Saliency Detection via CNN Coarse Learning and Compactness Based ELM Refinement

Ruirui Li(✉), Shihao Sun, Lei Yang, and Wei Hu

Beijing University of Chemical Technology,
No. 15, Beisanhuandong Road, Chaoyang District, Beijing, China
ilydouble@gmail.com, 472527311@qq.com, ylxx@live.com,
huwei@mail.buct.edu.cn

Abstract. Salient object detection has attracted a lot of research in computer vision. It plays a vital role in image retrieval, object recognition and other image processing tasks. Although varieties of methods have been proposed, most of them heavily depend on feature selection and fail in the case of complex scenes. We propose a processing framework for saliency detection which contains two main steps. It uses deep convolutional neural networks (CNNs) to find a coarse saliency region map that includes semantic clues. Then it refines the coarse saliency map by training an extreme learning machine (ELM) on a group of color and texture compactness features. To get final saliency objects, it synthesizes the coarse saliency region map and several multiscale saliency maps that are obtained by refining the coarse one together. The method achieves good experimental results and can be used to improve the existing salient object detection methods as well.

Keywords: CNNs · ELM · Saliency · Compactness · Multiscale

1 Introduction

Visual saliency aims at detecting salient attention-grabbing parts in an image. It has received increasing interest in recent years. Though early research primarily focus on predicting eye-xations in images, it has shown that salient object detection, which segments entire objects from images, is more useful and has been successfully applied in object recognition [22], image classification [29], object tracking [28], image compression [23], and image resizing [2]. Despite recent progress in deep learning, salient object detection remains a challenging problem that calls for more accurate solutions.

Without a rigorous definition of image saliency, traditional saliency detection methods rely on several saliency priors. The contrast prior is the most popular one, which can be further categorized as local contrast and global contrast [7,31]. Conventionally, the contrast based methods use hand-crafted features based on human knowledge on visual attention and thus they may not generalize well in different scenarios.

J. Yang et al. (Eds.): CCCV 2017, Part II, CCIS 772, pp. 445–460, 2017.
https://doi.org/10.1007/978-981-10-7302-1_37

Several researchers [19,26,33] propose CNN based approaches for saliency detection. Though better performance has been achieved, there are still two major issues of prior CNN based saliency detection methods [27]. Firstly, CNNs use limited sizes of local image patches as input. To consider the spatial consistency, the CNNs networks require carefully designing and become extremely complex. Secondly, saliency priors, which are shown to be elective in previous work, are completely discarded by most CNN based methods.

Because of this, in our work, we present a progressive framework for image saliency computation called CELM. We leverage the advantages of high-level semantically meaningful features from deep learning as well as hand-crafted features when inferring saliency maps. Specifically, the framework has two procedures and learns the saliency map in a coarse-to-fine manner. The coarse-level image saliency semantically identities rough regions for salient objects by the CNN. The CNN takes the whole images as input and trains a global model to measure the saliency score of each pixel in an image, generating a coarse-level saliency map in a lower resolution. The fine-level image saliency is achieved by an ELM-based classification. This step is guided by the coarse-level saliency map and the input RGB image from which we fetch compactness features. After that, the coarse-level saliency map and the refined saliency map are synthesis together to get the final result. Figure 1 shows some saliency results generated by our approach. Extensive experiments on the standard benchmarks of image saliency detection demonstrate that the proposed CELM has better performance compared with state-of-the-art approaches. In summary, this paper makes the following main contributions to the community:

- A progressive saliency framework is developed by integrating CNN and ELM, taking advantage of both semantic and hand-crafted features. This model is general to be extended to improve the current saliency detection method.
- A heuristic learning method based on ELM is proposed to refine the coarse saliency image. It utilizes a new compactness hypothesis on superpixels to find subtle structures which has the same color and texture distribution as the positive samples.
- A saliency image synthesis algorithm based on saliency priors is proposed to fusion multiple leveled saliency maps.

The remainder of the paper is organized as follows. Section 2 reviews related work and differentiates our method from such work. Section 3 introduces our proposed method. Extensive experimental results and comparisons are presented in Sect. 4. And Sect. 5 concludes this paper.

2 Related Works

2.1 Image Saliency Detection

Image saliency detection approaches can be roughly categorized into two groups: bottom-up and top-down. Bottom-up methods focus on the low level features

