Journal of Southeast University (English Edition) Dec. 2002 Vol.18 No.4 ISSN 1003—7985

An Approach to Unsupervised Character Classification Based on Similarity Measure in Fuzzy Model*

Lu Da[1]** Qian Yiping[1] Xie Mingpei[2] Pu Wei[1]

([1] Department of Physics, Changshu College, Changshu 215500, China)

([2] Department of Computer Science, Fudan University, Shanghai 200433, China)

Abstract: This paper presents a fuzzy logic approach to efficiently perform unsupervised character classification for improvement in robustness, correctness and speed of a character recognition system. The characters are first split into eight typographical categories. The classification scheme uses pattern matching to classify the characters in each category into a set of fuzzy prototypes based on a nonlinear weighted similarity function. The fuzzy unsupervised character classification, which is natural in the representation of prototypes for character matching, is developed and a weighted fuzzy similarity measure is explored. The characteristics of the fuzzy model are discussed and used in speeding up the classification process. After classification, the character recognition which is simply applied on a smaller set of the fuzzy prototypes, becomes much easier and less time-consuming.

Key words: fuzzy model, weighted fuzzy similarity measure, unsupervised character classification, matching algorithm, classification hierarchy

Although optical character recognition (OCR) technology is widely used today, problems such as efficiency, accuracy, and reliability within most OCR systems have been encountered because huge variations of font styles, character sizes, and special symbols are to be dealt with. Another problem is the unavoidable noise produces while the characters are digitized into binary images and the image distortion which is caused by the defect of optical image systems. Therefore, the postprocessing strategies, such as dictionary checking and semantic understanding, would seem the natural choice to be combined with OCR for improvement in correctness of a character recognition system[1-5].

Generally speaking, the postprocessing has been employed for raising the recognition rate, which makes it be worth applying OCR. However, the postprocessing cannot recover the character original information which were lost in the steps of character recognition processing such as digitizing images, feature selection, and statistical decision.

The character classification is performed prior to recognition to extract a set of representative prototypes into which the similar patterns of characters are grouped. Compared to the existing OCR systems, our classification reduces the scope of characters to be recognized significantly, and our fuzzy model is more robust in nature.

The procedures used in our document processing system are as follows: The office documents are first digitized and thresholded into binary images by a scanner. The preprocessing procedure, which includes the text block segmentation and classification and character isolation, generates a set of individual character images. The typographical categorization divides them into eight categories based on their typographical structures[6,7]. The unsupervised character classification adopts fuzzy logic to further classify them into a limited set of fuzzy prototypes based on a nonlinear weighted similarity function. The optical character recognition is employed to recognize the set of fuzzy prototypes. Finally, the postprocessing intends to correct errors by means of dictionary checking or semantics understanding. In this paper, we focus on the unsupervised character classification.

1 Statistical Fuzzy Model for Character Classification

The set of fuzzy prototypes is constructed based on statistical analysis by grouping similar patterns into a single class. An image of a fuzzy prototype is a matrix of pixels where each element is associated with a membership that represents the degree of the pixel that

Received 2002-07-31.

* The project supported by the China Scholarship Council Foundation (99832097) and the National Science of Jiangsu Education Commission (99KGB140009).

** Born in 1947, male, associate professor.

belongs to the object.

Definition 1 Let E^2 denote two-dimensional Euclidean space. A fuzzy model λ in E^2 is a matrix of ordered pairs such that $\lambda = \{(p_{ij}, \chi_{ij})\}$, where p_{ij} represents a pixel and χ_{ij}, in the range $[0,1]$, represents the grade of membership of the pixel belonging to the object.

Proposition 1 Assume a fuzzy prototype λ is composed of a set of merged binary images $\{A_1, A_2, \cdots, A_m\}$, where their elements are denoted by a_{ij}. The membership of each pixel in λ is computed by

$$\chi_{ij} = \Big(\sum_{a_{ij} \in A_1}^{A_m} a_{ij} \Big) \Big/ m \tag{1}$$

Definition 2 Let $\lambda = \{(p_1, \chi_1), (p_2, \chi_2), \cdots, (p_n, \chi_n)\}$ be a fuzzy prototype. The cardinality of λ denoted as σ_λ is a fuzzy number and can be formulated as $\sigma_\lambda = \{(i, \psi_i) \mid i = 0,1,2,\cdots,n\}$, where ψ_i denotes the membership of the cardinality being equal to i.

