Research on the Preprocessing of Tai Le Character Recognition

Hai Guoa, Doudou Yangb, Jingying Zhaoc, Yifan Liud

College of Computer Science and Engineering, Dalian Minzu University, Dalian, 116600, China

aguhai@dlnu.edu.cn, byddzjutc@163.com, czjy@dlnu.edu.cn, d1129550657@qq.com

Abstract—Tai Le character is a kind of alphabetic writing with a long history, and its character recognition research plays an important role in inheriting and protecting Tai Le. Preprocessing is an key part of the Tai Le character recognition system, the quality of the preprocessing directly affects the recognition rate in the recognition system. Image preprocessing is the premise of feature extraction and recognition, since the directly input image usually contains noise, it affects the recognition effect. Therefore, it is of great application value and far-reaching historical significance to study the preprocessing technology of Tai Le character recognition. This paper proposes the image preprocessing method of Tai Le character, including graying, binarization, de-noise, and edge detection, and achieves good preprocessing results, which lays a foundation for feature extraction and recognition of Tai Le character.

1. INTRODUCTION

The process of Tai Le character recognition mainly includes preprocessing, character segmentation, feature extraction, and Tai Le character recognition. Preprocessing plays an key role in Tai Le character recognition technology, it is the first condition for the research work to have a good preprocessing technology.

Jorge Calvo-Zaragoza et al. proposed a convolutional auto-encoders, which analyzes the document image and applies a global threshold. This method is superior to the existing binarization strategies in many document types[1]. Jinyuan Zhao et al. introduced conditional generative adversarial networks to solve the core problem of multi-scale information combination in binarization task. Compared with many classical and state-of-the-art approaches, this method gains promising performance in the accuracy and robustness of binarization[2]. Sauvola’s methods use local average and standard deviation of the pixels within a small neighborhood window to obtain a thresholding segmentation surface, and also have better performance but depend heavily on the sliding window size and hence the text stroke width[3]. Mitianoudis N et al. addressed the document image binarization problem using a Local Co-occurrence Mapping, local contrast and a two-state Gaussian Mixture Model[4]. Xiong W et al. proposed a support vector machine based method for degraded historical document image binarization, which is outperformed other state-of-the-art techniques in terms of F-measure, NRM, DRD, and MPM[5].

This paper mainly studies a series of digital image processing methods, e.g., image graying, binarization, de-noise and edge detection in Tai Le character recognition preprocessing.
2. **THE CHARACTERISTICS OF TAI LE CHARACTER**

The Tai Le character has a long history, profound heritage, and a wide range of uses. As a cross-border language, it is used in Myanmar, Thailand, Laos, and Yunnan Province, China. For a long time, the Chinese Dai people have cherished it because of its rich historical and cultural values, and it is the official language of the Jingpo Autonomous Prefecture in China.

With the rapid development of the times, the development of national culture has lagged behind. Studying Tai Le character is not only to show people the excellent culture of ethnic minorities, let more people feel the diligence and wisdom of the ancestors of the Dai people, feel the charm brought by this excellent culture, and more importantly, let people come to learn, cherish, inherit and carry forward and deeply understand the historical and cultural values contained in this culture, making the brilliant Chinese culture more eye-catching.

In the international standard, Tai Le contains a total of 35 characters and the coding space is U1950-U197F. Fig. 1 shows the Unicode of Tai Le characters.

![Figure 1. Tai Le characters Unicode](image)

Tai Le is a kind of alphabetic writing, which is made up of characters. Its characters are characterized by a small number of characters, but too many characters with similar forms, and the number of characters with similar shape accounts for more than 70% of the total number of characters. It can be seen that Tai Le has high similarity and low discrimination. In order to better distinguish the categories of Tai Le, it is necessary to preprocess the character images properly. Image preprocessing is the premise of character recognition, only by removing the interference part of the images and making the obtained information more reliable, can better realize the recognition of Tai Le. Next, we will discuss the preprocessing methods of Tai Le.

3. **TAI LE CHARACTER RECOGNITION PREPROCESSING FUNCTION MODULE**

Due to the uneven quality of the collected Tai Le character samples, in order to minimize the deviation and improve the recognition rate of Tai Le character, this paper adopts a series of image preprocessing methods to improve the image quality. In general, preprocessing algorithms generally include graying, binarization, de-noise and edge detection. Tai Le character preprocessing steps are shown in Fig. 2.

