Hough Transform with Guidance of Endpoints for the Purpose of Tangut Character Recognition

Yifei Meng, Xue Yuan, Xueye Wei, Feizhou Qin

***Abstract*—The Tangut script was a logographic writing system used for the extinct Tangut language of the Western Xia Dynasty spanned 1038 to 1227. To implement Tangut character machine recognition is a research issue claiming practical value. It is proposed in this paper to recognize Tangut character based on 4 corners code, and Hough Transform is suggested in the process of detecting the strokes of the character. Hough Transform is one of the most used procedures in morphology image processing, while for the purpose of character recognition, revised measure is needed. Hough Transform with Guidance of Endpoints (HTGE) was proposed in this paper to accommodate the character recognition. Improvements of HTGE: (1) Take advantage of endpoints to reduce the computational burden of Hough transform. (2) Detect the strokes with tolerance for the non-strictly straight line. (3) Judge the straight line with consideration of length of the line to avoid neglecting the short lines. Experiments results were presented to verify the improvement of the HTGE.**

***Keywords*—Tangut character recognition; four corners code; Hough transform; line detection**

I. Introduction

A. Tangut character

THE Tangut Empire, also known as the Xi Xia Empire, was an empire which has existed from 1038 to 1227 in what are now the northwestern Chinese provinces of Ningxia, Gansu, eastern Qinghai etc., measuring about 800,000 square kilometers[1-3]. Unique language and character have been used in Tangut Empire. The Tangut script, promulgated as the official script of the Tangut Empire, was in common used within the Tangut state for slightly more than 200 years. [4] With the extinction of the empire, the character has also been forgotten for nearly thousand years. Until recent 100 years, Tangut relic was discovered, and the missing world was unfolded with great amount of antique and script written in Tangut character.

Manuscript received Dec. 4, 2017; revised Jan. 15, 2018. This work is supported by the Natural Science Foundation of Ningxia Hui Autonomous Region No. NZ17261.

Yifei Meng is with the School of Electronic and Information Engineering of Beijing Jiaotong University and School of Physics and Electronic-Electrical Engineering Ningxia University (phone: 15825311852; e-mail: 10111045@bjtu.edu.cn).

Xue Yuan is with the School of Electronic and Information Engineering of Beijing Jiaotong University (e-mail: xyuan@bjtu.edu.cn).

Xueye Wei is with the School of Electronic and Information Engineering of Beijing Jiaotong University (e-mail: xywei@bjtu.edu.cn).

Feizhou Qin is with the School of Physics and Electronic-Electrical Engineering Ningxia University (e-mail: 329898420@qq.com)

B. Informatization and digitalization of Tangut character in Tangut research

During the 1907-1909 a Mongol-Sichuan expedition under the command of Pyotr Kuzmich Kozlov was launched by Asian Museum, St. Petersburg. In 1908, Kozlov made the historical discovery of Khara-Khoto. [5, 6] Over 2,000 books, scrolls and manuscripts in the Tangut language were uncovered. Considerable amount of Tangut script were discovered, content of these script range from sutra, statute book to secular document. From then on, the study of history and culture of the Tangut has been keeping on a hotspot.

Once the character can be recognized, all of these scripts will reveal a vivid society nearly one thousand years ago. But it is not a goal easy to be achieved. In the worldwide the number of the researchers who can master Tangut language intensively is less than ten, meaning while great deal of ancient Tangut scripts desiderate to be collated preserved and interpreted. So, the digitalization and informatization of Tangut character present an urgent demand. Optical character recognition (OCR) of the Tangut character. To compensate the scarcity of the specialist who can master the Tangut language, and to provide a helpful application for Tangut researchers, a practical Tangut OCR system is needed to be established. It will facilitate the work of Tangut researcher greatly on the aspect of ancient documents collating, character collection and comparison as while as script translation.

