Non-Linear Segmentation of Touched Roman Characters Based on Genetic Algorithm

Tanzila Saba, Ghazali Sulong and Amjad Rehman

Graphics and Multimedia Department
Faculty of Computer Science and Information Systems
University Technology Malaysia

Abstract

The segmentation accuracy of Roman cursive characters, especially touched characters, is essential for the high performance of Optical Character Recognition Systems. This paper presents a new approach for non-linear segmentation of multiple touched Roman cursive characters based on genetic algorithm. Initially, a possible segmentation zone is detected and then best segmentation path is evolved by genetic algorithm. The initial population is composed of each point column in possible segmentation zone. The individual coding, fitness function, crossover operator and mutation operator are also defined for this task. Experimental results on a test set extracted on the IAM benchmark database exhibit high segmentation accuracy up to 89.76%. Proposed approach can handle some complex types of touched cursive characters without special heuristic rules and recognition.

1. Introduction

Invent of modern technologies has brought significant changes in cursive handwriting. Touched character segmentation is a current issue for optical character recognition (OCR) systems. Although literature is replete with many character segmentation techniques, however their feasibility is for machine printed, hand printed and well written cursive script only. All these techniques fail for touched cursive handwritten character segmentation [1]. In this regard a detailed review can be viewed in [2-5]. On the other hand, character segmentation has two main strategies: linear and non-linear character segmentations. In linear segmentation simple vertical segmentation is performed. Consequently, either characters are deprived from their discriminative parts or get extra. Thus, it increases misclassification rate at later stage [10]. The nonlinear segmentation has the edge of finding the segmentation path.

In this regards, few work is reported in the literature. Ventzislav [6] introduces critical concave points (CP) to determine cutting candidate points. Afterward, all possible cutting points are verified using robust classifier to determine the segmented characters correctly recognize or misclassified. Although, experiments are conducted on printed characters, yet, authors claim accuracy of their method for hand-printed. However, detection of all critical concave points and contour tracing is time consuming. In addition, concave up-turned and down-turned are only suitable for single touching. Hence, it fails for long touching where concaves don't exists and multiple touching that consist of holes rather concave.

A recognition-based segmentation with help of contour and projection analysis is performed in [7]. To determine type of input image: single or string numerals, single-numeral classifier is adopted from [8]. Further processing is conducted for string numerals consists of more than one numeral. Candidate segmentation points are generated from corner point of contour and horizontal projection. Finally, the optimal segmentation result can be obtained according to the maximum a posterior (MAP) criterion with an imbedded classifier. Recently, touched Chinese characters are segmented using background thinning and fuzzy rules [9]. Additionally, an enhancement for segmentation of two Roman touched cursive handwritten characters is proposed in [1]. According to the authors, touching pairs can be divided into three regions (i.e left, right and middle region) and these regions acquire unique characteristics. By using self organizing map (SOM) the touching parts are identified based on their characteristics. However, the technique just identifies touching points and performs linear segmentation. Hence the segmented characters are deprived from their discriminative parts or get additional parts that increases misclassification rate.

Accordingly, this paper presents a nonlinear segmentation algorithm for cursive touched characters based on genetic algorithm. The paper is further organized into three sections. Section II presents detailed picture of the proposed technique. Experimental results are reported in Section III and finally, conclusion is drawn in Section IV.

2. Proposed Touched Character Segmentation Approach

This section elaborates proposed strategy for touched character segmentation based on genetic algorithm. Accordingly, the methodology composed of preprocessing, possible segmentation zone detection and non-linear segmentation of touched characters based on genetic algorithm.

2.1 Preprocessing.

More than two touching handwritten characters are rare therefore; two touched characters are extracted from IAM forms scanned in grayscale format. However, prior to character segmentation, digital images are binarized using Otsu method [11]. Additionally, core-region is detected to avoid ascenders and descenders of the overlapped characters [12]. Figure 1 exhibits preprocessing results.

