On Chinese Character Structure(CCS) Recognition Based On Bayesian Learning

Tiancai Liang   Youguo Pi

College of Automatic Science & Engineering, South China University of Technology, Guangzhou 510641, P.R. China

Abstract

Chinese character structure(CCS) is too complex to be represented or recognized by computer. In previous literature simple grid was proposed to be representation tools for CCS. In order to check effectiveness of CCS representation based on simple grid, the method of CCS recognition based on Bayesian learning is proposed by this paper. In the method, after acquiring candidate separate location between radicals under the guidance of simple gird, the optimal separate location is determined by Naive Bayesian classifier. The method proposed is simple and easy-to-use, and has good extension. The effectiveness of method proposed is proved by experimental results.

Keywords: Chinese character structure representation, Feature exaction, Simple grid, Naive Bayesian, Recognition

1. Introduction

Chinese language was record through graphics composition[1]. According to certain combinational rule, Chinese character can be pieced with basic symbols. These basic symbols is Chinese character prototype(CCP), piecing rule is Chinese character structure. Chinese character prototype may be radical [2], and may also be strokes [3], depending on the specific application. The piecing rule of phonetic system is simple, with the first letter as a starting point arranged in chronological order, the location of the letters, shape and size will remain unchanged [4]. However, when using CCP to generate Chinese characters, the position and shape and size of CCP are changing, according to different CCS. In addition, CCS is rather complex and diverse, International Standard ISO / IEC 16046 give 12 ideographic descriptor of CCS : left to right, above to below, left to middle and right, above to middle and below, full surround, surround from above, surround form below, surround from left, surround from upper left, surround from upper right, surround from lower left, overlaid. In all such structures also contain other structures, such as the "耱" is left to right structure, but it is surround from left structure for its right part. Such a complex relationship makes it difficult to use the computer to described or recognize. This may be the main reasons for the existing Chinese information processing system used to character library but abandon piecing mode.

CCS is basic element for piecing Chinese character, and informationization of CCS is necessary conditions for piecing Chinese characters. CCS representation is organic composition of feature of CCS, CCS recognition is the best way to check effectiveness of CCS representation. Therefore, the feature extraction of CCS is a very important, only to extract integrity, making it easier for digital features, CCS representation or recognition with computer can be achieved smoothly. As a representational tools, simple grid was proposed to represent CCS[5],but how to judge effectives of feature extracted with simple grid is still a problem. In this paper, the method of CCS recognition based on statistical learning theory is proposed to check effectiveness of CCS representation based on simple grid .

This paper is organized as follow, CCS representation based on simple grid is given in section II, method of CCS recognition and its implement is presented in Section III, anyway, experiment based on method proposed and experimental result is discussed in Section IV, finally, conclusion is given in Section V.

2. CCS representation based on simple grid[6]

As for shape structure , radical is key element of Chinese character, and radical can delineated Chinese characters more clearly than stroke[5]. Therefore, the CCP should be radical, and the “basic radical” come from GF3001-1997 Information Processing GB13000.1 Character Set Chinese Radical Norms which was build by the Chinese National Language committees, was taken as CCP to study the CCS representation. Simple grid and its use is presented in pattern literature[6], Mathematical description of simple grid and application to CCS representation based on simple grid is given in literature[7].

As for CCS representation, 2×2 uniform grid and 3×3 uniform grid act as basic, the other act as grid expansion, as shown in Fig.1 .

$N \times N$ uniform grid can be represented as follows:

$$A_{N \times N} = \begin{bmatrix} A^N{}_{11} A^N{}_{12} \bullet\bullet\bullet A^N{}_{1N} \\ A^N{}_{21} A^N{}_{22} \bullet\bullet\bullet A^N{}_{2N} \\ \bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet \\ A^N{}_{N1} A^N{}_{N2} \bullet\bullet\bullet A^N{}_{NN} \end{bmatrix}$$

Landscape orientation reticle can be represented $r^N{}_i$, portrait reticle can be represented as $h^N{}_i$.

