Jurnal Teknologi

Full Paper

OFF-LINE TEXT-INDEPENDENT WRITER RECOGNITION FOR CHINESE HANDWRITING: A REVIEW

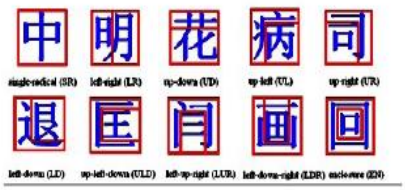
Gloria Jennis Tan, Ghazali Sulong*, Mohd Shafry Mohd Rahim

Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

Article history
Received
3 December 2013
Received in revised form
2 July 2014
Accepted
25 November 2014

*Corresponding author
ghazali@spaceutm.my

Graphical abstract



Abstract

This paper provides a comprehensive review of existing works including the characteristics of Chinese characters' complex stroke crossing and challenges, which is still a largely unexplored subject for off-line text-independent Chinese handwriting identification.

Keywords: Handwriting identification, text-independent writer recognition, Chinese handwriting writer recognition

1.0 INTRODUCTION

Personal identification based on handwriting [1] analysis is an important research area aimed at automatic identity recognition of the writer. It is receiving growing interest from both academia and industry because of the important role it plays in criminal justice [2][3]. There are numerous languages throughout the world. Each language poses a different challenge to the writer recognition problem depending on its characteristics. So, it is very clear that the problem associated with writer identification varies across multiple languages. It is also evident that the importance of writer recognition has become more significant these days. As a result, the number of researchers involved in studying this challenging problem is increasing.

The handwriting of an individual contains the following information; the content part (relating to the content of the handwritten text) and the style part (reflecting the individual writing style of the writer) [4]. The way writers' stroke while writing reveals their personality and the stability of their writing style. This helps experts to distinguish writings of different writers [5]. The main focus of writer recognition is to know more about the writer, rather than knowing what is written. So, the physical manner in which lines and loops are produced is of great importance in writer recognition. It is a fact that different people have different handwriting. Thus, the analysis of handwritten documents with the purpose of determining the writer can contribute greatly to the criminal justice system as well as a wide variety of fields ranging from security, forensics, financial activities to archeology (e.g. to identify ancient document writers).

This paper focuses on off-line text-independent writer recognition using images of Chinese handwritten texts. Text-independent approach is not limited by text content; therefore, it has received increased attention in recent years on extracting writing style features from the global writing text. However, this makes the problem more difficult in many cases (e.g. in criminal investigation) where writers and characters are indeterminable because no registration is available and off-line training is impossible [6]. Text-independent off-line Chinese handwriting identification is a rather challenging task [7][8] because many valuable writing features such as shape features and dynastic writing information are still not available. Writing features can only be extracted from the handwriting image; and that means that a lot of valuable writing information is lost. Before discussing writer recognition of Chinese handwriting, a summary of the properties of Chinese character structures which determines the requirements for writer recognition on off-line text-

independent Chinese Handwritten Script is provided as follows.

The stroke shapes and structures of Chinese characters whose handwriting characteristics are embedded are quite different from those of other languages [9]. This makes it more difficult to identify Chinese handwriting. Chinese characters are highly complex, and there are a high number of potential feature sets [10]. According to [11], in Chinese characters, the structures of Chinese characters are made complicated by the multi strokes of each character [12].

Stroke (笔画)	Names	Way of Writing	Example
一	横 (héng) Horizonal Stroke	Horizontally from left to right	二 two
丨	竖 (shù) Vertical Stroke	Vertically from top to bottom	十 ten
丿	撇 (piě) Left-falling Stroke	Softly from top to lower left	午 noon
乀	捺(nà) Right-falling Stroke	Softly from top to lower right	人 people
丶	点(diǎn) Dot Stroke	Dot to lower right and pause	太 very
㇀	提 (tí) Rising Stroke	Dot from bottom left to top right	习 practice

Figure 1 Basic Chinese writing strokes

Dependent Strokes		Names	Example
Turning Strokes	𠃌	横折 (héng zhé) Horizontal Turning Stroke	五 five
	𠃍	竖折 (shù zhé) Vertical Turning Stroke	山 mountain
	𠃋	撇折 (piě zhé) Left-falling Turning Stroke	云 cloud
Hook Strokes	亅	竖钩 (shù gōu) Vertical Hook Stroke	小 small
	𠄌	横钩 (héng gōu) Horizontal Hook Stroke	买 buy
	𠃎	斜钩 (xié gōu) Slanting Hook Stroke	划 scratch, paddle
	𠃊	弯钩 (wān gōu) Curved Hook Stroke	家 home
	𠃏	卧钩 (wò gōu) Lying Hook Stroke	心 heart

Figure 2 Turning strokes and hook stroke

2.0 CHINESE CHARACTER STRUCTURES

2.1 Strokes and Radical of Chinese Character

Chinese characters, which are in the shape of a square, are made up of strokes. These strokes fall into eight main categories: horizontal (一), vertical (丨), left-falling (丿), right-falling (丶), rising dot (、), hook (亅), and turning (𠂇, 乚, 乙, etc.), as shown in Figure 1 and Figure 2. Chinese characters are made up of two or more components (radicals). They have a total of 12 kinds of common layout structures according to the composing patterns of their radicals, as shown in Figure 3 and Figure 4. The most complicated Chinese character has 36 strokes [13][14], as shown in Figure 5. It is a good example to show that the complexity of structures makes structural description difficult.

