An Adaptive Approach to Extract Characters from Digital Ink Text in Chinese Based on Extracted Errors

Hao Bai(✉)

Beijing Language and Culture University, Beijing, China
baihao@blcu.edu.cn

Abstract. Extracting characters from digital ink text is an essential step which leads to more reliable recognition of text and also a prerequisite for structured editing. Casualness and diversity of handwriting input result in unsatisfied accuracy of extracted characters. Reprocessing the initial extracted characters based on context makes some considerable improvement. Therefore, this paper proposes an approach to adaptively extracting characters from digital ink text in Chinese based on extracted errors. The approach firstly classified the extracted errors in the primary extraction. According to different types of extracted errors, the approach gives different operations. Experimental data shows that the approach is effective.

Keywords: Digital ink · Character extraction · Error classification

1 Introduction

Comparing to keyboard, handwriting meets the needs of input conventionally and ergonomically, which provides the naturalness of writing and the rich expression of ink. With the development of such devices like digital pen, tablet PC and different pads, vast digital ink texts are accumulated and needed to be recognized. Extracting characters from text leads to more reliable recognition and availability of structured editing.

Self-adaptability of existing character extraction method is unsatisfactory in the results of processing. According to detailed analysis of the errors, many errors can be avoided in secondary extraction by rule-making. As a result, this paper proposes a method of self-adaptively extracting character from Chinese digital ink text based on error classification, firstly classifying and then processing the primary wrong extraction results with corresponding methods.

This paper is supported by Science Foundation of Beijing Language and Culture University(supported by "the Fundamental Research Funds for the Central Universities") (no.14YJ160202), and the National Natural Science Foundation of China (no. 61202249).

J. Li et al. (Eds.): NLPCC 2015, LNAI 9362, pp. 171–181, 2015.
DOI: 10.1007/978-3-319-25207-0_15

2 Related Works

In Chinese digital ink text, characters occupy a large proportion, including word, punctuation, number, letter, English word, etc. Based on the information used, character extraction methods can be divided into four types:

1. Method based on time and space distance of adjacent strokes

Subrahmonia et al [1] extracted words from English digital ink text based on time and spatial threshold, but did not specify the method for threshold determination; in addition, threshold ranges of different digital ink texts differ greatly. Zhang Xiwen et al [2] extracted characters from Chinese digital ink texts based on multi-level information. Zhang Shilong et al [3] extracted Chinese characters based on classification of stroke spacing. Han Yong et al [4] extracted characters according to minimum spanning tree (MST) of stroke, structural arrangement of Chinese characters and spatial arrangement of strokes. Su Rui et al [5] extracted characters based on histogram projection.

2. Method based on shape and structure

Tseng et al [6] calculated candidate character spacing with the minimum bounding box, firstly merged strokes according to structure of Chinese characters and finally used dynamic programming method to further merge candidate characters. Zhao Yuming et al [7] also calculated candidate character spacing with the minimum bounding rectangle, and gradually merged strokes, to extract separate Chinese character. Shilman et al [8] combined strokes with similar size and direction, and extracted words from English digital ink text following a bottom-to-top direction.

3. Method based on probability

Artieres [9] extracted words based on the probabilistic features grammars with genetic algorithm, considering the context information. Blanchard et al [10] put forward a method based on probability for English digital ink text, which takes the context and probabilistic features grammars into account.

4. Method based on the results of recognition

Subrahmonia et al [1] extracted letters from English digital ink text, considering all segmentation hypotheses, and took segmentations achieving the best effect as the satisfactory result. Chen et al. [11] extracted characters according to spacing of candidate characters, built grid based on the results of recognition, studied the best path according to scores of recognition and language model, to get results of Chinese characters extraction.

The above methods mostly extract words relying on information of the text. Due to the complexity of structure of Chinese characters and randomness of Chinese handwriting, the methods stated above have unsatisfactory adaptability for different languages and poorly structured data and need to integrate multi-level information besides the data, such as the information of primary extraction, etc. Therefore, this

paper presents a re-extraction method, which raises extraction accuracy on the basis of existing technical level.

3 Classification of the Extraction Errors

Figure 1 shows the results after extracting character from a digital ink text, and the circular circles mark the wrong characters extracted.

