2013 12th International Conference on Document Analysis and Recognition

An Empirical Comparative Study of Online Handwriting Chinese Character Recognition:Simplified v.s.Traditional

Yan Gao, Lianwen Jin+, Weixin Yang
School of Electronic and Information Engineering
South China University of Technology
Guangzhou, China
thegoldfishwang@163.com, +lianwen.jin@gmail.com, wxy1290@163.com

***Abstract*—There are two forms of Chinese character widely used in the world, simplified Chinese and traditional Chinese. Simplified Chinese is mainly used in Chinese mainland, Singapore, Malaysia and other Southeast Asian regions, while traditional Chinese is mainly used in Hong Kong, Macao, Taiwan, Japan and so on. The simplified Chinese is transformed from traditional Chinese by simplifying the character structure and reducing the stoke number of many traditional characters, while a few remain unchanged. The simplified Chinese reduces the memory storage and makes handwriting become more convenient, while it also brings the problem that many characters are so similar that they are difficult to be recognized. In this paper, an empirical study of the handwriting simplified/traditional Chinese character recognition are carried out in order to compare the difference between these two forms of Chinese characters. The experimental results based on SCUT-Couch2009 database show that the handwriting recognition accuracy of traditional Chinese is higher than simplified Chinese, for both unchanged part and changed part. This interesting finding may bring us some cues on the issue of confusable Chinese character recognition for further study.**

Keywords-handwriting recognition; simplified Chinese; traditional Chinese

I. INTRODUCTION

Chinese is one of the oldest writing languages in the world and may be the only ancient language still in use today [1]. As an ideographic language, Chinese inherits the long history and culture of Chinese nation during thousands of years. It is the language most widely spread and used in the world, especially has the far-reaching implications on China and the whole East Asian. The total number of Chinese character is about one hundred thousand. However, most of Chinese characters are variant and unfrequently used characters, only thousands of them are in daily use. According to statistics, 3,755 frequently used characters can cover more than 99% of the written material in China.

Nowadays, there are two forms of Chinese character widely used in the world, simplified Chinese and traditional Chinese. Simplified Chinese was used in the world since 1950s, which was proposed as the modern Chinese writing standard after the founding of New China. However, the traditional Chinese is still used in Hong Kong, Macao, Taiwan now. Simplified Chinese is generated from traditional Chinese by simplifying the character structure and reducing the stoke number of most traditional characters. Simplified Chinese has many advantages: it reduces the character stoke numbers; simplifies the character structure; and reduces the number of frequently used characters. All these advantages make Chinese easy to learn, read and write. Simplified Chinese reduces the memory storage difficulty and makes handwriting become more convenient. However, as an ideographic language, simplified Chinese reduces the inner meaning of Chinese characters, furthermore, it results in many similar characters difficult to recognize, such as “斤” and “厅”, “渚” and “诸”, “近” and “迈”, the traditional forms are“斤” and “廳”, “渚” and “諸”, “近” and “邁”, which is significantly easier to distinguish.

In recent years, as an important research direction in pattern recognition field, handwriting Chinese character recognition (HCCR) technology has made great progress and the recognition accuracy achieved more than 98% on certain constrained databases [2-3] and more than 95% on realistic unconstrained databases [4-6]. However, it was noted in [6] that, the recognition accuracy of simplified Chinese dataset is about 2% less than traditional Chinese dataset, although the two datasets are collected in the same experiment environment and contributed by same writers. To investigate this issue, we make a comprehensive comparative study on the issue of simplified HCCR against the traditional HCCR.

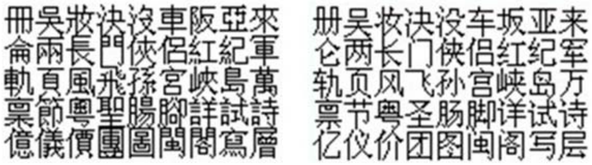
The rest of this paper is organized as follows: Section II introduces the simplified Chinese and traditional Chinese in details. Section III presents the handwriting character recognizer we used in our experiments. Section IV gives the experimental results and compares the difference between the simplified Chinese and traditional Chinese. Conclusions are summarized in Section V.