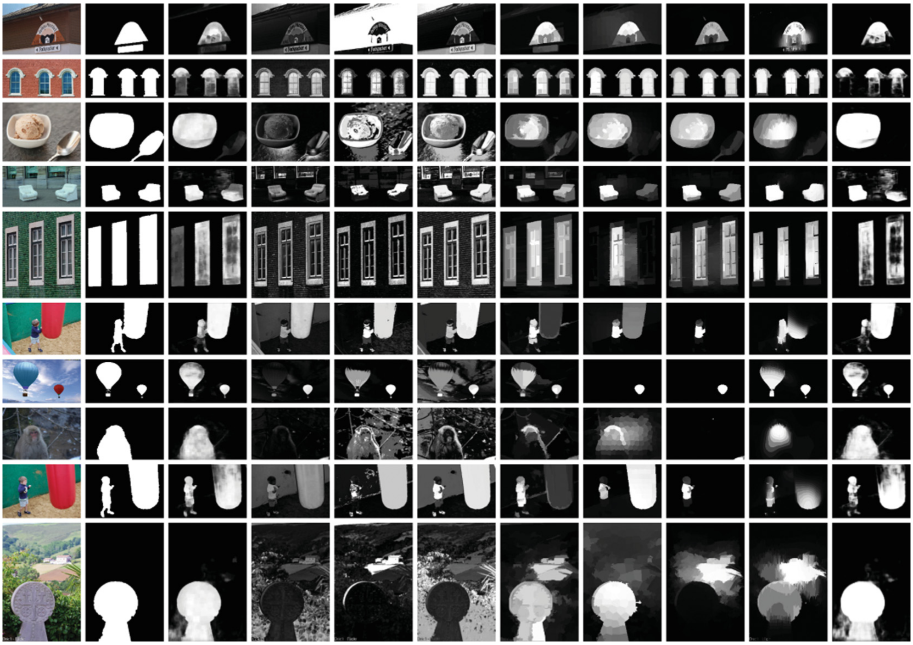

Fig. 1. Results from HKU-IS by different methods, from left to right: input, GT, our CELM, FT, GC, HC, DRFI, GMR, QCUT, PISA, DISC

(e.g. orientation, color, intensity, etc.). And the low level based methods can be further divided into local methods, global methods and hybrid of previous two according to the spatial scope of saliency computation. Local methods design saliency features by considering the contrast in a small neighborhood. As an example, in [16], multiscale image features (colors, intensity and orientations) are combined to generate saliency map. However, if there is a lot of high frequency noise in an image, local methods may result in a very poor performance. On the contrary, global methods compute saliency of an image region using its contrast over the entire image, which can tackle aforementioned problems. For example, Cheng et al. [9] used Histogram based Contrast (HC) and spatial information-enhanced Region based Contrast (RC) to measure saliency. But on the other hand, global methods ignore the details of local regions, leading to the blurring of the edges of the saliency. In order to combine the advantages of the complementary pair, Chen et al. [5] simultaneously integrate local and global structure information by designing a structure-aware descriptor based on the intrinsic bi-harmonic distance metric. Top-down methods move attention to high level features (e.g., faces, humans, cars, etc.), and are usually task-dependent. Yang et al. [32] proposed a top-down visual saliency model which incorporates a layered structure from top to bottom: CRF, sparse coding and image patches. Considering both importance of low and high features, Borji et al. [4] proposed a boosting model by integrating bottom-up features and top-down features.

2.2 Deep Learning for Saliency Detection

In recent years, with the development of deep learning, the methods of saliency detection based on deep learning have become a hotspot. Compared with the traditional manual extraction of features, the ones based on deep learning (such as convolution neural network) not only have better robustness, but also contains a higher level of semantic information, which is very important for salient object detection.

In [33], local and global context are integrated into a multi-context deep learning framework for saliency detection, whose performance was improved a lot, compared to many conventional approaches. Instead of using a fixed size local context, Li et al. [19] used a spatially varying one, which relies on the actual size of the surrounding regions. Furthermore, in order to dig more valuable information hidden inside the concatenated multiscale deep features, Li et al. [33] used neural network architecture at integrating stage. Though significant improvements have been made, the efficiency of deep feature extraction is not satisfied because of significant redundancy in computation and storage. In [10], rather than treat each region as an independent unit in feature extraction without any shared computation, Li et al. proposed an end-to-end deep contrast network which consists of a pixel-level fully convolutional stream and a segment-wise spatial pooling stream. In [6], Chen et al. also proposed an end-to-end deep hierarchical saliency network, whose architecture works in a global to local and coarse to fine manner. Our method also learns the saliency map in a progressive way. But different from Chens work, we use extreme learning machine to get fine saliency maps. While our approach leverages the advantages of high-level semantically meaningful features from deep learning, it also integrates handcrafted features when inferring saliency maps.

3 Proposed Method

3.1 Progressive Framework

The pipeline of the proposed method is summarized in Fig. 2. We train the CNN to get the coarse-level saliency map. It can be found that Pixels in the coarse-level saliency map are probably unconnected to form coherent regions. To maintain the spatial structure, we divided the original image into a number of superpixels. Then we statistically compute the labels of superpixels through the coarse-level saliency map. Combined with a group of extracted features of the superpixels, an ELM classier for the given image is trained, and its condense output for each super-pixel is used as a measure of saliency. This procedure is carried out in a multiscale way. Finally, the detected results of different scales and the coarse saliency map are synthesized to form a strong and fine saliency map result.

