Segmentation of Connected Chinese Characters Based on Genetic Algorithm^{*}

[†]Xianghui Wei, [‡]Shaoping Ma, [‡]Yijiang Jin

[†]*Institute of Software, CAS*

xianghuiwei2004@yahoo.com.cn

[‡]*State Key Lab of Intelligent Tech. & Sys., CST Dept, Tsinghua University,
Beijing 100084, P.R.China*

{msp, yjjin}@tsinghua.edu.cn

Abstract

The accuracy of segmenting Chinese character, especially connected Chinese characters, is essential for the performance of a Chinese character recognition system. In this paper, a new approach for segmenting connected Chinese characters based on genetic algorithm is proposed. The best segmentation path is evolved by genetic algorithm from a fixed area located in the middle of character image which is defined as Segmentation Path Zone (SPZ). The initial population is composed of each point line in SPZ. The individual coding, fitness function, crossover operator and mutation operator are also defined for this task. Experimental results on a dataset extracted from the Four Vaults show that our approach can get an average accuracy of 88.9% on test set and can handle some complex types of connected Chinese characters without special heuristic rules.

1. Introduction

Character segmentation is a main bottleneck in OCR system because most of them can only recognize well isolated character. Connected character is a more difficult situation for segmentation algorithms.

Many algorithms have been proposed to segment character. Lu [1] presents an overview of techniques in machine printed character segmentation. Lu and Shridhar [2] review hand-printed and handwritten character segmentation methods. Richard G. Casey and Eric Lecolinet [3] also give a survey of methods and strategies in character segmentation. G. Congedo et al. [4] introduces a digital character segmentation method

that simulates “drop-falling” process to segment digits. Zhongkang Lu et al. [6] utilizes feature points on background skeleton to construct segmentation path, then fuzzy rules generated from decision tree are used to rank all possible paths. Yi-Kai Chen et al. [7] combines background and foreground analysis to segment single- and multiple-touching handwritten numeral strings and then uses Mixture Gaussian Function to decide which one is the best among all possible segmentation paths. U. Pal [9] proposes a method for segmenting handwritten digital strings based on features obtained from the Water Reservoir.

Segmentation of connected Chinese characters is a more difficult task, in that a large number of Chinese characters (the primary set contains 3755 Chinese characters) and many different writing styles exist. Lin Yu Tseng [5] presents a method based on heuristic merging of stroke bounding boxes and dynamic programming. This method tends to fail when the characters are overlapped and depends on the effectiveness of stroke extracting algorithm. Shuyan Zhao [8] develops a two-stage algorithm to segment unconstrained Chinese characters using background-thinning and fuzzy rules.

In this paper, we propose an effective method for connected Chinese characters segmentation. Segmentation path is regarded as an array of points ordered by X-coordinate while their vertical expansion is limited in a fixed region located in the middle of character image. Genetic algorithm is applied to search this region to find the best segmentation path.

The detail of our method is described in Section 2. We report our experimental results in Section 3. Finally, we draw some conclusion in Section 4.

2. Segmentation of connected Chinese characters based on genetic algorithm

^{*} This work was supported by the National Natural Science Foundation of China (No.60223004, No.60303005) and the national “863” program (No.2001AA114082).

In digitalized image, segmentation path can be regarded as point array sorted by X-coordinate. It can be observed from data that the segmentation path hardly deviates from a fixed area in the middle of character image, which is defined as Segmentation Path Zone (SPZ) in our paper. Figure 1 shows an example. So the segmentation path can be defined as following

$$Path = \left\{ (x_i, y_i) \left| \begin{array}{l} 0 < x_i \leq ImageWidth, \\ x_{i+1} = x_i + 1, \\ Y_{upper-bound} < y_i \leq Y_{lower-bound} \end{array} \right. \right\}$$

Where

- $ImageWidth$ is the width of character image.
- $Y_{upper-bound}$, $Y_{lower-bound}$ are the upper and lower bound of SPZ defined as follow:

$$Y_{upper-bound} = Y_{center} - \frac{1}{5} \times ImageHeight$$

$$Y_{lower-bound} = Y_{center} + \frac{1}{5} \times ImageHeight$$

Where Y_{center} is the Y-coordinate of image center.

Figure 1. Segmentation Path Zone (SPZ).

As mentioned above, segmentation path is defined as a point array. So every row of points in SPZ can be regarded as a segmentation path. Normally, they are not the path to segment character image perfectly. To construct the best segmentation path, genetic algorithm is applied. Every point line in SPZ is defined as an individual, all of which compose the initial population. After individual coding, crossover operator, mutation operator and fitness function have been defined, genetic algorithm is applied to evolve the initial population. After specific generations (30 in our experiment), the individual with the highest fitness value in final population is regarded as the best segmentation path.

