Simulation of Stroke Number and Stroke Order Free Online Kanji Character Recognition

Jianxiong Dong

Character Recognition Laboratory
The Institute of Automation, Chinese Academy of Sciences Report A01

February 1998

Simulation of Stroke_Number and Stroke_Order Free Online Kanji Character Recognition

JianXiong Dong
(email:djx@hw.ia.ac.cn)

Character Recognition Laboratory

Institute of Automation, Chinese Academy of Sciences
February 1998

Abstract: This report describes simulation and improvement of Toru Wakahara online Recognition Algorithms. Toru Wakahara's method primarily solves the one_to_one stroke correspondence problem with both the stroke number and stroke order variations common in Chinese and Japanese handwriting. In this paper, it proposes two kinds of complementary algorithms: one dissolves excessive mapping and the other dissolves deficient mapping. Their joint use realizes stable optimal stroke correspondence without combinatorial explosion. Also, three kinds of inter_stroke distances are devised to deal with stroke concatenation or splitting and heavy shape distortion.

Table of Contents

1. General

For large alphabet language, like Japanese and Chinese, handwriting input using an online recognition technique is essential for input accuracy and speed. There exist difficulties in hand-written kanji and Chinese character recognition:

(1) A great number character categories

(2) Complicated character shapes

(3) Existence of similar character shapes

(4) A large variety of handwritten character shapes

In current character recognition, there are two trends: one is statistical recognition. It extracts the high dimension feature such as direction element Feature, which obtains 97.76% recognition rate for ETL9B[1]. The other is nonlinear pattern matching such as elastic matching [2], Dynamic Programming[3], LAT algorithms[4]. Moreover, stroke is the most basic unit of shape representation and discrimination[5] , online recognition techniques have the great advantage of being able to extract the stroke information.

2. The principle of this algorithms

The most important part of Toru wakahara's algorithms is Stroke Correspondence Determination. It effectively overcomes combination problem for searching true optimal correspondence, giving a good correspondence. As follows, the problem solved by this method is described.

Denote two sets of strokes by $R=\{R_1,...,R_m\}$ and $T=\{T_1,...,T_n\}$ $(M \geq N)$,then calculate a set of M times N interstroke distance between R and T:

$$d_{ij} = d(R_i, T_j) \quad (1)$$

$where.(1 \leq i \leq M, 1 \leq j \leq N)$

The optimal N-pairs stroke correspondence problem is the formulated as:

$$\sum_j d_{u(j)j} = \sum_j d(R_{u(j)}, T_j) \to \min .for.mapping..u \quad (2)$$
$$where..u(j) \in \{1,2,...,M\}, u(j) \neq u(k) for. j \neq k..and$$
$$\sum_j .the.summation. for. j = 1,2,..., N.$$

This is an assignment problem (AP). The AP has been solved by many methods including the Hungarian method and Simplex method [6,7,8]. AI-JIA HSIEH [9] implement the Hungarian method for online handwritten chinese recognition. It uses stroke_segment as the matching unit. But it is not good for cursive handwriting because stroke_segment is not stable. Toru Wakahara's algorithms are more robust.

3. Practical advice for obtaining good recognition rate

As for as Toru Wakahara's method is said, the key is stroke correspondence. First, the measure distance of stroke similarity is very important. If the measure distance is not designed well, the recognition rate is not high. Second, a problem is how to select reference effectively. Third, improving the classification accuracy is indispensable to facilitate candidate selection. One good choice is to reinforce the ability of clustering methods and statistical discriminate functions to the fullest extent in OCR technique. The improved Directional Element Feature of Ning Sun [1] and Teuvo Kohenon's SOM(Self_Organizing Mapping)[10] and LVQ [11] are all good method of coarse classification.

4. Installation of the program package

4.1 Installation in win95

At First, you should use winzip to extract the compressed file document.zip. Second, you should copy exectuting files into a directory. The steps are as follows:

```
md C:\dong2
copy dictionary.bin  C:\dong2
copy test.dat C:\dong2
copy mathtool.reg C:\dong2
copy strokeorderfree.exe C:\dong2
copy mlibvc0.dll C:\windows\system
```

4.2 Software and hardware environment requirement

Visual++5.0 in Win95 implements the program package. It must be executed in win95 or Windows NT. Because dictionary requires about 6M memory, the RAM should be more than 32M.