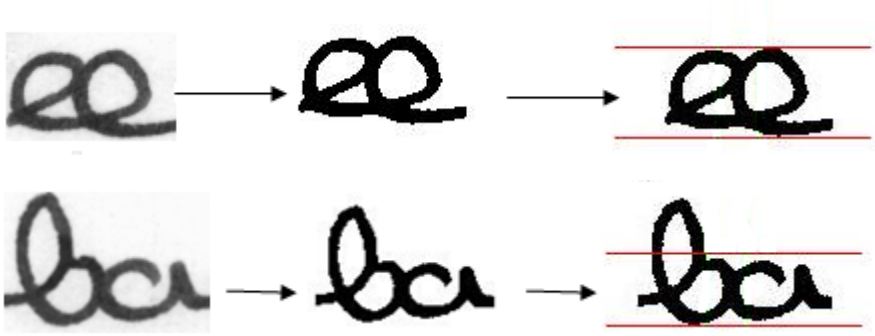

Figure 1: Preprocessing results

2.2 Possible Segmentation Zone Detection

Possible segmentation zone (PSZ) is defined as an area that occupies the segmentation path between the touching boundaries of the characters. In this research, possible segmentation zone is detected using vertical projection profile. Based on the experiments, the possible segmentation zone is evaluated to lie between two peaks in the vertical projection presented in Figure 2.

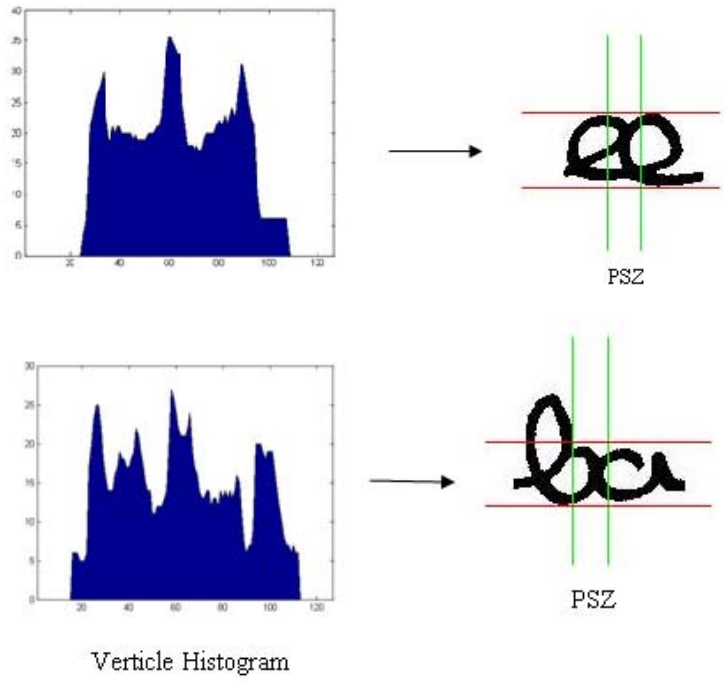

Figure 2: Possible segmentation zone detection

2.3 Non-linear Segmentation of Touched Characters Based on Genetic Algorithm

In binarized image, the possible segmentation zone (PSZ) is point array sorted by X co-ordinates. Therefore, every column of points in possible segmentation zone is regarded as segmentation path. However, this point array cannot perform non-linear segmentation. Therefore, genetic algorithm is employed to assists segmentation process.

Genetic algorithms are the most popular stochastic algorithms that use adaptive optimization criteria inspired by the genetic processes of biological organisms. This simulation has brought success for many real world problems particularly where heuristic solutions generally

lead to unsatisfactory results. Genetic algorithms use initial population, selection, crossover, mutation and fitness function [13]. In this research, binary representation of image is used for GA training such that each gene has the value of either 0 or 1. The initial population consists of all point columns in the possible segmentation zone. Following selection of points in each point column, crossover operator, mutation operator and fitness function are defined. Finally, genetic algorithm is applied to evaluate the initial population to find out non-linear segmentation path in the possible segmentation zone. Accordingly, several generations are performed to evaluate the individuals with maximum fitness value. In this research, 50 generations are found enough to find out individuals with maximum fitness value. Finally, the individuals with maximum fitness value are connected to develop non-linear segmentation path.