According to probability theory, it is plausible to formulate ψ_i as

$$\left.\begin{aligned} \psi_0 &= \prod_{i=1}^{n} \chi'_i \\ \psi_1 &= \sum_{i=1}^{n} \Big(\chi_i \prod_{\substack{j=1,\cdots,n \\ j \neq i}} \chi'_j \Big) \\ &\vdots \\ \psi_{n-1} &= \sum_{i=1}^{n} \Big[\Big(\prod_{\substack{j=1,\cdots,n \\ j \neq i}} \chi_j \Big) \chi_i \Big] \\ \psi_n &= \prod_{i=1}^{n} \chi_i \end{aligned}\right\} \tag{2}$$

where χ'_i and χ'_j denote the maximizing fitting value of χ_i and χ_j.

Eq.(2) can be expressed in a general form as

$$\psi_m = \sum_{i_1 \neq i_2 \neq \cdots \neq i_m} \Big(\chi_{i_1} \chi_{i_2} \cdots \chi_{i_m} \prod_{\substack{j \neq i_1 \\ j \neq i_2 \\ \vdots \\ j \neq i_m}} \chi'_j \Big) \tag{3}$$

The fuzzy cardinality has the property that $\sum_{i=0}^{n} \psi_i = 1$. The expected value of the cardinality can be derived as

$$E(\sigma_\lambda) = \sum_{i=1}^{n} i\psi_i = \sum_{i=1}^{n} \chi_i \tag{4}$$

Details of the derivation are omitted here. It is significant to simplify the cardinality of a fuzzy set to a unique value for analysis and computing.

Proposition 2 The cardinality of a fuzzy prototype λ in E^2 is equal to the sum of the membership values. That is $\sigma_\lambda = \sum \chi_{ij}$.

Similar to the derivation in cardinality, the centroid of a fuzzy prototype can be simplified to be a unique number.

Proposition 3 The centroid of the first moment of a fuzzy prototype λ in E^2 is formulated as

$$x_{C_\lambda} = \frac{\sum j\chi_{ij}}{\sum \chi_{ij}}, \quad y_{C_\lambda} = \frac{\sum i\chi_{ij}}{\sum \chi_{ij}} \tag{5}$$

where all summations are with respect to ij.

Proposition 4 The width of a fuzzy prototype λ in E^2 is the summation of the maximal membership in columns, and the height of a fuzzy prototype λ in E^2 is the summation of the maximal memberships in rows; that is

$$w_\lambda = \sum_j \max_i \chi_{ij}, \quad h_\lambda = \sum_i \max_j \chi_{ij} \tag{6}$$

Proof The proof is given by the induction hypothesis. For the case when $m = 1$, we have $w_\lambda = w_i$, which satisfies Eq.(7).

Assume that $m = x$ also satisfies Eq.(7) such that $w_\lambda \leqslant \frac{1}{x} \sum_{i=1}^{x} w_i$. Let $\alpha_j = \max_i \chi_{ij}^{\lambda}$ denote the maximal membership value of the j-th column in the fuzzy prototype λ, $\beta_j = \max_i \chi_{ij}^{A_{x+1}}$ denotes the maximal membership value of the j-th column in A_{x+1}, which is either 1 or 0, and let ρ_j denote the maximal membership value of the j-th column in the new fuzzy prototype λ'. When a new image A_{x+1} is merged, the following inequality is derived:

$$\rho_j \leqslant \frac{x\alpha_j + \beta_j}{x+1}$$

The width of the new fuzzy prototype will be

$$\begin{aligned} w_{\lambda'} = \sum \rho_j &\leqslant \frac{\sum x\alpha_j + \beta_j}{x+1} = \\ &\frac{x}{x+1} \sum \alpha_j + \frac{1}{x+1} \sum \beta_j = \\ &\frac{x}{x+1} w_\lambda + \frac{1}{x+1} w_{x+1} \leqslant \\ &\frac{x}{x+1} \Big(\frac{1}{x} \sum_{i=1}^{x} w_i \Big) + \frac{1}{x+1} w_{x+1} = \\ &\frac{1}{x+1} \sum_{i=1}^{x+1} w_i \end{aligned}$$

