http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (24): 8323-8327, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.8323.8327

Manchu Text Extract Based on Fuzzy Clustering

[1]Xu Shuang, [1]Li Min and [2]Zhu Man-Qiong
[1]Information and Communication Engineering, Dalian Nationalities University, Dalian Liaoning, 116600, China
[2]College of science, Beifang University of Nationalities, Yinchuan Ningxia, 750021, China

Abstract: Based on fuzzy clustering under the image background, this study proposes a new stroke extraction method of Manchu. Firstly, the digital image processing method is used to extract the Manchu word from an image background picture, do the de-noising, thinning, pruning and other pre-processed for fuzzy clustering. Then the mid-axis of Manchu character, termination point, interior points, intersection points is defined and finds the boundary point on mid-axis, use the stroke growth algorithm to separate the stroke so as to achieve the purpose of Manchu text stroke elaborates. To verify the feasibility of the method, this study carried out a text extraction on handwritten Manchu and printed Manchu. Simulation results show that the method is effective in separating the Manchu stroke and laid a solid foundation to the further study of Manchu identification.

Key words: Fuzzy clustering, manchu, stroke extraction, stroke growth

INTRODUCTION

Manchu as ever of the ruling class, a large number of political, cultural, economic, military, diplomatic, astronomy and other aspects of the data are recorded by the Manchu, it has a very high historical value, if it disappeared, these materials also loses its value. Zhang *et al.* (2007) Now the country have very few people can speak Manchu, Expert in the language of the people are much rarer, therefore research on Manchu character recognition system is very important for the protection of cultural heritage in the Qing Dynasty. Meanwhile, it is also a great contribution to the scanning to identify other Altai based languages, especially on the recognition of Mongolian and Sibo text. While the extraction of Manchu words is the key step of Manchu character recognition system, so do the extraction of Manchu languages is particularlyimportant (Chen, 2008; Wang and Zuo, 2011).

With the rapid development of computer technology, multimedia technology and communication technology, the multimedia information based on image, audio andvideo is rapidly becoming the mainstream on the exchange of information and services (Peng *et al.*, 2010). And the text in the image also reflects the important parts of the image. It is crucial to extract the text in the image accurately if we want to identify the image correctly (Hinchcliff *et al.*, 2012). And correct extraction of the Manchu language is an important part to improve the recognition rate, this study mainly studies the Manchu character extraction method of stroke growth based on Fuzzy clustering method under the image background.

FUZZY CLUSTERING ALGORITHM

Maximum membership degree principle: Let X be the recognition elements, $X = \tilde{A}_i \in F(X)(i = 1, 2, \cdots, n)$ is of n fuzzy pattern, $\tilde{A}_i \in F(X)(i = 1, 2, \cdots, n)$ is its corresponding membership functions. For any element x in X, we can determine which model it belongs according to the following principles:

$$\mu_{\tilde{A}_1}(x) = \max\left\{\mu_{\tilde{A}_1}(x), \mu_{\tilde{A}_2}(x), \cdots, \mu_{\tilde{A}_n}(x)\right\} \tag{1}$$

x is considered attributable to the kind represented by $\tilde{A}_1$. It is called the principle of maximum degree of membership. It represents the difference between the within class is smaller than the inter class (Abdullah *et al.*, 2013).

FCM Algorithms: We see the image pixel as the sample points of data set see the pixel of the feature (for grayscale image, namely gray) as sample points of the feature. According to the principle of maximum membership degree the target area will split from the image, first, the membership matrix and cluster centers of the data set is needed, find the maximum membership value of each class in the membership matrix; then extract the pixel which value equals to the maximum membership value, that is the classification of what we want; finally analysis the feature of pixels value so as to extract the target region. The step of determine the fuzzy clustering center; membership degree matrix with the method of FCM is as follows:

Corresponding Author: Li Min, School of Dalian Nationality University, Dalian, Liaoning, 116600, China

