A Hierarchical Visual Saliency Model for Character Detection in Natural Scenes

Renwu Gao[1](✉), Faisal Shafait[2], Seiichi Uchida[3], and Yaokai Feng[3]

[1] Information Sciene and Electrical Engineering, Kyushu University, Fukuoka, Japan
kou@human.ait.kyushu-u.ac.jp
[2] The University of Western Australia, Perth, Australia
[3] Kyushu University, Fukuoka, Japan

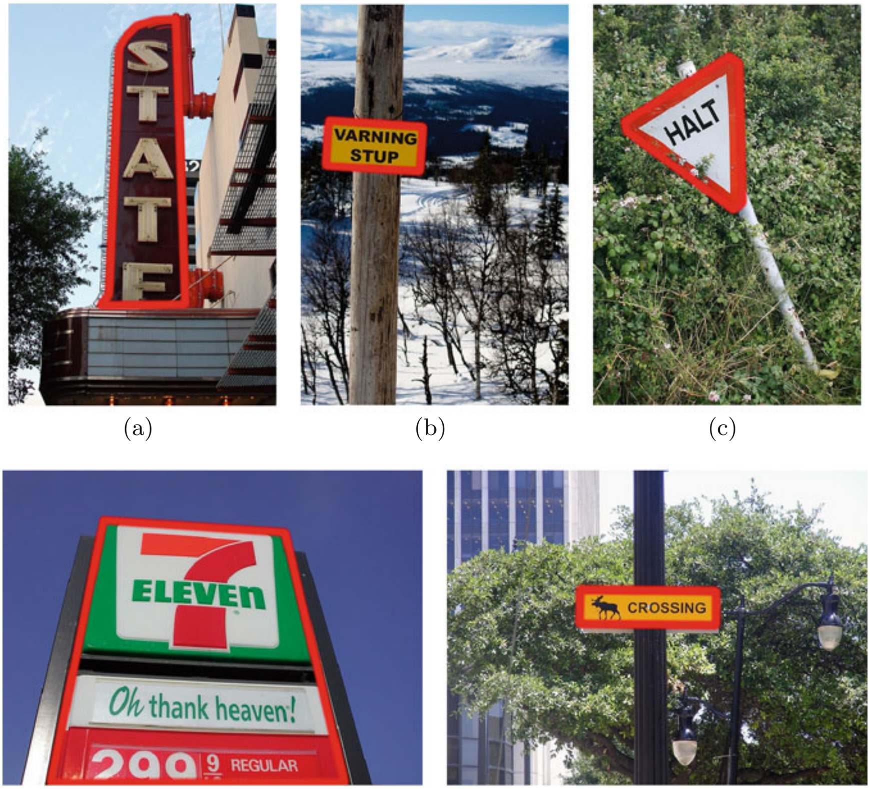
Abstract. Visual saliency models have been introduced to the field of character recognition for detecting characters in natural scenes. Researchers believe that characters have different visual properties from their non-character neighbors, which make them salient. With this assumption, characters should response well to computational models of visual saliency. However in some situations, characters belonging to scene text mignt not be as salient as one might expect. For instance, a signboard is usually very salient but the characters on the signboard might not necessarily be so salient globally. In order to analyze this hypothesis in more depth, we first give a view of how much these background regions, such as sign boards, affect the task of saliency-based character detection in natural scenes. Then we propose a hierarchical-saliency method for detecting characters in natural scenes. Experiments on a dataset with over 3,000 images containing scene text show that when using saliency alone for scene text detection, our proposed hierarchical method is able to capture a larger percentage of text pixels as compared to the conventional single-pass algorithm.