Therefore Eq.(7) for the width is proved. The proof for the height can be similarly derived.

Proposition 5 If a fuzzy prototype λ is composed of a set of merged binary images $\{A_1, A_2, \cdots, A_m\}$ with widths $\{w_1, w_2, \cdots, w_m\}$ and heights $\{h_1, h_2, \cdots, h_m\}$, then the width of λ is less than or equal to the average width of the set of the images and the height of λ is less than or equal to the average

height of the set of the images; that is

$$w_\lambda \leqslant \sum_{i=1}^{m} \frac{w_i}{m},\ h_\lambda \leqslant \sum_{i=1}^{m} \frac{h_i}{m} \tag{7}$$

Because the fuzzy prototype is group of similar patterns, the difference between w_λ and $\frac{1}{m}\sum_{i=1}^{m} w_i$ is small. Thus, the latter can be applied to approximate the width of the fuzzy prototype for simplicity.

2 Similarity Measure in Fuzzy Model

Similarity is an abstract concept of fuzziness that provides a quantitative measure to the relationship between two variables.

Proposition 6 Let $\lambda_1 = \{(p_{ij}, \chi_{ij}^{(1)})\}$ and $\lambda_2 = \{(p_{ij}, \chi_{ij}^{(2)})\}$ denote two fuzzy propotypes in E^2 and let $\gamma_{\lambda_1} = \{\gamma_{ij}^{(1)}\}$ and $\gamma_{\lambda_2} = \{\gamma_{ij}^{(2)}\}$ represent the weight functions associated with λ_1 and λ_2, respectively. The similarity measure of λ_1 and λ_2 is defined as

$$\zeta(\lambda_1,\lambda_2) = \frac{\sum\left(\chi_{ij}^{(1)} \wedge \chi_{ij}^{(2)} - \frac{1}{2}\gamma_{ij}^{(1)}\chi_{ij}^{(2)} - \frac{1}{2}\gamma_{ij}^{(2)}\chi_{ij}^{(1)}\right)}{\sqrt{\sum \chi_{ij}^{(1)^2} \sum \chi_{ij}^{(2)^2}}} \tag{8}$$

where $\wedge$ is the symbol for minimum representing the intersection on fuzzy sets; that is, $\lambda_1 \cap \lambda_2 = \{(p_{ij}, \chi_{ij}^{(1)}) \wedge \chi_{ij}^{(2)})\}$; where $\chi_{ij}^{(n)^2}(n = 1,2)$ denotes the self-intersection $\chi_{ij}^{(n)} \wedge \chi_{ij}^{(n)}$.

The reason why the denominator $\chi_{ij}^{(n)} \wedge \chi_{ij}^{(n)}$ of Eq.(8) is used instead of $\chi_{ij}^{(n)} \times \chi_{ij}^{(n)}$ comes from the viewpoint of fuzzy properties. Consider a membership $\chi_{ij} = 0.8$, which represents a concept that 80% of the area in a pixel p_{ij} belongs to the object (i.e., has value "1") and 20% of the area belongs to the background (i.e., has value "0"). Therefore, χ_{ij}^2 should be carried out as $\chi_{ij}^2 = 0.8 \times 1^2 + 0.2 \times 0^2 = 0.8$ instead of $0.8^2 = 0.64$. Moreover, the first term of the numerator is $\chi_{ij}^{(1)} \wedge \chi_{ij}^{(2)}$ instead of $\chi_{ij}^{(1)} \times \chi_{ij}^{(2)}$ because two fuzzy subsets are equal if their membership functions are equal; that is