5. File formats

5.1. Data file format

The input data is stored in binary form as a list of entries.

The online data format for a character is as follows:

.**C_i** : Category code of *i_th* data (unsigned short)

.**S_i** : The number of strokes of *i_th* data (unsigned short)

.**P_{ij}** :The number of points of *j_th* stroke in *i_th* data (unsigned short)

.**x_{ijk}**: *k-th* point's x-axis value of *j_th* stroke in *i_th* data (short)

.**y_{ijk}**: k-th point's y-axis value of j_th stroke in i_th data (short)

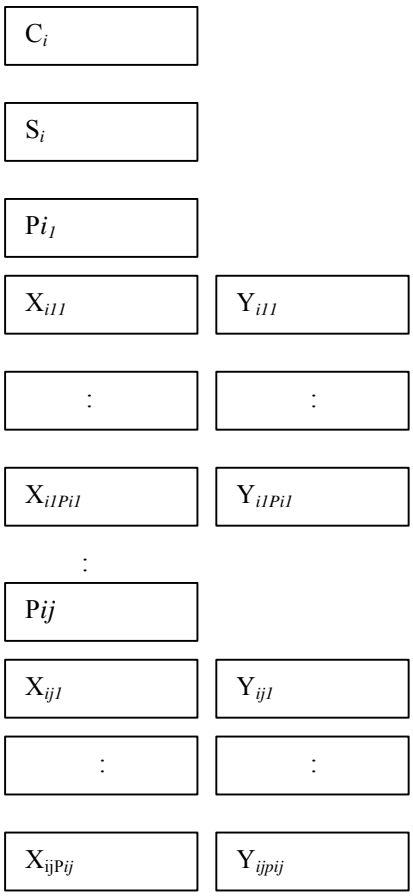


Fig.1 Online Data Format

5.2 Result file format

The output data is stored in binary form as a list of category code. For example, if 10 candidate output, top 10 category codes in increasing order of MeasureDistance are stored.

5.3 Error file format

It is the same as 5.1

6. Application of this package

6.1 Create dictionary

Click the menu bar **Procedure** of main menu, then popup menu appears. Select menu Item **Create Dictionary**. This process takes about two days in Pentium II . The flowchart of how to **Create Dictionary** is described in Readme.Doc

6.2 Display Chinese character after normalization

Click the menu bar **Procedure** of main menu, the popup menu appears. Select **Display Chinese**. A dialog appears, then you select button **Browse,** select your normalized file. You could compare the moment normalized character with both original character and nonlinear normalized character. Moreover, it can display stroke order of one character.

6.3 Demonstration of recognition

Click the menu bar **Procedure** of main menu , the popup menu appears, select **Recognition**. One dialog appears. There are three ways for your selection.

(1). Select radio button (**Recognition**),then click button **Browse**, An **OpenFile Dialog** appears, select your recognized file, then click button **Begin**. The program begin to recognition , the input sample and recognized result both appear in dialog.

(2). Select radio button (**Demonstration**) and radio button **Step_into**, then click button **Browse**, A **OpenFile Dialog** appears, select your recognized file. Click button **Begin**, finally click button **Step_Into**, you can observe the stroke matching of each loop of Excessive Mapping Matching, Deficient Mapping Matching, Advanced Linkage.

(3). Select radio button (**Demonstration**) and radio button **Step_Over**, then click button **Browse.** An **OpenFile Dialog** appears, select your recognized file. Click button **Begin,** finally click button **Step_Over**, you can observe the stroke matching of the whole

function of Excessive Mapping Matching, Deficient Mapping Matching, Advanced Linkage.

6.4 Test

Click the menu bar **Procedure** of main menu, then popup menu appears. Select **Test**, a dialog appears. Click the button **Begin**. After Test, the filename of Result file is "result.dat".

6.5 Statistical recognition rate

Click the menu bar **Procedure** of main menu, then popup menu appears. Select **Statistic Result**, a dialog appears, select the button **Begin.** After a while , Recognition Rate , 5rd Candidate Rate ,10 Candidate Rate are given.

