Techniques for Highly Accurate Optical Recognition of Handwritten Characters and Their Application to Sixth Chinese National Population Census

● Danian Zheng ● Jun Sun ● Hao Yu ● Satoshi Naoi

Highly accurate optical character recognition (OCR) of handwritten characters is still a challenging task, especially for languages like Chinese and Japanese. To improve the accuracy, we developed four techniques for enhanced recognition: character recognition based on modified linear discriminant analysis (MLDA), subspace-based similar-character discrimination, multi-classifier combination, and mutual-information-based adaptive rejection. They were applied by the Chinese government to the Sixth National Population Census in 2010. By combining address and nationality information, they achieved an accuracy of over 99% with a low rejection rate. This was the first time that optical recognition of handwritten Chinese characters had been used on a large-scale in the Chinese census project.

1. Introduction

China, the country with the largest population, has about 1.4 billion people. Every ten years, a population census is carried out by the National Bureau of Statistics of China (NBSC) to collect the latest population statistical data. In the Sixth National Population Census, which started in November 2010, millions of census takers spent tens of days going all over the country to collect census data, which was recorded by hand on paper. Such large-scale statistical data on a country's population is crucial for the development of its society and economy.

Handling paper input has been hugely labor intensive and costly. Recent improvements in the equipment used and the introduction of optical character recognition (OCR) techniques have reduced the amount of labor needed and the cost. Most of the manual data input work is now done using automatic machine-based processing—automatic scanning and automatic character recognition. Advanced optical techniques for recognizing handwritten characters developed at the Fujitsu Research & Development Center (FRDC) were used to support the Sixth National Population Census. An example household data collection form is shown in **Figure 1**. The handwritten numerals and Chinese characters to be recognized can number in the hundreds. Examples of the numerals and characters used for the person count, address, nationality, and name are shown in **Figure 2**. This was the first time that optical recognition of handwritten Chinese characters had been used on a large scale in the Chinese census project.

Out in the field, where a population census begins, good writing quality cannot be ensured. Moreover, handwritten Chinese character recognition (HCCR) is still a challenge in the OCR research field.[1)] The low quality of the handwritten Chinese characters makes the use HCCR for a census quite difficult. We have developed four techniques for enhancing the quality of HCCR: character recognition based on modified linear discriminant analysis (MLDA), subspace-based similar-character discrimination, multi-classifier combination, and

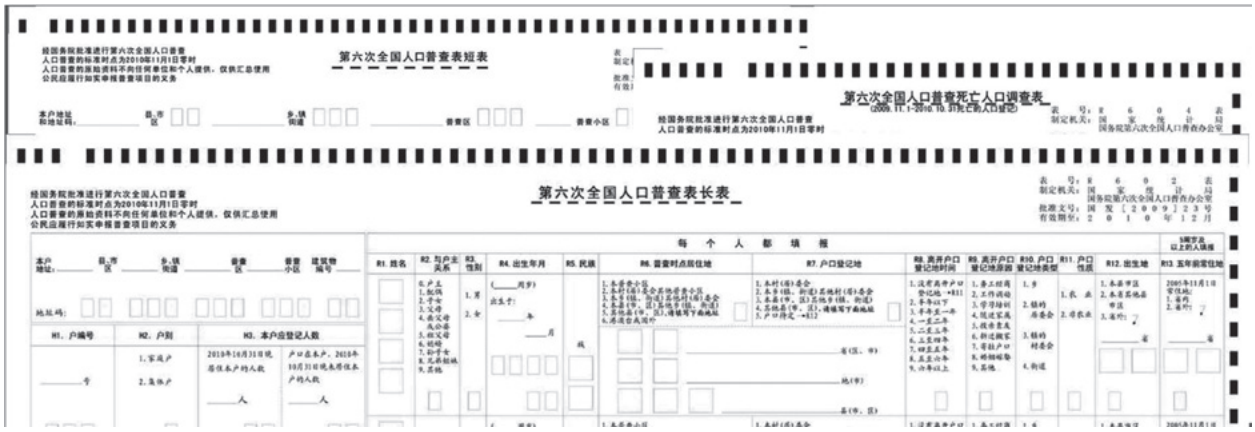

Figure 1
Form used to collect household data in Sixth Chinese National Population Census.

