Feature Extraction for Handwritten Chinese Character Recognition Using X-Y Graphs Decomposition and Haar Wavelet

J.C. Lee, T.J. Fong and Y.F. Chang
Department of Mathematical and Actuarial Sciences, University of Tunku Abdul Rahman
46200 Petaling Jaya, MALAYSIA
changyf@utar.edu.my

***Abstract-* In this paper, a new approach of feature extraction method for handwritten Chinese character recognition called *X-Y* graphs decomposition is presented. Central to the proposed method is the idea of capturing the geometrical and topological information from the trajectory of the handwritten character using two unique decomposed graphs: *X*-graph and *Y*-graph. For feature size reduction, Haar wavelet is applied on the graphs, in which this is a new attempt of wavelet transform. Features extracted using *X-Y* graphs decomposition with Haar wavelet not only cover both the global and local features of the characters, but also are invariant of different writing styles. As a result, the discrimination power of the recognition system can be strengthened, especially for recognizing similar characters, deformed characters and characters with connected strokes. Experimental results have proved the efficiency of our proposed method and it is superior to other representative traditional feature extraction schemes with high recognition rate of 95.5%, despite of small dimensionality between 64 (inclusive) and 128 (exclusive) and less processing time.**

Keyword- Feature extraction, handwritten chinese character recognition, graph decomposition, haar wavelet

I. INTRODUCTION

Defining a feature vector for handwritten Chinese character recognition is not an easy task since Chinese character set is of large size, complex in structure, contains many similarly shaped characters and variability of writing styles from different writers. To attain high recognition rate despite all the difficulties present, many feature extraction methods have been proposed over the years. In general, feature extraction schemes can be categorized into three approaches: structural, statistical and hybrid statistical-structural [1]. Structural approaches are the earliest methods used in extracting features for Chinese characters. Among them, Attributed Relational Graph (ARG) [2] and Fuzzy Attributed Relational Graph (FARG) [3] are the most widely used. Although high recognition rates can be achieved, the increase of character complexity (i.e. increase of stroke numbers) will increase the size of the feature model enormously. Hence, it is not practical to be applied to the memory limited devices. In recent year, structural approaches have been replaced by statistical approaches due to its computational efficiency. Among the statistical based feature extraction methods, the most famous one is the direction feature [4], [5]. It describes the number of occurrences for stroke directions of each character. However, this direction feature cannot tolerate well with character shape deformation and some of them are stroke number dependent [5]. On the other hand, Hidden Markov Model (HMM) [6], [7], which is based on statistical-structural approach, is considered as the most efficient way for temporal modeling. Nonetheless, the problems of huge time consumption (due to learning procedure) and complex computation become fatal in this method.

In this correspondence, a new method, called the *X-Y* graphs decomposition, is proposed for feature extraction of handwritten Chinese characters. From the trajectory of handwritten Chinese character, the sequence of points $\left(x_t, y_t\right), 1 \le t \le N$ where $t, N \in \mathbb{Z}^+$ obtained are transformed to an *X-Y* graph, in which this graph will be decomposed into two separated graph: (i) graph of *x*-coordinate versus time sequence (called *X*-graph) and (ii) graph of *y*-coordinate versus time sequence (called *Y*-graph). For the sake of size reduction for features, Haar wavelet transform is applied to obtain a new sequence of points with smaller dimensionality. This is a new attempt for *X-Y* graphs decomposition in the feature extraction of handwritten Chinese characters, and also a new approach to the application of wavelet transform. This proposed method not only tolerates well with the variance of handwritten Chinese characters, but also overcome the problem of having complex computation, as well as being time and space consuming.

The remaining parts of this paper are organized as follows. Methodology will be explained in Section 2. Section 3 shows the experimental results. Finally, we draw our conclusion in Section 4.

II. METHODOLOGY

In this paper, preprocessing and feature extraction of the recognition system which consists of two main steps: *X-Y* graphs decomposition and Haar wavelet will be discussed. The whole process of these is depicted diagrammatically in Fig. 1 and the detail of each step will be described in the following sub-sections.

