Int. Conf. on Robotic Sensor Networks (ROSENET), KITAKYUSHU, JAPAN, August 2018

Weighted Linear Multiple Kernel Learning for Saliency Detection

Quan Zhou[1,2], Jinwen Wu[3], Yawen Fan[1], Suofei Zhang[4], Xiaofu Wu[1], Baoyu Zheng[1], Xin Jin[5], Huimin Lu[6], and Longin Jan Latecki[7]

[1] National Engineering Research Center of Communications and Networking, Nanjing University of Posts & Telecommunications, Nanjing, P. R. China.
[2] State Key Lab. for Novel Software Technology, Nanjing University, P.R. China.
[3] College of Information Engineering, China University of Geosciences(Wuhan), P.R. China.
[4] School of Internet of Things, Nanjing University of Posts & Telecommunications, Nanjing, P.R. China.
[5] Department of Computer Science and Technology, Beijing Electronic Science and Technology Institute, Beijing, P. R. China.
[6] Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Kitakyushu, Japan.
[7] Department of Computer and Information Sciences, Temple University, Philadelphia, USA.

Abstract. This paper presents a novel saliency detection method based on weighted linear multiple kernel learning (WLMKL), which is able to adaptively combine different contrast measurements in a supervised manner. Three common-used bottom-up visual saliency operations are first introduced, including corner-surround contrast (CSC), center-surround contrast (CESC) and global contrast (GC). Then these contrast measures are fed into our WLMKL framework to produce the final saliency map. We show that the assigned weights for each contrast feature maps are always normalized in our WLMKL formulation. In addition, the proposed approach benefits from the advantages of the contribution of each individual contrast operation, and thus produces more robust and accurate saliency maps. The extensive experimental results show the effectiveness of the proposed model, and demonstrate the combination is superior than individual subcomponent.

Keywords: saliency detection · corner-surround contrast · center-surround contrast · global contrast · Multiple kernel learning.

1 Introduction

The human visual system (HVS) has an outstanding ability to quickly locate the most interesting parts in a given scene. Such image parts are considered as salient since it is assumed these parts attract greater attention than other parts by the HVS. The recent study of saliency approaches may reveal the attention mechanisms of visual biology to predict human fixation selection behavior.

Saliency detection plays a significant role in the fields of computer vision, and is involved in many visual applications, such as automatic image cropping [5], image thumbnailing [19], image/video compression [21], image segmentation [15], image quality assessment [17], and object detection/recognition [2].

The recent years have witnessed great progress in saliency detection, and it has received extensive attention by the researcher in the fields of psychologists and computer vision [5, 12, 6, 26, 30, 31]. As a pioneer work, Treisman [24] proposed a feature integration theory (FIT) which is composed by three main steps for HVS: (1) the bottom-up contrast computation based on simple low-level image stimuli signals, such as luminance, color, texture and orientation, which are driven from the input image [12]; (2) the integration process via fusing various bottom-up feature maps produced in first step[11, 8]; (3) the enhanced highlighting salient parts with the assistance of top-down priors if available[22, 3]. In spite of achieving promising results, these approaches are still suffered from the following limitations: The existing methods for feature map integration, such as average operation [12], selective fusing operation [8], max or min operation [28], are not flexible enough and adaptive sufficiently. They are not able to assign adaptive weights to predict visual saliency, which reflect the confidence level of each individual feature map.

This paper attempts to solve this problem using weighted linear multiple kernel learning (WLMKL) framework for the task of saliency detection. More specifically, our method firstly utilizes corner-surround contrast (CSC) [29] to measure the saliency for each pixel. Except computing the appearance contrast, this contrast operation also considers the relative location between center and surrounding regions, which enables us to predict more exact location of the salient parts. Thereafter, two types of common-used contrast measurement, named CESC [12, 28] and global contrast (GC) [5], are calculated as complementary feature maps. Finally, to further investigate the contribution of each feature map, a multi-cues integration framework is designed using our WLMKL scheme to predict visual saliency. To optimize our WLMKL model, we design an EM-like procedure to alternatively update the model parameters and combined feature weights, where a closed-form solution can be obtained for updating feature weights. In summary, the main contributions of this paper are mainly summarized as follows:

- We propose to use WLMKL paradigm to formulate visual saliency, motivated by assigning adaptive weights to integrate different feature maps. Due to the duality, an efficient algorithm is designed to solve our WLMKL problem with ℓ_2-norm regularization. The proposed model benefits from the advantages of each feature map, while keeping the assigned weights are always normalized.
- We evaluate our approach for the tasks of visual saliency detection. We compared our model with the mainstream models in terms of detection accuracy. The experimental results show that our method outperforms these top-ranked models that previous studies have shown to be significantly predictive of salient parts in natural scenes.