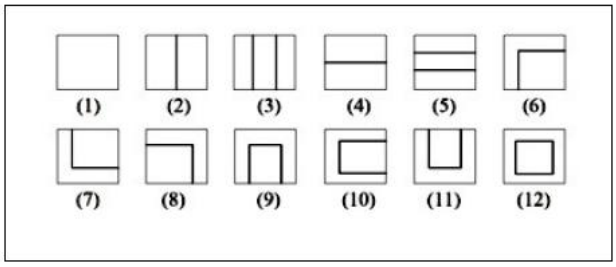


Figure 3 The common layout structures of Chinese characters: (1) single/radical; (2) left-right; (3) left-middle-right; (4) up-down; (5) up-middle-down; (6) up-left; (7) left-down; (8) up-right; (9) left-up-right; (10); up-left-down; (11) left-down-right; (12) enclosure

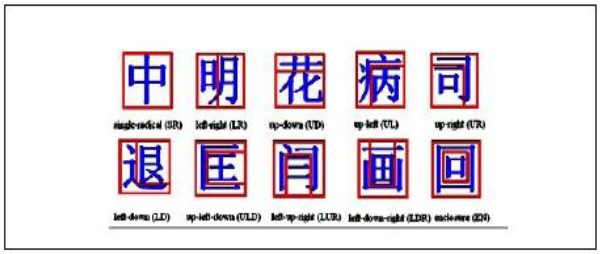


Figure 4 Examples of Chinese character structures

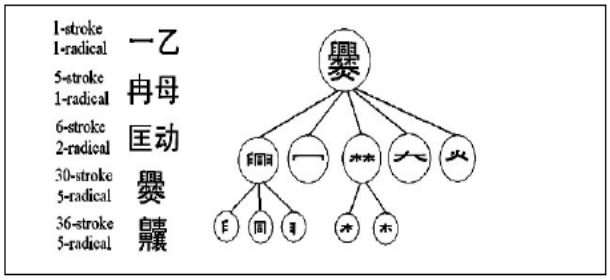


Figure 5 Examples of Chinese character structures. The right panel shows a complicated Chinese character with 5 radicals and 36 strokes, which can be further decomposed

In recent years, Western writer recognition has become an active research area. Western handwriting identification technology has been experimented on large handwriting databases and has shown practical effectiveness. But many identification methods proposed for Western handwriting are not suitable for Eastern handwriting, such as Chinese and Japanese [9], [15], [4], [4].

Eastern writer recognition technology and methods still face many challenges. The basic reason is that Western handwriting (like English) is alphabetic writing while eastern writing (like Chinese) is ideographic writing. English letters are simple and have plenty of curve connections between simple letters while Eastern handwriting is composed of separate characters with complex structures consisting strokes and radicals [16]. Western writing has lots of ink connection between individual letters. But Eastern writing, regardless of Chinese, Japanese or Korean, is composed of characters in separate block structures. Furthermore, words are written from left to right for Western handwriting while Chinese handwriting is written from top downwards [17]. But both sentences are written the same way from left to right, as shown in Figure 6.

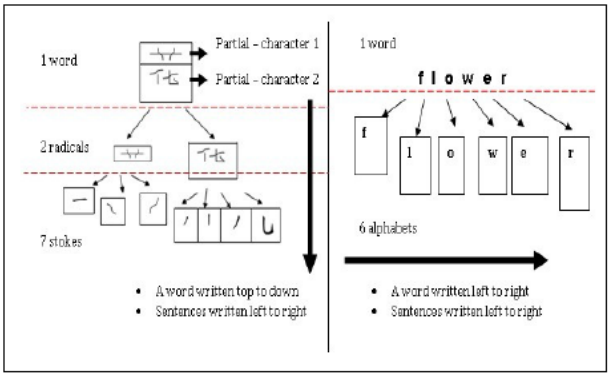


Figure 6 The construction principle of a Chinese character. The word depicted here means "flower"

Figure 7 Examples of the three major writing styles (from left to right: regular script, fluent script, cursive script)

Figure 8 Examples: the sentence " 我去北京" (wǒ qù běi jīng), which means "I am going to Beijing" presented in different writing styles by different writers

2.2 Chinese Handwriting Style

The writing styles of different people can be roughly divided into three categories: regular script (also called handprint), fluent script, and cursive script [13]. The intermediate style between regular and fluent is called fluent-regular, and the style between fluent and cursive is called fluent-cursive. Some examples of the three typical styles are shown in Figure 7. We can see that the strokes of regular script are mostly constructed in straight-line segments. The fluent script has many curved strokes and, frequently, successive strokes are connected. In cursive script, some character shapes differ drastically from the standard shape, as shown in Figure 8.

In off-line Chinese handwriting identification, many issues are still unresolved. Chinese characters have complex stroke crossing. Individual characters are made up of many different strokes, which are of different shapes, for example, dot, horizontal line, vertical line, downward left, downward right, left hook, and right hook. These strokes are used in different combinations to construct the different characters. Because the number of characters (classes) is very large, many characters have very complex structures that consist of strokes and radicals. The extraction of strokes and radicals is not easy, especially when they are connected within the characters. The joining of strokes in cursive style makes the precise capture of features and sub-structures (radicals or strokes) for recognition difficult, or even impossible. The method to decide the threshold of cursiveness for categorizing handwritten Chinese characters is still unregulated.

2.3 Chinese Database

Standard datasets play crucial roles in handwriting identification research [18][19][20][21]. With the large number [22][23][18] of training and testing data provided, they result in high model fit and reliable confidence in statistics[18]. They are also meant by

which evaluation among different recognition algorithms can be performed. More and more handwriting researchers are beginning to pay much attention to the dataset standardization and evaluate their work using standard datasets [24].

In this section, the main databases used for writer recognition of Chinese handwritten scripts are addressed. The Chinese character databases which vary in writing style, cursiveness and others are summarized in Table 1.

HCL2000 (Handwritten Character Library 2000): This database [25] facilitates handwritten Chinese recognition research. The word "2000" in HCL2000 indicates that the database contains 2000 samples collected in the year of 2000. The collection of HCL2000 was supported by the Chinese Government through China 863 high-tech project. HCL2000 contains 3,755 frequently used Chinese characters written by 1000 different writers. The writers' information, which was collected by Beijing University of Posts and Telecommunications for China 863 projects, was incorporated into the database to facilitate the testing on grouping writers with different backgrounds. HCL2000 has two characteristics: one is its large number (the total image sample number is 3,755,000; which is 3,755 frequently used simplified Chinese characters written by 1,000 different subjects); the other is that it contains information of the writers, which can help researchers to study the styles of calligraphy of different writers and writer identification. The writers' information includes the age, occupation, gender, education, address and others.

