Japanese Scene Character Recognition using Random Image Feature and Ensemble Scheme

Fuma Horie[1] and Hideaki Goto[2]
[1]*Graduate School of Information Sciences, Tohoku University, Sendai, Japan*
[2]*Cyberscience Center, Tohoku University, Sendai, Japan*

Keywords: Random Image Feature, Japanese Scene Character Recognition, Synthetic Scene Character Data, Ensemble Voting Classifier, Multi-Layer Perceptron.

Abstract: Scene character recognition is challenging and difficult owing to various environmental factors at image capturing and complex design of characters. Japanese character recognition requires a large number of scene character images for training since thousands of character classes exist in the language. In order to enhance the Japanese scene character recognition, we utilized a data augmentation method and an ensemble scheme in our previous work. In this paper, Random Image Feature (RI-Feature) method is newly proposed for improving the ensemble learning. Experimental results show that the accuracy has been improved from 65.57% to 78.50% by adding the RI-Feature method to the ensemble learning. It is also shown that HOG feature outperforms CNN in the Japanese scene character recognition.

1 INTRODUCTION

Recognition of text information in the scene, which is often referred to as scene character recognition, has some important applications such as automatic driving system and automatic translation. Scene character recognition is more difficult in comparison with printed character recognition as there are various factors such as rotation, geometric distortion, uncontrolled lighting, blur, noise and complex design of characters in the scene images. Japanese scene character recognition requires a large number of training data since thousands of character classes exist in the language. However, collecting a large number character image samples in real scenes is a hard task.

Some previous researches introduced a data augmentation method using Synthetic Scene character Data (SSD) which is randomly generated by some particular algorithms such as filter processing, morphology operation, color change, and geometric distortion from the font sets of printed characters (Jaderberg et al., 2014)(Ren et al., 2016)(Jiang and Goto, 2017)(Horie and Goto, 2018). Jader et al. and Ren et al. have shown that the accuracy of the deep neural network model can be improved by adding SSD to the training data. It has been proved that the augmentation methods are effective for improving the accuracy of the scene character recognition. Figure 1 shows some examples of the Japanese characters in natural scenes. In our previous work (Jiang and Goto, 2017)(Horie and Goto, 2018), we developed a training datasets consisting of both Real Scene character Data (RSD) and SSD. The ensemble scheme is used to improve the generalization ability of the classifier.

For further improvements of the generalization ability, Random Image Feature (RI-Feature) method is newly proposed in this paper. The RI-Feature method is to randomly process an image before extracting character features and it is applied to each classifier by different parameters. It is expected that the RI-Feature method will make the generalization ability higher. Moreover, we propose a new ensemble scheme using Multi-Layer Perceptron (MLP) in this paper. Experimental results show the effectiveness of RI-Feature method and MLP.

Convolutional Neural Network (CNN) has achieved a remarkable performance in various image recognition tasks including also the scene character recognition. However, the CNN needs a large number of high-quality training data. It is thought that CNN is not able to achieve high accuracy in the scene character recognition when it suffers from the shortage of training data. Especially, Japanese scene character datasets currently available are far from enough to train the CNN. Some previous scene character recognition systems use HOG feature since it has

Horie, F. and Goto, H.
Japanese Scene Character Recognition using Random Image Feature and Ensemble Scheme.
DOI: 10.5220/0007341904140420
In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, pages 414-420
ISBN: 978-989-758-351-3

been found that the HOG feature outperforms the other features (SIFT, DAISY, SURF, etc.) (Tian et al., 2016). In this paper, we compare the performances of CNN and HOG by some experiments using SSD.

This paper is organized as follows. Section II describes the ensemble scheme and RI-Feature method. Section III shows the process of experiments and the results. Conclusions and future work are given in section IV.

