A Feature Extraction Technique for Online Handwriting Recognition

Brijesh Verma[1], Jenny Lu[1], Moumita Ghosh[2], Ranadhir Ghosh[2]

[1]Faculty of Informatics & Communication, Central Queensland Universtiy, Rockhampton, QLD 4702, Australia

E-mail: {b.verma, j.lu}@cqu.edu.au

[2]School of Information Technology and Mathematical Sciences, University of Ballarat, Ballarat, VIC 3350, Australia

E-mail: {m.ghosh, r.ghosh}@ballarat.edu.au

ABSTRACT

The paper presents a feature extraction technique for on-line handwriting recognition. The technique incorporates many characteristics of handwritten characters based on structural, directional and zoning information and combines them to create a single global feature vector. The technique is independent to character size and it can extract features from the raw data without resizing. Using the proposed technique and a Neural Network based classifier, many experiments were conducted on UNIPEN benchmark database. The recognition rates are 98.2% for digits, 91.2% for uppercase and 91.4% for lowercase.

1. INTRODUCTION

Online handwriting recognition is one of the very complex and challenging problems [1][2][3] because of variability on size, writing style of hand-printed characters [4], and duplicate pixels caused by a hesitation in writing or interpolate non-adjacent consecutive pixels caused by fast writing. As mentioned in the literature [5] [6], the feature extraction plays an important role in the overall process of handwriting recognition. Many feature extraction techniques [5] [6] [7] [8] [9] [10] [11] [12] have been proposed to improve overall recognition rates; however most of them are depended on the size and slope of handwriting characters. They require very accurate resizing, slant correction procedure or technique otherwise they achieve very poor recognition rates. Also most of existing techniques use only one characteristic of a handwritten character. This research focuses on a new feature extraction technique that does not use resizing of a character and it uses novel characteristics of a character and combines them to create a global feature vector.

The remainder of this paper consists of three sections. Section two details the proposed technique. The discussion and analysis of the experimental results take place in section three. Section four presents the conclusion and the future research.

2. PROPOSED TECHNIQUE

Figure 1 outlines proposed technique that can be classified into eight modules such as dehooking, extract feature points, stroke extracting, calculate PEN-UP, extract zones and directions of start point and end point, extract changes in writing direction, calculate height/width ratio and extract zone information which creates a global feature vector and uses a back-propagation neural network based classifier. The models are described below in detail.

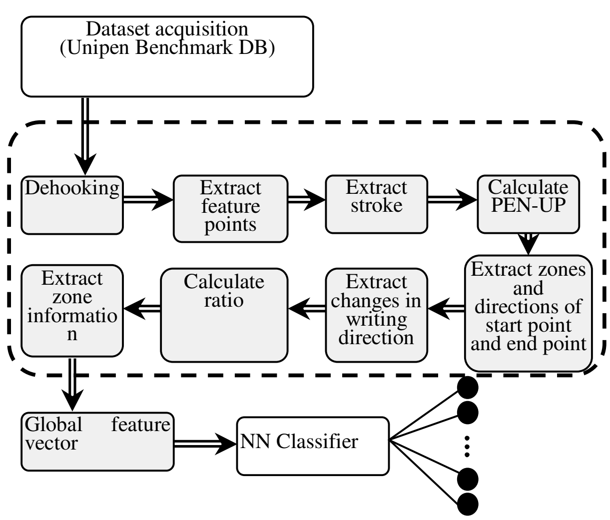

FIGURE 1. BLOCK DIAGRAM OF PROPOSED TECHNIQUE

2.1. DATASET ACQUISITION

The training and testing datasets used in this research were taken from a benchmark database (UNIPEN CD-ROM VERSION 7) [13].

2.2. FEATURE EXTRACTION

Dehooking: Hooks can occur at the beginning and end of strokes due to inaccuracies in pen-down detection and rapid or erratic motion in placing the stylus on, or lifting it off the tablet. Usually, hooks can be detected by their location, small size and large angular variation [14] [15]. In this research, we used the following approach to remove hooks. If the vector direction length is less than threshold, remove it from the dataset, otherwise keep it. The threshold was 3% of the diagonal line (DL). The diagonal line is calculated as follows:

$$DL = \sqrt{Height^2 + Width^2} \qquad (1)$$

Eight directions vector [16] was used in this research.

