LETTER

Learning a Saliency Map for Fixation Prediction*

Linfeng XU^{†a)}, Student Member, Liaoyuan ZENG[†], and Zhengning WANG[†], Nonmembers

SUMMARY In this letter, we use the saliency maps obtained by several bottom-up methods to learn a model to generate a bottom-up saliency map. In order to consider top-down image semantics, we use the high-level features of objectness and background probability to learn a top-down saliency map. The bottom-up map and top-down map are combined through a two-layer structure. Quantitative experiments demonstrate that the proposed method and features are effective to predict human fixation.

key words: saliency detection, human fixation prediction, bottom-up, top-down

1. Introduction

Human beings, as well as most primates have evolved the ability to pay attention to what they are interested in by using saccadic eye-movement which allows the information from complex scenes to be processed efficiently. Usually, it is believed that visual attention involves two main procedures: bottom-up and top-down [1]. Bottom-up attention indicates that the eye fixation is drawn to the places that contain discontinuities in image low-level features, such as color, texture, and brightness. A large amount of previous work concentrates on bottom-up saliency detection [1]–[4]. Top-down attention guides gaze in a task-dependent and goal-driven manner, which is dominated by object information [5]. In the last few decades, extensive research has been conducted to predict visual attention. The results not only help us better understand the mechanism of human cognitive process but also provide powerful tools for many applications such as object tracking [6], image retargeting [7], and automatic object segmentation [8], [9].

In this letter, we focus on learning saliency maps to predict human fixation. A set of low-level and high-level features are proposed. A linear support vector machine is used to train saliency models for the low-level and high-level features. Then the bottom-up and top-down saliency maps are learned by the corresponding model. The maps are combined to obtain the final saliency map through a meaningful two-layer structure. The saliency models, which are trained with different features, are compared to show the perfor-

mance enhancement of our method and the effectiveness of the proposed features.

2. Features Used to Learn Saliency Map

As used in [10], we use a learning based method to train a classifier from the human eye tracking dataset. However, the features used to learn the saliency map are totally dissimilar from [10]. During the learning process, an image is resized to 200×200 and the features for every pixel are computed.

2.1 Low-Level Features

In the field of bottom-up saliency detection, many algorithms have been proposed, which are usually based on various low-level features and attention hypotheses, such as center-surround contrast and visual rarity. Instead of directly using numerous low-level features of each pixel for learning as in [10], which may ignore the correlation of pixels, we use the saliency maps generated by several bottom-up algorithms as the features. In this manner, our model benefits the saliency computation of the existing methods, though the computational cost is increased. The experimental results demonstrate the effectiveness of these features. In our implementation, we choose seven state-of-the-art methods, i.e., IT [1], GBVS [11], SR [12], CA [7], MSS [13], SER [14], and RC [3]. As shown in Fig. 1, the feature vector of each pixel is formed by the saliency values from different methods.

Considering the discrepancies between previous methods in saliency detection, the pixels that are detected to be salient by most of the methods have a higher degree of confidence than the pixels that are considered to be salient by

Fig. 1 The low-level features by using the saliency maps of several bottom-up methods. The feature vector of each pixel is formed by the saliency values from different methods.

Manuscript received March 7, 2013.

Manuscript revised June 5, 2013.

[†]The authors are with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.

*This work was partially supported by NSFC (No. 61179060), and by National High Technology Research and Development Program of China (863 Program, No. 2012AA011503).

a) E-mail: lfxu@uestc.edu.cn

DOI: 10.1587/transinf.E96.D.2294

only a few methods. So we use the mean saliency value of these methods to form another low-level feature:

$$\text{mean}(p) = \frac{1}{N} \sum_{m=1}^N P_m(p) \quad (1)$$

where N denotes the number of methods (here, $N = 7$), $P_m(p)$ represents the saliency probability of pixel p detected by the m th method, which can be described as the normalized saliency value.

2.2 High-Level Features

In some scenes that contain complex backgrounds, the results from different bottom-up saliency algorithms vary significantly from each other and miss the objects of interest. Thus, combining high-level and low-level features to predict fixation seems inevitable to reach human performance.

In [10], Judd et al. consider that humans generally pay attention to pedestrians and faces, so they use the results of person and face detector as the high-level features. Being similar to Chang's work [15] which fuses the generic object template and visual saliency to construct a graphical model, we choose the general object detector to generate the objectness map. Following the prior work on salient object detection [16], we use the low rank matrix recovery to get the background map. The objectness and background maps are used as the high-level features in this letter.

