A New Segmentation Algorithm for Handwritten Word Recognition

A New Segmentation Algorithm for Handwritten Word Recognition

M. Blumenstein[1] and B. Verma[1,2]

[1]School of Information Technology
Griffith University-Gold Coast Campus
PMB 50, Gold Coast Mail Centre,
QLD 9726, Australia
Telephone: +61 7 5594 8738
Fax: +61 7 5594 8066
E-mail: {m.blumenstein, b.verma}@gu.edu.au

[2]Department of Computer Engineering
and Computer Science
University of Missouri-Columbia
Columbia, MO 65211, USA
Telephone: +1 573 8846464
Fax: +1 573 8828318
E-mail: bverma@ece.missouri.edu

Abstract

An algorithm for segmenting unconstrained printed and cursive words is proposed. The algorithm initially over-segments handwritten word images (for training and testing) using heuristics and feature detection. An Artificial Neural Network (ANN) is then trained with global features extracted from segmentation points found in words designated for training. Segmentation points located in "test" word images are subsequently extracted and verified using the trained ANN. Two major sets of experiments were conducted, resulting in segmentation accuracies of 75.06% and 76.52%. The handwritten words used for experimentation were taken from the CEDAR CD-ROM. The results obtained for segmentation can easily be used for comparison with other researchers using the same benchmark database.

1. Introduction

Although many highly accurate systems have been developed to recognise handwritten numerals and characters [1-5], their success has not carried onto the handwritten word recognition domain. This has been ascribed to the difficult nature of unconstrained handwriting, including the diversity of character patterns, ambiguity and illegibility of characters, and the overlapping nature of many characters in a word [6]. Many complex procedures are required to recognise unconstrained handwriting. One such procedure is that of character segmentation. Researchers have acknowledged the importance that segmentation plays in the handwriting recognition process [7-9]. This is precisely why more innovative and accurate methods need to be employed and compared to the work of other researchers.

This research attempts to integrate both heuristic and intelligent methods for the segmentation of cursive and printed handwritten words. For the initial task of segmentation, a feature-based heuristic algorithm is used to locate prospective segmentation points in handwritten words. An ANN trained with valid segmentation points from a database of handwritten words is used to assess the correctness of the segmentation points found by the algorithm.

The remainder of the paper is broken down into 6 sections. Section 2 discusses some previous research, Section 3 describes the proposed segmentation technique, Section 4 provides experimental results, a discussion of the results takes place in Section 5, Section 6 discusses future research and a conclusion is drawn in Section 7.

2. Previous Work and Related Research

Researchers have utilised many different approaches for both the segmentation and recognition tasks of word recognition. Some researchers have used conventional, heuristic techniques for both character segmentation and recognition [10, 11] while others have used heuristic techniques for segmentation followed by ANN based methods for the character/word recognition process [12, 13]. For printed and cursive handwriting, some of the most successful results have been obtained with the use of techniques that possess tightly coupled segmentation and recognition components [14]. These techniques are lexicon-directed and compute the best way segmented character images and sub-images (primitives) of a word can be assembled and matched to represent a possible string in a lexicon. These techniques do not employ complex segmentation algorithms. As a result, the number of segmentation points found can be relatively high.