(a) Numeral recognition (b) Address recognition (c) Nationality recognition (d) Name recognition

Figure 2
Example numeral and characters used for person count, address, nationality, and name.

mutual-information-based adaptive rejection. The handwritten character recognition level they achieved by combining address and nationality information met the strict population census requirements.

In this paper, we describe our handwritten character OCR techniques and their application to the Sixth Chinese National Population Census. In section 2, the form processing system to which these techniques were applied is described. Section 3 describes our evaluation of their performance. Section 4 concludes with a summary of the key points.

2. Handwritten character recognition system

2.1 System overview

The form processing system to which our techniques were applied comprises three main stages, as shown in **Figure 3**.

- Form scanning
- Character recognition
- Visual check (human verification)

In the form scanning stage, Fujitsu scanners are used to automatically create images of the paper forms at very high speed. Next, a character recognition module scans each form image and recognizes the handwritten characters. Very high accuracy is achieved in the third stage by having human checkers visually check the suspicious recognition results. Since only the suspicious results are checked, the amount of human labor needed is greatly reduced. The suspicious recognition results are detected by using a technique called "recognition rejection." Hence, character recognition and rejection are

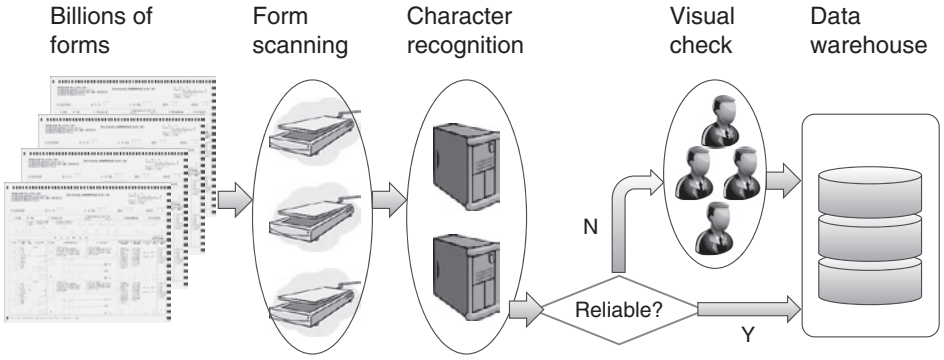

Figure 3
Form processing system used for population census.

equally important. Our technical target is to achieve the lowest error rate with minimum human verification.

2.2 Handwritten numeral recognition

Since most of the information collected is in numerical form, numeral recognition accuracy is crucial to quality for most parts of the collected data. In our handwritten numeral recognition, a support vector machine (SVM)[2), 3)] based numeral classifier and multi-classifier combination[4)] are used to enhance recognition. The final accuracy was higher than 99.9%, which guaranteed data quality.

2.3 Handwritten address recognition

Each handwritten address has three elements: province, city, and county, e.g., "河南 (province) - 信阳 (city) - 淮滨 (county)."

According to the latest Chinese standard address table, there are 2863 standard addresses. From this table and the user's requirements, 35 words were collected for the province lexicon, 345 words were collected for the city lexicon, and 2895 words (all standard county words plus some old ones) were collected for the county lexicon. Then a province dictionary with 51 Chinese characters, a city dictionary with 350 Chinese characters, and a county dictionary with 1162 Chinese characters were made from the province, city, and county lexicons. Since the address characters had been written in the specified square boxes on the form, character segmentation was not a problem.

Errors in filling out the form were taken into account. Of the 54 093 handwritten address samples, 7.8% had an error: incomplete address, incorrect province, incorrect or missing city, incorrect or missing county, etc. Therefore, the three elements in the address were processed independently, and then a most probable address result was deduced from the combined results.