Extract feature points: From the database we can extract all points of every character. However not all of the points are useful. Based on the idea that reducing the problem data into a manageable amount of information without discarding valuable or vital information, the first step is to remove most unwanted characteristics and to reduce the volume of data. At the same time, it can acquire the useful information.

Extract stroke: Stroke is defined as continuous path of the pen from the moment it is placed on the writing surface until the moment it is lifted up. In this case, stroke is the series of points from "PEN-DOWN" point to "PEN-UP" point. The feature calculated in this research is the number of strokes for one character. Thus we can simply get the stroke by counting how many "PEN_DOWN" occurred in the dataset for one character or digit.

Calculate PEN-UP: Only stroke is not enough, because most of the time different character may get the same number of strokes. Therefore, in this research, we also use PEN-UP as a feature to check how well the recognized character or digit matches to the standard one (the average for the same character in the database). This feature is calculated by using the average strokes of a specific letter or a digit as an input using the membership function as follows:

$$\text{PEN-UP} = e^{-|average - x|} \qquad (2)$$

where x is the real strokes for the specific character.

Zones and directions of start and end point: Using the lateral coordinates and the longitudinal coordinates of the first point and the second point, we calculate start point direction (SD) and the end point direction (ED).

FIGURE 2. 6 REGIONS ZONING INFORMATION

As for the zone information, we use the six regions to encode it. A comparative analysis [17] compared the results by using 3 regions, 6 regions and 9 regions and concluded that the 6 regions achieve the best result for recognition. Therefore the whole region of character was separated into six zones in this research.

Change of writing direction: On-line handwriting has achieved better results than off-line because it is a fact that more information may be captured in the case of on-line such as the direction, speed and the order of strokes of the handwriting. The change of writing direction is regarded as the changing from pen going up (down) to down (up) or going left (right) to right (left).

For one particular character or digit, the order of strokes may be very different, but the change of writing direction will be similar. Based on vector direction, we can get the jag point where the writing direction changed. Using the coordinates of continuous two jag points we can easily get how many times the direction is changed.

Calculate height/ width ratio: Distortions of original patterns are the big problem in handwriting recognition. Mostly, the distortion takes place during the normalization of an image such as the letter “l” as shown below in Figure 3.

FIGURE 3. SIZE NORMALIZATION OF PATTERN “L”

After the normalization the letter “l” looks like the letter “e”, thus this research try to use the width and height features rather than normalize the patterns. Whereas, the most important thing is the rate of width/height, using this rate we can easily know the difference of “m” and “f”.

Extract zone information: Zone information is the global characteristic and it can be potentially useful. In this research, it is defined as the clear boundaries of the character. We used the Ymax, Ymin, Xmax and Xmin to define the boundary of a character, and then separated it into 6 regions. So whatever the size of data, it will be included in 6 regions.

2.3. GLOBAL FEATURE VECTOR

A global feature vector is based on a number of characteristics as described in previous sections. The feature vector represents a single feature for the character. The format of feature vector is as follows:

PEN-UP	SZ	SD	EZ	ED	GD
GU	GL	GR	Stroke	Ratio	Z1
Z2	Z3	Z4	Z5	Z6	

where GD: writing direction going down, GU: writing direction going up, GL: writing direction going left, GR: writing direction going right, Z1, Z2, Z3, Z4, Z5 and Z6: 6 zones

2.4 BACK-PROPAGATION NEURAL NETWORK CLASSIFIER

A back-propagation neural network with a single hidden layer is used as a classifier.

3. EXPERIMENTAL RESULTS AND ANALYSIS

Various experiments were conducted using the proposed feature extraction technique and different parameter settings. The results are presented below.

3.1. CLASSIFICATION RATES

The following table shows the best results obtained using the proposed technique.