$$\lambda_1 = \lambda_2 \text{ iff } \chi_{ij}^{(1)} = \chi_{ij}^{(2)} \tag{9}$$

Therefore,

$$\zeta(\lambda_1,\lambda_1) = \frac{\sum\left(\chi_{ij}^{(1)} \wedge \chi_{ij}^{(1)} - \frac{1}{2}\gamma_{ij}^{(1)}\chi_{ij}^{(1)} - \frac{1}{2}\gamma_{ij}^{(1)}\chi_{ij}^{(1)}\right)}{\sqrt{\sum \chi_{ij}^{(1)^2} \sum \chi_{ij}^{(1)^2}}} \tag{10}$$

but

$$\zeta(\lambda_1,\lambda_1) \neq \frac{\sum\left(\chi_{ij}^{(1)}\chi_{ij}^{(1)} - \frac{1}{2}\gamma_{ij}^{(1)}\chi_{ij}^{(1)} - \frac{1}{2}\gamma_{ij}^{(1)}\chi_{ij}^{(1)}\right)}{\sqrt{\sum \chi_{ij}^{(1)^2} \sum \chi_{ij}^{(1)^2}}} \leqslant 1 \tag{11}$$

Proposition 7 Assume that λ, a fuzzy prototype is composed of a set of merged binary images $\{A_1, A_2, \cdots, A_m\}$ associated with the weight functions $\{\omega_{A_1}, \omega_{A_2}, \cdots, \omega_{A_m}\}$, The weight function γ of λ is defined as

$$\gamma_{ij} = \left(\sum_{\omega_{ij} \in \omega_{A_1}}^{\omega_{A_m}} \omega_{ij}\right) \Big/ m \tag{12}$$

where ω_{ij} s are the elements in $\omega_{A_1}, \omega_{A_2}, \cdots, \omega_{A_m}$.

3 Matching Algorithm and Classification Hierarchy

3.1 Matching algorithm

Given two fuzzy patterns A and B, the way to find the best matching is to shift A around B, calculate the correlation coefficient for every position, and select the highest value. However, it is not efficient. If two patterns are similar, they should have similar geometric properties. If two patterns are dissimilar, they are not the best matching. A simple way is to calculate the centroids of both patterns, and the similarity measure is calculated by matching the centroids. Some allowance must be considered due to noise. The algorithm is described below.

Step 1 Calculate C_A and C_B, which represent the centroids of A and B; that is

$$C_A = \left(\frac{\sum j\chi_{ij}^{(A)}}{\sum \chi_{ij}^{(A)}}, \frac{\sum i\chi_{ij}^{(A)}}{\sum \chi_{ij}^{(A)}}\right)$$

and

$$C_B = \left(\frac{\sum j\chi_{ij}^{(B)}}{\sum \chi_{ij}^{(B)}}, \frac{\sum i\chi_{ij}^{(B)}}{\sum \chi_{ij}^{(B)}}\right) \tag{13}$$

Step 2 Compute $\zeta(A,B)$ with minimum distance $\overline{C_A C_B}$. In other words, $\zeta_{\alpha\beta}(A,B)$ is derived from shifting pattern A with (α,β):

$$\zeta_{\alpha\beta}(A,B) = \frac{\sum\left(\chi_{ij}^{(A)} \wedge \chi_{i+\alpha,j+\beta}^{(B)} - \frac{1}{2}\gamma_{ij}^{(A)}\chi_{i+\alpha,j+\beta}^{(B)} - \frac{1}{2}\gamma_{i+\alpha,j+\beta}^{(B)}\chi_{ij}^{(A)}\right)}{\sqrt{\sum \chi_{ij}^{(A)^2} \sum \chi_{ij}^{(B)^2}}} \tag{14}$$

where $\alpha = \text{round}(x_{C_B} - x_{C_A})$ and $\beta = \text{round}(y_{C_B} - y_{C_A})$.

