Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2018, Article ID 1014595, 16 pages
https://doi.org/10.1155/2018/1014595

Research Article

Saliency Aggregation: Multifeature and Neighbor Based Salient Region Detection for Social Images

Ye Liang,[1,2] **Congyan Lang**,[1] **Jian Yu**,[1] **Hongzhe Liu**,[2] **and Nan Ma**[2]

[1]*School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China*
[2]*Beijing Union University, Beijing 100101, China*

Correspondence should be addressed to Ye Liang; liangye@buu.edu.cn

Received 30 June 2017; Revised 20 October 2017; Accepted 14 November 2017; Published 1 January 2018

Academic Editor: Anil Kumar

The popularity of social networks has brought the rapid growth of social images which have become an increasingly important image type. One of the most obvious attributes of social images is the tag. However, the sate-of-the-art methods fail to fully exploit the tag information for saliency detection. Thus this paper focuses on salient region detection of social images using both image appearance features and image tag cues. First, a deep convolution neural network is built, which considers both appearance features and tag features. Second, tag neighbor and appearance neighbor based saliency aggregation terms are added to the saliency model to enhance salient regions. The aggregation method is dependent on individual images and considers the performance gaps appropriately. Finally, we also have constructed a new large dataset of challenging social images and pixel-wise saliency annotations to promote further researches and evaluations of visual saliency models. Extensive experiments show that the proposed method performs well on not only the new dataset but also several state-of-the-art saliency datasets.

1. Introduction

Images and videos are two of the main ways for social entertainments and communications. With the popularity of photo sharing websites, social images have become an important type. The most obvious feature of social images is that they typically have several tags to describe the contents. How to use the tags for multimedia tasks, such as image indexing and retrieval [1, 2], has attracted increasing attention these days [3]. However, tags are seldom considered in state-of-the-art salient region detection models. Therefore, in this paper, we focus on salient region detection of social images using both appearance features and tag features.

With the development of saliency detection, a large number of saliency detection algorithms have been developed [4–6]. It has been found that only relying on low-level features cannot achieve satisfactory results. The researches have proved that the hierarchical and deep architectures [7–12] for salient region detection are very effective. Thus, a salient region detection method based on deep learning is proposed in this paper. In addition, various priors are also very important in salient region detection [13], for example, face [14–16], car [17], color [14], center bias [13], and objectness [18–20]. Intuitively, the tags could potentially be important high-level semantic cues for salient region detection [16, 21]. Thus, tags are incorporated into our salient region detection models.

It is observed that different methods perform differently in saliency analysis [22]. The performance of saliency varies with individual images. The problem also exists in deep feature based methods and handcrafted feature based methods. So handcrafted feature based detection methods can be considered as complementarities to deep feature based detection methods. However, the fusion process is without ground truth. It is nontrivial to determine which saliency map is better. The good saliency aggregation model should work on each individual image and be able to consider the performance gaps appropriately. Therefore, how to fuse saliency maps of different detection methods is a key issue to be solved in the paper.

The framework of salient region detection is shown in Figure 1. It includes two parts: deep learning based salient region detection and handcrafted feature based salient

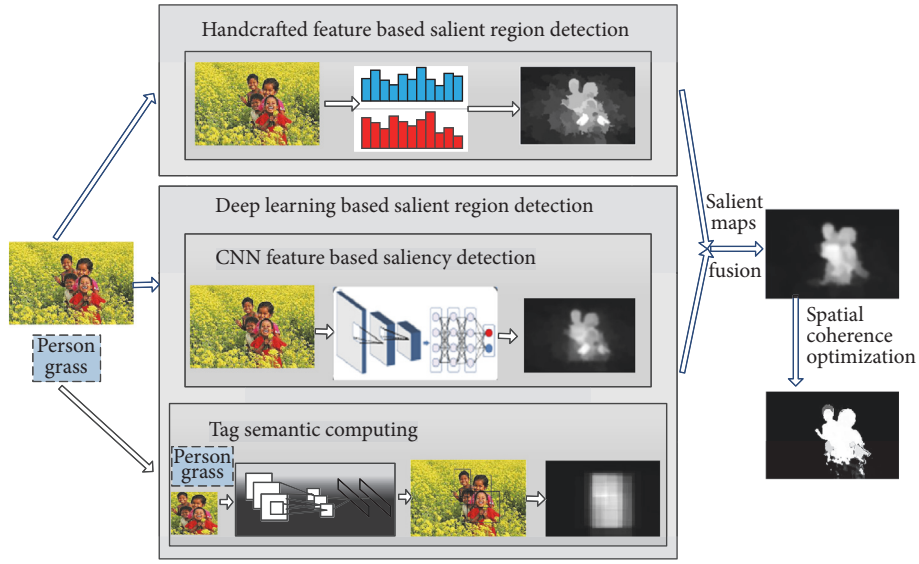

Figure 1: Framework.

region detection. Deep features include CNN (convolution neural network) features and tag features. Finally, the spatial coherence of saliency maps is optimized through the fully connected conditional random field model.

There are a variety of saliency detection benchmark datasets, either from saliency detection field [7, 8, 23–26] or from image segmentation field [27–29]. To promote further researches and evaluations on visual saliency detection for social images, it is necessary to construct a new dataset of social images.

The paper focuses on salient region detection of social images. The contributions of this paper are twofold. First, a deep learning based salient region detection method for social images is proposed, considering both appearance features and tag features. Second, tag neighbor and appearance neighbor based saliency aggregation method is proposed, which fuses state-of-the-art handcrafted feature based detection methods with our deep learning based detection method. The aggregation method is dependent on each specific individual image and considers the saliency performance gaps appropriately. So the detection model has fully taken advantage of image tags.

The rest of the paper is organized as follows. The deep learning based model is proposed in Section 2. Section 3 discusses the handcrafted feature based detection models. In Section 4, the saliency aggregation method is proposed. Spatial coherence optimization is discussed in Section 5. In Section 6, the new saliency dataset of social images is introduced. In Section 7, extensive experiments are performed and analyzed. Finally, conclusions are given in Section 8.

2. Deep Learning Based Salient Region Detection

Deep learning based salient region detection uses two types of features, appearance based CNN (convolution neural network) features and social image tag features. They are discussed in the following subsections.

2.1. CNN Based Salient Region Detection

2.1.1. Network Architecture. The deep network for appearance feature extraction has 8 layers [30] as shown in Figure 2. It includes 5 convolution layers, 2 fully connected layers, and 1 output layer. The bottom layer represents the input image and the adjacent upper layer represents the regions for deep feature extraction.

The convolution layers are responsible for the multiscale feature extraction. In order to achieve translation invariance, max pooling operation is performed after convolution operation. The learned feature is composed of 4096 elements. Fully connected layers are followed by ReLU (Rectified Linear Units) for nonlinear mapping. The dropout procedure is to avoid overfitting. ReLU performs the operation for each element in the following.

$$R\left(x^i\right) = \max\left(0, x^i\right), \tag{1}$$

where x is the feature of 4096 elements; if $x^i \geq 0$, then $\max(0, x^i) = x^i$; otherwise $\max(0, x^i) = 0$, $1 \leq i \leq 4096$.

The output layer uses softmax regression to calculate the probability of image patches being salient.