2.3.1 Initial Population

The GA starts by creating initial population that is composed of individuals [14]. The number of individuals is the population size. In this research, all point columns in the possible segmentation zone represent individuals. Therefore, initial population consists of all point columns in the possible segmentation zone. Hence the individual coding is represented as below.

$(p_1(x_1, y_1), p_2(x_2, y_2) \dots\dots\dots p_n(x_n, y_n))$---------- (i)

In equation (i), $p_i(x_i, y_i)$, $i = 1 \dots . n$, represents individual point in the possible segmentation zone. All points are sorted by X-coordinate in descending order and stored in an array. This sorted array serves as initial population for genetic algorithm (GA).

2.3.2. Selection Operator and Fitness Function

The key problems in the GA are the selection of fitness function and selection operator. The selection operator is used to select chromosomes called parents to generate new chromosomes called offspring. The selection of the fitness function is the critical problem, as it precisely quantifies the quality of the candidate solution [15]. Additionally, a fitness function is a designed function that enables the chromosomes to solve problem accurately. Therefore, inaccurate selection of the fitness function affects performance of the GA negatively [16]. The fitness function assigns an evaluation value to each individual in the population. Finally, based on these evaluation values, selection operator operates to select the best ones.

There are different selection approaches such as Roulette wheel selection, Rank selection, tournament selection, Gaussian selection and Elitism. In this research, Gaussian fitness selection is adopted as below.

$$P(x_n) = \sum_{m=1}^{M} \frac{c_m}{2\pi^{(D/2)}\sqrt{\prod_{d=1}^{D}\sigma_{m,d}}} \exp\left(-\sum_{d=1}^{D}\frac{(x_{n,d}-\mu_{m,d})^2}{2\sigma^2 m}\right)$$-- --------- (ii)

Here, D is the dimension of the feature vector and x_n represents the feature vector of each individual point. M represents Gaussian component size set to 8 experimentally. M $_{m,d}$ is the dth element of average vector μ_m. Similarly, $\sigma_{m,d}$ represents dth element of the standard deviation vector. c_m is the Gaussian component weight.

The feature vector is used to perform non-linear segmentation in the specific feature space. Finally, feature vector is taken as input vector in the fitness function to evaluate every path. In this research, six features are employed in the feature vector detailed below.

F1: The first feature represents ratio between the widths of the two segmented components.

F2: The second feature represents ratio between the heights of the two segmented components.

F3: The Third feature represents ratio between the widths to height for the two segmented components

F4: The fourth feature is the ratio between the counts of black pixels on the segmentation path and the width of the image.

F5: The fifth feature is the aspect ratio of the each segmented part.

F6: The sixth feature is X-coordinate covariance. It is used to compute the X-coordinate covariance of all points on the segmentation path.

Hence, the feature vector is composed of six features which are manipulated to get the fitness value by the selected fitness function.

2.3.3 The Crossover Operator

Crossover operator is a key operator in genetic algorithm. It occurs with user specified crossover probability p_c that ranges from 0.4 to 0.8 [17]. There are four types of crossover operators: single point crossover, two point crossover, uniform crossover and arithmetic crossover. In this research, two point crossover operator is used to construct segmentation path with $p_c = 0.4$ set experimentally. Accordingly, in the possible segmentation zone, two points crossover operator is implemented to select two positions randomly. Middle points of the parents are crossed to generate two new offspring. For example, consider two parents A and B with the following specifications.