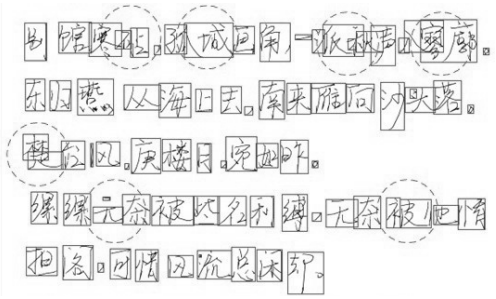

Fig. 1. Results after character extraction

Errors existing in character extraction can be divided into three categories: deficient extraction, beyond extraction and false extraction [12].

3.1 Deficient Extraction

Deficient extraction refers to the situation in which adjacent characters are deemed as one single character. The common errors of deficient extraction after primary extraction contain two types:

1. Adjacent characters are deemed as one single character due to the careless writing, undersize space between characters or low aspect ratio of rectangular bounding box. As shown in Fig. 2, the characters are extracted as one character because of the small bounding box and undersize space between characters.
2. Exaction error as a result of mixing punctuation with characters, as shown in Fig. 3, the comma after the character is taken as a stroke of the character.

Fig. 2. Deficient extraction of type 1

Fig. 3. Deficient extraction of type 2

3.2 Beyond Extraction

Beyond extraction refers to the situation in which a component or partial strokes are taken as another character, this kind of error covers two types after primary extraction:

1. A character in up-down structure is divided into two characters as shown in Fig. 4.
2. A character in left-right structure, as a result of the oversize spacing between strokes, is divided into two characters, as shown in Fig. 5.

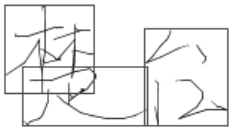

Fig. 4. Beyond extraction of type 1

Fig. 5. Beyond extraction of type 2

3.3 False Extraction

The reasons causing false extraction errors after primary extraction include: undersize spacing between adjacent characters or oversize spacing between the left and the right components, such as “被” and “他” as shown in Fig. 6, the left component of “被” is taken as a character while the right component “皮” and the left component of its adjacent “他” are extracted as the second character while the right component “也” is taken as the third character.

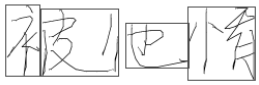

Fig. 6. False extraction

3.4 Classifying Rules

In view of structure of Chinese characters [13], even handwritten Chinese characters have some common features [14], such as aspect ratio of rectangular volume as the bounding box of a character, number of strokes of a character, and bounding box width of a character, etc.

1. Aspect ratio of rectangular bounding volume Rw/h=W/H, through analyzing 10 Chinese digital ink texts with an average of 300 characters, it was found that average aspect ratio of common Chinese characters ranges between 0.9187292 and 1.036495, thus the range of aspect ratio can be used to preliminarily judge whether a character extraction is correct or not, and the secondary processing will be conducted for false extraction.

2. Number of strokes (Ns): stroke of commonly used Chinese characters is averaged by 9.17 [14], and deficient extraction may occur if strokes of character extracted exceeds the average.
3. Width (W) of the bounding box: different from western words, Chinese characters are "square" shapes, and their width and line height (except for punctuations) show linear change, thus false extraction may occur if W exceeds the normal range of line height.
4. When one single extraction hypotheses is recognized as more than one characters, deficient extraction may exist in word segmentation.

Specifically, step 1, to recognize extraction hypotheses, if more than one character is recognized, deficient extraction occurs. Step 2, check Rw/h of rectangular volume and W of rectangular volume, if they exceed the normal range, false extraction may occur; Step 3, check the number of strokes (Ns) of single character, if Ns exceeds the normal range, deficient extraction may occur.

Based on analyzing extraction errors of the above types, it was found that secondary extraction of deficient extraction errors may result in beyond extraction errors; therefore, with the error classification-based method, the deficient errors should be firstly processed and then the beyond extraction errors.

4 Method for Deficient Extraction Errors

Unlike printed words, digital ink text always has poorly structures varying in size, spacing and internal distance, thus adjacent characters may easily be extracted as one due to the small spacing or undersize aspect ratio of bounding box. Punctuation marks may also be taken as a stroke of a character. In testing 10 Chinese digital ink texts with an average of 300 characters, the accuracy in primary extraction only reached an average of 80%. It was noticed that though handwriting is relatively arbitrary, basic positional arrangement of interior strokes is fixed; therefore, this paper suggests taking relative position of strokes of a character as the feature and conducting clustering analysis to process deficient errors, specifically,

1. Calculating central point of the strokes and take its horizontal projection value as value of the sample point;
2. Processing the data object with condensed hierarchical clustering algorithm [15], to obtain the number of clusters K;
3. Conducting clustering analysis against central points with K-means clustering algorithm [16] and the number of clusters K, to obtain each cluster as extraction.