7. Description of some parts of the program of this package

7.1 Normalization

The standard two-dimensional discrete moments m_{pq} of image *f(x, y)*

$$m_{pq} = \sum_{y}\sum_{x} f(x,y)\, x^{p}\, y^{q} \quad \text{.....(3)}$$

will vary for a given shape depending on the spatial position of the object . Translation invariance is obtained using the central moments[12]

$$u_{pq} = \sum_{x}\sum_{y} (x-\bar{x})^{p}\,(y-\bar{y})^{q} f(x,y) \quad \text{.....(4)}$$

where

$$\bar{x} = \frac{m_{10}}{m_{00}}$$

$$\bar{y} = \frac{m_{01}}{m_{00}}$$

They indicate the center of the object. Scaling invariant central moments are obtained by the

$$c_{pq} = \frac{u_{pq}}{(u_{00})^{r}}$$

$$r = \frac{p+q}{2} + 1 \quad \text{.....(5)}$$

$$p + q \geq 2$$

normalization [12].

In practice, according to the characteristic of online recognition, the following equations are given.

$$\bar{x} = \frac{\sum_{i=1}^{n} d_i(x_{i-1} + x_i)}{2 \times \sum_{i=1}^{n} d_i} \quad \text{.....(6)}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} d_i(y_{i-1} + y_i)}{2 \times \sum_{i=1}^{n} d_i} \quad \text{.....(7)}$$

$$m_{00} = \sum_{i=1}^{n} d_i \quad \text{.....(8)}$$

where d_I *is the length of I-th stroke segment . n is the number of stroke segment*

m_{10}, m_{01} are given in the document chinese.cpp

7.2 Feature point selection

Second moment of the pattern around the origin is set at the predetermined value 400. Feature point selection is performed using stroke segment length value equal to 10 in author's paper. However, By my experiment, stroke segment length equal to 4 is better than 10. If stroke segment length is smaller, recognition rate is not improved. In opposite, it increases computation time greatly.

7.3 Measure of stroke similarity

In Toru wakahara's method, measure of input stroke and reference stroke is defined by inter-stroke distance. Though it is good measure, it has a fault that doesn't describe the shape of the stroke. Moreover, it doesn't take consideration to structure information such Topology feature as hole, fork and so on.

7.4 Stroke correspondence

In the excessive mapping matching and deficient mapping matching algorithms, the importance of different stroke of one category is the same. In fact, some strokes are not important for a correct recognition. In opposite, it might result in misrecognition. Therefore, a new concept regarding stroke rank is necessary. The essence of stroke rank information is to flexibly represent

the role of each stroke, when written in the cursive style, while using templates formed in the block style [13]. This procedure may be done manually. But it is good to form stroke rank information by computer in training. Through stroke rank information, a weighted measure of strokes is also defined. Detailed information is described in a paper[14].

7.5 How to choose reference in dictionary

In Toru wakahara's method, "prototype" means a template generated by averaging x-y coordinate values of feature points over learning samples for each character. But it needs correct stroke-number and correct stroke order learning samples. For correct stroke order, there exist some difficulty in practice. In the package, good template is generated by using excessive mapping, deficient mapping and Linkage method. In addition, several template for each character is stored in dictionary. Details about these are described in Readme.Doc

7.6 Inter-pattern distance

In author's paper, the permissible maximal numbers of stroke splitting and concatenations per character were set at α=2 and β=10, respectively. In other words, it permits rejection. But in my experiment there is no rejection. In inter-pattern distance , I introduce a weight max(M,N)/min(M,N) .The inter-pattern distance of fine classification is defined as:

$$D_2 = \max(M,N)/\min(M,N).\sum_i \gamma_i.d_f(R_i^*,T_i)/\min(M,N) \quad (9)$$

Where M and N are stroke numbers of input pattern and reference pattern respectively.