2 JAPANESE SCENE CHARACTER RECOGNITION USING SSD AND ENSEMBLE SCHEME

2.1 Flow of the Recognition System

SSD and ensemble scheme were utilized in our previous work (Horie and Goto, 2018). The new system proposed in this paper is an extended version of the previous system, and Random Image Feature (RI-Feature) method is newly introduced. Figure 2 shows the flow of our recognition system. Let T be the number of classifiers. T subsets are created from the original font dataset by the bootstrap sampling (Breiman, 1996). Each subset consists of K images. These subsets are converted to an SSD set. RI-Feature sets are extracted from the generated SSD. Finally, T classifiers are created by learning the RI-Feature sets. At the recognition stage, T RI-Features are extracted from a query image, and each RI-Feature is put into each classifier. The answers obtained from every classifier are combined by plurality voting.

2.2 Synthetic Scene Character Generator

Synthetic Scene character Data (SSD) is used in order to increase the training data and to enhance the recognition system in this paper. Some previous researches have shown the effectiveness of SSD for the scene character recognition (Jaderberg et al., 2014)(Ren et al., 2016)(Jiang and Goto, 2017)(Horie and Goto, 2018). SSD is generated through some processes such as distortion, color change, morphology operation, background blending and various filters.

The SSD sets are created from image subsets. The i-th subset S_i is sampled from an original image set S_o by the bootstrap method. Normaly,

$$S_i = \{s_{ij}\},\ \ s_{ij} \in S_o = \{s_{oj}\},$$

where s_{ij} is randomly sampled with replacement.

The SSD are generated by the following processing in this paper.

- Affine Transformation.
  The process is defined by the following matrix:

$$\begin{bmatrix} 1 & 0 & C_x \\ 0 & 1 & C_y \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_1 & a_2 & 0 \\ a_3 & a_4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -C_x \\ 0 & 1 & -C_y \\ 0 & 0 & 1 \end{bmatrix},$$

$$a_1, a_3 \in [0.9, 1.1],\ a_2, a_4 \in [-0.1, 0.1],$$

  where (C_x, C_y) is the center coordinate of the image. a_1, a_2, a_3 and a_4 are chosen from uniformly-random numbers.

- Gaussian Filter.
  3×3 matrices are used as the kernels of Gaussian filter, and they are defined by the following formula:

$$K(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right),\ \ \sigma \in [0, 10]. \tag{1}$$

  σ is chosen with uniformly-random numbers.

- Morphology Operation.
  3×3 matrices are used as kernels of morphology operation. The operation mode is selected randomly from dilation, erosion, and none.

- Color Change.
  We chose 20 colors frequently appeared in various scene character images. Two colors are selected randomly as the background and foreground. The channel intensities of the output image are calculated by the following formula;

$$\begin{aligned} L'_R(i,\ j) &= \left[L(i,\ j) \times \frac{R_f - R_b}{255} + R_b + 0.5\right], \\ L'_G(i,\ j) &= \left[L(i,\ j) \times \frac{G_f - G_b}{255} + G_b + 0.5\right], \\ L'_B(i,\ j) &= \left[L(i,\ j) \times \frac{B_f - B_b}{255} + B_b + 0.5\right], \end{aligned} \tag{2}$$

  where $L(i,\ j)$ is the character image before processing, $L'_R(i,\ j)$, $L'_G(i,\ j)$, and $L'_B(i,\ j)$ are the matrices representing the processed images, R_f, G_f, and B_f are the foreground colors, R_b, G_b, and B_b are the background colors, respectively.

- Random Filter (RF).
  Random filter is proposed in our previous paper (Horie and Goto, 2018). The kernels of random filter are defined by the following formula;

$$K = (k_{n,n})_{1 \le n \le 3},\ k_{n,n} \in \mathbb{R},\ \sum_{m=1}^{3}\sum_{n=1}^{3} k_{m,n} = 1, \tag{3}$$

  where $k_{n,n}$ is randomly selected.

Figure 1: Example of Japanese scene characters.