C. Related work

In the field of Optical Character Recognition (OCR), large amount of technique and method have been presented. In the aspects of preprocessing and feature extraction, normalization-cooperated gradient feature (NCGF) was proposed in [7] to alleviate the effect of stroke direction distortion caused by shape normalization and provide higher recognition accuracies. In [8] a new technique of calculating twelve directional feature inputs depending upon the gradients was used to implement hand written character recognition. Zhao, Y.X. and Chou, C.H. put forward the neighborhood-relationship feature selection (NRFS) algorithm to identify rat electroencephalogram signals and recognize Chinese characters.[9]

All of above feature extracting method was proposed for the purpose either for universal character object or for kinds of general used character such as Chinese. In this paper, Tangut character was investigated as a specific recognition

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

object, and Hough Transform was introduced for the extraction of the character stroke feature.

The Hough Transform was patented as U.S. Patent by P.V.C. Hough in 1962, [10, 11]it is used to detect geometric features like straight lines in digital images, is likely one of the most widely used procedures in computer vision, as evidenced by more than 2500 research papers dealing with its variants, generalizations, properties and applications in diverse fields[12]. Modifications and variants have been made in one or more stages. Some of these algorithms have introduced features that are significantly distinct from the Standard Hough Transform (SHT). The variants include: Generalized HT (GHT), introduced by Ballard to detect non-parametric curves.[13] Probabilistic HT (PHT) by Kiryati et al. [14] and , Randomized HT (RHT) by Xu et al.[15] were proposed to achieve speeding up HT.

The HT has been utilized so widely, even after the golden jubilee year of existence. In[16] an integrated method is proposed based on the HT and Contour Detection to detach the overlapping circular objects efficiently. K. B. Ray et al. represented palm print with Hough peak features[17]. Hough Transform Statistics (HTS) was introduced to characterize the shape of objects in digital images in [18]. HT was also utilized in unmanned automobile and the automobile auxiliary driving system by Li, X., et al. for lane detection. [19]

D. Contribution of this paper

Though there are considerable amount of papers in the field of OCR, these papers scarcely are devoted to Tangut character. In this paper, Tangut character was investigated as a specific recognition object. The particularity way by which Tangut characters were composed, as well as the shape and structure of the character were intensively studied in this paper. As a result, character recognition based on the 4 corners code specifically for the Tangut character was proposed. The method was presented creatively in this paper with consideration of the particular nature of the Tangut character.

Hough Transform was used in the process of implement the Tangut character 4 corners component recognition. For the purpose to detect the stroke of Tangut character components SHT was revised to achieve better performance and efficiency. Given the feature of the Tangut character, Hough Transform with Guidance of Endpoints (HTGE) was proposed. HTGE features in aspect of (1) reduction of computational burden, (2) flexibility to detect none strictly straight line and (3) sensitivity to detect the trivial short line which may be ignored by SHT.

II. TANGUT CHARACTER RECOGNITION

A. Particularity of Tangut character for recognition

The Tangut character is remarkable for being the most inconvenient of all scripts, a collection of nearly 5,800 characters of the same kind as Chinese characters but rather more complicated. Very few are made up of as few as four strokes and most are made up of a good many more, in many cases nearly twenty. In some cases characters which differ from one another only in minor details of shape or by one or two strokes have completely different sounds and meanings.[20]

Fig. 1 are the words “signal processing” written in Tangut language[21].

Fig. 1 “signal processing” in Tangut character

Since much research devotion has been pay for the Chinese character recognition. And there is considerable similarity between Tangut and Chinese character, it is reasonable to analyze the difference and common ground of two kinds of character from the point view of character recognition to figure out a suitable method for Tangut character recognition.

The common ground includes:

1) Both Chinese and Tangut character are based on logographic writing system [1].

2) On the whole, their appearance both present square shape.

3) Chinese and Tangut character are composed with same basic strokes.

The difference between Chinese and Tangut character:

1) Tangut character is rather more complicated; most characters are made up of between four to twenty strokes. On average, a Tangut character is composed with 16 strokes approximately, while a Chinese character only less than 10[6].

2) Oblique strokes are more likely be used in Tangut character.

3) Unlike Chinese character as hieroglyphical, the shape of Tangut characters has little relationship with the things they present.

For the complexity and other particular feature of the Tangut character, recognition method suitable with these particularities is expected to be researched. In this paper four corners coding system of Tangut character was focused and studied to achieve efficient and accurate character recognition of Tangut character.