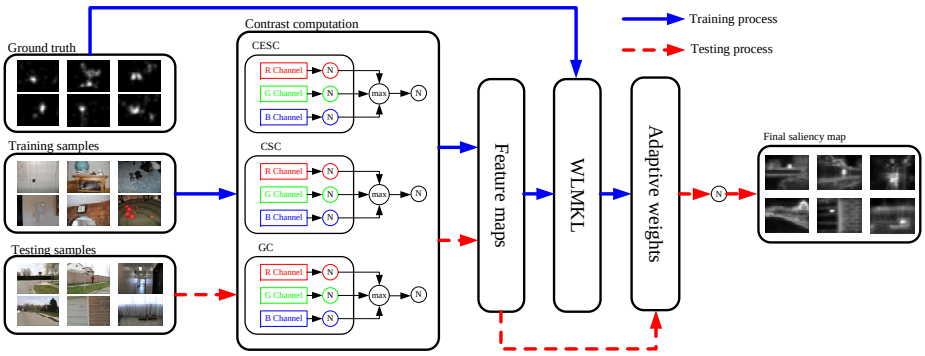

Fig. 1. Diagram of our saliency detection approach. In each channel of RGB color space, the CSC and CESC feature maps, which are the dissimilarity between a patch and its surroundings, and global feature map, which is based on rarity of an image patch with respect to the entire scene, are first computed and normalized. Then, the output feature maps and corresponding ground truth are used to train our WLMKL model. Finally, the feature maps of testing image are fed into the trained WLMKL model to generate final saliency map. The blue arrow denotes the training process and the dash red arrow indicates the testing process. (Best viewed in color)

This paper is organized as follows. We first elaborate on the detail of the proposed visual saliency detection method in Section 2. Section 3 reports the experimental results, and the conclusive remarks are given in Section 4.

2 The Proposed Method

The diagram of our approach is shown in Figure 1. The training and testing images undergo the same procedure of feature map computation and combination. We first introduce the image representation using sparse coding technique, and then present the aforementioned contrast calculation and WLMKL framework.

2.1 Image representation

From the perspective of human vision, a vision system should be adapted to the visual environment. As a supporting evidence for this theory, it has been shown that some neurons in V1 cortex resemble receptive fields that are learned via sparse coding algorithms [20]. From the perspective of computer vision, natural images are always with redundant structure, and thus can be sparsely represented by a set of localized and oriented filters [23]. To this end, we employ sparse coding technique to represent image patches, which has been demonstrated as an effective tool for saliency detection task [4, 10].

Preciously, the input image is first resized to $2^9 \times 2^9$ pixels. Suppose we have a series of image patches from the top-left to bottom-right of image with no

overlap. Then the reconstructive coefficients $\boldsymbol{\eta}_i$ are calculated to represent patch $\mathbf{p}_i$ using the sparse coding algorithms [20]. In our implementation, we extracted 500000 8×8 image patches randomly selected from color images using CIELab color space. Thus each basis in the dictionary is a $8 \times 8 = 64D$ vector, and we learn 200 dictionary functions. The sparse coding coefficients $\boldsymbol{\eta}_i$ are computed with the above learned basis using the LARS algorithm [18] implemented in the SPAMS toolbox [8]. Immediately below, we elaborate on the details of each blocks in Figure 1.

