Computer recognition research on handwritten Chinese characters

YuXiang Hu

School of North China Electric Power University,Baoding 071003,China

648223545@qq.com

Keywords: handwritten Chinese characters, pattern recognition, characters recognition

Abstract. Computer identification on handwritten Chinese characters is one of the most difficult problems in the field of pattern recognition.In order to improve the efficiency of identification of Chinese character recognition software and realize the automation and intelligenceof the software, it is necessary to realize computer automatic tracking identification on the handwritten Chinese characters. On the basis of introduction of general method and strategy and pretreatment process of characters recognition, this paper provides a good reference for the field of Chinese characters recognition research.

1. Introduction

Character recognition is an important branch of pattern recognition. Character recognition actually is a problem of characters classification. Usually, it is classified by feature discriminant and feature matching method.

Feature discriminant is classified through the common rules (for example, English or Chinese).It does not need to use a variety of specific knowledge of the word and according to the degree of feature extraction namely the knowledge use degree, itcompletes character recognition using structure analysis method by stages.

Matching method is conducted by shape matching method based on the knowledge of characters (called dictionary).According to the implementation technique, it can be divided into two kinds: one directly uses the whole domain matching between the input two-dimensional plane image andthe memory image indictionary; another only takes part of the image to match with dictionary.Then according to the shape and relative position, it compares with the knowledge stored in the dictionaryto identify each specific character.

Matching method is generally used in standardization of printed words, especially the same font of the printed words. Structure analysis method is usually used in handwritten character recognition.Generally speaking, the programming of matching method of is simple, the dictionary occupied space is large and recognition speed is fast. Structure analysis program is complex but it can deal with handwritten character deformation problems with the advantage of approximate words distinction, but it will be unstable when used for the initial classification.So, in handwritten word recognition, it tends to combine two methods.

1.1 The principle of characters recognition

Character recognition principle block diagram is shown in figure 1-1. The scannedby scanner or written words changeinto a certain gray value of digital sampling signal through the module inputting to computer. Pre-treatment generally includes eliminate noise, the binary, character segmentation, smooth, standardization, linear or nonlinear transformation and so on.After pretreatment, characters become normalized binary bitmap information (as shown in figure 1-2), in which “1” represents stroke and "0" represents background part. For binary bitmap, according to the requirements of the recognition method, it extracts characteristics on behalf of the word, and stores in a computer and compares with a known standard characters to feature matching and finally finds out the closest one to input character, which is considered to be the word recognition result.

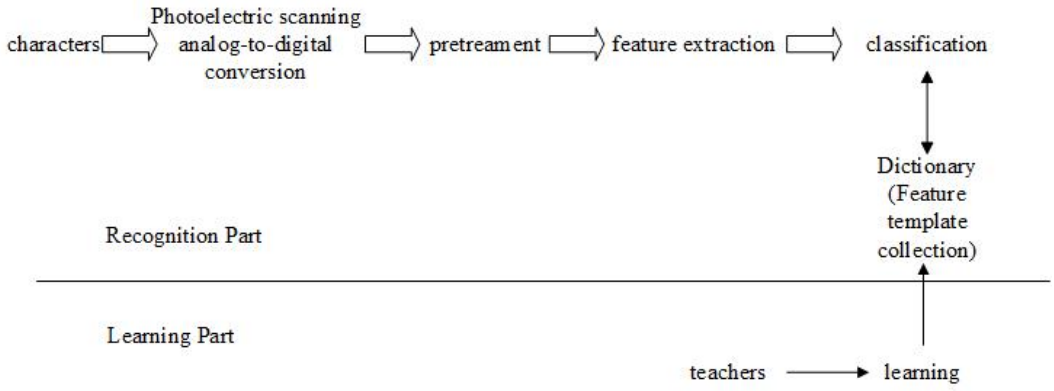

Figure 1-1 Character recognition principle block diagram

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

Figure 1-2 Normalized binary bitmap information

2. Summary of character recognition method

Characters recognition methods can be classified into two kinds. One is statistical approaches and the other is syntax-based approach. This paper mainly focuses on the former one. In the following, it introduces several statistical methods in details.

In general, it can obtain Ncharacteristics from the input text.Each set in N characteristics can be considered as a vector, known as the feature vector.The so-called classification problem is to define every possible vector into a specified model class [1]

2.1Fixed-point sampling method

It uses (i, j) to represent the coordination of character lattice, f (i, j) the point grey value. If the word calligraphy goes through the point (i, j), f (i, j) = 1; if not, f (i, j) = 0.In this way, the value of f (i, j) can be used to distinguish between k and k 'two words.

2.2Correlation method

The fixed-point sampling method is too idealistic, as long as the input character slightly deform or move, it will cause misjudgmenteven good printing quality words.If we are not only according to a few sampling points, but usinga n×nnormalized lattice characteristics as a dictionary, that is to say, if the stroke lattice $w_k(i, j)$ of word kis seen as a collection of characteristic vector and input word lattice is f (i, j), it can be calculated that

$$\sum_{i,j} f(i, j) \times w_k(i, j) \tag{2-1}$$

The greater the value, the better the consistency[2].

