LETTER

Multi Information Fusion Network for Saliency Quality Assessment*

Kai TAN[†a)], **Qingbo WU**[†b)], **Fanman MENG**[†], ***Nonmembers***, ***and*** **Linfeng XU**[†], ***Member***

SUMMARY Saliency quality assessment aims at estimating the objective quality of a saliency map without access to the ground-truth. Existing works typically evaluate saliency quality by utilizing information from saliency maps to assess its compactness and closedness while ignoring the information from image content which can be used to assess the consistence and completeness of foreground. In this letter, we propose a novel multi-information fusion network to capture the information from both the saliency map and image content. The key idea is to introduce a siamese module to collect information from foreground and background, aiming to assess the consistence and completeness of foreground and the difference between foreground and background. Experiments demonstrate that by incorporating image content information, the performance of the proposed method is significantly boosted. Furthermore, we validate our method on two applications: saliency detection and segmentation. Our method is utilized to choose optimal saliency map from a set of candidate saliency maps, and the selected saliency map is feeded into an segmentation algorithm to generate a segmentation map. Experimental results verify the effectiveness of our method.

key words: *saliency quality assessment, multi information, deep convolutional neural network, image content*

1. Introduction

Saliency quality assessment is a kind of non-reference image quality evaluation method [1], which estimates the quality of a saliency map without ground-truth. A good saliency quality assessment method can boost the saliency detection [2] performance from multi saliency algorithms by choosing optimal saliency maps from candidates [3].

Several approaches for saliency quality assessment have been proposed during the past decades. In [4], Mai et al. first proposes an algorithm to rank different saliency results. They design a range of hand-crafted features based on the attributes of salient object such as size, spatial and color compactness and so on. In [3], Tang et al. propose a CNN based feature learning method for saliency quality prediction, which boosts the performance. However, it is designed to only use the information from the salient map without considering image content, which makes it still far from enough to accurately predict the saliency quality. An example is shown in Fig. 1. We generate two same saliency maps corresponding to two images with different ground truth, which presents clearly different quality scores (0.6527 and 0.3155 respectively). However, the model [3] gives them same prediction scores 0.9075, which is clearly wrong. This inspires us that we should not only exploit the information from saliency map but also consider the information from image content to jointly represent saliency quality.

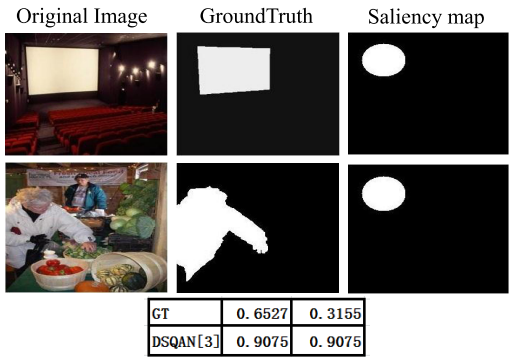


Fig. 1 For two same saliency maps corresponding two images with totally different ground truth, existing method give the same scores.

In this letter, we propose a multi-information fusion network (MIFN) to capture the information from the saliency map and image content simultaneously to represent saliency quality. The proposed deep network consists of two modules, a saliency map information module (SMI) and an image content information module (ICI). SMI operates on the saliency map to assess its compactness and closedness. ICI module is a siamese network utilizing the information from the original image to evaluate the consistence and completeness of foreground and the difference between foreground and background. At the end, a multi-information fusion module (MIF) is designed to combine all these information to output saliency quality scores.

Extensive experiments on two publicly available databases such as DUT-OMRON [5] and ECSSD [6], show that by incorporating image content information, the proposed model leads to significantly improved quality prediction accuracy. Furthermore, we also validate the applicability of the proposed method in selecting optimal saliency map from a set of candidates for improving the performance of saliency detection [7], [8] and segmentation [3], [9]. The results demonstrate that the optimal selection by our method can significantly outperforms the best saliency detection al-

Manuscript received January 4, 2019.
Manuscript revised February 10, 2019.
Manuscript publicized February 26, 2019.

[†]The authors are with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China.

*This work was supported in part by National Natural Science Foundation of China, under grant number 61601102 and 61871078. It was also supported by Sichuan Science and Technology Program (No. 2018JY0141).

a) E-mail: kaitanuestc@gmail.com
b) E-mail: qbwu@uestc.edu.cn (Corresponding author)

DOI: 10.1587/transinf.2019EDL8002

gorithms and existing quality assessment methods for both the saliency detection and saliency segmentation.

2. Multi Information Network

In this section, we present the details of the proposed multi-information fusion network. As shown in Fig. 2, our deep network consists of thee modules, Saliency Map Information (SMI), Image Content Information (ICI) and Multi Information Fusion (MIA) module. We will introduce these three modules, respectively.