Use simple grid, described the structure of Chinese characters can be transformed into mesh the relationship between the operator [6]. Based on the simple grid structure of Chinese characters depicted easily digitized, computer architecture is compatible with the appropriate computer.

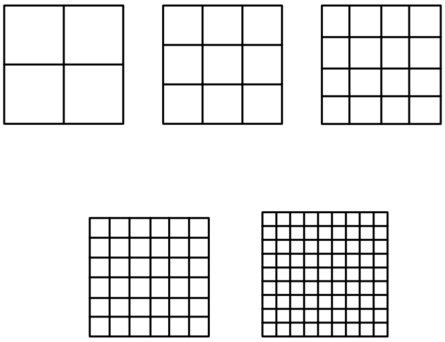

Fig.1: Simple grid and its extension

With simple grid, CCS representation can be can be transformed into operational relations between mesh A_{ij} [6]. Based on simple grid, CCS representation is easy to be implemented by computer.

3. Method of recognition based on Bayesian learning

Pattern recognition theory point out that pattern recognition is to match the feature of pattern [8]. Effectiveness of CCS representation refers to organize feature of CCS to represent CCS fully and clearly. Thus, CCS recognition is the best way to check effectiveness of CCS representation. CCS is defined by the location of radical used to piece Chinese character, anyway, these radical is independent. Therefore, CCS recognition refer to segment independent radical correctly. In this paper, a method based on statistical learning theory is proposed to study CCS recognition. Firstly, candidate segmentation position can be obtained from CCS representation based on simple grid, then Bayesian classifiers is used to determine the best candidate segmentation. Based on statistical learning theory, the model of CCS recognition can be divided into two parts, training and testing, as shown in Fig.2 .

3.1. Naive Bayesian classifier

Bayesian learning is a supervised learning method, based on Bayesian theorem, is implemented by solving maximum a posteriori probability under a priori probability and the conditional probability is known. Naïve Bayesian classifier[9]-[11] is one of the concrete implementation of Bayesian learning. Naive Bayesian classifier assumed that the feature vectors is relatively independent to decision variable. For the feature vector of the test samples.

$$X = [x_1, x_2, \bullet\bullet\bullet, x_d]^T$$

Conditional probability which belongs to the C_i category can be represented as follows:

$$\begin{aligned} P(C \mid X) &= P(X \mid C_i).P(C_i)/P(X) \\ &= \frac{P(C_i)}{P(X)} \prod_{j=1}^{d} P(x_i \mid C_i) \qquad (1) \end{aligned}$$

For each category, condition probability are calculated according to formula (1), the ultimate recognition results correspond to category which had maximal conditional probability . Even though the assumption that Naive Bayesian classifier is based on the assumption of independence is violated, Naive Bayesian classifier also shown considerable robustness and efficiency. It has been successfully applied to classification or clustering issues.

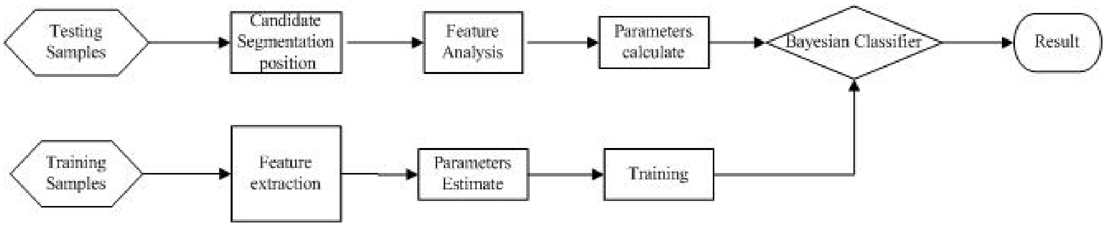

Fig2: Model of recognition.

