A new approach for synthesis and recognition of large scale handwritten Chinese words

Gang Liu, Lianwen Jin, Kai Ding, Hanyu Yan*

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, P.R.China
*E-Mail: lianwen.jin@gmail.com

Abstract

Lacking of dataset is still a serious problem for researchers who study on online handwriting word recognition (HWR). In this paper, a handwritten Chinese word synthesis method is proposed for the first time to generate a large scale handwritten Chinese word dataset. The distributions of shape and position characteristics, such as aspect radio, character interval and the angle of gravity center line in each word sample of the Word8888 dataset have been estimated respectively. Based on this, we synthesize as large as 44,208 categories of 8,311,104 unconstrained handwritten Chinese word samples. To verify the validity of the synthesized dataset, a practical rotation free handwriting Chinese word recognition system is presented based on a new holistic approach. Experimental results for randomly rotated word samples demonstrate that the holistic approach can achieve 91.96% recognition accuracy, which provides evidence for the effectiveness of our method.

Keywords: handwriting word recognition, word synthesis, holistic, rotation free

1. Introduction

Online handwriting Chinese recognition is attracting more and more attention among researchers in recent years [1]. And that is one key technology of input method for many popular portable devices, such as Personal Digital Assistant (PDA), smart mobile phone etc. However, most Chinese handwriting input methods now are based on recognition of single character. As the single Chinese character recognition has been solved at a certain extent, it would be advisable that more efforts be paid on the research of Chinese word recognition [2, 3].

Generally, most online handwriting word recognition approaches treat words as a collection of isolate characters which are recognized separately. And it is often called the analytical approach. In our paper, an alternative new method is introduced for rotation free online unconstrained Chinese word recognition through a holistic approach [4]. Inspired by results in cognitive psychology, the holistic approach makes no attempt to segment the word sample, but rely on the word-level features [5]. Compared with the analytical approach, the advantages of the holistic approach are obvious. On one hand, it can avoid the segmentation procedure which is still a challenging problem in the analytic approach. On the other hand, when the handwriting is so poor that individual characters can not be distinguished but the overall shape of the word is preserved, holistic features can also provide information about the whole word [4]. These advantages make the holistic approach more and more attractive recently.

However, unlike the analytical approach, the handwritten word dataset plays an important role on the performance of the holistic approach. But up to now, to our best knowledge, there are only two public online handwriting Chinese word datasets, ***Word8888*** and ***Word44208***, both of which are the subsets of SCUT-COUCH2009 [6], which is a revision of SCUT-COUCH2008 [7]. Moreover, it is widely know that the number of Chinese word phrases is huge. For example, there are near 44,208 words in "The Contemporary Chinese Dictionary" (the fourth edition) (TCCD4th), which is an official Chinese dictionary [8]. So it is obviously understood that building a comprehensive online handwritten Chinese word dataset means great amount of work.

Fortunately, in the meanwhile, it is also found that although Chinese word phrases is in great quantity, the number of characters which words are comprised of is much less. There are only 6,763 simplified characters in GB2312-80 standard, but it covers more than 99.9% daily usage in China. 44,208 Chinese words in TCCD4th which are employed in our experiments are all composed by them. Furthermore, several large scale handwritten Chinese single character databases have been published already [6, 9]. So our basic idea is that rather than collecting handwritten Chinese words completely with costing lots of efforts and time, we can generate the database by synthesizing isolated characters into Chinese words using the available handwritten Chinese character database.

Inspired by this, we proposed a word synthesis approach to generate a large scale Chinese word dataset

978-0-7695-4221-8/10 $26.00 © 2010 IEEE
DOI 10.1109/ICFHR.2010.94

using existed handwritten Chinese character dataset. First of all, elaborate analyses were carried out on some real collected word samples in the Word8888 subset of SCUT-COUCH2009, on the purpose of finding the characteristic of handwritten word samples and the difference between words and single characters. By estimating the distributions of shape and position features, such as aspect ratio, character interval and the angle of gravity center line, we explore the mechanism that how the handwritten Chinese words are synthesized from handwritten Chinese characters. In this way, a new handwritten Chinese word dataset, which consists of 44,208 categories of 8,311,104 handwritten Chinese word samples, is generated on the basis of SCUT-COUCH GB1 and GB2 subsets.