7.7 selective stroke linkage method

In author's paper, selective stroke linkage method is reasonable in most cases. But in some cases, I thought it has a fault. For example

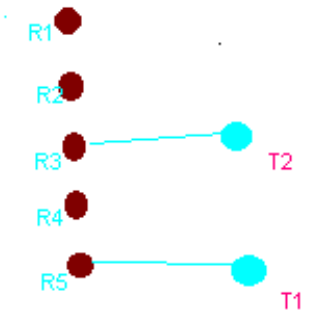
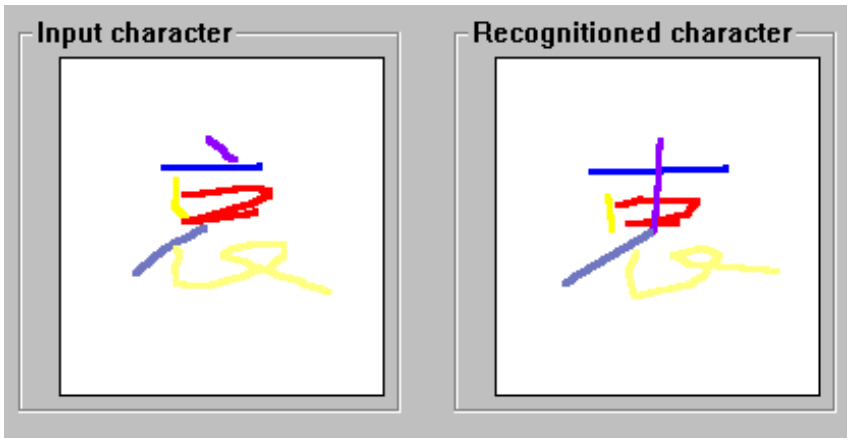


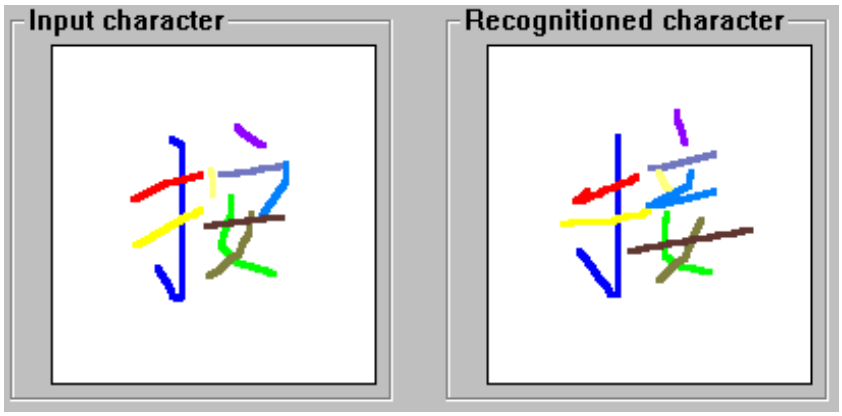
Fig 1. Selective Stroke linkage

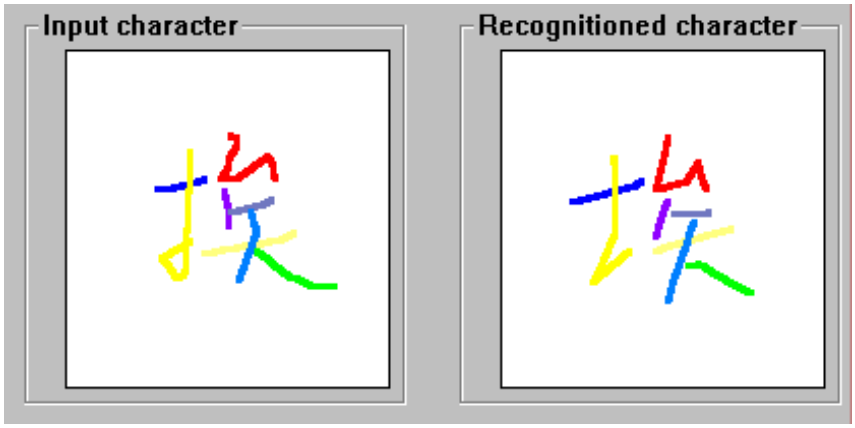
In author's method, R1, R2, R3 are concatenated. Whether R4 is concatenated into R3 or R5 is related with R3, R4, R5, T2, T1. But my method is followed: First R1,R2,R3 are concatenated into a stroke R123, then whether R4 is concatenated into R123 or R5 is determined by R123, R5, R4, T1,T2. In my experment, I compare their recognition rate. My improved method increases recognition rate by 0.7%.