2.2 Computing contrast feature maps

CESC Inspired from the well-established computational architecture of [12, 28], the CESC operation, denoted as $f_{ce}^{c}(\mathbf{p}_i)$ in our model, is defined as the average weighted dissimilarity between a center patch $\mathbf{p}_i$ and its surrounding M neighborhood patches:

$$f_{ce}^{c}(\mathbf{p}_i) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{W}_{im}^{-1} \mathbf{B}_{im} \tag{1}$$

where $\mathbf{W}_{im}$ is the Euclidean distance between the location of center patch $\mathbf{p}_i$ and the surround patch $\mathbf{p}_m$. $\mathbf{B}_{im}$ denotes the Euclidean distance between $\boldsymbol{\eta}_i$ and $\boldsymbol{\eta}_m$ in the feature space, vectors of coefficients for $\mathbf{p}_i$ and $\mathbf{p}_m$, respectively, where the Euclidean distance (ℓ_2 distance) is employed as distance measure. Superscript c denotes sub channels in RGB color space.

CSC It often happens that CESC may assign high saliency value to background, leading to incorrect detections. In order to overcome this shortcoming, we employ CSC operation $f_{c}^{c}(\mathbf{p}_i)$ to estimate visual saliency not only investigating the appearance difference but also relative location between center patch and its surrounding neighborhoods [29]. According to [29], four types of local contrast, namely bottom-right, bottom-left, top-right, and top-left templates, are defined. Let $f_{br}^{c}(\mathbf{p}_i)$, $f_{bl}^{c}(\mathbf{p}_i)$, $f_{tr}^{c}(\mathbf{p}_i)$ and $f_{tl}^{c}(\mathbf{p}_i)$ denote these four types of local contrast, respectively, the final CSC feature map is calculated as:

$$f_{c}^{c}(\mathbf{p}_i) = f_{br}^{c}(\mathbf{p}_i) \times f_{bl}^{c}(\mathbf{p}_i) \times f_{tr}^{c}(\mathbf{p}_i) \times f_{tl}^{c}(\mathbf{p}_i) \tag{2}$$

For one specific type of local contrast (e.g., $f_{br}^{c}(\mathbf{p}_i)$), we define it as well as CESC computation that encodes the discriminative difference between corner patch $\mathbf{p}_i$ and its surrounding region. The same operation then applies to $f_{bl}^{c}(\mathbf{p}_i)$, $f_{tr}^{c}(\mathbf{p}_i)$ and $f_{tl}^{c}(\mathbf{p}_i)$, separately.

From Eqn. (2), it is evidence that CSC assigns high value to patch $\mathbf{p}_i$ only when it is recommended by four type of local contrast simultaneously. Thus CSC is a more strict contrast operation than CESC, resulting in more effective to exclude outliers and inhibit background.

[8] http://www.di.ens.fr/willow/SPAMS/index.html

GC Sometimes, only using the local contrast operation may suppress areas within a homogeneous region, resulting in the uniformly highlighted salient regions and overemphasized object boundaries. Although the appearance cues of local patch may be similar to its neighbors, they are globally rareness with respect to the entire scene. To this end, we construct our global contrast feature map $f_g^c(\mathbf{p}_i)$ guided from the information-theoretic measure [4]. Instead of computing pixel saliency, here we calculate the inverse of probability $p(\mathbf{p}_i)$ for each patch over the entire scene as the global feature map:

$$f_g^c(\mathbf{p}_i) \propto -\sum_{j=1}^{n} \log(p(\eta_{ij})) \tag{3}$$

where η_{ij} is the j^{th} component of vector $\boldsymbol{\eta}_i$. The GC assumes that coefficients η_{ij} are conditionally independent from each other, which is to some extent guaranteed by the sparse coding algorithm [20]. To construct the probability density function ($p(\eta_{ij})$), we first calculate histogram distribution with B bins for each component η_{ij} among all image patches in the scene, then the distribution is normalized by dividing to its sum. If a patch is rare in one of the features, the product in Eqn. (3) will get a small value leading to high global contrast value for that patch overall.

2.3 WLMKL framework

In this section, we design a WLMKL framework to estimate visual saliency, where the model parameters and adaptive feature weights are learned simultaneously. From the perspective of WLMKL, the feature maps are explicitly encoded through a set of so-called basic kernel functions $\{k_m\}_{m=1}^{M}$, then a SVM objective is employed to optimal model parameters and kernel feature weights.