B. Four corners coding of Tangut character

In the Li Finsen’s *Xia-Han Dictionary*, it is creatively suggested to code the Tangut character with 4 corners coding rule. The main idea is to summarize 24 components of Tangut character and to classify the components into 9 classes, and furthermore, to code each character according to the components of the 4 corners of the character.

The 9 classes components are showed in Fig. 2.

For instance, the characters showed in Fig. 3 are coded according their 4 corners respectively. For the complexity of Tangut character, two more components in the bottom of character are accounted as sub code. Therefore a 6 digits number is used to stand for a Tangut character.

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

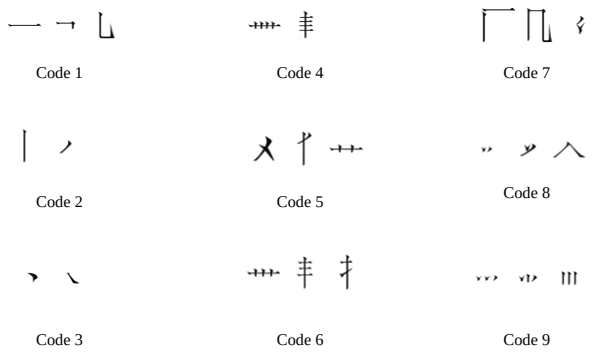

Fig. 2 24 character components classified into 9 classes

Fig. 3 Characters coded by the rule of 4 corners coding

C. Tangut Character Recognition based on four corners code

The issue of recognizing a Tangut character can be converted to recognizing the corner components of the character. The character is first divided into several sub images. For each sub image, a code is determined by calculating the specified image characteristics of this sub image. Each code represents a certain character component, and once a 6 digits number is extracted based on the character components, the correspondent character will be determined.

The procedure includes:

1) Preprocess the character image to get the skeleton of the character.
2) Divide the character into up, bottom, left and right part.
3) Check the component in each character part, and determine the corresponding code.
4) Determine the character according the 6 digits number code obtained in step 3.
5) Because there is case of repeated code in Tangut character 4 corners code, a 6 digits code may stand of several characters, so it is necessary to define the right one in characters that are corresponding to the 6 digits code obtained in step 4.

III. Preprocess and Restoration of Ancient Tangut Scripts Image

As previously mentioned, The Tangut relic had gone through for hundreds of years before they were excavated. There are great deals of phenomenon of character desquamation and strokes rupture, so the restoration of the ancient script image is necessary.

To fix these problems, morphological processing closing operation is employed.

The morphological closing operation of A by B, denoted $A \cdot B$, is a dilation followed by an erosion:

$$A \cdot B = (A \oplus B) \odot B \qquad (1).$$

In the formula (1), dilation of A by B is denoted as $A \oplus B$, and the erosion operation is denoted as $\odot$.

IV. Hough Transform with Guidance of Endpoints (HTGE)

A. Hough Transform

The simplest case of HT is detecting straight lines. In spatial space, a straight-line $y = mx + b$ can be represented as a point (b, m) in the parameter space also named as Hough space, while a point in spatial space project a line in Hough space presenting all the possible parameters of each line that pass the point. A set of points that form a straight line will produce lines in parameter plane which cross at the certain point (b, m) for that line. Thus, the problem of detecting collinear points can be converted to finding concurrent lines.

One problem is, for a vertical line, the slope parameter m would be infinity. Thus, for computational reasons, R. O. Duda and P. E. Hart resolved the issue by mapping a point in image space to a sinusoidal curve in parameter space[22]:

$$\rho(\theta) = x\cos(\theta) + y\sin(\theta) \qquad (2).$$

where ρ is length of the perpendicular line from the origin to the straight line, and θ is the angle between the x axis and the perpendicular line.

B. Reduce Processing Workload with HTGE

As stated above, to perform Hough transform to detect the lines in an image, every nonzero pixels should be supposed to be an online point, and the θ of the hypothetical line range from -90° to 90°. So every ρ corresponding to each hypothetical θ should be calculated with the formula (2). With the default $\Delta\theta$, 1°, 181 times calculation would be performed for each nonzero pixel. The workload of computing is enormous with the increasing of the image size and amount of nonzero pixel.