978-1-4244-5561-4/09/$26.00 ©2009

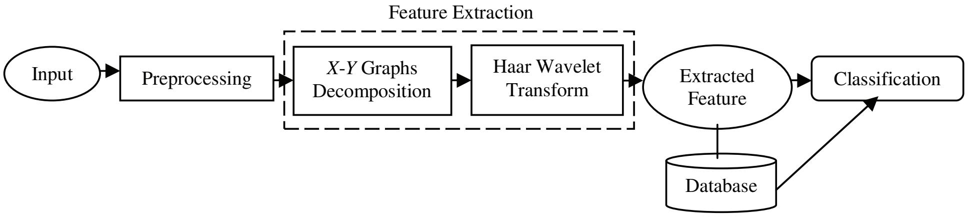


Fig. 1. Whole process of recognition system which includes preprocessing, feature extraction and classification

A. *Pre-processing*

For online handwriting recognition, the trajectory of the character is captured. The input from the digitizer corresponding to handwritten Chinese character is a sequence of points in the form of x and y coordinates, (x_t, y_t) with embedded pen-up and pen-down events when multiple strokes are involved. The Wacom Intuos®3 pen tablet is used as the digitizer in this research. For each character, 128 points are used to represent each stroke. Thus, a w-strokes character, for example, will have a total of $128\times w$ points.

In order to alleviate the negative influences caused by certain variations such as size and position variations between input characters and characters in database, preprocessing is necessary which includes two parts: (i) cropping and (ii) normalization.

(i) *Cropping*: Given the sequence of points for an input character, the maximum x and y coordinates, and also the minimum of them are determined. Then, a particular part of the original area of 256×256, that is the subarea, $[y_{\min}:y_{\max}, x_{\min}:x_{\max}]$ (from row $y_{\min}$ to row $y_{\max}$ and column $x_{\min}$ to column $x_{\max}$) is cropped.

(ii) *Normalization*: The sequence of points (x_t, y_t), which ranged within the cropped subarea are normalized to the size of 128×128, as shown below.

$$x_t^* = 127\left(\frac{x_t - x_{\min}}{x_{\max}-x_{\min}}\right)+1 \qquad (1)$$

$$y_t^* = 127\left(\frac{y_t - y_{\min}}{y_{\max}-y_{\min}}\right)+1 \qquad (2)$$

B. *X-Y Graphs Decomposition*

In the *X-Y* graphs decomposition, the sequence of preprocessed points (x_t^*, y_t^*); $1\le t\le N = 128\times w$ (refer to Eq. (1) and (2)) is transformed into two separated graph: (i) graph of x-coordinate versus time sequence (called *X*-graph) and (ii) graph of y-coordinate versus time sequence (called *Y*-graph), in which these two graphs are described in Fig. 2. The pattern of the graphs depends on how the character is written. As an example, the values in the *X*-graph rise while in *Y*-graph the values remain unchanged, when the second stroke, i.e. the horizontal stroke is written. The feature vectors are then constructed from the sequences of points in *X*-graph and *Y*-graph as $\left\{\left[x_1^*,\dots x_N^*\right]^T, \left[y_1^*,\dots y_N^*\right]^T\right\}$.

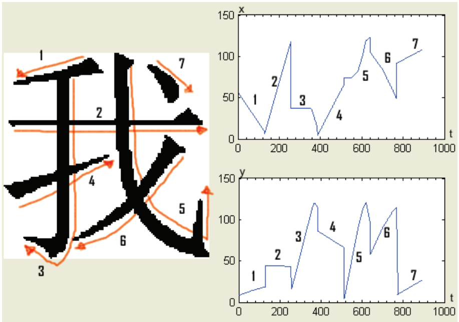
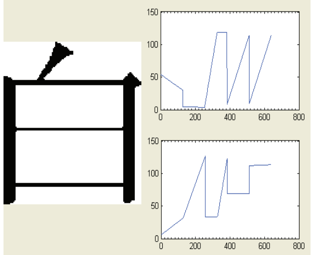


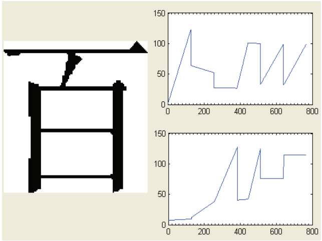
Fig. 2. *X*-graph (above) and *Y*-graph (below) of Chinese character '我' (means I or me)

The principle of *X-Y* graphs decomposition is to trace the pattern of each Chinese character based on the trajectory of handwriting. This new method is considered as holistic approach which extracts the character as a whole without any preliminary identification of strokes. It designed in accordance with the following properties:

(i) *Uniqueness*
The trajectory of each Chinese character is unique and in turn forms the unique *X*-graph and *Y*-graph. Even for similarly shaped characters, the graphs plotted are also of different shapes. An example is illustrated in Fig. 3. Hence, both the *X*-graph and *Y*-graph can be concluded to contain the most essential information of Chinese characters and have strong discriminative ability.



