Multiple Instance Learning Based Method for Similar Handwritten Chinese Characters Discrimination

Yunxue Shao, Chunheng Wang, Baihua Xiao, Rongguo Zhang, Yang Zhang
The Key Laboratory of Complex Systems and Intelligence Science
Institute of Automation Chinese Academy of Sciences
95 Zhongguancun East Road, 100190, BEIJING, CHINA
{ yunxue.shao, chunheng.wang, baihua.xiao, rongguo.zhang, yang.zhang }@ia.ac.cn

***Abstract*—This paper proposes a Multiple Instance Learning based method for similar handwritten Chinese characters discrimination. The similar handwritten Chinese characters recognition problem is first defined as a Multiple-instance learning problem. Then the problem is solved by the AdaBoost framework. The proposed method selects some self-adapting critical regions as weak classifiers, and therefore it is more suitable for the wide variability of writing styles. Our experimental results demonstrate that the proposed method outperforms the other state-of-the-art methods.**

Keywords- multiple instance learning; similar character recognition; self adapting critical region; critical instance

I. INTRODUCTION

The problem of offline handwritten Chinese character recognition has been investigated by many researchers over a long time, and great improvements have been achieved [1-4]. Despite the success of existing methods, there are still rooms for improvement. One of the problems is the classification of similar characters. Similar characters differ only in local details, e.g., "千" and "干", "王" and "玉", etc. The compound Mahalanobis function (CMF) method, proposed by Suzuki et al. [5], combines pair discrimination measures with class-wise Mahalanobis distance. This method is unique in that the pair discriminator has no extra parameters when using the MQDF as the baseline classifier. Tian-Fu Gao and Cheng-Lin Liu [6, 7] proposed a LDA-based compound distance to discriminate similar character pairs. They showed that under restrictive assumptions, the previous CMF is a special case of the LDA-based compound distance method. Their experiments demonstrated that the LDA-based compound distance method outperforms the previous CMF methods. The approach adopted by Gao and Liu [6, 7] is to use pair-wise discriminant functions for discriminating between similar character classes. A total of about 27,000-31,000 pairs of discriminant function were used in their experiment on the ETL-9B database. K.C.Leung and C.H.Leung [8] proposed the critical region analysis method to tackle the problem of similar character classes. Additional features are extracted from these critical regions and used to train the similar character set. Bo Xu et al. [9] proposed an Average Symmetric Uncertainty based critical region selection method and they showed that the critical regions selected by their method contain more discriminative information than by the method proposed in [8].

As we know, the critical region based methods [8][9] select some critical regions with the region's scale and position fixed for each similar character pair. However, both the critical region's scale and position change due to the wide variability of writing styles. The method proposed in this paper tries to tackle these shortages. As we can see in this paper, the proposed method selects some self adapting critical regions and uses them as weak classifiers. The MQDF [10] has been widely used to handwritten Chinese character recognition with great success. Most of the methods proposed in recent years use it as a baseline classifier. In this paper, we take it as the baseline classifier and discriminate the similar pairs by the proposed method.

In the following sections, a short introduction to the methods proposed in [8] [9] is given in Section 2, and in Section 3, the MIL based similar characters discrimination method is proposed, followed by experimental results in Section 4 and conclusions in Section 5.

II. TWO CRITICAL REGION SELECTION METHODS

In this section, two critical region selection methods are briefly reviewed. One is the Fisher's linear discriminant based critical region selection method [9]. The other is the Average Symmetric Uncertainty based critical region selection method [8].

A. Fisher's Linear Discriminant Based Critical Region Selection

Paper [9] exploits the fact that Fisher's linear discriminant can be used to find a projection axis that optimally separates two classes. The critical regions are located by examining the vector components of this projection axis.