TABLE 1. CORRECT CLASSIFICATION RATES

Hidden Units	Digit [%]		Upper[%]		Lower[%]	
	Train	Test	Train	Test	Train	Test
30	100	93.8	96.24	89.2	94.64	89.6
40	100	98.2	98.00	91.2	96.20	91.4
50	100	94.0	99.02	90.8	97.24	91.8

As can be seen, the best classification rate for digit was achieved using 40 hidden-units. The best classification rate for uppercase was achieved using 40 hidden units as well. The best classification rate for lowercase was achieved with 50 hidden units.

3.2. PROBLEMS WITH CLASSIFICATION RATES

During the experiment, we found that for some particular characters, the classification rate is poor. It was deduced that the following factors influenced some or most of the poor results obtained by novel feature extraction technique.

Database: the benchmark database used for experiments consists of more than 150 writers, they came from 5 different countries and have different educational background. It also involved both male and female, left and right hand writers. This leads to the input pattern vary from each other. Also the sample rate for database is from 10 points per second to 200 points per second. This made the original data looks very different to each other even for the same character.

Variations of character: The numerous variations lead to the number of styles to write a certain character can be considered as infinite. The order of the strokes always different for different writers. Moreover, the direction of a stroke varies if the writer tries to minimize the time that pen is lifted up or by some personal preferences. Also, the form of the strokes can be varied. For example, the straight strokes can be curved as bows.

FIGURE 4. CHARACTER "E", "G" AND "L"

Similarity of characters: After analyzing the result files that describe the target and actual outputs, we found that some classes got very high recognition rate, whereas some got the very low recognition rate. The top 3 recognition rates are as follow:

TABLE 2. TOP 3 CLASSIFICATION RATES

	Top1 [%]		Top2 [%]		Top3 [%]	
Digit	1	100	0	100	2	98.0
Uppercase	E	99.4	C	98.0	T	98.0
Lowercase	o	100	s	99.5	y	99.3

Some characters were easily recognized as other particular characters. Such as the capital P and capital D, lower u and lower v, lower l and lower e, and so on.

3.3. COMPARISON WITH THE EXISTING TECHNIQUE

The results described in [5] by Parizeau, Lemieux and Gagne were 97.0% for isolated digits and 85.6% for isolated lowercase. The results were conducted using a common neural network classifier, and trained with back-propagation as well. Most important thing is they obtained these results by using Unipen dataset. The comparison is valuable due to so many similar backgrounds. Also the LVQ system used in [18] achieved 95.4% for digits, 89.68% for uppercase and 88.28% for lowercase with Unipen dataset version 7.

In Figure 5, the black bars represent the results using the proposed technique, the others are the results of existing techniques described in [5] and [18] respectively. From the graph (Figure 5) we can see, compare with Parizeau et al., the classification rate was improved 1.2% for digits and 6.8% for lowercase by using the proposed feature extraction technique. For uppercase, there is no information available in ref. [5]. When comparing the classification rate between the results obtained by LVQ system described in [18] and the proposed technique, the classification rate has been improved 2.8% for digits, 1.5% for uppercase and 3.1% for lowercase by using the proposed feature extraction technique.

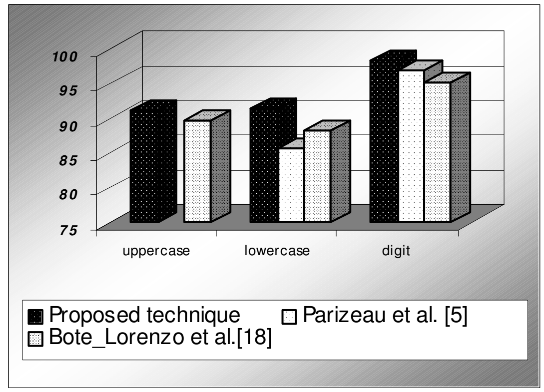

FIGURE 5. COMPARISON OF CLASSIFICATION RATE

4. CONCLUSIONS AND FUTURE RESEARCH

We have presented a novel approach that directly extracts many novel characteristics from the on-line handwritten character without resizing. The main focus of this paper was a feature extraction technique that uses structural, the change of writing direction, and zoning information to create a single global feature vector. Feature extractor in conjunction with back-propagation neural network was implemented and tested. The results were compared and analyzed. As can be seen from the results, the global feature vector is effective and promising. In our future research, we would like to optimize our global feature vector. The GA will be used to optimize bits in the global feature vector.