CASIA Off-line Chinese Handwriting databases [26] [27][28]: These were built by the Institute of Automation of Chinese Academy of Sciences (CASIA). Handwritten samples of isolated characters and handwritten texts (continuous scripts) were produced by 1,020 writers using Anoto pen on paper. A portion of on-line handwritten characters, in the dataset called CASIA-OLHWDB1 (now called as CASIA-OLHWDB1.0), was released at ICDAR 2009. The databases include six datasets of on-line data and six datasets of off-line data, in each case, three for isolated characters (DB1.0–1.2) and three for handwritten texts (DB2.0–2.2). All the data has been segmented and annotated at character level, and each dataset is partitioned into standard training and test subsets. The handwritten pages were scanned (in resolution of 300dpi) to obtain color images, including 3,866 Chinese characters and 171 alphanumeric and symbols. The databases are large in the sense that the number of writers is over 1,000 and the on-line and off-line isolated character datasets contain about 3.9 million samples of 7,356 classes, and the number of character samples in the on-line/off-line handwritten text datasets is about 1.35 million. All the data have been segmented and labeled at character level and partitioned into standard training and test subsets. The databases can be used for research tasks of handwritten document segmentation, character recognition, text line recognition, document retrieval, writer adaptation, and writer identification.

HIT-MW database [29][30] (HIT is the abbreviation of Harbin Institute of Technology, and MW means it is written by Multiple Writers): More than 780 participants (involving 240 writers) contributed their natural handwriting to this database. 853 legible Chinese handwriting samples were collected. There are 186,444 characters in total including letters and punctuation marks besides Chinese characters. These characters lead to 8,664 text lines. The database has at least three distinctions. First, the handwriting is naturally written with no rulers that can be used to make the text line straight by and large. This feature makes it suitable for conducting experiments on Chinese text line segmentation. Second, the underlying texts for hand-copying are sampled from People's Daily corpus in a systematic way and the writers are carefully chosen to give a balanced distribution. Third, the handwriting is collected by mail or middleman, not face-to-face, resulting in some real handwriting phenomena, such as miswriting and erasing. Besides text line segmentation, the HIT-MW is fit to research segmentation-free recognition algorithms, to verify the effect of statistical language model (SLM) in real handwriting situation, and to study the nerve mechanism of Chinese hand copying activity.

The ETL Character Database: This database was supplied by the Electrotechnical Laboratory (ETL) in Japan, and consequently by its reorganized successor the National Institute of Advanced Industrial Science and Technology (AIST). ETL character databases consist of ETL1-ETL9 gray-valued image data which contain about 1.2 million hand-written and machine-printed character images that include Japanese, Chinese, Latin and numeric characters for character recognition researches. ETL8 and ETL9 are binarized image data. ETL9 consists of 2,965 Chinese and 71 Hiragana, 200 samples per class written by 4000 people. ETL8 consists of 881 classes of handwritten Chinese and 75 classes of Hiragana, 160 samples per class written by 1,600 people. ETL8B and ETL9B are open to the public.

3.0 ANALYSIS OF EXISTING WORK

There are three main parts in the process of handwriting identification: pre-processing, feature extraction and classification (or matching). In literature of handwriting identification, the feature extraction and matching are the two major topics covered.

In the eastern text-independent writer recognition approach, any text can be used to establish identification. This method does not require the matching of the same characters but extract the writing style features from the global writing text. The global approach is based on texture analysis, where a writer's handwriting is regarded as a texture. Therefore, in the case of text-independent approach, different

handwriting images are considered as different textures. The goal of this approach is to determine the writer of a text among a number of known writers using images of their handwritings. This is popular in practical applications of writer identification. The methods based on texture analysis are used. In the field of image processing and identification, texture analysis is commonly used and accepted method.

In reference to [31] the multi-channel Gabor filter technique was introduced to analyze Chinese handwriting and achieved the highest identification accuracy of 95.7% using handwritings from 17 different people. Though Gabor filter is an effective method in handwriting-based writer identification, it has some disadvantages. One of the more serious disadvantages is the intensive computational cost because the 2-D Gabor filter has to convolute the whole image for each orientation and each frequency. Besides, the Gabor filter method does not consider the relation among the Gabor coefficients in each sub-band and only use the mean and standard derivation to represent a whole sub-band.

[32] outlined that the Independent Component Analysis method is adopted and achieved the highest identification accuracy of 91.3% with the database of 30 people. Independent component analysis (ICA) was developed as a de-correlation technique for high-order moment of input signal. In recent times, feature extraction and pattern recognition have also been prominent applications of ICA. ICA, based on the higher-order statistical correlation between data, extracted internal features of the image and made full use of the statistical characteristic of the input data. One of the advantages of the ICA is that it has greatly reduced the feature extraction time. However, ICA projections, whose generalization ability is generally worse than those derived by random projections, run the risk of producing misleading features; thus making further selection of the ICA features meaningless.

In reference [33], the Gabor filter is applied to feature extraction, after which multi-class Support Vector machine (SVM) is used to train and test the data from 87 persons. This method achieves the highest accuracy of 97.7%. SVM, which is known as suitable for small samples classification and high generalization performance, is adopted as the classifier. However, the identification process is rather troublesome. In the real application of handwriting identification, at least hundreds of samples in the database or even more are required. Previous texture methods have performed well even when the samples in the database are rather small. When the sample of handwriting in the database is increase, the identification accuracy of SVM is declines sharply.

In reference to [34], run-length measurements are also used for handwriting analysis and the highest accuracy of 91.3% was achieved from the database of 19 people. However, this method has only been tested in a database of small samples and it has not been used extensively. A grey level run is a set of consecutive, collinear picture points having the same grey level value.