Keywords: Scene character detection · Visual saliency models · Saliency map

1 Introduction

Detection of characters in natural scenes is still a challenging task. One of the reasons is the complicated and unpredictable backgrounds. Another reason is the variety of the character fonts. Many methods have been proposed with the hope of solving the above problems. Coates *et al.* [1] employed a large-scale unsupervised feature learning algorithm to solve the blur, distortion and illumination effects of fonts. Yao *et al.* [2] proposed a two-level classification scheme to solve the arbitrary orientation problem. Mishra *et al.* [3] presented a framework, in which the Conditional Random Field model was used as bottom up cues, and a lexicon-based prior was used as top down cues. Li *et al.* [4] employed adaboost algorithm to combine six types of feature sets. Epshtein *et al.* [5] use the Stroke Width Transform (SWT) feature, which is able to detect characters regardless

M. Iwamura and F. Shafait (Eds.): CBDAR 2013, LNCS 8357, pp. 18–29, 2014.
DOI: 10.1007/978-3-319-05167-3_2,



(a) (b) (c)

(d) (e)

Fig. 1. Examples of salient objects (bounded by red lines) containing characters in natural scenes (Color figure online).

of its scale, direction, font and language. In addition to those recent trials, many methods have been proposed [6].

Some other researchers have tried to employ visual attention models as features [7]. In the recent years, visual attention models have been employed for various object detection/recognition tasks [8–10]. Though the usage of visual attention models for character detection is still under-investigated, their effectiveness has been shown by Shahab *et al.* [11,12] and Uchida *et al.* [13]. Those researchers, who try to employ visual attention models for scene character detection, believe that the characters have different properties compared with their non-character neighbors (*pop-out*). This assumption is acceptable considering that the characters in natural scenes, such as those in Fig. 1 are used to convey "important" information *efficiently* to the passengers.

In some situations, characters are not salient when calculated by saliency-based method; instead, regions in which the characters are written are salient. However, when we only focus on those regions, characters become salient.

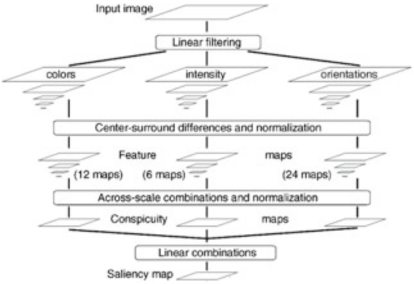


Fig. 2. A general architecture of Itti et. al.'s model. Reference from [14].

In this paper, we investigated how much those regions affect the task of character detection in natural scenes using visual saliency models. We also made a new assumption according to the investigation. The key contribution of this paper is the proposal of a new method for character detection and, compared to the conventional method, the proposed method obtained a better result.

2 Visual Saliency Models

In 1998, Itti *et al.* [14] proposed the first complete implementation and verification of the Koch & Ullman visual saliency model [15]. After that, several kinds of saliency models were proposed [16]. The visual attention models, most of which are directly or indirectly inspired by the mechanism and the neuronal architecture of the primate visual system, are studied to simulate the behavior of human vision [14]. These models provide a massively parallel method for the selection of the intesting objects for the later processing. Visual attention models have been applied to predict where we are focusing in a given scene (an image or a video frame) [17].

2.1 The Calculation of Saliency Map

Many implementations of visual saliency models have been proposed. In this paper, we employ the Itti *et al.*'s model [14] to detect characters in natural scenes. As shown in Fig. 2, three channels (Intensity, Color and Orientation) are used as the low level features [18] to calculate the saliency map as follows:

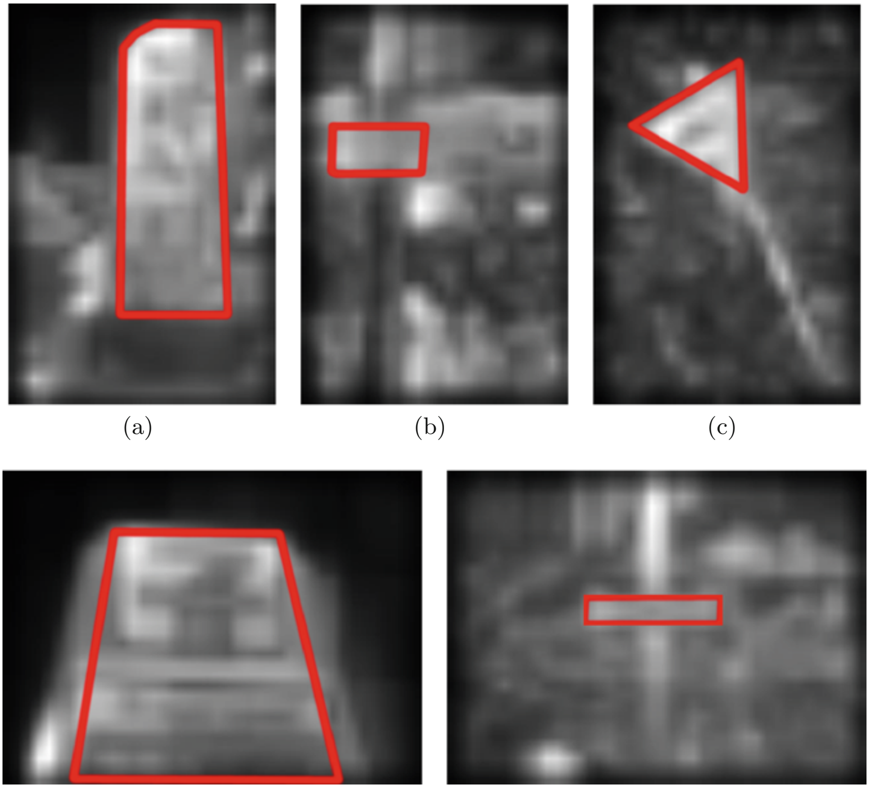
1. Feature maps are calculated for each channel via center-surround differences operation;
2. Three kinds of conspicuity maps are obtained by across-scale combination;
3. The final saliency map is built through combining all of the conspicuity maps.

Figure 3 shows saliency maps of scene images of Fig. 1, by Itti *et al.*'s models. All the visual saliency maps, in this paper, are calculated using Neuromorphic Vision C++ Toolkit (iNVT), which is developed at iLab, USC [19].

2.2 The Problem of Using Saliency Map for Scene Character Detection

From Fig. 3(a), we can find that the characters are salient as we expected. However, in the cases of Fig. 3(b) – (e), pixels belonging to the characters themselves are less salient as we expected; instead, objects (such as the signboards) containing those characters are salient enough to attract our attention.

The examples of Fig. 3 reveal that, in some situation, characters are not salient if we review the whole image; however, when we focus on the signboards, the characters become conspicuous. This means that signboards (or other objects on which the characters are written) are designed to be salient globally (compared to other parts of the image), whereas the characters are designed to be salient locally (compared to their surrounding region).



(a) (b) (c)

(d) (e)

Fig. 3. The corresponding saliency maps of Fig. 1. The surrounding regions are bounded by red lines (Color figure online).

2.3 A Hierarchical-Saliency Model

Based on the above observation, we now have a new assumption: characters are prominent compared to their near non-character neighbors, although they may not be so in a global view of the image. In other words, characters are often *locally salient* inside their possibly *globally salient* surrounding region. For example, characters on a car license number plate may be less prominent than the license number plate when we look at the entire car, but they become prominent if we only look at the number plate.

Correspondingly, a new approach for detecting characters in natural scenes is proposed in this study (called the hierarchical-saliency method) which is briefly introduced below:

- First step (extraction of globally salient region):
  1. A saliency map S is calculated from input image I;
  2. The regions of interest (ROIs) of S are evaluated (the procedure of the evaluation will be provided later) and all pixels are automatically classified into two categories to obtain mask M: the globally salient region (1) and the rest (0);
  3. Multiply the mask M with the input image I to calculate filtered image I';
- Second step (evaluation of local saliency inside the globally salient region): Use I' to obtain a new saliency map S', which is the final map we want.

It is very important to note that though we use the same saliency model to calculate the saliency map in both first and second step, the first saliency value and the second value are different even for the same characters. This is simply because the areas subjected to the model are different.