Contrast to the SHT, the HTGE applied for the character reorganization take the advantage of stroke information. For the endpoints of stroke of character can be easily detected, and with the given stroke endpoints, the workload of computing the ρ and θ will be reduced.

As shown in Fig, after preprocessing, the endpoints of the skeletonized components are detected.

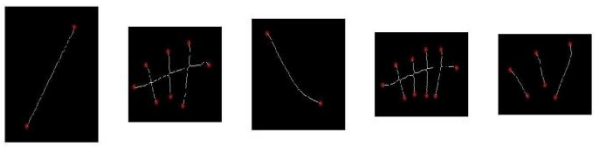

Fig. 4 Skeletonized character components and the endpoints

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

With these endpoints as guiding information, the workload of HTGE computation can be reduced greatly. The method is described as follows:

1) Hypothesize possible straight line between every two endpoints.
2) Compute ρ and θ of each possible line, to get the result ρ_i and θ_i; $i \in 1{\sim}n$; $n = C_p^2$; p is the number of the endpoints, n is the combination number of 2 from p.
3) In the process of Hough transform, for each nonzero point, instead of assigning θ with value from -90° to 90° to get ρ, the value range of θ can be shrink to $\theta_i \pm \theta_{thresh}$. The θ_{thresh} is decided by how march.

C. Line detection with tolerance

In SHT line detecting process, the target is limited to the strictly straight line. As shown in Fig. 5, a1, b1, c1, d1 are 4 lines of different curvature, a2, b2, c2, d2 are SHT result for each line, the results are shown as bold red line in figure.

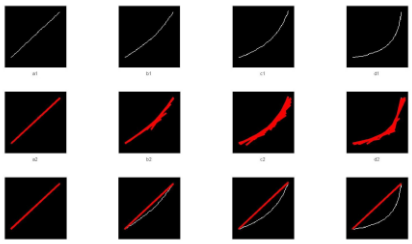

Fig. 5 Deferent curved lines and the corresponding Hough transform result

With increase of the curvature, the curves in b1, c1, d1 in Fig. 5 are interpreted as several strait lines with SHT. While in the process of character stroke detection, there are always the cases that the stroke is not strictly straight line, so the method of line detection with tolerance is suggested in this paper.

The process is presented as follow:

1) To detect endpoints of the character skeleton, and hypothesize there be line $L_i, (i \in 1{\sim}I$ I is the number of hypothesized line) between each pair of endpoints.
2) To calculate θ_i and ρ_i $(i \in 1{\sim}I,)$ according every L_i.
3) All the nonzero points in skeleton image are supposed to be on the line $L_n, (n \in 1{\sim}N$, N is number of nonzero points). The angle between line L_n and x axis is θ_i. ρ_n of L_n is calculated.
4) To check the amount of ρ_n which meet the condition $\{\rho_n \mid \rho_i - rdd < \rho_n < \rho_i + rdd\}$. The amount above the threshold means that hypothesized L_i can be accounted a real line.

In Fig. 5, a3 b3 c3 d3 are the HTGE detection results with process described above. All curves are interpreted as a straight line.

D. Houghpeak with length factor

In SHT process, parameter space is discrete into an accumulative matrix, points that form a line in image is reflected to corresponded accumulative matrix in parameter space. So, extraction of the peak value in parameter space will lead to definition of the line in spatial space.

While in SHT absolute value of accumulative matrix in parameter space is used as standard to determine if there is a corresponding line in spatial space. So, the amount of pixels which form a line in image is the key factor determining whether they can be interpreted as a line. The defect of the method is that short line may be ignored. As show in Fig. 6, a is the image with 4 straight lines, b is the result of SHT, 2 shorter lines are not detected.

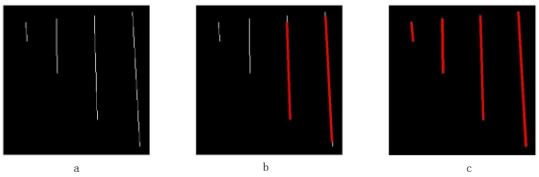

Fig. 6 Lines of different length and Hough transform result.