2.1.2. Multiscale CNN Feature Computation. In an image, salient regions have uniqueness, scarcity, and obvious difference with their neighborhoods. Inspired by literature [8], in order to effectively compute the saliency, three types of differences are computed, that is, the difference between the region and its neighborhoods, the difference between the region and the whole image, and the difference between the region and image boundaries. To compute these differences, four types of regions are extracted: (1) rectangle

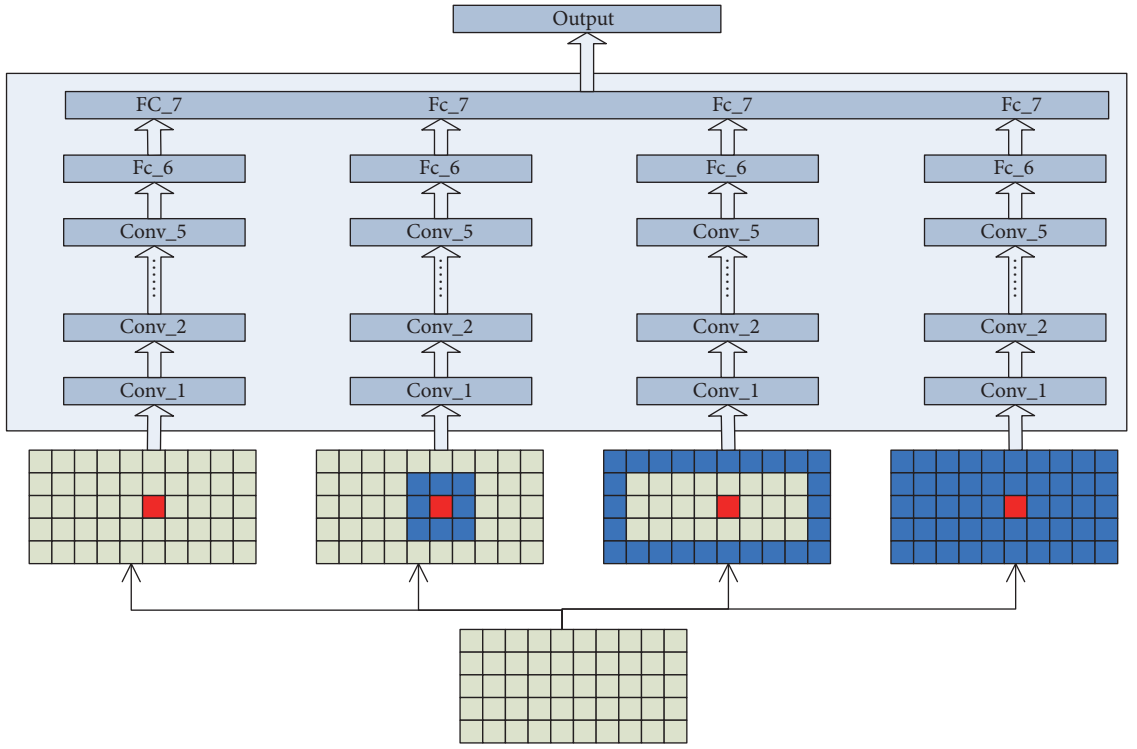

Figure 2: Architecture of network.

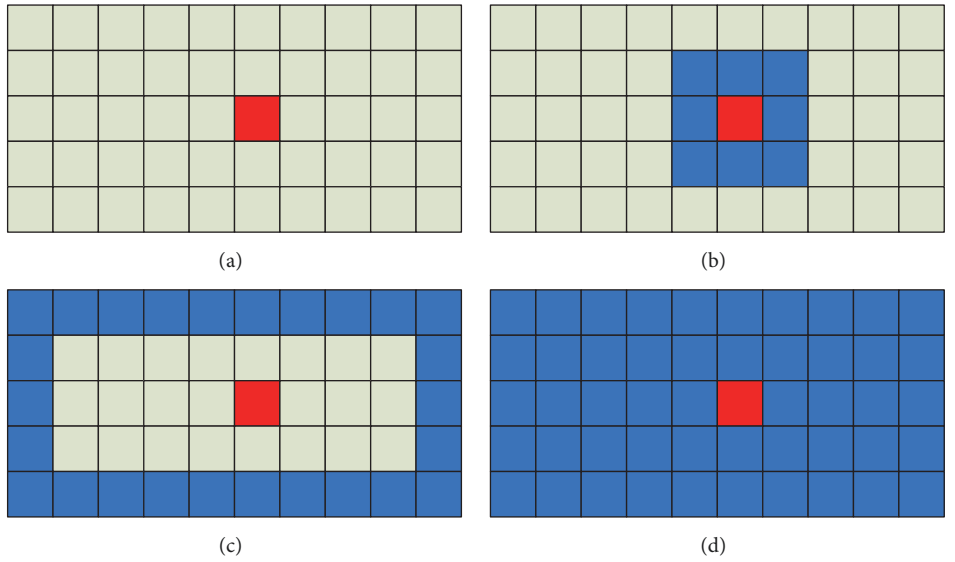

Figure 3: Four types of regions. (a) The red region denotes rectangle sample, (b) The blue regions denote neighborhoods of rectangle sample, (c) The blue regions denote boundaries of the image, (d) The blue regions denote image area except rectangle sample.

sample in a sliding window fashion; (2) neighborhoods of rectangle sample; (3) boundaries of the image; (4) image area except rectangle sample. Four types of regions are shown in Figure 3.

2.1.3. Training of CNN Network. Caffe [30], an open source framework, is used for CNN training and testing. The deep convolution neural network is originally trained on the ImageNet dataset. We extract multiscale features for each region and fine-tune the network parameters. For each image in the training set, we crop samples into 51×51 RGB patches in a sliding window fashion with a stride of 10 pixels. To label the sample patches, if more than 70% pixels in the example are salient, then this sample label is 1; otherwise it is 0. Using this annotation strategy, we obtain sample regions $\{B_i\}$ and corresponding labels $\{l_i\}$.

In fine-tuning process, the cost function is the softmax loss with weight decay given by

$$L(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=0}^{1} l\{l_i = j\}\log P(l_i = j \mid \theta) + \lambda\sum_{k=1}^{8}\|W_k\|_F^2, \quad (2)$$

where θ is the learnable parameter of convolution neural network, including the bias and weights of all layers; $l\{\cdot\}$ is the indicator function; $P(l_i = j \mid \theta)$ is the probability of the ith sample being salient; λ is the parameter of weight decay; W_k is the weight of the kth layer. We use stochastic gradient descent to train the network with batch size $m = 256$, $\lambda = 0.0005$. The initial learning rate is 0.01. When the cost is stabilized, the learning rate is decreased by a factor of 0.1. 80 epochs are repeated for the training process. The dropout rate is set to 0.5 to avoid overfitting.

2.2. Tag Semantic Feature Computation. Due to the fact that objects are closely related to salient regions, we use object tags to compute semantic features. The probability that a region is a particular object reflects the possibility being a salient region to some extent. Therefore, the probabilities that regions are specific objects can be regarded as priors.

RCNN (Regions with CNN) [31] is based on deep learning and has been widely used because of its excellent object detection accuracy. In the paper, RCNN is used to detect objects; thus tag semantics are transformed into RCNN features.

Suppose there are X object detectors. For the kth detector, the detection process is as follows.

(1) Select N proposals which are more likely to contain the specific object.

(2) Compute the ith proposal probability ${p_k}^i$ of the ith proposal being the kth object, $1 \le k \le X$, $1 \le i \le N$. At the same time, each pixel in the ith proposal also has the same probability ${p_k}^i$.

(3) For N proposals, each pixel has the score $\sum_{i=1}^{N} {p_k}^i * {f_k}^i$ being the kth object. If the pixel is contained by ith proposal, then ${f_k}^i = 1$, else ${f_k}^i = 0$.

X dimension feature is obtained for each pixel after X objects detector detection. X dimension feature is normalized as f, $f \in R^X$. Each dimension of f indicates probability being a specific object.

2.3. Fusion of CNN Based Saliency and Tag Semantic Features. Assume that the saliency map is S_D and RCNN based semantic features is T; the fusion is

$$S = S_D \cdot \exp(T). \quad (3)$$

Tags are priors and play weights in fusion. S represents the fused saliency map.

3. Handcrafted Feature Based Salient Region Detection

It is observed that different methods perform differently in saliency analysis [22]. Although the overall detection effect based on deep features is better than that based on handcrafted features, the differences still exist on individual images. So handcrafted feature based salient maps can be considered as complementarities to deep feature based saliency maps. In Figure 4, the first column shows the original social images; the second shows the ground truth masks; the third shows the salient maps of DRFI method [25] which is based on handcrafted features; the last represents the salient maps of MDF method [8], which are based on deep features. We can see that the last column includes incomplete parts, unclear boundaries, and false detections. So in the paper, some state-of-the-art salient region detection methods based on handcrafted features are selected as complementarities to our proposed deep detection method.