Letting the above classification thought into abstract, mathematically speaking, classification problem can be conducted with the aid of discriminant function.Using $\omega_1, \omega_2, ...\omega_m$ to indicate m model class need to be identified and let:

$$X = \left| x_1, x_2 ... x_m \right|^T \tag{2-2}$$

X indicates the characteristic vector where x_i describes the measurement of ith characteristic. Using $D_j(x)$ represents the discriminant function related with model class $w_j (j = 1, 2...m)$. If the input model of characteristic vector X is in w_i, denoted as $x - w_i$,so the value of $D_i(x)$ must be the largest one, namely for all of $x - w_i$

$$D_i(x) > D_j(x) \quad i = 1, 2, ...m, i \neq j \tag{2-3}$$

The region edges related to class w_i and w_j are called discriminant edge which can be described as follows.

$$D_i(x) - D_j(x) = 0 \tag{2-4}$$

2.3 Minimum distance classifier

Minimum distance classifier is a linear classifier.It usesthe distance between the input words and some reference vector or some model points in the feature space as classification criteria.

Supposed giving m reference vector: $G_1, G_2, ...G_m$, the minimum distance means that when $|X - G_i| = \min$, $X - w_i$. Let X as the characteristic vector of input characters $X = (x_1, x_2, ...x_m)$, G represents a certain vector of standard words in dictionary $G = (g_1, g_2, ...g_m)$. In the pattern recognition field, the following distances are commonly used.

(1) Minkowski Distance

$$D(X, G) = \left[\sum_{i=1}^{m} |x_i - g_i|^q \right]^{1/q} \tag{2-5}$$

(2)Mahalanobis distance

When two m dimensional vector X, G are normally distributed and with the same covariance matrix, the Mahalanobis distance can be described as:

$$D(X, G) = \left[(X - G) \sum^{-1} (X - G)^T \right]^2 \tag{2-6}$$

When using the minimum distance in character recognition, it respectively calculates the distance between characters characteristic vector X and character vector G_i ,namely $D(X, G_1), D(X, G_2)...D(X, G_m)$, finding the minimum one $D(X, G_i)$ so it can discriminant input words belong to class ω_i.

3. Image preprocessing in Chinese character recognition

Preprocessing is an important part of the whole Chinese character recognition. Situations like paper is not put straight when the manual and scan for writing Chinese characters will cause Chinese character size position uncertainty, which causesthe identification cannot directly conducted on the original image.So, before a single Chinese character recognition, first it needs to preprocessthe original image whose performance will directly affect the quality and result of the performance of the whole Chinese character recognition system.The process of pretreatment method is as shown in figure 2-1.

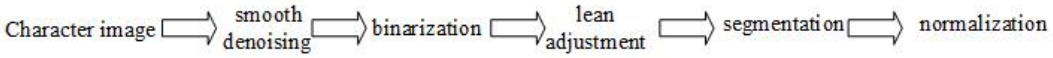

Figure2-1 the process of preprocessing in character recognition

3.1Smoothing processing

Smoothing processing is one of the image enhancement technology.It has two purposes: one highlightsuseful information in an imageaccording to the specific needs. The other is to eliminate the mixednoisein order to adapt to the needs of computer processing.The requirement of image smoothing processing has two: one is it cannot damage image edge profile and lines and other important information;the other is to make the image clear to verify the good visual effect.

3.2 Image binaryzation

Image binarization is to remove the unnecessary information in the scanned image to improve the speed of recognition and lays the foundation for later Chinese character feature extraction.Usually the Chinese character image used to be identifiedis grayscale level 256 or binary figure.Addingbinary functionof grayscale image in pretreatment not only can improve the quality of binarization and greatly compress image data, but can reduce the storage capacity and improve the software adaptability

3.3 Skew Correction of character image

In the process of scanning, there are various reasons to cause the tilt phenomenon, however, Chinese character recognition is much sensitive to image tilt. The image tilt correction generally can be divided into manual correction and automatic correction.Manual correction refers to the system providing a means of human-computer interaction to help the user to specify the image tilt angle, then using corresponding algorithm for image rotation to the correction.Automatic correction, that is, by the system automatically calculate the angle of the image through the analysis of the lattice of image information, and thus automatically to realize correction.

3.4 Line segmentation of the image

We usually use the integral projection method to finish line, character segmentation processing. The method has a simple algorithm with the advantages of small amount of calculation but is insufficient and poor adaptability.But before word segmentation, we have already smoothed the Chinese character image and skew correction, so the integral projection method is feasible [3].

3.5Normalization

After line cutting processing and character segment, it needs to conduct normalized processing to eliminate the change of location and sizeof Chinese characters because of layout, font size, font changes.Normalized processing mainly include the location and size normalization, sometimes includes stroke (thickness) normalized.

4. Summary

Character recognition is an important branch of pattern recognition.It’s easy to be wrong when choosing from a large section of the handwritten manuscriptsbesides,picking handwriting is a complicated thing, the workload is also big.This topic research isChinese characters recognition and it’s principle and its procedures which provides a good reference for the further research.

References

[1].N.W.Strathy, C.Y.Suen, A.Krzyzak. Segmentation of hand writtend igitsusing contour features. Proceedings of the Second International Conference on Document Analysis and Recognition, 1993, PP.577-580.

[2].M,Suters,H.Yan.Connectedhandwrittendigitsseparationusingexternalboundarycurvature.J.Electron.Imaging,1994,3(3):251-256.

[3]. Cunhua Li.Character image skew correctionbased on the contour projection method.Chinese journal of image and graphics,2001,10.