4.1 Eigenvectors Extraction

Considering the left-right structure of character writing, obvious spacing generally exists between characters, which makes extraction of horizontal projection value from central points of strokes feasible. Taking the deficient extraction as shown in Fig. 7, projection values of abscissa of stroke centers are shown in dashed lines, specifically,

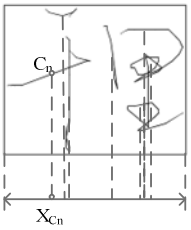

Fig. 7. Projection diagram of stroke center

1. Calculating central point C_n of the bounding boxes;
2. Calculating horizontal projection value of the central point X_{Cn};
3. Get date set $Cluster_0$ as the objects of X_{Cn}.

4.2 Hierarchical Agglomeration Clustering Algorithm

In this paper, K-means algorithm is used in clustering analysis; however, as the number of clusters K is required to be given in advance when this algorithm is used and inappropriate K can produce non-ideal clustering results, thus K should be knownbefore using K-means clustering algorithm. This research adopts hierarchical agglomeration to determine the number of clusters, obtain an initial clustering and improve it

Table 1. Similarities among Clusters (top 40 characters)

No.	Characters	Similarities	No.	Characters	Similarities	No.	Characters	Similarities
1	0	0	15	正	0.12	29	江	0.2
2	4	0.07	16	致	0.32	30	似	0.41
3	可	0	17	国	0.22	31	冻	0.33
4	桂	0.16	18	港	0.3	32	,	0
5	枝	0.31	19	秋	0.37	33	革	0.15
6	香	0.15	20	了	0	34	绎	0.3
7	王	0.01	21	天	0.31	35	如	0.2
8	安	0.12	22	气	0.19	36	预	0.39
9	石	0.29	23	初	0.31	37	0	0
10	登	0.24	24	弟	0.29	38	归	0.35
11	临	0.25	25	0	0	39	虹	0.34
12	送	0.23	26	于	0.13	40	支,	0.25
13	目	0.17	27	里	0.38			
14	,	0	28	澄	0.28			

with K-means. Defining the similarity of clusters as a termination condition is the key to the algorithm. This research takes internal and external distance ratio to represent similarity of clusters, the higher the value, the higher the degree of polymerization of data objects in the clusters. Table 1 shows the top 40 characters out of 134 from a Chinese digital ink text and whose extracted errors was manually corrected. According to the results of testing text data of 10 Chinese ink texts averaged by 300 characters, if the similarity was below 0.5, the target cluster was obtained. Internal distance of a cluster adopts average spacing of projection values (X_{Cn}) of stroke center in a cluster, while cluster spacing adopts spacing of projection values of stroke center between clusters.

4.3 K-means Clustering Algorithm

With the value of K obtained in hierarchical agglomeration algorithm and the initial data set $Cluster_0$ in calculating the eigenvector, K-means algorithm can be used to iterate and improve $Cluster_0$, in order to obtain the optimal results of extraction, as shown in Fig. 8. Fig. 9 shows comparison of a sample before and after application of deficient extraction algorithm, and dotted line marks the characters with varying results in extraction.