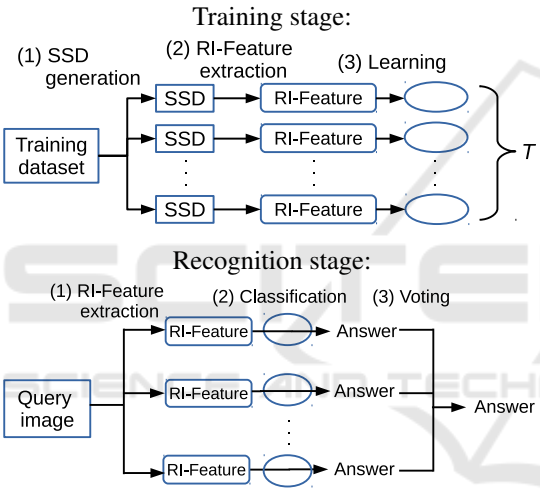


Figure 2: Flow of ensemble scheme.

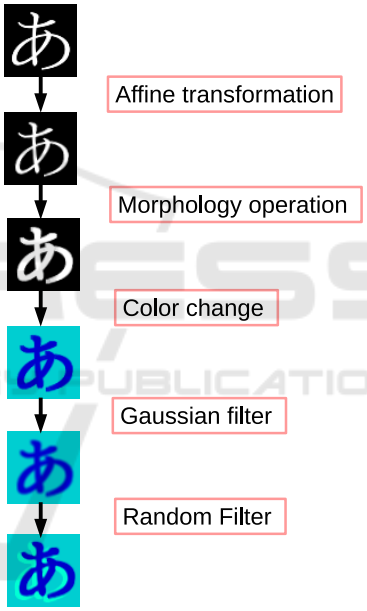


Figure 3: Flow of SSD generation.

Each process is applied to each image using different parameters. Normally, all SSD are different with each other.

Figure 3 shows the flow of the above SSD generation. Figure 4 shows some examples of SSD generated by the above processing.

2.3 Random Image Feature

Ensemble scheme has been used to improve the generalization ability of classifiers in some previous work (Jiang and Goto, 2017)(Horie and Goto, 2018). We have shown that the ensemble scheme effectively improves the accuracy of scene character recognition. For further improvement of the recognition accuracy, we propose RI-Feature method.

RI-Feature method applies some random image processing to the character image before extracting the character features. The random processing is applied for each classifier using different parameters. Let D_i be the i-th SSD set. The processing is as follows.

- Multi-Scale Resizing (MSR).
  Images of D_i are resized to the following size;

$$s_i = \begin{cases} 16 \ (1 \leq i \leq \frac{T}{3}) \\ 32 \ (\frac{T}{3} \leq i \leq \frac{2T}{3}) \\ 64 \ (\frac{2T}{3} \leq i \leq T) \end{cases}, \quad (4)$$

  where T is the number of classifiers.

  MSR was proposed in the previous paper(Jiang and Goto, 2017). It was demonstrated that MSR

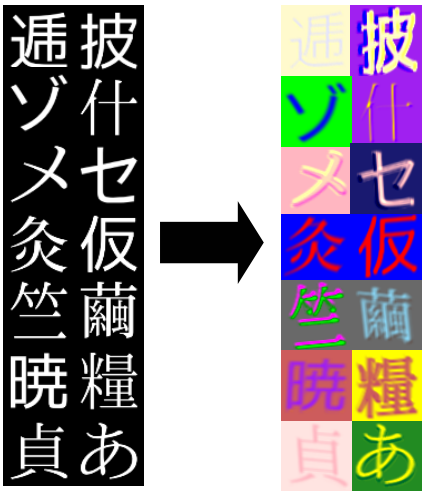


Figure 4: Examples of SSD used in the training.

effectively improve the ensemble recognition accuracy.