2.1 Saliency Map Information

For Saliency Map Information (SMI), we aim to design a convolutional network [10] to model the information from a saliency map. The existing convolutional neural network VGG16 [11] is used as the basic model of SMI and it preserves all convolutional layers up to 'pool5'. Thus, the feature tensor F_{SM} output from the last pooling layer (pool5) is used to represent the information feature of the saliency map.

2.2 Image Content Information

The basic architecture of ICI is a siamese network with two weights-shared streams. The first stream processes the information from foreground which is used to evaluate its completeness and consistence. The second stream models the information from background [12] which is designed to assess the difference between foreground and background when combined with the first stream.

Specially, the input of the two streams is a saliency image I_s and a reverse-saliency image I_{rs}, respectively. The saliency (reverse-saliency) image is defined as an element-wise multiplication of an image I and its corresponding saliency map S (reverse-saliency map $1 - S$).

$$\begin{aligned} I_s &= I \otimes S \\ I_{rs} &= I \otimes (1 - S) \end{aligned} \tag{1}$$

These two streams also use VGG16 as the basic model. Each stream consists of all convolutional layers up to 'pool5'. After the last pooling layer (pool5), we obtain two feature tensors F_d and F_{ud} which are fused to output the final feature $F_{IC} \in 7\times7\times512$ with a concat layer, a convolutional layer and a Relu layer. The convolutional layer is used to reduce feature dimension to output a $7\times7\times512$ feature tensor F_{IC}. We set its kernel size as $3\times3\times1024\times512$ with a [1 1] stride and [1 1 1 1] padding.

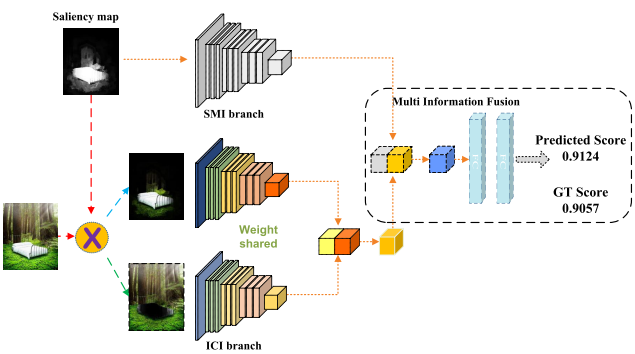


Fig. 2 The proposed multi information fusion network, consisting of SMI module, ICI module and MIF module.

2.3 Multi Information Fusion

Given F_{SM} and F_{IC}, a fusion module is designed to fuse these two feature tensors, which combines the information from both saliency map and image content to output a saliency quality score. F_{SM} and F_{IC} is first stacked by a concat layer to get a $7 \times 7 \times 1024$ tensor, and we reduce its dimension into 512 by a convolution layer with kernel size 3×3. Then, two fc blocks are added to capture the global feature. Each fc block consists of a fc layer, Batch Normalization (BN), Relu and dropout layer. After these two fc blocks, a fc layer is added to output a prediction which is mapped into the rang of [0, 1] by a sigmoid layer.

2.4 Training

In our proposed network, ICI is initialized with VGG-16 pretrained on ImageNet. SMI is initialized randomly. The learning rate is set to 0.0001 and the weight decay is 0.0005. Considering the GPU memory, we set the bach size as 16. The training epoch is 30. Our method is implemented in Matlab with the MatConvNet toolbox. We augment the training data by horizontal flipping to reduce overfitting. Stochastic gradient descent (SGD) optimization method is used for training. We use Euclidean loss to calculate the error between the predicted saliency quality score and groundtruth score.

We adopt the most commonly used metric F-measure to measure the quality of a saliency map. It is defined as the balanced mean of precision and recall:

$$S_F = \frac{(1+\beta^2) \times precision(p) \times recall(p)}{\beta^2 \times precision(p) + recall(p)} \tag{2}$$

where p represents the adaptive threshold which is used to binarize the saliency map. It is defined as twice the mean value of the saliency map, $p = \frac{2}{N}\sum S(i)$, where N is the number of the image pixels and $S(i)$ represents the saliency value of the i_{th} pixel. Following existing work [8], β^2 is set to 0.3 to place more emphasis on precision. The higher the S_F, the better the quality of a saliency map.

3. Experiments

We train and test our method on the well-known saliency detection dataset: DUT-OMRON [5], which has 5168 images. DUT-OMRON is randomly splited into training, validation and testing set with 3000, 1000 and 1168 images, respectively. The training set is used to train our model. To further verify the generation of our method, we also test

Table 1 The performance of saliency detection. Green numbers represent the best performance among singe salient object detection algorithms, and red numbers are the best performances.