II. SIMPLIFIED CHINESE AND TRADITIONAL CHINESE

Due to historical and political reasons, simplified Chinese character is currently mainly used in mainland China while traditional Chinese character is used in areas like Taiwan, Hong Kong and Macao. GB2312-80 is one standard of simplified Chinese characters, while BIG5 is the standard of

1520-5363/13 $26.00 © 2013 IEEE
DOI 10.1109/ICDAR.2013.176

traditional characters. Although most of the simplified Chinese characters are simplified from their traditional characters, quite a few parts of the characters are unchanged. Examples of some simplified and traditional characters are shown in Figure 1. Examples of some unchanged characters in both sets are shown in Figure 2.

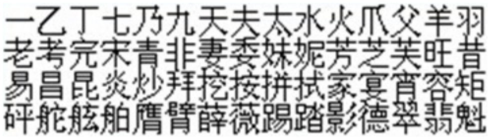


冊吳妝決沒車阪亞來
侖兩長門俠侶紅紀軍
軌頁風飛孫宮峽島萬
稟節粵聖腸腳詳試詩
億儀價團圖閩閣寫層

册吴妆决没车坂亚来
仑两长门侠侣红纪军
轨页风飞孙宫峡岛万
禀节粤圣肠脚详试诗
亿仪价团图闽阁写层

(a) traditional Chinese (b) simplified Chinese

Figure 1. The characters in GB and BIG5



一乙丁七乃九天夫太水火爪父羊羽
老考完宋青非妻委妹妮芳芝芙旺昔
易昌昆炎炒拜挖按拼拭家宴宵容矩
砰舵舷舶膺臂薛薇踢踏影德翠翡魁

Figure 2. The unchanged characters in GB and BIG5

Simplified Chinese is simplified from traditional Chinese, in order to make Chinese easy to learn, read and write. There are four methods mainly used in the process of generating the simplified Chinese characters [7].

The first method is simplifying the structure of traditional characters, including several aspects such as following. (1) A character is replaced by another existing character with same sound. (2) Some characters preserve the basic outline. (3) Considering the general cursively writing habits, some of the characters are replaced by their cursive forms. (4)Some characters are simplified by replacing or omitting a complex component of the character. The new component can be either a simple symbol or a near-sound component. (5) Some new characters are created to replace the traditional one. (6) Some ancient forms and variations are adopted. Table I gives some examples in order to show how to simplify the structure of traditional characters.

TABLE I. EXAMPLES OF STRUCTURAL SIMPLIFICATION METHOD

Detail	Forms	Examples				
Same Sounds	Simplified	谷	丑	苹	松	只
	traditional	穀	醜	蘋	鬆	隻
Similar basic outline	Simplified	飞	龟	齿	夺	门
	traditional	飛	龜	齒	奪	門
Use cursive form	Simplified	书	长	当	韦	乐
	traditional	書	長	當	韋	樂
Replace component	Simplified	对	观	叹	邻	蜡
	traditional	對	觀	嘆	鄰	蠟
Omit component	Simplified	广	习	宁	业	气
	traditional	廣	習	寧	業	氣
Create a new character	Simplified	护	惊	艺	响	泪
	traditional	護	驚	藝	響	淚
Adopt ancient form	Simplified	尘	从	众	云	与
	traditional	塵	從	眾	雲	與

Second, some characters are derived based on simplified character components. Because of the similarity of characters, simplified characters can be created by systematically simplifying components. For instance, “單” is simplified to “单”, so from “彈”, “嬋” and “驒”, “弹”, “婵” and “骣” can be made. “頁” is simplified to “页”, then “顏”, “順” and “額” convert to “颜”, “顺”, “额”. And “食” is simplified to “饣”, thus “飯”, “飽” and “餃” convert to “饭”, “饱” and “饺”.