Our approach to handwritten address recognition is illustrated in **Figure 4**. The characters in the province, city, and county elements are recognized in parallel. For every character, the recognition confidences for all candidates in the related dictionary are stored in a character confidence list. The word confidences for all candidates in the related lexicon are calculated for each element and stored in one word confidence list. The word confidence is equal to the average of the character confidences. Then, the address confidences for all candidates in the address table are calculated and stored in one address confidence list. Address confidence is defined as the weighted sum of the province word, city word, and county word confidences. The weights are proportional to the number of characters in the elements. Address confidences are sorted from largest to smallest, and the address with the highest confidence is output as

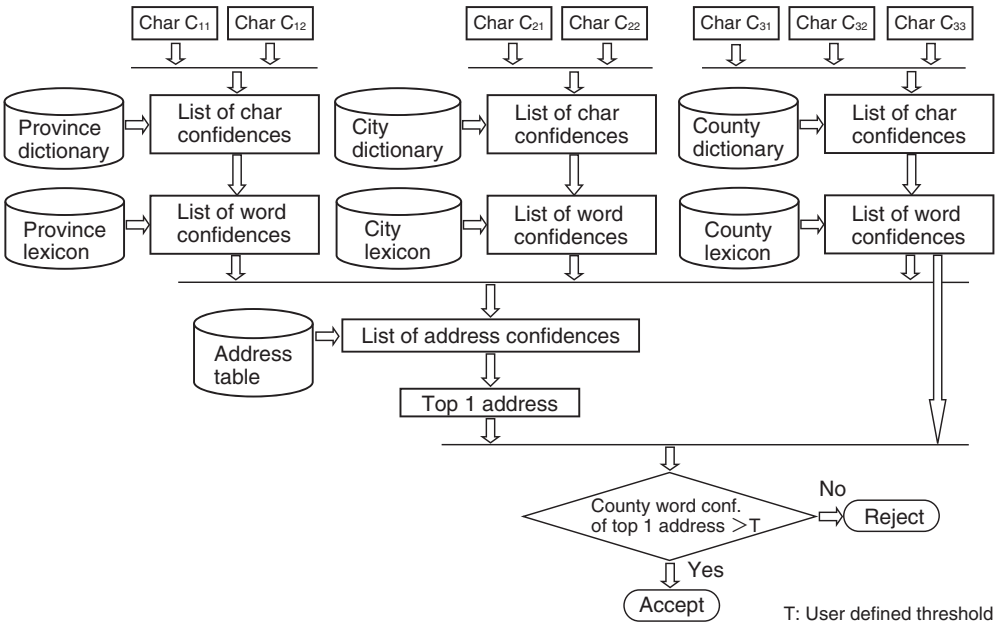

Figure 4
Approach to handwritten address recognition.

the final recognition result.

The rule used for rejection is critical to achieving a high recognition rate with a low rejection rate. Whether to accept or reject the top 1 address candidate is decided by comparing its county word confidence with a user defined threshold. If the confidence is less than the threshold, the recognition result is rejected; otherwise it is accepted. Comparison of this rejection rule with rejection based on the address confidence and rejection based on the minimum of the word confidences of the top 1 address showed that the rule based on the county word confidence produces the best results.

Figure 5 (a) shows a standard, complete address and the results of recognition (the top 5 addresses with the county word confidence the highest for the top 1 address. Typical errors in handwritten addresses are shown in **Figure 5 (b)–(f)**. **Figure 5 (b)** shows an incomplete address with the county word missing. **Figure 5 (c)** shows another incomplete address with the city "荆州" missing between the province "湖北" and the county "洪湖." **Figure 5 (d)** shows a mismatched address, where the city should be "烟台," not "东营." **Figure 5 (e)** shows an address with the county entered as the city, where "青州" is a county and "黄楼镇" is a town in that county. **Figure 5 (f)** shows a modified address, where even without the modified city, "祈州," the address was correctly recognized. From **Figure 5 (b)–(f)**, we can see that our recognition system can handle typical errors in a practical application.

