ONLINE RECOGNITION OF CHINESE HANDWRITING USING A HIERARCHICAL FUZZY CLUSTERING APPROACH

*Ming-Yen Tsai, Leu-Shing Lan**, and *Wei-Tzen Pao*

Department of Electronics Engineering
National Yunlin University of Science and Technology, Taiwan

ABSTRACT

In this paper, we propose a fuzzy clustering approach to improve the template matching method described in [1] that demonstrated high recognition rates. Our focus is on the reduction of recognition time. The key idea of the presented scheme is to organize the template pool in a structured manner so that matching can be performed more quickly. Through the hierarchical fuzzy clustering, the original template pool has the form of a decision tree where each internal node in the tree is represented by a cluster center. Since the number of class templates in each subcluster is less than the original cluster, the recognition time can be effectively reduced. An approximate formula for the number of distance calculations needed for the presented scheme is derived. The use of fuzzy clustering is to provide a mechanism for cluster overlapping so that in the middle of the recognition process, almost 100% hit ratio can be obtained. In the experiments conducted, we obtained approximately a 3.4 times reduction of recognition time (using the simplified formula we derived) at only a 0.2% average loss of recognition rate.

1. INTRODUCTION

A successful online handwriting recognizer exhibits the properties of high recognition rate, low computational complexities, and low on-board memory usage. A variety of methods have been proposed for online Chinese handwriting recognition. These may roughly be classified into several categories: (1) template matching [1], (2) deformation-based template matching [2], (3) statistical methods [3], (4) structural methods [4], (5) neural network based methods [5], (5) hidden Markov model (HMM) based methods [6], (6) multiple expert methods [7], etc. Each category of methods has its own merits and shortcomings. In this work, we propose a fuzzy clustering approach to improve the template matching method described in [1] that demonstrated high recognition rates. Our focus is on the reduction of recognition time. The key idea of the presented scheme is to organize the template pool in a structured manner so that matching can be performed more quickly. For method 1 in [1], each stroke of a Chinese character is represented by three feature points that are uniformly sampled along the stroke trajectory. Assuming correct number of strokes, the search space is confined to the character database with the same number of strokes. The recognition is performed using the minimum distance rule. Through the hierarchical fuzzy clustering, the original template pool will have the form of a decision tree where each internal node in the tree is represented by a cluster center. Since the number of class templates in each subcluster is less than the original cluster, the recognition time can be reduced effectively. The use of fuzzy clustering is to provide a mechanism for cluster overlapping so that in the middle of the recognition process, almost 100% hit ratio can be obtained. This relies on fine tuning of the fuzzy clustering results.

The rest of this paper is organized as follows. In Section 2, the proposed method is described in detail. In Section 3, computational complexities and recognition rates are discussed. Experimental results are shown in Section 4, which is followed by the conclusions section.

2. THE HIERARCHICAL FUZZY CLUSTERING APPROACH FOR ONLINE CHINESE HANDWRITING RECOGNITION

In non-cursive online Chinese handwriting recognition, a popular approach is to choose a certain amount of feature points from each stroke and use the minimum distance rule for classification. This method is well documented in [1] and was shown to provide high recognition rates. As is true for all minimum distance – based recognition methods (or 1-nearest neighbor methods), the primary computational cost lies in the search process for the nearest class template. To speedup the recognition process, an essential method is to organize the standard templates in the database in an efficient way so that a match can be easily and quickly found. Clustering is one of the tools that can be useful for this purpose. If clusters do exist in the template pool, they should be clearly identified so that a tree search strategy can be employed to reduce the search time. Empirical evidence reveals such a clustering tendency inside Chinese characters. Inside each cluster, a

*This work was supported by the National Science Council of Taiwan, R.O.C., under grant no. NSC92-2213-E-224-046.
*email of the corresponding author: lanls@yuntech.edu.tw

representative template is constructed by calculating the mean of all those character templates in this cluster in a certain sense. Ideally, these representative templates should be far apart from one another so that mis-clustering of the input test pattern seldom occurs. On the other hand, the decision regions inside a certain cluster should also be clearly separable so that the input pattern will not be mis-classified. In practice, these two goals may not be simultaneously attained. Thus a tradeoff does exist between the recognition rate and recognition speed. A carefully designed clustering, which is performed offline before the actual recognition is done, should be performed so that the maximal computational saving is obtained at the minimum cost of recognition rate degradation.