Third, eliminate the variants of the same characters. A set of variant characters often sound the same or share the same meaning, so the simplest one in form is chosen to represent this set of characters and the rest are abandoned. For example, “虖”, “嘑” and “謼” are replaced by “呼”. And “災”, “烖” and “菑” are replaced by “灾”. And “獃” and “騃” are replaced by “呆”.

Fourth, adopt new standardized character forms. With this method, characters appear slightly simpler than the old forms, and are as such mistaken as structurally simplified characters mentioned in the first method. Some example are shown as following: “粵” convert to “粤”, “溫” to “温”, “虛” to “虚”, “靜” to “静”, and “換” to “换”.

With these four simplified methods, the average character stroke number per character is reduced from 16 to 10. Less time and energy will be spent in learning and writing these simplified characters. However, on the other hand, the energy-saving effect is limited in the Internet Age and the simplified characters increase ambiguity and lose the original meaning of traditional characters. Simplified characters also cause many similar characters difficult to be recognized.

III. HANDWRITING CHARACTER RECOGNITION TECHNOLOGIES

The general steps of handwriting recognition include preprocessing, feature extraction, dimension reduction, classification and so on [8]. In this paper, we use the compact MQDF handwriting character recognizer [9] in the experiments. The recognizer uses elastic mesh normalization and 8-directional feature extraction. The feature dimensionality is reduced from 512 to 160 by Fisher linear discriminant analysis (LDA). Finally, modified quadratic discriminant function (MQDF) is used for classification.

A. 8-Directional Feature Extraction

As an effective feature extraction method, the 8-directional Feature [10] is widely used in the field of online handwriting Chinese character recognition. The 8-directional feature is computed through the 8-directional vectors of each sampling point. Then 8 directional pattern images are generated accordingly, and the blurred directional features are extracted at 8 × 8 uniformly sampled locations using a Gaussian filter. Finally, a 512-dimensional vector of raw features is formed.

B. Linear Discriminant Analysis

LDA [11] is a supervised learning method that can select the lower dimensional sub-space features with the most discriminating information. Mathematically speaking, LDA can seek directions for efficient discrimination through maximizing the between-class scatter while minimizing the within-class scatter.

C. MQDF classifier

The MQDF classifier proposed by Kimura et al. [12] is widely used in handwriting recognition for its high classification performance. The quadratic discriminant function (QDF) is based on Bayesian decision rule, under the assumption of multivariate Gaussian density for each class. The MQDF is obtained by making a modification to the QDF with PCA transformation and smoothing the minor eigenvalues by a constant, which is shown in (1).

$$g_1(x,\omega_i)=\frac{1}{\delta_i}\left\{\left\|x-\bar{x}_i\right\|^2-\sum_{j=1}^{K}(1-\frac{\delta_i}{\lambda_{ij}})[\phi_{ij}^T(x-\bar{x}_i)]^2\right\}$$
$$+\sum_{j=1}^{K}\log\lambda_{ij}+(D-K)\log\delta_i \qquad (1)$$

where $\bar{x}_i$ denotes the mean vector of class ω_i ; λ_{ij} and ϕ_{ij} denote the eigenvalues and eigenvectors respectively of the covariance matrix of class ω_i ; D is the dimension of $\bar{x}_i$; K is the number of dominant eigenvectors and δ_i is a constant.

IV. Experiments and Analysis

The dataset we used is SCUT-COUCH2009 [6,13], which is a comprehensive database that consists of 11 subsets, including simplified and traditional Chinese characters, words, pinyins, letters, digits, symbols and so on. SCUT-COUCH2009 is collected with PDA (Personal Digit Assistant) and smart phones with touch screens, contributed by more than 190 different persons, resulting in more than 3.6 million handwritten samples. It is worth to note that SCUT-COUCH2009 is currently the only public available database that including the 5,401 BIG5 traditional Chinese character set.

In this paper, we use the GB subset and BIG5 subset of SCUT-COUCH2009 as the simplified Chinese and traditional Chinese datasets respectively. The two datasets are collected in the same experiment environment. GB subset contains 188 sets of handwritten samples of 6,763 Chinese characters in GB2312-80 standard, while BIG5 subset contains 65 sets of 5,401 frequently characters in BIG5 standard. In order to ensure the experiment environment consistent, we randomly select 60 sets for training and 5 sets for testing in the both subsets respectively.