4. Saliency Aggregation

4.1. Main Idea. It is observed that if a salient region detection method has good effects on a social image, this method has great possibility to get sound effect on similar images. The main idea of aggregation is based on this assumption.

In training process, sort lists of all detection methods on all images can be achieved. Sort lists can be seen as priors in testing.

In testing process, we search *KNN (K nearest neighbors)* images similar to the test image in the training set. Moreover, sort lists of KNN images are known in the training stage. KNN images can vote for detection methods through sort lists. Thus, the test image is able to obtain its sort list based on voting. Salient map of test image can be computed by aggregating its salient maps of different methods using sort lists.

Training process and testing process are shown in Figures 5 and 6.

4.2. Training Process. Given an image I in the training set, its ground truth is given by G; its salient maps using different detection methods is denoted as $S = \{S_1, S_2, S_3, \ldots, S_i, \ldots, S_M\}$. In this saliency map set, M is the number of detection methods, and S_i is the salient map of the ith method.

For every detection method, its salient maps can be compared with ground truth G and yield AUC (Area under ROC Curve) values. The greater the AUC value, the better the saliency detection performance. After AUC value computation, sort lists of all methods can be obtained.

For convenience, it is assumed that there are four detection methods. Sort lists are shown in Figure 7. The data structure is single linked list. Data domain of header node denotes image and pointer domain of header node points to data node. Nonheader node includes three domains: the first domain is the AUC value, the second domain is the method index, and the last domain is a pointer.

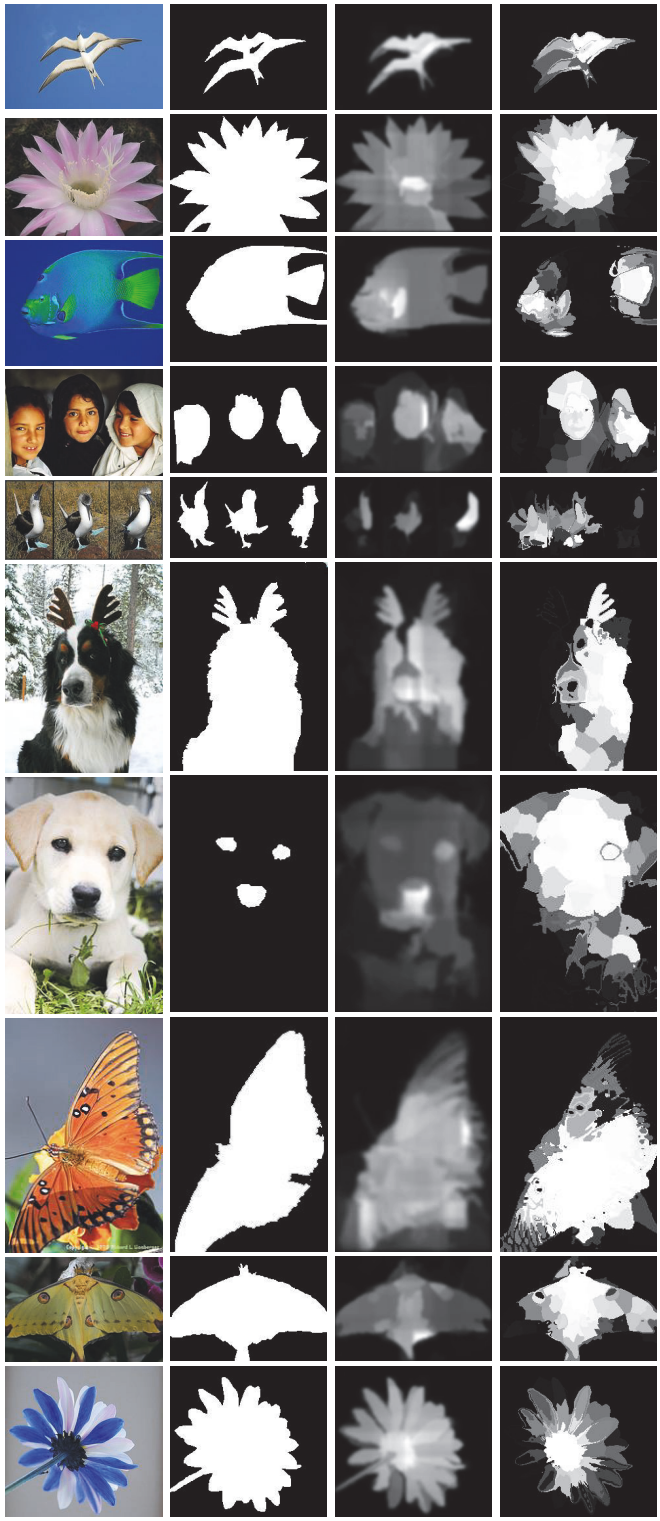

Figure 4: Examples of saliency detection results. Images in each column are original images, ground truth masks, salient maps of method DRFI [25], and salient maps of method MDF [8], respectively.

4.3. Testing Process. A social image has two parts: image and corresponding tags. In the testing set, image I and its tag set $T = \{t_1, t_2, \ldots, t_i, \ldots, t_N\}$ are given, where N is the number of tags. We search its neighbors through tag semantics and image appearance. Sort lists of neighbors can vote for salient maps of image I.

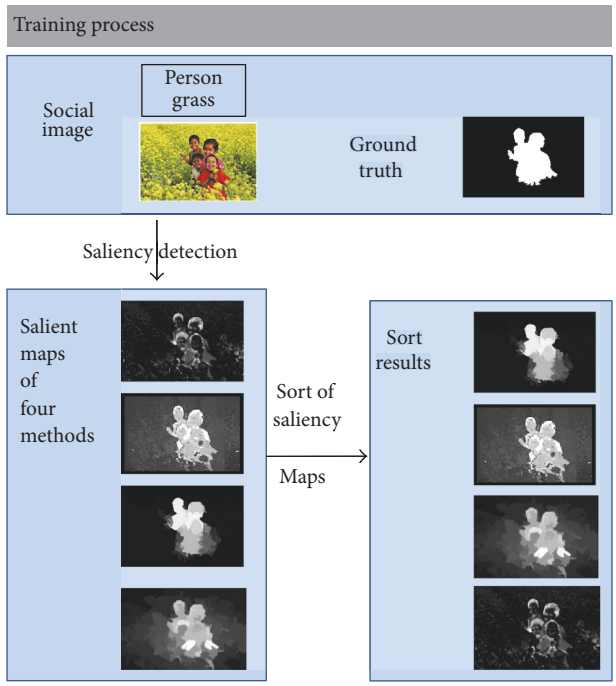

Figure 5: Training process.

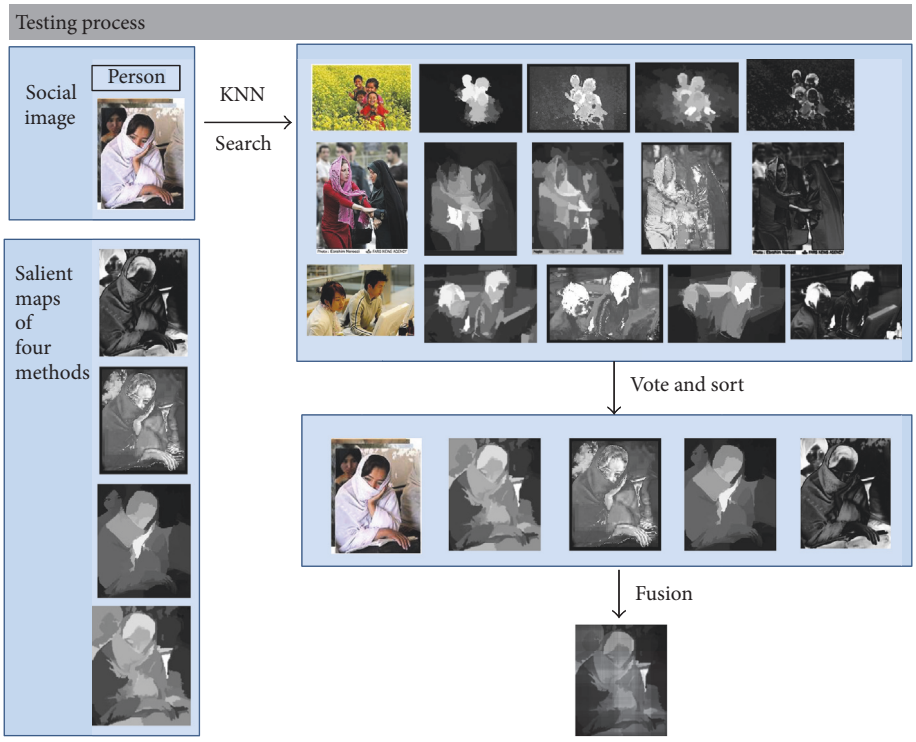

Figure 6: Testing process.