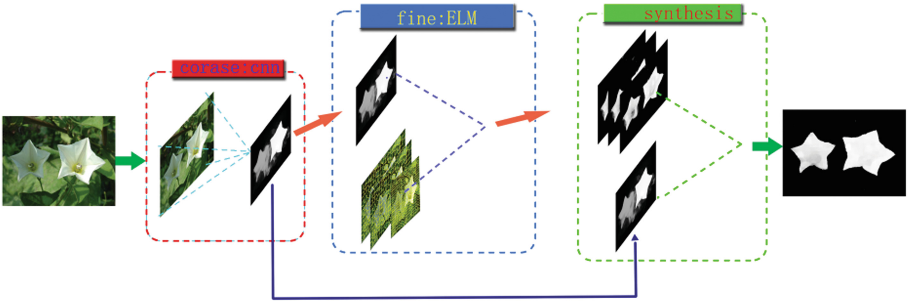

Fig. 2. Saliency detection framework

3.2 Coarse Saliency Map

We use all the images in the MSRA10K dataset and their labeled ground truth to train the coarse-level CNN. Its architecture is similar to the general AlexNet [18] proposed by Krizhevsky et al. Fig. 3 depicts the overall architecture of our CNN which contains six layers. The first five are convolutional and the last one is fully connected. All of them contain learnable parameters. The input of our CNN contains three RGB channels of whole images and one channel of average map as Fig. 4 illustrated. They are all resized to 256 × 256 and then filtered by the first convolutional layer (COV1) with 96 kernels of size 11 × 11 × 4 with a padding of 5 pixels. The result 62 × 62 × 96 feature maps are then sequentially given to a rectified linear unit (ReLU1) followed by a LRN and a max-pooling layer (MAXP1) which performs max pooling over 3 × 3 spatial neighborhoods with a stride of 2 pixels. The output of the MAXP1 is 31 × 31 × 96 features and then passed to the second convolutional layer (COV2). The number of filter kernels is changed from COV2 to COV5. They are set to 5, 3, 3, and 3 respectively. According to the parameter configuration of each layer, the architecture of the CNN can be described concisely by layer notations with layer sizes:

COV1(62 × 62 × 96)→RELU1→MAXP1→
COV2(31 × 31 × 96)→RELU2→MAXP2→
COV3(15 × 15 × 256)→RELU3→
COV4(15 × 15 × 384)→RELU4→
COV5(15 × 15 × 256)→RELU5→MAXP5→
FC1(7 × 7 × 256)→4096.

The last fully connected layer serves like a SVM. It computes the linear transformations of the feature vector and outputs 4096 saliency scores. The 4096 values are later re-arranged to a 64 × 64 coarse-level saliency map.

3.3 Saliency Refinement

The coarse-level saliency detection takes the whole image as input. It mainly considers the global saliency region in the image. Since it pays less attention

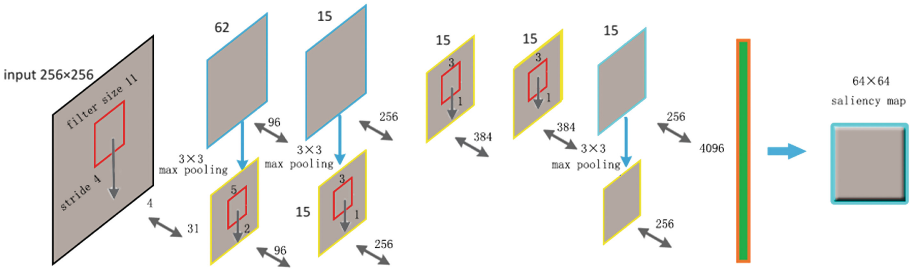

Fig. 3. CNN architecture

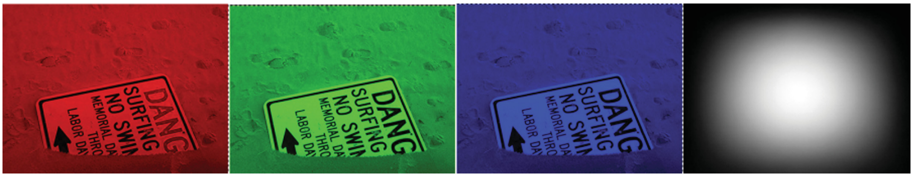

Fig. 4. Four input channels for CNN

to local context information, the salient pixels in coarse saliency map may be unconnected and may mistakenly lose subtle salient structures. To further refine the saliency result, we train a binary classier through Extreme Learning Machine [13] (ELM). The extreme learning machine (ELM) is proposed by Huang et al. [13,14]. It is a kind of machine learning algorithm for the single-layer feed-forward neural network (SLFN) [15] and its architecture is given in Fig. 5. The ELM contains three layers: the input layer, the hidden layer, and the output layer. The ELM randomly initializes the weights between input layer and hidden layer as well as the bias of hidden neurons. It analytically determines the weights between the hidden layer and the output layer using the least-squares method. Compared with Neural networks (NN), support vector machines (SVM) and other popular learning methods, the ELM has several significant advantages, such as real-time learning, high accuracy, least user intervention. In our experiments, sigmoid neurons are chosen for the training.

We use both the coarse saliency image and the original given image as input and propose a multiscale superpixel-based statistic method to label the saliency image. The procedure of saliency refinement is shown as Fig. 6. The original given RGB image is converted into the CIE LAB color space first, and is efficiently segmented into multi-leveled sets of superpixels by the SLIC algorithm [1]. We use two thresholds, T_h and T_l ($T_h > T_l$) to collect the training samples from the coarse saliency map. If average saliency value of the superpixel is higher than T_h, the superpixel is labeled as positive sample. If average saliency value of the superpixel is lower than T_l, the superpixel is labeled as negative sample.