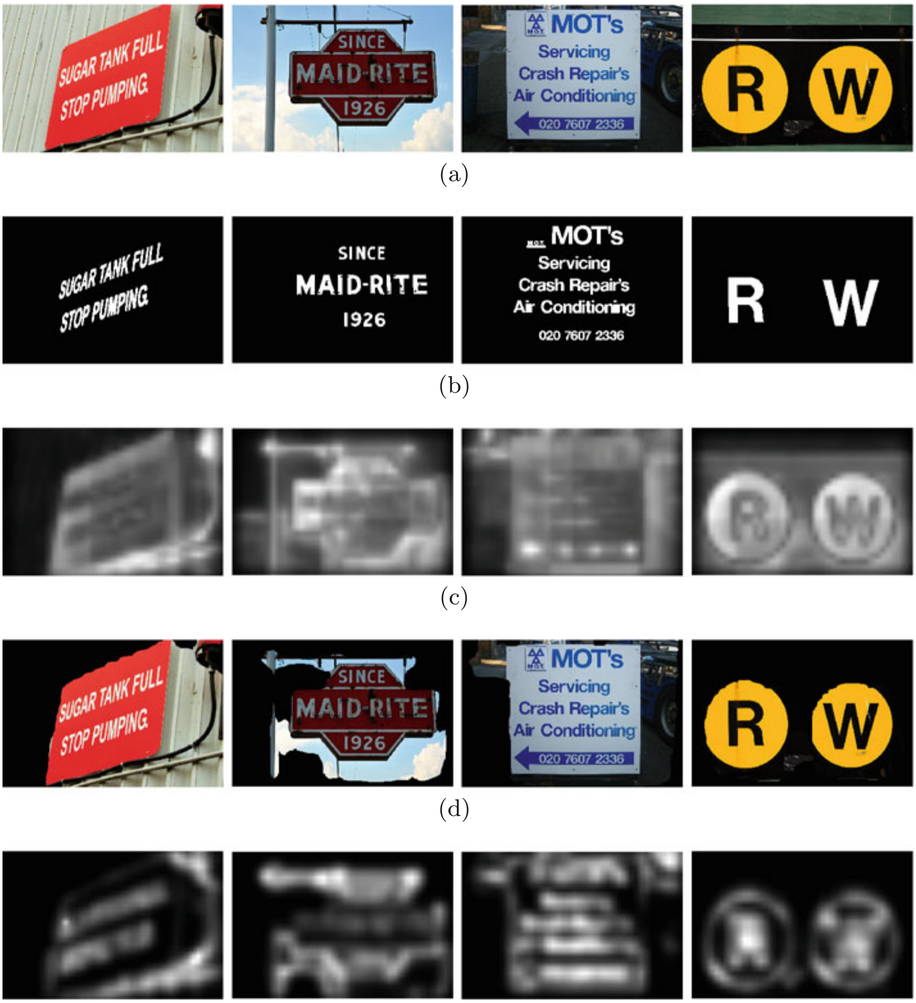
3 Experimental Results

Two experiments were included: (1) in order to investigate how much the salient regions where characters were written affect the task of scene character detection, we firstly arbitrarily selected 101 images from the database and cropped them manually, then calculated the saliency maps for all the 101 images using Itti's saliency model; (2) in order to give a comparison of the performance between the conventional method and the hierarchical-saliency method, we used the whole database with 3,018 images to calculate both the *global* and *local* saliency map, and the salient regions were automatically cropped using Otsu's method and/or Ward's hierarchical clustering method in the process of extracting the ROI regions.

3.1 Database

The scenery image database containing 3,018 images of different sizes has been prepared by our laboratory[1]. All these images were collected from the website

[1] We are planning to make the database freely available in near feature.



(a)

(b)

(c)

(d)

(e)

Fig. 4. (a) input images; (b) ground-truth images; (c) Itti *et al.*'s visual saliency maps, calculated using the whole images (Intensity, Color and Orientation); (d) cropped ROI images calculated with (c); (e) Itti *et al.* 's visual saliency maps, calculated within (d) (Intensity, Color and Orientation) (Color figure online).

"flickr". For each image of our database, pixels of characters were labeled in the corresponding image (ground truth image) and the character information (for example, the bounding-box of the character) was stored into a separate text file.

3.2 Extraction of the ROI

How to extract the ROI (signboards, banners, etc.) from the global saliency map S for calculating the local saliency map is an important problem, because the results of the second step depends on it. In this paper, the Otsu's global thresholding method [20] and the Ward's hierarchical clustering method [21] were employed for a trial (see Fig. 4(d)).