4.3.1. Tag Based Neighbor Search. There are two types of tags: object tags and scene tags. Because objects are closely related to salient regions, object tags are used in semantic search.

There are 37 object tags in the new dataset, including animal, bear, birds, cat, fox, zebra, horses, tiger, cow, dog, elk, fish, whale, vehicles, boats, cars, plane, train, person, police, military, tattoo, computer, coral, flowers, flags, tower, statue, sign, book, sun, leaf, sand, tree, food, rocks, and toy.

In these categories, animal has super class and subclass relationship with bear, birds, cat, fox, zebra, horses, tiger,

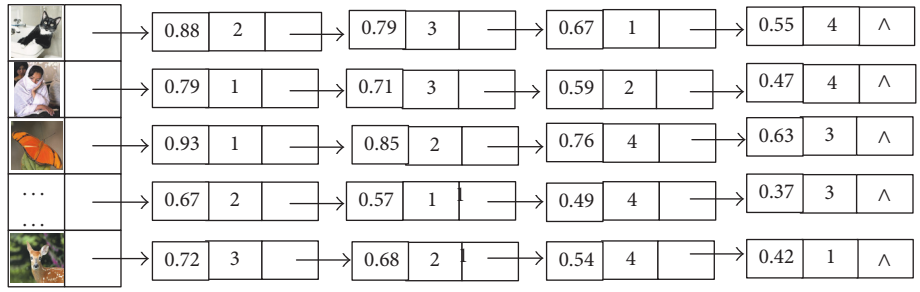

Figure 7: Images and their sort lists.

cow, dog, elk, fish, and whale; vehicles have super class and subclass relationship with boats, cars, plane, and train; person has super class and subclass relationship with police, military, and tattoo.

Although super class and subclass have great relevance in the class definition, many subclasses have a variety of differences in environment and appearance. So, for animal class, subclasses need exact matching to find neighbors; for vehicles class, subclasses need exact matching to find neighbors; because of particularity of class people, if there is no exact matching of subclass, matching can be performed at person level.

4.3.2. Appearance Based Neighbor Search. 256 dimensional histogram of RGB color space is used and χ^2 distance is computed.

4.4. Vote Based Saliency Maps Aggregation. Suppose the test image is I, the number of tag neighbors is k, and the number of appearance neighbors is k.

After tag based search in the training set, the detected neighbor number is y. If y is bigger than k, then k images are selected according to appearance similarities from y images. Finally, tag based neighbor set is given as

$$\operatorname{Im} g^T = \left\{\operatorname{Im} g_1^T, \operatorname{Im} g_2^T, \ldots, \operatorname{Im} g_i^T, \ldots, \operatorname{Im} g_x^T\right\}, \quad (4)$$

where x is the final number of neighbors; if $y \geq k$, then $x = k$; otherwise, $x = y$.

After appearance based similarity computation in the training set, k nearest neighbors are selected as

$$\operatorname{Im} g^A = \left\{\operatorname{Im} g_1^A, \operatorname{Im} g_2^A, \ldots, \operatorname{Im} g_i^A, \ldots, \operatorname{Im} g_k^A\right\}. \quad (5)$$

Merge sets (4) and (5) and get the set as

$$\operatorname{Im} g = \left\{\operatorname{Im} g_1, \operatorname{Im} g_2, \ldots, \operatorname{Im} g_x, \ldots, \operatorname{Im} g_{x+k}\right\}. \quad (6)$$

Each neighbor image has a sort list and contains the AUC values of all detection methods. The AUC values can vote for each detection method. Vote weights are summed as

$$\text{auc} = \left[\sum_{i=1}^{x+k} \text{auc}_i^1, \sum_{i=1}^{x+k} \text{auc}_i^2, \ldots, \sum_{i=1}^{x+k} \text{auc}_i^j, \ldots, \sum_{i=1}^{x+k} \text{auc}_i^M\right]. \quad (7)$$

In $\sum_{i=1}^{x+k} \text{auc}_i^j$, i is the ith neighbor and j is the jth detection method. M is the number of detection models.

The salient map set of image I is

$$S(I) = \left[S_1(I), S_2(I), \ldots, S_i(p), \ldots, S_M(I)\right], \quad (8)$$

where $S_j(I)$ is the saliency map of the jth detection method.

The fused saliency map can be computed as follows.

$$S_F(I) = S(I) \cdot \text{auc}^T. \quad (9)$$

5. Spatial Coherence Optimization

In saliency computations, the spatial relationship of adjacent regions is not considered, so it will result in noises on salient regions. In the field of image segmentation, the researchers use fully connected CRF (conditional random field) model [49] to achieve better segmentation results. Therefore, we use the fully connected CRF model to optimize the spatial coherence of saliency maps.

The objective function is defined as follows.

$$S(L) = -\sum_i \log P(l_i) + \sum_{i,j} \theta_{ij}(l_i, l_j), \quad (10)$$

where L is the binary variable being salient or not. $P(l_i)$ is the probability of pixel x_i being salient. Initially, $P(1) = S_i$, $P(0) = 1 - S_i$. S_i is the saliency of the pixel i.

$\theta_{i,j}$ is defined as follows.

$$\theta_{i,j} = u(l_i, l_j)\left[\omega_1 \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_1^2} - \frac{\|I_i - I_j\|^2}{2\sigma_2^2}\right) + \omega_2 \exp\left(-\frac{\|p_i - p_j\|^2}{2\sigma_3^2}\right)\right]. \quad (11)$$

If $l_i \neq l_j$, then $u(l_i, l_j) = 1$, or else 0.

Both position information and color information are considered in $\theta_{i,j}$.

p_i is the position of pixel i and p_j is the position of pixel j.

I_i is the color of pixel i and I_j is the color of pixel j.

$\omega_1 \exp(-\|p_i - p_j\|^2/2\sigma_1^2 - \|I_i - I_j\|^2/2\sigma_2^2)$ suggests that adjacent pixels with similar colors should have similar saliency. σ_1 and σ_2 control color similarity and distance proximity.

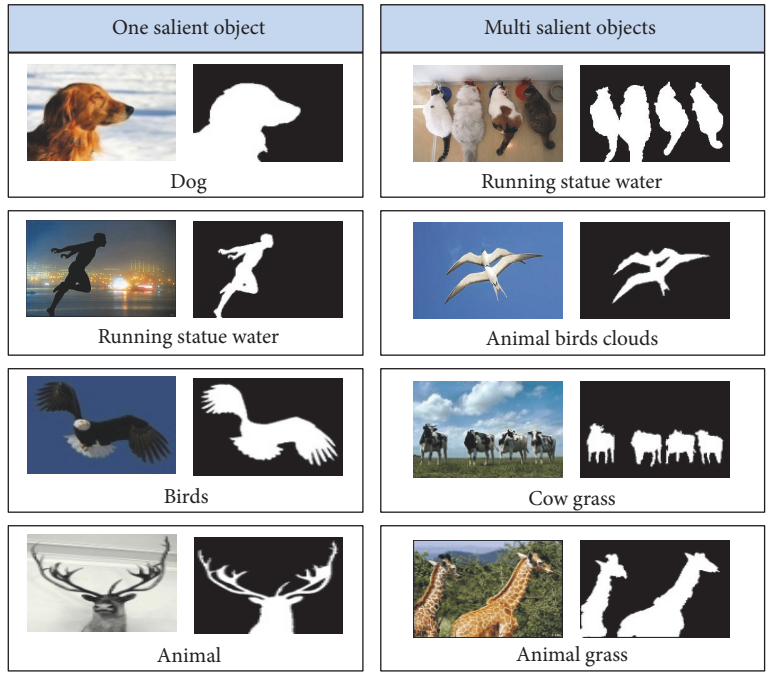

Figure 8: Images with one or multiple salient regions.