In reference to [35], the combination of Gabor Filter and auto-correlation function has shown better performance and higher stability than either one's single algorithm. The texture images are divided into 4 sub-images from input handwriting. This division improves the identification accuracy greatly. Then the Gabor filter is applied on each. Mean and standard deviation are then used as extracted features for writer identification. An advantage of these filters is that they satisfy the minimum space-bandwidth product per uncertainty principle and provide simultaneous optimal resolution in both the space and spatial-frequency domains. Weighted Euclidean Distance (WED) classifier is used to match the extracted features. The traditional autocorrelation function method is modified to decrease the computational cost while still maintains high identification accuracy. The experiments were carried out on 50 people's handwritings and the result was achieved. In future experiments, a larger number of handwriting is needed.

To solve [35] problem, the Fast Fourier Transform (FFT) referred to by [36] is used to extract textural features. The original image is proposed to form a binary image with black foreground and white background after image denoising. Separated "character images" (image with single character extracted from the binary image) are obtained for writer identification. Then the computer chooses parts of the character images to form texture images, after which FFT is applied for feature extraction. All the required feature vectors are fused to get their mathematical expectations. The Weighted Euclidean Distance (WED) classifier is then implemented to identify the writer. The feasibility of this method was tested on 200 samples of handwriting of 100 different people. Each person presents two samples; one for establishing the database and the other for testing. These samples of handwriting had only one requirement, that is, they should contain at least 200 characters. The content and writing tools used were not important. This experiment started out with a database of 100 persons and then increased to 500 persons respectively. It achieves the highest accuracy of 98% concerning Top 10 and 95% concerning Top 25. The result shows that though this method can obtain positive results, its accuracy is not as high as expected. This is mainly because the fluctuation of textural features leads to random differences between handwriting for tests and handwriting in the database. The three main differences are:

- The irregularities in handwriting characteristics: One's handwriting can vary depending on the writing tool used, the writing environment and the state of the writer at the time of writing.

- The arbitrary nature of the written content: The contents of handwriting samples vary at a large scale in real application. Since the size of the texture images is limited, a single

texture image cannot contain all the characters. Different characters will definitely lead to textural feature differences.

- The random positions of the written characters: According to the method of constructing texture images proposed above, the characters in a given texture image may be positioned randomly, thus influencing the stability of the textural features.

According to [37], Wavelet-based GGD features to replace the traditional 2-D Gabor filters. Wavelet transform is a tool that cuts up data or functions or operators into different frequency components, and then studies each component with a resolution matched to its scale. Compared with the Gabor filter, 2-D wavelet can decompose the image into sub-bands with different frequencies and orientations. The method is to overcome the intensive computational cost, because the 2-D Gabor filter has to convolute the whole image for each orientation and each frequency. The GGD method is firstly decomposes the handwriting image via wavelet transform at three levels, and then apply the GGD model on the wavelet decomposition sub-bands except the HH sub-band at the finest scale. Twenty Chinese handwriting samples written by 10 persons were used in the experiments, with each one of them contributing a sample of training handwriting and another sample of testing handwriting. Based on the experiments, the 2-D Gabor filter is a good method for handwriting-based writer identification but the new method achieves better results and reduces the elapsed time greatly.

In [38] a new method using Hidden Markov tree model (HMT) in wavelet domain for off-line, text-independent handwriting identification was presented. Experiments on 1000 Chinese handwritings and 4000 sub-handwritings provided by 500 persons indicate that the new method was satisfactory and outperformed the 2-D Gabor model, the representative of the existing methods for off-line, text-independent writer identification, on both identification accuracy and computational efficiency. Firstly, the handwritings are decomposed into a series of wavelet sub-bands at different resolutions via wavelet transform, and only the wavelet coefficients within sub-bands of interest are considered. With the 2-D Gabor model, the entire handwriting is convoluted with 2-D Gabor filters for each frequency and each orientation and the convolution must be redone when either frequency or orientation is changed. This greatly decreases the computational efficiency. In addition, mean and standard derivation, two statistical parameters used in the 2-D Gabor model, certainly cannot describe well the statistical properties of Gabor coefficients within each Gabor sub-band. On the contrary, the HMT model provides an accurate description for the statistical distribution of wavelet coefficients. The accurate model naturally brings about a better identification result. The disadvantage of the Gabor method is that it does not consider the relationship between the Gabor coefficients in each sub-band and only uses the mean and standard derivation to represent a whole sub-band. Mean and standard derivations are not accurate statistical descriptions of one sub-band. To capture the relationship among wavelet coefficients well, the hidden Markov tree (HMT) is an ideal model.

[39] proposed an efficient method on texture feature for text-independent writer recognition. The method decomposes the image into sub-bands with different frequencies and orientations by using modified 2-D Gabor filter to extract texture features. It transforms an image into four sub-images by 2-D separable Gabor transform. The modified 2-D Gabor filter gives a total of 48 output images. The feature vector is the mean and variance of each output image. A total of 96 features are extracted from a given image. This modified method can overcome high computational cost in traditional 2-D Gabor filter by extracting features from specified sub-bands but not from the whole handwriting image. This experiment was conducted on a self-compiled Chinese handwriting database. The database contains 609 address images, which are in three pages written by each of 203 writers. Each image contains about 20 Chinese characters. The Weighted Chi-square classifier is implemented to identify the writer. The accuracy obtained from this modified method was 90.3% and from the traditional method was 76.1%. The limitation of this method is that it was not tested on different standard handwriting databases and needs real-world implementation.

Another method for off-line Chinese handwriting identification based on stroke shapes and structures is proposed by [40]. To extract the features embedded in Chinese handwriting characters, two special structures were explored according to the trait of Chinese handwriting characters. These two structures are the bounding rectangle and the TBLR quadrilateral. Sixteen features were extracted from the two structures, which were used to compute the unadjusted similarity, and the other four commonly used features were also computed to adjust the similarities. The final identification was performed on the similarities. Experimental results on the SYSU generated and collected from 950 Chinese characters and HanjaDB1 databases (the 800 most frequently used characters classes in names of Korean, which covers 96.6% of usage, were collected) validated the effectiveness of this proposed method. It obtained both the lowest FAR and FRR on the SYSU database.

