Characters Feature Extraction based on Neat Oracle Bone Rubbings

Lei Guo
School of Computer and Information Engineering, Anyang Normal University
Anyang 455000, Henan, China
e-mail: 1071958514@qq.com

Abstract

In order to recognize characters on the neat oracle bone rubbings, a new mesh point feature extraction algorithm was put forward in this paper by researching and improving of the existing coarse mesh feature extraction algorithm and the point feature extraction algorithm. Some improvements of this algorithm were as followings: point feature was introduced into the coarse mesh feature, the absolute address was converted to relative address, and point features have been changed grid and position relationship integrating into the feature vector. The recognition effect has been improved greatly using this algorithm to recognize oracle characters on the neat bone rubbings. At the same time, it could supply some help to recognize words of the neat handwriting instruments.

Keywords: *oracle bone inscription, neat characters, feature extraction, character recognition, relative address*

1. Introduction

In current, the total number of rubbings engraved with oracle bone inscriptions was more than 10 million in Shang dynasty, including more than 8000 neat writing oracle bone rubbings. The number of characters was 5000 on oracle bone rubbings, however the number of recognized characters was only more than 1000 [1].

At present there have been some recognition systems of oracle bone rubbings character in China, which can recognize the majority of characters on oracle bone rubbings. But for these systems, their recognition rate of these systems is low and recognition speed is slow, at the same time, they also require manual labors to assist recognition. The number of rubbings with neat writing is only more than 8000, but the number of characters in these rubbings is has accounted for about half of the total number of all oracle bone inscriptions [2]. Therefore, it is very important for the digitization of oracle bone to recognize neat characters on oracle bone rubbings. This algorithm mainly realized the efficient and rapid identification to neat characters on oracle bone rubbings.

2. Feature Extraction Algorithm

The basic task of feature selection and extraction is how to find out the most important characteristics from many features. So the first step of process for any recognition accomplished either by computer or manual is analyzing the effectiveness of the various features and selecting the most typical characteristics.

Good features should have some characteristics, such as distinguishable, reliability, independence, amount small, etc. [3]. That is, these features should have the obvious difference for different categories of objects and should be more similar for similar categories of objects. Each characteristic should be unrelated to each other, and the characteristics of high correlation should be combined to reduce the noise interference.

The business processes of characters recognition engineering on oracle bone rubbings can be divided into four modules: rubbing binary, characteristics segmentation, feature extraction of oracle bone inscriptions, comparison and recognition, as shown in Figure 1.

Received January 8, 2013; Revised June 15, 2013; Accepted June 25, 2013

The feature extraction of oracle bone inscriptions is accomplished by the flowing processes. First, the characters of segmentation binary would be normalized, second, the strokes of normalized characters would be thinned, then the coarse meshing features and point features of thinning characters would be extracted respectively, and finally, these features would be superimposed to form the final composite features. As shown in Figure 2.

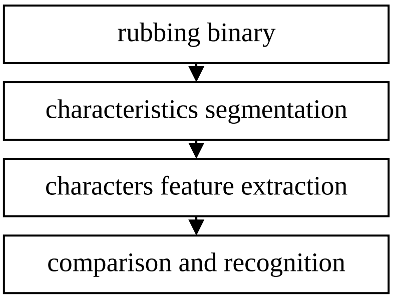

rubbing binary
characteristics segmentation
characters feature extraction
comparison and recognition

Figure 1. The Recognition Flow Chart of Oracle Bone Inscriptions

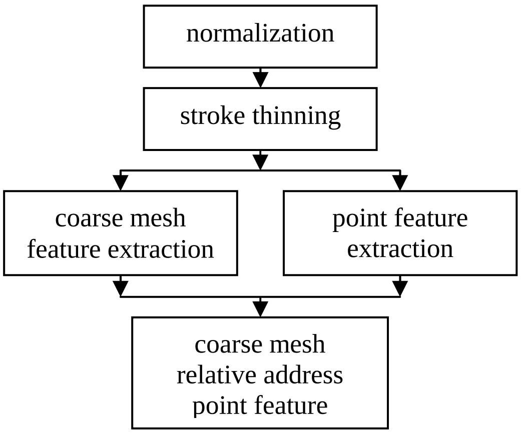

normalization
stroke thinning
coarse mesh feature extraction
point feature extraction
coarse mesh relative address point feature