$\omega_2 \exp(-\|p_i - p_j\|^2/2\sigma_3^2)$ only considers position information. The purpose is to remove small areas.

6. Construction of Saliency Dataset of Social Images

The paper focuses on salient region detection of social images, so it is necessary to construct a new dataset of social images to promote further researches and evaluations of visual saliency models. The following will be discussed in detail.

6.1. Data Source. NUS-WIDE dataset [50] is a web image dataset constructed by NUS lab for media search. The images and the tags of this dataset are from Flickr which is a popular social web site. We randomly select 10000 images from NUS-WIDE dataset. The images come from thirty-eight folders of NUS-WIDE dataset, including carvings, castle, cat, cell phones, chairs, chrysanthemums, classroom, cliff, computers, cooling tower, coral, cordless cougar, courthouse, cow, coyote, dance dancing, deer, den, desert, detail, diver, dock, close-up, cloverleaf, cubs, doll, dog, dogs, fish, flag, eagle, elephant, elk, f-16, facade, and fawn.

6.2. Salient Region Annotation. Since the bounding boxes for salient regions are rough and can not reveal region boundaries, we adopt the pixel-wise annotation. In annotation process, nine subjects are asked to specify the attractive regions according to their first glance at the image.

To reduce label inconsistency of the annotation results, the pixel consistency score is computed. A pixel can be considered salient if 50% of subjects have selected it [23].

Finally, two subjects use Adobe Photoshop to segment salient regions.

6.3. Image Selection. First, 10000 images are randomly selected from NUS-wide dataset. Then, the images are further selected by the following criteria.

(1) The color contrast of any salient region and corresponding image is less than 0.7.

(2) Salient regions are rich in size. The proportion of salient regions to the corresponding image covers 10 grades, $[0, 0.1)$, $[0.1, 0.2)$, $[0.2, 0.3)$, $[0.3, 0.4)$, $[0.4, 0.5)$, $[0.5, 0.6)$, $[0.6, 0.7)$, $[0.7, 0.8)$, $[0.8, 0.9)$, $[0.9, 1]$.

(3) At least ten percent of the salient regions connected with the image boundaries.

After 5 rounds of selecting, the dataset contains 5429 images.

In the new dataset, the images have one or more salient regions; the positions of salient regions are not limited to image centers. The sizes of salient regions are varied. A great deal of images have complex/cluttered backgrounds. There are 78 tags which come from 81 tags of NUS-WIDE dataset. All these will bring challenges to salient region detection.

6.4. Typical Images of the New Dataset. In this section, typical examples of images, ground truth masks, and tags are listed below. Images can have one or multiple salient regions in Figure 8. The images may have cluttered and complex backgrounds in Figure 9. The sizes of salient regions are rich in Figure 10.

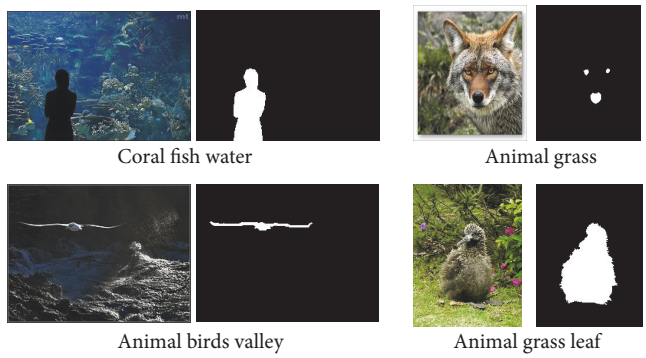

Figure 9: Images with cluttered and complex backgrounds.

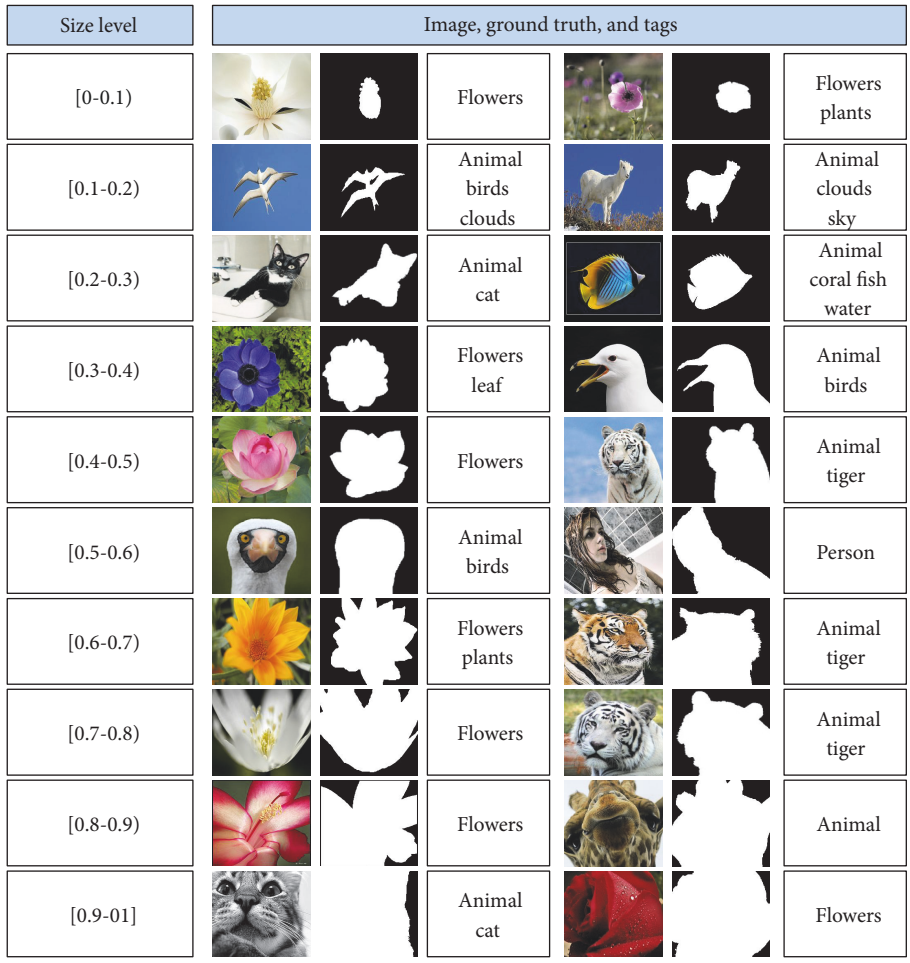

Figure 10: Images in various size levels.

7. Experiments

7.1. Experimental Setup

7.1.1. Experiments on the New Dataset. The aim of the paper is to solve salient region detection of social images. So the main experimental dataset is our new dataset, which is abbreviated as TBD (Tag Based Dataset).

We selected 20 object tags, including bear, birds, boats, buildings, cars, cat, computer, coral, cow, dog, elk, fish, flowers, fox, horses, person, plane, tiger, train, and zebra. Correspondingly, 20 RCNN object detectors were chosen to

extract RCNN features. Top 1000 proposals of each detector were used to compute RCNN features.