Parent A= 11**0010**10

Parent B= **01**1100**11**

Offspring A= 11110010

Offspring B= **01001011**

2.3.4 The Mutation Operator

The mutation operator is applied after the crossover operator in order to search new area/path in the search space (possible segmentation zone). Additionally, it operates as background operator and adds diversity to the initial population of chromosome. Mutation probability is problem dependent and typically it is from 0.1 to 0.3[18]. In this research piece mutation is applied as the mutation operator. The length of the mutation operator is defined as below.

$$length_{mutation} = value\ (0.1{\sim}0.3) * w \text{----------(iii)}$$

In equation (iii), value is the mutation value selected randomly in the range between 0.1 to 0.3. w is the character width derived heuristically. Mutation operator points are selected from zero to w. Finally, X-coordinates are limited within the possible segmentation zone. The mutation operation is exhibited in the Figure 3.

Figure 3: The 8 bit mutation operator

3. Results and Discussion

Touch characters in cursive handwriting is a special case and therefore, there is no benchmark database for touched cursive script. However, to perform training and testing of GA, touched characters are identified and extracted from IAM benchmark database [19]. IAM benchmark database is publicly available and meanwhile has been used by several research groups [20]. Five hundred writers have contributed their handwriting in this database, consisting of 1500 pages of scanned text, 10,000 isolated and labeled text lines and 100,000 isolated and labeled cursive script words. All text is scanned in grayscale format at 300 dpi. For experimentation, 450 touched characters are extracted while the test set consists of 150 touched characters. Few extracted touched cursive characters are presented in Figure 4.

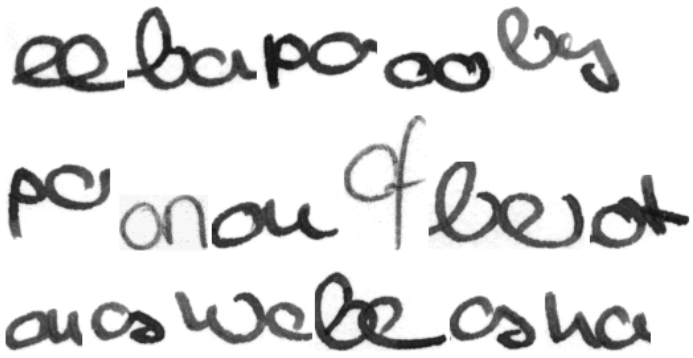

Figure 4: Cursive touched characters extracted from IAM benchmark database [19].

In order to avoid the possible variation occurred due to randomness of the genetic algorithm , its training and testing is repeated seven times [17]. Finally, average segmentation accuracy on training set and testing set is counted to evaluate the performance of our proposed segmentation algorithm. In our experiment, the number of evolutional generations is empirically set to 45, the crossover probability set to 0.3, and the mutation probability set to 0.05. The proposed approach attains touched character segmentation accuracy on test set up to 89.76%. Training and testing results are tabulated in Table 1.

Table 1: Touched character segmentation accuracy based on GA

	1	2	3	4	5	6	7	Mean
Train	92.75%	93.10%	92.83%	92.71%	92.49%	93.07%	93.12%	92.86%
Test	87.10%	88.52%	89.01%	90.11%	90.75%	91.03%	91.84%	89.76%

Relationship between number of generations and fitness function is demonstrated in Figure 5. It is evaluated that despite of an increase in the number of generations, fitness function is stable.

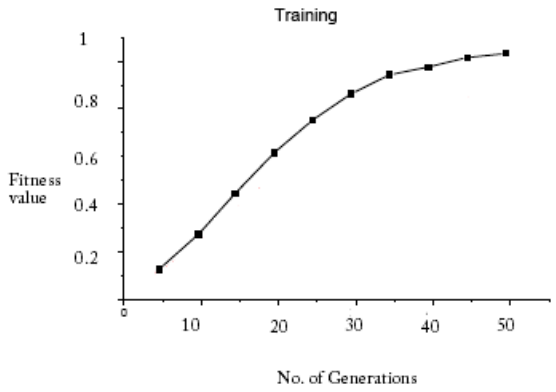

Figure 5: Relationship between number of generation and fitness value.