Traditionally, clustering is used as a non-supervised classification tool. In this work, we use it in a different way, aiming to organize the template pool in a more structured manner to promote search efficiency. Here the sample pattern is no longer an unlabeled data point, but rather a representative template for a character class. Since the recognition rate will be deteriorated due to incorrect clustering, we adopt the fuzzy k-means clustering algorithm [8] as our clustering device. To provide nearly perfect clustering results, the overlapping among clusters seems to be a natural choice. As such, those object patterns located close to class decision boundaries can simultaneously belong to all those classes sharing the intersection of decision regions. For fuzzy clustering, each sample pattern belongs to a specific cluster in a certain degree, which is formally described as membership value. Consider a k-cluster problem, each sample pattern owns k membership values, associated with all the clusters. If the highest membership value is much larger than the others, the decision of cluster residence is obvious. However, in cases that the differences among membership values are not large enough, the same sample pattern may then be claimed to belong to those clusters with larger membership values at the same time. By this arrangement, the mis-clustering effect can be effectively reduced.

We now describe the overall structure of the proposed hierarchical fuzzy clustering approach for online Chinese handwriting recognition. First the online handwritten character trajectories are stored and analyzed for its features. These features mainly encompass the number of strokes, the written stroke order, and the feature point sequence. The extracted feature point vector first matched against the representative template of each cluster using the minimum distance principle. Once a match is found (corresponding to the nearest representative template), the search is transferred to that specific cluster. Inside this cluster, we repeat the same fuzzy-clustering search procedure. This process can then be repeated for a few number of levels, thus creating a hierarchical fuzzy-clustering search scheme. Since the number of templates inside a subcluster is less than that of the original, the search time (or recognition time) can be effectively reduced. However, if the clustering is incorrect at any stage, the recognition will fail and this causes degradation in recognition rate.

The idea and benefits of this hierarchical fuzzy-clustering recognition procedure can be illustrated clearly graphically, as illustrated in Figure 1(a) through 1(e). Figure 1(a) shows the non-clustered case. In this case, there are totally 17 class templates. To reach a direct recognition decision, we need to perform 17 distance calculations in order to find the nearest neighbor. Figures 1(b) through 1(e) show the fuzzy-clustered case. Figure 1(b) shows the result using the first-level fuzzy clustering. The two clusters have 8 and 11 class templates, respectively, with two class templates in common. Since the input test pattern is closer to the center of the second cluster, the search continues in the second cluster. Figure 1(c) shows the result using the second-level fuzzy clustering. The two clusters have 7 and 5 class templates, respectively, with one class template in common. Since the input test pattern is closer to the center of the first cluster, the search continues in the first cluster. Figure 1(d) shows the result using the third-level fuzzy clustering. The two clusters both have 5 class templates, with no class template in common. Since the input test pattern is closer to the center of the first cluster, the search continues in the first cluster. Finally, in Figure 1(e) no further fuzzy clustering is used, the recognition is performed by direct search. The numbers of distance calculations from Figures 1(b) through 1(e) are 2, 2, 2, and 3, respectively, which add up to 9. Therefore in this illustrative example, the number of distance calculations, or effectively the recognition time, is reduced from 17 to 9. The reduction of computational complexities will be discussed in more detail in the succeeding section.

3. COMPUTATIONAL COMPLEXITIES AND RECOGNITION RATES

Assume there are altogether N character classes, and we have performed Q-level fuzzy clustering in advance. In every level of the clustering process, each cluster is divided into K subclusters. For simplicity of derivation, we further assume that the number of overlapped class templates at each clustering level is very small which can be neglected. The number of distance calculations needed in the proposed hierarchical fuzzy-clustering search scheme is then given by

$$KQ + \frac{N}{K^Q} \qquad (1)$$

In contrast, for the non-clustered direct search approach the number of distance calculations is equal to N. For very large N, small K and Q, the asymptotic value of (1) can be approximated by N/K^Q. We see clearly that for large values of N, the savings in the distance calculations are obvious.

In principle, the probability of recognition error can be mathematically formulated as

$$P(\text{error}) = 1 - \sum_{k=1}^{N} P(\omega_k) \int_{\Re_k} p(\mathbf{x} \mid \omega_k) d\mathbf{x} \quad (2)$$

For an n-dimensional feature vector, since the decision regions are irregularly shaped, the integral in Eq.(2) is extremely difficult to exactly evaluate. Here, we illustrate the change in recognition rate graphically. In Figures 2(a) through 2(c), each figure contains 5 class templates. Figure 2(a) shows the decision boundaries for the non-clustering case. Figure 2(b) shows the decision boundaries for the crisp-clustering case. The fuzzy clustering case is shown in Figure 2(c). It is seen that the decision region of the center class template in Figure 2(b) has changed greatly compared to that of Figure 2(a). This will result in observable degradation in recognition rates. In contrast, the decision regions between Figures 2(a) and 2(c) are quite similar. This indicates similar recognition rate performance between the two schemes.