2.2.1 Objectness Map

We use the general object detection methods [17], [18] to generate the probability map of objectness. In the detectors, a lot of windows and the corresponding scores to be an object are returned. For each returned window, a map of objectness is obtained, in which the pixels in the window are assigned the corresponding score and the other pixels are set to zero. All the obtained maps are accumulated and normalized to generate the final probability map of objectness $\bar{O}$, which can be formulated as

$$\bar{O} = \mathcal{N} \left(\sum_n O_n \right) \quad (2)$$

where O_n denotes the objectness map obtained from the n th returned window, $\mathcal{N}$ represents the normalization operation. The value of each pixel p in the map O_n is computed as

$$O_n(p) = \begin{cases} \text{score}_n & \text{if } p \text{ is in the } n\text{th window} \\ 0 & \text{others} \end{cases} \quad (3)$$

where score_n is the returned objectness score corresponding to the n th window. An example of the obtained objectness map is shown in Fig. 2.

2.2.2 Background Map

Based on the recent success of low-rank matrix recovery,

Fig. 2 An example of the objectness map. (a) Original image. (b) The objectness map obtained from [17]. (c) The objectness map obtained from [18].

we generate the probability map of the background which is generally the repeated and homogeneous regions and can be represented by a low-rank matrix.

We first oversegment the image into superpixels using the TurboPixels segmentation algorithm [19]. As described in [16], we extract the 53-dimensional feature vector d_i for the i th superpixel, which includes the color, edge and texture features. The feature vectors of all the superpixels are stacked into a data matrix $\mathbf{D} = [d_1, d_2, \dots]$.

Then, the data matrix $\mathbf{D}$ is decomposed into $\mathbf{L} + \mathbf{E}$, where $\mathbf{L}$ is a low-rank matrix and $\mathbf{E}$ is the related sparse error. The low-rank matrix recovery problem is formulated as

$$(\mathbf{L}^*, \mathbf{E}^*) = \arg \min_{\mathbf{L}, \mathbf{E}} \text{rank}(\mathbf{L}) + \lambda \|\mathbf{E}\|_0 \quad \text{s.t. } \mathbf{D} = \mathbf{L} + \mathbf{E}. \quad (4)$$

Since the above optimization problem is NP-hard, Wright et al. [20] solve the following formulation to make it tractable:

$$(\mathbf{L}^*, \mathbf{E}^*) = \arg \min_{\mathbf{L}, \mathbf{E}} \|\mathbf{L}\|_* + \lambda \|\mathbf{E}\|_1 \quad \text{s.t. } \mathbf{D} = \mathbf{L} + \mathbf{E} \quad (5)$$

where the nuclear norm $\|\mathbf{L}\|_*$ approximates the rank of $\mathbf{L}$, the l_0 -norm $\|\mathbf{E}\|_0$ is replaced by the l_1 -norm $\|\mathbf{E}\|_1$. To solve the optimization problem in (5), we use the method of inexact augmented Lagrange multipliers (ALM) [21] for its computational efficiency.

After performing low rank matrix recovery, the sparse error matrix $\mathbf{E} = [e_1, e_2, \dots]$ is obtained. Since the background is visually homogeneous, a background superpixel should have a low-rank characteristic with small error. We use the l_1 -norm of the vector e_i in the matrix $\mathbf{E}$ to estimate the probability of the superpixel s_i to be background. If $\|e_i\|_1$ is small, s_i is assigned a larger value in the background map B , which is formulated as

$$B(s_i) = \frac{1}{\|e_i\|_1}. \quad (6)$$

Afterwards the map is normalized to a gray-scale image.

2.3 Other Features

Due to the basic rule of photographic composition, the salient objects are naturally placed near the center of the image by most photographers. So the image center is generally important, which is called ‘‘center prior’’. We also take the distance to the image center for each pixel as an important feature as [10] used.

3. Learning Saliency Maps

The saliency maps are learned by the classification model to predict human fixation. We evaluate our method on the MIT dataset [10] that includes 1003 color images as well as the eye fixation data obtained from multiple subjects. Following the method in [10], we divide the set of images into 903 training images and 100 testing images. For each training image 10 positive pixels and 10 negative pixels are randomly chosen from the top 20% and bottom 70% salient locations to generate training samples.