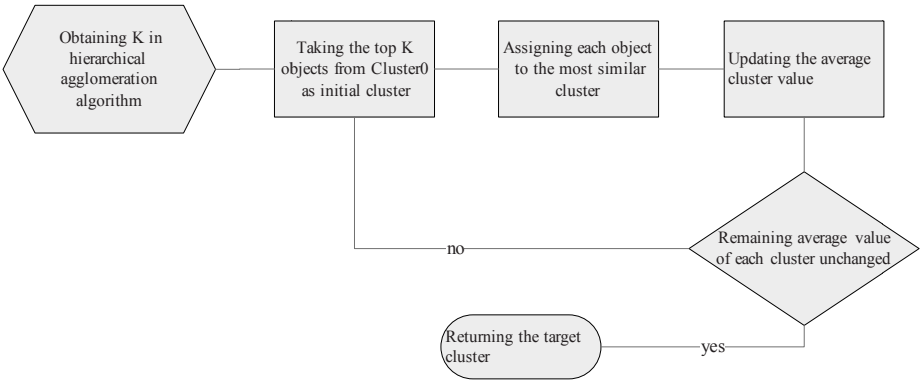

Fig. 8. K-means algorithm flow chart

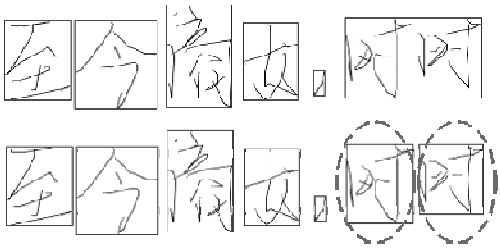

Fig. 9. Results comparison

5 Method for Beyond Extraction

According to the habit of Chinese handwriting, centers of characters in a line form a relatively stable straight line, as shown in Fig. 10. For two adjacent characters in a row, their horizontal angle of the center point connection is small (excluding punctuations). Table 2 lists horizontal angles of the center point connection of adjacent characters in a 134-character text extraction of which is corrected. Considering the structure of characters, the algorithm can be proceeded by,

1. Calculating center of each character;
2. Calculating horizontal angle of the center point connection;
3. If the angle exceeds the threshold (above 30°), take them as a character.

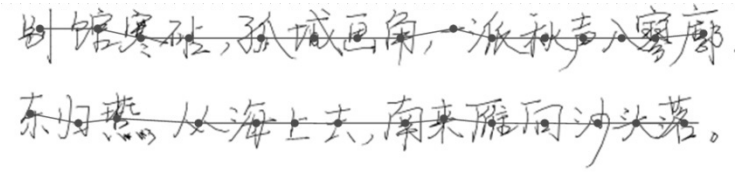

Fig. 10. Center line of characters

Table 2. Horizontal angles

No.	Characters	Angles (°)	No.	Characters	Angles (°)	No.	Characters	Angles (°)
1	0	12.25	15	正	3.02	29	江	1.58
2	4	5.1	16	致	5.14	30	似	6.62
3	可	3.13	17	国	0.75	31	冻	13.45
4	桂	3.32	18	港	3.97	32	,	0
5	枝	1.5	19	秋	28.51	33	革	0
6	香	0	20	了	22.27	34	绎	9.65
7	王	3.31	21	天	2.93	35	如	0.94
8	安	6.52	22	气	0.87	36	预	21.44
9	石	0	23	初	12.08	37	0	21.86
10	登	1.84	24	弟	12.78	38	归	0.79
11	临	2.51	25	0	0	39	虹	1.68
12	送	8.73	26	于	1.02	40	支,	12.78
13	目	20.58	27	里	0.76			
14	,	14.5	28	澄	3.49			

Through testing 10 Chinese digital ink texts averaged by 300 characters, it was found that apart from left-right structure, this algorithm effectively avoided extraction error for characters with up-down structure and half-investing structure. Punctuation

identification (period, comma, etc.) is the key point of the algorithm, to avoid taking punctuation as beyond extraction. In this research, structural feature of punctuation and its positional feature in the text were combined: firstly, number of strokes of punctuation is less than 2; secondly, width and height of punctuation are far smaller than the height of the line. Fig. 11 shows the comparison of results before and after using the beyond extraction respectively, and the dotted lines mark the characters with varying results of extraction.

Fig. 11. Examples of beyond extraction

6 Performance Test

Based on the proposed approach, in this research, a prototype system was developed using C# programming language and development platform of Microsoft visual studio 2005. This system operates on PC with Windows XP SP3. In the following part, quantitative analysis of results of a large number of Chinese digital ink texts is conducted to determine performance of the method proposed in this paper. Writings of six undergraduates, collected with digital pen produced by Swedish Anoto [17], were taken as Chinese digital ink test data of this research, and the ink data was extracted and rendered by prototype system developed with MS Tablet PC SDK [18]. Based on the experimental data, three performance indicators are put forward:

1. Extraction efficiency: the ratio between total time consumption and the total number of characters;
2. Initial accuracy: the ratio between the initial number of extraction hypotheses and the total number of characters;
3. Classification accuracy: the ratio between the number of extraction hypotheses and the total number of characters;

Partial statistical results of the algorithm proposed in this paper are shown in Table 3, and the experimental data shows that (1) the method for initial results of extraction could effectively improve accuracy; (2) for data with lower initial extracting accuracy, the method reached obviously higher accuracy; while for data with higher accuracy, the method is less effective; (3) as the method is a secondary extraction based on primary extraction, with the increase in number of text characters, the time consumption will be multiplied.