- Random Filter (RF).
  Images of D_i are calculated by RF using the following kernel;

$$K_i = (k_{in,n})_{1\leq n\leq 3},\ k_{in,n} \in \mathbb{R},$$
$$\text{where } \sum_{m=1}^{3}\sum_{n=1}^{3} k_{im,n} = 1. \tag{5}$$

  RF is used more than once and by combined with Mean Filter (MF). We expect that RF is useful for adding some variations to the features as it introduces various effects to the image. MF is expected to simulate image blur and also to reduce image noise.

- Random Affine (RA).
  Images of D_i are deformed by affine transformation using the following matrix;

$$\begin{bmatrix} 1 & 0 & C_x \\ 0 & 1 & C_y \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a_{i1} & a_{i2} & 0 \\ a_{i3} & a_{i4} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -C_x \\ 0 & 1 & -C_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{6}$$
$$a_{i1}, a_{i3} \in [0.8, 1.2],\ a_{i2}, a_{i4} \in [-0.2, 0.2],$$

  where (C_x, C_y) is the center coordinate of the image.

K_i, a_{i1}, a_{i2}, a_{i3} and a_{i4} are randomly determined.

RI-Feature method is considered to make an overall correlation of classifiers smaller. Ensemble learning theory is considered in some researches. Tumor and Ghosh indicated the following formula (Tumor and Ghosh, 2016).

$$err_{add}(H) = \frac{1+\theta(T-1)}{T}\overline{err}_{add}(h), \tag{7}$$

where err_{add} is the overall error rate of classifiers, $\overline{err}_{add}(h)$ is the mean of error rates of classifiers, θ is an overall correlation of classifiers, and T is the number of classifiers. The overall error rate becomes smaller by making $\overline{err}_{add}(h)$ or θ smaller or increasing T. For example, Random Forest is the ensemble learning considered an overall correlation of classifiers (Breiman, 2001). It is expected that the RI-Feature method makes θ smaller by adding some fluctuations to the input data.

2.4 Recognition Stage

To recognize a query image, T RI-Features are extracted from the query image at first. Some different combinations of random parameters and kernels, s_i, K_i, a_{i1}, a_{i2}, a_{i3}, and a_{i4} are used to extract the i-th RI-Feature. Second, the RI-Features are put into each classifier, and each classifier produces the class label as output. An answer vector is created as follows when the i-th classifier outputs an answer r_i.

$$\begin{aligned} \boldsymbol{A}_i &= (a_{i1}, a_{i2}, ..., a_{iN_c}), \\ a_{ij} &= \begin{cases} 1 & \text{if } j = r_i \\ 0 & \text{if } j \neq r_i \end{cases}, \end{aligned} \tag{8}$$

where N_c is the number of classes. The final answer R is calculated by the plurality voting as follows.

$$\begin{aligned} \boldsymbol{A}_f &= \sum_{i=1}^{T} \boldsymbol{A}_i,\ \ \boldsymbol{A}_f = (a_{f_1}, a_{f_2}, ..., a_{f_{N_c}}), \\ R &= \text{argmax}_x\{f(x) \mid f(x) = a_{f_x}\}. \end{aligned} \tag{9}$$

3 PERFORMANCE EVALUATION OF ENSEMBLE SCHEME

3.1 Experimental Environment

Experimental environment is as follows:

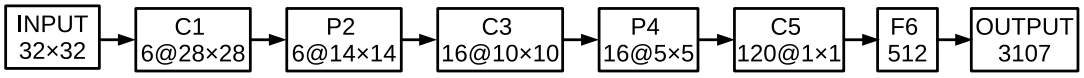
- CPU: Intel Core i7-3770 (3.4 GHz)
- Memory: 16 GB
- Development language: C/C++, Python

3.2 Dataset

We have created a new Japanese scene character dataset which is based on the dataset compiled in

Table 1: Parameters of a HOG feature.

Image size	Cell size	Block size	Orientation	Dimension
16×16	2	16	5	320
32×32	4	32	5	320
64×64	8	64	5	320



CNN

HOG-3LP

Figure 5: The architecture of CNN and HOG-3LP.