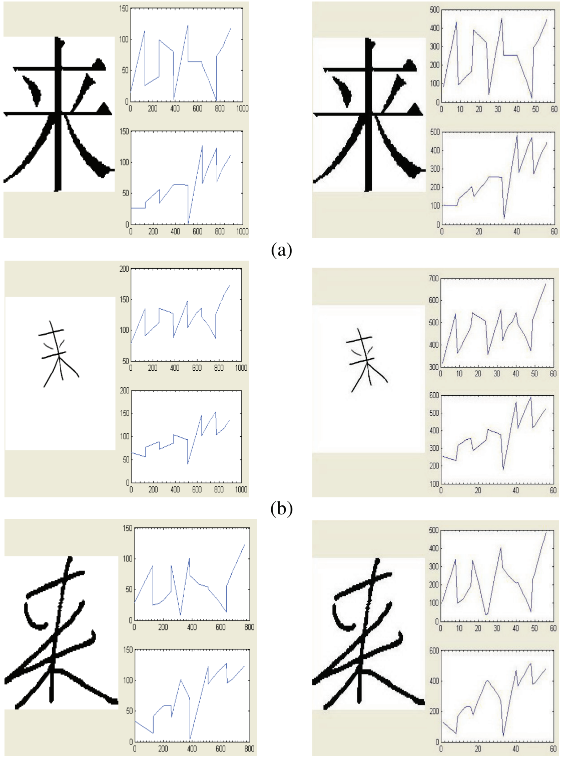
(a)



(b)

Fig. 3. *X*-graphs (above) and *Y*-graphs (below) plotted for two similar characters: (a) '白' (means white) (b) '百' (means hundred)

(ii) *Invariant of different writing styles*
Different writing style is the main problem faced in handwriting recognition. To address this problem, *X-Y* graphs decomposition is an excellent solution since *X*-graph and *Y*-graph plotted for the Chinese characters are invariant of size and position of the written character. In other words, the pattern of both *X*-graph and *Y*-graph will be retained for the same characters despite of variable size and position. Besides, the preservation of graph patterns for deformed characters can helps in handling cursive characters or characters with connected strokes. The instances of these are presented in Fig. 4.



(a)

(b)

(c)

Fig. 4. *X*-graphs (above) and *Y*-graphs (below) plotted for character '来' (means come). Left shows the graphs before Haar wavelet transform while right shows the graphs after Haar wavelet transform. (a) Regular character in database (b) Characters written in variable size and position (c) Character written with connected strokes

(iii) Simplicity
In *X-Y* graphs decomposition, neither complicated process nor heavy computation is involved. Decomposing the character coordinates into two separated *X-Y* graphs and obtaining the feature vectors from the corresponding graphs are the only tasks that required implementing. As a result, the simplicity of this approach will boost the efficiency and speed of the recognition system.

Therefore, due to the above properties, the decomposed *X*-graph and *Y*-graph are good representation for Chinese character.

C. Haar Wavelet

Among different types of wavelets, Haar wavelet is the most fundamental and widely used due to its simplicity and efficiency. Thus, it is utilized in order to reduce the size of the feature vector by converting the feature vectors into new sequences of points $\boldsymbol{a_i} = [a^x_i, a^y_i]$, $1 \le i \le M$, $2^5 \le M < 2^6$, which is the approximation coefficients of Haar transform as defined in the following.

$$a^x_i = \frac{x^*_{2i-1} + x^*_{2i}}{\sqrt{2}} \qquad (3)$$

$$a^y_i = \frac{y^*_{2i-1} + y^*_{2i}}{\sqrt{2}} \qquad (4)$$

Notice that Haar wavelet transform is performed in several stages or levels and in each level, the sequence of points will be reduced to half of its size. The new extracted feature $\boldsymbol{a_i}$ is then used for classification. Similarly, $\boldsymbol{b_i} = [b_{xi}, b_{yi}]$ represents the extracted feature of the characters in database. It has been proved that the extracted features still retain the important information of the characters after size reduction by Haar wavelet transform, in which the examples are demonstated in Fig. 4. Notice that the graph patterns are still preserved after undergoing Haar wavelet transform.