Thereafter, to show the validity of the synthesized handwritten Chinese word samples, we presented a rotation free holistic handwriting Chinese word recognition approach by using the synthesized word samples as training dataset. Experimental results testing on another dataset, Word44208, which contains 44,208 categories of 221,040 real collected handwritten Chinese words, demonstrate that the classifier trained with synthesized word samples works very well.

The rest of this paper is organized as follows: the experimental database and proposed word synthesis approach are presented in Section 2. Section 3 describes the rotation free holistic handwriting Chinese word recognition approach. The experimental results and analysis are given in Section 4. Section 5 concludes the paper.

2. Chinese word Synthesis approach

2.1. Existed datasets used for analysis of word characteristics

In order to generate a large scale handwritten Chinese word database and evaluate the performance of our proposed holistic approach, the GB1 and GB2 character subsets (henceforth referred to as ***CouchGB*** dataset), of SCUT-COUCH 2009 database, which contains 188 writer's 6763 categories of Chinese characters, is used as the material for synthesizing new words. And the ***Word8888*** subset is used to estimate the distribution of real handwritten Chinese words' shape and position features. The ***Word44208*** subset, which is comprised of 5 sets of 44,208 categories real handwritten Chinese word samples, is employed to evaluate the performance of proposed word synthesis approach and the holistic handwriting Chinese word recognition approach. Up to now, the Word8888 subset covers 130 writers' samples of 8,888 categories of most frequently used handwritten words (Total 130×8888 handwriting word samples). A part of handwritten word samples of Word8888 subset is illustrated in Figure 1.

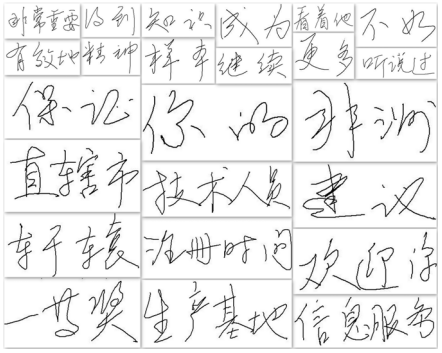

Figure 1. A part of samples in ***Word8888*** subset

2.2. Handwritten word Synthesis algorithm

In this section, we will propose a new handwritten Chinese word synthesis approach using available isolated handwritten Chinese character dataset. From Figure 1 we can see that, the aspect ratio of the handwritten Chinese words, which is computed from their bounding boxes, depends on the numbers of characters the words are comprised of. Even if the numbers of characters in two words are the same, the aspect ratios of them may also be different. So we make a statistics on the aspect ratios of all the words in the Word8888 dataset in order to investigate the distribution of the aspect ratio, and the result is shown in Figure 2. From Figure 2, it is found that, given the number of the handwritten Chinese word's composition characters, the distribution of the word's aspect ratio approximates the Gaussian distribution.

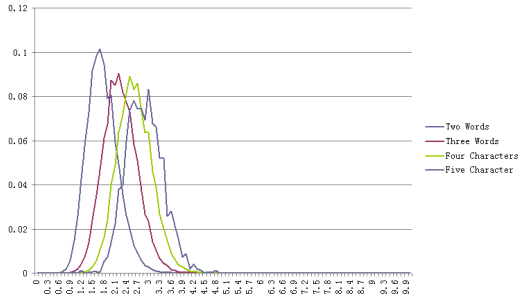

Figure 2. The distribution of real handwritten Chinese word's aspect ratio

On the other hand, because handwritten Chinese word samples are occasionally slant, so another experiment is designed to study the slant characteristic of handwritten Chinese words. First, the word sample is partitioned into left and right parts by the vertical line through the gravity center, and then the gravity centers of each part are also calculated. After that, the inclined angle of θ of the line, which connected the two gravity centers, in relation to horizontal line is counted. To estimate the distribution of θ, an experiment on 1,137,664 unconstrained Chinese word

data is conducted and the results are shown in Table 1. From Table 1, we can see that, the intra-angles of most of words are very small, less than $\pm 5^{\circ}$.

