Early-Modern Printed Character Recognition using Ensemble Learning

Kaori Fujimoto, Yu Ishikawa, Masami Takata, Kazuki Joe
Department of Advanced Information and Computer Science
Graduate School of Humanity and Sciences, Nara Women's University
Nara, Japan
(fujimoto-kaori1329, ishikawa-yu0804, takata, joe)@ics.nara-wu.ac.jp

Abstract – *In recent years, archiving printed books has gained attention. The archiving means the extraction of text so that extracted text data is used as big data on the internet. The current DTP publishing includes text itself, so there is no problem to extract text while old printed books published over half a century ago are difficult to extract text data because of font difference compared with the current. We have studied the early-modern printed character recognition for the past ten years. In this paper, we present the improvement of the method using ensemble learning.*

Keywords: Character recognition, Ensemble learning, Early-modern books, Multiple features

1 Introduction

As knowledge acquisition methods on the Internet, web-mining or text-mining by intelligent agents are widely known. The target data include SNS, LOD and contents searchable books such as Google books, and they increase dramatically. Especially, the need to easily search book content as text is rapidly increasing. It is expected that further innovation will progress with this trend. Currently, books are mainly produced by DTP, and text conversion is performed simultaneously with the production while text conversion of books published in the past is roughly divided into manual and OCR. Although OCR has been put to practical use and is now widely used, it takes manual work to convert very important books into texts because the possibility of miss recognition cannot be excluded depending on the preservation state of the books and the scanning situation. However, what is important for knowledge acquisition by intelligent agents is large-scale automatic text conversion using OCR. The most severe constraint on automatic text conversion of books is their publication time. It is difficult to use existing OCR for early-modern books with so-called typographical printing before the current standardized type fonts [4]. The early-modern books are to be published in Japan from the Meiji period (1868-1912) to the early Showa era (1926-1945). The National Diet Library [1] allows public access to 350,000 early-modern books as image data on the National Diet Library's digital collection [2][3] web site.

Since the beginning of the century, storage density has dramatically improved, it becomes possible to digitize and save book contents as images. This tendency is remarkable in recent years, and it is possible to record image data of several thousands to several tens of thousands books at an enough resolution capable for OCRs on an HDD. At present, archiving early-modern books in various fields is the primary purpose of keeping valuable materials. If there is something different type of knowledge from the current Internet searchable knowledge, the needs of OCR for early-modern books will be dramatically enlarged. However, since research on character recognition specialized for early-modern Japanese books has not yet been conducted, we have collaborated with the National Diet Library for the research of automatically generating texts from images of early-modern books. Since early-modern books are published with letterpress printing without any unified standard fonts, the recognition rates of applying existing OCRs to early-modern books are not so good [4]. Furthermore, it is reported that fonts are different not only when the typeface is different by publisher but also when the same publisher makes prints in different times [5]. Because of these reasons, we proposed the multi-font Kanji recognition method based on hand-written Kanji recognition methods for early-modern printed books [6][7]. In this method, PDC features are used as feature quantities of characters, and SVM, which is one of supervised machine learning, is used as a classifier. We further add two kinds of feature quantity of weighted direction index histogram features and cell features to carry out recognition experiments with the three kinds of feature quantity. We have reported that the misrecognition rate of just cell features is the worst of 26.8% and the misrecognition rate of the three feature quantities is the second worst of 18.7%. We analyzed the cause of misrecognitions to find the following cases. In the case of misrecognitions of all the three feature quantities, the influence of the image itself such as the thickness and blurring of the character image is quite large while there are no misrecognized characters for each of the three feature quantities. We concluded that some recognition method with multiple feature quantities would improve the recognition system [8].

In this paper, we apply an ensemble learning method to improve the recognition rate for early-modern printed books.