3.2. Recognition based on naive Bayesian classifier

For CCS recognition, it can be assumed that there are n kinds of candidate results, corresponding to n kinds of CCS. For the same CCS, radical of Chinese character has m kinds of a candidate segmentation position, corresponding to m kinds of candidate segmentation results, and the effectiveness of segmentation is measured with d kinds of features. These d kinds of feature can be expressed as follows,

$$X_{\mathrm{j}} = [x_{j1}, x_{j2}, \bullet\bullet\bullet, x_{jd}]^T \, (\mathrm{j} = 1,2,\bullet\bullet\bullet,\mathrm{m})$$

When there are $\mathrm{C_i}\,(i = 1,2,\bullet\bullet\bullet,n)$ kinds of CCS, the probability to candidates segmentation position j can be represented as follows,

$$\begin{aligned} & P(X_{\mathrm{j}}, \mathrm{C_i}) \\ & = P(X_{\mathrm{j}} | C_i).P(C_i)/P(X) \\ & = \frac{P(C_i)}{P(X)} \prod_{\mathrm{k}=1}^{d} P(x_{jk} | C_i) \\ & (\mathrm{i} = 1,2,\bullet\bullet\bullet,\mathrm{n};\ \mathrm{j} = 1,2,\bullet\bullet\bullet,\mathrm{m}) \qquad (2) \end{aligned}$$

Certainly, the formula (2) not only need to determine a candidate structure types, but also need to identify the candidate segmentation position that is the most appropriate. Therefore, the maximum value of $P(X_{\mathrm{j}} | \mathrm{C_i})$ corresponds with the best segmentation position, while $\mathrm{C_i}$ corresponds with CCS which is recognition result.

3.3. Conditional probability estimation

Application of Naive Bayesian classifier need to estimate conditions probability. The feature of CCS are discrete, therefore, conditional probability of CCS recognition is discrete. According to assumption of independence, there is formula as follows

$$P(X_{\mathrm{j}} | C_i).= \prod_{\mathrm{k}=1}^{d} P(x_{jk} | C_i) \qquad (3)$$

Obviously, Naive Bayesian classifier conditional probability density function can be solved by estimating conditional probability density of feature, that is $P(x_{jk} | C_i)$. The conditional probability density of each feature can be calculated with the following formula,

$$P(x_{\mathrm{j_k}} | C_i) = [(\mathrm{z}_{\mathrm{jk}} + 1)/\mathrm{Z_i}]/(Z_i / Z) \qquad (4)$$

Where $\mathrm{z_{jk}}$ is numbers of training sample which category have feature $\mathrm{x_{jk}}$ and derive from $\mathrm{C_i}$ category , Where $\mathrm{Z_i}$ is numbers of training samples of $\mathrm{C_i}$ category , Z is numbers of all training samples.

4. Implement

For 6763 Chinese characters deriving from GB2312-80, left to right structure is the most proportion [12]. Therefore, , representation of left to right structure based on simple grid was taken as research object to illustrate the method proposed in section III. As only to study one structure, there is $n = 1$. The the maximum value of $P(X_{\mathrm{j}} | \mathrm{C}\ )$ should be calculated in order to obtain best segmentation location.

4.1. Representation of left to right structure based on the simple grid

The structure disciplinarian of Chinese character was analyzed completely through putting Chinese character into simple grid, as shown in figure 3.

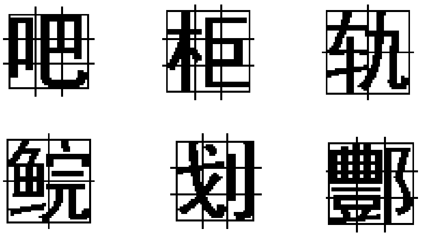

Fig.3: Analysis of CCS based on simple grid

Left to right structure of Chinese characters is formed by relatively independent left radical and right radical, that is a typical cross-vertical structure. The distribution of these relatively independent radicals in simple grid were shown in Figure 4.

