修士学位論文要約（令和４年３月）

人工データを用いた高精度な日本語シーン文字認識に関する研究

龍　永臻
指導教員：菅沼 拓夫,　学位論文指導教員：後藤 英昭

High-Accuracy Japanese Scene Character Recognition Using Synthetic Data

Yongzhen LONG
Supervisor: Takuo SUGANUMA,　Research Advisor: Hideaki GOTO

Text in the scene is generally affected by various environmental factors. Conventional Optical Character Recognition (OCR) systems have been used to recognize scanned documents successfully for many years. However, these systems cannot be directly applied to the scene text. This study takes the Japanese scene character recognition task as the research object, and solved the problem of non-uniformity and incompletion of training dataset in conventional approach. Based on the previous synthetic data generation framework, two improved image processing are proposed in order to improve the diversity of training data. The convolutional neural network is employed as feature extractor. Experiments results show the effectiveness of the improved recognition framework.

1. Introduction

Scene character recognition has been a hot topic in the computer vision community in recent years. A robust scene character recognition is expected to deal with scene characters with large distortion, complex background and other environment effects. Chinese characters as well as Kanji characters in Japanese have a huge number of classes and most of them have more complex strokes than English letters. In order to achieve the ability to recognize all the characters in the language, a complete and balanced training dataset is necessary. Besides, diversity and complexity of training dataset is also important for the robustness of the model to various complex environmental factors.

As a matter of fact, the recognition performance of Japanese and Chinese scene characters is still far from a satisfactory level. The main objective of this research is to develop a Synthetic Scene Character Data (SSCD) generation engine, which can generate a large number of high-quality SSCD for training recognition model without any real sample. In addition, the performance of using convolutional neural network as feature extractor is also discussed.

2. Synthetic Scene Character Data Generator

The methods of synthetic data creation belong to data augmentation technology. For computer vision tasks, data augmentation is realized by applying image processing methods to existing data. For synthetic scene character data, the object of applying image processing methods is font image data.

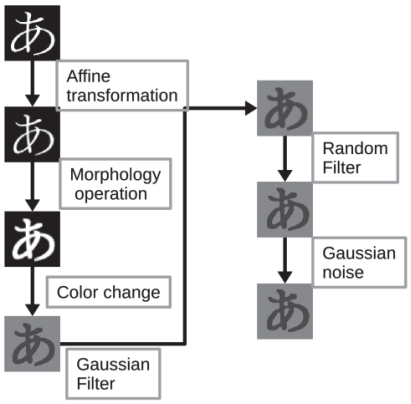

Figure 1. Flow of SSCD generation by Horie and Goto

The generator of SSCD proposed by Horie and Goto [1] is shown in Figure 1. Based on this procedure, two optimized image processing methods are developed to improve the diversity of SSCD. Parity binary morphological transformation is used to improve the diversity of character structure in SSCD.

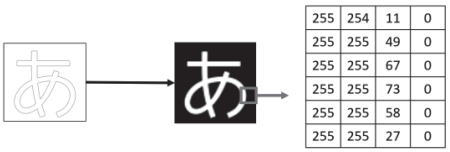

Figure 2. Pixels adjacent to the outline in font images have intermediate value.

As shown in Figure 2, when creating an 8-bit grayscale font image with the highest brightness, the pixels adjacent to the outline have intermediate value based on linear interpolation. It can realize an unsmooth binarization process based on this characteristic. Modified pixel values will be updated to black or white according to the parity of the original pixel value, the stroke structure of the character can be preserved, and the edge of the stroke becomes unsmooth. The flow of processing is shown in Figure 3. After binarization process, binary morphological transformation with specifically defined structural elements is applied to further improve the diversity of SSCD. Besides, improved color change will randomly generate the color of background and color of character from the fully 8-bit RGB space instead of a pre-defined color library. Compared with traditional methods, these improved transformations are expected to improve the diversity of SSCD.

Figure 3. Flow of parity binary morphological transformation.

3. Shallow Convolutional Neural Network For Scene Character Recognition

Convolutional Neural Network (CNN) is a representative of deep learning methods. Increasing the depth of the network can improve the ability to extract abstract features. Different from general object recognition task such as animal classification, characters are symbols invented by human beings, using stroke structure to distinguish different characters. Considering this characteristic, I adopted shallow CNN. The number of weighted layers in the network is reduced as much as possible. The introduction of the shallow network structure is to downsample the feature map after each convolutional layer, and each downsample will halve the width of the feature graph. In my research, the width of inputted images size is 64, which is a reasonable value and can ensure the clear structure of complex Japanese characters. The width of the feature graph is set to 4, which means that there are only 4 convolutional layers in the network structure. The structure of proposed CNN is shown in Figure 4.

Figure 4. Structure of proposed CNN.

4. Experiments and Discussion

In experiments, the dictionary used to generate SSCD is JIS level-1 Kanji (2,965 classes) addition with Katakana and Hiragana (142 class). JPSC1400 [1] is used to evaluate the new scene character recognition framework.

Comparative experiments are carried out firstly by replacing morphological transformation (MO) and color change (CC) with improved transformation methods respectively. The results are shown in Table 1, where '*' means conventional method. The results show that the improved transformation can greatly enhance the diversity of data and help the CNN model learn more robust feature extraction ability.

Table 1. Comparation of proposed processing methods.

Transformation	Accuracy(%)
MO*-CC*	85.64
MO*-CC	87.92
MO-CC*	87.28
MO-CC	**89.07**

Additionally, performance evaluation of proposed CNN models and other deeper CNNs are executed. The results are shown in Table 2. From the results, the recognition accuracy using proposed shallow CNN are higher than conventional model and deeper CNN models, which proves that the shallow CNNs with only four convolutional layers is enough to extract the features of scene character.

Table 2. Accuracy of proposed model and other model.

Model	Accuracy(%)
Horie and Goto, 2020	84.30
MobileNetV2	86.42
ResNet34	88.07
Shallow CNN	**89.07**

5. Conclusion

The single CNN model trained by SSCD without using any real data got 89.07% accuracy on evaluation dataset JPSC1400. Compare with previous method, about 5% accuracy gap is obtained. The improved SSCD generator can improve the diversity of SSCD. Besides, the proposed CNN model has more suitable learning ability for scene character recognition task.

References

1) Fuma Horie and Hideaki Goto. Japanese scene character recognition using random image feature and ensemble scheme. In ICPRAM, pages 414–420, 2019.