We use PDC features, weighted direction index histogram features and cell features as the multiple feature quantities, and use SVM and OLVQ1 as classifiers. Namely, six week classifiers are composed with three feature quantities and two classifiers for the ensemble learning to recognize early-modern printed characters.

The rest of the paper is organized as follows. In section 2 we present a method for early-modern printed character recognition using ensemble learning. In section 3 we describe some preliminary experiment results for the ensemble learning to analyze the misrecognitions by feature quantity in order to get better ensemble learning. We show the experiment results of the ensemble learning with six week classifiers in section 4.

2 Recognition Method

In this section, we describe three features and two classifiers, which are used for weak classifiers of an ensemble learning method. The ensemble learning method we propose is explained.

2.1 Features

We describe three features: PDC features, weighted direction index histogram features and cellular features which are extracted from image data of early-modern Japanese printed characters.

The PDC (Peripheral Direction Contributivity) [9] is a feature quantity focused on the direction of character-lines. The PDC features reflect four statuses of character-segments: complexity, direction, connectivity and relative position. The complexity of character-segments is represented by line density. The relative position is represented by the peripheral form. The direction and the connectivity are represented by direction contributivity.

The weighted direction index histogram features [10] are obtained to focus on contour lines of character-segments. First, an input image character is smoothed in order to extract contour lines. Then, the contour lines are divided by smaller regions. Four direction histograms in each region are obtained. The four directions are determined from the location of the pixel next to the target pixel. The direction index of the region is put into much smaller regions using the two dimension Gaussian filter. The feature quantity is obtained from weighted direction index histograms of each small region.

The cellular features [11] are calculated to focus on the edge directions that indicate rapid changes of luminance in a character image. First, microscopic features are represented by the edge direction and size of a 3 × 3 local region using the target and its eight neighborhood pixels. Then, the determined edge directions are quantized in eight directions. Next, features of cell space are obtained to integrate the microscopic features around four pixels. Thereafter, it defines a fan-shaped area with ± 45 degree in order to determine the extension of the local region for each direction. The feature quantity is determined using features of the target cell and the cells in the fan-shaped.

2.2 Classifiers

We apply the three kinds of feature quantities extracted from early-modern Japanese printed character images to two classifiers for recognition. The two classifiers are SVM (Support Vector Machine) [12] and OLVQ1 (Optimized-learning-rate Learning Vector Quantization) [13]. OLVQ1 is a method to improve the learning convergence characteristics of LVQ1 by optimizing the learning coefficients. The LVQ is to make discrimination by dividing input space into a finite number of labeled reference vectors. Each time test data are given, the reference vectors learn them sequentially and are updated. The recognition result is the label of the nearest reference vector to the given test data. In OLVQ1, the learning coefficient is optimized. Note that the upper limit of the learning coefficient is adequately defined in this algorithm since the learning coefficient should be 1 or less.

2.3 Ensemble learning

Ensemble learning is also called collective learning. First, several classifiers learning from given data separately are generated. The generated classifiers are defined as weak classifiers. By appropriately selecting and combining a set of weak classifiers, a new classifier is generated. Ensemble learning often improves the accuracy of the generated classifier compared with any single weak classifiers.

In ensemble learning, a weight is given to each weak classifier when they are combined. If some weights are not properly set, the recognition results of each weak classifier are not correctly reflected to the total recognition result of the ensemble learning. To efficiently improve the accuracy, the weights need to be adjusted with analyzing weak classifiers.

In this paper, we adopt the ensemble learning of Adaboost.M1 [14] that is a method to compensate weaknesses. The distribution of weights for training data is adaptively changed at each step of learning so that the data to be intensively learned is determined. Calculating the weights using Adaboost.M1 to decide the class by a majority vote of the weights, the recognition result is obtained.

2.4 Proposed method

In this paper, we propose an early-modern Japanese printed character recognition method using ensemble learning.

