Synthetic Scene Character Generator and Ensemble Scheme with
the Random Image Feature Method for Japanese and Chinese
Scene Character Recognition

Fuma HORIE (a), Student Member, Hideaki GOTO (b), and Takuo SUGANUMA (c), Members

SUMMARY Scene character recognition has been extensively investigated for a couple of decades because it has a great potential in many applications including automatic translation, signboard recognition, and reading assistance for the visually-impaired. However, scene characters are difficult to recognize at sufficient accuracy owing to various noise and image distortions. In addition, Japanese scene character recognition is more challenging and requires a large amount of character data for training because thousands of character classes exist in the language. Some researchers propose training data augmentation techniques using Synthetic Scene Character Data (SSCD) to compensate for the shortage of training data. In this paper, we propose a Random Filter which is a new method for SSCD generation, and introduce an ensemble scheme with the Random Image Feature (RI-Feature) method. Since there has not been a large Japanese scene character dataset for the evaluation of the recognition systems, we have developed an open dataset JPSCI1400, which consists of a large number of real Japanese scene characters. It is shown that the accuracy has been improved from 70.9% to 83.1% by introducing the RI-Feature method to the ensemble scheme.

key words: character recognition, Japanese scene character recognition, synthetic scene character data, ensemble system, scene character dataset

1. Introduction

Scene character recognition has great potential for many applications including automatic translation, signboard recognition, and reading assistance for the visually-impaired. Scene character recognition is challenging and more difficult compared with printed character recognition owing to some environmental factors such as rotation, geometric distortion, uncontrolled lighting, blur and noise at image capture. The accuracy of current OCR systems has not become high enough, although some effective methods have been proposed in the past. In addition, Japanese and Chinese scene character recognition is more challenging and requires a large amount of character data for training because thousands of character classes exist in the language. Figure 1 shows some examples of Japanese characters in natural scenes. This paper aims to improve the accuracy of Japanese and Chinese scene character recognition when real scene character images are insufficient for training.

Researchers have proposed scene character recognition methods using Synthetic Scene Character Data (SSCD) [1]. Synthetic data were shown to be useful for improving the accuracy of recognizing real scene characters. SSCD are generated from the font data of printed characters by applying processes such as color change, image filtering, and geometric deformation. The data augmentation by SSCD is expected to be effective especially for languages consisting of thousands of character classes since the required training data need to be increased.

Scene character recognition systems have been developed for a wide variety of languages. Regarding the English alphabet, Gao et al. achieved state-of-the-art recognition accuracy by introducing a Spatiality Embedded Dictionary (SED) into the recognition system [16]. Ali et al. discussed some problems in the cursive scene text recognition and devised a CNN classification system for Urdu scene character recognition [3]. He and Hu developed a Chinese scene character recognition system based on the linear SVM [4]. Tian et al. applied the scene character recognition system to three different languages [13]. In our study, we focus on Japanese scene characters consisting of Hiragana, Katakana and Kanji.

Some researchers have studied ensemble techniques combining multiple classifiers to improve prediction accuracy. Breiman proposed Random Forests which is a combination of tree predictors [5]. Bay showed that multiple Nearest Neighbor Search (NNS) classifiers are able to reduce both bias and variance components of the error rate [6]. Kim et al. demonstrated that the SVM ensemble scheme outperforms the single SVM in terms of the classification accuracy.
The accuracy of current Japanese and Chinese scene character recognition systems has not become high enough, and there is as yet no large real scene character dataset for training in Japanese and Chinese yet. Therefore, the objective of this research is to improve the accuracy of Japanese and Chinese scene character recognition by using an ensemble strategy with SSCD. For the training data augmentation, we propose a Random Filter for generating SSCD effectively. For further improvements of the generalization ability, the Random Image Feature (RI-Feature) method is proposed in this paper. The RI-Feature method randomly processes an image before extracting character features. It is expected that the RI-Feature method will make the recognition system less sensitive to noise and geometric distortions and improve the generalization ability. These proposed methods contribute to improving recognition accuracy when the number of real scene character images for training is limited. As the recognition accuracy has not been sufficient for use in various applications, we prioritize the improvement of the accuracy more than computational complexity.

To evaluate the recognition system, we develop an open Japanese scene character dataset, JPSC1400 dataset, https://www.imlab.org/db/, as no existing large Japanese dataset was known to us. The dataset consists of 1,400 Japanese scene characters and 523 classes collected by taking natural scene images in various environments. This dataset is intended to aid the developments and improvements of Japanese scene character recognition.