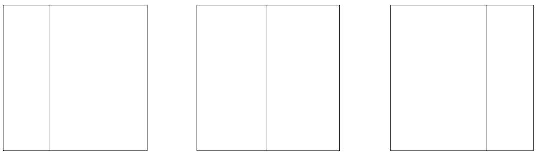

Fig.4: Distribution of left to right structure in simple grid sapce

From Fig3, the distribution in simple grid have three type , the representation of left to right structure can be expressed as follows,

(a) $\Omega_{C1}=\left(\bigcup_{i=1}^{3}A^{3}{}_{i1}\right)$； $\Omega_{C2}=\left(\bigcup_{i=1}^{3}A^{3}{}_{i2}\right)\cup\left(\bigcup_{i=1}^{3}A^{3}{}_{i3}\right)$

(b) $\Omega_{C1}=\left(\bigcup_{i=1}^{2}A^{2}{}_{i1}\right)$； $\Omega_{C2}=\left(\bigcup_{i=1}^{2}A^{2}{}_{i2}\right)$

(c) $\Omega_{C1}=\left(\bigcup_{i=1}^{3}A^{3}{}_{i1}\right)\cup\left(\bigcup_{i=1}^{3}A^{3}{}_{i2}\right)$； $\Omega_{C2}=\left(\bigcup_{i=1}^{3}A^{3}{}_{i3}\right)$

Where Ω_{C1} is grid subspace on which left radical located, and Ω_{C2} is grid subspace on which right radical located.

C1 C2

The 2×2 uniform grid and 3×3 uniform grid were used to analyze left to right structure. The above analysis shows that the segmentation position between left radical and right radical have the following three types : one is portrait reticle area of 2×2 uniform vertical grid, h_1, others is landscape orientation reticle area of 3×3 uniform vertical grid, h_1 and h_3.

The recognition of left to right structure refer to find segmentation position between left radical and right radical correctly. According to the representation of left to right structure, there is three segmentation position. Therefore, The recognition of left to right structure is decision-making about three category, that is $m=3$. In preprocess of recognition, the vertical projection below certain threshold of experience was chosen to be candidate segmentation position.

4.2. Feature analysis

Chinese character is a special graphic character [13], representation of left to right structure based on the simple grid shown that Chinese character deriving from left to right structure can be pieced with left graphic and right graphic. Then, automatic segmentation between left radical and right radical depend on graphics features of Chinese character, especially graphics features in the segmentation position. Generally, separate ditch will be appeared at the segmentation position, stroke of radical can not leap over others radical. As the basic unit of Chinese character, most of CCP is single connectivity , so, connectivity should be taken into account during recognition.

After studying vertical projection of left radical and right radical, strokes, closed connective region and its combined properties deeply, it can get the following conclusions : Segmentation of the vertical projection should be less than the threshold value of experience, Segmentation position does not exist a single stroke which leap over two relatively independent radical, Segmentation position should not present within closed connective region, the single connective region happening to both sides of segmentation position could be combined .

Based on the foregoing analysis, Naive Bayesian classifier should have four features : vertical projection x_{j1}, stroke leaping x_{j2}, closure of connective region x_{j3}, combination of connective region x_{j4}. Above our features can be obtained by image preprocessing, because of space limitations, implementation of feature extraction will given in another paper.

C1 C2 C1 C2

(a) (b) (c)

4.3. Experiment and analysis

For experimental study of the method proposed, sample sets of 3950 Chinese characters of left to right structure , which were chosen from GB2312-80, were involved. All samples were transform into bitmap image as a training and test images. Among them, boldface font was chosen. Experimental procedures and results were shown as follows :

Step1: Design of Naive Bayesian classifier. Choose training samples correspondent to three segmentation position from the Chinese characters images, and then estimate conditional probability distribution function with feature distribution of training sample .

Step2: According to recognition model shown in Figure 2, design testing algorithm and build CCS recognition system within the computer .

Step3: Take all sample sets as test input of CCS recognition system for the test. Compare with artificial classification results and machine classification result into the computer, and record the results of comparision .

Step4: If artificial classification results is the same as machine classification result, then record number of correct recognition. Take numbers of correct recognition for the proportion of test sample as correct recognition rate, then calculate the correct recognition rate.

In accordance with the above steps to perform experiment, correct recognition rate reach above 95.5%. The main reason affect correct recognition rate are as follows :

(1) Due to misclassify non-simply connective CCP. Such as Chinese character "八", "非", "川", "兆",and so on. They are semantic-based independent entity, but as for graphics , they are made up of two single connectivity region which are separated completely. Semantic independence but Graphics separation lead to find error segmentation position between radicals which are used to piece these Chinese characters. To resolve such problems, standard templates of CCP should be made , and then corresponding components can be cover-match to achieve recognition .