If the correlation coefficient is higher than the threshold, e.g. 0.9, A and B are considered to be the

same.

Step 3 If ζ is the critical range, e.g. 0.8 to 0.9, then the values of $\zeta_{\alpha\beta}(A,B)$ are also calculated with α and β in the range of

$$\left|\alpha - (x_{C_B} - x_{C_A})\right| \leqslant 1$$

and

$$\left|\beta - (y_{C_B} - y_{C_A})\right| \leqslant 1 \qquad (15)$$

If the values of $x_{C_B} - x_{C_A}$ and $y_{C_B} - y_{C_A}$ are not integers, there are more than three positions for both α and β to be matched in this step. If there is a match, i.e., higher than 0.9, A and B are set to the same class. Otherwise, the two patterns are considered as distinct objects.

Step 3 represents fuzzy reasoning because the similarity is ambiguous in a critical range. The similarity between two similar patterns can be measured in this range because of the distortion of the centroid of the image due to noise, and a better match can be obtained by shifting one pixel for pattern. Fig.1 illustrates an example to two primitive images. The ambiguous similarity calculated in step 2 is shown in Fig.1(c), and a better match found by step 3 is shown in Fig.1(d). Note that Fig.1(c) and (d) show the superposition of images a_1 and a_2 in their matching positions. The symbols "#", "1" and "2" are used, respectively to represent the common pixel of two images, the pixel in image a_1 only, and the pixel in image a_2 only.

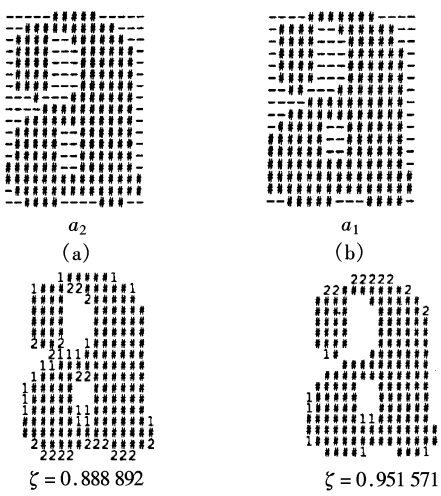

Fig. 1 An example to two primitive images. (a) The sample image a_2; (b) The sample image a_1; (c) A critical similarity is measured; (d) A good matching is found by shifting a_1 one pixel down

3.2 Classification hierarchy

It is inefficient if the matching is performed against all the character prototypes in an arbitrary sequence. In this section, a hierarchical-tree approach is adopted.

Compared with sequential classification, hierarchical approach reduces the searching time for matching. Another advantage of the hierarchical classification is the capability of being processed in parallel. Two kinds of parallelism can be carried out. First, each node in a classification tree can be performed in parallel. Second, the merging to two fuzzy subsets can be implemented in parallel because all the elements in a fuzzy subset are distinct and can be processed at the same time.

A text block after segmentation typically represents a line of characters. Usually, the sizes of characters in a line are uniform. Therefore, a text line is considered as a unit in classification. The set of fuzzy prototypes associated with each line is hierarchically grouped. This facilitates the comparison of the character sized. If the sized of two text lines are obviously different, the merged prototype set is split into two disjoint subsets correspondingly and the matching is not performed.

4 Rules of Preclassifier for Grouping the Fuzzy Prototypes

For the lower levels in the hierarchical classification tree, the representative subsets of fuzzy prototypes contain only a few elements. Therefore, it is simple to search for similar objects in two subsets. However, in the higher levels it becomes impractical to perform the matching one by one because the size of the libraries increases. Therefore, Casey et al. applied a binary decision network[8]. However, the reliability of this method is questionable because the reliable pixels of each prototype are different.