Denote the mean vectors of the two classes by m_1 and m_2, and the within-class and between-class scatter matrices by S_W and S_B respectively. Let w denote the optimal projection axis, then the criterion function $J(w)$ that measures the separability between the two classes can be defined as:

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.203

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \quad (1)$$

It can be shown that the w that maximizes $J(w)$ can be directly computed from the equation:

$$w = S_W^{-1}(m_1 - m_2) \quad (2)$$

The critical regions of a similar pair are located by identifying the components of w that have large absolute magnitudes.

B. Average Symmetric Uncertainty Based Critical Region Selection

In paper [8], the Average Symmetric Uncertainty (ASU) is defined as the Symmetric Uncertainty between a region and the class label. Formula (3) gives the ASU of the *i*th region:

$$ASU_i = \frac{1}{n}\sum_{j=1}^{n} SU_{ij} \quad (3)$$

Where n is the number of features in the region and

$$SU_{ij} = 2\left[\frac{I(X_{ij};Y)}{H(X_{ij}) + H(Y)}\right] \quad (4)$$

Where X_{ij} is the *j*th feature in the *i*th region and Y is the class label. $I(X;Y)$ is the information gain of X and Y and $H(X)$ is entropy of X.

The mean of ASUs in all regions is given as:

$$MASU = \frac{1}{N}\sum_{i=1}^{N} ASU_i \quad (5)$$

Regions are selected if the ASUs in them are higher than a threshold T. Where T is computed by:

$$T = \alpha * MASU, \alpha > 0 \quad (6)$$

III. Character Discrimination Based on Multiple Instance Learning

In this section, a short induction to the multiple instance learning is given. Then the definition of our problem and a framework for solving this problem is proposed.

A. Multiple Instance Learning

Multiple Instance Learning (MIL) is a variation of supervised learning for problems with incomplete knowledge about the labels of examples. In MIL, all training samples are given as bags of instances. Only the bags are labeled. A bag is labeled positive if at least one instance in the bag is positive, and a bag is labeled negative if all the instances in it are negative. The multiple-instance learning model is first introduced by Dietterich et al. [11]. After that, a lot of methods have been proposed for solving multiple-instance learning problems, such as diverse density (DD) [12] and extended Citation kNN [13]. Recently, a number of researchers have modified the boosting algorithm to perform MIL [14] [15].

B. Problem Definition

If a character image is denoted as an image bag, and patches in the image as instances within the bag, then the problem of similar handwritten Chinese character discrimination can be defined as a MIL problem. Given two character classes C^+ and C^-, B_i^+ represents the *i*th bag in C^+, B_{ij}^+ represents the *j*th instance in the bag B_i^+. Likewise, B_i^- represents *i*th bag in C^- and B_{ij}^- represents the *j*th instance in the bag B_i^-. A bag is labeled positive if at least one instance in it is positive, and a bag is labeled negative if all the instances in it are negative. In this paper, an instance that can discriminate the two classes is called as a Critical Instance (CI). Fig.1 shows an example of CI between "王" and "玉".

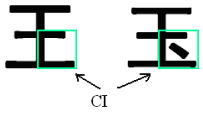

Figure 1. an example of critical instance

The other critical region based methods are based on the priori that the CIs are fixed in position and scale. So they try to find CIs for a similar pair as accurately as possible and use them to discriminate this similar pair. However, because of the wide variability of writing styles, a CI's position often shifts in a certain scope and its scale often changes, too. This is the motivation of the proposed MIL based method.

The original multiple-instance learning model does not restrict the CI's position and scale at all. For the sake of more suitable for our problem, the original multiple-instance learning model is modified as follows: B_i^+ represents a image bag in class C^+. $I_{ij}^+(x_1, y_1)$ represents the *j*th instance in bag B_i^+ and the instance's position in the previous image is (x_1, y_1). Divide the image bag B_i^+ into several small bags. $B_{ij}^+(x_c, y_c, s)$ represents the *j*th small bag in B_i^+, where x_c and y_c represent the center position of $B_{ij}^+(x_c, y_c, s)$ in the previous image and s represents the small bag's scale. An instance $I_{ij}^+(x_1, y_1)$ belongs to $B_{ij}^+(x_c, y_c, s)$ if

$$\sqrt{(x_1 - x_c)^2 + (y_1 - y_c)^2} < s \quad (7)$$

The distance between an instance I_0 and a small bag $B_{ij}^{+}(x_c, y_c, s)$ is given as:

$$d(I_0, B_{ij}^{+}(x_c, y_c, s)) = \min_{I \in B_{ij}^{+}(x_c, y_c, s)} d(I_0, I) \quad (8)$$