FCM put a data set X = {x1, x2, ..., xn} contains vector n into C categories, among them xj = (x1j, x2j, ..., xsj). Data set of n vectors belongs to C categories of membership can be classified as the following membership function matrix:

$$U = \begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ & & \cdots & \\ \mu_{c1} & \mu_{c2} & \cdots & \mu_{cn} \end{bmatrix} \tag{2}$$

μij, i = 1, 2, ..., c; j = 1, 2, ..., n, shows of the membership the j-th vector xj belongs to the i-th class. It is should meet:

$$\sum_{i=1}^{c} \mu_{ij} = 1, j = 1, 2, \cdots, n \tag{3}$$

The definition of fuzzy clustering center is V = {v1, v2, ..., vc}, where vi represents the i-th cluster center of fuzzy set μi, i = 1, 2, .., c; the distance between the j-th input vector xj and the i-th fuzzy cluster center vi is defined as:

$$d_{ij} = \left\| v_i - x_j \right\| = \sqrt{\sum_{l=1}^{s} \left(v_{il} - x_{jl}\right)^2}, i = 1, 2, \cdots, c; j = 1, 2, \cdots n$$

$$d_{ij} \ge 0, i = 1, 2, \cdots, c; j = 1, 2, \cdots, n \tag{4}$$

Set:

$$J(U, V) = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}{}^{m} d_{ij}{}^{2}$$

among them, μij and dij is obtained by using the above (3) and (4) $m \in [1, \infty)$ is a weighting coefficient. The purpose of the algorithm is to find a suitable U and V so that J (U, V) is minimized. For this problem we can construct a Lagrange function:

$$L = J(U, V) + \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{c} \mu_{ij} - 1 \right) \tag{5}$$

where, $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n)^T \in R$ is the Lagrange multiplier. We can get the fuzzy clustering center according to the formula as follows:

$$\frac{\partial J(U, V)}{\partial v_i} + \sum_{j=1}^{n} \lambda_{ij} \frac{\partial \left(\sum_{i=1}^{c} \mu_{ij} - 1 \right)}{\partial v_i} = 0, \quad i = 1, 2, \cdots, c \tag{6}$$

The fuzzy clustering center is as follows:

$$v_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{m} x_i}{\sum_{j=1}^{n} \mu_{ij}^{m}}, i = 1, 2, \cdots, c \tag{7}$$

We can obtain the membership function according to the formula:

$$\frac{\partial J(U, V)}{\partial \mu_{ij}} + \sum_{J=1}^{N} \lambda_{ij} \frac{\partial \left(\sum_{i=1}^{c} \mu_{ij} - 1 \right)}{\partial \mu_{ij}} = 0 \tag{8}$$

While the membership function is as follows:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{\frac{2}{m-1}}} \tag{9}$$

If the data set X, clustering number c and weighting coefficient m is known, then the optimal membership matrix and fuzzy clustering center can obtain by formula (7) and (9). Algorthm of iterative steps is as follows:

Firstly, Initialization: use a random number between (0,1) to initialize the membership function matrix and pay attention to the constraints:

$$\sum_{i=1}^{c} \mu_{ij} = 1, j = 1, 2, \cdots, n$$

Then, accurate the fuzzy clustering center:

$$v_i(1) = \frac{\sum_{j=1}^{n} \mu_{ij}^{m}(1) x_i}{\sum_{j=1}^{n} \mu_{ij}^{m}(1)}, i = 1, 2, \cdots, c$$

Furthermore, calculate the value of function :

$$j(U, V)(1) = \sum_{i=1}^{c} \sum_{i=1}^{n} \mu_{ij}{}^{m}(1) d_{ij}{}^{2}$$

and judge whether it is less then ε, which ε is a small positive real number. We will stop the calculating and getting the fuzzy clustering center and membership degree matrix if the value is less then ε; otherwise, continue to the next step.