In the Ward's hierarchical clustering method, the error sum of squares (ESS) was given as a loss function F:

$$F = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

where x_i is the score of the ith individual and n donates the number of the individulas in the set. This method reduces n sets to $n-1$ mutually exclusive sets by considering the union of all possible pairs and selecting a union having a minimal value for the loss function F. Assume there are 3 numbers: {1,2,8} and we want to group them into 2 sets. In the Ward's method, all the combinations are considered: {(1,2),(8)}, {(1),(2,8)}, {(1,8),(2)}. Then the loss F are calculated for each combination:

$$\begin{aligned}
F_{\{(1,2),(8)\}} &= F_{\{(1,2)\}} + F_{\{(8)\}} = 0.5 + 0 = 0.5 \\
F_{\{(1),(2,8)\}} &= F_{\{(1)\}} + F_{\{(2,8)\}} = 0 + 18 = 18 \\
F_{\{(1,8),(2)\}} &= F_{\{(1,8)\}} + F_{\{(2)\}} = 24.5 + 0 = 24.5
\end{aligned}$$

The combination which made the minimal value of loss function is selected, so the final result is {(1,2),(8)}. This process is repeated until k groups remain. (please refer to [21] for more details).

3.3 Evaluation Protocol

In the first experiment, we used three low level channels to calculate the saliency map S. While doing the second experiment, in the first step, we also used three low level channels to calculate the saliency map. However, in the second step, saliency map was calculated using different combinations (7 kinds) of channels for each image I, with the purpose of figuring out the best features for character detection. Thresholds t_n ($n \in [0, 255]$) from 0 to 255 were obtained by step 1. Given the corresponding ground truth image I_{GT} with the number of character pixels G_T and the number of non-character pixels G_B, thresholds t_n were applied to evaluate:

1. The number of pixels that matches between saliency map I' (salient pixels) and ground truth I_{GT} (character pixels), $|S_T|$
2. The number of pixels that are salient in the saliency map I', but belong to the non-character regions in the ground-truth image I_{GT}, $|S_B|$.