III. Experimental Results

In order to evaluate the performance of our feature extraction method in recognition system, database is setup for testing. It is based on Jun Da's modern Chinese character frequency list [8], which he generated from a large corpus of Chinese texts collected from online sources. Our database is composed of the first 3000 frequently used simplified Chinese characters listed in [8] and 10 digits from 0 to 9. These Chinese characters in database are written in the font style of *songti*, due to its widely used, and each character sample is normalized to the size of 128 × 128. For testing, 20 different Chinese characters are collected from 10 different writers. The writers are requested to write with stroke order restriction. In our recognition system, minimum distance (MD) classifier is applied for classification.

For comparison with our proposed method, some representative traditional feature extraction schemes which are (i) Attributed Relational Graph (ARG) (ii) whole character-based hidden Markov model (HMM) and (iii) directional feature densities (DFD) are selected. The efficiency of our methods will be judged from the following three different perspectives.

A. *Recognition Rate*

In order to validate our new designed feature extraction method, accuracy is the major aspect that has to be concerned about.

Table 1 shows the recognition rates of the three existing feature extraction schemes studied from [2], [7] and [4] respectively, and the current methods in this paper. Although different database and testing samples are used, it is still sufficient to prove that our method is superior with higher recognition rate, since it has stronger discriminative ability when dealing with similar and deformed characters.

TABLE I
RECOGNITION RATES FOR FOUR DIFFERENT FEATURE EXTRACTION METHODS

Method	Database	Testing Samples	Recognition Rate (%)
Attributed Relational Graph (ARG) [2]	320 frequently used Chinese Characters each with stroke number between 9 and 11	375 Chinese characters collected from 8 writers	94.20 (with connected strokes) 98.9 (with correct strokes)
Whole Character-Based Hidden Markov Model (HMM) [7]	Kuchibue database [7]	5 writers × 881 Kanji characters.	90.00
Directional Feature Densities (DFD) [4]	2965 Kanji characters with 380 samples each	10 writers × 2965 Kanji characters.	91.78
X-Y Graphs Decomposition with Haar Wavelet	3000 frequently used simplified Chinese characters and 10 digits	10 writers × 20 Chinese characters.	95.5

Remark: Kanji characters in Japanese are the same as Chinese characters.

B. *Feature Size*

The size of the features extracted determines the memory space required for the recognition system. For commercial purpose, it is ideal to have a small memory requirement, so that it can be directly embedded into hand-held devices.

From Table 2, it implies that feature extraction using *X-Y* graphs decomposition with Haar wavelet gives the smallest feature size which is most practical and appropriate for memory limited devices.

C. *Processing Time*

The speed of the recognition system is one of the important factors for a good-quality recognition system, in which reduction of processing time is necessary.

Table 3 indicates that our proposed method results in the least processing time compared with other feature extraction schemes. This is mainly because of the simplicity and efficiency of our algorithm, in which no heavy computation is involved totally. In this research, the testing platform is on a Dell Vostro 1400 N-Series notebook of Intel(R) Core(TM)2 Duo Processor T5470 and 1GB (2×512 MB) 667MHz Dual Channel DDR2 SDRAM. The processing time for feature extraction proposed in this paper is 0.1 second per character.

TABLE II
FEATURE SIZES FOR FOUR DIFFERENT FEATURE EXTRACTION METHODS

Methods	Feature Size (Dimension)
Attributed Relational Graph (ARG) [2]	$(stroke\ number)^2$, increase of stroke number will increase the feature size massively. For characters with 22 strokes, dimension = $22^2 = 484$.
Whole Character-Based Hidden Markov Model (HMM) [7]	Sum of the size of parameters $\{a_{ij}\},\{b_{ik}^1\},\{b_{il}^2\},\{\pi_i\}$ and *N*, where $\{a_{ij}\},\{b_{ik}^1\}$ and $\{b_{il}^2\}$ are matrices, $\{\pi_i\}$ is vector and *N* is scalar.
Directional Feature Densities (DFD) [4]	$8 \times 8 \times 4 = 256$
X-Y Graphs Decomposition with Haar Wavelet	Between $2^5 \times 2 = 64$ (inclusive) and $2^6 \times 2 = 128$ (exclusive)