Finally, let l = l+1, calculate the membership function as follows and go to the step 2.

$$\mu_{ij}(1) = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}(1-1)}{d_{kj}(1-1)}\right)^{\frac{2}{m-1}}}$$

MANCHU TEXT EXTRACT UNDER THE IMAGE BACKGROUND

Image Segmentation Based on Fuzzy Cluster:In this study the approach of color image segmentation is based on the fuzzy maximum degree of membership. Before the image segmentation based on principle of fuzzy maximum degree of membership, first, observe the pixel feature of target area from the histogram, then the pixel values are classified by FCM algorithm, calculate the membership for each pixel belongs to each category in the color image, finally, according to the principle of maximum degree of membership to determine those pixels of each class of fuzzy sets belonging, thus completing the image segmentation. Image segmentation is showed in Fig. 1.

Manchu text preprocessing: Do the pre-processed on the Manchu words after segmentation with the fuzzy membership method, we can remove the noise around the words. Preprocessing mainly completes the work of denoising, thinning and pruning which is showed Fig. 2.

The specific process is as follows:

- **Denoising:** Because Manchu words have more details characteristics on the left side of mid-axis, while right feature less, this study uses morphological processing to filter out noise. First set a structure element, we find using the 1*2 rectangular structure element is the best through experiments. Corrosion Manchu word with the structural elements and then do the closing operation, however, we don't achieve the desired results from the image, so set the new structural elements according to the need, do the further corrosion process on the unsatisfactory area. Remove the noise after morphological filtering. As it shows in Fig. 2a-c
- **Skeletonized:** After the noise treatment Manchu word needs to be refined. Due to the refinement will refine the word into a ring, changed the original shape, so this study adopts the method of skeletonized, it can remains the important information of original object shape. As it shows in Fig. 2d
- **Pruning:** As skeletonized usually produce irrelevant "burr" or parasitic components, pruning can eliminate the "burr", here a pruning is enough. As it is showed in Fig. 2e.

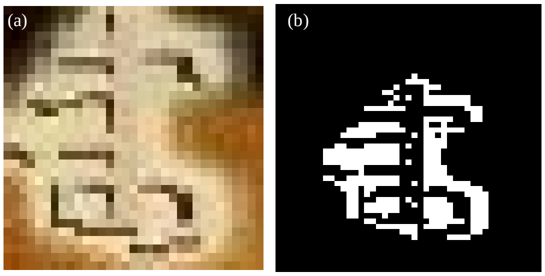

Fig. 1(a-b): Image segmentation (a) Manchu character and (b) Segment image

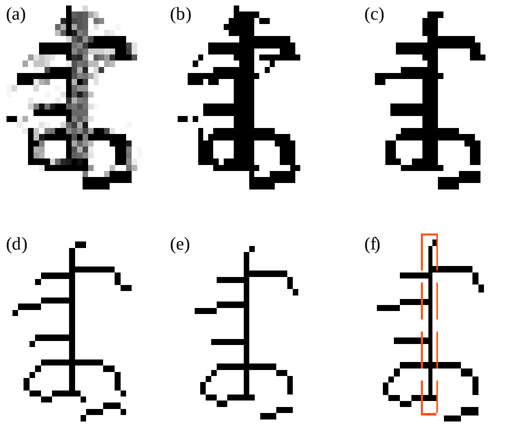

Fig. 2(a-e): Manchu text preprocessing (a) Wipe out, (b) Binaryzation (c) Denoising (d) Thinning (e) Pruning and (f) Mid-axis

Although Manchu is composed of letters, it is not like the English words which have a very clear boundary between letters. Manchu words have no space between letters and different letters in different positions have different writing. This makes the division of Manchu stroke very difficult.

This study uses classification method of stroke primitives, through the analysis we found that most of the Manchu writing body is vertical. Font structure is about structure so each Manchu word has a mid-axis as the backbone; all the strokes are outward expansion from the mid-axis. The strokes in the left of the mid-axis and connected with mid-axis is called left connecting strokes, unconnected strokes is called left free strokes; Similarly, the strokes in the right of the mid-axis and connected with mid-axis is called right connecting strokes, unconnected strokes is called right free strokes. Thus the strokes can be divided into four categories, respectively is the left connecting strokes, right connecting strokes, left free strokes and right free strokes.

Extract Mid-axis: After pretreatment, do the rank scanning on the extraction text and find the most effective pixel column as a mid-axis, as shown in Fig. 2f. The longest one in the middle is mid-axis. In order to reduce the main shaft caused by writing reason or the image acquisition process, study horizontal extension of the mid-axis to the right and left, the extension width is 1/10 of text width.