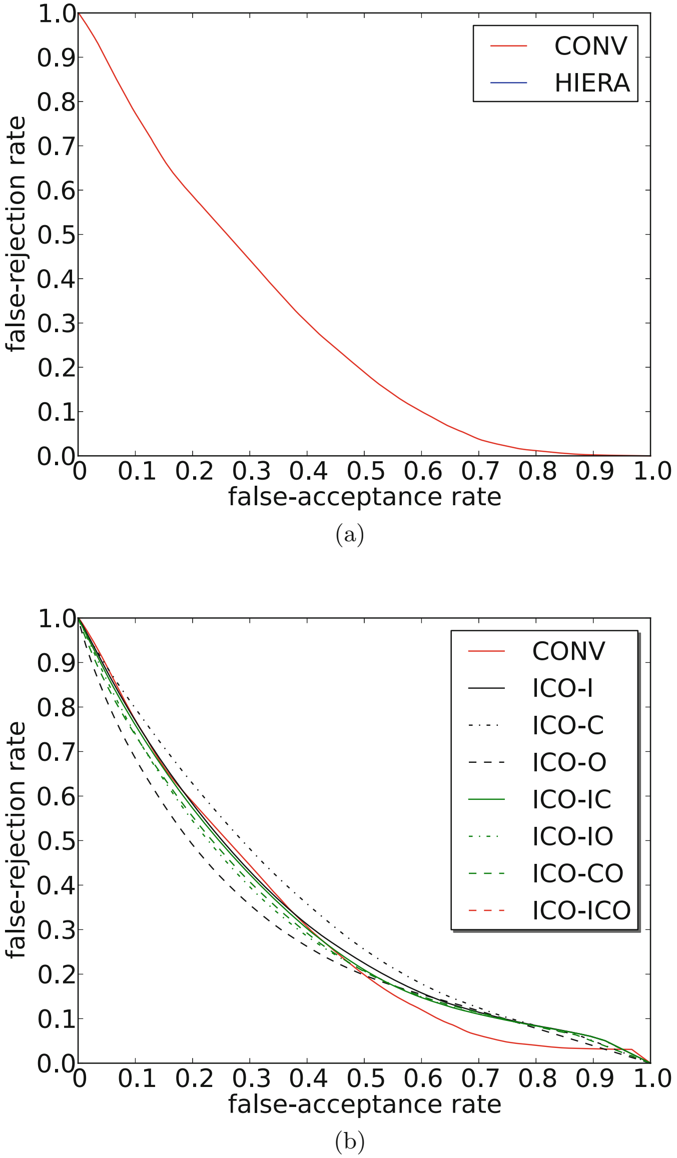


Fig. 5. ROC curves performance comparison. (a) the conventional method (*CONV*) vs. the proposed hierarchical-saliency method with manually cut images; *CONV* represents the conventional method, in which method Itti's model is used only once (first step in our method), *HIERA* represents our proposed method. (b) a comparison between the conventional method and the hierarchical-saliency method with Ward's method cut images (In the second step, we applied all the combination of the low level features); *I / C / O* represent the low level features (*Intensity / Color / Orientation*). Using Otsu's global thresholding method instead of Ward's method gave similar results.

For each threshold, the following performance metrics were calculated:

$$FAR = \frac{|S_B|}{|G_B|} \tag{1}$$

and

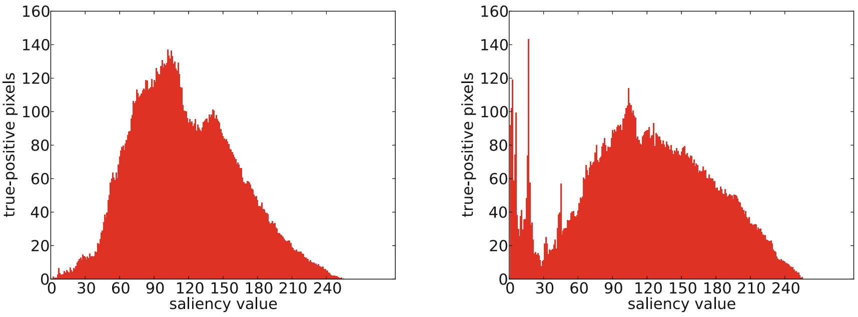
$$FRR = \frac{|G_T| - |S_T|}{|G_T|} \tag{2}$$

Receiver operator characteristic (ROC) curves were employed to evaluate the performance. Figure 5(a) shows the result of comparison. False acceptance rate (FAR) and false rejection rate (FRR) are plotted on the x and the y axis respectively for the range of threshold values. In Fig. 5(a), the closest curve line to the origin represents the best performance algorithm since it has the lowest equal error rate.