(2) Some Chinese characters have serious overlap and adhesion, feature extraction based on digital image processing technology may extract wrong feature.

(3) The GF3001-1997 Information Processing GB13000.1 Character Set Chinese Radical criterion is only a guiding radical criterion, and the representation of CCS deriving from ISO / IEC 16046 is only a ideographic descriptor . Therefore, in accordance with the above criterion, artificial classification must exists misclassification.

5. Conclusions

As for human being, cognitive process happened to the object is to find information characterization of essential attribute. When the object can be represented by attribute found and representation of the object can distinguish others objects, cognitive process of the object can be finished. Above cognitive process is just a process of pattern recognition. Feature of pattern, that is essential attribute of the object, its completeness can be reflected through the recongiton of the object as well. Therefore, representation of the pattern is the same as recognition of pattern.

Simple grid is a description tool for CCS. For features which were extracted with a simple grid of , its completeness can be reflected through the CCS recogniton. In order to check effectiveness of CCS representation based on the simple grid, the method of CCS recognition based on Bayesian learning is proposed by this paper. Experimental results showed that simple grid is effective tools to represent CCS. Experimental results indicated that method proposed can extract the features of CCS correctly, has limited capacity and easy-to-use, and good extension. Anyway, the method proposed can achieve purpose of complete examination .

It can be found that the definition of CCS and representation of radical is only guidance in nature, and not conducive to the computer. There is also not uniform standard for CCS and radical. Whatever, CCS are linked to CCP closely. Therefore, it is possible to solve problem through establishing uniform standard for CCS and CCP based on computer architecture .

Acknowledgement

The author would like to thank ZhiWen Qiu and SuLi Wang for their helpful work in the experiment, and also thank JianPin Lu for his helpful comment on this work.

References

[1] E.J. Dao, Cultural function of chinese characters based on physical structure. *Chinese Character Culture*, 2:39-43, 2002.

[2] N. Wang, Chinese configurational basis and modern radical resolution. *Language Planning* , 3:4-9, 1997.

[3] Y. Xia, X.Z. Zhang, The expressions of Chinese characters used in machine recognition and learning. *Acta Automatica Sinica*, 12:312-314, 1986.

[4] Y. WU, X.Q. Ding, *Prinple of Chinese character recognition and its implementation*, High Education Press, BeiJing , 1992.

[5] Z.W. Feng, Description of Chinese character structure by context free grammar. *Linguistic Sciences*, 5:14-23, 2006.

[6] G.Y. Pi, Z.B. Mu, Simple grid of representing Chinese character within computer and its descriptive method. *China patent,* 200410015239.2, 2004.

[7] T.C. Liang, Z.W. Qiu, Y.G. Pi, Simple grid based on cognitive mechanism and application research on description for structure of Chinese character. *Proceedings of the 26th Chinese Control Conference* , pp.689-693,2007.

[8] A.K. Jain, Duin, R.P.W, J.C. Mao, Statistical pattern recognition a review. *IEEE Transaction On Pattern Analysis and Machine Intelligence*, 1:4-37, 2000, 2005.

[9] B.C. Li, S.S. Wan, L.M. Wang, The simplified expression of Bayesian network. *Chinese Journal of Scientific Instrument*, 10:1070-1073,2005.

[10] F.Y u, Y.F. Jiang, A feature selection method for NB-based classifier. *Acta Scientiarum Naturalium Universitatis Sunyatseni*, 5:118-120, 2006.

[11] Y.H. Bai, M. Chen, J.Q. Wang, Naive Bayesian approaches for anomaly detection. *Computer Engineering and Applications*, 34:131-133, 2005.

[12] X.C. Guo, Analysis of Chinese characters and experimental sampling strategies. *Chinese Ergonomics*, 3:14-18, 1999.

[13] G.Y. Wang, Chinese configuration system and its development stage. *Journal of Renmin University of China*, 1:104-108, 1999.