The latest research, [41] proposed a different approach for off-line, text-independent Chinese writer identification based on the handwriting style. The scheme based on the edge structure coding (ESC) distribution feature and non-parametric discrimination of the sample has been experimented on HIT-MW database (240 text documents provided by 240 writers, and each text documents consists of at least 200 characters). ESC distribution feature can

efficiently model the writing style of each writer and overcome the limitation of statistical approaches using a single scale. This approach first requires the extraction of the fragmented edge structure coding distribution feature of each writer, and each writer is represented as a code-based structural probability distribution. And then, to improve the performance of writer identification, a simple feature selection method is used to reduce the dimensions of features. Finally, the non-parametric discrimination of sample and prototype distributions based on a chi-square statistical measure of the dissimilarity of histograms is employed. This method is free of any possible erroneous and assumptions about feature distributions. Future work will focus on testing on different handwriting databases, computation complex analysis, features combination and real-world implementation. A summary of existing methods up till 2012 is given in Table 2.

4.0 CHALLENGES

The technology for Western handwriting identification has been experimented on a large handwriting database and is proven to be effective. However, the same technology is not suitable for Eastern handwriting (Chinese and Japanese) writer recognition. At this stage, the technology and methods for Eastern handwriting writer recognition are still lacking.

The main challenges in Chinese handwriting writer recognition are as follows:

- Western handwriting (like English) is alphabetic writing while Eastern handwriting (like Chinese) is ideographic writing. English letters are simple while Chinese characters have complex stroke crossing. Western handwriting includes lots of ink connection (linking) between individual letters, but Eastern handwriting, regardless of Chinese, Japanese or Korean, is composed of characters in separate block structures.

- The strokes, shapes and structures of Chinese handwriting are highly complex. There is a big number of potential feature sets. The extraction of strokes and radicals is complicated and difficult, especially when radicals and strokes are connected within characters.

- The Chinese writing styles of different writers can be divided into three categories: regular script (also called handprint), fluent script and cursive script. The strokes of the regular script are mostly in straight lines. The fluent script consists of many curved strokes and, frequently, successive strokes are connected. In the cursive script, character shapes differ drastically from the standard shape. There are more strokes in cursive style, making the extraction of cursive stroke a complicated and difficult task.

- Writer recognition distinguishes writers based on the shape or individual style of writing while ignoring the meaning of the word or character written. The shape and style of writing are different from one person to another. Even for one person, they are different at times. However, everyone has his/her own style of writing and it is individualistic. There is a unique feature that can be generalized as significant individual features through the handwriting shape.

- Chinese handwriting is complex in structure and the relation of character, shape and the style of writing is also different one from another. Chinese characters have unique structures compared to Western characters and this uniqueness poses technical challenges to writer recognition. Additionally, the stroke shape and structure of Chinese characters are quite different from those of other languages [11] thus making it more difficult to identify Chinese handwriting writers.

- Acquiring the features that reflect the author of various styles of handwriting is difficult. In the feature selection phase, threshold selection can affect the feature dimensions and ultimately the identification results[41]. Among these features there exist significant individual features which are directly unique to the individual. High frequency patterns better reflect a writer's writing style because a writer does not write the same strokes in the same way every time.

- Handwriting features are generated from a size-adjustable sliding window to overcome statistical features obtained using a single scale [41]. Previous research used various window sizes to obtain the best identification results. The selection of window size will directly affect the identification performance. We understand that the selection of window is important to acquire the features that reflect the author of handwriting, whether the extracted features are optimal or near-optimal to identify the author. The features may not be independent of each other or even redundant. Besides, there may be features that do not provide any useful information for the task of writer identification. Extracted features may include many garbage features. Such features are not only useless in classification, but sometimes degrade the performance of a

classifier designed on a basis of a finite number of training samples.

- Previous researches have not been tested on different handwriting databases. To solve this, new methods needs to be tested with existing standard dataset and databases, computation complex analysis, features combination and real world implementation so that their results can be comparable and will not be ambiguous.

5.0 CONCLUSION

A brief overview of the work including its strengths and weaknesses, a summary and comparison of the various on-going methods and main challenges faced in this area has been presented in this paper. Chinese handwriting identification is complicated because there are many different characters in the Chinese language and writing styles, different approaches utilize different varieties of features, and yield different accuracies and goals. The journey to fully develop a writer recognition system that is practical for widespread use is still need further research to finding the best and most appropriate writing feature sets based on the characteristics features of each language to represent handwriting image, and the best practice off-line text-independent Chinese writer recognition are real challenging issues.

Acknowledgement

This research was funded by the Universiti Teknologi Malaysia through Flagship-COE and manage by Research Management Centre under Vot No. Q.J130000.2428.02G28.

References

[1] Zhu Y., T. Tan, and Y. Wang. 2000. Biometric Personal Identification Based on Handwriting. In Proceedings of the 15th International Conference on Pattern Recognition (ICPR 2000). Barcelona. September 3-7, 2000. 2: 797–800. IEEE.

[2] Kore S. and S. Apte. 2012. The Current State of Art: Handwriting a Behavioral Biometric for Person Identification and Verification. In Proceedings of the International Conference on Advances In Computing, Communications and Informatics (ICACCI 2012). Chennai, India. August 3-5, 2012. 925–930.

[3] Saranya K. and M. S. Vijaya. 2013. An interactive Tool for Writer Identification based on Offline Text Dependent Approach. In International Journal of Advanced research in Artificial Intelligence. 2(1): 33–40.

[4] Li X., X. Ding, and X. Wang. 2008. Semi-text-independent Writer Verification of Chinese Handwriting. In Proceedings of the 11th International Conference on Frontiers of Handwriting Recognition (ICFHR 2008). 100–105.