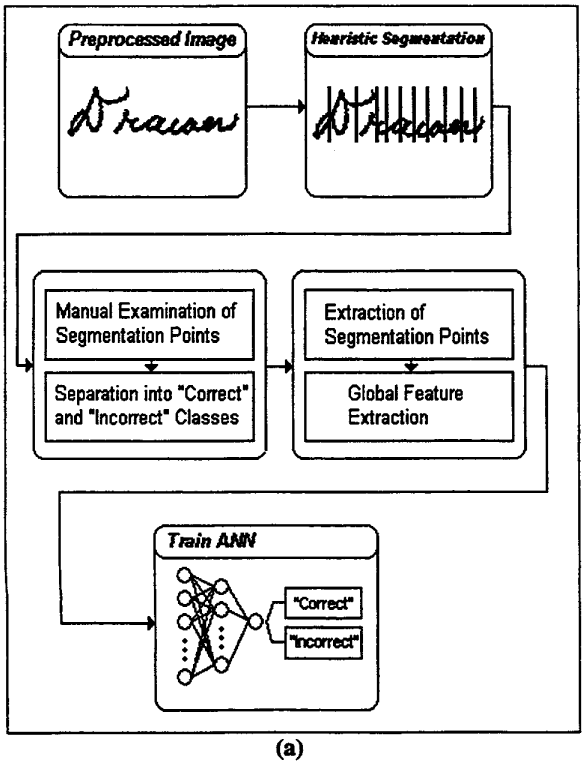
0-7803-5529-6/99/$10.00 ©1999 IEEE

By employing more powerful segmentation techniques, it is possible to reduce the number of false segmentation points found. This in turn can increase the speed and efficiency of the system by reducing the number of primitive combinations that need to be assembled and checked against the lexicon of words.

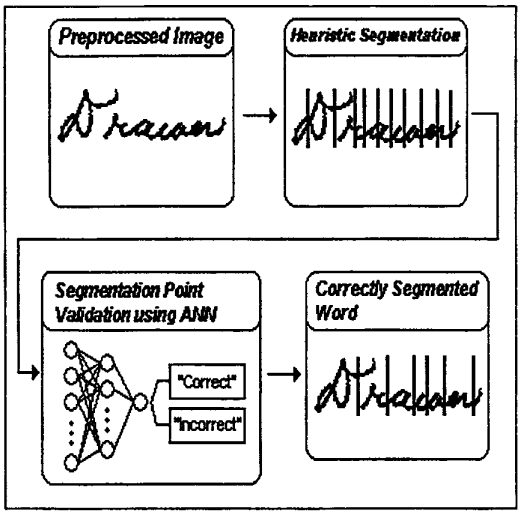
Recently, some researchers have turned to ANNs to assist in the segmentation process [15, 16]. Unfortunately there have only been a small number of authors that have detailed their findings for segmentation of cursive words. Due to the fact that most segmentation techniques are usually explained in the context of a complete system, researchers tend to measure the success of their system by their findings from the character or word recognition phases only. Cursive word segmentation deserves particular attention as it has been acknowledged as the most difficult of all handwriting segmentation problems [17].

3. Proposed Segmentation Technique

This section addresses the steps required to segment the handwritten words using the proposed technique. An overview of the technique is provided in Figure 1.



(a)



(b)

Figure 1. Proposed Segmentation Technique (a) Stage 1: Training Phase (b) Stage 2: Testing Phase

3.1 Preprocessing

Prior to segmentation and recognition, it was necessary to preprocess all word images. Initially the images were all in a grey-level format. Otsu's thresholding algorithm [18] was used to binarise the images. Many of the cursive and even some of the printed words were slanted at various angles, it was therefore necessary to employ a slant detection and correction technique [10].

3.2 Overview of the Heuristic Algorithm

For both training and recognition phases, an heuristic feature detection algorithm is used to locate prospective segmentation points in handwritten words. Each word is inspected in an attempt to locate characteristics representative of segmentation points. The object of the algorithm is to oversegment all the words (Figure 2).

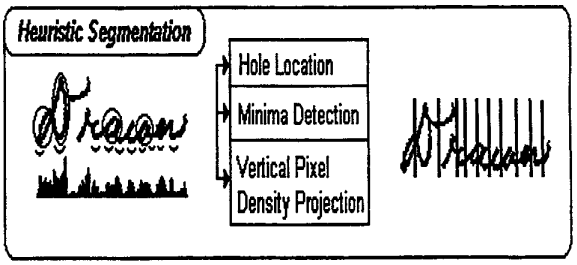


Figure 2. Heuristic Segmentation Algorithm

3.3 Character Width Estimation

Firstly, the average character width of each word is estimated by locating segregated characters and calculating their average width. If no segregated characters are found within a particular word, the average word height is used to obtain an estimate of character width. It is understood that the width of a character in most cases is less than its height. Therefore as an approximate character width estimate, we use a percentage of the average word height to provide a rough solution.

3.4 Contour Extraction

Next, upper and lower word contours are determined to enable the location of upper and lower minima in the word (possible ligatures in cursive writing). The contour information is analysed to find these "valleys" or "saddles" that may represent ligatures. Each pixel on the contour is analysed to see whether there is a slope change associated with the pixels immediately to its left or right. Possible minima in a word (as found by the algorithm), may be seen in Figure 2.