It is suggested in this paper that length factor be introduced in the peak search process of Hough transform.

$$q = \frac{a}{\sqrt[2]{(x_{e1}-x_{e2})^2+(y_{e1}-y_{e2})^2}} \qquad (3)$$

In formula (3), a is the absolute peak value of the accumulative matrix, ratio of a and distance of the two endpoints (x_{e1}, y_{e1}), (x_{e2}, y_{e2}) of the line make the relative value q. The method to search in the parameter space with q as criterion, will mend the defect of ignoring the short line.

As show in Fig. 6, c is the result processed with recommended method.

V. Experiment Result and Analysis

A. Accuracy and running time of line detection by HTGE

Several experiments were set to verify the accuracy and efficiency of HTGE. 3 sets of images were created for experiments. The size of each image is 200*200. Image set 1 contains 100 images of straight lines, for each image, there are 5 straight lines with random length, position and direction, as shown in Fig. 7 a. Image set 2 contains 100 images of arcs, for each image, there are 5 arcs with random radian, position and random radius range from 50 to 400, as shown in Fig. 7 b. Image set 3 contains 100 images of arcs and strait lines, for each image, there are 3 random straight lines and 3 random arcs as shown in Fig. 7 c.

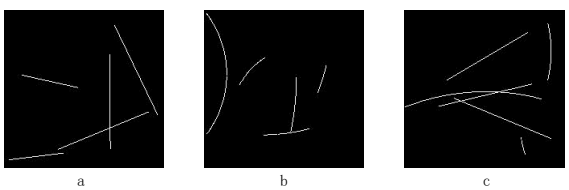

Fig. 7 Different line sets for experiment

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

3 sets of images were processed with SHT and HTGE respectively. The test results were shown in the Table I.

TABLE I
RESULT OF LINE DETECTION EXPERIMENT WITH HTGE AND SHT

Experiment	Different methods	Truthful lines	Detected lines	Run time(s)	Accuracy
1	HTGE	500	426	0.684	85.2%
	SHT		365	1.956	73.0%
2	HTGE	500	364	0.770	72.8%
	SHT		1636	3.807	
3	HTGE	600	501	0.841	83.5%
	SHT		336	2.322	56.0%

In experiment 2, 1636 lines are detected with SHT, while the truthful line number is 500. The case can be explained with Fig. 5 b2, c2 and d2. One curve was interpreted as several strait lines with SHT.

B. Tangut Character component Recognition with HTGE

With HTGE, the stroke in the Tangut character component can be efficiently detected, and the results of the line detection provide key discriminatory information for the Tangut character recognition.

Some Tangut character components and their HTGE result are shown in Fig. 8

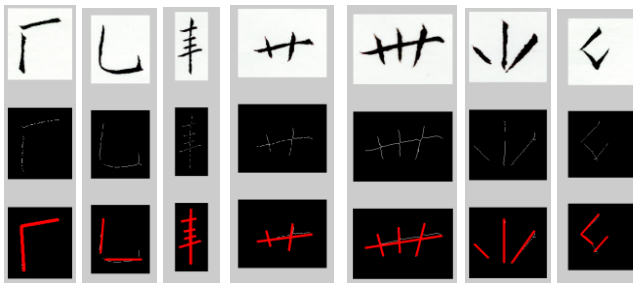

Fig. 8 Line detection result with HTGE

In Fig. 7, first row of image are the handwriting Tangut character components, second row are skeletons of the components, the third row are the result of HTGE, the strokes are detected as straight line and shown as bold red lines.

The slop and number of the lines that compose a Tangut character component can be used to be a decisive discriminatory feature in the character recognition. In the process of recognition, the lines information, such as slop and amount, are extracted from the recognition object character component, and compared with the sample to match with the most similar one.

VI. Defection and Prospection

The practical program code that implements Tangut character component recognition as well as statistics and analysis of the recognition result will be presented in further literature.

There are some problems not yet solved in this paper, such as how to detach the character components when they are overly mixed with other strokes.