2.4 Handwritten nationality recognition

The 1.4 billion people in China are categorized into 56 nationalities. The names of these nationalities comprise 1 to 4 Chinese characters, not including the suffix "族" (Zu); the nationalities include "汉族" (Han), "蒙古族" (Mongolian), and "乌兹别克族" (Uzbek). To save space, the form has only two vertical rectangular boxes for nationality, and only one or two of the characters for nationality must be entered. Therefore, the recognition character set contains less than 100 characters, which improves recognition accuracy. However, abbreviation

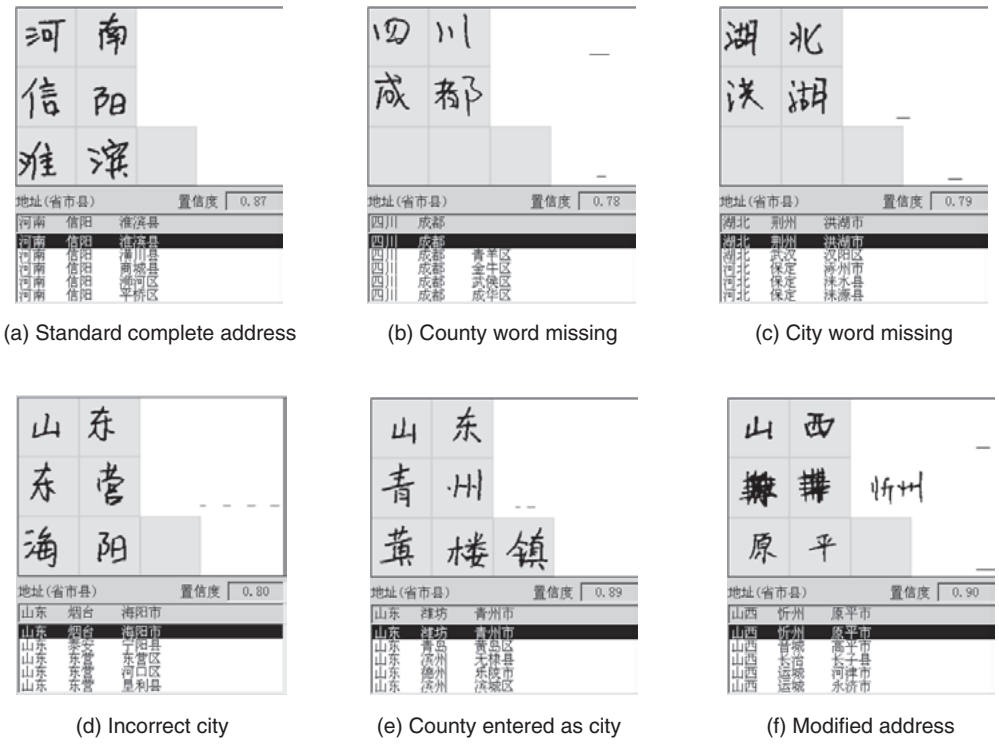

Figure 5
Examples of handwritten addresses and recognition results.

and alias recognition increases recognition difficulty. For example, the abbreviations for "蒙古族" (Mongolian) and "维吾尔族" (Uyghur) are "蒙" and "维." "摩梭族" (Mosuo) is an alias of "纳西族" (Naxi). Although the suffix "族" (Zu) is not supposed to be entered on the form, it often is. Therefore, its recognition must be supported. **Figure 6** shows examples of characters input for nationality.

There are thus two cases for handing nationality strings: single-character strings and double-character strings. To enable these two cases to be handled, the nationality names are divided into two lexicons. One lexicon contains words of single-character nationality names, and the other contains words of double-character nationality names. When only one character is entered into the nationality field, the matching is done using the single-character lexicon, and the result depends entirely on single character recognition. When two characters are entered, the double-character lexicon is used. If "族" (Zu) is found to be in the recognition candidate list for the second character, both the single- and double-character lexicons are used for nationality matching, as diagrammed in **Figure 7**. For the census data, the nationality recognition accuracy was greater than 99.9%.