Saliency formulation Considering a training set Ω containing N samples, where each sample is characterized by M kinds of feature descriptors. Let $\Omega = \{(\mathbf{F}_n(\mathbf{x}), y_n \in \pm 1)\}_{n=1}^{N}$, $\mathbf{F}_n(\mathbf{x}) = \{f_{n,m}(\mathbf{x})\}_{m=1}^{M}$, where $f_{n,m}(\mathbf{x})$ is the feature map value for pixel $\mathbf{x}$, y_n is the target label for pattern $\mathbf{F}_n(\mathbf{x})$, where $+1$ denotes the pixel $\mathbf{x}$ is salient, and -1 indicates not. We use $\mathbf{F}_n$ to represent $\mathbf{F}_n(\mathbf{x})$ and $f_{n,m}$ to represent $f_{n,m}(\mathbf{x})$ for notation simplicity, then the saliency formulation for pixel $\mathbf{x}$ is defined as follows:

$$s(\mathbf{x}) = \sum_{n=1}^{N} \alpha_n y_n \mathbf{K}(\mathbf{F}_n, \mathbf{F}) + b \tag{4}$$

where α_n and b are model coefficients required to be learned from training samples, while $\mathbf{K}(\cdot, \cdot)$ represents a given ensemble kernel functions which are symmetric and positive definite. Our formulation considers that the kernel function

$\mathbf{K}(\mathbf{F}_n, \mathbf{F})$ is a convex combination of basic kernels:

$$\begin{aligned} \mathbf{K}(\mathbf{F}_n, \mathbf{F}) &= \sum_{m=1}^{M} \beta_m k_m(\mathbf{F}_n, \mathbf{F}) \\ \beta_m \geq 0, &\qquad \sum_{m=1}^{M} \beta_m = 1 \end{aligned} \tag{5}$$

where $M = 3$ since we employ three different kind of feature maps (CESC, CSC, and GC) to predict visual saliency, and β_m is the feature weight for the mth feature map. Keeping in mind that different feature maps may have different proportion of contribution for final saliency map, we thus constrain the feature weights in Eqn. (5) are nonnegative and always normalized. Substitute Eqn. (5) to Eqn. (4), we get our final discriminative saliency model as:

$$\begin{aligned} s(\mathbf{x}) &= \sum_{n=1}^{N} \alpha_n y_n \sum_{m=1}^{M} \beta_m k_m(\mathbf{F}_n, \mathbf{F}) + b \\ \beta_m \geq 0, &\qquad \sum_{m=1}^{M} \beta_m = 1 \end{aligned} \tag{6}$$

WLMKL primal learning problem Actually, Eqn. (6) corresponds to a standard support vector machine (SVM) formulation under MKL framework [7]. In order to identify the model parameters, we thus propose to address following convex optimal problem, which we refer to as our primal WLMKL problem:

$$\begin{aligned} \min_{b,\ \boldsymbol{\xi},\ \boldsymbol{\beta},\ \mathbf{K}} \quad & \frac{1}{2} \sum_{m} \frac{1}{\beta_m} ||k_m||^2 + C \sum_{n} \xi_n \\ s.t. \quad & y_n [\sum_{m} k_m(\mathbf{F}_n) + b] \geq 1 - \xi_n, \forall n \\ & \xi_n \geq 0, \forall n \\ & \beta_m \geq 0, \qquad \sum_{m} \beta_m = 1 \end{aligned} \tag{7}$$

where $\boldsymbol{\beta} = \{\beta_m\}_{m=1}^{M}$, $\boldsymbol{\xi} = \{\xi_n\}_{n=1}^{N}$ are slack variables, and C is the tradeoff parameter between training error and margin. It is clear that Eqn. (7) is a primal learning problem involved in a weighted ℓ_2-norm regularization, where β_m controls the shape of the objective function.