This paper is organized as follows. Section 2. describes the ensemble scheme methods, the procedure of generating SSCD, and the RI-Feature method. Although we presented our initial ideas of the SSCD generation and the RI-Feature method in our conference paper [11], this paper newly introduces the projective transformation and Gaussian noise into the SSCD generation aiming to improve the generalization ability further. Section 3. details the Japanese scene character dataset developed in this work. Section 4. shows the experimental procedures and the results. The experiments and evaluations were quite limited in [11]. In this paper, we conduct comprehensive comparisons with many other recognition systems and also add Chinese scene character dataset. Conclusions and future work are given in Sect. 5.

2. Scene Character Recognition Using SSCD and the RI-Feature Method

2.1 Flow of the Recognition System

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 method in [8]. Each subset consists of $K$ images. These subsets are converted to an SSCD set. RI-Feature sets are extracted from the generated SSCD. Finally, $T$ classifiers are created by a learning method using 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 SSCD Generation

There are thousands of character classes in Japanese and Chinese. For example, the JIS (Japanese Industrial Standards) Level 1 character set consists of 2,965 kanji characters. Although it may be better to collect training data from real scene images in order to obtain higher recognition accuracy, data acquisition is a hard task. Collecting all classes of scene characters is quite challenging owing to the sheer number of character classes in Japanese and Chinese. In order to enhance the recognition accuracy, some researchers have used augmentation methods with Synthetic Scene character Data (SSCD) randomly generated by particular algorithms from some font sets of printed characters. SSCD is useful for solving the problem of insufficient amounts of training data.

We propose an SSCD generator including the following processes. A set of random parameters is chosen uniquely for each character class and for each sample.

1. Projective Transformation (PT)

Scene images have geometric distortions as they are captured with various angles. Projective transformation is utilized during part of the SSCD generation in some previous studies [1], [2] since the transformation is able to express the distortions.

We assume an image of $W \times H$ pixels. The vertices of the image $(1,1), (W,1), (1,H), (W,H)$ are transformed to $(1+d_1, 1+d_2), (W+d_3, 1+d_4), (1+d_5, H+d_6), (W+d_7, H+d_8)$, where $d_i \in [-5,5]$. $d_i$ is chosen from uniformly-random numbers.

2. Morphology Operation (MO)

There are fonts with various stroke thickness in scene
characters. Morphology operation changes the stroke thickness of SSCD image. 
3×3 matrices are used as the kernels of morphology operation. The type of the operation is randomly selected from [no applying, dilation, erosion].
3. Gaussian Filter (GF)
Scene images are often affected by blurs when the images are captured by a camera. Gaussian filter simulates various blurs. In some previous papers, Gaussian filter was utilized for the SSCD generation [1], [2], [10].
3×3 matrices are used as the Gaussian filter kernels defined by the following formula;

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

where \((x, y)\) is the offset from the kernel center, and \(\sigma\) is a uniformly-random number. Although the kernel becomes almost the same as the mean filter for larger \(\sigma\), we chose the range so that the value can be smaller and the filter covers a wide variety of blurring effects.
4. Color Change (CC)
Scene characters, unlike printed and handwriting characters, have a wide variety of colors.
We chose 20 colors frequently appearing in various scene character images. Two colors are selected randomly as the background and foreground. The intensity of an image is calculated by the following formula;

\[
\begin{align*}
L'_{f}(x, y) &= L(x, y) \times \frac{R_f - R_b}{255} + R_b + 0.5, \\
L'_{g}(x, y) &= L(x, y) \times \frac{G_f - G_b}{255} + G_b + 0.5, \\
L'_{b}(x, y) &= L(x, y) \times \frac{B_f - B_b}{255} + B_b + 0.5, 
\end{align*}
\]

where \((x, y)\) is the coordinate in the image; \(L(x, y)\) is the character image before processing; \(L'_f(x, y), L'_g(x, y), \) and \(L'_b(x, y)\) are the matrices representing the processed images; \(R_f, G_f, \) and \(B_f\) are the foreground colors; and \(R_b, G_b, \) and \(B_b\) are the background colors.
5. Gaussian Noise (GN)
Image noise often appears as images are captured with a camera. In order to simulate this noise, Gaussian noise is used for the SSCD generation in the previous research [1], [2], [10].
Gaussian noise is random noise following a Gaussian distribution that can be added to images. Let \(z\) be the Gaussian random variable, and \(p(z)\) be a probability density functions. The formula of the Gaussian distribution is expressed as follows:

\[ p(z) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{z^2}{2\sigma^2}\right), \]

where \(\sigma \in (0, 5]\) is chosen from uniformly-random numbers.
In the program, the Gaussian noise \(z\) can be generated by the Box-Muller’s formula;

\[ z = \sqrt{-2\log X} \sin 2\pi Y, \]

where \(X\) and \(Y\) are chosen from uniformly-random numbers in \((0, 1)\) separately.
The operation is not applied to the all pixels in the image so as not to destroy the character structure. Let \(W \times H\) be the image size, then we randomly select \(\sqrt{W \times H}\) pixels from the image, and apply the Gaussian noise processing to the selected pixels. For each pixel and for each color channel (Red, Green, and Blue), the pixel intensity is modified by adding \(\sigma z\), and the new intensity is obtained by clipping the modified value by 0 and 255.
6. Random Filter (RF)
Details of the Random Filter are described in this section.
Figure 3 shows some examples of the SSCD generated by the above processing methods. Figure 4 shows the flow of the above SSCD generation.
2.3 Random Filter in the SSCD Generation

Scene characters may suffer from some blurring, shadow, and noise when they are captured with a camera. Some decorrelated characters may have some enhanced edges and artificial shadows. In this work, we focus on these factors rather than shape variation as conventional character recognition methods are generally designed to tolerate some deformations. We generate a 3x3 kernel with uniformly-random numbers. Each element can be positive or negative, and it is normalized so the sum becomes 1.0. That is,

\[
K_{RF}^{m,n} = (k_{RF}^{m,n})_{1 \leq m,n \leq 3},
\]

where

\[
K_{RF}^{m,n} = (k_{RF}^{m,n})_{1 \leq m,n \leq 3}, \quad k_{RF}^{m,n} \in \mathbb{R},
\]

\[
S_{RF} = \sum_{m=1}^{3} \sum_{n=1}^{3} k_{RF}^{m,n},
\]

\[
k_{RF}^{m,n} = \frac{1}{S_{RF}} k_{RF}^{m,n},
\]

and the uniformly-random number \( k_{RF}^{m,n} \) is selected from the set of all real number \( \mathbb{R} \).

Figure 5 shows some images obtained by different kernels.

2.4 Random Image Feature

The ensemble scheme has been used to improve the generalization ability of recognition systems in previous work [9], [10]. We have shown that the ensemble scheme effectively improves the accuracy of scene character recognition. For further improvement of the recognition accuracy, we propose the RI-Feature method.

In our previous papers [10], [11], Multi-Scale Resizing (MSR) is presented, and the RI-Feature method includes the MSR as an image processing. The revised RI-Feature method in this paper includes Random Filter, Affine transformation, and Mean filter as shown below.

The 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 SSCD set. The processing is as follows.

- **Multi-Scale Resizing (MSR)**
  Images of \( D_i \) are resized to \( s_i \times s_i \) pixels, where \( s_i \) is defined by the following:

\[
s_i = \begin{cases} 
16 & (1 \leq i \leq T_{16}) \\
32 & (T_{16} < i \leq T_{16} + T_{32}) \\
64 & (T_{16} + T_{32} < i \leq T_{16} + T_{32} + T_{64}) 
\end{cases},
\]

where \( T_{16}, T_{32}, \) and \( T_{64} \) are the number of classifiers. Thus, the total number of classifier is \( T = T_{16} + T_{32} + T_{64} \). The aspect ratio of the input image may change when the image is not a square.

MSR was proposed in previous papers [10], [11], which demonstrated that MSR effectively improves the ensemble recognition accuracy.

- **Random Filter (RF)**
  Images of \( D_i \) are calculated by the RF using the following kernel:

\[
K_i = (k_{im,n})_{1 \leq m,n \leq 3},
\]

where

\[
K'_i = (k'_{im,n})_{1 \leq m,n \leq 3}, \quad k'_{im,n} \in \mathbb{R},
\]

\[
S_{i} = \sum_{m=1}^{3} \sum_{n=1}^{3} k'_{im,n},
\]

\[
k_{im,n} = \frac{1}{S_{i}} k'_{im,n},
\]

and the uniformly-random number \( k'_{im,n} \) is selected from the set of all real number \( \mathbb{R} \).