$d(I_0, I)$ is a chosen distance metric. Similar definitions can be given to class C^{-}. A bag is labeled positive if at least one instance in a chosen small bag $B_{ij}^{+}(x_c, y_c, s)$ is positive, and a bag is labeled negative if all instances in the chosen small bag are negative. The chosen small bag is called critical region (CR), and the positive instance in it is called critical instance (CI). Fig.2 gives an illustration of our problem definition. Image bag A and B belong to the same class. The gray circle represents the chosen small bag CR. Rectangles in the image bag represent the instances and the blue one indicates the CI within the CR. The CI in different bags may changes both in position and scale. Our goal is to find all CIs and CRs and use them to discriminate C^{+} and C^{-}.

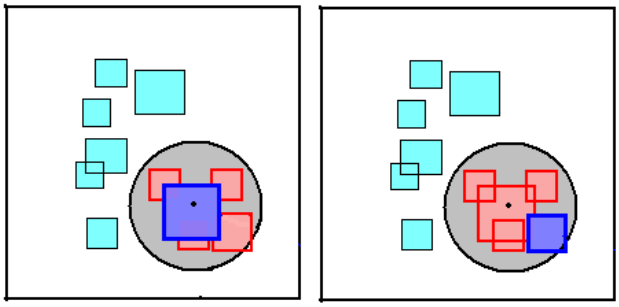

Figure 2. Illustration of the problem definition for MIL based handwritten Chinese character discrimination.

C. *Solving the Problem by Boosting.*

In subsection A, we know that our goal is to find a critical region and a critical instance that can best discriminate the two class C^{+} and C^{-} under the current distribution. This can be done by a boosting approach. In our experiment, the AdaBoost [17] framework is adopted to select the best CI and CR and update the sample's distribution. The framework of AdaBoost is:

- Initialize the distribution D.
- Find a weak classifier h that minimizes the error with respect to the current distribution D.
- Update the distribution according to h.
- Repeat 2-3 step and stop while a stopping criterion reached.

The weak classifier used in this paper is:

$$h(I, B(x, y, s)) = \begin{cases} 1, & d(I, B(x, y, s)) \times p_w < T_w \times p_w \\ -1, & \text{otherwise} \end{cases} \quad (9)$$

Where I is an instance, $B(x, y, s)$ represents a small bag, T_w is a threshold and $p_w \in \{1, -1\}$.

Given an instance set S_I and a small bag set S_B, a instance and small bag pair $(I_0, B_0(x, y, s)) \in S_I \times S_B$ and its corresponding parameters T_w and p_w that minimize the error at the current distribution D can be selected, and then the weak classifier $h(I_0, B_0(x, y, s))$ is used to update the sample's distribution. In each iteration, the $(I_0, B_0(x, y, s))$ pair might not be the best one in the whole space. However, it must be the pair in the candidate set which is nearest to the perfect one. This is enough for solving the problem.

D. *Instance Extraction*

Each character image is normalized to 64*64 pixels by a normalization method. Using a sliding window, 141 windows are extracted for each image bag. The window's width s is set to 16, 24 and 32, and one third of the window's width is used as the sliding step. Each window is then divided into 2*2 sub blocks, and within each sub block 8-direction gradient features are extracted, resulting in a 32 dimensional instance. The instance's position (x,y) is the sliding window's row and column coordinate. 141 instances are extracted for each image bag. Integral images can be used to speed up this process.