The features of the training set are whiten to have zero mean and unit variance. The same whiten parameters are used to normalize the test data. Then a linear-kernel support vector machine (SVM) [22] is used to predict saliency probabilities for the test images. When the model is trained, the positive samples are labeled as class 1, while the negative samples are labeled as class 0. The instance matrix is constructed by stacking the features of the samples. Using the trained model to predict the pixels in a test image, a saliency map is obtained by using the estimated probability to be class 1. When the bottom-up features are used, a bottom-up saliency map *BU* is obtained. When the top-down features are used, a top-down saliency map *TD* is generated.

The saliency maps of *BU* and *TD* need to be combined to yield the final saliency map in an effective manner. According to the research in vision [23], objects are better predictors of human fixation than early bottom-up saliency. On the other hand, if the top-down saliency is relatively uniform for an image, human attention is attracted by the bottom-up cues, such as high contrast and vivid color. Accordingly we use the two-layer saliency structure in [24] to construct the final saliency from two layers, i.e., basic layer and enhancement layer, which is defined as follows:

1. Basic layer is designed based on the top-down saliency that is obtained from the high-level and center prior features.
2. Enhancement layer is designed based on the bottom-up saliency that is learned from the low-level cues.

Based on the two layers, the final saliency value of a pixel p is defined as

$$S(p) = TD(p)(1 + \rho BU(p)) \quad (7)$$

where $TD(p)$ and $BU(p)$ represent the saliency value of pixel p in the top-down and bottom-up saliency map, respectively. The parameter ρ is a weight factor which adjusts the extent of emphasis on the bottom-up saliency. From our empirical study, good performance can be achieved when ρ is set to the values in the range of [0.8, 1.8]. So in our experiments, the parameter is set as $\rho = 1$. An example of the bottom-up, top-down, and combined saliency maps is shown in Fig. 3.

4. Experimental Results

In this section, we apply the proposed method on the MIT-

Fig. 3 An example of the combined saliency map. (a) Original image. (b) The bottom-up saliency map. (c) The top-down saliency map. (d) The combined saliency map by using the two-layer saliency structure.

Fig. 4 Examples of saliency maps over the MIT-1003 dataset. (a) Original images. (b) Judd's saliency maps [10]. (d) Our saliency maps. (f) Human fixation density maps. (c), (e) and (g) Thresholded maps of Judd's map, our map, and human map, respectively. All the thresholded maps show the top 10% salient locations.

1003 dataset to evaluate the performance of human fixation prediction. The dataset contains 1003 natural images of varying dimensions (the maximal dimension of the width and height is 1024 pixels), along with human fixation data from 15 subjects. The human fixation in the dataset is used as groundtruth to measure the predicting performance. The qualitative and quantitative assessments of the predicting results are reported.

Figure 4 qualitatively shows the comparison results of saliency maps for some test images of the MIT-1003 dataset. The original images are shown in Fig. 4 (a), while the saliency maps obtained by Judd's method [10] and our method are given in Fig. 4 (b) and Fig. 4 (d), respectively. The fixation density maps are shown in Fig. 4 (f), which are generated from the sum of 2D Gaussians corresponding to each fixation point of all the subjects. The thresholded maps of Figs. 4 (b), (d), and (f) are illustrated in Figs. 4 (c), (e), and (g), respectively, in which the top 10% salient locations are shown. The results show that the predicted areas from our maps are more consistent with the human fixation.

In order to evaluate the quality of our features, we perform a quantitative comparison by computing the salient degree between the extracted saliency maps and the human fixation. We threshold the saliency maps at $n = 1, 3, 5, 10, 15, 20, 25$, and 30 percent of the image to obtain binary attention maps. For each binary map, we compute the percentage of human fixations within the salient areas of the map and plot ROC curve to evaluate the performance of fixation prediction. It is seen from Fig. 5 (a) that the mean saliency value feature performs better than most of the other low-level features. Furthermore, the obtained bottom-up saliency maps outperform all the low-level features. Fig-

Fig. 5 The ROC curve for performance evaluation using SVM models trained on each set of features individually and combined together. (a) The comparison between the low-level features and the bottom-up saliency map. (b) The comparison of the high-level features, mean map of all the feature maps, Judd's saliency map, all features without two-layer model, our bottom-up, top-down, and combined saliency map. The human performance and chance are also plotted.

ure 5 (b) shows that our combined saliency maps using the two-layer structure give about 2.4% performance enhancement over Judd's method. If all features are used to train a single classifier without the two-layer model, the result is inferior to the combined map slightly. The top-down saliency plays an important role because the dataset contains many semantic objects, such as people, faces, and text, which are not modeled by bottom-up saliency [10]. In the high-level features, the objectness is effective, while the background map performs not so well. In Fig. 5 (b), if all the feature maps (the background maps are inverted) are simply averaged to get the mean map, it is comparable to the learned map using center feature. However, with the training process and two-level combination, the true positive rate of the combined map reaches about 9% higher than the mean map.