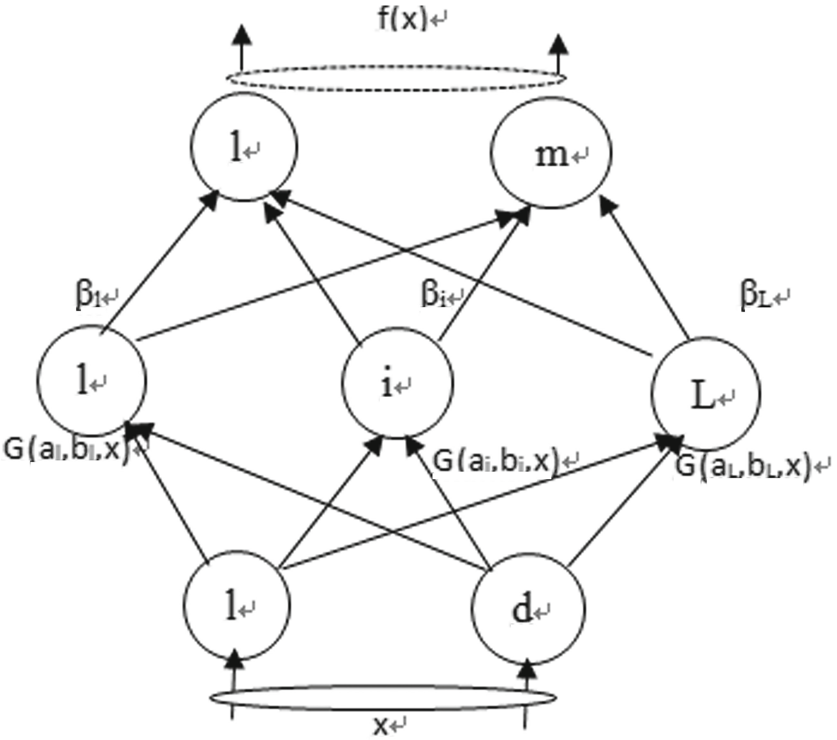

Fig. 5. ELM architecture: single layer feed forward neural network

Those whose values are between T_h and T_l are discarded. The thresholds are statically computed by the formula as:

$$\begin{cases} T_h = \min[(1+a)*(M+b), 0.9] \\ T_l = \min[(1-a)*(M+b), 0.1] \end{cases} \tag{1}$$

where M is the mean grayscale value of the coarse saliency map; a is the scope ratio parameter; b is the saliency ratio parameter. In our experiments, a is generally set to be 0.8 and b is computed by 0.5 subtraction of the mean gray scale value of average map.

The classification is based on a group of 16-dimensional feature of compactness. According with boundary prior assumption, we depart the superpixels into two groups, the center group and the boundary group. As the Fig. 7 illustrated, the areas which are masked by the transparent red color are the boundary group, and the left areas are the center group. For both groups, we compute compactness values on both color and texture.

The method to compute the compactness features is similar to the work [12]. For a superpixel v_i, the compactness degree Θ are computed by the Eq. 2 where W is a weight factor, θ is the scatter degree, c is an element in the set of { L, a, b, Lab}. The compactness degree is the summation of weighted scatter degree.

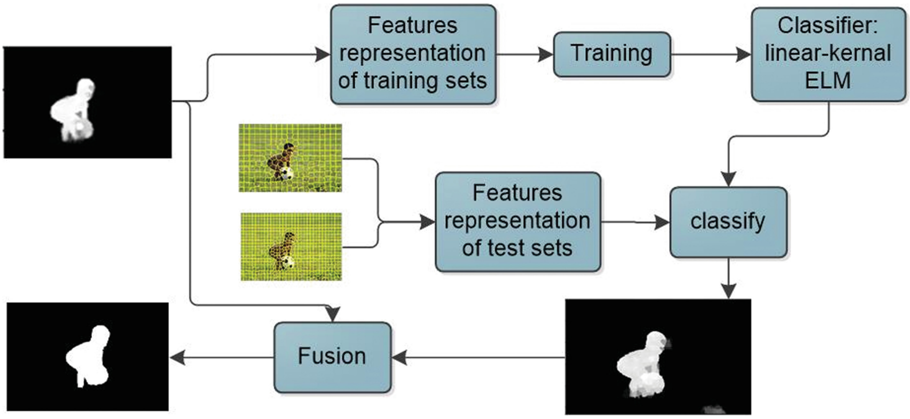
Features representation of training sets
Training
Classifier: linear-kernal ELM
Features representation of test sets
classify
Fusion

Fig. 6. Procedure of saliency refinement

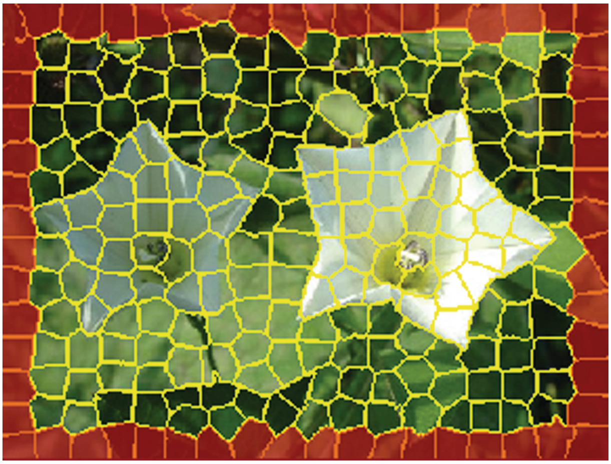

Fig. 7. The Center group and the boundary group

The weight value is computed according to the pattern of Gaussian functions by Eq. 4. The scatter degree for v_i is computed by Eq. 3 which is the reciprocal form of weighted linear combination of spatial distance factor $D(p_i, p_j)$. We use the Euclidean distance to measure the spatial distance.