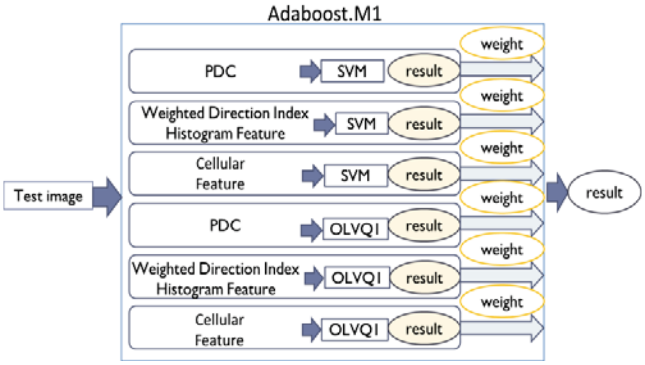

Fig. 1 recognition model using ensemble learning

Figure 1 shows the recognition model using ensemble learning. First, three features including PDC features, weighted direction index histogram features and cellular features are extracted from the images of early-modern Japanese printed characters. The extracted features are recognized with two classification methods: SVM and OLVQ1. It is known that ensemble learning improves the recognition capability by increasing the number of weak classifiers. In this paper, we add new classifier OLVQ1 which has a different mechanism from SVM. Therefore, we aim to improve the recognition rate by increasing weak classifiers to perform ensemble learning. Experiments are performed by ensemble learning using six weak classifiers generated by combining the three features and the two classification methods.

According to AdaBoost.M1, which is a well-known ensemble learning method, training data is given to six weak classifiers, and weights used for the experiments are updated from the recognition results. Adaboost.M1 updates the weights to compensate the weaknesses based on the misrecognition of each weak classifier. Adaboost.M1 makes each weak classifier learn the training data to calculate the recognition results of a set of test data. Then, we obtain the next recognition rates using the weights for each weak classifier which are calculated from the previous results. At this point, the calculated weights are used for judging each weak classifier. In the case of the same recognition result, the weights of the weak classifiers are added. The final recognition result is obtained so that the largest weights added by each weak classifier are found.

3 Preliminary experiment

In this section, we perform preliminary experiments for the experiments using ensemble learning described in the next section. In subsection 3.1, we explain recognition experiments by weak classifier. In subsection 3.2, we analyze the misrecognition of each weak classifier to investigate the tendency of the misrecognition.

3.1 Recognition experiments by weak classifier

For the experiments in this paper, we use 2,678 types of Japanese characters: JIS level-1 Kanji, JIS level-2 Kanji [15], and Hiragana. Each data set consists of 2,678 character images, and different six data sets are prepared. The character image data included in the data sets are collected from different eras and publishers. The publishers include Hiyoshi-dou, Shinshin-dou, Shun'you-dou, Heimin-sha, Okura-shoten, Iwanami-shoten, Shuei-kaku, Sakuma-shobo, Shincho-sya and Jitsugyononihon-sha.

We explain the dimension numbers of three features which are extracted from character images. In PDC features, the range of projected axis is divided into 16, and the character images are scanned from eight ways of vertical, horizontal and diagonal directions. By scanning the third contour point of each character stroke, direction contributivity in four directions is calculated. So the dimension number of PDC feature is 16×8×4×3=1,536. In weighted direction index histogram features, four directions histogram is aggregated into 18×18 regions using a two-dimensional Gaussian filter. The dimension number of weighted direction index histogram feature is 18×18×4=1,296. In cellular features, since the size of a cell space is 148×148, the dimension number of cellular feature is 1,152.

When SVM is used as a classification method, appropriate parameter choice is done by performing grid search on the feature space. We use an RBF (Radial Basis Function) kernel in SVM because the kernel is effective for recognizing the features with a large number of dimensions. When OLVQ1 is used as a classification method, we set the reference vector number and the initial learning coefficient to four and 0.5, respectively. When the reference vector number is five per class, any misrecognition does not occur for learning data, and Adaboost.M1 is not applicable. Therefore, OLVQ1 with five reference vectors cannot be a weak classifier.