Table 1. Statistical distribution of

Range of θ	Proportion (%)
$0 \leqslant \lvert\theta\rvert < 3^{\circ}$	49.84
$3 \leqslant \lvert\theta\rvert < 5^{\circ}$	21.89
$5 \leqslant \lvert\theta\rvert < 10^{\circ}$	22.47
$\lvert\theta\rvert \geqslant 10^{\circ}$	5.80

The third characteristic of handwritten Chinese words is the intervals between the two adjacent composition characters. Conveniently, the character intervals are considered as a constant value in a given word sample.

After exploring the characteristics of real handwritten Chinese word samples, a handwritten Chinese word synthesis approach by using available isolated handwritten Chinese character samples is proposed as follows. Given a Chinese word, we first select its composition characters, which are from the CouchGB subset. Then, using the Word8888 dataset, the Gaussian distribution parameters, which are average and variance, of the handwritten Chinese word samples' aspect ratio are estimated. The aspect ratio of the given synthesized word is calculated according to:

$$AR = Rand(N(\mu_{charnum}, \delta_{charnum})) \quad (1)$$

We set $\mu_{charnum}$ and $\delta_{charnum}$ as the estimated average and variance of aspect ratio respectively in the case that the number of the composition characters is *charnum*, *AR* and *charnum* be the aspect ratio and the number of the word's composition characters respectively. And $Rand(N(\mu_{charnum}, \delta_{charnum}))$ can generate a random number according to the Gaussian distribution $N(\mu_{charnum}, \delta_{charnum})$.

Next, given the height and the aspect ratio of the synthesized word, the width of the synthesized word can be counted according to:

$$W = \max(H \times AR, \sum_{i=1}^{charnum} W_i) \quad (2)$$

Let *W* be the estimated width of the synthesized word, *H* be the maximum height among the selected composition characters. And *W* means the width of the i^{th} composition character.

As mentioned previously, since the intervals between two adjacent characters are set as a constant value, the interval can be computed as follows:

$$Interval = (W - \sum_{i=1}^{charnum} W_i)/(charnum - 1) \quad (3)$$

Then, in each synthesized word, the horizontal coordinate of each composition character's gravity center can be calculated as follows:

$$GC_{ix} = \begin{cases} GC_{1x} & i = 1 \\ \sum_{j=1}^{i-1} W_j + (i-1) \times Interval + GC_{1x} & 1 < i < charnum \end{cases} \quad (4)$$

Where GC_{ix} is the horizontal coordinates of the i^{th} composition character's gravity center. Specially, the zero point for them is the left down corner of the synthesized word sample.

After that, a random angle θ whose range is from -5 to 5 is used as the synthesized word's intra-angle. Therefore, the vertical coordinate of each composition character's gravity center can be calculated as follows:

$$GC_{iy} = \begin{cases} GC_{1y} & i = 1 \\ GC_{1y} + \tan(\theta) \times (GC_{ix} - GC_{1x}) & 1 < i < charnum \end{cases} \quad (5)$$

At last, the synthesized word can be created by moving each handwritten Chinese character according to their gravity centers' coordinate. Figure 3 shows the comparison between real collected word samples (from the Word44208 subset) and some synthesized word samples. It should be noted that, all of the word samples are written in one stroke to unify the writing style.

(a)

(b)

Figure 3. Word samples from: (a) collected data; (b) synthesized data

3. Holistic approach for HWR

After the word synthesis experiment, in order to prove that synthesized samples are good substitutes, a set of experiments for rotation free online Chinese word recognition are performed through a holistic approach. By utilizing a novel gravity center balancing method [2], the rotation ranging from 0° to 360° of handwritten words can

be detected. Based on the elastic meshing technique, the directional feature of the whole word has been extracted. At last, holistic feature is recognized by the classifier. The overall flowchart of our method is shown in Figure 4.

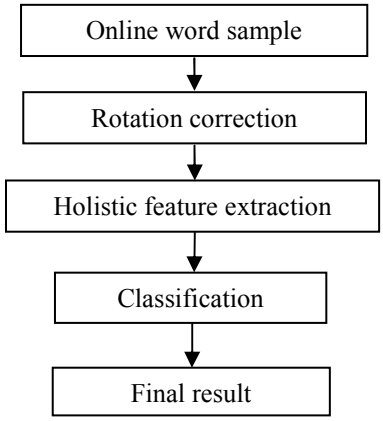

Figure 4. Overall flowchart

First, it is found that many character based recognition methods can achieve a relatively high accuracy on regular handwriting words, but they would get a considerable performance penalty when slant samples are provided. In our paper, to overcome this issue, we use the gravity center balancing method proposed by Teng L, and LW.J [2] for the rotation correction of handwritten Chinese words.