The characters of GB standard are simplified from BIG5 standard. The two character standards can be divided into three parts respectively: (1) unchanged part, where characters in this part are same in both GB and BIG5; (2) synonym part, where characters in this part of GB and BIG5 have the same meaning but different writing form; (3) disjoint part, where in this set, characters in BIG5 does not appear in GB set and vice versa. The details of character number of each part are given in Table II.

TABLE II. CHARACTER NUMBER OF EACH PART

Character set	unchanged part	synonym part	disjoint part	total
GB	3,247	1,687	1,829	6,763
BIG5	3,247	1,687	467	5,401

In the following experiments, we compare the recognition accuracy of the GB and BIG5 datasets, and analysis the difference between the two datasets. Due to that the disjoint parts of GB and BIG5 are unrelated with each other, the experiments are mainly focused on the unchanged part and synonym part to make a fair comparison.

Table III shows the recognition accuracies of GB and BIG5 datasets, where we only combine the unchanged part and synonym part as the training and testing datasets. From Table III it can be seen that the total accuracy of BIG5 is 2.57% higher than that of GB dataset. For both unchanged and synonym parts, the accuracies of BIG5 set are significant higher than that of GB set.

TABLE III. RECOGNITION RATE OF UNCHANGED AND SYNONYM PART

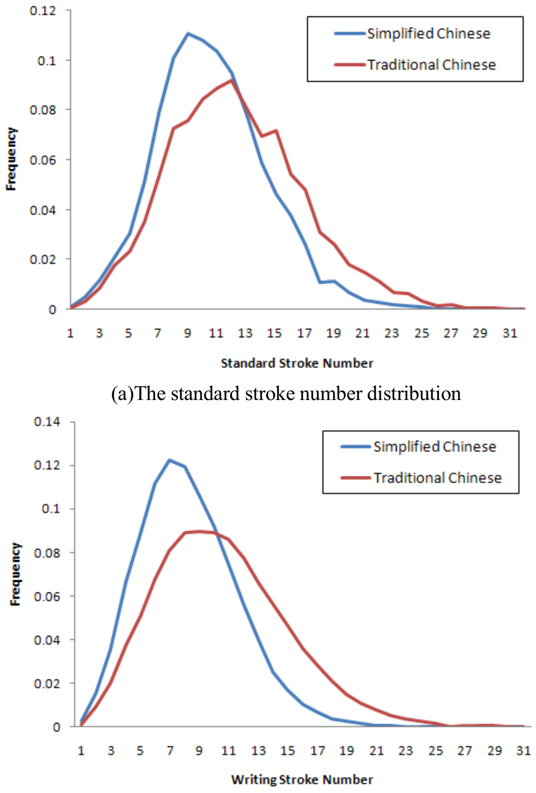
Character set	unchanged part	synonym part	Average rate
GB	95.12%	95.73%	95.34%
BIG5	**97.86%**	**98.01%**	**97.91%**

Besides, we also carry the experiments by using the total 6,763 categories of GB set and the 5,401 categories of BIG5 set. The similar results are observed, as shown in Table IV.

TABLE IV. RECOGNITION RATE OF THE TOTAL DATASETS

Character set	unchanged part	synonym part	disjoint part	Average rate
GB	94.82%	95.19%	96.25%	95.05%
BIG5	**97.59%**	**97.83%**	**96.66%**	**97.61%**

One main difference between simplified and traditional Chinese is that the stroke number is different. Figure 3 shows the statistical distributions of the stroke number of GB and BIG5 datasets, for standard GB and BIG5 Chinese sets and for handwriting subsets form SCUT-COUCH 2009, respectively. Figure 3(a) shows the standard stroke number distribution. It can be seen that the average stroke number of BIG5 is more than GB dataset. We can see that the average stroke number of the handwritten samples in Figure 3(b) is less than that of standard characters in Figure 3(a), due to that many Chinese characters are written cursively, with two or more strokes connected. However, the average stroke number of BIG5 is still more than GB.