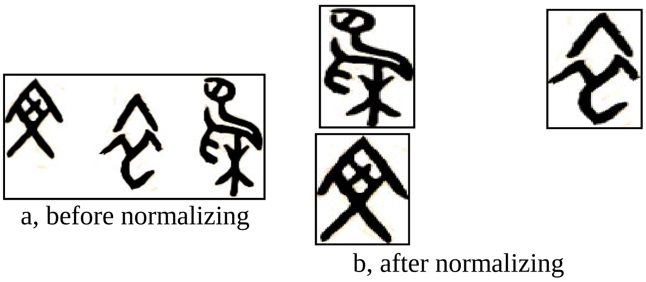
Figure 2. The Feature Extraction Algorithm Process of Oracle Bone Inscriptions

2.1. Normalization

Before features were extracted, first rubbings would binary processing, and then characters would be processed using the normalization method. For original oracle bone inscription, supposing the range of height was [ha, hb], the range of width was [wa, wb], and the transformed characters would be represented by matrix C (i, j), the barcentric coordinates (X, Y) of characters as shown in formula 1 and 2.

$$X = \frac{\sum_{i=ha}^{hb}\sum_{j=wa}^{wb} i \cdot c(i,j)}{\sum_{i=ha}^{hb}\sum_{j=wa}^{wb} c(i,j)} \quad (1)$$

$$Y = \frac{\sum_{i=ha}^{hb}\sum_{j=wa}^{wb} j \cdot c(i,j)}{\sum_{i=ha}^{hb}\sum_{j=wa}^{wb} c(i,j)} \quad (2)$$


a, before normalizing

b, after normalizing

Figure 3. Normalized Diagrams

Assuming that the character size is W × H, taking the center of characters (W/2, H/2) as the standard, moving the gravity of characters (X, Y) and other pixels, the normalized results would be gotten, as shown in Figure 3.

2.2. Stroke Thinning

Stroke thinning refers to the process through which width lines of the normalized word image would be translated into line with only one pixel width [4]. The pixel lines of thinning algorithm must be located in the center of the original line width to ensure the stability of topological structure for original oracle bone inscription. Image thinning not only eliminates the personalized features of handwritten but also keeps basic topology of oracle bone inscription invariant. So, these laid a foundation for feature extraction of oracle bone inscription. Thinning effect is shown in Figure 4.

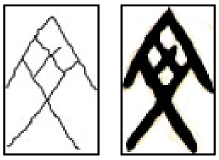


Figure 4. Thinning Effect

Noted: In order to make the thinned image in the grab more clearly, the refining width of the second figure is 5 pixels rather than 1 pixel.

2.3. Coarse Mesh Feature Extraction

The coarse mesh features belong to the local feature of statistical characteristics. It divided characters of binary image into A× B dimensional network and statistics the number of particular pixels of each grid.

Let g (I, J) shown in formula (3) represented the binary image of the W×H size.

$$g(i,j)=\begin{cases}0(withepixel)\\1(blankpixel)\end{cases} \tag{3}$$

Each grid can reflect a part of features of oracle bone inscription. In the recognition stage, each grid would be combined as statistic features of oracle bone inscription.

Coarse mesh features reflect the overall shape distribution of oracle bone inscription, but the position of character must according to requirements. Because the recognition principle of coarse mesh feature extraction algorithm is comparison would be carried out between recognizing characters and corresponding grid of standard grid, after normalization the recognizing character must is same with standard word of library in direction layout of position. That is, grid address of coarse mesh feature extraction is absolute address and comparison operation has completely dependent on absolute address. The great influence would be produced on the recognition effect if the recognizing words changed (tilt, shift and so on)[5,6].

2.4. Point Feature Extraction

The stroke of thinning oracle bone inscription is one track formed by connecting pixel points. These points can be divided into three categories according to positional relationship with surrounding points: end point, connection point, crossing point. For crossing points, the minimum intersection is three, theoretical the maximum is eight. For all points, the number of connection points is the most, but they are not significant for recognition, while end points and crossing points have important significance for the oracle bone inscription recognition.

In theory, there should be a kind of point: isolated point. That is, the number of surrounding points is zero. In reality, this kind of point does not exist in topology analysis of oracle bone inscription, because any stroke is not corresponding to only a pixel point in the refining process. The dots of the strokes would be corresponding to track with a pixel point [7, 8]. Therefore, each point of the thinning image would belong to one of three points.

Any pixel can be represented using 3 x 3 matrixes shown in Figure 5. The center point of matrix is the analyzing point, and around8 points are its surrounding pixels.