The recognition method presented in this paper is assumed to determine the 4 corners code of the character. For the problem of repeated code in Tangut character 4 corners code, further research are expected to achieve the target of defining the right one from several characters that correspondent to the same 4 corners code.

References

[1] W. Tianshun, The Battle History of Western Xia. Ningxia People's Press, 1993.
[2] B. Ren, Western Xia: the kingdom lost in historical memories. Beijing: Foreign Language Press, 2005.
[3] L. Fanwen, Comprehensive History of Western Xia. Beijing, Yinchuan: People's Press, Ningxia People's Press, 2005.
[4] L. Kwanten, "THE STRUCTURE OF THE TANGUT HSI-HSIA CHARACTERS," (in English), Toung Pao, Article vol. 75, no. 1-3, pp. 1-42, 1989.
[5] (29 Nov. 2016). Khara-Khoto. Available: https://en.wikipedia.org/wiki/Khara-Khoto#cite_note-Kychanov-7
[6] E. Kychanov, "Wen-Hai Bao-Yun: The book and its fate," Manuscripta Orientalia, vol. 1, no. 1, pp. 39-44, 1995.
[7] C. L. Liu, "Normalization-cooperated gradient feature extraction for handwritten character recognition," Ieee Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 8, pp. 1465-1469, Aug 2007.
[8] D. Singh, S. K. Singh, and M. Dutta, "Hand Written Character Recognition using twelve Directional feature Input and Neural Network.pdf," International Journal of Computer Applications, vol. 1, no. 3, 2010.
[9] Y. X. Zhao and C. H. Chou, "Feature Selection Method Based on Neighborhood Relationships: Applications in EEG Signal Identification and Chinese Character Recognition," Sensors (Basel), vol. 16, no. 6, Jun 14 2016.
[10] P. V. C. Hough, "Method and means for recognizing complex patterns," U.S. Patent 3069654, 1962.
[11] P. E. Hart, "How the Hough Transform Was Invented," (in English), Ieee Signal Processing Magazine, Editorial Material vol. 26, no. 6, pp. 18-22, Nov 2009.
[12] P. Mukhopadhyay and B. B. Chaudhuri, "A survey of Hough Transform," (in English), Pattern Recognition, Article vol. 48, no. 3, pp. 993-1010, Mar 2015.
[13] D. H. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," Pattern Recognition, vol. 13, no. 2, pp. 111-122, 1981/01/01 1981.
[14] N. Kiryati, Y. Eldar, and A. M. Bruckstein, "A probabilistic Hough transform," Pattern Recognition, vol. 24, no. 4, pp. 303-316, 1991/01/01 1991.
[15] L. Xu, E. Oja, and P. Kultanen, "A new curve detection method: Randomized Hough transform (RHT)," Pattern Recognition Letters, vol. 11, no. 5, pp. 331-338, 1990/05/01 1990.
[16] J. Ni, Z. Khan, S. Wang, K. Wang, and S. K. Haider, "Automatic detection and counting of circular shaped overlapped objects using circular hough transform and contour detection," in 2016 12th World Congress on Intelligent Control and Automation (WCICA), 2016, pp. 2902-2906.
[17] K. B. Ray and R. Misra, "Palm Print Recognition Using Hough Transforms," in 2015 International Conference on Computational Intelligence and Communication Networks (CICN), 2015, pp. 422-425.
[18] G. B. de Souza, A. N. Marana, and Ieee, "HTS: a New Shape Descriptor Based on Hough Transform," 2013 Ieee International Symposium on Circuits and Systems (Iscas), pp. 974-977, 2013.
[19] X. Li, Q. Wu, Y. Kou, L. Hou, and H. Yang, "Lane detection based on spiking neural network and hough transform," in 2015 8th International Congress on Image and Signal Processing (CISP), 2015, pp. 626-630.
[20] C. Gerard, "The Future of Tangut (Hsi Hsia) Studies," Asia Major, vol. 11, no. 1, pp. 54–77, 1964.
[21] C. You. 古今文字集成. Available: http://www.ccamc.co/index.php
[22] R. O. Duda and P. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," Communications of the ACM vol. 15, no. 1, pp. 11-15, Jan 1972.

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)