Scenery Character Detection with Environmental Context

Yasuhiro Kunishige, Feng Yaokai, and Seiichi Uchida
Kyushu University, Fukuoka, 819-0395, Japan

***Abstract*—For scenery character detection, we introduce environmental context, which is modeled by scene components, such as sky and building. Environmental context is expected to regulate the probability of character existence at a specific region in a scenery image. For example, if a region looks like a part of a building, the region has a higher probability than another region like a part of the sky. In this paper, environmental context is represented by state-of-the-art texture and color features and utilized in two different ways. Through experimental results, it was clearly shown that the environmental context has an effect of improving detection accuracy.**

***Keywords*-scenery character detection, environmental context, feature, random forest**

I. Introduction

The main contribution of this paper is to show the usefulness of *environmental context*, which is modeled by scene components, such as sky, ground, building, etc., for scenery character detection. Consider a small region (a block or a connected component (CC)) in a scenery image. If the region (and its surroundings) look like a part of the sky, we can say that a character existence probability at this region is low. This is because characters never exist in the sky. This simple example indicates that environmental context is useful for detecting scenery characters.

Past trials on scenery character detection (e.g., [1]–[5]) have never utilized environmental context. They tried to employ various features just for representing character areas. Unfortunately, the past trials suffer from many false detections due to complicated non-character areas. One naive strategy to suppress the false detections is to apply more strict conditions to the detectors; however, this strategy will result in the situation that various characters (such as decorated characters) are overlooked. To summarize, there is a severe trade-off when we try to detect characters only by character features. Our trials will utilize not only features for representing characters but also features for representing environmental context, in order to relax the trade-off.

Among several methodologies of utilizing environmental context in scenery character detection, we will examine two methods, later called **Method 2** and **Method 3**. The former is a rather straightforward method where environmental context features are concatenated with character features and then a character/non-character discrimination is performed. The latter is a more elaborated method where the likelihoods of all scene component categories (such as "sky") are first calculated and the discrimination is performed using the likelihoods as features. Experimental results will show the better performance of those methods over the simple method without environmental context, which is called **Method 1** later.

The remainder of this paper is organized as follows. After a brief description of the preprocessing step in Section II, the extraction of not only character features but also environmental features is described in Section III. Then, the three character/non-character discrimination methods using those features are described in Section IV. In Section V, the performance of character detection are evaluated qualitatively and quantitatively through experiments using real ground-truthed scenery images. Finally, Section VI draws our conclusion and future work.

II. Preprocessing

In this paper, scenery character detection is formulated as a character/non-character discrimination problem at each connected component (CC). Accordingly, the first-step is the decomposition of the entire image into non-overlapping CCs. CC is a better unit for our character detection problem than a fixed-size block because the size of a character (and environmental context) varies largely in the target scenery image and any fixed-size block cannot deal with this large variation.

For the CC decomposition, trinarization by Niblack's method [6] is used. Its two thresholds are determined at each pixel by local mean and standard deviation values. Precisely speaking, we use an improved version [3], [7] of the Niblack's original method. This improvement is effective to increase the robustness against noises and shadings around characters. Note that very small CCs (less than 5 pixel size) are always treated as non-characters because they are unreadable even if they are actually characters.

III. Feature extraction

A. Extraction of Character Features

Many researcher have proposed their own character features which are expected to be suitable for evaluating the likelihood that a target CC (or block, or pixel) is a character. In this paper, by referring very recent trials [3], [4], the following twelve features (C1-12) are extracted for each CC: the ratio between areas of the CC and the entire image, the ratio between width of the CC and the entire image, the aspect ratio of the CC, contour roughness, the number of holes, the ratio between area and squared contour length of the CC. the ratio between areas of the CC and its bounding

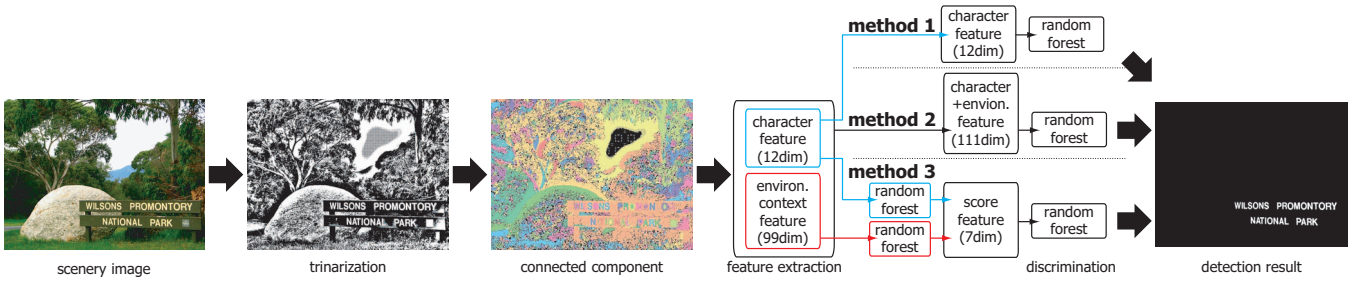


Figure 1. Overview of the proposed method. There are three different methods to be compared.

box, mean of stroke width, standard deviation of stroke width, the ratio between areas of a dilated CC and the entire image, the ratio between contour lengths of original and the dilated CCs, and edge contrast.