E. *Determination of the Candidate Instance Set and the Small Bag Set.*

In subsection *B*, solving the MIL based character discrimination problem needs to determine an instance set S_I and a small bag set S_B for each similar pair. In this paper, a uniform candidate instance set S_I is constructed for all similar pairs by clustering instances in all training bags based on the following reasons:

- In handwritten Chinese character recognition, the number of similar pairs is very large. If we store the candidate instance set and weak classifiers for each similar pair, it will require a lot of memory when recognition.
- Experience and experimental results tell us that the number of patches that compose the Chinese character is much fewer than the number of Chinese characters.

Using this public instance set, every weak classifier just needs to store its error rate, T_w, p_w and the corresponding small bag's position (x,y) and the corresponding instance's index. About 12 Bytes memory is required for every weak classifier. The memory requirement for each similar pair is $12 \times N_w$ (Bytes), where N_w is the average number of weak classifiers for each similar pair. Our experiments demonstrate that this strategy can save a lot of memory while keeping a good performance.

The clustering algorithm used in this paper is given in Table 1. M represents the number of character classes, N represents the number of image bags in each class, and P

represents the number of instances in each image bag. There are M*N*P instances. σ is a threshold, $d(I,e)$ is the Euclidean distance between instance I and e. $cnt(e)$ represents the number of instances falling into the ball centered at e.

TABLE I. THE CLUSTERING ALGORITHM FOR CONSTRUCTING THE CANDIDATE INSTANCE SET.

Initialize: candidate instance set $C = \phi$
Input: M*N*P instances set S
For each instance $I \in S$
  For each element $e \in C$
    If $d(I,e) < \sigma$
      Update the element $e = (e \times cnt(e) + I)/(cnt(e)+1)$
      and $cnt(e) = cnt(e)+1$.
    Else
      $C = C \cup \{e\}$ and $cnt(e) = 1$.
    End If
  End For
End For
Output C.

Using this clustering method, a uniform candidate instance set containing 609 instances is got from the experimental database.

For constructing the small bag set S_B for each similar pair, the small bag $B(x,y,s)$'s position (x,y) and scale s should first be determinate, and then find all instances belonging to the small bag. If the small bag's scale is too small, it will not suitable for the wide variability of writing styles in the same class. If the scale is too big, it may not distinguish some similar pairs well enough. In our experiments, the small bag's scale is set to 16 pixels, and the position (x,y) is:

$$x = 8*idx, idx = 1,2,...,7 \quad (10)$$

$$y = 8*idy, idy = 1,2,...,7 \quad (11)$$

After a small bag's position and scale have been fixed, we can find instances in each image bag that belong to this small bag by formula (7).

IV. EXPERIMENTAL RESULTS

We evaluate our method on the CASIA database. The CASIA database, which is collected by the institute of automation, Chinese academy of sciences, contains 3755 Chinese characters, 300 samples per class. 290 samples per class are chosen for training and the remaining 10 samples for testing. Each character image is normalized by two normalization methods. One is the line-density normalization method and another is the linear normalization method. In the following sections, DB1 and DB2 represent the database one and database two generated from the CASIA database using the line-density normalization method and the linear normalization method respectively. Some examples in DB1 and DB2 are listed in Fig.3.

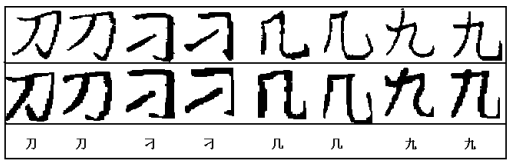

Figure 3. Examples of characters using different normalization method. Characters in the first and second row are normalized by the linear normalization method and the line-density normalization method, respectively. Characters in the third row are the corresponding printed characters.

Just for experiment, 3849 similar pairs are selected by the error rate on the training set DB1 using the linear discriminant function (LDF) classifier. S_p will be used to represent the similar pair set.