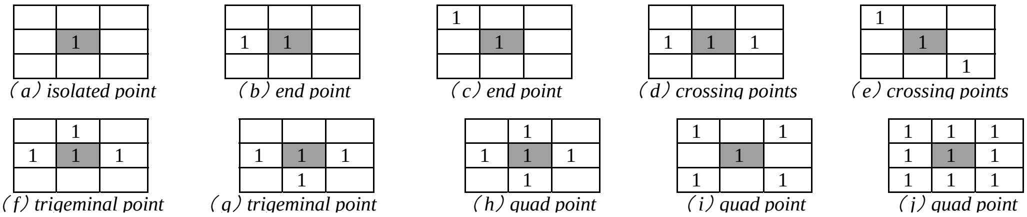
P1	P2	P3
P8	P0	P4
P7	P6	P5

Figure 5. Point Feature Matrix

The value of point feature can be calculated by formula (4).

$$PF = \frac{1}{2}\left[\sum_{k=2}^{8} | P_k - P_{k-1} | + | P_1 - P_8 |\right] \quad (4)$$

If PF=0, then this point is an isolated point but is not exist in topology analysis, as shown in Figure 6 (a). If PF=1, then this point is an end point, as shown in Figure 6 (b) and 6 (c). If PF=2, then this point is a connecting point, as shown in Figure 6 (d) and 6 (e). If PF=3, then this point is a trigeminal point of crossing points, as shown in Figure 6 (f) and (g). If PF=4, then this point is a quad point of crossing points, as shown in Figure 6 (h) and 6(i). if PF=8, then this point is a eight fork point of crossing points but only a theoretical value. As shown in Figure 6 (j).



(a) isolated point (b) end point (c) end point (d) crossing points (e) crossing points (f) trigeminal point (g) trigeminal point (h) quad point (i) quad point (j) quad point

Figure 6. Point Feature Legend

The specific implementation measures of point feature extraction method include two parts. First, point feature of each pixel would be determined by lattice progressive scan to binary refining image. Then, the numbers of all end points and crossing points in whole lattice diagram and their relative positions would be recorded. These would be treated as text features to analyze oracle bone inscription.

For the hand carved body, such as oracle bone word, the arbitrary is very large. So, the point feature extraction algorithm abandoned the dependence of absolute address, the recognition rate is low.

2.5. Coarse Mesh Relative Address Point Feature Extraction Algorithm

As mentioned earlier, because grid address of coarse mesh feature extraction is absolute address and comparison operation has completely dependent on absolute address, the great influence would be produced on the recognition effect if the recognizing words changed (tilt, shift and so on). While the recognition rate is low using point feature extraction algorithm, because the relative position of oracle bone inscription is very large. In Coarse mesh relative address point feature extraction algorithm, point feature was introduced into the course mesh feature, the absolute address was converted to relative address, point features have been changed grid and position relationship integrating into the feature vector. The recognition effect has been improved using this algorithm to recognize oracle characters on the neat bone rubbings.

First, coarse mesh feature value of thinning binary image would be extracted and the binary feature value of each grid pixel would be recorded. As shown in formula (5).

$$f(i,j)=\{x \mid 0 \le i \le w, 0 \le j \le H, x \ge 0\} \tag{5}$$

Second, the point feature of all points which value is 1 would be recorded by scanning progressively image [9, 10]. Then, the point value features would be introduced into the coarse mesh feature, which referred to the number of various types of feature points in each grid. Point feature can be represented using a multi-dimensional vector, as shown in formula (5).

$$P=(pf1, pf2, pf3, pf4, pf5, pf6, pf7, pf8) \tag{6}$$

Among them, pf1 represented end point, pf2 represented connection point, and pfn represented n fork point of the crossing point.

Note that, in the coarse grid feature and point feature, the relationship shown in formula (7) is true.

$$x = pf1 + pf2 + \cdots + pf8 \tag{7}$$

The relative position would be calculated between each feature point with the gravity according to the formula (8).

$$cf = pf - pg \tag{8}$$

Among them, PG focus feature vector of the gravity.

Then the relative position is introduced to the mesh feature values. The positional relationship may also be represented by a multidimensional vector, as shown in formula (9).

$$L=(cf1, cf2, cf3, cf4, cf5, cf6, cf7, cf8) \tag{9}$$

In other words, the grid feature not only recorded binary features, but also joined the point features. In this case, the grid feature can also be represented by a multidimensional vector, as shown in formula (10).

$$G=(x,P,L) \tag{10}$$

Comparison of coarse mesh feature would be processed by comparing G value of corresponding grid. Oracle bone inscription also is turned that shown in the formula (11).

$$\mathrm{ch(i,j)}=\{\mathrm{G} \mid 0 \le \mathrm{i} \le \mathrm{w}, 0 \le \mathrm{j} \le \mathrm{H}, \mathrm{G}=(\mathrm{x},\mathrm{P},\mathrm{L})\} \tag{11}$$

3. Experiment

The objects of data analysis are ten pieces of the neat oracle bone inscription rubbings in this experiment, and 1316 oracle bone inscription have been involved. For the 10 pieces of rubbings, recognitions have been processed separately using coarse mesh feature extraction algorithm and the improved mesh point feature extraction algorithm, and a better recognition effect has been gotten. Table 1 shown recognition data of five pieces of these rubbings.