Five out of six data sets are used as training data, and the remaining data set is used as test data. By changing the test data, six recognition experiments are performed as cross validation. First, the three features are extracted from six data sets. Next, the feature vectors obtained from five data sets for training data are given to the two classification methods of SVM and OLVQ1, and the feature vectors obtained from the remaining data set for test data are given to the trained classifiers. Table 1 shows the average recognition rates of six recognition experiments for each weak classifier.

Tab.1: Average recognition rates by weak learner

Weak classifier (feature, classification method)	Average recognition rate
PDC feature, SVM	87.47%
Weighted direction index histogram feature, SVM	85.55%
Cellular feature, SVM	81.01%
PDC feature,OLVQ1	75.94%
Weighted direction index histogram feature, OLVQ1	78.42%
Cellular feature, OLVQ1	79.38%

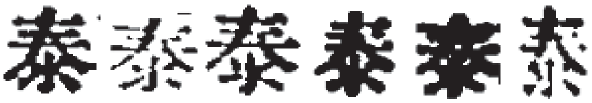

Fig.2 : images of a character recognized by OLVQ1 but misrecognized by SVM

From Tab.1, in the case of SVM, the better recognition rates are obtained from the feature vectors with higher dimension numbers. On the other hand, in the case of OLVQ1, the better recognition rates are obtained from the feature vectors with smaller dimension numbers. Therefore, it is clear that SVM is effective when the feature vector has the large number of dimensions and OLVQ1 is effective when the feature vector has the small number of dimensions. In addition, the recognition rates by SVM are higher than by OLVQ1. Thus, we confirm that SVM is better than OLVQ1 as a classification method.

3.2 Misrecognition analysis for each weak classifier

We analyze the misrecognition results in the case of SVM as the classification method. It is reported that the weak classifier with PDC features and SVM is easily affected by blurred character images and is not easily affected by character images with different thickness lines. It is also reported that the weak classifier with weighted direction index histogram features and SVM is easily affected by character images with similar center parts or blurred lines. Furthermore, the weak classifier with cellular features and SVM is affected by the difference such as line thickness and blur and connection of curves in hiragana [8].

Next, we analyze the misrecognition results in the case of OLVQ1 as the classification method. The recognition rate by OLVQ1 among the misrecognized characters by SVM is approximately 71.80%. We observe that OLVQ1 efficiently detect stroke lines to recognize characters. The stroke lines can be discriminated even if images with thick stroke lines and images with blurred stroke lines are mixed among six image data sets prepared for each character type. Figure 2 shows examples of a character type correctly recognizable by OLVQ1 but misrecognized by SVM.

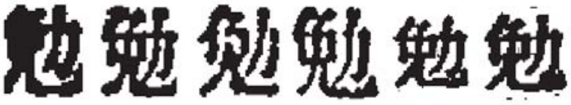

Fig.3 : images of a character of which part is blurred by ink bleeding

The weak classifier with PDC and OLVQ1 approximately recognizes 38.79% characters among the characters misrecognized by SVM. The recognition rate by PDC is the lowest among the three weak classifiers with OLVQ1. In addition, the weak classifier with PDC and OLVQ1 is not easily affected by the stroke line thickness of characters as compared with weak classifiers with SVM.

The weak classifier with weighted direction index histogram features and OLVQ1 recognizes about 42.04% characters among the characters misrecognized by SVM. The recognition rate with the weak classifier for the characters misrecognized by SVM is the highest among three weak classifiers with OLVQ1.

The weak classifier with cellular features and OLVQ1 recognizes about 37.59% characters among the characters misrecognized by SVM. The weak classifier is not easily affected by characters partly blurred by ink. Figure 3 shows a character type including character images partly blurred by ink.