5. REFERENCES

[1] Plamondon, R., & Srihari, S., "*On-line & Off-line Handwriting Recognition: A Comprehensive Survey*", IEEE PAMI, vol. 22, no.1, pp. 63-84, 2000.

[2] Xiaolin, L., and Yeung, D.-Y., "*On-line Handwritten Alphanumeric Character Recognition Using Dominant Points In Strokes*", Pattern Recognition, vol. 30, no. 1, pp. 31-44, 1997.

[3] Plamondon, R., Lopresti, D.P., Schomaker, L.R.B.and Shrihari, R., "*Online Handwriting Recognition*", Wiley Encyclopedia of Electrical and Electronics Engineering, John G. Webster, vol.15, pp.123-146, John Wiley & Sons, 1999.

[4] Casey, R., "*Moment Normalization Of Handprinted Characters*", IBM Journal of Research Development, pp. 548-57, 1970.

[5] Parizeau, M., Lemieux, A., and Gagne, C., "*Character Recognition Experiments Using Unipen Data*", Proceedings, Sixth International conference on Document Analysis and Recognition, pp. 481-485, 2001.

[6] Jaeger, S., Manke, S., Reichert, J., and Waibel, A., "*Online Handwriting Recognition: The NPen++ Recognizer*", International Journal on Document Analysis & Recognition, vol. 3, no. 3, pp. 169-180, 2001

[7] Devijver, P. A., and Kittler, J., "*Pattern Recognition: A Statistical Approach*", London: Prentice-Hall, 1982.

[8] Chakraborty, B., and Chakraborty, G., "*A New Feature Extraction Technique For On-line Recognition Of Handwritten Alphanumeric Characters*", Information Sciences, vol. 148, no. 1, pp. 55-70, 2002.

[9] Ping, Z., and Lihui, C., "*A Novel Feature Extraction Method And Hybrid Tree Classification For Handwritten Numeral Recognition*", Pattern Recognition Letters, vol. 23, no. 1, pp. 45-56, 2002.

[10] Gomez Sanchez, E., Gago Gonzalez, J.A., Dimitriadis, Y.A., Cano Izquierdo, J.M., and Lopez Coronado, J., "*Experimental Study Of A Novel Neuro-fuzzy System For On-line Handwritten UNIPEN Digit Recognition*", Pattern Recognition Letters, vol. 19, no. 3, pp. 357-364, 1998.

[11] Trier, Ø. D., Jain, A.K., and Taxt, T., "*Feature Extraction Methods For Character Recognition – A Survey*", Pattern Recognition, vol. 29, no. 4, pp. 641-662, 1996.

[12] Hammandlu, M., Murali Mohan, K. R., Chakraborty, S., Goyal, S. and Choudhury, D. R., "*Unconstrained Handwritten Character Recognition Based On Fuzzy Logic*", Pattern Recognition, vol. 36, no. 3, pp. 603-623, 2003

[13] http://hwr.nici.kun.nl/unipen/

[14] Guerfali, W., and Plamondon, R., "*Normalizing And Restoring On-line Handwriting*", Pattern Recognition, vol. 26, no. 3, pp. 419-431, 1993.

[15] Mandler, E., "*Advanced Preprocessing Technique For On-line Script Recognition Of Nonconnected Symbols*", Proc. 3rd Int. Symp. On Handwriting and Computer Applications, pp. 64-66, 1987.

[16] Freeman, H. "*Computer Processing Of Line-drawing Images*", Computing Surveys, vol. 6, no. 1, pp. 57-97, 1974.

[17] Li, X., Plamondon, R. and Parizeau, M., "*Model-based On-line Handwritten Digit Recognition*", Fourteenth International Conference on Pattern Recognition, vol.2, pp. 1134-1136, 1998.

[18] Bote_Lorenzo, M., Dimitriadis, Y., and Gomez-Sanchez, E., "*Automatic Extraction Of Human-recognizable Shape And Execution Prototypes Of Handwritten Characters*", Pattern Recognition, vol.36, no. 7, pp. 1605-17, 2003.