4. EXPERIMENTAL RESULTS

The 4-stroke group of characters was considered in our experiments. There are altogether 91 characters inside this character group. Five sets of test characters with different writing characteristics were collected.

4.1 Fuzzy Clustering

We empirically determined a threshold, 0.2, to choose those overlapped class templates among clusters. Table 1 illustrates this operation. In Table 1, the two largest membership values have a difference less than 0.2, and therefore the class template “支” is viewed as being a true member of both clusters 1 and 3.

4.2 Clustering Correctness

The clustering correctness is defined as the percentage that the input test patterns are classified into the correct clusters. Table 2 lists the results of clustering correctness. Almost perfect clustering was obtained, with only one exception from test set 5. The correctness of clustering is important in the sense to maintain the original recognition rate of the non-clustered scheme.

4.3 Recognition Performance

Without considering character rejection, the recognition rate is defined as the ratio between the number of correctly recognized characters and the number of characters to be recognized. For the purpose of comparison, the recognition rates of the conventional non-clustering method are also shown here. Table 3 lists the results of recognition rates. From this table, we find that the recognition rate performance of the proposed hierarchical fuzzy-clustering scheme is nearly the same as the non-clustered one. By carefully fine tuning of the fuzzy clustering, the loss in recognition rates can be kept at an unobservable value.

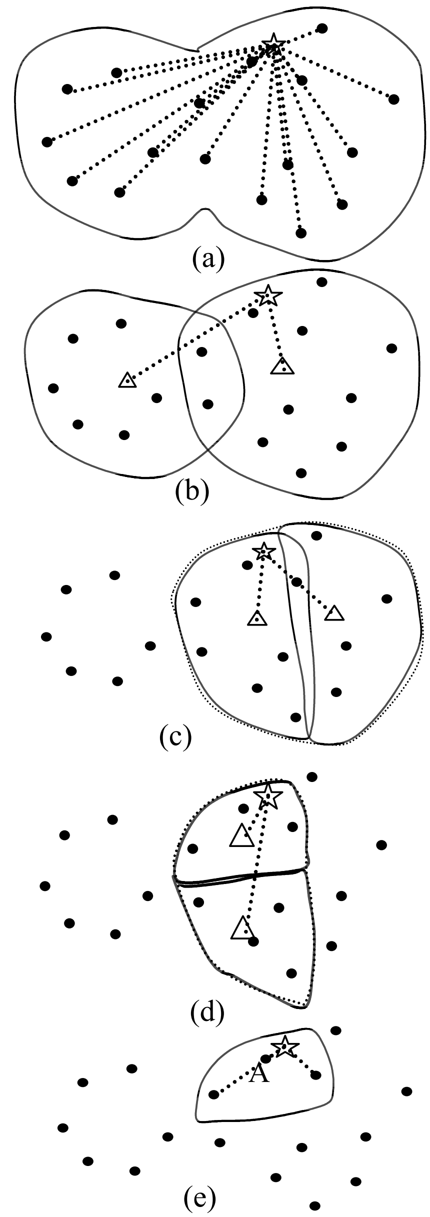

Figure 1. An illustrative example of recognition using the hierarchical fuzzy clustering approach. In each figure, the star represents the input test pattern, and the triangles represent cluster centers. The “A” in Figure 1(e) is the final recognition result.

5 CONCLUSIONS

In this paper, we propose a fuzzy clustering approach to improve the template matching method based on feature point sequences described in [1] that demonstrated high recognition rates. The focus is on the reduction of