2.5 Low-quality handwritten Chinese character recognition

In the HCCR tasks described above, address and nationality information is used to ensure the feasibility of the solutions. The bases for each task are the same: highly accurate character recognition. This is not easy because of the large number of categories in HCCR. Moreover, an actual census will include many cases of cursive handwriting and noise. The millions of Chinese census takers had greatly different writing styles, so uniform writing quality was impossible to obtain. All of these are uncontrollable quality

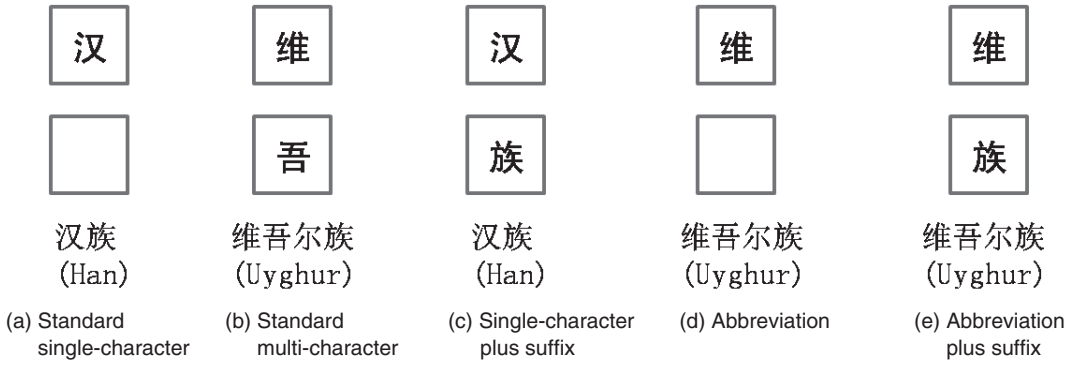

Figure 6
Examples of characters input for nationality.

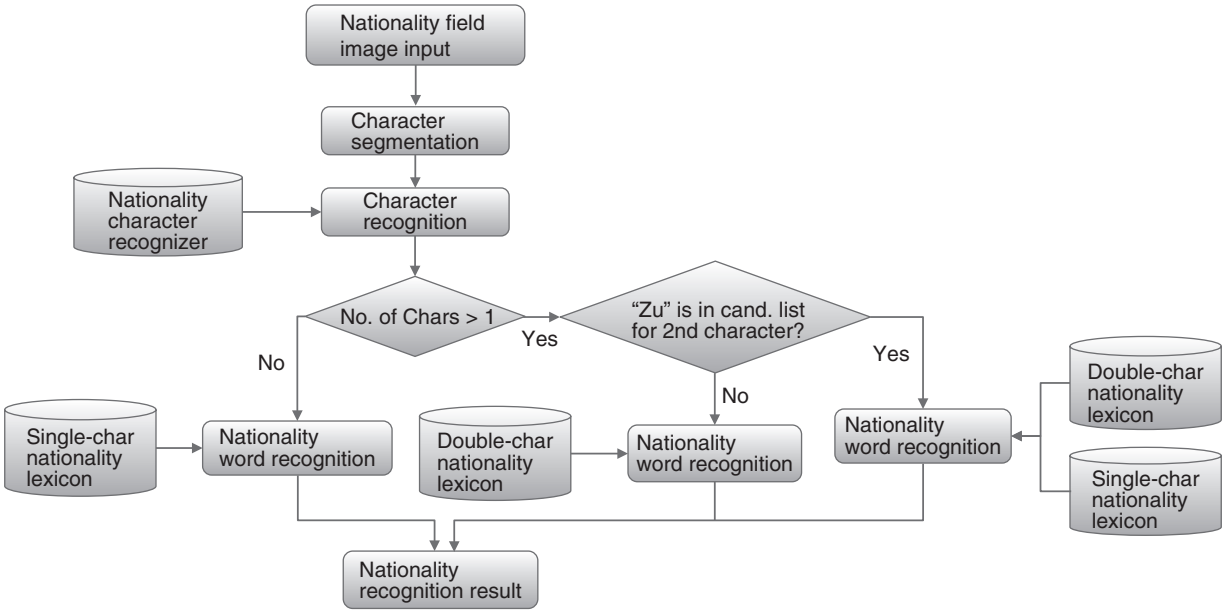

Figure 7
Approach to nationality recognition.

factors. Improving the performance of low-quality handwritten Chinese character recognition is thus indispensable. The overall performance of HCCR for a census depends on both single character recognition and character recognition rejection. To enhance both components, four main techniques were researched and applied.