Finally, some segmentation results are presented in Figure 6. Accordingly, non-linear segmentation of multiple touching cursive characters is performed successfully. It is very hard to compare segmentation accuracy with others in the literature due to different dataset used, experimental setup and so on. However, for the sake of evolution of the proposed approach a few comparisons are detailed. These comparisons are selected as they are relevant and latest in the state of the arts.

Background and foreground analysis is performed for touched character segmentation [21]. Additionally, different heuristic rules are applied to character image of different touching types, which complicates the segmentation algorithm and cannot cover all touching situations. Additionally, all experiments are performed on touched numeric and 96% segmentation accuracy is claimed. Self organizing feature map is applied to identify touching position of two touched characters [1]. However, no segmentation path is evaluated. In the proposed approach, following fitness function evaluation, all touching types are handled using relatively simple algorithm without any heuristic rules. Moreover, proposed method can identify correct nonlinear segmentation path for some types of touching character that were not included in the training set. In few cases due to incorrect detection of possible segmentation zone, proposed algorithm could not identify the correct segmentation path. Additionally, its speed is slow. Few failure results are exhibited in Figure 7. However, the overall efficiency is good.

Figure 6: Touched character segmentation based on GA

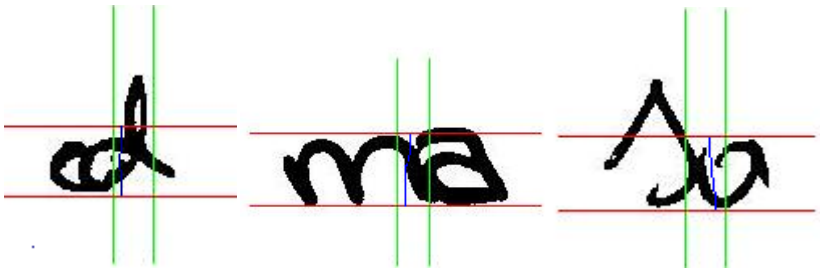

Figure 7: Inaccurate character segmentation

4. Conclusion

Correct segmentation of touched characters is mandatory for their successful recognition. Accordingly, this paper has presented a new algorithm for cursive handwritten touched character segmentation based on genetic algorithm. The algorithm handles successfully single and multi touching characters. Proposed algorithm searches within the search area only that makes its search narrow and fast. Finally, genetic algorithm produces non-linear segmentation path based on its training. Additionally, different parameters related with GA such as individual coding, initial population, crossover operator, mutation operator and fitness function are also explained. The promising experimental results are achieved up to 89.76% segmentation accuracy on test set without using special heuristic rules. This technique can be implemented for touched character segmentation of different languages provided that fitness function is trained on relevant dataset. However, due to the randomness in genetic algorithm, its speed is slow. Moreover, this algorithm can segment two touching characters only. In future research efforts will be carried out for more than two touched handwritten characters segmentation. The proposed approach will be optimized by proposing new parameters such as individual coding, fitness function etc and training as the future research. Finally, the proposed approach will be applied to different datasets to increase its validity.

References

[1] Fajri Kurniawan, Amjad Rehman, Dzulkifli Mohamed and Siti Mariyam (2010). Self Organizing Features Map with Improved Segmentation to Identify Touching of Adjacent Characters in Handwritten Words. IEEE Ninth International Conference on Hybrid Intelligent Systems, 2010. HIS 2010.China pp. 475-480.

[2] Y. Lu., “Machine Printed Character Segmentation: An Overview”, Pattern Recognition, 28, 1995, pp. 67-80.

[3] Y. Lu, M. Shridhar, “Character Segmentation in Handwritten Words-An Overview”, Pattern Recognition, 29, 1996, pp. 77-96.