[5] Bensefia A., T. Paquet, and L. Heutte. 2004. Handwriting Analysis for Writer Verification. In Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR 2004). October 26-29, 2004. 196–201.

[6] Liu C.L., R. W. Dai, and Y. J. Liu. 1995. Extracting Individual Features from Moments for Chinese Writer Identification. In Proceedings of the 3rd International Conference on Document Analysis and Recognition. Montreal, Que. August 14-16, 1995. 1: 438–441.

[7] Tan J., J. H. Lai, C. D. Wang, and M. S. Feng. 2010. Off-Line Chinese Handwriting Identification Based on Stroke Shape and Structure. In 2010 2nd International Conference on Information Engineering and Computer Science (ICIECS). Wuhan. December 25-26, 2010. 1–4. IEEE.

[8] Tan J., J. H. Lai, W. S. Zheng, and M. Zhong. 2013. Chinese Handwritten Writer Identification based on Structure Features and Extreme Learning Machine. In M. I. Malik, M. Liwicki, L. Alewijnse, M. Blumstein, C. Berger, R. Stoel & B. Found (Eds.), AFHA. 21-25.

[9] Tan J., J. H. Lai, C. D. Wang, and M. S. Feng. 2011. A Stroke Shape and Structure Based Approach for Off-line Chinese Handwriting Identification. International Journal of Intelligent Systems and Applications (IJISA). 3(2): 1–8.

[10] Srihari S. N., X. Yang, and G. R. Ball. 2007. Offline Chinese Handwriting Recognition: An Assessment of Current Technology. In Front. Comput. Sci. China. 1(2): 137–155.

[11] Cheng F. H. 1998. Multi-Stroke Relaxation Matching Method for Handwritten Chinese Character Recognition. In Pattern Recognition. 31(4): 401-410.

[12] Leng W. Y. and S. M. Shamsuddin. 2010. Writer Identification for Chinese Handwriting. In Int. J. Adv. Soft Comput. Appl. 2(2): 143–173.

[13] Dai R., C. Liu, and B. Xiao. 2007. Chinese Character Recognition: History, Status and Prospects. In Front. Comput. Sci. China. 1(2): 126–136.

[14] Tang, Y. Y., L. T. Tu, J. Liu, S. W. Lee, and W. W. Lin. 1998. Offline Recognition of Chinese Handwriting by Multifeature and Multilevel Classification. IEEE Transactions on Pattern Analysis and Machine Intelligence. 20(5): 556–561.

[15] Bulacu M. and L. Schomaker. 2007. Text-independent Writer Identification and Verification using Textural and Allographic Features. IEEE Transactions on Pattern Analysis and Machine Intelligence. 29(4): 701–717.

[16] Li X. and X. Q. Ding. 2009. Writer Identification of Chinese Handwriting using Grid Microstructure Feature. In Proceedings of the 3rd International Conference on Advances in. Biometrics (ICB 2009). 1230–1239.

[17] S. N. Srihari, X. Yang, and G. R. Ball. 2007. Offline Chinese Handwriting Recognition : A Survey. Frontiers of. Computer. Science in China. 1–32.

[18] T. Su, T. Zhang, and D. Guan. 2007. Corpus-based HIT-MW Database for Offline Recognition of General-purpose Chinese Handwritten Text. International Journal of Document Analysis and Recognition. 10(1): 27–38.

[19] Dimauro G., S. Impedovo, Modugno, and Pirlo G. 2002. A New Database for Research on Bank-check Processing. In Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition. 524–528.

[20] Marti U. V. and H. Bunke. 1999. A Full English Sentence Database for Off-line Handwriting Recognition. In Proceedings of the 5th International Conference Document Analysis and Recognition (ICDAR 1999). Bangalore. September 20-22, 1999. 705 – 708.

[21] Mezghani A., S. Kanoun, M. Khemakhem, and H. El Abed. 2012. A Database for Arabic Handwritten Text Image Recognition and Writer Identification. In Proceedings of the 2012 International Conference on Frontiers on Handwriting Recognition (ICFHR 2012). Bari. September 18-20, 2012. 399–402. IEEE.

[22] Van der Maaten L. 2009. A New Benchmark Dataset for Handwritten Character Recognition. Tilburg University. 2-5.

[23] Zimmermann M. and H. Bunke. 2002. Automatic Segmentation of the IAM Off-line Database for Handwritten English Text. In Proceedings of the 16th International