- Character recognition based on MLDA
- Subspace-based similar-character discrimination
- Multi-classifier combination
- Mutual-information-based adaptive rejection for character recognition

With the help of these techniques, a recognition system based on algorithms developed by the FRDC outperforms one using a conventional method. **Figure 8** illustrates how the new techniques improve the overall performance. Not only are more characters correctly recognized in the recognition stage, but also more incorrectly recognized characters are rejected in the rejection stage. The four techniques are described in detail below.

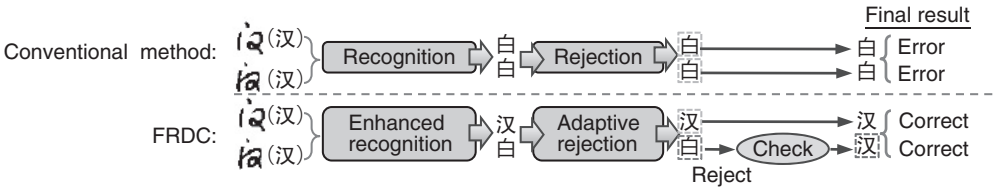

Figure 8
Improved handwritten Chinese character recognition.

2.5.1 MLDA-based character recognition

In character classification, LDA[5] is usually applied to feature selection to reduce the character feature from hundreds of dimensions to fewer dimensions. This removal of redundant information increases computation efficiency. Moreover, discriminative information is enhanced in feature selection, which improves classification accuracy. Conventional LDA takes the maximization of the global between-class variance as the optimization objective function, which results in some classes being difficult to separate after transformation. Thus, we introduce character similarity as one factor in covariance calculation, which tends to push similar characters away from each other after feature transformation. This means that better separability between similar characters is preserved in the new feature space, so better classification is obtained.

As **Figure 9** illustrates, two classes are mixed after conventional LDA transformation, but all three classes are separable along the MLDA projection axis. Higher classification accuracy is thus ensured.

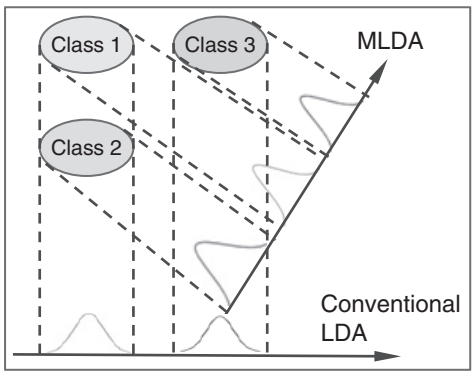

Figure 9
Modified linear discriminant analysis.

2.5.2 Subspace-based similar-character discrimination

Analysis of the character recognition errors revealed that most of the classification errors were among similar characters. Typically, the performance of similar-character discrimination determines the accuracy of the final recognition. The normal character recognizer is expected to have good overall performance for the entire character set. However, Chinese character recognition is a large category recognition problem. Global consideration for a large category is unable to handle all similar-character patterns well. By identifying potential similar-character patterns from recognition and then designing special classifiers for these similar-character patterns, we can convert large category recognition on the entire character set to small category recognition among similar characters. This should result in better recognition of similar characters, which will improve overall character recognition performance.

We observed that the correct class label could generally be found in the first two recognition candidates. On the basis of this observation, we designed a pair-wise similar-character classifier to handle suspicious similar-character pairs in the first two recognition candidates. In particular, when a similar-character pair was found in the first two candidates, the pair-wise classifier for that pair was applied. In this way,

large-category recognition was converted to two-class recognition. The corresponding recognizers were respectively dubbed "global recognizer" and "local recognizer." Since local discriminative information is used, the local recognizer always outperforms the global recognizer for a similar-character pair. Combining the result of the local recognizer with the incorrect classification result of the global recognizer corrects the error, as shown in **Figure 10**.