[4] Richard G. Casey, Eric Lecolinet, “A Survey of Methods and Strategies in Character Segmentation”, IEEE Trans. PAMI, 18, 1996, pp. 690-706.

[5] G. Congedo, G. Dimauro, S. Impedovo, G. Pirlo, “Segmentation of Numeric Strings”, Proc. 3rd ICDAR, 1995, pp. 103801041.

[6] Ventzislav, A. (2004). Using critical points in contours for segmentation of touching characters. Proceedings of the 5th international conference on Computer systems and technologies. Rousse, Bulgaria, ACM.

[7] Yun, L., Liu, C.S., Ding, X.Q. and Qiang, F. (2004). A recognition based system for segmentation of touching handwritten numeral strings. Ninth International Workshop on Frontiers in Handwriting Recognition, 294-299.

[8] Zhang, J.Y. and Ding, X.Q. (2000). Multi-Scale Feature Extraction and Nested-Subset Classifier Design for High Accuracy Handwritten Character Recognition. Proc. 15th ICPR, Barcelona, Spain, 581-584.

[9] Shuyan Zhao, Zheru Chi, Penfeu Shi, Hong Yan, “Two stage Segmentation of Unconstrained Handwritten Chinese Characters”, Pattern Recognition, 2003, pp. 145-156.

[10] Jayarathna, U.K.S. Bandara, G.E.M.D.C. New Segmentation Algorithm for Offline Handwritten Connected Character Segmentation Proceedings of the First International Conference on Industrial and Information Systems, First International,2006: 540-546Jayarathna, U.K.S. Bandara, G.E.M.D.C. New Segmentation Algorithm for Offline Handwritten Connected Character Segmentation Proceedings of the First International Conference on Industrial and Information Systems, First International,2006: 540-546.

[11] Otsu, N. (1979). A Threshold Selection Method from Gray level Histograms, IEEE Transactions on Systems, Man and Cybernetics Vol. 9(1), 63-66.

[12] Amjad Rehman, Dzulkifli Mohammad, Ghazali Sulong and Tanzila Saba (2009). Simple and Effective Techniques for Core Zone Detection and Slant Correction in Script Recognition. The IEEE International Conference on Signal and Image Processing Applications (ICSIPA’09), pp. 15-20.

[13] Banzhaf, Wolfgang; Nordin, Peter; Keller, Robert; Francone, Frank (1998) Genetic Programming – An Introduction, Morgan Kaufmann, San Francisco, CA.

[14] Cha, Sung-Hyuk; Tappert, Charles C (2009). "A Genetic Algorithm for Constructing Compact Binary Decision Trees". Journal of Pattern Recognition Research 4 (1): 1–13.

[15] 1Omar Al Jadaan, 2Lakishmi Rajamani, 3C. R. RaoIMPROVED SELECTION OPERATOR FOR GAJournal of Theoretical and Applied Information Technology, 269-277

[16] R. K. Belew and M. D. Vose (1997). Foundations of Genetic Algorithms 4. San Francisco, CA: Morgan Kaufmann.

[17] M. Mitchell (1996). An Introduction to Genetic Algorithms. Cambridge, MA: MIT Press

[18] D. B. Fogel (1995). Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. IEEE Press.

[19] Marti, U., and Bunke, H. (2002). The IAM database: An English Sentence Database for Offline Handwriting Recognition. International Journal of Document Analysis and Recognition, Vol.15, 65-90.

[20] Vamvakas, G, Gatos, B., Pratikakis, I., Stamatopoulos, N., Roniotis, A., Perantonis, S.J. (2007). Hybrid Offline OCR for Isolated Handwritten Greek Characters. Proceedings of Fourth IASTED international conference on Signal Processing, Pattern Recognition and Applications, 197-202

[21] Yi-Kai Chen, Jhing-Fa Wang, “Segmentation of Single or Multiple-Touching Handwritten Numeral String Using Background and Foreground Analysis”, IEEE Trans. PAMI, 2000, pp. 1304-1317