To solve the aforementioned problems, a rule-based preclassifier is used. Considering two fuzzy patterns λ_1 and λ_2, the similarity measure is computed using Eq.(8). The correlation coefficient ζ can be divided into two terms of equality measure E and inequality measure I:

$$\left.\begin{aligned} E &= \frac{\sum \chi_{ij}^{(1)} \wedge \chi_{ij}^{(2)}}{\sqrt{\sum {\chi_{ij}^{(1)}}^2 {\chi_{ij}^{(2)}}^2}} = \frac{\sigma_{\lambda_1 \cap \lambda_2}}{\sqrt{\sigma_{\lambda_1 \cap \lambda_2}}} \\ I &= \frac{\gamma_{ij}^{(1)} \chi_{ij}^{(2)} + \gamma_{ij}^{(2)} \chi_{ij}^{(1)}}{2\sqrt{\sigma_{\lambda_1} \sigma_{\lambda_2}}} \end{aligned}\right\} \qquad (16)$$

where σ_{λ_1}, σ_{λ_2} and $\sigma_{\lambda_1 \cap \lambda_2}$ are the cardinalities of the fuzzy prototypes λ_1, λ_2 and $\lambda_1 \bigcap \lambda_2$, respectively.

Let ζ_t be the threshold of the similarity measure, let λ_1 be the input model, and let λ_2 be the fuzzy variable in the library. For $\sigma_{\lambda_1} > \sigma_{\lambda_2}$, the best equality measure happens if $\lambda_1 \supset \lambda_2$, i.e., $\chi_{ij}^{(1)} \geqslant \chi_{ij}^{(2)}$. Therefore, λ_2 is a possible-match prototype only if

$$E(\lambda_1, \lambda_2) = \frac{\sigma_{\lambda_1 \cap \lambda_2}}{\sqrt{\sigma_{\lambda_1}\sigma_{\lambda_2}}} = \frac{\sigma_{\lambda_2}}{\sqrt{\sigma_{\lambda_1}\sigma_{\lambda_2}}} = \sqrt{\frac{\sigma_{\lambda_2}}{\sigma_{\lambda_1}}} \geqslant \zeta_t \tag{17}$$

Similarly, for $\sigma_{\lambda_1} < \sigma_{\lambda_2}$, λ_2 is a possible-match prototype only if

$$\sqrt{\sigma_{\lambda_1}/\sigma_{\lambda_2}} \geqslant \zeta_t \tag{18}$$

Therefore, the first rule is concluded using Eqs. (17) and (18).

Rule 1 λ_2 is a possible-match prototype of λ_1 iff $\zeta_t^2 \sigma_{\lambda_1} \leqslant \sigma_{\lambda_2} \leqslant \sigma_{\lambda_1}/\zeta_t^2$.

Since similar prototypes possess similar features, additional heuristic rules based on the features of the prototypes are described as follows.

Rule 2 Two fuzzy prototypes are impossible to match if the difference between their widths exceeds a threshold w_t.

Note that the height is not taken into consideration because the prototypes in the same typographical category have similar heights.

Rule 3 Two fuzzy prototypes are impossible to match if the difference of two prototypes in the total number of columns of the left or right region to the centroid is greater than a threshold C_1.

Rule 4 Two fuzzy prototypes are impossible to match if the difference of two prototypes in the total number of rows of the upper or lower region to the centroid is greater than a threshold C_2.

In our system, each library of prototypes is sorted by its cardinalities. The set of prototypes to be possibly matched is extracted by Rule 1. Rules 2, 3 and 4 filter out the prototypes that are impossibly matched. Finally, a rough estimation of the similarity measure based on the projection profile is applied to extract the prototypes to be possibly matched prior to the two-dimensional pattern matching.

Rule 5 Let γ_i^{λ} denote the summation of the membership values on the i-th column of fuzzy prototype λ. Two fuzzy prototypes λ_1 and λ_2 are possibly matched iff the following condition holds:

$$\zeta_t \sqrt{\sigma_{\lambda_1}\sigma_{\lambda_2}} \leqslant \sum (\gamma_i^{\lambda_1} \wedge \gamma_{i+\alpha}^{\lambda_2})$$

$$\text{for} \quad \left| \alpha - (x_{C_B} - x_{C_A}) \right| \leqslant 1 \tag{19}$$

where $\wedge$ denotes the symbol of minimum.