In following sections, we will discuss how to define some important components in genetic algorithm for its application in segmentation of connected Chinese characters, which include individual coding, initial

population, crossover operator, mutation operator and fitness function.

2.1 Individual coding

As mentioned above, a segmentation path is an array of points. So individual coding can be defined as following:

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

Where p_i represent a point on segmentation path, of which x_i , y_i are the X- and Y-coordinate. All points are stored in a array which is sorted by X-coordinate ascendingly.

2.2 Initial population

Each row of points in SPZ is defined as an individual and the initial population is composed of all point rows.

2.3 Fitness function

The fitness function of genetic algorithm gives every individual in population an evaluation value upon which the selection procedure is done. In our method, the Mixture Gaussian Function is adopted as the fitness function which used in [7]. The Mixture Gaussian Function is defined as following:

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

Where x_n is the feature vector for each individual, D is the dimension of feature vector, M is the size of Mixture Gaussian component, $\mu_{m,d}$ is the d th element of mean vector μ_m , $\sigma_{m,d}$ is the d th element of the standard deviation vector σ_m , c_m is the Mixture Gaussian component weight. In our experiment, $D = 8$ and $M = 12$.

μ_m , σ_m and c_m are parameters that must be determined from training set. The Modified K-means algorithm is adopted to get these parameters as described in [7].

The feature vector represents the segmentation path in specific feature space and then taken as input vector in the fitness function to evaluate every path. In our method, eight features are used to represent the segmentation path, six of which are adopted from [7]:

- G_1 : ratio between the heights of the two separated parts.

- G_2 : ratio between the widths of the two separated parts.
- G_4 , G_5 : ratio between the width to height for the two separated parts respectively.
- G_6 : ratio between the vertical length of any overlap of the two separated parts and the smaller of the heights of the two separated parts.
- G_8 : ratio between the counts of black pixels on the segmentation path and the width of the image.

A new feature, called the convex-hull ratio, is also introduced which firstly appeared in [10]. Every character image which contains characters connected vertically can be segmented into two parts: P_{upper} -the upper part and P_{lower} -the lower part. For every segmentation path, the convex-hull ratio can be defined as following:

$$ConvexHull = \frac{\min(\text{CONVEX}(P_{upper}), \text{CONVEX}(P_{lower}))}{\max(\text{CONVEX}(P_{upper}), \text{CONVEX}(P_{lower}))}$$

Where the function CONVEX calculates the number of pixels that fall into the convex hull built from the foreground points in this segmented part.

Another feature called the Y-coordinate covariance is also used, which calculates the Y-coordinate covariance of all points on the segmentation path.

All these eight features compose a feature vector, which is calculated by the Mixture Gaussian function to get the fitness value.

2.4 The Crossover operator

Crossover is the essential operator to construct new segmentation path. The Two-point crossover operator is adopted in our approach as shown in Figure 2.

Figure 2. The Two-point crossover operator.

2.5 The Mutation operator

A new mutation operator called Piece Mutation is introduced in our method, in which a piece of points rather than a single one is operated. The length of mutation section is defined as following:

$$Length_{mutate} = \text{Rnd}(0.1, 0.3) \times ImageWidth$$

Where $\text{Rnd}(0.1, 0.3)$ is a random number between 0.1 and 0.3, $ImageWidth$ is the width of the character image. The start point of this section is selected randomly between 0 and $ImageWidth$.

The Y-coordinate of the mutated section is limited in the scope of SPZ, that is to say the Y-coordinate of mutated section must be smaller than the lower-bound and larger than the upper-bound of SPZ. The mutation process is shown in Figure 3.

Figure 3. The Piece Mutation operator.

3. Experimental results and discussion

428 Chinese character images (each image includes two Chinese characters) that are cut from the Four Vaults are used as the data set in our experiments. The Four Vaults is a famous collection of Chinese ancient books. The characters in this collection are written by hand in the column mode. Thus the two Chinese characters in a cut image are connected in the vertical direction. In our experiments, 200 images are selected randomly for training the Mixture Gaussian function and other 228 images are used for testing.

The experiment of training and testing has been repeated five times to overcome the possible fluctuations that are caused by the randomness of genetic algorithm. The average segmentation accuracy on both the training set and testing set is counted to evaluate the performance of our proposed segmentation approach. In our experiment, the number of evolutionary generations is empirically set to 30, the crossover probability set to 0.7, and the mutation probability set to 0.02. Table 1 shows the results. Our approach achieves the accuracy of 88.9% on the testing set.