Stroke classifications: After refining the pixels can be divided into the following categories:

- **Critical point:** We have find the boundary of mid-axis, on the boundary scan to find the pixel which value is not zero and then store it. These points can be divided into left and right critical point due to the different of cross boundary
- **Interior points:** Progressive scan the word, if the value of the points is 1 and the sum of its eight neighboring pixels is 2, stored it until the scanning is completed. That is Rx = 1 and Rx = 1. According to the border area, the interior points are divided into internal points outside the boundary and interior point within the boundary. The formula is as follows:

$$B_x = \sum_{j-1}^{j+1}\sum_{i-1}^{i+1} \mathrm{img}(i,j) - \mathrm{img}(i,j) \qquad (10)$$

$$R_x = \mathrm{img}\ (I, j) \qquad (11)$$

The i, j is the current location of pixel values. img is the image which is preprocessed, Bx represents the sum of eight neighborhood of current pixel, Rx is the current pixel value.

- **Intersection point:** If scanned a effective pixel and the sum of its 8 neighborhood values is greater than 2, that is to say Bx>2, Rx = 1, then the point is the intersection point
- **Termination point:** If scanned a effective pixel and the sum of its 8 neighborhood values is 1, that is to say Bx = 1, Rx = 1, so the point is the termination point
- **Stroke growths:** After the Manchu text pixels are classified, we use the stroke growth method for text stroke extraction, at this time you should pay attention to the seed pixels, growth criterion and growth stop criterion. Specific steps are as follows:
- First generate a zero matrix which have the same size with the original image and generate the mid-axis in them, then carries on the anti color processing. The first left boundary point was regarded as the seed point to grow

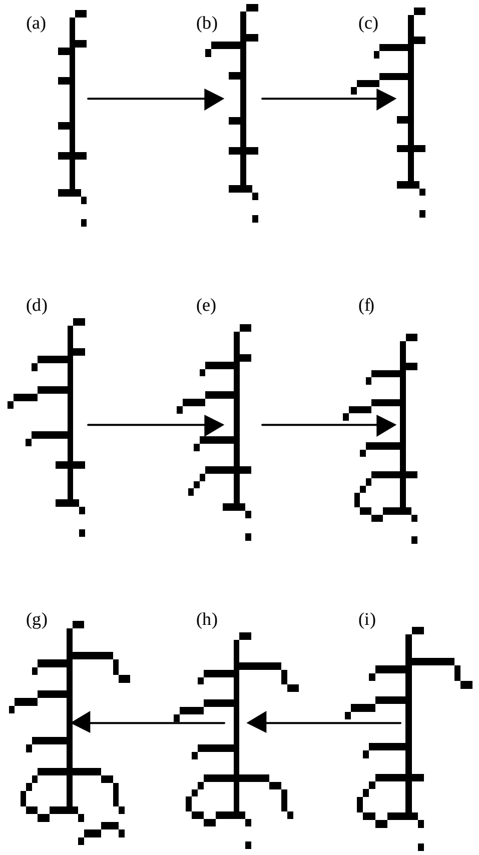

Fig. 3(a-g): Manchu character extraction process with stroke growth

- Depending on the size of pixel values around the left boundary point set a threshold value, since the image has been binarized, we set the threshold 0. The initial number of points will meet the region's growing conditions set to 1, track the effective pixel on the left which is adjacent to the boundary point. If the difference between the tracking pixel and the boundary pixel is less than the threshold we set, then set the track point as valid point to get the new pixel seed so as to continue to grow; otherwise the growth stops
- After the trace is complete, gather all the effective pixels then we will get the strokes by the growth
- Then scan the other boundary points sequentially on the left and get the corresponding connecting strokes. The same approach to the right ones. Grow from the left boundary points was called left connecting strokes, grow from the right boundary points was called right connecting strokes. Thus, the off-line handwritten Manchu character strokes were extracted, the extraction process is shown in Fig. 3a-i