Table 3. Extraction algorithms based on error classification

Data sample	001	007
Total number of characters	114	309
Extracting time (sec)	12.81	48.8
Extraction efficiency (sec/words)	0.11	0.16
Primary accuracy	78.95%	93.20%
Accuracy achieved in this research	87.72%	94.50%

7 Conclusions

Due to the subjectivity and differentiation of handwriting text, single-pass character extraction always fails to reach satisfactory accuracy; through reprocessing the results of single-pass extraction, considerable improvement can be made. For this reason, this paper proposed an approach of extracting character from Chinese digital ink text based on error classification: firstly classify different types of errors and then processing the primary wrong extraction results with corresponding methods. As the method is a secondary extraction based on primary extraction, with the increase in number of text characters, the time consumption will be multiplied. However, considering the short time of extraction and the improvement of hardware performance, the extraction accuracy for texts in different length can be maintained in an acceptable range, thus the method proposed in this paper is effective in improving accuracy of single-pass extraction.

Reference

1. Subrahmonia, J., Zimmerman, T.: Pen computing: challenges and applications, vol. 2, pp. 60–66 (2000)
2. Xiwen, Z., Xiujuan, G., Guozhong, D.: Adaptive Character Extraction from Continuous Handwriting Chinese Text. Computing Technology and Automatics **3**, 73–77 (2003)
3. Shilong, Z., Xiwen, Z.: Adaptive character extraction from continuous handwriting Chinese textbased on classifying between - stroke gaps. Information Technology **8**, 80–82 (2005)
4. Yong, H., De, X., Guo-Zhong, D.: Using MST in Handwritten Chinese Characters Segmentation. Journal of Software **3**, 403–409 (2006)
5. Rui, S., Xi-wen, Z., Yong-quan, L., Guo-zhong, D.: Intelligent Editing of Handwriting Based on Structure Understanding. Journal of System Simulation **z1**, 371–373 (2006)
6. Yu, L., Rung, T., Chen, C.: Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming. Pattern Recognition Letters **8**, 963–973 (1998)
7. Yuming, Z., Xingzhi, J., Pengfei, S.: Algorithm for off-line handwritten Chinese character segmentation based on extracting and knowledge-based merging of stroke bounding boxes. Infrared and Laser Engineering **1**, 23–27 (2002)
8. Shilman, M., Wei, Z., Raghupathy, S., Simard, P., Jones, D.: Discerning structure from freeform handwritten notes, pp. 60–65 (2003)

9. Artières, T.: Poorly structured handwritten documents segmentation using continuous probabilistic feature grammars, pp. 5–8 (2003)
10. Blanchard, J., Artieres, T.: On-line handwritten documents segmentation.: frontiers in handwriting recognition. In: Ninth International Workshop on IWFHR-9, pp. 148–153 (2004)
11. Chen, H., Loudon, G., Yimin, W., Zitserman, R.: Segmentation and recognition of continuous handwriting Chinese text. International Journal of Pattern Recognition and Artificial Intelligence **1998**, 223–232 (1998)
12. Kun, Z., Xi-wen, Z.: Comparison of features and classifiers for detailedly classifying handwriting characters in Chinese ink text. Application Research of Computer, pp. 3486–3489 (2008)
13. Donghan, L.: Character Structure and Evolution. Shanghai Education Publishing House, pp. 73–75(1959)
14. Lian, Z., Chenxiao, W., Jicang, H.E., et al.: The effects of font, stroke and contrast on the reading speed of Chinese characters. Chinese Journal of Optometry & Ophthalmology, pp. 96–99 (2008)
15. Dunham, M.H.: Data Mining Introductory and Advanced Topics. Prentice Hall, pp. 112–115 (2003)
16. Han, J., Kamber, M.: Data Ming: Concepts and Techniques. Elsevier Inc., pp. 263–269 (2006)
17. Anoto Inc. http://www.anoto.com
18. Microsoft Windows XP Tablet PC Edition Software Development Kit 1.7. http://www.microsoft.com/downloads/details.aspx?familyid=b46d4b83-a821-40bc-aa85-c9ee3d6e9699&displaylang=en