From the above considerations, it is clear that the weak classifiers using OLVQ1 make up the misrecognition of the weak classifiers using SVM. Therefore, it turns out that the weak classifiers with SVM and OLVQ1 are effective for ensemble learning.

3.3 The number of misrecognizing weak classifiers against character types

In Adaboost.M1, the weight for the data misrecognized by a weak classifier is increased to obtain the weighted error rate. A weak classifier with the smallest weighted error rate is chosen to be weighted. Then, the weight for the data misrecognized by the selected weak classifier is increased. Since the weak classifier selected in the next step also has the smallest weighted error rate, only the weak classifier which is selected first is chosen many times when misrecognition ranges of weak classifiers are similar. Therefore, the misrecognitions of weak classifiers need to be dispersed. We investigate the number of the weak classifiers which misrecognize each character of 2,678 types for each data set to be recognized and consider whether each weak classifier is appropriate to be applied to AdaBoost.M1.

Tab. 2 average recognition rates by six weak classifiers and recognition results by Adaboost.M1

Recognition target	Average recognition rate by each of weak classifiers	Adaboost. M1
testdata1	77.81%	86.86%
testdata2	83.51%	90.14%
testdata3	84.25%	90.70%
testdata4	84.01%	91.60%
testdata5	82.47%	88.46%
testdata6	79.44%	85.55%

The number of weak classifiers which misrecognize characters among six weak classifiers for each character of 2,678 types is examined. When each of six test data sets is recognized, the average rates of the character types misrecognized by six, five, four, three and two weak classifiers are 3.78%, 3.34%, 5.06%, 5.62% and 9.64%, respectively. The average rate is 14.25% for only a single classifier. From the results, it is clear that the rate of the character types misrecognized by one weak classifier is the largest.

The average number of the weak classifiers which misrecognize each character type is 1.12, and the average of their standard deviations is 1.44. From the results, it is consider that the character types misrecognized by six weak classifiers are dispersed enough. Therefore, the recognition using Adaboost.M1 with the six weak classifiers is effective.

4 Experiments

In this section, we explain the recognition experiments using the ensemble learning. In subsection 4.1, we describe the experiment method and results. In subsection 4.2, we analyze the misrecognition.

4.1 Experiment results

The weak classifiers used in Adaboost.M1 are six types to combine three features (PDC, weighted direction index histogram and Cellular) and two classification methods (SVM and OLVQ1). The recognition target, the dimension number of three features, and the parameter choices of classification methods are the same as the preliminary experiment. Among six image data sets, five data sets are used as the training data, and the one data set is used as the test data. The experiments to recognize each image data set are performed six times. Each test data is recognized by using weights calculated based on the algorithm of Adaboost.M1 from training data. Table 2 shows the average recognition rates for test data by six weak classifiers and the recognition results by Adaboost.M1. From Tab.2, the recognition results by Adaboost.M1 are better than the averages of the recognition results by each weak classifier. The recognition rate is improved, and the average of the recognition result by Adaboost.M1 is 88.88%.

Tab. 3 the numbers of misrecognition times and misrecognized character types

Misrecog. in all expt.	0/1,228
Misrecog. in 5 expt.	14/1,228
Misrecog. in 4 expt.	15/1,228
Misrecog. in 3 expt.	93/1,228
Misrecog. in 2 expt.	271/1,228
Misrecog. in 1 expt.	835/1,228

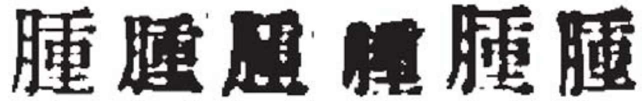

Fig.4 : a character type where a part of the image is filled by bleeding of ink

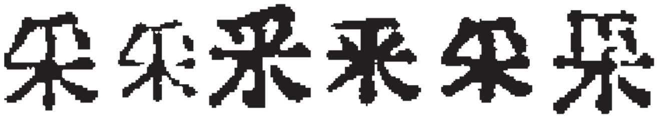

Fig.5 : a character type where a part of the character is different because of difference publishers

4.2 Misrecognition analysis

In the recognition experiments, the number of misrecognized character types is 1,228. Table 3 shows the number of times to misrecognize each character type and the number of the misrecognized character types.