3.1 **Graying**

The color image does not play any role in the follow-up Tai Le recognition, but increases the amount of calculation. Therefore, it is necessary to convert the Tai Le image into a grayscale image before subsequent preprocessing. Color images are three channels, and the pixel value of each channel ranges from [0, 255]. When the values of the three channels are equal, the color is expressed as a grayscale color, and this value is called the grayscale value. There are several ways to image graying:

1) **Component method**

The brightness of the three components R, G, B in the color image is taken as the grayscale value of the three gray images, and one of the gray images can be selected according to the specific situation.
2) **Maximum method**

The maximum value of the brightness of the three components R, G, B in the color image is taken as the grayscale value of the gray image, that is:

\[
g_{\text{gray}}(i, j) = \max\{R(i, j), G(i, j), B(i, j)\}
\]

3) **Average method**

Take the average of the brightness of the three components R, G, B in the color image as the grayscale value of the gray image, that is:

\[
g_{\text{gray}}(i, j) = (R(i, j) + G(i, j) + B(i, j)) / 3
\]

4) **Average method**

Any color can be obtained by mixing the three primary colors in different proportions. Multiply the brightness of the three components R, G, B in the color image by different coefficients, and then add them as the grayscale value of the gray image, that is:

\[
g_{\text{gray}}(i, j) = \alpha R(i, j) + \beta G(i, j) + \gamma B(i, j), \quad \alpha + \beta + \gamma = 1
\]

where \( R(i, j), G(i, j), B(i, j) \) are the pixel values in the red, green, and blue channels, respectively. Because the human eye is the most sensitive to green and the least sensitive to blue, we selects \( \alpha = 0.229, \beta = 0.587, \gamma = 0.114 \) as the scale coefficient, \( g_{\text{gray}}(i, j) \) is the grayscale value obtained.

Because the maximum method will produce a high brightness gray image, the average method will produce a softer gray image, and the weighted average method is better than other methods in terms of the overall and local color and brightness level distribution characteristics of the character image. Therefore, the weighted average method is selected for image graying.
3.2 Binarization

After the image graying, binarization is carried out. Seba Susan et al. proposed adaptive thresholding mimics the ability of the human eye to iteratively adjust to varying light intensities through iterative gamma correction followed by contrast stretching so that the text becomes well defined against the background clutter[6]. In the process of Tai Le character recognition, we should pay attention to the stroke features of characters, not the color.

Therefore, in order to better extract effective features and reduce the amount of calculation, Tai Le character images need to be binarized. The binary expression is as follows:

\[
p_b(i, j) = \begin{cases} 
0, & p(i, j) < T \\
1, & p(i, j) \geq T 
\end{cases} 
\]  

(7)

where the gray value of the original image is expressed as \( p_{ij} \), the gray value of the pixel after binarization is \( b_{ij} \), and \( T \) is the threshold value. Otsu method was proposed by Nobuyuki Ostu in 1979[7]. Ostu binarization is based on the statistical characteristics of the whole image to achieve automatic threshold selection. It divides the image into foreground pixel and background pixel, and calculates the best threshold which can separate the two classes, so that the between-class variance is maximum.

Thus, the threshold \( T \) can be computed as follows:

\[
T = \max_{t \in [0,255]} \sigma_B^2, \sigma_B^2 = \frac{[\mu_t \omega(t) - \mu(t)]^2}{\omega(t)[1-\omega(t)]}
\]

\[
\mu_T = \sum_{i=0}^{255} t ip_t, \mu(t) = \sum_{i=0}^{255} t ip_t, \omega(t) = \sum_{i=0}^{255} p_t
\]  

(8)

where \( \sigma_B^2 \) is the between-class variance, \( \mu_T \) is the mathematical expectation of the whole image, \( \mu_t \) and \( \omega(t) \) respectively represent the mathematical expectation and probability sum of pixel values in the range \([0,t]\) . Ostu method is widely used, and satisfactory results can be obtained no matter whether the histogram has obvious bimodal histogram.

Fig. 3 shows the comparison of the grayscale image of Tai Le character and binary image by Otsu method.

![Figure 3](image)

Figure 3. Binarization effect of Tai Le character image. (a) Grayscale image. (b) Binary image by Ostu method.

3.3 De-noise

After binarization, some noise may remain in the Tai Le character image. In order to improve the recognition rate of Tai Le character, the image needs to be filtered. There are many commonly used filtering methods, such as morphological denoising, median filtering, median filtering, Gaussian filtering, and Laplacian denoising. Among them, the median filtering is a commonly used nonlinear...
smoothing filter. This method arranges all pixel values in the convolution kernel in descending order, and assigns the median value of the sequence to the central pixel of the convolution kernel.