The RF may be applied to the image more than once and in combination with Mean Filter (MF) which is defined by a 3x3 convolution kernel whose elements are all 1/9. It is expected that the RF is useful for adding some variations to the features as it introduces various effects to the image. In addition, the MF is expected to simulate image blur and also reduce image noise.

- **Random Affine (RA)**
  Images of \( D_i \) are deformed by affine transformation using the following matrix:

\[
\begin{bmatrix}
1 & 0 & C_x \\
0 & 1 & C_y \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
a_{11} & a_{12} & 0 \\
a_{13} & a_{14} & 0 \\
0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
1 & 0 & -C_x \\
0 & 1 & -C_y \\
0 & 0 & 1
\end{bmatrix}
\]

where \( (C_x, C_y) \) is the center coordinate of the image. \( a_{11}, a_{12}, a_{13}, \) and \( a_{14} \) are randomly determined.

After applying the above image processing, an image feature is extracted from the image. Various feature extraction methods have been proposed for scene character recognition. The Histogram of Oriented Gradients (HOG) feature is utilized in a wide range of fields including object detection. Yi et al. demonstrated that the HOG feature is
effective for Chinese scene character recognition in comparison with other popular features (SIFT, SURF, BRIEF, ORB and DAISY) [12]. Although the HOG feature is utilized in this paper, the feature extraction method in the RI-Feature scheme can be replaced with another one.

The RI-Feature method is considered to make the overall correlation of classifiers smaller. Ensemble learning theory is considered in some earlier research. Tumor and Ghosh indicated the following formula [9].

$$err_{add}(H) = \frac{1 + \theta(T - 1)}{T} \overline{err}_{add}(h),$$  (15)

where $err_{add}$ is the overall error rate of classifiers, $\overline{err}_{add}(h)$ is the mean of error rates of classifiers, $\theta$ 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 $\theta$ smaller or by increasing $T$. For example, Random Forest is a kind of ensemble learning that takes into account the overall correlation of classifiers [5]. It is expected that the RI-Feature method makes $\theta$ smaller by adding some fluctuations to the input data.

2.5 Recognition Stage

In the recognition stage, $T$ RI-Features are first extracted from the query image. Some different combinations of random parameters and kernels, $s_i$, $K_i$, $a_{i1}$, $a_{i2}$, $a_{i3}$, and $a_{i4}$ are used to calculate 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 produces an answer $r_i$.

$$A_i = (a_{i1}, a_{i2}, ..., a_{iN}),$$

$$a_{ij} = \begin{cases} 1 & \text{if } j = r_i \\ 0 & \text{if } j \neq r_i \end{cases},$$  (16)

where $N_i$ is the number of classes. The final answer $R$ is determined by plurality voting as follows.

$$A_f = \sum_{i=1}^{T} A_i, \quad A_f = (a_{f1}, a_{f2}, ..., a_{fN}),$$

$$R = \underset{a_f}{\text{argmax}} \{ f(x) \mid f(x) = a_f \}.$$  (17)

2.6 Classifiers in the Ensemble Scheme

For the base classifier in the ensemble scheme, we compare the following configurations:

- NNS: Linear Nearest Neighbor Search
- SVM: Support Vector Machine
- MLP: Multi-Layer Perceptron

Some conventional methods employ NNS, SVM or MLP as the classifiers in the ensemble scheme in order to improve the limited classification performance [6], [7], [17]. They demonstrated some degree of validity for the NNS, SVM and MLP ensemble schemes.

3. JPSC1400: Japanese Scene Character Dataset

Real scene characters are required for evaluating the performance of character recognition systems, and some scene character datasets have already been built for various languages such as English and Chinese. Pan+ChiPhoto [13] and C-SVT [14] are Chinese scene character datasets consisting of images taken in some Chinese cities. IC-DAR2003 [18], Chars74K [19] and IIIT 5K-Word [20] are English scene character datasets that have been widely used to evaluate recognition systems. Regarding Japanese, Iwamura et al. developed a large dataset of scene text captured on shopping streets in the downtown of Osaka [21]. However, the datasets [14], [21] are not suitable for evaluating individual-character recognition methods since each character is not cropped or annotated. No open dataset for Japanese scene character recognition was suitable for our evaluations although some open Japanese font or handwriting datasets exist.