3.4 Results and Discussion

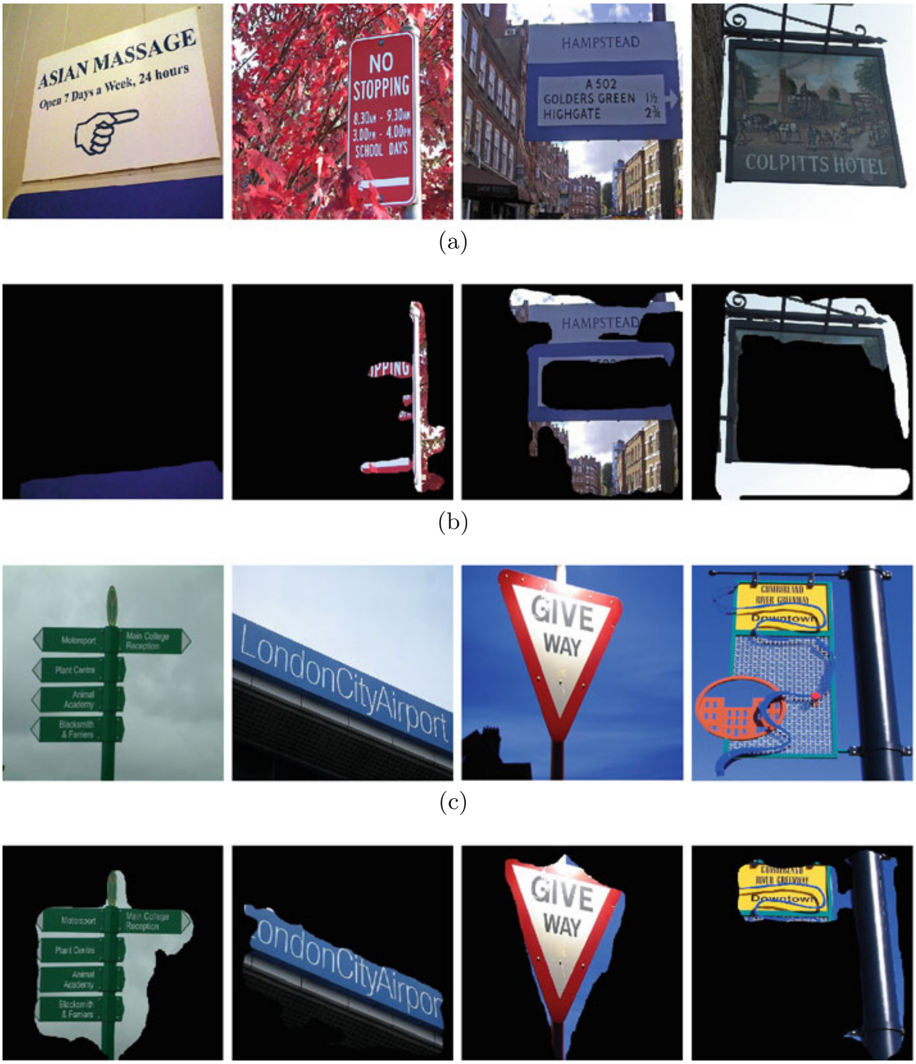
According to Fig. 5(a), we can clearly observe that, compared to the conventional method (using Itti's saliency model once), the proposed hierarchical-saliency method has a better performance. This indicates that the assumption we made in this paper is acceptable. In this section, we give a brief explanation to this. In order to investigate the reason of this result, we built histograms of the true positive pixels for both methods with feature combinations (see Fig. 6). From Fig. 6, we can find that, at the high threshold side, the pixels of characters detected by the hierarchical-saliency method are more salient compared to those detected by the conventional method. On the other hand, the non-salient regions, most of which are non-characters, are suppressed by cropping those salient objects.

From Fig. 5(b) we can see that using orientation as the low level feature in the second step for scene character detection produced the best results. This is



(a) (b)

Fig. 6. Histograms of the true positive pixels where x-axis represents threshold to decide whether pixels belong to character, and y-axis represents the average number of true positive pixels per one image. (a) histogram calculated using conventional method; (b) histogram calculated using the proposed hierarchical-saliency method



(a)

(b)

(c)

(d)

Fig. 7. Examples of the fail and successful images for cropping the surrounding regions. (a) the input images result in the failed results; (b) failed results calculated from (a); (c) the input images result in the successful results; (d) successful results calculated from (c).

mainly because the background in the ROIs is generally simple with few orientation features, whereas the characters have strong orientation features (such as edge features). This makes the characters respond better to the orientation-based visual saliency model and they are easier to detect.

When only using color as the low level feature in the second step, performance became the worst. A possible explanation for this effect is that in natural scenes, the character carriers (the ROIs) are usually designed to be colorful with the purpose of attracting people's attention, which makes them globally salient. As a result, in the procedure of the second step, both the background and the characters respond well to the visual saliency models. Hence characters cannot be distinguished reliably from the background based on saliency alone.

A key issue in our method is how to determine the shape of the salient objects in the first step. We employed Otsu's thresholding algorithm and a simple clustering method (the Ward's hierarchical clustering method) and compared their performance with the conventional method. Though the results of both Otsu' and Ward's method for cropping the salient object were not always good (please refer to Fig. 7 for some successful and failed examples), we still got a better result than the conventional method. It is believable that our method can be used for scene character detection.