Conference on Pattern Recognition. Quebec, Canada. August 11-15, 2002. 4: 35-39. IEEE.
[24] Marti U.V. and H. Bunke. 2002. The IAM-database: An English Sentence Database for Offline Handwriting Recognition. International Journal on Document Analysis and Recognition (IJDAR). 5(1): 39–46. Springer-Verlag.
[25] Zhang H., J. Guo, G. Chen, and C. Li. 2009. HCL2000-A large-Scale Handwritten Chinese Character Database for Handwritten Character Recognition. In Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR 2009). Barcelona. July 26-29, 2009. 286–290. IEEE.
[26] Liu C. L., F. Yin, Wang D. H. and Wang Q. F. 2011. CASIA Online and Offline Chinese Handwriting Databases. International Conference on Document Analysis and Recognition (ICDAR 2011). Beijing. September 18-21, 2011. 37–41. IEEE.
[27] Liu C. L., F. Yin, D. H. Wang and Q. F. Wang. 2013. Online and Offline Handwritten Chinese Character Recognition: Benchmarking on New Databases. In Journal of Pattern Recognition. 46(1): 155–162.
[28] Liu C. L., F. Yin, D. H. Wang and Q. F. Wang. 2011. ICDAR 2011 Chinese Handwriting Recognition Competition. In International Conference on Document Analysis and Recognition. Beijing. September 18-21, 2011. 1464–1469.
[29] Su T., T. Zhang and D. Guan. 2006. HIT-MW Dataset for Offline Chinese Handwritten Text Recognition. In Proceedings of the 10th International Workshop on Frontiers in Handwriting Recognition. La Baule, France. October 23-26, 2006.
[30] Su T. 2013. Chinese Handwriting Recognition: An Algorithmic Perspective. London: Springer Science & Business Media.
[31] Zhu Y., Tan T. N and Y. H. Wang. 2001. Writer Identification Based on Texture Analysis. ACTA Automatica Sinica. 17(2): 229–234.
[32] Huang Y., E. Chen and S. Luo. 2003. Writer Recognition Based on Independent Component Analysis. Journal of Chinese Information Processing. 17(4): 52–58. (In Chinese).
[33] Hong L., G. Cui, J. Li and S. Tang. 2003. Writer Identification using Support Vector Machines and Texture Feature. Journal of Computer-Aided Design & Computer Graphics. 15(12): 1479-1484. (In Chinese).
[34] Zi-hua Y., W. Min and L. Jiang-rong. 2004. Handwriting Identification System Based on Texture Analysis. Journal of Hunan Institute of Engineering. 14(2): 67-69. (In Chinese).
[35] He Z. Y. and Y. Y. Tang. 2004. Chinese Handwriting-based Writer Identification by Texture Analysis. In Proceedings of the International Conference on Machine Learning and Cybernetics. Shanghai, China. August 26-29, 2004. 6: 3488–3491.
[36] Chen Q., Y. Yan, W. Deng, and F. Yuan. 2008. Handwriting Identification Based on Constructing Texture. In 1st International Conference on Intelligent Networks and Intelligent Systems (ICINIS 2008). Wuhan, China. November 1-3, 2008. 523–526.
[37] He Z., B. Fang, J. Du, and Y. Y. Tang. 2005. A Novel Method for Offline Handwriting-based Writer Identification. In Proceedings of 8th International Conference on Document Analysis and Recognition. Seoul, Korea. August 29-September 1, 2005. 1: 242–246.
[38] He Z., X. You, and Y. Y. Tang. 2008. Writer Identification of Chinese Handwriting Documents using Hidden Markov Tree Model. In Journal of Pattern Recognition. 41(4): 1295–1307.
[39] Wang Y., D. Zhang and W. Luo. 2009. Writer Identification Based on the Distribution of Character Skeleton. In Proceedings of the 3rd International Conference on Teaching and Computational Science (WTCS 2009). AISC 117: 305–309.
[40] Tan J., J. H. Lai, C. D. Wang, and M. S. Feng. 2010. Off-Line Chinese Handwriting Identification Based on Stroke Shape and Structure. In Proceedings of the 2nd International Conference on Information Engineering and Computer Science (ICIECS 2010). Wuhan, China. 1–4.
[41] Wen J., B. Fang, J. L. Chen, Y. Y. Tang, and H. X. Chen. 2012. Fragmented Edge Structure Coding for Chinese Writer Identification. Neurocomputing. 86: 45–51.
[42] Senior A. W. and A. J. Robinson. 1998. An Off-line Cursive Handwriting Recognition System. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 20(3): 309–321.

Table 1 Chinese character database

Database Name + Collected by	Language	Dataset	# writer	#Character sample			Remark
				Total	Symbol	#Chinese / Class #	
HCL2000 + Beijing University of Posts & Telecommunication for China 863 project	Chinese	700 training sets xx001-xx700 300 testing sets hh001-hh300	1000	3,755,000	3755 simplified Chinese character		• Collection was supported by the Chinese Government through China 863 high-tech project. • Contains information the writers' information includes the age, occupation, gender, education, address and others. • Can help researchers to study the styles of calligraphy of different writers and writer identification.
CASIA + Institute of Automation of Chinese Academy of Sciences	Chinese	**ISOLATED CHARACTER** OLHWDB1.0 OLHWDB1.1 OLHWDB1.2 HWDB1.0 HWDB1.1 HWDB1.2 **HANDWRITTEN TEXT** OLHWDB2.0 OLHWDB2.1 OLHWDB2.2 HWDB2.0 HWDB2.1 HWDB2.2	420 300 300 420 300 300 420 300 299 419 300 300	1,694,741 1,174,364 1,042,912 1,680,258 1,172,907 1,041,970 2,098 1,500 1,494 2,092 1,500 1,499	71,806 51,232 51,181 71,122 51,158 50,981 20,573 17,282 14,365 20,495 17,292 14,443	1,622,935/3,866 1,123,132/3,755 991,731/3,319 1,609,136/3,866 1,121,749/3,755 990,989/3,319 540,009/1,214 429,083/2,256 379,812/1,303 538,868/1,222 429,553/2,310 380,993/1,331	**Image quality :** Handwritten pages were scanned (in resolution of 300 DPT) to obtain colour images
HIT-MV + Harbin Institute of Technology,	Chinese	DGR_test Document Level DL1-DL10 Experimental Validation Train Test Textline Segmentation Experiments LGT1-LGT10 Textlines TL1-TL10	240 >780 participants	• 853 legible Chinese handwriting • Total 186444 character including letters, punctuation beside Chinese characters • Character lead to 8664 text lines • By simple computation, get following statistics: Each sample has 10.16 text lines; Each text line has 21.51 characters;			**Image quality** : Images are binarized using Otsu algorithm and saved as BMP files with no compression. The average storage space of each image is about 260K bytes **Character image size/pattern resolution:** Each writing block of legible forms is scanned into computer by Microtek ScanMaker 4180. The resolution is set to 300dpi.
ETL + Electrotechnical Laboratory under Cooperation with Japan Electronic Industry Development Association, Universities and other research organizations	• Japanese • Chinese • Latin • Numeric character	ETL1-ETL9 ETL9 ETL8	1.2 million 4000 1600	• 2965 Chinese • 71 Hiragana • 200 sample per class • 881 classes of Chinese • 75 classes of Hiragana • - 160 sample per class			**Image quality :** • ETL1-ETL9 are gray-valued image data • ETL8, ETL9: binarized image data **Character image size/pattern resolution:** 60X60 pixels 64X63 pixels 72X76 pixels 128X127 pixels