TABLE III
PROCESSING TIMES FOR FOUR DIFFERENT FEATURE EXTRACTION METHODS

Methods	Processing Time
Attributed Relational Graph (ARG) [2]	Preprocessing time + defining nodes and relations for each stroke time + constructing ARG time + computing generalized relation matrix time
Whole Character-Based Hidden Markov Model (HMM) [7]	Preprocessing time + quantization time + computing parameters time + computing joint distribution time + learning time
Directional Feature Densities (DFD) [4]	Preprocessing time + computing directional feature vector time + defining vector for square areas time + dimension condensation time
X-Y Graphs Decomposition with Haar Wavelet	Preprocessing time + X-Y graph decomposition time + wavelet transform time

Obviously, from the experimental results, *X-Y* graphs decomposition associated with Haar wavelet leads to a better performance in term of recognition rate, feature size and processing time than the other three existing feature extraction methods. Therefore, it is undoubted that our new approach is more efficient for handwritten Chinese character recognition.

IV. CONCLUSION

Making use of the rich geometrical and topological characteristic of the handwritten character trajectory, this paper presents a new pattern descriptor, *X-Y* graphs decomposition, to extract informative features. To reduce the size of features, new approach of Haar wavelet transform is proposed, in which it is applied on the graphs instead of images. The unique *X*-graph and *Y*-graph used to represent each Chinese character cover both the global and local features of the characters and so enhance the discrimination power. In addition, the most beneficial thing is that both the graphs are invariant of different writing styles. It tolerates well not only for the regular handwritings, but also to natural writings with shape deformation and stroke connection. On the other hand, due to small dimensionality of features and simplicity of algorithm, the performance of the recognition system can be boosted in term of memory space and speed. Experimental results show that features extracted using *X-Y* graphs decomposition with Haar wavelet can achieve a promising recognition rate of 95.5% with dimensionality between 64 (inclusive) and 128 (exclusive) in less processing time. Therefore, this new idea of feature extraction method which gives a great improvement in the performance of the recognition system has provided a new inspiration for this research area. However, the limitation of this method is that it is stroke order dependent and will affect the shape of the graphs severely. Rearranging the pattern of the graphs will allow the stroke order variation but it is absolutely a difficult task. These will be left for further investigation.

ACKNOWLEDGMENT

Y. F. Chang wishes to thank H. V. Chen and W. S. Ng at UTAR for their useful comments. This work was supported by E-Science Grant (No: 01-02-11-SF0053) of Ministry of Science, Technology and Innovative, Malaysia.

REFERENCES

[1] C.L. Liu, S. Jaeger, M. Nakagawa, Feb. 2004: Online Recognition of Chinese Characters: The-State-of-the Art, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, No. 2, pp. 198-213.

[2] J.Z. Liu, W.K. Cham, M.M.Y. Chang, 1996: Online Chinese Character Recognition Using Attributed Relational Graph Matching, *IEE Proc. Vision Image Signal Processing*, vol. 143, no, 2, pp. 125-131.

[3] J. Zheng, X. Ding, Y. Wu, 1997: Recognizing On-Line Handwritten Chinese Character via FARG Matching, *Proceedings of the 4th International Conference on Document Analysis and Recognition*, vol. 2, pp. 621-624.

[4] A. Kawamura et al., 1992: On-Line Recognition of Freely Handwritten Japanese Characters Using Directional Feature Densities, *Proceeding of the 11th International Conference on Pattern Recognition,* vol. 2, pp. 183-186.

[5] F. Kimura, T. Wakabayashi, S. Tsuruoka, Y. Mayake, 1997: Improvement of Handwritten Japanese Character Recognition Using Weigthed Direction Code Histogram, *Pattern Recognition*, vol. 30, no. 8, pp. 1329-1337.

[6] H. Shimodaira, T. Sudo, M. Nakai, S. Sagayama, 2003: On-line Overlaid-Handwriting Recognition Based on Substroke HMMs, *Proceedings of the 7th International Conference on Document Analysis and Recognition*, vol. 2, pp.1043.

[7] K. Takahashi, H. Yasuda, T. Matsumoto, 1997: A Fast HMM Algorithm for On-Line Handwritten Character Recognition, *Proceedings of the 4th International Conference on Document Analysis and Recognition,* pp. 369-375.

[8] Ju Dan's WebCentral, Chinese Text Computing, 2004: Modern Chinese Character Frequency List, http://lingua.mtsu.edu/chinesecomputing/statistics/char/list.php?Which=MO