Method	HC	GR	ORB	DSR	AMC	DSS	UCF	CS	DSQAN	SMI	ICI	MIFN
DUTOMRON	31.67	47.34	51.32	51.25	52.39	66.60	59.68	61.97	67.71	67.14	68.13	69.37
ECSSD	41.54	51.03	67.64	68.99	69.84	85.32	83.42	80.56	85.62	85.54	85.52	85.78

Table 2 Compared our model and its variants with the other state-of-the-art method on two well-known saliency detection datasets.

(a) DUT-OMRON

Method	SROCC	PLCC	RMSE
CS	0.5499	0.5392	27.73
DSQAN	0.7179	0.7103	23.17
SMI	0.6957	0.6883	23.88
ICI	0.7779	0.7745	20.83
MIFN	**0.8009**	**0.7980**	**19.84**

(b) ECSSD

Method	SROCC	PLCC	RMSE
CS	0.6518	0.6522	21.52
DSQAN	0.7676	0.7332	19.31
SMI	0.7433	0.7256	19.54
ICI	0.7521	0.7158	19.83
MIFN	**0.7870**	**0.7583**	**18.51**

our method on the other widely used dataset: ECSSD [6], which has 1000 images. The saliency map is generated by seven state-of-the-art saliency detection algorithms, which includes two deep learning based methods, i.e., DSS [13], UCF [14], and five non deep learning methods, i.e., HC [8], DSR [15], AMC [16], ORB [17], GR [18].

3.1 Comparison with State-of-the-Art Methods

We compare the proposed model with two existing state-of-the-art saliency quality assessment methods: CS [4] and DSQAN [3]. We implement CS method by ourself (since the code is not provided) and use the source code provided by the author to train DSQAN model.

We adopt three most commonly used performance measures Spearman's rank ordered correlation (SROCC), Pearson's linear correlation coefficient (PLCC), and Root mean-squared error (RMSE) to evaluate our method. SROCC and PLCC are used to measure the nonlinear correlation and linear correlation between the predicted score ranking and groundtruth ranking. RMSE is used to measure the error between the predicted score and groundtruth score. The higher scores of SROCC, PLCC, and the lower score of RMSE indicate the better performance of the quality metrics.

Table 2 shows the results of quantitative comparison with state-of-the-art methods. It can be observed that our model significantly and consistently outperforms the competing methods in terms of SROCC, PLCC and RMSE on both two datasets. For example, the proposed MIFN is 2%–8% higher than DSQAN and 13%–25% higher than CS in terms of SROCC on the two dataset.

3.2 Ablation Study

Our deep network consists of two complementary modules, SMI subnet and ICI sunbet. To show the effectiveness and necessity of these two components, we compare the quality score predicted from SMI, ICI and MIFN, respectively. Before training the SMI and ICI network separately, we add two fc blocks, a fc layer and sigmoid layer after the last layer in SMI and ICI. Euclidean loss is used as the loss function for these two networks, respectively.

Table 2 shows the performance of our model and variants on the two datasets. According to the results, we have following two observations: First of all, ICI performs better than SMI on two datasets, which demonstrates that image content plays a very important role in saliency quality assessment. Second, the information from the saliency map and image content are complementary to each other since the combination of SMI and ICI (i.e., MIFN) improves the performances in terms of the two single network in all cases. Specifically, the combination of SMI and ICI clearly boosts the performance, and has about 3%–4% performance gain in terms of SROCC and PLCC across two datasets.

3.3 Applications

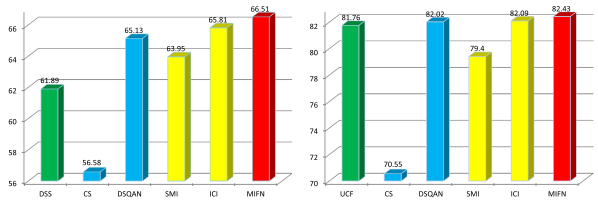
To further evaluate the effectiveness of the proposed MIFN, we validate our method and the other quality assessment methods CS [4] and DSQAN [3] on two applications: saliency detection and segmentation.

3.3.1 Saliency Map Selection

For each image, we use quality assessment method to choose the best quality saliency map from a set of candidate saliency maps generated from seven state-of-the-art saliency detection algorithms. Particularly, we select the saliency map with the highest predicted quality score as the optimal saliency map for the corresponding image.

Table 1 shows the quantitative performance of the different optimal selection algorithms, which includes our model MIFN, its variants SMI and ICI, and two state-of-the-art saliency quality assessment methods CS and DSQAN. The numbers with green color represent the best performance within 7 saliency detection algorithm. The numbers with red color represent the best performance within the optimal selection algorithms.