$$\varTheta^c(v_i) = \sum_{j=1}^{N} W^c(v_i, v_j) \cdot \theta^c(v_j) \tag{2}$$

$$\theta^c(v_j) = \frac{1}{\sum_{j=1}^{N} D(p_i, p_j) \cdot W^c(v_i, v_j)} \tag{3}$$

$$W^c(v_i, v_i) = \frac{1}{\Omega_i} \exp(-\frac{(D(c_i, c_j)^2}{2\sigma^2}) \tag{4}$$

The algorithm computes the weight values for individual color channel and the whole color space. For color weights computation, we simply use the mean color value of the superpixel to represent v_i and v_j. For texture weights computation, we use a uniform LBP histogram stated on the superpixel. It is commonly a 59-dimensional vector.

The 16 features of superpixels and their saliency labels gotten from the coarse saliency map are all sent to the ELM as input. According to our approach, an ELM classier is trained. The output of ELM is confidence value of each superpixel and it is used as a measure of saliency. We perform the process in multiple scales of superpixels. In our experiments, most of the images in dataset are in about 400 × 400 pixels resolution. Thus we choose three numbers for superpixels to control the scales; they are [150, 250, 750].

3.4 Image Synthesis

The refined saliency map is generated by integrating several rened saliency maps. For each saliency map, we first perform a smoothing operation through the GrabCut algorithm. The GrabCut algorithm [24] was designed by Rother et al. from Microsoft Research Cambridge, UK. It extracts foreground using iterated graph-cuts. The GrabCut requires a mask map in which all pixel values are set with a value in 0, 1, 2, 3 which means foreground, background, probably foreground, and probably background. We compute two thresholds by Eq. 1 to set the mask values using the following rules:

$$M(i,j) = \begin{cases} 0, & M(i,j) < T_l \\ 1, & M(i,j) > T_h \\ 2, & T_l < M(i,j) < (T_h + T_l)/2 \\ 3, & T_h > M(i,j) > (T_h + T_l)/2 \end{cases} \tag{5}$$

The quality of the saliency map should be evaluated before synthesis. This step is carried out based on two hypotheses: the compactness and the variances. It is observed that salient regions are usually distributed close and the pixels in saliency map probably have high variance. Thus we compute the assessment weight $\psi(S_i)$ by

$$\psi(S_i) = norm(\frac{1}{\sum_{i=1}^{H} \sum_{j=1}^{W} S(i,j) D(S(i,j), S_{GC})^2}) \tag{6}$$

where $S(i,j)$ is the pixel with position(i,j), S_{GC} is the saliency gravity, and $norm()$ is the normalization operation. Multiple saliency maps then are integrated in a linear weighted sum form of involved saliency map according to its corresponding assessment score.

4 Experiment Result and Analysis

4.1 Experiments Setup

To evaluate the effectiveness of the proposed method, we perform experiments on five different types of datasets publicly available, which are MSRA10K [9], ECSSD [30], PASCAL-S [21], DUT-OMRON [31], and HKU-IS [19] datasets. All of the dataset are annotated by people and have pixel-wise groundtruth. The feature of each dataset is listed in Table 1.

Table 1. The description of the five datasets

Dataset	Size	Source	Description
MSRA10K	10000	MSRA	Only one salient object
ECSSD	1000	Internet	Structurally complex images
DUT-OMRON	5168	Dalian University of Technology	Controversial annotations
PASCAL-S	850	VOC2010	Contain 12 subjects
HKU-IS	4447	Hong Kong University	Contain multiple salient objects

4.2 Evaluation Metrics

In the comparison experiments, we use Precision-Recall (PR) curves, $F_{0.3}$ metric, AUC, and Mean Absolute Error (MAE) to evaluate the proposed method. With saliency value in the range [0,255], the P-R curve is obtained by generating the binary map when the threshold varies from 0 to 255, and comparing the binary result with the ground-truth. The F-measure is defined as,

$$F = \frac{(1+\beta)^2 \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{7}$$

where β^2 is set to be 0.3 as suggested in [1]. AUC is the area under ROC. As indicated in [11], PR curves, AUC and F metric provide a quantitative evaluation, while MAE provides a better estimate of the dissimilarity between the saliency map and binary ground truth. The MAE computes the average pixel-wise difference between saliency map S and the binary ground truth G.

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)| \tag{8}$$

4.3 Performance Comparison

In this subsection, we evaluate the proposed method on MSRA10K, ECSSD, OMRON, PASCAL-S, HKU-IS dataset and compare the performance with 7 state-of-the-art algorithms, including HC [9], GC [8], GMR [31], PISA [25],

Table 2. Comparison between our CELM and other methods

Dataset	Metrics	HC	GC	GMR	PISA	QCUT	DRFI	DISC	Ours
MSRA10K	MAE	0.215	0.150	0.126	0.101	0.116	0.126	0.044	0.039
	F_β	0.677	0.766	0.846	0.861	0.873	0.877	0.940	0.952
ECSSD	MAE	0.331	0.233	0.187	0.150	0.173	0.170	0.119	0.121
	F_β	0.455	0.598	0.738	0.766	0.766	0.782	0.799	0.838
OMRON	MAE	0.310	0.217	0.189	0.141	0.126	0.150	0.119	0.095
	F_β	0.381	0.496	0.612	0.631	0.684	0.664	0.659	0.711
PASCAL-S	MAE	0.354	0.264	0.233	0.192	0.198	0.207	0.156	0.151
	F_β	0.423	0.536	0.643	0.657	0.683	0.688	0.725	0.744
HKU-IS	MAE	0.292	0.215	0.175	0.128	0.143	0.145	0.103	0.087
	F_β	0.493	0.580	0.712	0.752	0.766	0.776	0.784	0.834