Collecting Japanese scene characters is considered much harder than collecting alphabet characters because thousands of classes exist in Japanese. We compiled a dataset of Japanese scene characters, JPSC1400 [22], in order to provide an open dataset to be used in the research community. JPSC1400 consists of Hiragana, Katakana, and Kanji (1,400 images and 523 classes) taken in real scenes in and around Sendai, Japan. JPSC1400 does not cover all character classes and the average number of samples per class is limited owing to the difficulties in data collection. Some Japanese characters are used only in some limited areas, and it is hard to collect real images. All character images are in color and have arbitrary sizes. The images are not necessarily square and the size ranges from $25 \times 25$ to $616 \times 715$ pixels. Figure 6 shows some examples of the character images in the dataset.

The character images included in JPSC1400 were all collected by taking natural images, and individual character images were manually cropped out from the natural images since there is no character segmentation method that can automatically and precisely crop natural images having complex backgrounds. The character images were manually...
labeled with their character codes. We took the character images with various environmental factors such as rotation, blur, shadows, and uncontrolled lighting. In addition, we intentionally collected as many different fonts as possible. The character images collected included guide signboards, product labels, and leaflets.

Table 1 shows the recognition accuracy obtained using some conventional methods. We used the training dataset consisting of seven Japanese fonts and HOG features in this experiment. The images in the seven Japanese fonts dataset are not SSCD but grayscale font data. The accuracy is thought to be too low for practical use. Moreover, some characters in the dataset have remarkable distortion which is mainly due to a large angle at image capture. JPSC1400 is supposed to be used for test data, and may not be suitable for the training of classifiers because it does not cover all character classes.

## 4. Performance Evaluation and Comparison

### 4.1 Experimental Environment

The experimental environment is as follows:
- CPU: Intel Core i7-3770 (3.4 GHz)
- Memory: 16 GB
- Development language: C/C++, Python

We evaluate the developed ensemble method in the two languages:
- Japanese (JP): We use seven Japanese fonts (3,107 classes per font, total 21,749) for the training. All character images are in 64 x 64 pixel size. We create an SSCD set from the seven fonts. JPSC1400 is used as the evaluation dataset.
- Chinese (CH): The SCUT-SPCCI dataset [23] consisting of Chinese font images is used for training, and we use a part of SCUT-SPCCI (3,775 classes, 100 fonts) [10] in this evaluation. Also, we create an SSCD set from the SCUT-SPCCI dataset as we did for the Japanese dataset. Regarding the evaluation, the system recognizes the real scene character images in Pan+ChiPhoto [13] consisting of the Chinese scene characters (1,410 classes, 10,065 images).

With regard to the feature extraction, we employed the HOG. The parameters of HOG feature are shown in Table 2.

### 4.2 Effect of SSCD Generator

We compare the following seven kinds of SSCD generators.
- PT-MO-CC-GF-RF-GN

### 4.3 Performance of RI-Feature Method with Different Classifier

We compare the following six kinds of RI-Feature extraction.
- (a) None
- (b) MSR
- (c) MSR-RA
- (d) MSR-RF
- (e) MSR-RA-RF-RF-RF
- (f) MSR-RA-RF-MF-RF-MF-RF

In each SSCD generator, each image processing is applied in the written order. For instance, “PT-MO-CC-GF-RF-GN” indicates the SSCD generator shown in Fig. 4. We use $T = 120$ and $K = 10000$ in the ensemble scheme, where $T$ is the number of classifiers, and $K$ is the number of SSCD images in each subset. For the comparison of the SSCD generators, the RI-Feature method is not applied in this section. NNS is used as the ensemble classifier. We use the Japanese dataset for the evaluation.

Table 3 shows the results. All processing contributes to the improvement of recognition accuracy, and projective transformation has the largest contribution. Thus, we use PT-MO-CC-GF-RF-GN in the following steps. Regarding the parameter values of $K$ and $T$, we have confirmed that there is no significant change in recognition accuracy as long as the values are sufficiently large as we chose.