The proposed deep based detection method is abbreviated as DBS (Deep Based Saliency). DBS method was compared with 27 state-of-the-art methods in Section 7.2.1. 27 state-of-the-art methods are CB [34], FT [23], SEG [44], RC [14], SVO [17], LRR [39], SF [45], GS [37], CA [33], SS [47], HS [7], TD [48], MR [24], DRFI [25], PCA [41], HM [38], GC [36], MC [40], DSR [35], SBF [43], BD [42], SMD [46], BL [32], MCDL [9], MDF [8], LEGS [10], and RFCN [11]. These methods not only are very popular but also cover many types.

In addition, we also verify the performance of the aggregation method in Section 7.2.2.

7.1.2. Experiments on State-of-the-Art Datasets. We also carried out the experiments on six state-of-the-art datasets to validate our method. These datasets are MSRA1000 [23], DUT-OMRON [24], ECSSD [7], HKU-IS [8], PASCAL-S [51], and SOD [27]. In these datasets, SOD [27] is a dataset which is from segmentation field; others are from saliency field. Because these datasets have no image level tags, we extract objectness feature [19] of these datasets. Objectness is a kind of high-level semantic cues, so objectness cue is similar to tag feature. Compared with the method DBS, the method using objectness feature instead of tag feature is abbreviated as OBS (Objectness Based Saliency).

OBS method was compared with 11 state-of-the-art methods, including FT [23], RC [14], SF [45], HS [7], MR [24], DRFI [25], GC [36], MC [40], BD [42], MDF [8], and LEGS [10].

7.1.3. Evaluation Criteria. We adopted popular performance evaluations to quantitatively evaluate the results, including PR (Precision Recall) curves, ROC (Receiver Operating Characteristic) curves, F-measure value, AUC (Area under ROC Curve) value, and MAE (Mean Absolute Error) value, respectively.

7.2. Experiments on the New Dataset TBD

7.2.1. Experiments of Deep Learning Based Detection Method. DBS is compared with 27 state-of-the-art methods. The results are given in Table 1 and Figure 11.

Among the 28 methods in Table 1, the top four methods are all deep learning based methods, including MCDL [9], RFCN [11], MDF [8], and DBS. To some extent, deep learning based detection methods are better than handcrafted feature based methods, in terms of both completeness and accuracy of saliency maps. AUC value of DBS method is the highest. F-measure value of DBS method is slightly lower than RFCN [11]. MAE value of DBS is third low. The overall performance of DBS method is good.

Typical saliency maps are shown in Figure 11.

7.2.2. Experiments of Aggregation Method. The handcrafted feature based detection methods used as complementarities to DBS are DRFI [25], SMD [46], BL [32], and MC [40].

Table 1: F-measure, AUC, and MAE of DBS and 27 state-of-the-art methods.

	F-measure	AUC	MAE
CB	0.5472	0.7971	0.2662
SEG	0.4917	0.7588	0.3592
SVO	0.3498	0.8361	0.409
SF	0.3659	0.7541	0.2077
CA	0.5161	0.8287	0.2778
TD	0.5432	0.8081	0.2333
SS	0.2516	0.6714	0.2499
HS	0.5576	0.7883	0.2747
DRFI	0.5897	0.8623	0.2063
HM	0.4892	0.7945	0.2263
BD	0.5443	0.8185	0.1955
BL	0.5823	0.8562	0.266
MR	0.5084	0.7753	0.229
PCA	0.5392	0.8439	0.2778
FT	0.3559	0.6126	0.2808
RC	0.5307	0.8105	0.3128
LRR	0.5124	0.7956	0.3067
GS	0.5164	0.8136	0.2056
SMD	0.6033	0.8437	0.1976
GC	0.5063	0.7511	0.2596
DSR	0.5035	0.8139	0.2105
MC	0.574	0.8427	0.2287
SBF	0.493	0.848	0.2325
MCDL	0.6559	0.8813	0.1457
LEGS	0.6124	0.8193	0.1844
RFCN	0.6768	0.8803	0.1476
MDF	0.6574	0.8483	0.1556
DBS	0.6621	0.8917	0.1505

In neighbor searching, the number of tag neighbors is 4 and the number of appearance neighbors is 4.

In order to verify the effect of neighbors, appearance neighbor based method and tag neighbor based method are carried out, respectively. Appearance neighbor based aggregation method is abbreviated as ABS (Appearance Based Saliency). Tag neighbor based aggregation method is abbreviated as TBS (Tag Based Saliency). Tag neighbor and appearance neighbor based aggregation method is abbreviated as FBS (Fusion Based Saliency).

The detection performances of DBS, ABS, TBS, and FBS are compared in Table 2.

The performance of TBS is better than the performance of ABS. The reasons are as follows. ABS method is based on appearance feature based neighbor search. Appearance similar images cannot guarantee similar saliency maps. However, TBS method uses object information. The same or similar

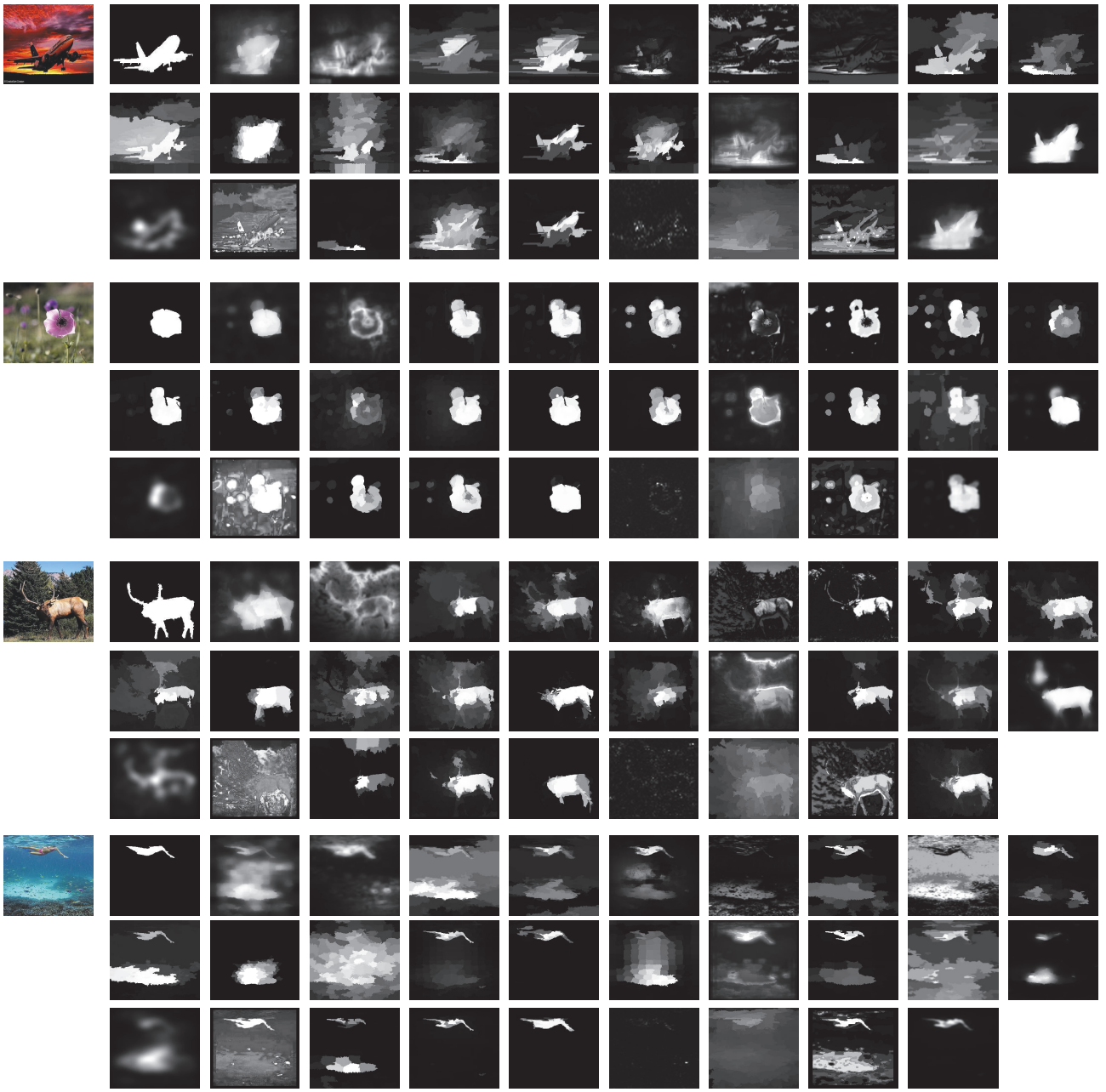

Figure 11: Visual comparisons of DBS with 27 state-of-the-art methods. The order of images are original image, ground truth mask, BL [32], CA [33], CB [34], DRFI [25], DSR [35], FT [23], GC [36], GS [37], HM [38], HS [7], LEGS [10], LRR [39], MC [40], MCDL [9], MR [24], PCA [41], BD [42], RC [14], RFCN [11], SBF [43], SEG [44], SF [45], SMD [46], MDF [8], SS [47], SVO [17], TD [48], and DBS.

objects can ensure similar salient regions to some extent. So the performance of TBS is better.