QCUT [3], DRFI [17], and DISC [6]. For fair comparison, we use the original source code provided by authors or the detection results provided by the corresponding literatures. The quantitative comparisons are shown in Figs. 8 and 9, and Table 1. We train the coarse CNN model on MSRA10K and test the model on other datasets to prove the generalization performance (Table 2).

The existing CNN based model DISC uses two CNN to train an end-to-end saliency detection model in a coarse to fine manner. Different from it, our model choose the ELM to refine the saliency map combining deep features and handicraft features. Our CELM based model improves the F-measure achieved by the DISC [6] by 1.3%, 4.9%. 7.9%, 2.6% and 6.8% respectively on MSRA10K,

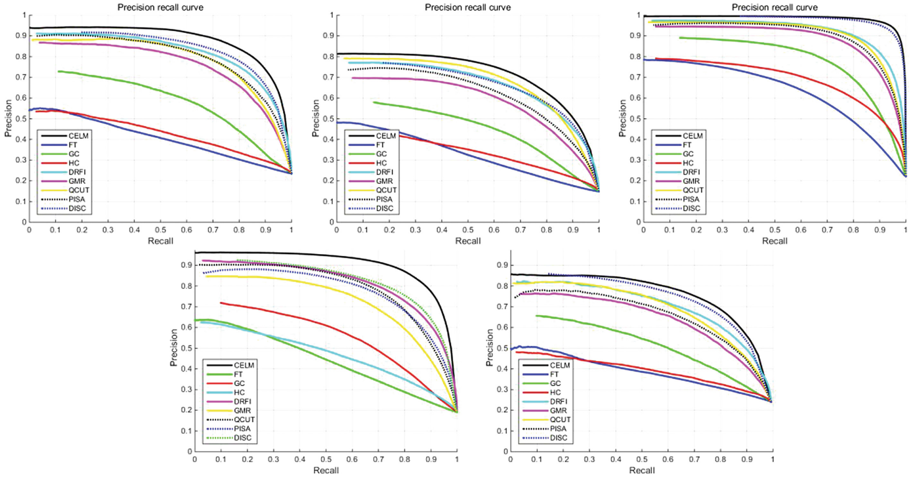

Fig. 8. Precision-Recall Curves on datasets, from left to right, up to down: ECSSD, OMRON, MSRA10K, HKU-IS, PASCAL-S

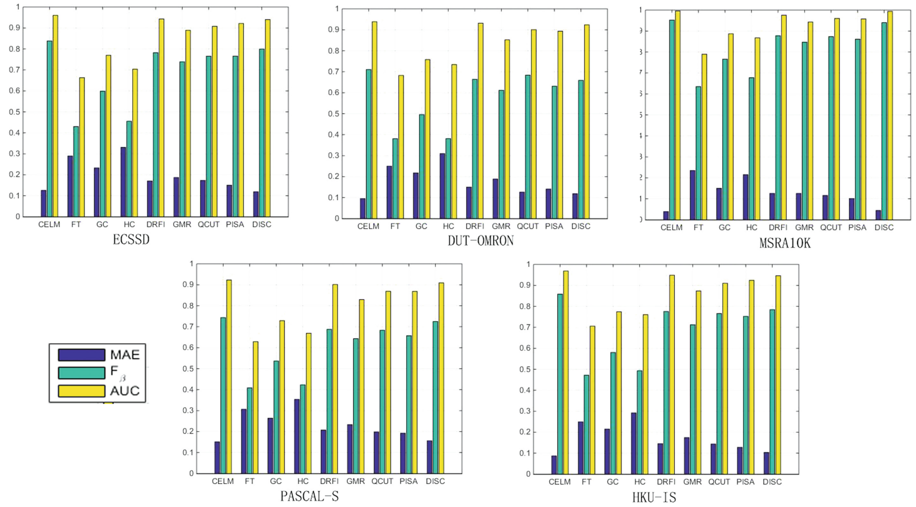

Fig. 9. MAE, F-measure and AUC values of compared methods on five datasets, from left to right, up to down: ECSSD, OMRON, MSRA10K, PASCAL-S, HKU-IS

ECSSD, OMRON, PASCAL-S, HKU-IS. At the same time, our CELM based model lowers the MAE by 11.4%, 20.2%, 3.2%, 15.5% on MSRA10K, OMRON, PASCAL-S and HKU-IS. Our method outperforms all the seven previous methods based on the three evaluation metrics.