B. Extraction of Environmental Context Features

Again, the main purpose is to introduce environmental context into the character detection problem. For this purpose, we introduce 99 environmental context features of each CC.

Among them, 84 features are *texton features* [8] of the target CC. Texton features have been recently proposed for describing textures. In [9], [10], it was proved that they can provide a good performance of environmental objects, such as sky, road, green, etc. Texton features are based on 21 responses from 9 Gaussian filters , 4 Laplacian of Gaussian (LoG) filter, and 8 Gaussian derivative filters. Then for each response the mean, standard derivation, Kurtosis, and skewness within the CC are calculated. ($84 = (9+4+8)\times 4$.) Hereafter, the features from the Gaussian filters, the LoG filter, and the Gaussian derivative filters are denoted as E1-36, E37-52, E53-84, respectively. It is important to note that on the above calculation, 10-pixel margin is added around the CC. By this margin, we can incorporate environmental context around the target CC.

The remaining 15 features are employed to represent non-texture characteristics of environmental context. Specifically, 12 features are L*a*b* color features, 2 features are positional features, and 1 feature is an area feature. The color features are extracted as the mean, standard derivation, Kurtosis, and skewness for each of L* (E85-88), a* (E89-92), and b* (E93-96) values of the target CC. The positional features (E97,98) are the x and y-coordinates of gravity point of the CC. The area feature (E99) is the number of pixels of CC. Note that the positional features are employed because they can represent an environmental context [10]; for example, sky will be generally located in an upper part of the image.

IV. Character/Non-Character Discrimination with Environmental Context

A. Random Forest

The three discrimination methods (**Methods 1∼3**) commonly employ random forest [11]. As shown in Fig. 2,

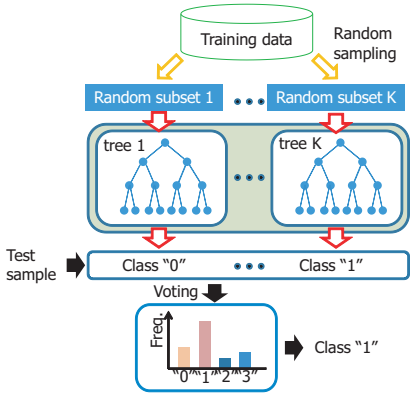


Figure 2. Random forest.

random forest is an ensemble of K classifiers. Each classifier is a decision tree and trained by a random subset of whole training data. During the training of each decision tree, effective features are selected automatically. The final classification result is determined by the majority voting of K decision results. In this paper, K and the depth of each tree are fixed at 500 and 10, respectively.

There are several merits of random forest for our detection task. First, because of its random sampling strategy and ensemble framework, random forest is robust to outliers in training data. Since there are very special character shapes in scenery images, overfitting will become serious without this robustness. Second, its feature selection mechanism is suitable for the task because we do not know useful features in advance. Third, we can investigate the "importance" of each feature on the discrimination by random forest. (In Section V, we will observe the importance of each feature introduced in Section III.) Other more general merits are its fast computation and multi-class recognition ability.

B. Three Discrimination Methodologies

Method 1 uses only the character features (C1-12). A random forest is trained with those features, that is, in the 12-dimensional feature space. If "votes for character" are more than "votes for non-character" among K =500 votes, the target CC is detected as a character CC. Detection results by Method 1 will indicate an up-to-date accuracy of scenery character detection, where environmental context is not considered. In other words, we can consider Method 1

Table I
DETECTION RATES. PARENTHESIZED VALUES ARE PIXEL-WISE DETECTION RATES.

	Recall	Precision	F-value
Method 1	54.6 (67.8)	54.0 (63.1)	54.3 (65.4)
Method 2	62.2 (73.3)	64.8 (58.3)	63.5 (64.9)
Method 3	65.6 (75.3)	66.2 (70.9)	65.9 (73.0)

as one of the best past trials.

Method 2 uses not only the character features (C1-12) but also the environmental context features (E1-99). A random forest is trained in 111-dimensional feature space. Detection results by Method 2 will indicate whether the environmental context features are useful.