(a)The standard stroke number distribution

(b)The writing stroke number distribution

Figure 3. The statistics of the stroke number of GB and BIG5

Figure 4 shows the histogram relationship between recognition accuracy and the average stroke number. It can be seen that the average stroke number for both of BIG5 and GB datasets are increasing with the increasing of recognition accuracy, and the average stroke number of BIG5 is larger than GB when recognition accuracy is high. It is worth to note that the recognition accuracies of some few-stroke characters in GB are quite lower.

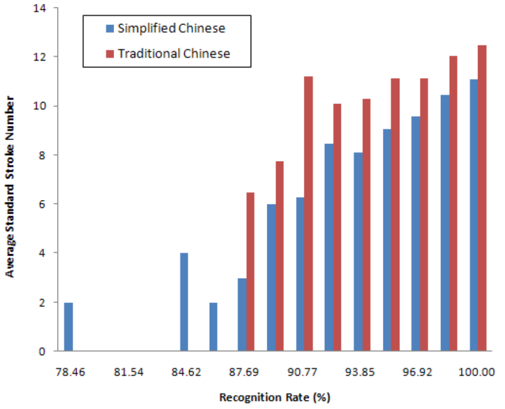


Figure 4. The relationship between recognition accuracy and the average stroke number

It is known that the confusable similar character recognition is the bottleneck to further improve the performance of handwritten Chinese character recognition. In order to analyze the recognition difference of BIG5 and GB, we select some typical characters and compare the recognition accuracies of their corresponding similar characters. The details are shown in Table V and Table VI. Table V shows the recognition rate of 10 characters in the unchanged part. It is obviously that the character accuracies of BIG5 are all higher than GB, although the writing forms are same. Furthermore, for the corresponding similar characters of the 10 given characters in Table V, we can see that the similar characters are much more difficult to be correctly recognized in GB dataset, such as“斤” and “厅”,“方” and “万”, and so on. However, these similar character-pairs in BIG5 are written in traditional forms as “斤” and “廳”,“方” and “萬” and so on, which are much easier to be recognized.

Table VI shows the similar results of the recognition rate of 10 characters in synonym part of both sets. The traditional characters with more stokes not only provide more feature information, but also reduce the similar characters and results in higher recognition accuracy. It is worth to note that the two components “氵” and “讠” in GB dataset are usually written in similar way and hard to distinguish and the recognition accuracies of characters with these components are usually lower than that of the corresponding traditional characters.

V. Conclusions

In this paper, we have made an empirical comparative study of the handwriting simplified and traditional Chinese character recognition. Some interesting finding is observed, which may be useful to bring us some cues on the issue of confusable Chinese character recognition for further study.

The experiment results show that the average handwriting stroke number of traditional Chinese is higher than that of simplified Chinese, and the character with more stroke number usually result in higher recognition accuracy. Furthermore, we compare the recognition accuracy of some characters in simplified Chinese and traditional Chinese. It is obviously that the simplified characters have more similar characters that are difficult to be recognized, which leads to performance decrease.

Simplified Chinese makes Chinese easy to learn, read and write. However, as an ideographic language, simplified Chinese reduces the inner meaning of Chinese characters, and brings many similar characters difficult to recognize. Chinese character is still in developing and it will become more convenient for people to understand and use. It is suggested that we should also pay attention on the traditional Chinese, because traditional Chinese inherits the long history and culture of Chinese nation.