5 Experimental Results

The proposed model was implemented in C language on the SUN workstation 4/490 under a UNIX operating system. The raw image of a document is captured and thresholded into binary code by the scanner. The characters in each textual block are extracted and typographically analyzed and categorized.

The unsupervised character classification is performed on each text line and a subset of fuzzy prototypes is generated correspondingly. Note that a fuzzy set contains eight subsets of the typographical categories. The subsets with similar height are grouped hierarchically. Finally, several sets of fuzzy prototypes corresponding to different sizes of characters are produced. Fig.2 shows the details of a few prototypes examples, where the notations "#" and "–" represent membership greater than 0.95 and less than 0.05, respectively, and an integer "i" represents the membership between $i/10 - 0.05$ and $i/10 + 0.05$. Some of the merged characters are illustrated in Fig.3. It is observed that the membership values of the linking pixels are lower when more patterns are included, and they can be separated. Those characters that do not have lower values on the linking pixels because the prototype contains too few patterns can be split more easily by means of a splitting algorithm which was raised by Ref. [9] or can be split by partial matching and dictionary look-up.

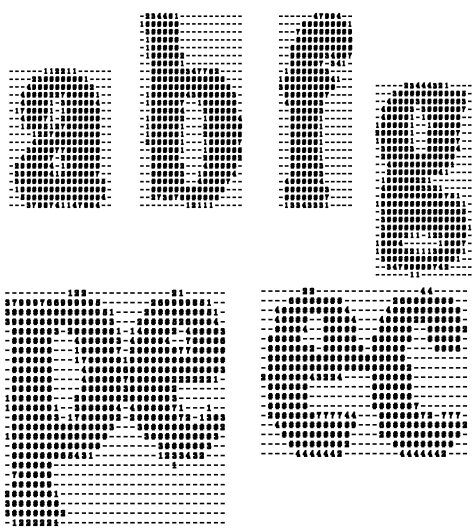

Fig.2 Examples of fuzzy prototypes with merged characters

The experiment has been divided into two sections. First, various character sizes from 5 to 12 points, each consisting of 4?000 text lines containing

307?221 characters have been tested. The results are shown in Tab.1. Where N denotes the test of character typographical structure classification which was raised by Ref.[6], and U denotes the test using typographical structure classification and unsupervised character classification.

Second, the experiment also tested a text which was polluted by typography. There are 28?600 characters in this text including 212 merged characters and 103 defective characters. The results are shown in Tab.2.

Tab.1 Experimental results from our character classification

Character size	Ambiguous character		Average runtime/(μs · ch^{-1})	
	N/%	U%	N	U
5	0.50	0.06	85	106
6	0.27	0.03	82	101
7	0.10	0	84	102
8	0.21	0	84	101
9	0.19	0	81	98
10	0.62	0	81	98
11	0.68	0	79	95
12	0.41	0	80	94

Tab.2 Experimental results for merged characters from using unsupervised character classification and using typographical structure classification

Recognition method	Rate for merged character recognition/%	Rate for character recognition/%	Speed of character recognition/(ch · s^{-1})
Using unsupervised character classification	79.24	99.72	0.270×10^{-2}
Using typographical structure classification	80.35	99.84	0.257×10^{-2}

From the results of experiment, we can see that the problem of classification is the selection of a suitable threshold for the similarity measure. For the larger character size, the boundary between distinct character categories is clear. However, when characters are smaller, the scope of different categories can be overlapped. There are two reasons. First, for smaller size the cardinality of the patterns is relatively small. However, the tolerance of noise, which can be caused by scanning or printing, is not affected by the character size. Therefore, the possible minimal equality measure for the same category will be lower. Second, the inequality measure can be smaller for different categories because the different parts of two different patterns in smaller size is smaller, and the weight is linear to the distance from the object. For example, the inequality measure between character "i" and "l" in smaller size is usually less than that in larger size for the same font because the latter contains more pixels of difference and has a higher weight. Therefore, the threshold must be adaptive to the character size. One approach is to emphasize the inequality measure such as