### 4.4 Table 1: Recognition accuracy of conventional methods.

| Method       | Accuracy [%] |
|--------------|--------------|
| HOG & NNS    | 47.0         |
| HOG & SVM    | 44.3         |

### 4.5 Table 2: Parameter of HOG features.

| Image size | Cell size | Block size | Orientation | Dim. |
|------------|-----------|------------|-------------|------|
| 16x16      | 2x2       | 16x16      | 5           | 320  |
| 32x32      | 4x4       | 32x32      | 5           | 320  |
| 64x64      | 8x8       | 64x64      | 5           | 320  |

### 4.6 Table 3: Effect of each operation in SSCD generation.

| PT | MO | CC | GF | RF | GN | Accuracy [%] |
|----|----|----|----|----|----|--------------|
| o  | o  | o  | o  | o  | o  | 64.7         |
| o  | o  | o  | o  | o  | o  | 55.8         |
| o  | o  | o  | o  | o  | o  | 60.9         |
| o  | o  | o  | o  | o  | o  | 63.9         |
| o  | o  | o  | o  | o  | o  | 61.8         |
| o  | o  | o  | o  | o  | o  | 63.9         |

- MO-CC-GF-RF-GN
- PT-CC-GF-RF-GN
- PT-MO-CC-RF-GN
- PT-MO-CC-GF-GN
- PT-MO-CC-GF-RF

In each SSCD generator, each image processing is applied in the written order. For instance, “PT-MO-CC-GF-RF-GN” indicates the SSCD generator shown in Fig. 4. We use $T = 120$ and $K = 10000$ in the ensemble scheme, where $T$ is the number of classifiers, and $K$ is the number of SSCD images in each subset. For the comparison of the SSCD generators, the RI-Feature method is not applied in this section. NNS is used as the ensemble classifier. We use the Japanese dataset for the evaluation.
Table 4 Performance of RI-Feature method with different classifiers

| Method    | NNS  | SVM  | MLP  |
|-----------|------|------|------|
| RI-Feature method | Accuracy [%] | Accuracy [%] | Accuracy [%] |
| (a) None     | 64.7 | 68.1 | 70.9 |
| (b) MSR      | 71.4 | 69.4 | 77.3 |
| (c) MSR-RA   | 74.5 | 74.2 | 77.4 |
| (d) MSR-RF   | 73.2 | 75.2 | 78.1 |
| (e) MSR-RA-RF-RF-RF | 76.6 | 79.0 | 81.6 |
| (f) MSR-RA-RF-MF-RF-MF-RF | 79.9 | | 83.1 |

3 Layers: 320-512-(3,107 or 3,775)

We use the Japanese dataset for evaluation. Table 4 shows the results. The combination of RF and MF makes an especially high contribution to the accuracy improvement. This is probably because the MF effectively reduces the image noise. Moreover, it is thought that the RA processing makes the ensemble learning robust against geometric distortions. In terms of classifiers, it is shown that MLP is superior to NNS and SVM. Thus, we use the RI-Feature extraction (e) and the MLP hereafter.

4.4 Comparison with Other Methods

We compare our developed system with some other scene character recognition methods. The ensemble parameters $K = 500,000$ and $T = 120$ are used.

- Evaluation with Japanese dataset:
  - We evaluate the following five methods:
    - a) HOG-MLP: MLP trained from HOG features. The structure is shown in Fig. 7.
    - b) LeNet [24]: Seven-layer CNN. The structure is shown in Fig. 8.
    - c) VGG16 (Visual Geometry Group) [25]: 16-layer CNN. The structure is shown in Fig. 9.
    - d) Our previous method [11]
    - e) Proposed method
  - Methods a), b) and c) are trained using 2,174,900 SSCD images. Table 5 shows that our proposed method is superior to the other methods.

- Evaluation with Chinese dataset:

Table 5 Comparison with JPSC1400 dataset.

| Method            | Accuracy [%] |
|-------------------|--------------|
| a) HOG+MLP        | 71.7         |
| b) LeNet          | 65.1         |
| c) VGG16          | 80.6         |
| d) Our previous method [11] | 67.2 |
| e) Proposed       | 85.2         |

Table 6 Comparison with Pan+ChiPhoto dataset.

| Method            | Training data | Accuracy [%] |
|-------------------|---------------|--------------|
| CNN [26]          | RSCD          | 61.5         |
| CNN+softmax [27]  | RSCD          | 53.5         |
| CNN+SVM [26]      | RSCD          | 62.5         |
| Co-HOG [13]       | RSCD          | 64.5         |
| Co-HOG,NewOffset [13] | RSCD         | 65.4         |
| ConvCoHOG [13]    | RSCD          | 68.4         |
| ConvCoHOG,NewOffset [13] | RSCD      | 71.2         |
| MCA-FV [28]       | RSCD          | 76.7         |
| CSM-FV [29]       | RSCD          | 75.3         |
| BCA-FV [29]       | RSCD          | 77.3         |
| FV-CCA [30]       | RSCD          | 79.5         |
| Jiang&Goto [10]   | RSCD+SSCD     | 81.5         |
| Proposed          | SSCD          | 84.0         |

We compare our method with other methods that were evaluated previously in other studies. Table 6 shows the results. In the table, SSCD means Synthetic Scene Character Data, and RSCD means Real Scene Character Data. Our proposed method is superior to the other methods. In addition, recognizable classes in the other methods are limited because they use the RSCD as training data covering only some of the character classes. It is thought that the proposed system is available in a wide range of situations for languages with many classes.