For these 10 pieces of rubbings, the whole word recognition rate is 78.83% and the accuracy rate is 78.83% using coarse mesh feature extraction algorithm, while the whole word recognition rate is 70.84%% and the accuracy rate is 88.57% using point feature extraction algorithm. The average recognition rate increased by 22%, and the accuracy rate increased by 10%. As shown in Table 2.

Table 1. The Recognition Result of Partial Rubbings

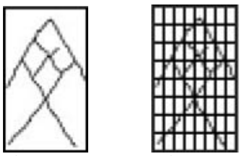
rubbing	comparison index	coarse mesh feature extraction algorithm	mesh point feature extraction algorithm
1	character recognition rate	53.62%	77.87%
	the correct rate	82.63%	91.35%
2	character recognition rate	50.21%	69.33%
	the correct rate	79.24%	88.62%
3	character recognition rate	32.68%	60.37%
	the correct rate	71.95%	83.47%
4	character recognition rate	63.53%	79.60%
	the correct rate	88.05%	92.56%
5	character recognition rate	46.58%	72.64%
	the correct rate	77.29%	89.81%

Table 2. The whole Recognition Data

comparison index	coarse mesh feature extraction algorithm	mesh point feature extraction algorithm
character recognition rate	48.26%	70.84%
the correct rate	78.83%	88.57%

Following, the specific experimental data sets has been given taking a rubbing of them an example.

The number of complete oracle bone inscription contained in this rubbing. One of oracle bone inscriptions was shown in Figure 7 (a). This word was translated for 8 ×8 grid, as shown in Figure 7 (b).



(a) Before meshing (b) After meshing

Figure 7.The Experimental Oracle Bone Inscription

0	0	0	0	6	0	0	0
0	0	2	6	0	8	0	0
0	2	4	4	9	1	8	0
1	3	7	6	2	9	5	8
4	0	0	7	5	1	0	0
0	0	1	7	9	0	0	0
0	3	5	0	0	8	0	0
5	1	0	0	0	0	7	0

Figure 8.The Coarse Grid Feature Matrix

3	3	0	0	0	0	0	0
0	0	0	0	0	0	0	0
2	3	2	1	0	1	0	0
0	2	0	0	0	0	0	0
1	3	1	0	0	0	0	0
0	3	2	1	3	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Figure 9. The Point Feature Matrix

1	0	-2	-1	0	-1	0	0
-2	-3	-2	-1	0	-1	0	0
0	0	0	0	0	0	0	0
2	1	2	1	0	1	0	0
1	0	1	1	0	1	0	0
2	0	0	0	-3	0	0	0
2	3	2	1	0	1	0	0
2	3	2	1	0	1	0	0

Figure 10. The Position Relationship Matrix

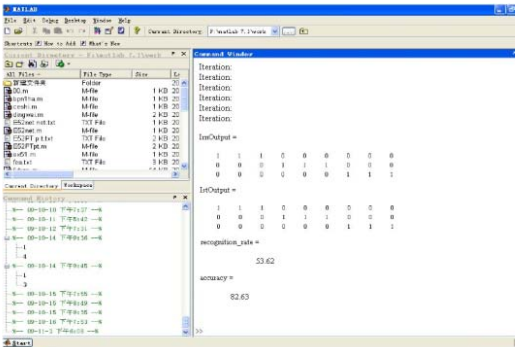
The coarse grid feature matrix of this word has been shown in Figure 8.

Taking the fifth column of this grid an example, the point feature matrix has been shown in Figure 9.

The position relationship matrix of this column has been shown in Figure 10.

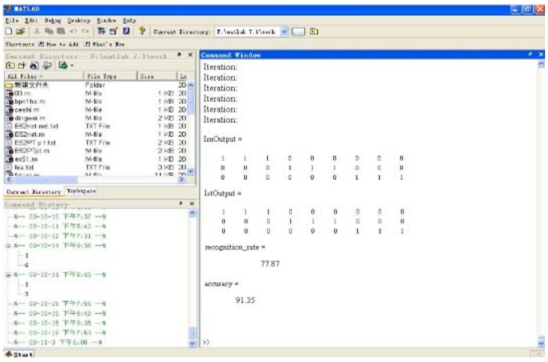
The recognition results of rubbings have been gotten using coarse mesh feature extraction algorithm, as shown in Figure 11. The recognition rate is 53.62%, and the accuracy rate is 82.63%.