Method 3 is the most elaborated for utilizing environmental contexts more explicitly. The likelihoods of several scene component categories are first calculated and then used as feature values for the discrimination. Precisely, Method 3 is organized in the following two-step manner.

1) As the first step, we calculate seven *score features*, which are comprised of one character score feature and six scene component score features. The former is derived as the number of "votes for character" in the random forest of Method 1. The latter six features are derived by another random forest trained to classify the CC into one of six scene component categories, "sky", "green", "sign(board)", "ground", "building", and "others". For example, the sky score feature is derived as the number of votes for the sky category in the random forest.
2) As the second step, the final character/non-character discrimination is done by a single random forest and the above seven score features.

V. EXPERIMENTAL RESULTS

A. Dataset

A scenery image dataset were prepared for our experiments. Using Google Image Search™, top 300 photo images (each of which contains some characters and has a size around 640×480) were first collected. The keywords used in the search were "park" and "sign." Those 300 images were then decomposed into a training dataset (150 images) and a test dataset (150 images).

For each image, a ground-truth was attached manually. Specifically, for each CC, character label or non-character label was assigned by a human operator. For a CC which contains both character and non-character regions, non-character label was assigned. For Method 3, the scene component category was also assigned to the CC. Figure 3 shows an example of the ground-truth of environmental context.

Figure 3. Ground-truth of scene component category.

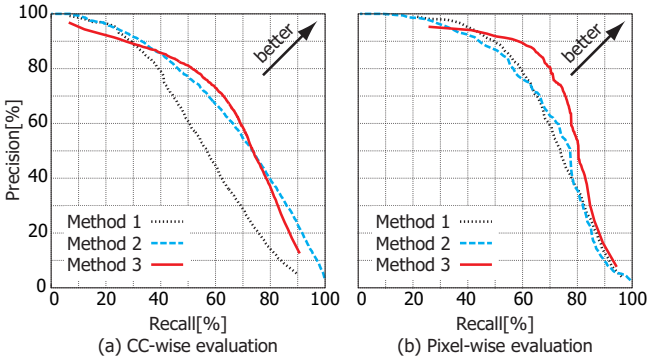


Figure 4. Detection accuracy.

B. Quantitative Evaluation

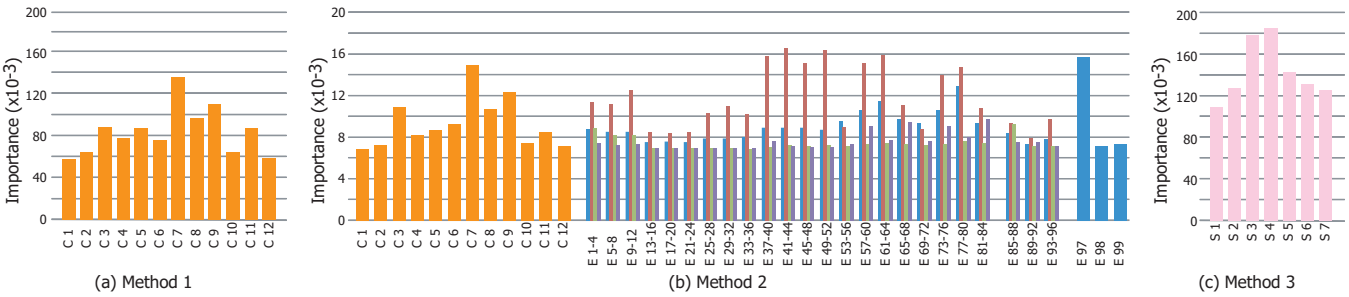
Table I and Fig. 4 show the detection accuracy of the three methods. Figure 4 is an ROC curve showing the change of recall and precision according to a parameter k ($1 \leq k \leq K = 500$) of the random forest; if the number of votes for "character" (against "non-character") exceeds k, the target CC is detected as a character region. The evaluation has been done in two ways; CC-wise evaluation and pixel-wise evaluation.

The results by the CC-wise evaluation clearly show the superiority of Methods 2 and 3 over Method 1. Again, Method 1 is one of the best conventional detection methods, where the environmental context is not used. Thus, the superiority confirms that the environmental context is very useful to scenery character detection.

Methods 2 and 3 have no significant difference by CC-wise evaluation. This fact is interesting at the following point; Method 3 uses 7-dimensional score features and its final decision totally relies on these seven features. This indicates that score features, the likelihoods of the main scene components, are very compact and sufficient representations of the environment for scenery character detection.

The pixel-wise evaluation showed that Method 3 is better than Method 2; this indicates that Method 2 failed at some larger CCs. A typical example of this failure will be shown in Section V-C. This is also the reason why Method 2 could not outperform Method 1 by the pixel-wise evaluation as shown in Fig. 4(b).