5. Conclusion

In this letter, we train the saliency model to learn a bottom-up saliency map and a top-down saliency map for an image, which are further combined through the two-layer structure to obtain the final saliency map for fixation prediction. To learn the bottom-up saliency map, the low-level features are constructed by using the results of several bottom-up methods, as well as their mean. To learn the top-down saliency map, we use the high-level features of objectness map and background map, as well as the center prior. The objectness map is generated from the scores of the general object detection results. The background map is obtained by low-rank matrix recovery. The experimental results demonstrate that the learned saliency maps are effective to predict human fixation.

References

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual

attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.20, no.11, pp.1254–1259, Nov. 1998.

[2] H. Li and K.N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol.20, no.12, pp.3365–3375, Dec. 2011.

[3] M.M. Cheng, G.X. Zhang, N.J. Mitra, X. Huang, and S.M. Hu, "Global contrast based salient region detection," *Proc. CVPR*, pp.409–416, June 2011.

[4] L. Xu, H. Li, L. Zeng, and K.N. Ngan, "Saliency detection using joint spatial-color constraint and multi-scale segmentation," *J. Vis. Commun. Image R.*, vol.24, no.4, pp.465–476, 2013.

[5] A. Oliva, A. Torralba, M. Castelhana, and J. Henderson, "Top-down control of visual attention in object detection," *Proc. ICIP*, pp.253–256, Sept. 2003.

[6] H. Li and K.N. Ngan, "Automatic video segmentation and tracking for content-based applications," *IEEE Commun. Mag.*, vol.45, no.1, pp.27–33, Jan. 2007.

[7] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *Proc. CVPR*, pp.2376–2383, June 2010.

[8] H. Li and K.N. Ngan, "Learning to extract focused objects from low dof images," *IEEE Trans. Circuits Syst. Video Technol.*, vol.21, no.11, pp.1571–1580, Nov. 2011.

[9] F. Meng, H. Li, G. Liu, and K.N. Ngan, "Object co-segmentation based on shortest path algorithm and saliency model," *IEEE Trans. Multimedia*, vol.14, no.5, pp.1429–1441, Oct. 2012.

[10] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," *Proc. ICCV*, pp.2106–2113, Sept.-Oct. 2009.

[11] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in Neural Information Processing Systems 19*, pp.545–552, MIT Press, Cambridge, MA, 2007.

[12] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," *Proc. CVPR*, June 2007.

[13] R. Achanta and S. Süsstrunk, "Saliency detection using maximum symmetric surround," *Proc. ICIP*, pp.2653–2656, Sept. 2010.

[14] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," *Proc. CVPR*, pp.2368–2375, June 2010.

[15] K. Chang, T. Liu, H. Chen, and S. Lai, "Fusing generic objectness and visual saliency for salient object detection," *Proc. ICCV*, pp.914–921, Nov. 2011.

[16] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," *Proc. CVPR*, pp.853–860, June 2012.

[17] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?," *Proc. CVPR*, pp.73–80, June 2010.

[18] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," *Proc. ICCV*, pp.1052–1059, Nov. 2011.

[19] A. Levinstein, A. Stere, K.N. Kutulakos, D.J. Fleet, and S.J. Dickinson, "Turbopixels: Fast superpixels using geometric flows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.31, no.12, pp.2290–2297, 2009.

[20] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," *Advances in Neural Information Processing Systems 22*, pp.2080–2088, 2009.

[21] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *tech. rep.*, UIUC UILU-ENG-09-2215, Nov. 2009.

[22] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin, "Lib-linear: A library for large linear classification," *J. Mach. Learn. Res.*, vol.9, pp.1871–1874, June 2008.

[23] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vision*, vol.8, no.14, pp.1–26, 2008.

[24] H. Li, L. Xu, and G. Liu, "Two-layer average-to-peak ratio based saliency detection," *Signal Processing: Image Communication*, vol.28, no.1, pp.55–68, Jan. 2013.