Then, the feature we used for HWR in this paper is the 8-directional feature proposed by ZL.Bai and Q.Huo [10]. But, we replace the nonlinear shape normalization (NSN) [11] with elastic meshing (ELM) technique [3, 12]. ELM is a non-uniform region partition for character images with imaginary grids, whose principle is that after partitioning, adjacent regions should have equal number of character pixels. In our experiments, the ELM technique is employed and a comparison of recognition accuracies of the ELM and NSN are shown in section 4.1.

Third, at the classification stage, both of linear discriminant analysis (LDA) classifier and modified quadratic discriminant function (MQDF) classifier are employed. After the 8-directional feature is extracted, LDA algorithm is implemented to find the linear projections of them and the dimension of the feature is reduced to 256. Then, the minimum Euclidean distance classifier is used to classify. For the MQDF classifier, when the LDA classifier finishes coarse classification, the first ten candidates are fed into the MQDF classifier to output the final recognition result.

4. Experiments and results

To evaluate the performance of both our proposed word synthesis and holistic word recognition approach, the synthesized subset is used for training, and the Word44208 subset is chose for testing.

4.1. A comparison of recognition accuracies between elastic meshing and nonlinear shape normalization

The first experiment is designed to compare the performance of elastic meshing (ELM) technique and the nonlinear shape normalization (NSN) technique using non-rotated handwritten word samples in the Word44208 dataset. The results given in Table 2 shows that the proposed holistic handwritten that the ELM technique significantly outperforms the NSN technique. Therefore, the ELM technique is used in all of the rest experiments of this paper.

Table 2. Performance comparison of ELM with NSN

	LDA classifier	MQDF classifier
NSN	75.68%	85.68%
ELM	**87.22%**	**94.37%**

4.2. A comparison of recognition accuracies on rotated data

To demonstrate the robustness of the synthesized data and the efficiency of rotation correction method, this set of experiments is designed to compare the performances of non-rotation correction method with the rotation correction method. The results are shown in Table 3. (In this table, "With" and "Without" stand for with and without rotation correction respectively). From the results, it is seen that: (1) When the rotation angle exceeds the range from -10° to $+10^{\circ}$, the recognition accuracy decreases dramatically if the rotation correction method is not employed, and this method can achieve an significant performance on rotated word data; (2) Using the rotation correction method, the recognition accuracies are competitive within an acceptable range. This encouraging result indicates that our method is very useful and the problem of lacking dataset can be solved by the proposed word synthesis approach.

4.3. A comparison of recognition accuracies using holistic approach vs. analytic approach

T. LONG, LW. JIN [2] proposed an analytic approach for rotation free online unconstrained cursive handwritten Chinese word recognition. Table 4 illustrates the results about the comparison of recognition accuracies using holistic approach vs. analytic approach, where testing data is rotated randomly from 0° to 360°.

Specifically, the analytic approach proposed by T. LONG, LW. JIN [2] can only handle handwritten Chinese Words whose number of characters is from two to four.

Therefore, 930 categories of handwritten words whose number of character is more than four in each set of the synthesized and the ***Word44208*** dataset are excluded. The first two results in Table 4 are conducted on just 43,278 categories of samples. For comparison, the third

Table 3. A comparison of recognition accuracies of without and with rotation correction

Rotated angle		30 º	20 º	15 º	10 º	5 º	0 º	-5 º	-10 º	-15 º	-20 º	-30 º	random
Without	LDA	0.09	4.47	21.10	53.23	78.64	87.21	80.47	51.70	17.77	3.79	0.16	5.30
	MQDF	0.25	10.55	37.25	71.44	89.78	94.37	90.34	68.43	29.48	7.26	0.28	6.69
with	LDA	84.90	85.06	85.28	85.45	85.59	84.38	85.59	85.31	85.03	84.95	84.73	84.98
	MQDF	91.85	92.07	92.23	92.37	92.49	91.47	92.34	92.21	92.03	91.90	91.72	91.96

recognition accuracy in Table 4 is based on the whole 44,208 categories of handwritten Chinese word samples.

