments

A. Evaluation Data and Metrics

To fully evaluate the research presented in this paper, we conduct experiments on the RGB-T saliency detection benchmark dataset [1]. The benchmark dataset is large and challenging enough, which includes 821 pairs of spatially aligned RGB and thermal images and their ground truth annotations for saliency detection purpose. The image pairs in the dataset are recorded in approximately 60 scenes with different environmental conditions. In this work, PR curves and F-measure metric are used for quantitative comparison, whose definitions are in [17], The F-measure (F) is a combined metric of precision and recall.

B. Baseline Methods

For comprehensively validating the effectiveness of our approach, we qualitatively and quantitatively compare the proposed approach with 12 state-of-the-art approaches, including SSKDE [7], MCI [8],NIF [9], GMR [2], GR [10], CA [11], RRWR [12], MST [13], MDF [14], MCDL [15], MILPS [16], MTMR [1]. Comparing with above methods with RGB or thermal inputs, we could justify the effectiveness of complementary benefits from different modalities of our approach. To further demonstrate the importance of our fusion strategy, we also compare our approach with kinds of baseline methods with both RGB and thermal inputs, and the implementation details of multi-modal extension please refer to [1]. The extended methods include SSKDE, MCI, NIF, GMR, GR, CA and RRWR.

C. Experimental Settings

For fair comparison, we fix all parameters and other settings of our approach in the experiments. In graph construction, the number of superpixels on K different scales are $\{\pi,\pi_1,\pi_2,\pi_3\}=\{200,100,200,300\}$ with $K=3$. The graph edges are strengthened by the parameter γ, and we set it to 40. The proposed model involves four parameters, and we set them as: $\{\lambda,\lambda_1,\lambda_2,\lambda_3\}=\{0.01,0.6,0.02,0.01\}$.

D. Comparison Results

Overall performance. On the entire dataset, we first compare our M3S-NIR against other methods mentioned in Section IV-B on the aspects of precision (P), recall (R)

Table I

Average precision, recall, and F-measure of our method against different kinds of baseline methods on the dataset of [1], where the baselines are with RGB and thermal inputs. The code type and runtime (second) are also presented. The bold fonts of results indicate the best performance, and "M" and "P" are the abbreviations of MATLAB and Python.

Algorithm	RGB			Thermal			RGB-T			Code Type	Second
	P	R	F	P	R	F	P	R	F		
SS-KDE [7]	0.581	0.554	0.532	0.510	0.635	0.497	0.528	0.656	0.515	M&C++	0.94
MCI [8]	0.526	0.604	0.485	0.445	0.585	0.435	0.547	0.652	0.515	M&C++	21.89
NIF [9]	0.557	0.639	0.532	0.581	0.599	0.541	0.564	0.665	0.544	M	12.43
GMR [2]	0.644	0.603	0.587	**0.700**	0.574	0.603	0.733	0.653	0.666	M	1.11
GR [10]	0.621	0.582	0.534	0.639	0.544	0.545	0.705	0.593	0.600	M&C++	2.43
CA [11]	0.592	0.667	0.568	0.623	0.607	0.573	0.645	0.668	0.609	M	1.14
RRWR [12]	0.642	0.610	0.589	0.689	0.580	0.596	0.695	0.617	0.628	C++	2.99
MST [13]	0.627	0.739	0.61	0.665	0.655	0.598	-	-	-	C++	0.53
MDF [14]	0.692	0.699	0.654	0.631	0.585	0.549	-	-	-	M&C++	20.19
MCDL [15]	**0.701**	**0.751**	**0.689**	0.606	0.663	0.588	-	-	-	P&C++	2.41
MILPS [16]	0.637	0.691	0.612	0.643	**0.680**	**0.612**	-	-	-	M&C++	165.48
MTMR [1]	-	-	-	-	-	-	0.716	**0.713**	0.680	M &C++	1.39
M3S-NIR	-	-	-	-	-	-	**0.785**	0.671	**0.724**	M&C++	1.50