4.5 Example of Recognized Characters

Figure 10 shows some examples of test character images. The proposed ensemble system is able to recognize the characters affected by some rotation, blur, lighting, noise and various fonts as shown at (a). It is necessary to improve
the SSCD in order to enable recognition of characters like those shown at (b). The incorrectly recognized results show that some correct results exist as the 2nd or 3rd candidates. This implies that the proposed method leaves some possibility for further accuracy improvement by natural language post-processing [31].

5. Conclusions

Japanese and Chinese scene character recognition involves two major difficulties: some image degradations and distortions affect the recognition, and there is a shortage of training data due to the thousands of character classes in the language. We have proposed a training data augmentation technique using some image filters and a scene character recognition system based on an ensemble voting scheme. We have proposed a Random Filter as a part of SSCD generation. Moreover, our recognition system has adopted the RI-Feature to make the system less sensitive to noise and geometric distortions in the image.

A new open Japanese scene character dataset, JPSC1400, has been presented. This dataset consists of character images affected by various environmental factors. We have evaluated our proposed methods using the new dataset.

The experiments have shown that all processes in the SSCD generation contributes to higher accuracy and multi-scale scheme is able to make ensemble learning better. In addition, the accuracy was improved from 70.9% to 83.1% by introducing the RI-Feature method to the ensemble scheme.

Our future work includes further analyses and optimization of the system. Although the evaluations will not be easy owing to the high computational cost, it would also be worth trying Co-HOG or ConvCo-HOG instead of HOG features to achieve higher recognition accuracy with the proposed framework.