Since this primal formulation is convex and differentiable, it provides a simple derivation of the dual problem [25]. By simply setting zero to the derivatives of the Lagrangian function for Eqn. (7) with respect to the primal variables, we

derive the associated dual problem as follows:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{\beta}} J(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{1}{2} \sum_{n,n'} \alpha_n \alpha_{n'} y_n y_{n'} \sum_m \beta_m k_m(\mathbf{F}_n, \mathbf{F}_{n'}) + \sum_n \alpha_n \\ &s.t. \qquad \sum_n \alpha_n y_n = 0 \qquad 0 \le \alpha_n \le C \quad \forall n \\ &\qquad\qquad \beta_m \ge 0, \qquad \sum_m \beta_m = 1 \end{aligned} \tag{8}$$

where $\boldsymbol{\alpha} = \{\alpha_n\}_{n=1}^N$. Optimizing the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is one particular form of the proposed WLMKL problems. Our approach utilizes such optimization to yield more flexible feature integration for visual saliency estimation.

Optimization Directly optimizing Eqn. (8) is difficult, we thus resort to an iterative, EM-like strategy to alternately optimize $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, separately. In each iteration, one of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is optimized while the other is fixed, and then the roles of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are switched. The whole iterations are repeated until convergence is reached.

On optimizing $\boldsymbol{\alpha}$. Suppose we are given the optimized parameter $\boldsymbol{\beta}^*$, the optimization problem of Eqn. (8) becomes:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} J(\boldsymbol{\alpha}) = -\frac{1}{2} \sum_{n,n'} \alpha_n \alpha_{n'} y_n y_{n'} \sum_m \beta_m^* k_m(\mathbf{F}_n, \mathbf{F}_{n'}) + \sum_n \alpha_n \\ &s.t. \qquad \sum_n \alpha_n y_n = 0 \qquad 0 \le \alpha_n \le C \quad \forall n \end{aligned} \tag{9}$$

which is identified as the standard SVM dual formulation using the combined kernel $\mathbf{K}(\mathbf{F}_n, \mathbf{F}) = \sum_m \beta_m^* k_m(\mathbf{F}_n, \mathbf{F})$. Thus the objective value $J(\boldsymbol{\alpha})$ can be obtained by any SVM algorithm.

On optimizing $\boldsymbol{\beta}$. Suppose we are given the optimized parameter $\boldsymbol{\alpha}^*$, the optimization problem of Eqn. (8) becomes:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) = -\frac{1}{2} \sum_{n,n'} \alpha_n^* \alpha_{n'}^* y_n y_{n'} \sum_m \beta_m k_m(\mathbf{F}_n, \mathbf{F}_{n'}) + \sum_n \alpha_n^* \\ &s.t. \qquad \sum_n \alpha_n^* y_n = 0 \qquad 0 \le \alpha_n^* \le C \quad \forall n \\ &\qquad\qquad \beta_m \ge 0, \qquad \sum_m \beta_m = 1 \end{aligned} \tag{10}$$

which is actually a non-linear objective function with constraints over the simplex. With our positivity definition on the kernel functions, $J(\boldsymbol{\beta})$ is convex and differentiable. Thus we solve this problem using a reduced gradient method. By simple differentiation of the objective function of Eqn. (10) with respect to β_m, we have:

$$\nabla J = \frac{\partial J(\boldsymbol{\beta})}{\partial \beta_m} = -\frac{1}{2} \sum_{n,n'} \alpha_n^* \alpha_{n'}^* y_n y_{n'} k_m(\mathbf{F}_n, \mathbf{F}_{n'}) \quad \forall n \tag{11}$$

Algorithm 1: The training procedure of our algorithm

Input: Training data: $\mathbf{F}_1, \mathbf{F}_2, \cdots, \mathbf{F}_N$; Associated data label: $y_1, y_2, \cdots, y_N \in \{+1, -1\}$;
Initial kernel weights: $\boldsymbol{\beta} = \{\beta_1, \beta_2, \cdots, \beta_M\}$, where $\beta_m = \frac{1}{M}$, $m = \{1, \cdots, M\}$;
Initial temp weights: $\boldsymbol{T} = \mathbf{0}$;
Basic kernel: Gaussian kernek; Step size: γ;
Stopping parameters: ε;
Output: Model coefficients: $\boldsymbol{\alpha}$; Basic kernel weights (feature map weights): $\boldsymbol{\beta}$

```
for ||T − β||_2 ≥ ε do
    Save current β as T = β;
    E-step: Optimize α*
        Compute α* using a standard SVM solver with fixed β and
        k(F_n, F) = Σ_m β_m k_m(F_n, F);
    M-step: Optimize β*
        Compute descent direction ∇J for β using Eqn. (11);
        Update β* as β* ← β + γ∇J;
        Normalize β* to satisfy the equality constraint in Eqn. (10);
end
return α and β;
```

Once the gradient of $J(\boldsymbol{\beta})$ is computed, $\boldsymbol{\beta}$ is updated using a descent direction ∇J as $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \gamma \nabla J$, where γ is the step size. Recalling from Eqn. (10), the non-negative and normalized constraint are also required to be satisfied after $\boldsymbol{\beta}$ are updated.