Table 2: Results of Comparison between CNN and HOG.

Method	Recognition accuracy [%]
CNN	63.21
HOG-3LP	**66.21**

(Horie and Goto, 2018) for testing. The dataset consists of Hiragana, Katakana and Kanji (1,400 images and 523 classes) taken in real scenes. All character images are in color and in arbitrary size. Seven Japanese fonts (3,107 classes, total 21,749 characters) are used for training. The training dataset does not include real scene characters since it is difficult to collect characters of all classes in Japanese.

3.3 Comparison of CNN and HOG

Although CNN has been reported to achieve high-level accuracy in character recognition, it is thought that CNN is not effective in a situation of learning only SSD. In order to confirm it, the following two architectures are compared.

- CNN: CNN based on LeNet(LeCun et al., 1998) which consists of 512 nodes in F6 layer and 3,107 in the output layer.
- HOG-3LP: Three-Layer Perceptron (3LP) which consists of 320 nodes in the input layer, 512 in the hidden layer and 3,107 in the output layer, and the learning HOG feature of 320 dimension. ReLU is used as the activation function in the hidden layer (Glorot et al., 2011).

Figure 5 shows the architectures. The parameters of the HOG feature are shown in Table 1. 434,980 grayscale synthetic scene character images are used as training data in either cases.

Table 2 shows the results. It is shown that CNN is inferior to HOG. Thus, we use the HOG feature hereinafter.

3.4 Evaluation of RI-Feature

The following RI-Feature structures are compared in order to evaluate the effects of the RI-Feature method.

- MSR (Horie and Goto, 2018)
- MSR-RF
- MSR-RA
- MSR-RF-MF-RF-MF-RF
- MSR-RA-RF-MF-RF-MF-RF

Regarding the parameters of the ensemble learning, $T = 90$ and $K = 9000$ are used in all cases. Nearest Neighbor Search (NNS) is utilized as the classifier of the ensemble scheme. We have chosen NNS in order to see the system's behavior in an environment which is as simple as possible in this early stage of development. Although using some other classifiers would be quite interesting, it should be included in our future work.

Table 3 shows the results. Particularly, the combination of RF and MF greatly improves the accuracy. This is probably because the MF effectively decreases the image noise. Moreover, it is thought that the RA makes the ensemble learning robust against geometric distortions.

Figure 6 shows the evaluation results about the number of classifiers. Our system using the RI-Feature outperforms the previous system in a condition of $T > 15$.

3.5 Improvement of Classifiers

It is expected that more SSD make the recognition system better. Time complexity and space complexity of NNS are both $O(n)$, where n is the number of image samples. On the other hand, the complexities of Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) are $O(1)$. Thus, SVM and

Table 3: Comparison among methods of different RI-Feature.

RI-Feature method	Recognition accuracy [%]
MSR (Horie and Goto, 2018)	65.57
MSR-RF	70.14
MSR-RA	71.64
MSR-RF-MF-RF-MF-RF	76.00
MSR-RA-RF-MF-RF-MF-RF	**78.50**

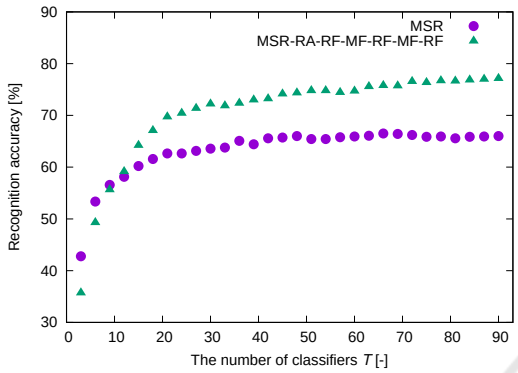


Figure 6: The evaluation about the number of classifiers.

Table 4: Evaluation of different classifiers in the ensemble learning.