4.4 Analysis of Propose Method

For most images, the approach achieves good salient results as Fig. 1 shows. However, the final saliency map heavily depended on the results gotten from CNN. Our CELM based model may fail if the coarse CNN output totally wrong region as Fig. 10 illustrated. Because the saliency map generated by CNN indicates the location of the salient object, if this saliency map indicates the wrong location of the salient map, the detection will failed. To further improve the performance, we could replace the AlexNet CNN with the state-of-the-art CNN network,

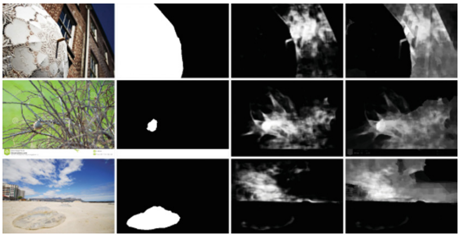

Fig. 10. Failed cases, from left to right are Input, groundtruth, coarse map, CELM

i.e. the DCL network [20]. The DCL network is proposed in CVPR 2016. It is one of the best models to detect salient object.

We use the same refining and synthesis framework as Subsects. 3.3 and 3.4 described. We compare our proposed CELM-DCL method with the original CELM, the DCL, and the other seven state of the art methods on three datasets: HKU-IS(1446), ECSSD and OMRON as Fig. 11 illustrated. This is because the DCL only provides results on these three datasets, and only 1446 results are provided in the HKU-IS dataset. For quantitative evaluation, we show comparison results with PR curves and F-measure scores in Table 3 and Fig. 12.

Table 3. Comparison between CELM-DCL and other methods

Dataset	Metrics	GMR	PISA	QCUT	DRFI	DISC	CELM	DCL	CELM-DCL
ECSSD	MAE	0.187	0.150	0.173	0.170	0.119	0.126	0.068	0.067
	F_β	0.738	0.766	0.766	0.782	0.799	0.838	0.902	0.907
OMRON	MAE	0.189	0.141	0.126	0.150	0.119	0.116	0.080	0.075
	F_β	0.612	0.631	0.684	0.664	0.659	0.711	0.756	0.768
HKU-IS(1446)	MAE	0.175	0.128	0.143	0.145	0.103	0.102	0.048	0.057
	F_β	0.712	0.752	0.766	0.776	0.784	0.834	0.907	0.913

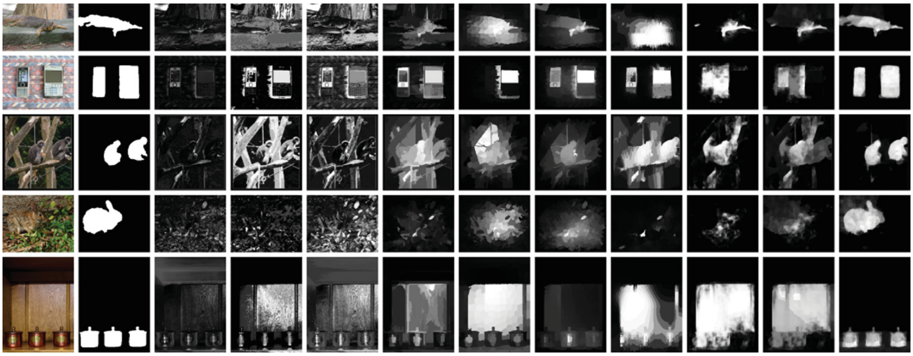

Fig. 11. Results from HKU-IS dataset by different methods, from left to right: input, groundtruth, FT, GC, HC, DRFI, GMR, QCUT, PISA, DISC, CELM, CELM-DCL

In Table 3, the best results are marked red and the second ones are marked green. Our CELM-DCL based model improves the F-measure achieved by the DCL by 0.5%, 1.6% and 0.7% respectively on ECSSD, OMRON, and HKU-IS. At the mean-time, the CELM-DCL based model lowers the MAE by 1.5% and 6.25% on ECSSD and OMRON.

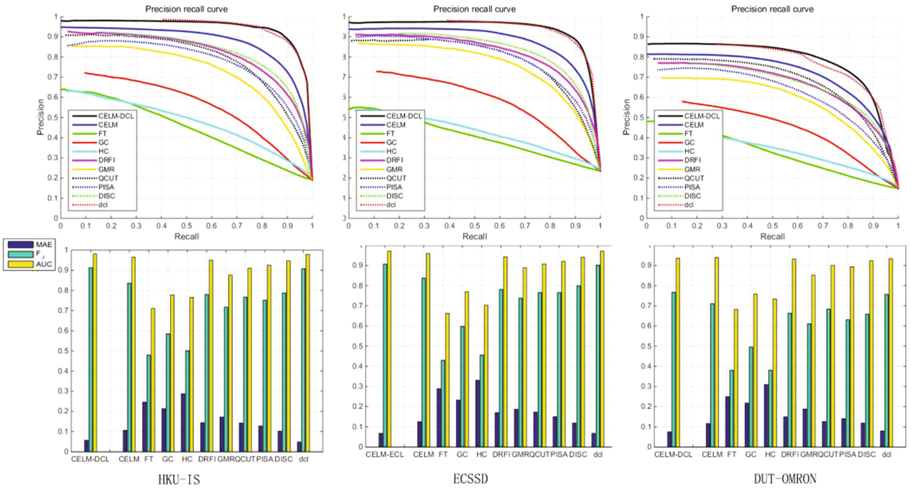

Fig. 12. Comparison of GMR, PISA, QCUT, DRFI, DISC, CELM, CDL, CELM-CDL on HKU-IS, ECSSD and DUT-OMRON; up: Precision-Recall curves; down: MAE, F-measure, and AUC