References

[1] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Synthetic data and artificial neural networks for natural scene text recognition,” Workshop on Deep Learning, NIPS, 2014.
[2] X. Ren, K. Chen, and J. Sun, “A CNN Based Scene Chinese Text Recognition Algorithm With Synthetic Data Engine,” arXiv preprint arXiv:1604.01891, 2016.
[3] A. Ali, M. Pickering, and K. Shafi, “Urdu Natural Scene Character Recognition using Convolutional Neural Networks,” Proc. ASAR 2018, 2018.
[4] S. He and X. Hu, “Chinese Character Recognition in Natural Scenes,” Proc. ISCI 2016, vol.2, pp.124–127, 2016.
[5] L. Breiman, “Random Forests,” Machine Learning, vol.45, pp.5–32, 2001.
[6] S.D. Bay, “Combining Nearest Neighbor Classifiers Through Multiple Feature Subsets,” Proc. the 15th Int. Conf. Machine Learning, pp.37–45, July 1998.
[7] H. Kim, S. Pang, H. Je, D. Kim, and S.Y. Bang, “Constructing support vector machine ensemble,” Pattern Recognit., vol.36, no.12, pp.2757–2767, Dec. 2003.
[8] L. Breiman, “Bagging Predictors,” Machine Learning, vol.24, pp.123–140, 1996.
[9] K. Tumor and J. Ghosh, “Theoretical foundations of linear and order statistics combiners for neural pattern classifiers,” Technical Report TR-95-02-98, Computer and Vision Research Center, University of Texas, Austin, 1995.
[10] L. Jiang and H. Goto, “Ensemble classifier with dividing training scheme for Chinese scene character recognition,” Proc. IVCNZ 2017, 2017.
[11] F. Horie and H. Goto, “Synthetic scene character generator and multi-scale voting classifier for Japanese scene character recognition,” Proc. IVCNZ 2018, 2018.
[12] C. Yi, X. Yang, and Y. Tian, “Feature representations for scene text character recognition,” Proc. ICDAR 2013, pp.907–911, 2013.
[13] S. Tian, U. Bhattacharya, S. Lu, B. Su, Q. Wang, X. Wei, Y. Lu, and C.L. Tan, “Multilingual scene character recognition with co-occurrence of histogram of oriented gradients,” Pattern Recognit., vol.51, pp.125–134, March 2016.
[14] Y. Sun, J. Liu, W. Liu, J. Han, E. Ding, and J. Liu, “Chinese street view text: Large-scale Chinese text reading with partially supervised learning,” Proc. ICCV 2019, pp.9085–9094, 2019.
[15] C. Yang and Y. Yang, “Improved local binary pattern for real scene optical character recognition,” Pattern Recognit. Lett., vol.100, no.C, pp.14–21, Dec. 2017.
[16] S. Gao, C. Wang, B. Xiao, C. Shi, W. Zhou, and Z. Zhang, “Scene text recognition by learning co-occurrence of strokes based on spatiality embedded dictionary,” IET Comput. Vis., vol.9, no.1, pp.138–148, 2015.
[17] L.K. Hansen and P. Salamon, “Neural network ensemble,” IEEE Trans. Pattern Anal. Mach. Intell., vol.12, no.10, pp.993–1001, Oct. 1990.
[18] R.M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, “ICDAR 2003 robust reading competitions,” Proc. ICDAR 2003, pp.682–687, 2003.
[19] T.E. de Campos, B.R. Babu, and M. Verma, “Character recognition in natural images,” Proc. VISAPP, vol.2, pp.273–280, 2009.
[20] A. Mishra, K. Alahari, and C.V. Jawahar, “Scene text recognition using higher order language priors,” Proc. 23rd BMVC, 2012.
[21] M. Iwamura, T. Matsuda, N. Morimoto, H. Sato, Y. Ikeda, and K. Kise, “Downtown Osaka scene text dataset,” Proc. ECCV 2016, pp.440–455, 2016.
[22] JIPSCI1400, https://www.imglab.org/db.
[23] Z. Zhong, L. Jin, and Z. Feng, “Multi-font printed Chinese character recognition using multi-pooling convolutional neural network,” Proc. ICDAR, pp.96–100, 2015.
[24] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. the IEEE, vol.86, no.11, pp.2278–2324, Nov. 1998.
[25] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2015.
[26] M. Jaderberg, A. Vedaldi, and A. Zisserman, “Deep features for text spotting,” Proc. ECCV 2014, pp.512–528, 2014.
[27] C. Shi, Y. Wang, F. Jia, K. He, C. Wang, and B. Xiao, “Fisher vector for scene character recognition: a comprehensive evaluation,” Pattern Recognit., vol.72, pp.1–14, Dec. 2017.
[28] Y. Wang, C. Shi, C. Wang, B. Xiao, and C. Qi, “Multi-order co-occurrence activations encoded with fisher vector for scene character recognition,” Pattern Recognit. Lett., vol.97, pp.69–76, Oct. 2017.
[29] Z. Zhang, H. Wang, S. Liu, and T.S. Durrani, “Bilateral convolutional activations encoded with fisher vectors for scene character recognition,” IEICE Trans. Inf. & Syst., vol.E101-D, no.5, pp.1453–1456, May 2018.
[30] Z. Zhang, H. Wang, S. Liu, and B. Xiao, “Consecutive Convolutional Activations for Scene Character Recognition,” IEEE Access, vol.6, pp.35734–35742, 2018.
[31] M. Généreux and D. Spano, “NLP challenges in dealing with OCR-ed documents of derogated quality,” https://www.researchgate.net/publication/281112670_NLP_challenges_in_dealing_with_OCR-ed_documents_of_ derogated_quality, 2015. (accessed on Nov. 13, 2020)

Fuma Horie received the M.Sc. degree in information sciences in 2019 from Tohoku University in Japan. He is currently a graduate student in the doctoral course. His research interests include character recognition, image processing, and machine learning.

Hideaki Goto received the M.Sc. degree in physics in 1992 and the Ph.D. degree in engineering in 1995 from Tohoku University in Japan. He has been an Associate Professor at Tohoku University since 2001. His research interests include pattern recognition, document image analysis and recognition, identity federation, wireless network, and information security. He is a member of IEICE in Japan, IPSJ, and IEEE Computer Society.

Takuo Suganuma received the M.S. and Ph.D. degrees in engineering from the Chiba Institute of Technology, Japan, in 1994 and 1997, respectively. He is currently a Professor and the Director with the Cyberscience Center, Tohoku University, Japan. His main research interests include agent-based computing, flexible network middleware, network management, symbiotic computing, green ICT, and Disaster-resistant communications. He received the UIC-07 Outstanding Paper Award in 2007, etc. He is a member of the IEICE and the IPSJ.