PR and ROC curves are shown in Figures 12 and 13. PR and ROC curves of FBS are higher than 27 state-of-the-art methods.

The examples of typical saliency maps of FBS method and DBS method are shown in Figure 14. It can be seen that the aggregation results are more complete and the details are better.

7.3. Experiments on State-of-the-Art Datasets. The experiment results are given in Table 3. We can see that AUC values of OBS are the highest on all datasets, F-measure values of OBS are the highest on all datasets, and MAE values are the lowest or the second lowest. The performance of OBS is the best. However, the improvements of OBS are not so obvious because objectness feature is not the accurate tag feature. Thus we believe that the results will be improved obviously if we use accurate tag annotation of images.

Table 2: F-measure, AUC, and MAE of DBS, ABS, TBS, and FBS.

	DBS	ABS	TBS	FBS
F-measure	0.6621	0.6652	0.6688	0.6712
AUC	0.8917	0.9061	0.9113	0.9166
MAE	0.1505	0.1497	0.1474	0.1452

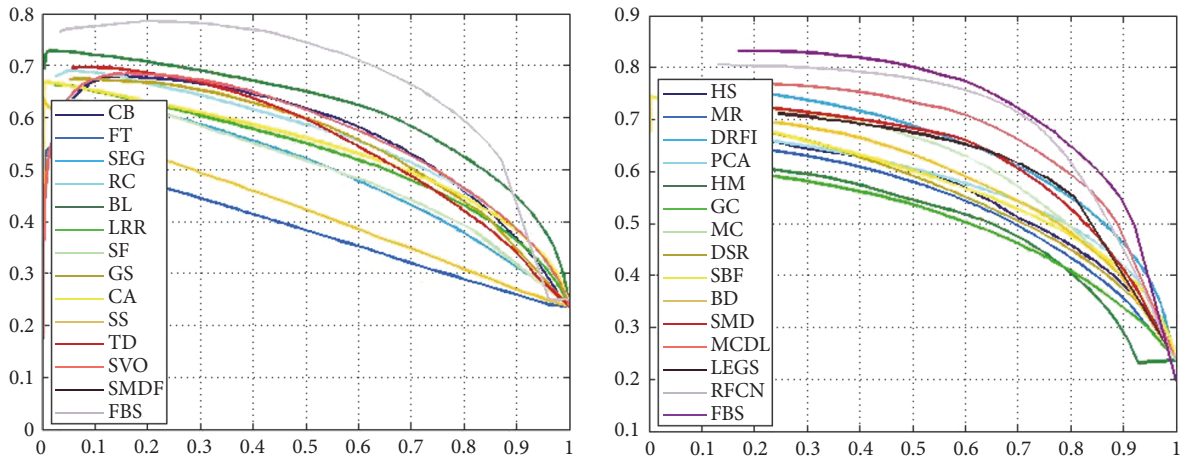

Figure 12: PR curves of FBS and 27 state-of-the-art methods.

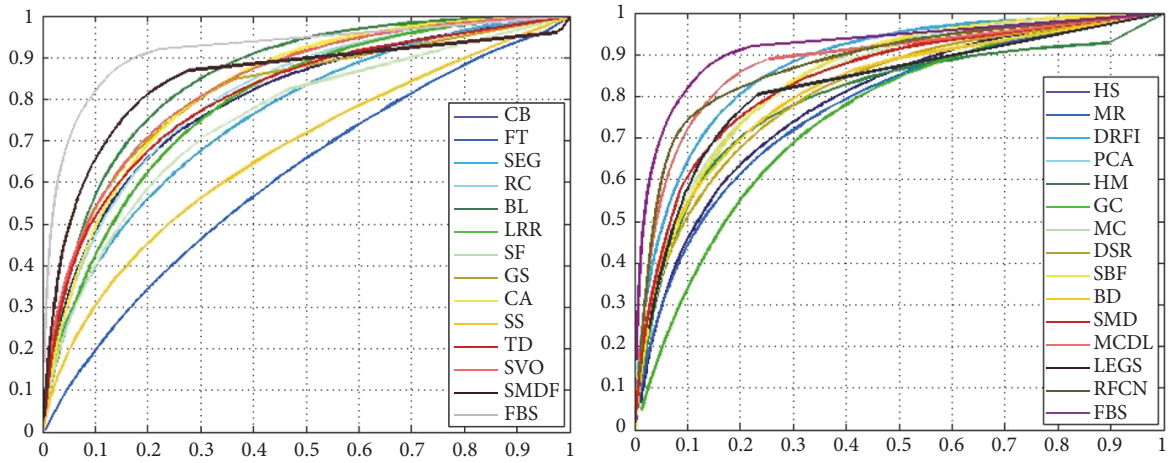

Figure 13: ROC curves of FBS and 27 state-of-the-art methods.

Experiments on state-of-the-art datasets validate the effectiveness of our proposed method DBS.

8. Conclusions

The paper focuses on salient region detection of social images. First, the proposed deep learning based salient region detection method considers both appearance features and tag features. Tag features are detected by RCNN models. Second, tag neighbor features and appearance neighbor features are added to the saliency aggregation model. Finally, a new database of challenging social images and pixel-wise saliency annotations is constructed, which can promote further researches and evaluations of visual saliency model.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Figure 14: Visual comparisons of FBS with DBS. The order of images is original image, ground truth mask, FBS, and DBS.

Table 3: F-measure, AUC, and MAE of OBS and 11 state-of-the-art methods on six state-of-the-art datasets.