The convolution kernel size is $3 \times 3$ in this paper, as shown in Fig. 4. The median filtering algorithm in two dimensions is defined as: Let $\{p(i, j)\}$ be the gray value set of each point in the convolution kernel $K$, and the convolution kernel $K$ size is $S = (2k + 1) \times (2k + 1)$, $k$ is a positive integer, $g(i, j)$ is the median value of the gray value set in the convolution kernel $K$, then:

$$g(i, j) = \text{med}[p(i - k, j - k), \cdots, p(i, j), \cdots, p(i + k, j + k)],$$

$k$ is a positive integer. (9)

where $\text{med}[]$ is the median function.

Gaussian filtering is different from median filtering. When averaging pixels in the neighborhood, it gives different weights to pixels in different positions. The expression is as follows:

$$G(i, j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}}$$

(10)

where $G(i, j)$ represents the weight assigned to the pixel $(i, j)$. Gaussian filtering calculates the weighted average of the pixels in the convolution kernel, and the pixels near the center have a larger weight value. The convolution kernel slides on the image, and after traversing each pixel of the image, the Gaussian filtering of the entire image is completed. The Tai Le character images using median filtering and Gaussian filtering are shown in Fig. 5. The results show that the Gaussian filtering can better remove the noise in the image, which is beneficial to the subsequent preprocessing steps.
3.4 Edge Detection
The edge of an image is generated between two areas with different gray values, the edge detection method is based on this feature, and the continuity of gray values is detected by calculating the differential value of each pixel to determine the image edge. Canny edge detection operator is a multi-level edge detection algorithm developed by John F. Canny in 1986[8], the specific steps are as follows:

- Use the first-order partial derivative of the image gray value \( G_x, G_y \) to calculate the gradient amplitude and direction:

\[
|G| = \sqrt{G_x^2 + G_y^2}
\]

\[
\theta = \arctan\left(\frac{G_y}{G_x}\right)
\]

- Non-maximum suppression of gradient amplitude. The traversal method is used along the gradient direction of the edge, which can be roughly divided into four types (horizontal, vertical, 45° direction and 135° direction), and compared with different neighboring pixels to determine the local maximum.

- Use double threshold algorithm to detect and connect edges.

![Figure 6. Tai Le character images detected using a fixed threshold Canny operator](image)

The experimental results are shown in Fig. 6. The results show that the edges of the character detected by the Canny algorithm are completely continuous, can identify the clear edges of the character. The edge of the character recognized by the Canny operator can well express the detail of the character. Even if a fixed threshold is used, there is no interference information such as blurry and pseudo edges, which can better complete the follow-up research work of feature extraction and recognition of Tai Le character.

4. SUMMARY
In this paper, the image graying, binarization, de-noise and edge detection of the image in the preprocessing of Tai Le character recognition are studied, and the algorithms is implemented in Python, which can get a good preprocessing image, and prove the effectiveness and stability of the algorithms, laid a solid foundation for the subsequent feature extraction and character recognition of Tai Le characters.

ACKNOWLEDGMENT
This work was supported only by the Science Foundation of Ministry of Education of China (No.18YJCZH040).
REFERENCES

[1] Jorge Calvo-Zaragoza, Antonio-Javier Gallego. A selectional auto-encoder approach for document image binarization[J]. Pattern Recognition, 2017, 86:37-47.

[2] Jinyuan Zhao, Cunzhao Shi, et al. Document image binarization with cascaded generators of conditional generative adversarial networks[J]. Pattern Recognition, 2019, 96.

[3] J. Sauvola, M. Pietikäinen. Adaptive document image binarization[J]. Pattern Recognition, 2000, 33(2):225-236.

[4] Mitianoudis N, Papamarkos N. Document image binarization using local features and Gaussian mixture modeling[J]. Image & Vision Computing, 2015, 38(jun.):33-51.

[5] Xiong W, Jingjing X, Xiong Z, et al. Degraded historical document image binarization using local features and support vector machine (SVM)[J]. Optik, 2018:218-223.

[6] Susan S, Rachna Devi K M. Text area segmentation from document images by novel adaptive thresholding and template matching using texture cues[J]. Pattern Analysis & Applications, 2019.

[7] Otsu N. A thresholding selection method from gray-level histogram[J]. IEEE Trans.syst.man. & Cybern, 1979, 9(1):62-66.

[8] J. Canny. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 8: 679-698.