8. Analysis of misrecognition cause

There are two main cause of misrecognition in the cursive style. The first, shown in Fig.2(a), is the existence of similarly shaped but different chinese characters. For this problems Toru Wakahara thinks that it is possible to devise a kind of nonlinear shape normalization technique that will enhance the discriminative shape differences. But a nonlinear normalization has been used to test, recognition rate is not improved. Therefore I think that a similar measure should be devised. Combined with the topology feature of character image, it maybe improves the ability of discrimination. The other is excessive shape distortion as shown in Fig..2(b). This problem is more difficult because the cursive style strongly depends on the individual's writing style. Hence, it is promising to develop a learnable or custom_built online handwriting recognizer for each individual [5].


Input character
Recognitioned character


Input character
Recognitioned character


Input character
Recognitioned character

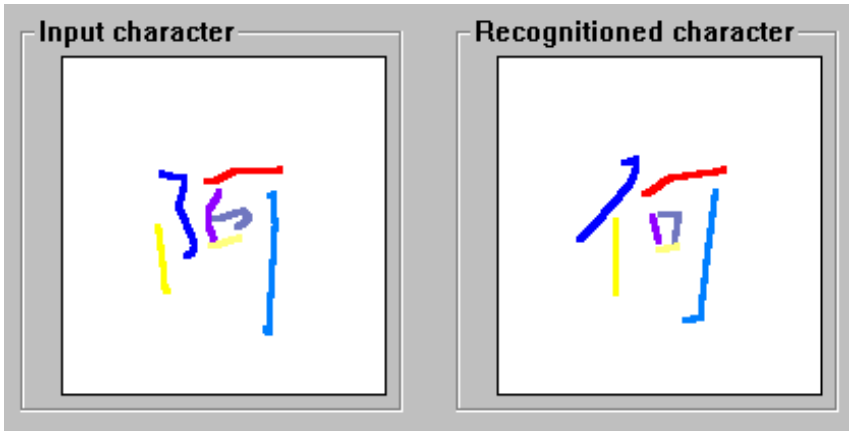
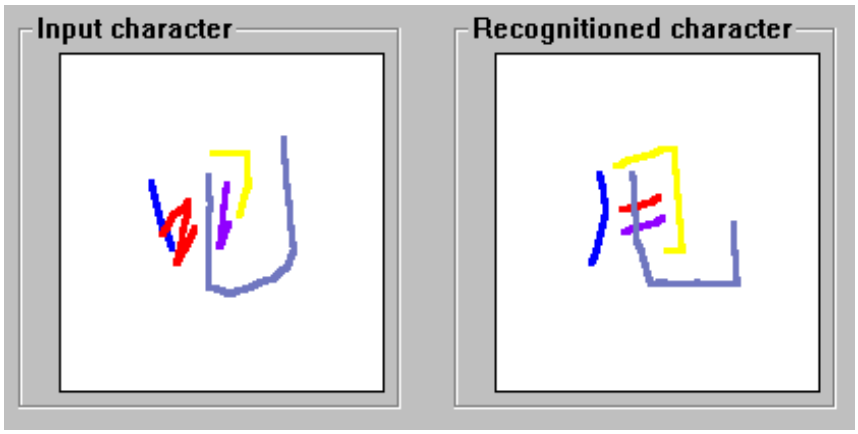
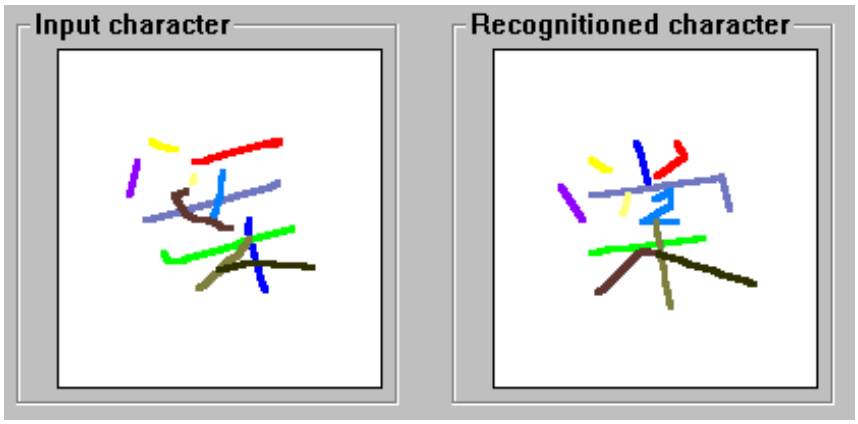