3.5 Determining Vertical Pixel Density and Hole Location

A histogram of vertical pixel densities is calculated for each word. The histogram is obtained by calculating total runs of vertical pixels for each column of the word image where black pixels exist. The histogram is examined for minima (low vertical pixel density) which may further confirm the location of possible segmentation points in the word: Figure 2.

Words are also scanned for possible holes i.e. areas in a word that may be occupied by an "o", "a", "b" etc. The search for "holes" only proceeds within areas in a word that are suspected of having a segmentation point. The area to be examined spans a distance half the average character width immediately to the left and right of the suspected segmentation point. The contour of the word segment in this area is inspected to determine if there exist any totally enclosed regions. If any regions are found to exhibit "hole"-like characteristics they are marked as being inappropriate to accommodate possible segmentation points: Figure 2.

3.6 Segmentation Point Distribution

Finally, the word is further analysed to determine whether segmentation points have been properly distributed throughout the word. Clusters of proximate segmentation points are analysed and are reduced in number so that only small collections of more likely points representing a particular area may exist. To conclude, areas in a word which are lacking segmentation points are examined. Therefore, if an area with a width larger than that of the calculated average character width has a sparse distribution of segmentation points, a segmentation point is forced in the most likely area of the word segment. The result is a set of over-segmented words that await ANN verification.

3.7 Training Phase of Segmentation Technique

Prior to ANN training, the heuristic algorithm is used to segment all words that shall be required for the training process. The segmentation points output by the heuristic feature detector are manually analysed so that the x-coordinates may be categorised into "correct" and "incorrect" segmentation point classes. For each segmentation point in a particular word (given by its x-coordinate), a matrix of pixels is extracted and stored in an ANN training file. Each matrix is first normalised in size, and then significantly reduced in size by a simple feature extractor. The feature extractor breaks the segmentation point matrix down into small windows of equal size and analyses the density of black and white pixels. Therefore, instead of presenting the raw pixel values of the segmentation points to the ANN, only the densities of each window are presented. As an example, if a window exists which is 5x5 in dimension, and contains 12 black pixels, then a single value of 0.48 (Number of black pixels/25) is written to the training file to represent the value of the window. An example of density feature extraction is shown in Figure 3. Accompanying each matrix the desired output is also stored in the training file (0.1 for an incorrect segmentation point and 0.9 for a correct point) ready for ANN training.

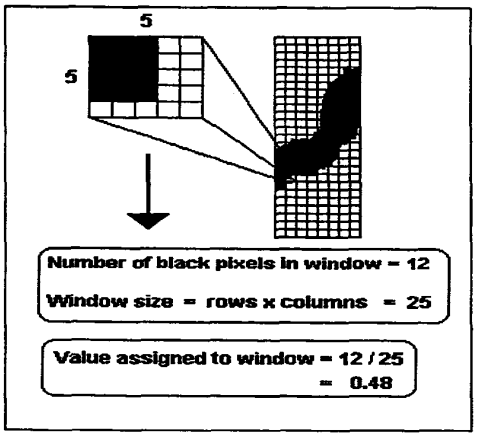


Figure 3. A window of 5x5 in dimension is extracted from a segmentation point matrix

3.8 Testing Phase of Segmentation Technique

Following ANN training, the words used for testing are also segmented using the heuristic feature-based algorithm. However, for testing there is no manual processing. The segmentation points are automatically extracted and are fed into the trained ANN. The ANN then verifies which segmentation points are correct and which are incorrect. Finally, upon ANN verification, each word used for testing should only be marked with valid segmentation points, which can then be used for further processing.

4. Experimental Results

For experimentation of the techniques detailed in Section 3, we used samples of handwritten words from the CEDAR benchmark database [19]. In particular we used samples from words contained in the "BD/cities" directory of the CD-ROM. Some examples of handwritten words used in the experiments are shown in Figure 4.

All segmentation experiments were conducted using an ANN trained with the backpropagation algorithm. Two major experiments were conducted. In the first experiment, segmentation point patterns for both training and testing were taken from words contained in the "BD/cities" directory, totaling 3620 and 385 respectively.