The recognition results of rubbings has been obtained using mesh point feature extraction algorithm, as shown in Figure 12. The recognition rate is 77.87%, and the accuracy rate is 91.35%.



```
Iteration:
Iteration:
Iteration:
Iteration:
Iteration:
Iteration:

ImOutput =

     1     1     1     0     0     0     0     0     0
     0     0     0     1     1     1     0     0     0
     0     0     0     0     0     0     1     1     1

IstOutput =

     1     1     1     0     0     0     0     0     0
     0     0     0     1     1     1     0     0     0
     0     0     0     0     0     0     1     1     1

recognition_rate =

          53.62

accuracy =

     82.63

>>
```

Figure 11. The Recognition Results of Coarse Mesh Feature Extraction Algorithm



```
Iteration:
Iteration:
Iteration:
Iteration:
Iteration:
Iteration:

ImOutput =

     1     1     1     0     0     0     0     0     0
     0     0     0     1     1     1     0     0     0
     0     0     0     0     0     0     1     1     1

IstOutput =

     1     1     1     0     0     0     0     0     0
     0     0     0     1     1     1     0     0     0
     0     0     0     0     0     0     1     1     1

recognition_rate =

          77.87

accuracy =

     91.35

>>
```

Figure 12. The Recognition Results of Grid Point Feature Extraction Algorithm

Based on the experiment results, the recognition effect has been improved obviously. The recognition rate increased from 53.62% to 77.87%, and the correct rate is improved from 82.63% to 91.35%. As shown in Table 3.

Table 3. The Comparison Result between Original Algorithm and Improved Algorithm

comparison index	coarse mesh feature extraction algorithm	mesh point feature extraction algorithm
character recognition rate	53.62%	77.87%
the correct rate	82.63%	91.35%

These 10 pieces of rubbings have been selected randomly from more than relatively neat 8000 oracle bones, so they are very strongly general. That is to say, the coarse mesh point feature extraction algorithm is universal to recognize oracle bone inscription on the neat rubbings.

4. Conclusion

Compared with the traditional coarse grid feature extraction and point feature extraction, the coarse mesh point feature extraction algorithm is more suitable to recognize oracle bone inscription on neat rubbings, which can have greatly improved various performances. At the same time, it can be used to recognize neat regular handwritten document, and we can get quite good effect. However, the recognition effect is not very good for scattered text. Then we would improve this algorithm to increase application, and the main breakthrough points are segmentation and normalization of oracle bone inscription of rubbings.

Acknowledgements

This work was supported by the National Natural Science Funds of China (NO: 60973051, the editing and encoding technology of oracle bone inscriptions).

References

[1] Huayan Jiang, Qian Zhu. The classification and source of the Yin Ruins Oracle said. *Journal of Wuhan Institute of Technology.* 2010; 25(6): 35-39.
[2] Shangting Lu, Yaoqing Liu. Opened the mysterious veil of Oracle. *Consultative Forum*. 2005; 33(8): 52-55.
[3] Deren Liu, Zailin Ge. The contact line electric locomotive wear based on digital image processing. *Modern manufacturing engineering*. 2006; 21(3): 12-14.
[4] Fan Han, Yongmei Zhang. Improved method of Chinese characters recognition and Rosen thinning algorithm. *Journal of North China Institute of Technology*. 1997; 23(1): 35-40.
[5] Shongtao Liang. Research of video caption recognition based on feature compensation. *Computer applications and software*. 2010; 31(11): 22-26.
[6] Jingzhong Wang. Study on Application of normalization algorithm in character recognition system. *Computer applications and software.* 2011; 35(3): 41-43.
[7] Xuefang Zhu, Houjie Bi. A method of a variety of printed Chinese characters recognition. *Journal of Nanjing University of Posts and Telecommunications*. 2012; 7(4): 2-5.
[8] Qichun Wang, Guangli Guo, Jianfen Za. Research on adaptive parameter extraction operator point feature based on image gray. *Journal of combustion science and technology*. 2012; 32(6): 52-55.
[9] Wenbing Li, Ping Feng. Displacement measurement based on digital image processing. *Journal of Zhejiang University of Technology.* 2012; 23(6): 13-16.
[10] Wei M, Xu J, Yun H, et al. Ontology-based home service model. *Computer Science and Information Systems*. 2012; 33(2): 813-838