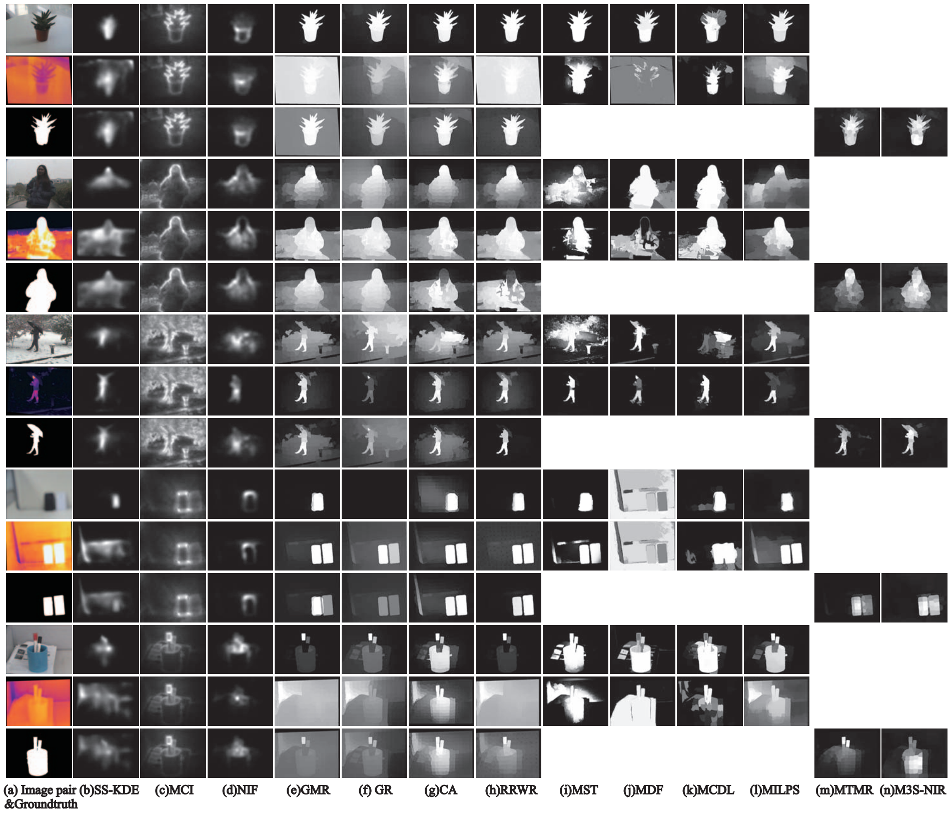

Figure 3. Sample results of the proposed approach and other baseline methods with different modality inputs. (a) Input RGB and thermal image pair and their ground truth. (b-h) Results of the baseline methods with RGB, thermal and RGB-T inputs. (i-l) Results of the baseline methods with RGB and thermal inputs. (m) and (n) Results of RGB-T inputs, including MTMR and our M3S-NIR.

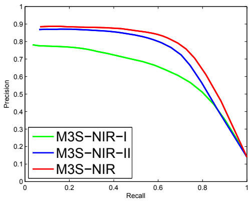

Figure 2. Evaluation results of the proposed approach with its variants on the entire dataset.

and F-measure (F), shown in Fig. 1 and Table I. From the quantitative evaluation results, we can observe that the M3S-NIR algorithm implements a good balance of precision and recall, then obtain better F-measure value over all baseline methods with a clear margin. It demonstrates that our M3S-NIR can detect the salient objects more accurately than other methods. The visual comparison are shown in Fig. 3. We notice that our M3S-NIR obtains weaker performance in recall comparing with deep learning based methods, i.e., MDF and MCDL. However, M3S-NIR has the following advantages over deep learning based methods. i) It does not require laborious pre-training or a large training set. ii) It does not need to save a large pre-trained deep model. iii) It is easy to implement as each subproblem of the proposed model has a closed-form solution. iv) It performs favorably against MDF and MCDL in terms of efficiency on a cheaper hardware setup.

Component Analysis. To justify the significance of the main components of the proposed approach, we implement two special versions for comparative analysis, they are: i) M3S-NIR-I, that removes noise-insensitive terms in our ranking model, and ii) M3S-NIR-II, that removes multiple scales in our ranking model. The results are presented in Fig. 2, through observations, we can see that our method substantially outperforms M3S-NIR-I. This demonstrates the significance of the introduced optimiztion algorithm that infers optimal background seeds or queries. The complete algorithm achieves superior performance than M3S-NIR-II, validating the effectiveness of multi-scale information integration in RGB-T saliency detection.

V. Conclusion

In this paper, we propose a novel and general algorithm for RGB-T saliency detection. We performs multi-modal multi-scale manifold ranking to achieve saliency calculation, in which the cross-modal and cross-scale collaborations are both considered to integrate both information from RGB image and Thermal image at multiple scales. We also introduce an intermediate variable to infer the optimal ranking seeds, to handle noises and/or corruptions caused by registration errors or cross-boundary of salient objects. And we present a unified ADMM-based optimization framework to solve the ranking model efficiently. In the future, we will improve the robustness of our approach by studying other prior models and graph construction.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (No. 61602006,61702002). This work was supported in part by the Anhui Key Research and Development Plan under Grant (No 1804a09020101), and also supported by Guangdong Province Science and Technology Plan projects (No. 2017B010110011).

References

[1] C. Li, G. Wang, Y. Ma, A. Zheng, B. Luo, and J. Tang, “A unified rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach,” 2017.

[2] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang, “Saliency detection via graph-based manifold ranking,” in *Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[4] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, “Slic superpixels compared to state-of-the-art superpixel methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[5] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[6] Q. Wang, W. Zheng, and R. Piramuthu, “Grab: Visual saliency via novel graph model and background priors,” in *Computer Vision and Pattern Recognition*, 2016, pp. 535–543.

[7] H. R. Tavakoli, E. Rahtu, and J. Heikkilä, “Fast and efficient saliency detection using sparse sampling and kernel density estimation,” in *Proceedings of the Scandinavian Conference on Image Analysis*, 2011.

[8] S. Goferman, L. Zelnik-Manor, and A. Tal, “Context-aware saliency detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[9] E. Erdem and A. Erdem, “Visual saliency estimation by nonlinearly integrating features using region covariances,” *Journal of vision*, vol. 13, no. 4, 2013.

[10] C. Yang, L. Zhang, and H. Lu, “Graph-regularized saliency detection with convex-hull-based center prior,” *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 637–640, 2013.

[11] Y. Qin, H. Lu, Y. Xu, and H. Wang, “Saliency detection via cellular automata,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[12] C. Li, Y. Yuan, W. Cai, and Y. Xia, “Robust saliency detection via regularized random walks ranking,” pp. 2710–2717, 2015.

[13] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, “Real-time salient object detection with a minimum spanning tree,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[14] G. Li and Y. Yu, “Visual saliency based on multiscale deep features,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463.

[15] R. Zhao, W. Ouyang, H. Li, and X. Wang, “Saliency detection by multi-context deep learning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[16] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, “Salient object detection via multiple instance learning,” *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1911–1922, 2017.

[17] G. Hripcsak and A. S. Rothschild, “Agreement, the f-measure, and reliability in information retrieval,” *J Am Med Inform Assoc*, vol. 12, no. 3, pp. 296–298, 2005.