The second experiment was conducted using a larger sample of training and testing patterns: 8375 and 724 respectively. Table 1 shows the top experimental results of verified segmentation points for the smaller set of patterns, while Table 2 shows results for the larger set. Many experiments were performed varying settings such as the number of iterations, the number of hidden units, learning rate and momentum. For each experiment the number of inputs remained constant: a 14x3 matrix of pixel densities (42 inputs). These dimensions produced optimal results in preliminary tests. The number of outputs was always set to 1.

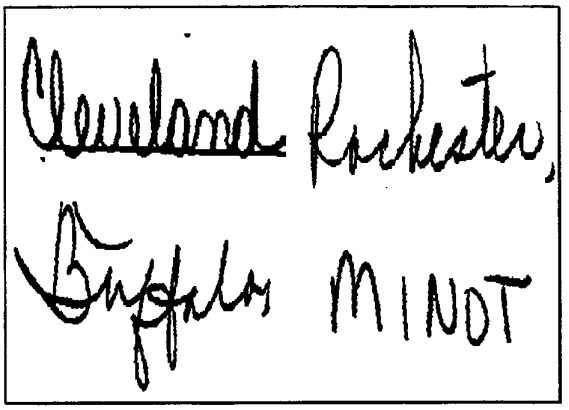


Figure 4. Handwriting samples used for training/testing

Table 1. Segmentation point results using 3620 training patterns

Iterations	Hidden Units	Learning Rate	Momentum	Classification Rate For Test Set	Classification Rate [%] Test Set
100	4	0.3	0.3	289/385	75.06
200	5	0.6	0.3	286/385	74.29
100	3	0.3	0.3	285/385	74.03

Table 2. Segmentation point results using 8375 training patterns

Iterations	Hidden Units	Learning Rate	Momentum	Classification Rate For Test Set	Classification Rate [%] Test Set
300	20	0.1	0.1	554/724	76.52
200	20	0.1	0.1	553/724	76.38
200	10	0.1	0.1	552/724	76.24

5. Discussion of Results

Many researchers mention segmentation as a part of their overall systems, however few report their findings at the segmentation level. This research has focussed on this very important area and has produced commendable results which can easily be compared to other researchers in the field. The neuro-heuristic algorithm obtained results of up to 76.52% for a test set of segmentation point patterns. Eastwood et al. [16] presented an ANN-based method for the segmentation of cursive and printed handwriting from the CEDAR CD-ROM, detailing a segmentation accuracy

of 75.9%. Han and Sethi [20], achieved an 85.7% accuracy using an heuristic algorithm for the segmentation of words on 50 envelopes from real mailpieces. Finally, Yanikoglu and Sandon [21] reported that 97% of letter boundaries from 750 words were correctly located. It must be noted that they did not use a benchmark database of real-world unconstrained words for their experiments. The results for segmentation achieved in this research compare favourably with other researchers.

6. Future Research

In future work, the segmentation technique will be improved in a number of ways. Firstly, the heuristic component of the segmentation system will need to be enhanced further. Originally, one of the main aims of the heuristic algorithm was to keep the number of incorrect segmentation points to a minimum, so that errors and processing time could be reduced. As a result, under-segmentation was noticeable in some words. Therefore, the algorithm shall be modified so that it will be possible to detect a smaller number of incorrect segmentation points, while at the same time recovering more correct segmentations. This can be achieved by looking for more features or possibly enhancing the current feature detection methods. In particular, a postprocesser shall be added to the segmentation phase. The postprocessor will be used to detect substantially difficult segmentation points usually found when either large uppercase characters cross over into regions occupied by lowercase characters or when two characters are tightly coupled.

In the neural component, a more robust, structural feature extraction technique shall be used to better exemplify information from segmentation zones in the handwritten words. More patterns shall also be used in training and testing, and finally the technique shall be integrated into a complete handwriting recognition system.

7. Conclusion

An intelligent segmentation technique has been presented in this paper, producing good results. It was used to segment difficult cursive and printed handwritten words from the CEDAR database. With some modifications, more testing shall be conducted to allow the technique to be used as part of a larger system. It has been noted that there are very few researchers that have published their segmentation results for handwritten word recognition when discussing a complete system. It is therefore hoped that further research can be dedicated to analysing and improving the results of this very important procedure.