2.5.3 Multi-classifier combination

To further improve character classification accuracy and robustness, we use a multi-feature classifier combination to construct a stronger classifier. In general, the larger the complementariness among the features chosen, the higher the performance after combination. On the basis of the principle of diversity theory, we chose three features: contour direction feature (CDF),[6] pseudo-gray gradient direction (GGRD),[6] and weighted direction code histogram (WDCH).[7] After a weighted voting rule is applied to the combination, not only classification accuracy, but also generalization capability, is enhanced significantly. This is very important for actual applications. **Figure 11** show an example application of the multi-classifier combination technique.

2.5.4 Mutual-information-based adaptive rejection

A population census has a very strong requirement for data precision, and automatic character recognition alone is not sufficiently accurate. Although the error rate for human verification is very close to zero, its cost is unacceptable. By identifying suspicious recognition results and using human verification only for those results, an appropriate trade-off between data precision and cost can be obtained. The mutual-information-based adaptive rejection technique is used to identify suspicious recognition results. The magnitude of the suspicious results identified corresponds to the workload for human verification, so the objective of rejection is to achieve the lowest error rate with the lowest number of rejections.

Our proposed mutual-information-based adaptive rejection technique[8] differs from

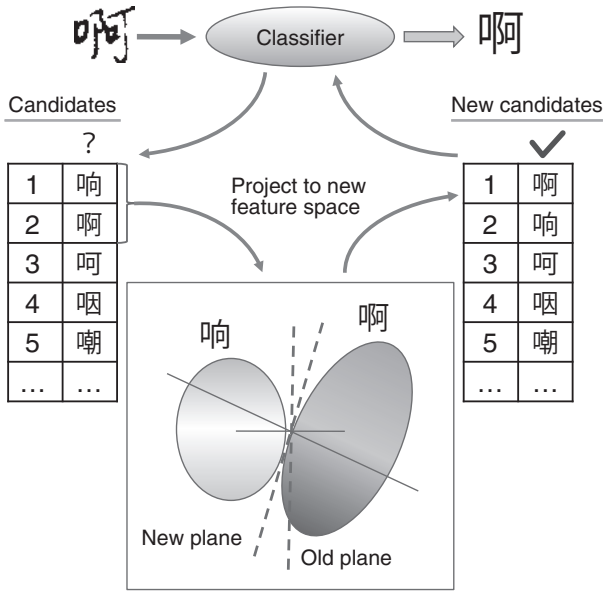

Figure10
Subspace-based similar-character discrimination.

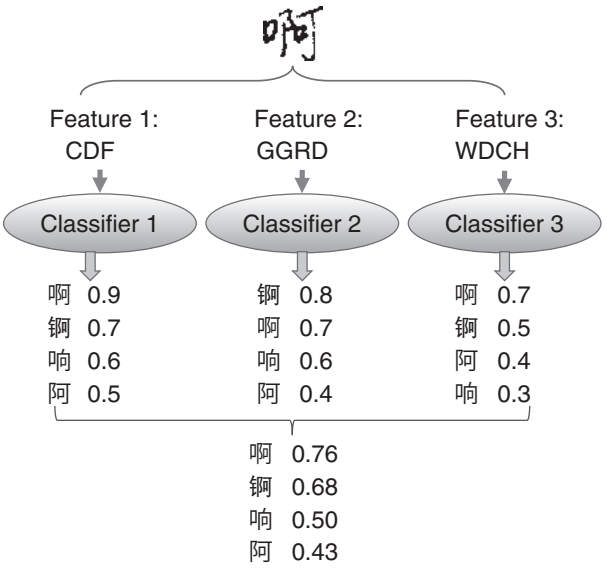

Figure11
Example application of multi-classifier combination.