Table 2 A summary of existing methods for off-line text-independent writer recognition of Chinese handwriting

References +Year	Brief Description	Database	Feature	Classifier	Accuracy/ Result	Remark	
[31] + 2001	Gabor filter is an effective method was first introduced to analyse Chinese handwriting but it involving intensive computational cost because 2-D Gabor filter has to convolute the whole image for each orientation and each frequency. The method also does not consider the relation among the Gabor coefficients in each sub-band, only use the mean and standard derivation to represent a whole sub-band.	Handwritings collected from 17 different people	Multi-channel Gabor	Weighted Euclidean Distance (WED)	95.7%	Future	In the real application of handwriting identification there are actually at least hundreds of samples in the database or even more.
						Weakness	Intensive computational cost
						Strength	An effective method in handwriting-based writer identification
[32] + 2003	Independent Component Analysis method (ICA) based on the higher-order statistical correlation between data, extracted internal features of the image and made full use of the statistical characteristic of the input data and greatly reduced the feature extraction time but it also can produce misleading features.	Database collected from 30 people	Independent Component Analysis	Weighted Euclidean Distance (WED)	91.3%	Future	Methods need to be tested in the large scale of database
						Weakness	Producing misleading features
						Strength	Greatly reducing the feature extraction time.
[33] + 2003	Gabor filter and multi-class Support Vector machine (SVM) is adopted as the classifier, known as suitable for small samples classification and high generalization performance but the identification process is troublesome when the samples increase.	Data collected from 87 persons	Gabor filter	Multi-class Support Vector machine (SVM)	97.7% concerning Top 8	Future	Troublesome identification process need to be improved
						Weakness	Tested in a database of small samples
						Strength	The classifier suitable for small samples classification and high generalization performance
[34] + 2004	Run-length measurements are also used for handwriting analysis but only been tested in a database of small samples and it has not been used extensively.	Data collected of 19 people	Run-length measurements	Euclidean distance	91.3%	Future	Need to be tested in a large scale of database
						Weakness	Tested in a database of small samples and it has not been used extensively
						Strength	A good start of feature to be experimented
[35] + 2004	Combination of Gabor Filter and auto-correlation function has shown better performance and higher stability than either one's single algorithm. The texture images are divided into 4 sub-images from input handwriting. This division improves the identification accuracy greatly.	Carried out on 50 people's handwritings	Gabor Filter + auto-correlation function	Weighted Euclidean Distance (WED)	N/A	Future	A larger number of handwriting is needed.
						Weakness	Limitation is not tested on different standard handwriting databases
						Strength	Decrease the computational cost while still maintains high identification accuracy
[37] + 2005	Wavelet-based GGD features method achieves better results and reduces the elapsed time greatly to overcome the 2-D Gabor filter intensive computational cost, because 2-D wavelet can decompose the image into sub-bands with different frequencies and orientations.	20 Chinese handwriting samples written by 10 persons are collected in their own database	Wavelet-based GGD features	Kullback-Leibler Distance	97.8% concerning Top 25	Future	Limitation is not tested on different standard handwriting databases
						Weakness	Tested in a database of small samples
						Strength	To overcome the intensive computational cost by traditional 2-D Gabor filters
[36] + 2008	Fast Fourier Transform (FFT) is used to extract textural features which can obtain positive result but its accuracy is not high enough because of the fluctuation of the textural features which leads to the random differences between handwriting for test and handwriting in the database.	200 handwriting collected from 100 different people Enlarge data to 500 people	Fast Fourier Transform (FFT)	Weighted Euclidean Distance (WED)	98% concerning Top 10 95% concerning Top25	Future	Solving fluctuation of the textural features problem.
						Weakness	Fluctuation of the textural features leads to the random differences between handwriting for test and handwriting in the database.
						Strength	Solving [35] problem.

[38] + 2008	Hidden Markov tree model (HMT) model provides an accurate description for the statistical distribution of wavelet coefficients. It can obtain better identification result to overcome Gabor method problem which is does not consider the relationship between the Gabor coefficients in each sub-band and only uses the mean and standard derivation to represent a whole sub-band.	Handwritings written by 500 persons	Hidden Markov Tree Model	Support Vector machine (SVM)	95.4% concerning Top 25	Future	Method limitation is not tested on different standard handwriting databases
						Weakness	The method is not tested to standard database for result comparison
						Strength	Greatly decreasing the computational efficiency
[40] + 2011	Method based on stroke shapes and structure is proposed to extract the features embedded in Chinese handwriting characters, two special structures were used. Four commonly used features were computed to adjust the similarities and another sixteen features were extracted to compute the unadjusted similarity.	HanjaDB1 and SYSU databases	Bounding rectangle and TBLR quadrilateral	Weighted sum	results are encouraging - lowest FAR and FRR	Future	Method limitation is not tested on different standard handwriting databases
						Weakness	Validating the effectiveness of this proposed method accuracy for comparison with others method
						Strength	Used to compute the unadjusted similarity
[39] + 2012	Modified 2-D Gabor filter method can overcome high computational cost in traditional 2-D Gabor filter by decomposing the image into sub-bands with different frequencies and orientations and extracting features from specified sub-bands but not from the whole handwriting image.	Handwriting database collected by their own data set.	Modified 2-D Gabor filter	Weighted chi-square distance	90.3%	Future	Need to be tested to others standard database.
						Weakness	Limitation is not tested on different standard handwriting databases and need real-world implementation.
						Strength	Overcome high computational cost in traditional 2-D Gabor filter
[41] + 2012	Edge structure coding (ESC) distribution feature based on handwriting style and non-parametric discrimination can efficiently model the writing style of each writer and overcome the limitation of statistical approaches using a single scale.	HIT-MW database	HIT-MW database	Chi-square statistical measure	95.4%	Future	Need to be tested with others standard database for comparison
						Weakness	Limitation is not tested on different standard handwriting databases and need real-world implementation.
						Strength	Efficiently model the writing style of each writer and overcome the limitation of statistical approaches using a single scale