Input character
Recognitioned character

Fig.2(a)


Input character
Recognitioned character



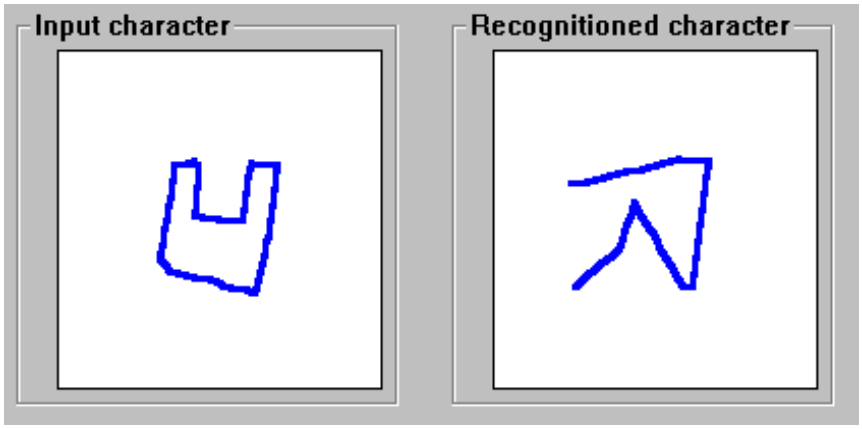


Fig.2(b)

9. Training and test condition

Training and test are based on online simplified Chinese database of Character Recognition Laboratory. There are 3213 categories in training and test.

Training sample: 100 patterns for one category

Testing sample : 30 patterns for one category

10. Result of Recognition

Fig.2 shows the obtained cumulative classification rates by stroke matching approach through moment normalization, where the cumulative classification rate at the *n-th* order is the rate that the correct category is included within the top *n* candidates.