$$E - I \geqslant \zeta_t + CI \text{ or } E - (C+1)\, I \geqslant \zeta_t \quad (20)$$

where C is an coefficient. For the sample image, $C = 4$ and $\zeta_t = 0.86$ are adopted in classifying the main text with about 30-pixel height (point size 7). By emphasizing the inequality measure, the different categories are guaranteed to be separated, which is the principle of the classification. Although patterns belonging to the same category can be split due to noise and the fuzzy set can contain more prototypes than it really has, we can perform the similarity comparison again for the fuzzy set. Because the noise effect is reduced for the fuzzy set, the similar prototypes that were separated before probably will be merged.

6 Conclusion

This paper proposes a fuzzy model of unsupervised classification for preclassifying characters in the document analysis system. A fuzzy model of prototypes is defined and several propositions of the features of the fuzzy model are given. The existing similarity equations for matching are investigated and a nonlinear weighted similarity function is proposed and extended to the similarity measure of the fuzzy model. The hierarchy of the prototype grouping saves computational time as compared to sequential grouping, and it also has the advantage of parallel processing. The emphasis of inequality measure for small characters guarantees no misclassification, but a little redundancy is encountered on the fuzzy prototype set. This redundancy can be removed by self-grouping of the final prototype set. The propositions and algorithms have been tested with satisfactory performance.

References

[1] Akiyama T, Hagita N. Automated entry system for printed documents[J]. *Pattern Recognition*, 1990, **23**(11): 1141 - 1154.

[2] Schurmann J, Bartneck N, et al. Document analysis-from pixels to contents[A]. In: *Proceedings of IEEE*[C]. 1992, **80**(7): 1101 - 1119.

[3] Srihari S N. *Computer text recognition and error correction*[M]. Silver Spring, MD, USA: Computer Science Press, 1985.

[4] Impedovo S, Ottaviano L. Optical character recognition — a survey[J]. *Pattern Recognition*, 1991, **5**(1,2): 1 - 24.

[5] Bokser M. Omni document technologies[A]. In: *Proceeding of the IEEE*[C]. 1992, **80**(7): 1066 - 1078.

[6] Lu Da, McCane B, Pu Wei. Character preclassification based on fuzzy typographical analysis[A]. In: *Proceeding of the 6 th International Conference on Document Analysis and Recognition* [C]. Seattle, Washington, USA: IEEE Press, 2001. 74 - 78.

[7] Lu Da, Pu Wei, Xie Mingpei. Precise detection algorithm for locating the baseline of a text line[J]. *Mini-Micro System*, 2000, **21**(7):726 - 728.(in Chinese)

[8] Prakash M, Murty M. Growing subspace pattern recognition methods and their neural-network models[J]. *IEEE Transactions on Neural Networks*, 1997,**8**(1): 161 - 168.

[9] Lu Da, Xie Mingpei. A segmentation method of topographic approach for merged character images based on skeletonization [J]. *Journal of Chinese Information Processing*, 1999,**13**(2): 40 - 45.(in Chinese)

一种基于模糊模型相似测量的字符无监督分类法

卢 达[1] 钱忆平[1] 谢铭培[2] 浦 炜[1]

([1] 常熟高等专科学校物理系,常熟 215500)
([2] 复旦大学计算机科学系,上海 200433)

摘 要 提出了一种能有效完成对无监督字符分类的模糊逻辑方法,以提高字符识别系统的速度,正确性和鲁棒性.字符首先被分为 8 种印刷结构类,然后采用模式匹配方法将各类字符分别转换成基于一非线性加权相似函数的模糊样板集合.模糊无监督字符的分类是字符匹配的一种自然范例并发展了加权模糊相似测量的研究.本文讨论了该模糊模型的特性并用以加快字符分类处理,经过字符分类,在字符识别时由于只需针对较小的模糊样板集合而变得容易和快速.

关键词 模糊模型, 加权模糊相似测量, 字符无监督分类, 匹配算法, 分级归类

中图分类号 TP391