The whole training process is shown in Algorithm 1, the procedure of our method requires an initial guess to $\boldsymbol{\beta}$ in the alternating optimization, where each entry of $\boldsymbol{\beta}$ is initialized with equal weights. The whole algorithm is terminated when a stopping criterion is achieved. Here a simple stopping criterion is adopted based on the variation of $\boldsymbol{\beta}$ between two consecutive iterative steps.

3 Experiments

This section first describes our implementation details and experimental setup. Then, we compare our method with state-of-the-art methods in the literature.

3.1 Experimental setting

Dataset To evaluate the performance of our method, we employ two widely used datasets, including TORONTO [4] and MIT [13]. The first dataset contains 120 color images with resolution of 511×681 pixels from indoor and outdoor environments. Images are presented at random to twenty human subjects for 3 seconds with 2 seconds of gray mask in between. For the second dataset, it is a larger dataset containing 1003 images (resolution from 405×1024 to 1024×1024 pixels) collected from Flicker and LabelMe datasets. There are 779 landscape and 228 portrait images. The ground truth saliency maps are generated using the

Table 1. Performance comparison of the baseline methods and our approach on two datasets in terms of AUC and sAUC.

	TORONTO [4]		MIT [13]	
Methods	AUC	sAUC	AUC	sAUC
IT [12]	0.739	0.627	0.725	0.614
US [14]	0.815	0.670	0.804	0.658
CS [16]	0.817	0.659	0.814	0.656
FT [1]	0.534	0.447	0.515	0.422
SR [9]	0.516	0.409	0.544	0.437
SUN [27]	0.670	0.505	0.722	0.609
CESC	0.691	0.671	0.677	0.603
GC	0.816	0.690	0.808	0.676
CSC	0.811	0.694	0.816	0.670
Ours	**0.827**	**0.702**	**0.843**	**0.697**

eye fixation data collected from fifteen human subjects, where each subject are asked to freely view images for 3 seconds with 1 seconds delay in between.

Baselines To show the advantages of our approach, we selected 6 state-of-the-art models as baselines, including spectral residual saliency (SR [9]), attention measure (IT [12]), unified saliency (US [14]), nature statistic saliency (SUN [27]), frequency-tuned saliency (FT [1]), and Co-bootstrapping saliency (CS [16]). Besides, we directly borrow three feature maps (CESC, CSC and GC) as baselines for comprehensive comparison.

Evaluation metrics We utilize receiver operating characteristic (ROC) curve to evaluate our system. Under this criteria, each predicted saliency map is thresholded to generate the final map. The pixels with larger saliency values than the threshold are identified as salient (positive samples), and the other pixels are considered as non-salient (negative samples) [4]. The ROC curve is plotted with the true positive rate against the false positive rate under varying threshold. After that, we also compute the area under ROC curve (AUC) score for direct comparison. As discussed in [27], however, there is always a center bias that our HVS always prefers to the center of an image. Therefore, we turn to the shuffled AUC (sAUC) score [27] as an alternative metric.