Figure 5 shows an importance for each feature at the character/non-character discrimination by random forest. The importance was evaluated by the "Out of Bag" method, where the feature to be evaluated is replaced by another



(a) Method 1 (b) Method 2 (c) Method 3

Figure 5. Importance of each feature on discrimination by random forest.

Original Method 1 Method 2 Method 3

Figure 6. Detection results.

Figure 7. Visualization of score feature value. From left to right, original image, and, character, sky, green, sign(board), ground, and building features.

feature and the accuracy degradation by the replacement is measured. If a larger degradation is observed, the feature is more important.

Figure 5 (a) shows that no character feature has very low importance and the ratio between areas of the CC and its bounding box feature (C7) has the top importance among the 11 character features. A more interesting fact is that as shown in Fig. 5 (b), both of the character features and the environmental context features are important in Method 2. This fact supports that Method 2 utilizes the environmental context features for better detection accuracy than Method 1. Figure 5 (c) shows that Method 3 utilizes all of the seven score features and especially, green (S3) and signboard (S4) features are important for scenery character detection.

C. Qualitative Evaluation

Figure 6 shows character detection results by Methods 1∼3 on four scenery images. In Fig. 6 (a) and (b) , Method 1 produced many false detections around sky and green regions. Those false detections decrease drastically by using environmental context.

Figure 7 (a) is a visualization of the score feature of Fig. 6 (b). The false detections by Method 1 are found in the sky region (Fig. 6 (b)) and the region has high sky feature values (Fig. 7 (a)). This fact indicates that the high sky score feature values could suppress the false detection and thus environmental context is useful to improve character detection performance.

Figure 6 (c) is an example that environmental context could not improve the result. In this scenery image, thin grasses are overlapped on a signboard. Accordingly, most CCs of characters are severely broken and far different from CCs of normal characters. In the present framework, if a CC of a character is broken, it is difficult to recognize it as a character correctly, even with environmental context features.

Figure 6 (d) is an example showing a difference between Methods 2 and 3. Both methods could successfully remove false detections in the green region, Method 2, however, wrongly detect the CC of a signboard region as a (large) character. One reason of this false detection is existence of the edges caused by the characters on the signboard. Method 2 was badly affected by these edges. In contrast, as shown in Fig. 7 (b), Method 3 could give larger signboard feature values correctly for this signboard region and finally give a correct detection result as a signboard.

VI. Conclusion

Environmental context, which is modeled by scene components, was utilized for scenery character detection. The environmental context was represented by many features, which are mainly of texture features and color features, and utilized in two different methods. One method utilized the environmental context features directly in character/non-character discrimination. The other utilized them as scores showing the likelihoods with respect to environmental components, such as "sky" and "green". Both methods could show clear superiority over the detection method without environmental features in experimental results.

Future work will focus on the following points. (i) A better CC segmentation method is necessary since it affects detection performance severely. (ii) The environmental context features of CCs around the target CC should be utilized. (iii) Another kind of context features, such as geometric context [12] (which can estimate flat areas) and visual saliency [13], will be useful because areas with higher flatness and/or saliency will have a higher probability of existing characters.

Acknowledgment

The authors greatly thank Prof. K. Kise and Prof. M. Iwamura (Osaka Pref. Univ., Japan) and Prof. S. Omachi (Tohoku Univ., Japan) for their precious comments. This research was partially supported by JST, CREST.

References

[1] X. Chen, and A. L. Yuille, "Detecting and reading text in natural scenes," *CVPR*, 2004.

[2] S. Lucas, et al., "ICDAR 2003 robust reading competitions: entries, results, and future directions," *IJDAR*, 2005.

[3] K. Zhu, F. Qi, R. Jiang, and L. Xu, "Automatic character detection and segmentation in natural scene images," *J. Zhejiang University, Science A*, vol. 8, No. 1, 2007.

[4] L. Xu, H. Nagayoshi, and H. Sako, "Kanji character detection from complex real scene images based on character properties," *DAS*, 2008.

[5] H. Goto, "Redefining the DCT-based feature for scene text detection," *IJDAR*, 2008.

[6] W. Niblack, *An Introduction to Digital Image Processing,* Prentice Hall, 1986.

[7] L. L. Winger, J. A. Robinson, and M. E. Jernigan, "Low-complexity character extraction in low-contrast scene images," *IJPRAI*, 2000.

[8] T. Leung, and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *IJCV*, 2001.

[9] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," *ECCV*, 2006.

[10] Y. Kang, K. Yamaguchi, T. Naito, and Y. Ninomiya, "Road scene labeling using SfM module and 3D bag of texton," *ICCV-3dRR*, 2009.

[11] L. Breiman, "Random forests," *Machine Learning*, 2001.

[12] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," *ICCV*, 2005.

[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *PAMI*, 1998.