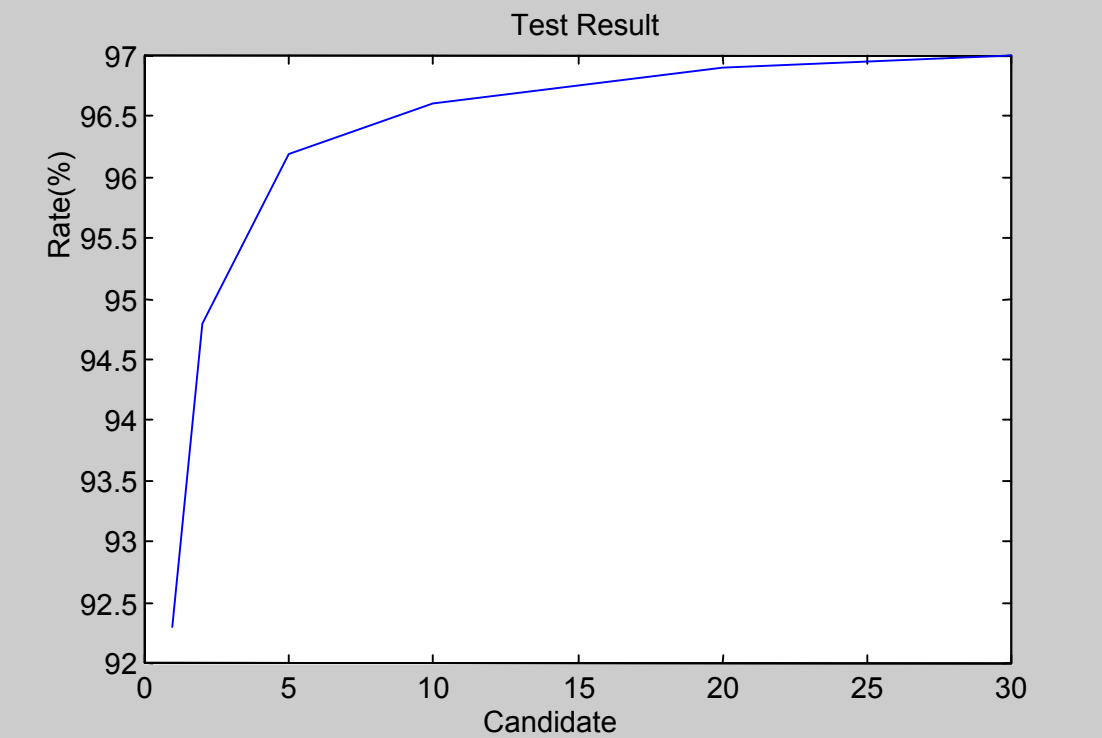


Table 1 shows the exact result of recognition rate.

	1	2	3	4	5	10	20	30	50	100	200	300	500
Rate	92.3	94.8	95.6	95.9	96.2	96.6	96.8	96.9	97.1	97.7	99.7	99.8	99.8

Table 1 cumulative recognition rate(%) of moment normalization

The following Table is recognition result through nonlinear normalization [15].

	1	2	5	10
Cumulative rate	92.4%	94.7%	96.8%	97.0%

Table 2. Result of recognition Using nonlinear normalization

11. Reference

[1] N.sun, M.Abe and Y.Nemoto, "a handwritten character recognition system by using improved directional elemental feature and subspace method", Trans, IEICE Japan, J78-D-II(6), 922-930(1995) (in Japanese)

[2] K.Yoshida and H.Sakoe " online handwritten character recognition for a personal computer system ", IEEE Trans ,Consumer Electron, Vol.CE-28, pp202-209, Aug.1982

[3] H.Yamada, "contour DP matching method and its application to handprinted chinese character recognition " Proc .7th ICPR. pp389-392, 1984.

[4] Toru Wakahara "Shape Matching Using LAT and its Application to Handwritten Numberal Recognition ", IEEE PAMI Vol.16, No.6, June 1994

[5] Toru Wakahara, Akira Suzuki, Naoki Nakajima, Sueharu Miyahara ,and Kazumi Odaka, "Stroke –Number and Stroke-Order Free Online Kanji character Recognition as One-to-One Stroke Correspondence Problem" , IEICE Trans. INF& SYST, Vol.E79-D, No.5, MAY 1996

[6] E.L.Lawler, Combinatorial Optimization: Networks and matroids . Holt. Rinehart&Winston , New York(1976)

[7] C.H. Papadimitrion and K.Steiglitz, Combinatorial Optimization: Algorithms and Complexity. Prentice –Hall , Englewood Cliffs , New Jersey(1982)

[8] H.W.Kuhn, The Hungarian method for the assignment problem. Naval Research Logistics Quarterly 2, 83-97(1955)

[9] AI-JIA, HSIEH, KUO-CHIN. Fan Tzu-I Fan " bipartite weighted matching for online handwritten chinese character recognition", Pattern Recognition, Vol.28, No.2, pp143-151,1995

[10] Teuvo Kohonen, Self-Organizing Maps, Springer-verlag Heidelberg 1995

[11] Teuvo Kohonen ,Improved versions of Learning Vector Quantization (LVQ) in Proceedings of the International Joint Conference On Neural Networks, Pages I 545-550, San Diego, June 1990

[12] R.C.Gonzalez and R.E.Woods, Digital Image Processing Addison-wesley (1992)

[13] Toru Wakahara, Hiroshi Murase and Kazumi odaka," online handwritting Recognition ", Proceedings of the IEEE, Vol.80, No.7, July 1992

[14] Adam Krzyzak " Handwriting Recognition Using weighted Elastic Matching" E_mail: krzyzak@cs.concordia.ca

[15] J.Tsukumo and H.Tanaka, Classification of handprinted Chinese Characters using nonlinear normalization methods. Proc.9th Int. Conf. on Pattern Recognition. Rome, Italy.pp168-171, November (1988).

12. Acknowledgements

First of all, I would like to express my gratitude to Professor YingJian Liu, for supervising me during the simulation research. I am also indebted to Professor XianLi Wu ,Dr MinJing Li and associate general Liqing Zhang who give me good advice.

I would like to thank everyone else in character recognition group, including HongYu Que, GaoTao and so on. Finally, I also thank Toru wakahara that his papers provide me with good illumination.