3.2 Results and analysis

Table 1 shows the performance comparison between the proposed WLMKL method and the baseline methods in terms of the AUC and sAUC. The corresponding ROC curves are illustrated in Figure 2. Results show that our WLMKL method outperforms the state-of-the-art approaches. From Table 1, our method

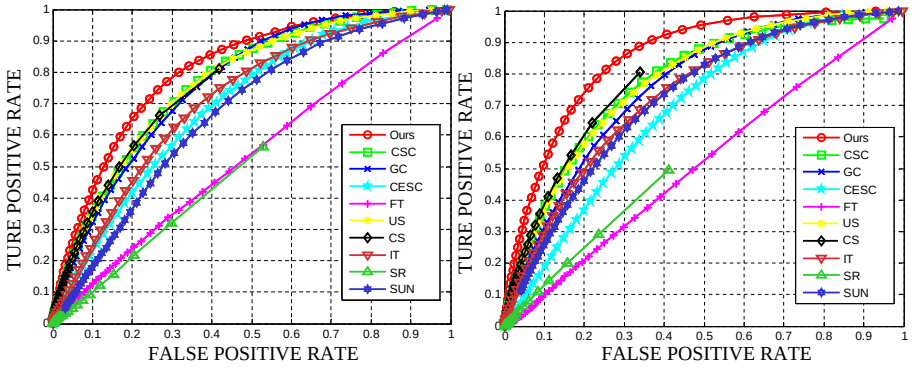

Fig. 2. ROC curve comparison between our method and other basline approaches. From left to right are the results on the TORONTO and MIT datasets. (Best viewed in color)

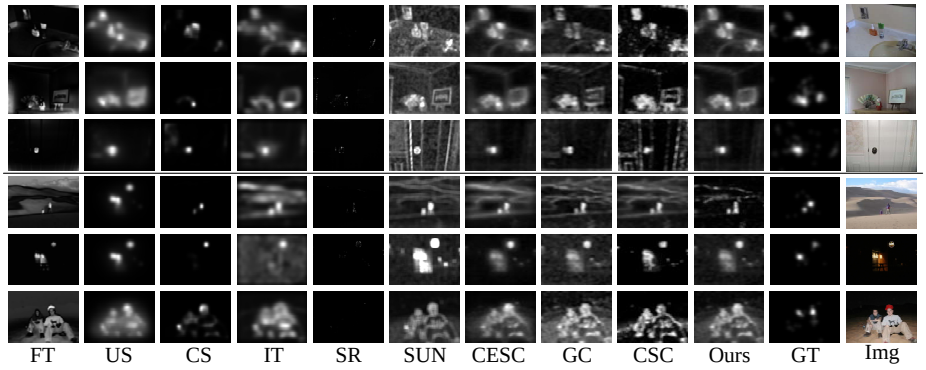

Fig. 3. Visual comparison between our method and other basline approaches. From top to bottom are some examples of predicted saliency maps on the TORONTO and MIT datasets. The columns from left to right, respectively, show estimated saliency maps produced by FT, US, CS, IT, SR, SUN, CESC, GC, CSC, and our methods, with corresponding ground truth and original images. (Best viewed in color)

averagely improves the results with 0.020 and 0.035 in terms of AUC and sAUC, and outperforms the best performing baselines with a margin of 0.031 and 0.043, respectively. We also show the promising results of each contrast feature map (CSC, GC, and CESC) over two datasets, especially the CSC almost gains higher performance than CESC, achieving comparable results with GC contrast measure. In particular, on the TORONTO dataset, CSC achieves an AUC of 0.827 and a sAUC value of 0.702, while on the MIT dataset, CSC achieves an AUC of 0.854 and a sAUC value of 0.697.

Some examples of the saliency maps produced from our WLMKL and the baseline methods are shown in Figure 3. One can observe that WLMKL pro-

duces saliency maps more consistent with the ground truth, compared with other baselines. These results clearly demonstrate the effectiveness of WLMKL in combining the contrast feature maps to perform visual saliency detection. It is worth noting that the proposed WLMKL does not require any preprocessing such as over-segmentation, nor any assistance from the top-down priors.

4 Conclusion

In this paper, a WLMKL framework is proposed for visual saliency detection. WLMKL learns adaptive weights to incorporate three contrast feature maps, namely, CSC, CESC and GC, respectively. Our WLMKL model enables each contrast feature map contributes to predict pixel saliency via preserving salient features and suppressesing the nonsalient features. Extensive experiments well validate the effectiveness of our framework on TORONTO and MIT benchmark datasets. In the future, we would like to explore more feature space (e.g., texture feature and edge strength) to further enhance performance.