Metric	AUC	F-measure	MAE
Dataset		MSRA1000	
FT	0.766	0.579	0.241
DRFI	0.966	0.845	0.112
RC	0.937	0.817	0.138
GC	0.863	0.719	0.159
HS	0.93	0.813	0.161
MC	0.975	0.894	0.054
MR	0.941	0.824	0.127
SF	0.886	0.7	0.166
BD	0.948	0.82	0.11
MDF	0.978	0.888	0.066
LEGS	0.958	0.87	0.081
OBS	0.984	0.893	0.061
Dataset		HKU-IS	
FT	0.71	0.477	0.244
DRFI	0.95	0.776	0.167
RC	0.903	0.726	0.165
GC	0.777	0.588	0.211
HS	0.884	0.71	0.213
MC	0.928	0.798	0.102
MR	0.87	0.714	0.174
SF	0.828	0.59	0.173
BD	0.91	0.726	0.14
MDF	0.971	0.869	0.072
LEGS	0.907	0.77	0.118
OBS	0.976	0.871	0.078
Dataset		PASCAL-S	
FT	0.627	0.413	0.309
DRFI	0.899	0.69	0.21
RC	0.84	0.644	0.227
GC	0.727	0.539	0.266
HS	0.838	0.641	0.264
MC	0.907	0.74	0.145
MR	0.852	0.661	0.223
SF	0.746	0.493	0.24
BD	0.866	0.655	0.201
MDF	0.921	0.771	0.146
LEGS	0.891	0.752	0.157
OBS	0.927	0.778	0.141
Dataset		ECSSD	
FT	0.663	0.43	0.289
DRFI	0.943	0.782	0.17
RC	0.893	0.738	0.186
GC	0.767	0.597	0.233
HS	0.885	0.727	0.228
MC	0.948	0.837	0.1
MR	0.888	0.736	0.189
SF	0.793	0.548	0.219
BD	0.896	0.716	0.171
MDF	0.957	0.847	0.106
LEGS	0.925	0.827	0.118
OBS	0.968	0.856	0.112
Dataset		DUT-OMRON	
FT	0.682	0.381	0.25
DRFI	0.931	0.664	0.15
RC	0.859	0.599	0.189
GC	0.757	0.495	0.218
HS	0.86	0.616	0.227
MC	0.929	0.703	0.088
MR	0.853	0.61	0.187
SF	0.81	0.495	0.147
BD	0.894	0.63	0.144
MDF	0.935	0.728	0.088
LEGS	0.885	0.669	0.133
OBS	0.943	0.731	0.091
Dataset		SOD	
FT	0.607	0.441	0.323
DRFI	0.89	0.699	0.223
RC	0.828	0.657	0.242
GC	0.692	0.526	0.284
HS	0.817	0.646	0.283
MC	0.868	0.727	0.179
MR	0.812	0.636	0.259
SF	0.714	0.516	0.267
BD	0.827	0.653	0.229
MDF	0.899	0.793	0.157
LEGS	0.836	0.732	0.195
OBS	0.907	0.801	0.163

Acknowledgments

This work was supported in part by the Program Project of Beijing Municipal Education Commission (KM201511417008), the National Natural Science Foundation of China (Grant no. 62372148), the National Natural Science Foundation of China (Grant no. 61272352), and Beijing Natural Science Foundation (4152016).

References

[1] S. Pare, A. Kumar, V. Bajaj, and G. Singh, "An efficient method for multilevel color image thresholding using cuckoo search algorithm based on minimum cross entropy," *Applied Soft Computing*, vol. 61, pp. 570–592, 2017.

[2] S. Pare, A. Bhandari, A. Kumar, and G. Singh, "An optimal color image multilevel thresholding technique using grey-level co-occurrence matrix," *Expert Systems with Applications*, vol. 87, pp. 335–362, 2017.

[3] P. J. McParlane, Y. Moshfeghi, and J. M. Jose, "Collections for automatic image annotation and photo tag recommendation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 8325, no. 1, pp. 133–145, 2014.

[4] W. Zhang, A. Borji, Z. Wang, P. Le Callet, and H. Liu, "The application of visual saliency models in objective image quality assessment: a statistical evaluation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1266–1278, 2016.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[6] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[7] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 1155–1162, IEEE, Portland, Ore, USA, June 2013.

[8] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR '15*, pp. 5455–5463, 2015.

[9] 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, CVPR 2015*, pp. 1265–1274, IEEE, Massachusetts, Mass, USA, June 2015.

[10] L. Wang, H. Lu, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR '15*, pp. 3183–3192, IEEE, Massachusetts, Mass, USA, June 2015.

[11] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 9908, pp. 825–841, 2016.

[12] H. Li, J. Chen, H. Lu, and Z. Chi, "CNN for saliency detection with low-level feature integration," *Neurocomputing*, vol. 226, pp. 212–220, 2017.

[13] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 2106–2113, IEEE, Kyoto, Japan, October 2009.

[14] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[15] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, pp. 438–445, USA, June 2012.

[16] G. Zhu, Q. Wang, and Y. Yuan, "Tag-Saliency: combining bottom-up and top-down information for saliency detection," *Computer Vision and Image Understanding*, vol. 118, pp. 40–49, 2014.

[17] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 914–921, Barcelona, Spain, November 2011.

[18] J. Hosang, R. Benenson, P. Dollar, and B. Schiele, "What makes for effective detection proposals?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 4, pp. 814–830, 2016.

[19] Y. Jia and M. Han, "Category-independent object-level saliency detection," in *Proceedings of the 2013 14th IEEE International Conference on Computer Vision, ICCV '13*, pp. 1761–1768, IEEE, Sydney, Australia, December 2013.

[20] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by UFO: uniqueness, focusness and objectness," in *Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV '13)*, pp. 1976–1983, IEEE, Sydney, Australia, December 2013.

[21] W. Wang, C. Lang, and S. Feng, "Contextualizing tag ranking and saliency detection for social images," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface*, vol. 7733, no. 2, pp. 428–435, 2013.

[22] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: a data driven approach," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1131–1138, IEEE, Oregon, Ore, USA, June 2013.

[23] R. Achantay, S. Hemamiz, F. Estraday, and S. Süsstrunky, "Frequency-tuned salient region detection," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 1597–1604, June 2009.

[24] C. Yang, L. H. Zhang, H. C. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 3166–3173, 2013.

[25] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: a discriminative regional feature integration approach," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2013.

[26] T. Liu, Z. Yuan, J. Sun et al., "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[27] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of the 8th International Conference on Computer Vision*, pp. 416–423, July 2001.

[28] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 315–327, 2012.

[29] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR '10*, pp. 3169–3176, IEEE, California, Calif, USA, June 2010.

[30] Y. Jia, E. Shelhamer, J. Donahue et al., "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*, pp. 675–678, ACM, Orlando, Fla, USA, November 2014.

[31] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 580–587, Columbus, Ohio, USA, June 2014.

[32] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR '15*, pp. 1884–1892, IEEE, Massachusetts, Mass, USA, June 2015.

[33] S. Goferman, L. Manor, and A. Tal, "Context-aware saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 1915–1926, 2010.

[34] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *Proceedings of the British Machine Vision Conference*, pp. 1–12, BMVA Press, 2011.

[35] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV '13)*, pp. 2976–2983, Sydney, Australia, December 2013.

[36] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV '13)*, pp. 1529–1536, Sydney, Australia, December 2013.

[37] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proceedings of the 12th European conference on Computer Vision (ECCV '12)*, pp. 29–42, Florence, Italy, October 2012.

[38] X. Li, Y. Li, C. Shen, A. Dick, and A. V. D. Hengel, "Contextual hypergraph modeling for salient object detection," in *Proceedings of the 2013 14th IEEE International Conference on Computer Vision, ICCV '13*, pp. 3328–3335, December 2013.

[39] X. H. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix Recovery," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 853–860, 2012.

[40] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing Markov chain," in *Proceedings of the 14th IEEE International Conference on Computer Vision (ICCV '13)*, pp. 1665–1672, IEEE, Sydney, Australia, December 2013.

[41] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 1139–1146, IEEE, Oregon, Ore, USA, 2013.

[42] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '14)*, pp. 2814–2821, June 2014.

[43] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 4, pp. 996–1010, 2013.

[44] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Computer Vision—ECCV 2010*, vol. 6315 of *Lecture Notes in Computer Science*, pp. 366–379, Springer, Berlin, Germany, 2010.

[45] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: contrast based filtering for salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, pp. 733–740, June 2012.

[46] H. Peng, B. Li, R. Ji, W. Hu, W. Xiong, and C. Lang, "Salient object detection via Low-rank and Structured sparse Matrix Decomposition," in *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI '13*, pp. 796–802, July 2013.

[47] X. Hou, J. Harel, and C. Koch, "Image signature: highlighting sparse salient regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 194–201, 2012.

[48] C. Scharfenberger, A. Wong, K. Fergani, J. S. Zelek, and D. Clausi, "Statistical textural distinctiveness for salient region detection in natural images," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013*, pp. 979–986, June 2013.

[49] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with Gaussian edge potentials," *Advances in Neural Information Processing Systems*, pp. 109–117, 2011.

[50] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: a real-world web image database from national university of singapore," in *Proceedings of ACM International Conference on Image and Video Retrieval, CIVR '09*, 2009.

[51] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14*, pp. 280–287, IEEE, Massachusetts, Mass, USA, June 2014.