5 Conclusions

In this paper, we propose a saliency detection framework through the combination of CNN and ELM. To carry it out, we statically label the coarse map, extract the compactness features on two groups, and synthesis the multiple saliency map based on their qualities. Experiments show that the CELM get excellent salient results on all five datasets. Further improvements are also made by replacing the AlexNet with the network that the DCL uses. Extent experiments prove that the CELM-DCL outperforms the state of the art. It is proved that our approach can be used not only as a complete method, but also as a lifting method for current CNN based method. Future works include extending our work to a pixel-wise and accuracy approach as well as exploring better CNN networks.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Patt. Anal. Mach. Intell. **34**(11), 2274 (2012)
2. Avidan, S., Shamir, A.: Seam carving for content-aware image resizing. ACM Trans. Graph. **26**(3), 10 (2007)
3. Aytekin, Ç., Ozan, E.C., Kiranyaz, S., Gabbouj, M.: Visual saliency by extended quantum cuts. In: IEEE International Conference on Image Processing (2015)
4. Borji, A.: Boosting bottom-up and top-down visual features for saliency estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 438–445 (2012)

5. Chen, C., Li, S., Qin, H., Hao, A.: Structure-sensitive saliency detection via multilevel rank analysis in intrinsic feature space. IEEE Trans. Image Process. **24**(8), 2303–16 (2015)
6. Chen, T., Lin, L., Liu, L., Luo, X., Li, X.: DISC: deep image saliency computing via progressive representation learning. IEEE Trans. Neural Netw. Learn. Syst. **27**(6), 1135 (2016)
7. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Trans. Patt. Anal. Mach. Intell. **37**(3), 569–582 (2015)
8. Cheng, M.M., Warrell, J., Lin, W.Y., Zheng, S., Vineet, V., Crook, N.: Efficient salient region detection with soft image abstraction, pp. 1529–1536 (2013)
9. Cheng, M.M., Zhang, G.X., Mitra, N.J., Huang, X., Hu, S.M.: Global contrast based salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 409–416 (2011)
10. He, S., Lau, R.W., Liu, W., Huang, Z., Yang, Q.: SuperCNN: a superpixelwise convolutional neural network for salient object detection. Int. J. Comput. Vis. **115**, 330–344 (2015)
11. Hornung, A., Pritch, Y., Krahenbuhl, P., Perazzi, F.: Saliency filters: contrast based filtering for salient region detection. In: Computer Vision and Pattern Recognition, pp. 733–740 (2012)
12. Hu, P., Wang, W., Zhang, C., Lu, K.: Detecting salient objects via color and texture compactness hypotheses. IEEE Trans. Image Process. **25**(10), 4653–4664 (2016)
13. Huang, G.B., Zhou, H., Ding, X., Zhang, R.: Extreme learning machine for regression and multiclass classification. IEEE Trans. Syst. Man Cybern. Part B **42**(2), 513–529 (2012)
14. Huang, G.B., Ding, X., Zhou, H.: Optimization Method Based Extreme Learning Machine for Classification. Elsevier Science Publishers B.V., Amsterdam (2010)
15. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing **70**(1), 489–501 (2006)
16. Itti, L., Koch, C.: Computational modelling of visual attention. Nat. Rev. Neurosci. **2**(3), 194 (2001)
17. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: a discriminative regional feature integration approach. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2083–2090 (2013)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: International Conference on Neural Information Processing Systems, pp. 1097–1105 (2012)
19. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Computer Vision and Pattern Recognition, pp. 5455–5463 (2015)
20. Li, G., Yu, Y.: Deep contrast learning for salient object detection, pp. 478–487 (2016)
21. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 280–287 (2014)
22. Lin, L., Wang, X., Yang, W., Lai, J.H.: Discriminatively trained and-or graph models for object shape detection. IEEE Trans. Patt. Anal. Mach. Intell. **37**(5), 959–72 (2015)
23. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Tenth ACM International Conference on Multimedia, pp. 533–542 (2002)

24. Rother, C., Kolmogorov, V., Blake, A.: GrabCut: interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. (TOG) **23**(3), 309–314 (2004)
25. Wang, K., Lin, L., Lu, J., Li, C., Shi, K.: PISA: pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence. IEEE Trans. Image Process. **24**(10), 3019–3033 (2015)
26. Wang, L., Lu, H., Xiang, R., Yang, M.H.: Deep networks for saliency detection via local estimation and global search. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3183–3192 (2015)
27. Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency detection with recurrent fully convolutional networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 825–841. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_50
28. Wu, H., Li, G., Luo, X.: Weighted attentional blocks for probabilistic object tracking. Vis. Comput. **30**(2), 229–243 (2014)
29. Wu, R., Yu, Y., Wang, W.: SCaLE: supervised and cascaded laplacian eigenmaps for visual object recognition based on nearest neighbors. In: Computer Vision and Pattern Recognition, pp. 867–874 (2013)
30. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Computer Vision and Pattern Recognition, pp. 1155–1162 (2013)
31. Yang, C., Zhang, L., Lu, H., Xiang, R., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Computer Vision and Pattern Recognition, pp. 3166–3173 (2013)
32. Yang, J.: Top-down visual saliency via joint CRF and dictionary learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2296–2303 (2012)
33. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: Computer Vision and Pattern Recognition, pp. 1265–1274 (2015)