Classifier	Recognition accuracy [%]
NNS	78.50
SVM	77.93
3LP	**80.71**

MLP are able to learn a large number of training data. The previous methods (Jiang and Goto, 2017)(Horie and Goto, 2018) utilized SVM in the ensemble learning. We have introduced MLP in order to improve the recognition accuracy.

Following three classifiers are compared.

- NNS: Nearest Neighbor Search ($K = 9,000,\ T = 90$)
- SVM: Linear Support Vector Machine ($C = 1$) ($K = 200,000,\ T = 90$)
- 3LP: Three-Layer Perceptron ($K = 200,000,\ T = 90$)

Figure 7 shows the architecture of 3LP. ReLU is used as the activation function in the hidden layer (Glorot et al., 2011). MSR-RF-MF-RF-MF-RF-MF-RF-RA is used as the RI-Feature method in all cases.

Table 4 shows the results. It is shown that 3LP is superior to NNS and SVM. It is considered that the number of training data and the kind of classifier are important for the ensemble scheme. Figure 8 shows that our proposed system is able to recognize characters which has some rotation, blur, lighting, noise and various fonts. It is shown that the correct answers of some incorrectly recognized characters are included in the second or third candidate. We expected that some of the characters are correctly recognized by combining the natural language processing or any other processes. Our system can not recognize some characters having great geometric distortion, complex background, and extraordinary design.

4 CONCLUSION

We have proposed RI-Feature method for improvement of the ensemble scheme proposed in our previous work. RI-Feature method is to randomly process an image before extracting the character features. We have also proposed to introduce MLP in the ensemble.

Experimental results have shown that HOG outperforms CNN in the case of using only SSD. It is also shown that the accuracy has been improved from 65.57% to 78.50% by the newly introduced RI-Feature method in the ensemble scheme.

Our future work includes to examine the appropriate feature in the ensemble scheme learning the Japanese synthetic scene characters.

REFERENCES

Breiman, L. (1996). Baggin predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Glorot, X. et al. (2011). Deep sparse rectifier neural networks. In *Proceedings of Machine Learning Research*, volume 15, pages 315–323.

Horie, F. and Goto, H. (2018). High-accuracy japanese scene character recognition using synthetic scene characters and multi-scale voting classifier. In *DAS2018 Short Paper*.

Jaderberg, M. et al. (2014). Synthetic data and artificial neural networks for natural scene text recognition. *Workshop on Deep Learning, NIPS*.

Jiang, L. and Goto, H. (2017). Ensemble classifier with dividing training scheme for chinese scene character recognition. In *2017 International Conference on Image and Vision Computing New Zealand (IVCNZ)*.

LeCun, Y. et al. (1998). Gradient-based learning applied to document recognition. In *Proc. of the IEEE*.

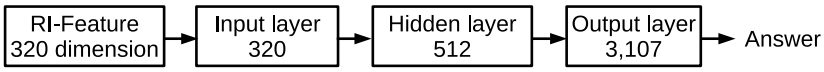


Figure 7: The architectures of 3LP.

Correctly recognized characters

1st candidate	団	偽	康	肪	へ
2nd candidate	口	悌	東	訪	毒
3rd candidate	司	お	味	防	キ

1st candidate	電	審	こ	世	ど	且
2nd candidate	馳	毒	ニ	櫨	と	甘
3rd candidate	億	泰	ご	纏	橡	日

Incorrectly recognized characters

Figure 8: Recognition examples.

Ren, X. et al. (2016). A cnn based scene chinese text recognition algorithm with synthetic data engine. *arXiv preprint arXiv: 1604.01891*.

Tian, S. et al. (2016). Multilingual scene character recognition with co occurrence of histogram of oriented gradients. *Pattern Recognition*, 51:125–134.

Tumor, K. and Ghosh, J. (2016). Theoretical foundations of linear and order statistics combiners for neural pattern classifiers. In *Technical Report TR-95-02-98, Computer and Vision Research Center, University of Texas, Austin*.