TABLE V. RECOGNITION ACCURACY OF UNCHANGED PART

The given characters			The Similar characters			
characters	Rate in GB	Rate in BIG5	simplified	Rate in GB	traditional	Rate in BIG5
斤	95.38%	100.00%	厅	95.38%	廳	100.00%
方	96.92 %	100.00%	万	96.92%	萬	100.00%
近	96.92 %	98.46%	迈	96.92%	邁	100.00%
枚	96.92 %	100.00%	极	95.38%	極	98.46%
洞	96.92%	100.00%	泪	95.38%	淚	100.00%
湛	98.46%	100.00%	谌	98.46%	諶	100.00%
汕	98.46%	100.00%	汹	98.46%	洶	100.00%
河	96.92%	100.00%	诃	96.92%	訶	98.46%
江	95.38%	100.00%	讧	96.92%	訌	100.00%
渚	98.46%	100.00%	诸	93.82%	諸	96.92%

TABLE VI. RECOGNITION ACCURACY OF SYNONYM PART

The given characters				The Similar characters			
GB	Rate in GB	BIG5	Rate in BIG5	simplified	Rate in GB	traditional	Rate in BIG5
尝	95.38%	嘗	100.00%	党	96.92%	黨	100.00%
经	96.92%	經	98.46%	径	93.85%	徑	100.00%
奋	96.92%	奮	100.00%	备	98.46%	備	100.00%
钢	95.38%	鋼	100.00%	铜	98.46%	銅	96.92%
凉	95.38%	涼	100.00%	谅	96.92%	諒	98.46%
沦	95.38%	淪	100.00%	论	96.92%	論	96.92%
渎	96.92%	瀆	98.46%	读	93.85%	讀	100.00%
沟	95.38%	溝	100.00%	询	96.92%	詢	100.00%
清	96.92%	淸	100.00%	请	98.46%	請	100.00%
泽	95.38%	澤	100.00%	译	96.92%	譯	100.00%

ACKNOWLEDGMENT

This research was partially supported by NSFC (no. 61075021), the National Science and Technology Support Plan (2013BAH65F01-2013BAH65F04), GDNSF (No. S2011020000541) and GDSTP (No.2012A010701001). .

REFERENCES

[1]. DeFrancis, John, " The Chinese Language: Fact and Fantasy". University of Hawaii Press. ISBN 0-8248-1068-6.

[2]. H. Liu and X. Ding, "Handwritten character recognition using gradient feature and quadratic classifier with discriminative feature extraction", ICDAR2005, 2005, pp. 19-23.

[3]. C. Liu, "High accuracy handwritten Chinese character recognition using quadratic slassifiers with discriminative feature extraction", ICPR2006, 2006, pp. 942-945.

[4]. Cheng-Lin Liu, Fei Yin, Da-Han Wang, Qiu-Feng Wang, "CASIA Online and Offline Chinese Handwriting Databases", Proc. 11th ICDAR, Beijing, China, 2011.

[5]. C. Liu, Fei Yin, Da-Han Wang, Qiu-Feng Wang, Online and Offline Handwritten Chinese Character Recognition: Benchmarking on New Databases, Pattern Recognition, 46(1): 155-162, 2013.

[6]. L. Jin, Y. Gao, et al, "SCUT-COUCH2009–A Comprehensive Online Unconstrained Chinese Handwriting Database and Benchmark Evaluation, International Journal on Document Analysis and Recognition", vol. 14, no.1, pp53-64,2011.

[7]. Wikipedia: http://zh.wikipedia.org/wiki/

[8]. C. Liu, S. Jaeger, M. Nakagawa, "Online Recognition of Chinese Characters: The State-of-the-Art", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 2, pp. 198-213, 2004.

[9]. T. Long, L. Jin, "Building compact MQDF classifier for large character set recognition by subspace distribution sharing", Pattern Recognition, Vol. 41, No. 9, pp. 2916-2925, 2008.

[10]. Z. Bai and Q. Huo, "A Study On the Use of 8-Directional Features For Online Handwritten Chinese Character Recognition", ICDAR2005, 2005, pp. 232-236

[11]. R.A. Fisher,"The Use of Multiple Measurements in TaxonomicProblems." Annals of Eugenics 1936, Vol. 7, pp. 179-188.

[12]. F. Kimura, K. Takashina, et al., "Modified quadratic discriminant functions and the application to Chinese character recognition", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 9, no. 1, pp. 149–153, 1987.'

[13]. SCUT-COUCH, Website: http://www.hciilab.net/data/