From Table 1, we have following three observations: First, the optimal selection by our MIFN clearly boots the performance of saliency detection. For example, our MIFN outperforms the best single saliency detection method by



(a) DUT-OMRON (b) ECSSD

Fig. 3 The segmentation performances on DUT-OMRON and ECSSD dataset. Green bar represents the best performance within 7 saliency detection algorithms. Blue bars are the performances of existing selection algorithms. Red and yellow bars are the performances of our method and its variants, respectively.

3% in terms of F-measure. Second, our MIFN significantly outperforms the existing stat-of-the-art saliency quality assessment method, which is about 8.8% and 2% better than CS and DSQAN. Third, the information of saliency map and image content are useful and complementary in saliency map selection. Both SMI and ICI improves the performance of saliency detection, which outperform the best single saliency detection method by 0.5% and 1.5%, respectively. Moreover, the combination of SMI and ICI, i.e., MIFN, further boosts the performance of saliency detection, which is about 2.2% and 1.2% better than SMI and ICI.

3.3.2 Salient Object Segmentation

After saliency map selection, we apply a salient object segmentation method [3] to generate the segmentation result from the selected saliency map. We use the standard segmentation metric, mean intersection-over-union (IOU), to compare the segmentation performances of different salient object detection algorithms. The results are shown in Fig. 3. From Fig. 3, we can see that our method MIFN remarkably improves the performance of saliency segmentation. The mean IOU of saliency selection segmentation by using MIFN achieves 66.51%, outperforming the best single saliency detection (i.e., DSS) by 5% on DUT-OMRON. Furthermore, our MIFN also significantly outperforms the existing saliency quality assessment methods CS and DSQAN on both datasets.

4. Conclusion

In this letter, we propose a multi information fusion network for saliency quality assessment, where the formation from the saliency map and image content are all incorporated. Experiments demonstrate that the proposed method achieves state-of-the-art quality prediction accuracy and the significant improvement is gained by exploiting the image content information. As the applications of saliency quality assessment, we apply our method to both the saliency map selection and saliency segmentation. The experimental results prove the effectiveness of the proposed method.

References

[1] Q. Wu, H. Li, F. Meng, K.N. Ngan, B. Luo, C. Huang, and B. Zeng, "Blind image quality assessment based on multichannel feature fusion and label transfer," IEEE Trans. Circuits Syst. Video Technol., vol.26, no.3, pp.425–440, March 2016.

[2] Q. Zhou, J. Cheng, H. Lu, Y. Fan, S. Zhang, X. Wu, B. Zheng, W. Ou, and L.J. Latecki, "Learning adaptive contrast combinations for visual saliency detection," Multimedia Tools and Applications, pp.1–29, 2018.

[3] L. Tang, Q. Wu, W. Li, and Y. Liu, "Deep saliency quality assessment network with joint metric," IEEE Access, vol.6, pp.913–924, 2018.

[4] L. Mai and F. Liu, "Comparing salient object detection results without ground truth," European Conference on Computer Vision, Lecture Notes in Computer Science, vol.8691, pp.76–91, Springer, 2014.

[5] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.3166–3173, 2013.

[6] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.1155–1162, 2013.

[7] Q. Zhou, "Object-based attention: Saliency detection using contrast via background prototypes," Electronics Letters, vol.50, no.14, pp.997–999, 2014.

[8] M.-M. Cheng, N.J. Mitra, X. Huang, P.H.S. Torr, and S.-M. Hu, "Global contrast based salient region detection," IEEE Trans. Pattern Anal. Mach. Intell., vol.37, no.3, pp.569–582, 2015.

[9] Q. Zhou, B. Zheng, W. Zhu, and L.J. Latecki, "Multi-scale context for scene labeling via flexible segmentation graph," Pattern Recognition, vol.59, pp.312–324, 2016.

[10] J. Pan, Y. Yin, J. Xiong, W. Luo, G. Gui, and H. Sari, "Deep learning-based unmanned surveillance systems for observing water levels," IEEE Access, vol.6, pp.73561–73571, 2018.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[12] J. Xiong, X. Long, R. Shi, M. Wang, J. Yang, and G. Gui, "Background error propagation model based RDO in HEVC for surveillance and conference video coding," IEEE Access, vol.6, pp.67206–67216, 2018.

[13] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.5300–5309, 2017.

[14] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," 2017 IEEE International Conference on Computer Vision (ICCV), pp.212–221, 2017.

[15] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," 2013 IEEE International Conference on Computer Vision, pp.2976–2983, Dec 2013.

[16] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing Markov chain," Proc. IEEE International Conference on Computer Vision, pp.1665–1672, 2013.

[17] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp.2814–2821, 2014.

[18] C. Yang, L. Zhang, and H. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," IEEE Signal Process. Lett., vol.20, no.7, pp.637–640, 2013.