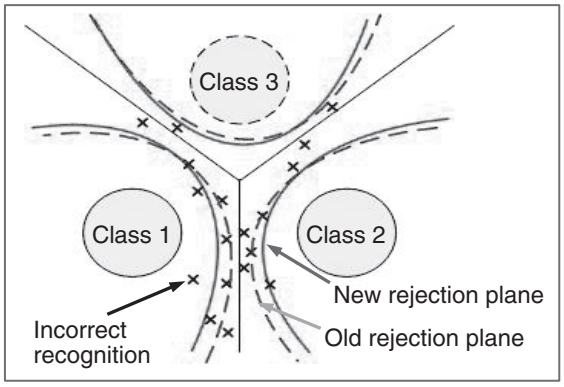

Figure 12
Difference between adaptive and conventional rejection methods.

conventional rejection methods that use the same rejection threshold for all samples. The proposed technique uses the local information of samples to optimize the rejection threshold. Samples with the same class pair in the first two recognition candidates are treated as being in the same sample category. Mutual-information-based correlation is used to select suitable sample categories. Then, category-dependent rejection parameters are optimized on the basis of maximum entropy. This means that the rejection threshold is adaptive to each sample, which results in the optimal rejection result. **Figure 12** illustrates the difference between the adaptive and conventional rejection methods. More incorrect recognition samples are rejected with the adaptive method for the same rejection rate as with the conventional method.

3. Evaluation

We evaluated our algorithms and techniques by applying them to a large data set. **Table 1** summarizes the evaluation results. Compared with HCCR, handwritten numeral recognition had the best performance due to the smaller number of categories for recognition. The corresponding error rate and rejection rate were both very low. The address and nationality recognition results indicate the significant benefit of using address and nationality information. Although single-character recognition is difficult for HCCR, utilization of this information enables highly accurate address and nationality recognition. In short, the recognition performance for each task met the customer's requirement.

Table 1
Handwritten recognition evaluation results.

Task	Recognition rate (%)	Rejection rate (%)	Data set size (no. of samples)
Numeral recognition	99.98	0.62	1.0 M
Address recognition	99.60	4.10	0.5 M
Nationality recognition	99.92	2.70	1.5 M

4. Conclusion

The FRDC developed advanced handwritten character optical recognition techniques to support the Sixth Chinese National Population Census, which required recognition of handwritten numerals, addresses, nationalities, and names. To handle the low quality of the handwritten Chinese characters, four main techniques were developed: character recognition based on linear discriminant analysis, subspace-based similar-character discrimination, multi-classifier combination, and mutual-information-based adaptive rejection. By using address and nationality information, they achieved an accuracy of over 99% with a low rejection rate, thus meeting the strict requirements for a population census. This is the first time that Chinese character recognition technology has been used on a large scale in the Chinese census project. These endeavors have helped to greatly increase the efficiency and cut the cost of performing a population census.

References

1) M. Cheriet, N. Kharma, C. Liu, and C. Suen: Character Recognition Systems: A Guide for Students and Practitioners. Wiley-Interscience, 2007.
2) V. Vapnik: The Nature of Statistical Learning Theory. Springer, New York, 1995.

3) C. Burges: A tutorial on support vector machines for pattern recognition. (In *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp. 121–167), Kluwer Academic Publishers, Boston, 1998.
4) C. Liu: Classifier combination based on confidence transformation. *Pattern Recognition*, Vol. 38, No. 11, pp. 11–28 (2005).
5) R. O. Duda et al.: Pattern Classification. Wiley, 2000.
6) C. Liu et al.:Handwritten digit recognition: investigation of normalization and feature extraction techniques. *Pattern Recognition*, Vol. 37, No. 2, pp. 265–279 (2004).
7) F. Kimura et al.: Improvement of handwritten Japanese character recognition using weighted direction code histogram. *Pattern Recognition*, Vol. 30, No. 9, pp. 1329–1337 (1997).
8) Y. Zhu et al.: Rejection Optimization Based on Threshold Mapping for Offline Handwritten Chinese Character Recognition. The 12th International Conference on Frontiers in Handwriting Recognition, 2010, pp. 72–77.

Danian Zheng
Fujitsu Research and Development Center Co., Ltd.
Mr. Zheng is engaged in the research and development of document-image-processing and optical-character-recognition technologies.

Hao Yu
Fujitsu Research and Development Center Co., Ltd.
Mr. Yu is engaged in the research and development of information-processing and smart-grid technologies.

Jun Sun
Fujitsu Research and Development Center Co., Ltd.
Mr. Sun is engaged in the research and development of document-image-processing and optical-character-recognition technologies.

Satoshi Naoi
Fujitsu Research and Development Center Co., Ltd.
President