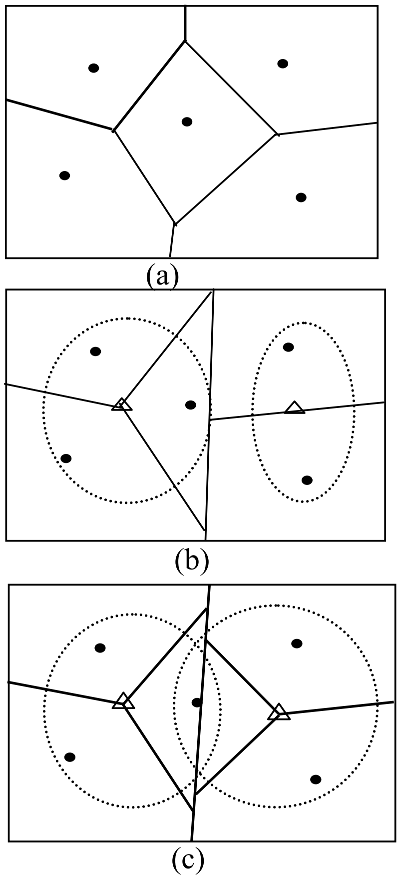

Figure 2. An illustrative example demonstrating the changes in decision regions for various schemes. (a) the conventional non-clustered scheme. (b) the crisp-clustering scheme. (c) the fuzzy-clustering scheme.

recognition time. The key idea of the presented scheme is to organize the template pool more structurally so that matching can be performed easily to speedup the recognition process. An input pattern first matches against the representative templates of clusters, and then searches in the specific subcluster. Since the number of standard templates in each subcluster is less than the original non-clustered template pool, the recognition time can be reduced. An approximate formula for the number of distance calculations needed for the presented scheme is also derived. The use of fuzzy clustering is to provide a mechanism for cluster overlapping so that in the middle of the recognition process, almost 100% hit ratio can be obtained. This relies on fine tuning of the fuzzy clustering results.

Although not mathematically verified, we examine the possible degradation in recognition rates by inspecting the decision boundaries graphically. The changes of decision regions will directly influence the recognition rates. For the example given in Section 3, it was shown that the differences between the conventional non-clustered scheme and the proposed hierarchical fuzzy clustering scheme may be non-noticeable. We then expect a strictly minor loss in the performance of recognition rate. The experiments conducted justified our intuition. For the 4-stroke character group, there is only 0.2% loss of recognition rate in average, while we obtained approximately a 3.4 times reduction of recognition time (using the simplified formula given in Eq. (1)).

5. REFERENCES

[1] T. Wakahara, H. Murase, and K. Odaka, "On-Line Handwriting Recognition," *Proc. IEEE*, Vol. 80, No. 7, pp. 1181-1194, July 1992.

[2] T. Wakahara, and K. Odaka, "On-Line Cursive Kanji Character Recognition Using Stroke-Based Affine Transformation, "*IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 19(12), pp. 1381-1385, Dec. 1997.

[3] M. Umeda, "Advances in Recognition Methods for Handwritten Kanji Characters," *IEICE Tr. Information and Systems*, Vol. 79-D, No. 5, pp.401-410, 1996.

[4] K. J. Chen, K. K. Li, and Y. L. Chang, "A System for On-Line Recognition of Chinese Characters," *Int'l J. Pattern Recognition and Artificial Intelligence*, Vol. 2, pp. 139-148, 1988.

[5] H. J. Kim, J. W. Jung, and S. K. Kim, "On-line Chinese character recognition using ART-based stroke classification," *Pattern Recognition Letters*, Vol. 17 (12), pp. 1311-1322, Dec. 1996.

[6] J. Hu, M. K. Brown, and W. Turin, "HMM based online handwriting recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 18(10), pp. 1039 ~ 1045, Oct. 1996.

[7] V.D. Mazurov, A. I. Krivonogov, and V.L. Kazantsev, "Solving of optimization and identification problems by the committee methods," *Pattern Recognition*, Vol. 20(4), pp. 371 ~ 378, 1987.

[8] J. C. Bezdek, *Fuzzy Mathematics in Pattern Classification*, PhD thesis, Applied Math. Center, Cornell University, 1973.

Table 1. An example demonstrating how a class template simultaneously belong to more than one subclusters (membership values are shown)

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
支	0.4475	0.1057	0.2898	0.1569

Table 2. Clustering correctness for the 4-stroke character group

	Test 1	Test 2	Test 3	Test 4	Test 5
1st-level	100%	100%	100%	100%	100%
2nd-level	100%	100%	100%	98.9%	100%

Table 3. Comparison of recognition rates of the proposed hierarchical fuzzy clustering scheme and the non-clustered scheme

	Test 1	Test 2	Test 3	Test 4	Test 5	Avg
proposed scheme	90.1%	91.2%	89.0%	86.8%	87.9%	89.0%
non-clustered scheme	90.1%	91.2%	89.0%	87.9%	87.9%	89.2%