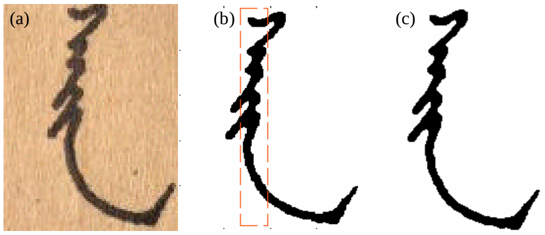

Fig. 4(a-c): Handwritten Manchu character extraction (a) Handwritten original image, (b) Handwritten mid-axis positioning (c) Character after stroke growth

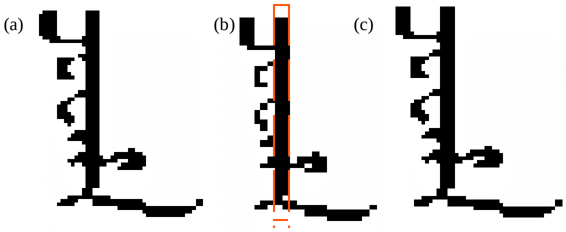

Fig. 5(a-c): Printed Manchu character extraction (a) Printed original (b) Mid-axis image positioning and (c) Character after stroke growth

EXPERIMENTAL ANALYSIS

In order to verify the effect of stroke growth on the text extraction, this study does a experiment on the image of handwritten Manchu. Text extraction on a image which contains 1000 handwritten Manchu character that not have free strokes, from the result we can find that all the Manchu characters are extracted, the extraction rate is 100%, as shown in Fig. 4. (a) For the handwritten original image, (b) For the handwritten Manchu mid-axis positioning, (c) Is the character after stroke growth. Obviously have a high accuracy. We also extract character from an image containing 1000 printed Manchu words with stroke growth method, the extraction rate also reached 100%, as shown in Fig. 5, (a) For the printed original image, (b) For the mid-axis positioning, (c) Is the character after stroke growth. From the experimental result we can see the stroke growth method is effective for the extraction of Manchu text.

CONCLUSION

In this study we use stroke growth method to extract character on handwritten Manchu, printed Manchu and Manchu under color image background, using digital image processing method for image pre-processing, achieve denoising effect by dilation and erosion, and then thinning and pruning on the image after denoising. Find the mid-axis and the seed point on it; finally do the stroke growth so as to achieve the purpose of text extraction. Experiments shows that the method can accurately extracted Manchu character in color images, it's a good text extraction method, thereby lay a good foundation for the recognition of Manchu language.

ACKNOWLEDGMENT

This study are partly supported by the Science and Technology Research Project of Liaoning Provincial Department of Education (L2012479), Research Project of State Ethnic Affairs Commission (12DLZ018) and the Fundamental Research Funds for the Central Universities (DC120101132).

REFERENCES

Abdullah, A., A. Hirayama, S. Yatsushiro, M. Matsumae and K. Kuroda, 2013. Cerebrospinal fluid image segmentation using spatial fuzzy clustering method with improved evolutionary expectation maximization. Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, July 3-7, 2013, Osaka, Japan, pp: 3359-3362.

Chen, F.X., 2008. Based on region growing image segmentation techniques. J. Sci. Technol. Inform., 15: 58-59.

Hinchcliff, M., E. Just, S. Podlusky, J. Varga, R.W. Chang and W.A. Kibbe, 2012. Text data extraction for a prospective, research-focused data mart: Implementation and validation. BMC Med. Inform. Decis. Making, Vol. 12. 10.1186/1472-6947-12-106

Peng, D.Q., J.Q. Li and Y.Q. Lin, 2010. FCM image segmentation based on the spatial restrained fuzzy membership. Comput. Sci., 10: 257-259.

Wang, Z.R. and C.L. Zuo, 2011. Method of the image feature extraction. J. Jishou Univ. (Nat. Sci. Edn.), 5: 43-47.

Zhang, G.Y., J.J. Li and A.X. Wang, 2007. Separation and recognition method for off-line handwritten Manchu character strokes. Comput. Eng., 22: 200-202.