Table 1. Segmentation Accuracy

	1	2	3	4	5	Ave
Train	93.5%	91.5%	92.0%	92.0%	92.5%	92.3%
Test	88.6%	88.6%	89.0%	89.0%	89.5%	88.9%

Figure 4 shows the relationship between the evolution generation and the mean and maximum fitness value in training. It indicates that the fitness values tend to be steady as the number of generation increases. And Figure 5 shows an example of evolving procedure of the best segmentation path. It is observed that the segmentation path gets better and better in our defined genetic algorithm. The path of the No.10 generation never goes through either character while

that of No.5 generation does. And further the path of the No.15 generation is close to the ideal.

Figure 4. Relationship between the average/max fitness value and the number of generation.

Figure 6 shows some segmentation results of our algorithm. As shown in the figure, single- and multiple- touching Chinese characters are correctly segmented. Single-touching with overlap is also segmented into correct Chinese characters. In other methods, such as background and foreground analysis in [7], different heuristic rules are applied to character image of different touching types, which complicate the segmentation algorithm and can not cover all touching situations. In our approach, all touching types can be handled in a relatively simple algorithm after fitness function has been determined. No special heuristic rules are required for constructing segmentation path under different touching situations. Even for some types of connected character that do not exist in training set, our method can also get correct results. For other datasets which are composed of different language character or digit, our method can easily transfer to it after the specific fitness function on this dataset has been trained.

Figure 5. Example of evolving procedure of the best segmentation path.

(a)

(b)

(c)

Figure 6. Examples of separated connected Chinese characters. (a) single-touching. (b) single-touching with overlap. (c) multiple-touching.

4. Conclusion

In this paper, a new method for segmenting connected Chinese characters based on genetic algorithm is proposed. The individual coding, initial population, crossover operator, mutation operator and fitness function are also defined respectively. The experimental result shows our method can handle

many complex types of connected character and achieves a segmentation accuracy of 88.9% in test set without using special heuristic rules. If the fitness function is trained based on other dataset which contain different language character or digit, our method can be easily applied to it.

From another point of view, our method can be looked upon as a searching process applied in a state space which is defined as the Segmentation Path Zone in our method. The genetic algorithm searches this space and produces the final optimized path. Thus, other searching algorithms can also be introduced to construct segmentation path that gives a new idea in this filed.

Because of the randomness in genetic algorithm, our algorithm's speed and efficiency is lower than expectation. In future research efforts, we will try to improve our algorithm by optimizing its parameters and introducing other methods, such as in individual coding. We will also apply our algorithm in various dataset to expand its application.

References

- [1] Y. Lu., "Machine Printed Character Segmentation: An Overview", *Pattern Recognition*, 28, 1995, pp. 67-80.
- [2] Y. Lu, M. Shridhar, "Character Segmentation in Handwritten Words-An Overview", *Pattern Recognition*, 29, 1996, pp. 77-96.
- [3] Richard G. Casey, Eric Lecolinet, "A Survey of Methods and Strategies in Character Segmentation", *IEEE Trans. PAMI*, 18, 1996, pp. 690-706.
- [4] G. Congedo, G. Dimauro, S. Impedovo, G. Pirlo, "Segmentation of Numeric Strings", *Proc. 3rd ICDAR*, 1995, pp. 103801041.
- [5] Lin Yu Tseng, Rung Ching Chen, "Segmenting Handwritten Chinese Characters Based on Heuristic Merging of Stroke Bounding Boxes and Dynamic Programming", *Pattern Recognition Letters*, 1998, pp. 963-973.
- [6] Zhongkang Lu, Zheru Chi, Wan-Chi Siu, Pengfei Shi, "A Background-thinning-based Approach for Separating and Recognizing Connected Handwritten Digit Strings", *Pattern Recognition*, 1999, pp. 921-933.
- [7] 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.
- [8] Shuyan Zhao, Zheru Chi, Penfeu Shi, Hong Yan, "Two-stage Segmentation of Unconstrained Handwritten Chinese Characters", *Pattern Recognition*, 2003, pp. 145-156.
- [9] U. Pal, A. Belad, Ch. Choisy, "Touching Numeral Segmentation Using Water Reservoir Concept", *Pattern Recognition Letters*, 2003, pp. 261-272.
- [10] Xianghui Wei, Shaoping Ma, "Segmentation of touching Chinese character based on convex hull ratio feature", *Journal of Chinese Information Processing*, 2005, pp. 91-96.