There is no character type misrecognized in all six experiments. It is considered that the misrecognition of one or two times is caused by difference of the training data. Next, we analyze the character types which are misrecognized in more than half of the six experiments.

The character types misrecognized in three experiments include those in which two or more images out of the six images are partially missing due to blur or those in which parts of the images are collapsed due to bleeding. Figure 4 shows a character type in which a part of the images are collapsed due to bleeding of ink. In the case of poor scanning, the misrecognition occurs because some character cannot be correctly learned.

The character types misrecognized in four experiments include those in which three or more images are not readable by eyes due to bleeding of ink or those in which some parts of character are different because of the font difference. Figure 5

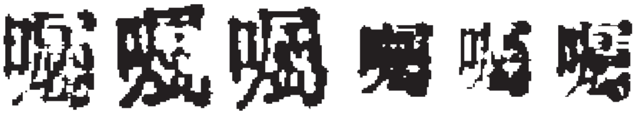

Fig.6 : a character type where most characters are hard to be recognized

shows examples of a character type in which a part of characters is different because of different publishers. In the case of the character types with different publishers, it is considered that recognition of test data is difficult because of learning in different axis line. In addition, in the case of the character types which are not readable by eye due to bleeding of ink, it is considered that the recognition is difficult because the major contents of the features are not learned enough.

The character types misrecognized in five experiments include those in which there are extremely resembling but different characters and those in which most characters are hard to be recognized because the character image is largely deformed due to bleeding and blurring of ink among six image data. Figure 6 shows a character type in which most characters are hard to be recognized. In the case of extremely resembling character types, it is considered that misrecognition occurs because the difference by character stroke thickness does not clearly appear on the character image data. In addition, in the case of the character types in which characters in which majority account for characters in which most characters are hard to be recognized, it is considered that the misrecognition occurs because the features common to the character types are not correctly extracted.

It is considered that major character types misrecognized by Adaboost.M1 are due to the image quality such as the thickness of strokes, ink bleeding, and blurring in character image data. Compared with the common misrecognition which occurs when PDC feature, weighted direction index histogram feature, and cellular feature are used by SVM, the number of misrecognitions is decreasing while the misrecognition range is overlapped [10]. Therefore, the recognition rate is improved by including OLVQ1 into Adaboost.M1. Moreover, when OLVQ1 is used as a classification method, it is considered that there is a range in which it cannot cover the misrecognition by SVM. Thus, it is conceivable that the recognition rate improves by preparing other weak classifiers which can recognize character images with poor quality to perform better ensemble learning using Adaboost.M1.

To deal with a character type with poor image quality, it is necessary to expand the feature space by learning the same character types in other books. Therefore, we should increase the learning data to improve the recognition rate.

5 Conclusions

In this paper, to further improve the recognition rate of early-modern Japanese printed characters, we proposed the early-modern Japanese printed character recognition method by using ensemble learning. As the features of character image, three features including PDC, weighted direction index histogram feature and Cellular are applied, and two classification methods including SVM and OLVQ1 are used. The early-modern Japanese printed characters are recognized by performing the ensemble learning according to Adaboost.M1 algorithm using six weak classifiers composed of three features and two classification methods.

As the preliminary experiment, we performed recognition experiments by each weak classifier. When the three features and OLVQ1 are used, the rate of the character types recognized by using OLVQ1 among the character types misrecognized by using SVM is approximately 71.80%. In addition, it became clear that OLVQ1 can recognize the character types which can distinguish the stroke lines even when the quality of the prepared image data is not good. From the experiment, it is considered that the weak classifiers with OLVQ1 compensate for the misrecognition of the weak classifiers with SVM. Therefore, the six weak classifiers are effective for ensemble learning.