Table 4. A comparison of recognition accuracies using holistic approach vs. analytic approach on randomly rotated word data

Approach	MQDF(%)
Analytic [2]	63.16
Holistic (43,278 categories)	93.50
Holistic (44,278 categories)	91.96

From the results, we can see that when the handwritten samples are very cursive and in large categories, the recognition of the analytic approach is far from good enough. By contrast, the proposed holistic method works well, with the increase of recognition rate by about 30%.

5. Conclusion

In this paper, a handwritten Chinese word synthesis approach is proposed to generate a large scale handwritten Chinese word dataset using available isolated handwritten Chinese character database. In order to explore the mechanism that how the handwritten Chinese words are comprised of handwritten Chinese characters, the distributions of shape and position characteristics, such as aspect radio, character interval and the angle of gravity center line in each word sample of the Word8888 dataset have been estimated respectively. Thereafter, we create as large as 44,208 categories of 8,311,104 unconstrained handwritten Chinese word samples. Then, to verify the validity of the synthesized dataset, a practical rotation free handwriting Chinese word recognition system is presented through a holistic approach. From the experimental results, we can conclude that: (1) the problem of lacking large scale dataset for large classes based holistic handwriting Chinese word recognition can be solved by the proposed word synthesis approach. (2) The holistic approach can provide much higher accuracy than analytic method on such large scale classification task.

Acknowledgments

We would like to thank all anonymous reviewers for their valuable suggestions. This work is supported in part by the research funding of NSFC (no.U0735004, 60772216), GDSTP (no.07118074, 2007B010200048, 2008A050200004, 2009B090300394).

6. References

[1] C.L. Liu, S. Jaeger, M. Nakagawa, "Online recognition of Chinese characters: the state-of-the-art", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Volume 26, Issue 2, pp. 198-213, Feb 2004.

[2] T. Long, L.W. Jin, "A Novel Orientation Free Method for Online Unconstrained Cursive Handwritten Chinese Word Recognition", *Proceedings of the 2008 19th International Conference on Pattern Recognition*, 2008.

[3] K. Ding, L.W. Jin, X. Gao, "A New Method for Rotation Free Method for Online Unconstrained Handwritten Chinese Word Recognition: A Holistic Approach", *proceedings of the 2009 10th International Conference on Document Analysis*, 2009, pp. 1131-1135.

[4] S. Madhvanath, V.Govindaraju, "The Role of Holistic Paradigms in Handwritten Word Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, pp. 149-164, 2001.

[5] J. Ruiz-Pinales, Jaime-Rivas etc., "Holistic cursive word recognition based on perceptual features", *Pattern Recognition Letters*, Vol. 28, No. 13, pp. 1600-1609, 2007.

[6] L.W. Jin, Y. Gao, G. Liu, Y.Y. Li, K. Ding, "SCUT-COUCH2009----A Comprehensive Online Unconstrained Chinese Handwriting Database and Benchmark Evaluation", *International Journal of Document Analysis and Recognition*, 2010. Website:http://www.hcii-lab.net/data/scutcouch/

[7] Y.Y. Li, L.W. Jin, X.H. Zhu, and T. Long, "SCUT-COUCH2008: A Comprehensive Online Unconstrained Chinese Handwriting Dataset", *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, pp. 165-170, 2008.

[8] http://www.cp.com.cn/xh/

[9] D.H. Wang, C.L. Liu, J.L. Yu, X.D. Zhou, "CASIA-OLHWDB1: A Database of Online Handwritten Chinese Characters", *proceedings of the 2009 10th International Conference on Document Analysis*, 2009, pp. 1206-1210.

[10] Z.L. Bai, Q. Huo, "A Study On the Use of 8-Directional Features For Online Handwritten Chinese Character Recognition", *proceedings of the 2005 8th International Conference on Document Analysis*, 2005, pp. 232-236.

[11] C.L. Liu, "Handwritten Chinese Character Recognition: Effects of Shape Normalization and Feature Extraction", *Lecture Notes in Computer Science*, Vol. 4768, 2008, pp. 104-127.

[12] L.W. Jin, G. Wei, "Handwritten Chinese Character Recognition with Directional Decomposition Cellular Features", *Journal of Circuit, System and Computer*, 1998, 8(4), pp. 517-524.