4 Conclusion

In this paper, we discussed the problem of applying the Itti *et al.*'s visual saliency model to the task of character detection in the natural scenes, and proposed a new method (called hierarchical-saliency method). We first gave a view of how much the surrounding regions affect character detection, then proposed the Otsu's method and the Ward's hierarchical clustering method to crop the salient objects. In order to investigate the validity of our proposal, we made a performance comparison between the two methods. From the result we can conclude that though the clustering method is not good enough, the hierarchical-saliency method (using orientation feature in the second step) still achieved a better result.

References

1. Coates, A., Carpenter, B., Case, C., Satheesh, S., Suresh, B., Wang, T., Wu, D., Ng, A.: Text detection and character recognition in scene images with unsupervised feature learning. In: International Conference on Document Analysis and Recognition (ICDAR), pp. 440–445 (2011)
2. Yao, C., Bai, X., Liu, W., Tu, Z.: Detection texts of arbitrary orientations in natural images. In: Computer Vision and Pattern Recognition (CVPR), pp. 1083–1090 (2012)
3. Mishra, A., Alahari, K., Jawahar, C.V.: Top-down and bottom-up cues for scene text recognition. In: International Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2687–2694 (2012)
4. Lee, J.J., Lee, P.H., Lee, S.W., Yuille, A., Koch, C.: AdaBoost for text detection in natural scene. In: International Conference on Document Analysis and Recognition (ICDAR), pp. 429–434 (2011)
5. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: Computer Vision and Pattern Recognition (CVPR), pp. 2963–2970 (2010)

6. Uchida, S.: Text localization and recognition in images and video. In: Doerman, D., Tombre, K.(eds.) Handbook of Document Image Processing and Recognition (to be published in 2013)
7. Sun, Q.Y., Lu, Y., Sun, S.L.: A visual attention based approach to text extraction. In: International Conference on Pattern Recognition (ICPR), pp. 3991–3995 (2010)
8. Walther, D., Itti, L., Riesenhuber, M., Poggio, T.A., Koch, Ch.: Attentional selection for object recognition - a gentle way. In: Bülthoff, H.H., Lee, S.-W., Poggio, T.A., Wallraven, Ch. (eds.) BMCV 2002. LNCS, vol. 2525, pp. 472–479. Springer, Heidelberg (2002)
9. Elazary, L., Itti, L.: A Bayesian model for efficient visual search and recognition. Vision. Res. **50**(14), 1338–1352 (2010)
10. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: International Conference on Computer Vision (ICCV), vol. 1, pp. 273–280 (2003)
11. Shahab, A., Shafait, F., Dengel, A.: Bayesian approach to photo time-stamp recognition, In: International Conference on Document Analysis and Recognition (ICDAR), pp. 1039–1043 (2011)
12. Shahab, A., Shafait, F., Dengel, A., Uchida, S.: How salient is scene text?. In: International Workshop on Document Analysis Systems (DAS), pp. 317–321 (2012)
13. Uchida, S., Shigeyoshi, Y., Kunishige, Y., Feng, Y.K.: A keypoint-based approach toward scenery character detection. In: International Conference on Document Analysis and Recognition (ICDAR), pp. 918–823 (2011)
14. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **20**(11), 1254–1259 (1998)
15. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiol. **4**, 219–227 (1985)
16. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **35**(1), 185–207 (2013)
17. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: International Conference on Computer Vision, Kyoto, Japan, pp. 2016–2113 (2009)
18. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cogn. Psychol. **12**(1), 97–136 (1980)
19. http://ilab.usc.edu/toolkit
20. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Sys. Man Cybern. **9**(1), 62–66 (1979)
21. Ward Jr, J.H.: Hierarchical grouping to optimize an object function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)

Springer

http://www.springer.com/978-3-319-05166-6

Camera-Based Document Analysis and Recognition
5th International Workshop, CBDAR 2013, Washington, DC,
USA, August 23, 2013, Revised Selected Papers
Iwamura, M.; Shafait, F. (Eds.)
2014, VIII, 187 p. 117 illus., Softcover
ISBN: 978-3-319-05166-6