From the recognition results, the average of the recognition results by Adaboost.M1 was 88.88%. The recognition rate is improved by recognizing by Adaboost.M1 with OLVQ1 added. Moreover, the results of misrecognition analysis reveal that the character image quality is often the cause of misrecognition. When OLVQ1 is used as a weak classifier, there is a range in which it cannot cover the misrecognition by the weak classifier of SVM. For that reason, in future, it is conceivable that recognition rate improves by preparing other weak classifiers which can recognize the character images with poor quality to perform ensemble learning using Adaboost.M1.

Our future work include that we investigate other classifiers that recognize characters with poor printing quality and we obtain other learning data with high variety of different publishers and publishing years.

Acknowledgment

This work is partially supported by Grant-in-Aid for scientific research from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) No.26280119.

References

[1] National Diet Library http://www.ndl.go.jp/

[2] Digital Library From the Meiji Era http://kindai.ndl.go.jp/

[3] National Diet Library Digital Collections http://dl.ndl.go.jp

[4] National Diet Library full text conversion demonstration experiment report http://www.ndl.go.jp/jp/aboutus/digitization/fulltextreport.html (in Japanese)

[5] Fukuo, M., Takata, M. and Joe, K.:"The Kanji character recognition evaluation for the modern book of the same publisher"(in Japanese). The Information Processing Society of Japan. Mathematical Modeling and Problem Solving(MPS), 26:1–6 (2012).

[6]Ishikawa, C.， Ashida, N.， Enomoto, Y.， Takata, M.， Kimesawa, T. and Joe, K : Recognition of Multi-Fonts Character in Early-Modern Printed Books，Proceedings of International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA2009)，Vol. Ⅱ， pp． 728-734(2009).

[7] Fukuo, M.， Enomoto, Y.， Yoshii, N.， Takata, M.， Kimesawa, T. and Joe, K : Evaluation of the SVM based Multi-Fonts Kanji Character Recognition Method for Early-Modern Japanese Printed Books，Proceedings of The 2011 International Conference on Parallel and Distributed Processing Technologies and Applications (PDPTA2011)， Vol.Ⅱ， pp． 727-732(2011).

[8] Kosaka, K, Fujimoto, K, Ishikawa, Y, Takata, M and Joe, K． : Comparison of Feature Extraction Methods for Early-Modern Japanese Printed Character Recognition (PDPTA2016)，pp.408-414(2016).

[9] Hagita, N., Naito, S. and Masuda, I.: Handprinted Chinese Characters Recognition by Peripheral Direction Contributivity Feature (in Japanese), IEICE, Vol.J66-D, 10, pp.1185-1192(1983).

[10] Tsuruoka. S, Kurita, M., Harada, T., Kimura, F. and Miyake, Y: Handwritten “KANJI” and “HIRAGANA” Character Recognition Using Weighted Direction Index Histogram Method, IEICE, and Vol.J70-D，No.7，pp.1390 – 1397 (1987) .

[11] Oka, R.: Handwritten Chinese-Japanese Characters Recognition Using Cellular Features (in Japanese), IEICE, Vol.J66-D, No.1, pp.17–24 (1983).

[12] Cristianini，N． and Shawe-Taylor，J. : Support vector machine Introduction，Kyoritsu Publisher (2005).

[13] Kohonen, T., Hynninen, J., Kangas, J., Laaksonen, J. and Torkkola, K.: LVQ_PAK: The Learning Vector Quantization Program Package, Technical Report A30, Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finland (1996).

[14] Freund, Y. and Schapire, R. E.: Experiments with a new boosting algorithm, Proceedings of the Thirteenth International Conference on Machine Learning, pp.148-156 (1996).

[15] National Institute of Informatics ： https ://www.jisc.go.jp