References

[1] C. Y. Suen, and R. Legault, C. Nadal, M. Cheriet, and L. Lam, "Building a New Generation of Handwriting Recognition Systems", *Pattern Recognition Letters*, Vol. 14, 1993, pp. 305-315.

[2] S-W. Lee, "Multilayer Cluster Neural Network for Totally Unconstrained Handwritten Numeral Recognition", *Neural Networks*, Vol. 8, 1995, pp. 783-792.

[3] H. I. Avi-Itzhak, T. A. Diep, and H. Garland, "High Accuracy Optical Character Recognition using Neural Networks with Centroid Dithering", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 17, 1995, pp. 218-224.

[4] S-W. Lee, "Off-Line Recognition of Totally Unconstrained Handwritten Numerals Using Multilayer Cluster Neural Network", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 18, 1996, pp. 648-652.

[5] S-B. Cho, "Neural-Network Classifiers for Recognizing Totally Unconstrained Handwritten Numerals", *IEEE Trans. on Neural Networks*, Vol. 8, 1997, pp. 43-53.

[6] P. D. Gader, "Fusion of Handwritten Word Classifiers", *Pattern Recognition Letters*, Vol. 17, 1996, pp. 577-584.

[7] S. N. Srihari, "Recognition of Handwritten and Machine-printed Text for Postal Address Interpretation", *Pattern Recognition Letters*, Vol. 14, 1993, pp. 291-302.

[8] M. Gilloux, "Research into the New Generation of Character and Mailing Address Recognition Systems at the French Post Office Research Center", *Pattern Recognition Letters*, Vol. 14, 1993, pp. 267-276.

[9] R. G. Casey and E. Lecolinet, "A Survey of Methods and Strategies in Character Segmentation", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 18, 1996, pp. 690-706.

[10] R. M. Bozinovic, and S. N. Srihari, "Off-Line Cursive Script Word Recognition", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 11, 1989, pp. 68-83.

[11] N.W. Strathy, C.Y. Suen, and A. Krzyzak, "Segmentation of Handwritten Digits using Contour Features", *ICDAR '93*, 1993, pp. 577-580

[12] B. A. Yanikoglu, and P. A. Sandon, "Off-line cursive handwriting recognition using style parameters", *Tech. Report PCS-TR93-192*, Dartmouth College, NH., 1993.

[13] J-H. Chiang, "A Hybrid Neural Model in Handwritten Word Recognition", *Neural Networks*, Vol. 11, 1998, pp. 337-346.

[14] Gader, P., Whalen, M., Ganzberger, M., Hepp, D., "Handprinted Word Recognition on a NIST Data Set", Machine Vision Applications, Vol. 8, 1995, pp 31-40.

[15] G. L. Martin, M. Rashid, and J. A. Pittman, "Integrated Segmentation and Recognition through Exhaustive Scans or Learned Saccadic Jumps", *Int'l J. Pattern Recognition and Artificial Intelligence*, Vol. 7, 1993, pp. 831-847.

[16] B. Eastwood, A. Jennings, and A. Harvey, "A Feature Based Neural Network Segmenter for Handwritten Words", *Int'l Conf. Computational Intelligence and Multimedia Applications*, Gold Coast, Australia, 1997, pp. 286-290.

[17] Y. Lu, M. Shridhar, "Character Segmentation in Handwritten Words – An Overview", *Pattern Recognition*, Vol. 29, 1996, pp. 77-96.

[18] N. Otsu, "A threshold selection method from gray level histograms", *IEEE Trans. Systems, Man and Cybernetics*, Vol SMC-9, 1979, pp. 62-66.

[19] J. J. Hull, "A Database for Handwritten Text Recognition", *IEEE Transactions of Pattern Analysis and Machine Intelligence*, Vol. 16, 1994, pp. 550-554.

[20] K. Han, I. K. Sethi, "Off-line Cursive Handwriting Segmentation", *ICDAR '95, Montreal, Canada*, 1995, pp. 894-897.

[21] B. Yanikoglu, P. A. Sandon, "Segmentation of Off-line Cursive Handwriting using Linear Programming", *Pattern Recognition*, Vol. 31, 1998, pp. 1825-1833.