Acknowledgment

This work was partly supported by the National Science Foundation (Grant No. IIS-1302164), the National Natural Science Foundation of China (Grant No. 61881240048, 61571240, 61501247, 61501259, 61671253), China Postdoctoral Science Foundation (Grant No. 2015M581841), Open Fund Project of Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education (Nanjing University of Science and Technology) (Grant No. JYB201709, JYB201710), and Natural Science Foundation of Jiangsu Province, China (BK20160908), NUPTSF (Grant No. NY214139).

References

1. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: CVPR. pp. 1597–1604 (2009)
2. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR. pp. 73–80 (2010)
3. Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: CVPR. pp. 438–445 (2012)
4. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: NIPS. pp. 155–162 (2006)
5. Cheng, M., Zhang, G., Mitra, N., Huang, X., Hu, S.: Global contrast based salient region detection. In: CVPR. pp. 409–416 (2011)
6. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. TPAMI **34**(10), 1915–1926 (2012)
7. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. JMLR **12**, 2211–2268 (2011)
8. Gopalakrishnan, V., Hu, Y., Rajan, D.: Salient region detection by modeling distributions of color and orientation. TMM **11**(5), 892–905 (2009)

9. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: CVPR. pp. 1–8 (2007)
10. Hou, X., Zhang, L.: Dynamic visual attention: Searching for coding length increments. In: NIPS. pp. 681–688 (2008)
11. Hu, Y., Xie, X., Ma, W.Y., Chia, L.T., Rajan, D.: Salient region detection using weighted feature maps based on the human visual attention model. In: Advances in Multimedia Information Processing-PCM, pp. 993–1000 (2005)
12. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. TPAMI **20**(11), 1254–1259 (1998)
13. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV. pp. 2106–2113 (2009)
14. Kruthiventi, S.S.S., Gudisa, V., Dholakiya, J.H., Babu, R.V.: Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In: CVPR. pp. 5781–5790 (2016)
15. Li, J., Li, X., Yang, B., Sun, X.: Segmentation-based image copy-move forgery detection scheme. IEEE TIFS **10**(3), 507–518 (2015)
16. Lu, H., Zhang, X., Qi, J., Tong, N., Ruan, X., Yang, M.H.: Co-bootstrapping saliency. IEEE TIP **26**(1), 414–425 (2017)
17. Ma, Q., Zhang, L.: Image quality assessment with visual attention. In: ICPR. pp. 1–4 (2008)
18. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. JMLR **11**, 19–60 (2010)
19. Marchesotti, L., Cifarelli, C., G., C.: A framework for visual saliency detection with applications to image thumbnailing. In: ICCV. pp. 2232–2239 (2009)
20. Olshausen, B.A., et al.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature **381**(6583), 607–609
21. Pan, Z., Zhang, Y., Kwong, S.: Efficient motion and disparity estimation optimization for low complexity multiview video coding. IEEE TB **61**(2), 166–176 (2015)
22. Shen, X., Wu, Y.: A unified approach to salient object detection via low rank matrix recovery. In: CVPR. pp. 853–860 (2012)
23. Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. Annual review of neuroscience **24**(1), 1193–1216 (2001)
24. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive psychology **12**(1), 97–136 (1980)
25. Vapnik, V.: The Nature of Statistical Learning Theory. Springer (1993)
26. Yu, J.G., Xia, G.S., Gao, C., Samal, A.: A computational model for object-based visual saliency: Spreading attention along gestalt cues. TMM **18**(2), 273–286 (2016)
27. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. Journal of vision **8**(7), 32 (2008)
28. Zhou, Q., Chen, J., Ren, S., Zhou, Y., Chen, J., , Liu, W.: On contrast combinations for visual saliency detection. In: ICIP. pp. 2665–2669 (2013)
29. Zhou, Q., Li, N., Yang, Y., Chen, P., Liu, W.: Corner-surround contrast for saliency detection. In: ICPR. pp. 1423–1426 (2012)
30. Zhou, Q.: Object-based attention: saliency detection using contrast via background prototypes. EL **50**(14), 997–999 (2014)
31. Zhou, Q., Cai, S., Zhu, S., Zheng, B.: Salient object detection using window mask transferring with multi-layer background contrast. In: ACCV. pp. 221–235 (2014)