For the baseline MQDF classifier, 8-direction gradient features [16] are extracted on DB1 and then compressed to 256 dimensions by LDA. For each input pattern x, the MQDF gives two top-rank candidate classes ω_i and ω_j. The MQDF distances from x to ω_i and ω_j are d_i and d_j, respectively. If $|d_i - d_j| \le T_c$, where T_c is a threshold, then we check if ω_i and ω_j is a similar pair in set S_p, if it is then we discriminate ω_i and ω_j with the methods proposed in [8] [9] and the method proposed in this paper on database DB1 and DB2, respectively. Table 2 shows the number of similar pairs with different T_c.

TABLE II. NUMBERS OF SIMILAR PAIRS WITH DIFFERENT THRESHOLD.

T_c	20	40	60	80	100
N	867	1559	2219	2974	3731

The AdaBoost's stopping criterion used in our experiments is "iterating T times then stop training". The proposed method is trained on both DB1 and DB2 and use MIL-DB1 and MIL-DB2 to denote the classifiers trained on DB1 and DB2, respectively. We compare MIL-DB1 and MIL-DB2 with the method proposed in [8] [9] which are denoted as Fisher-DB1, Fisher-DB2, ASU-DB1 and ASU-DB2 respectively.

Fig.4 shows the test accuracies of MIL-DB1, MIL-DB2, Fisher-DB1, Fisher-DB2, ASU-DB1 and ASU-DB2 on database DB1 and DB2 with different threshold T_c. The AdaBoost's iterating times T is set to 31. Fisher-DB1, Fisher-DB2, ASU-DB1 and ASU-DB2 are implemented as in paper [9]. From Fig.4, we can see that the accuracy of MIL based method outperforms the Fisher and ASU based methods on both DB1 and DB2. All methods trained on DB2 outperform the corresponding method trained on DB1. The

reason probably is the non-linear normalization method distorts the character shape excessively.

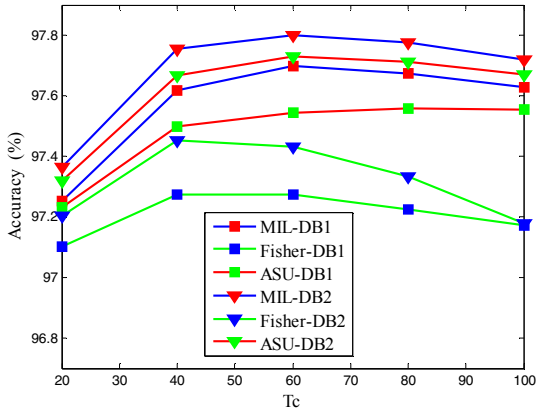

Figure 4. Test accuracies on DB1 and DB2 using MQDF as baseline classifier.

Fig.5 shows the test accuracy of MIL-DB2 with different T for T_c =60. From Fig.6 we can see that the recognition accuracy improves with a large T though accompanied with increased storage.

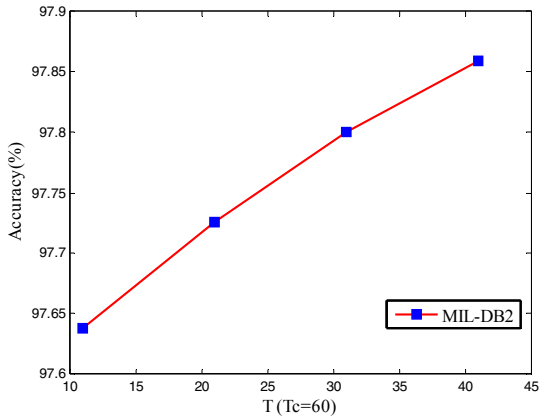

Figure 5. Test accuracies of MIL-DB2 using MQDF as baseline classifier with varying number T.

V. CONCLUSION

This paper proposes a Multiple Instance Learning based method for similar handwritten Chinese characters discrimination. Our experimental results demonstrate that the proposed method requires very little memory and is more suitable for the wide variability of writing styles. More detailed analysis about the parameters used in the proposed method will be concerned in our future work.

ACKNOWLEDGMENT

This work is supported by National Science Foundation of China (NSFC) under grants no.60802055, no.60835001 and no.60933010.

REFERENCES

[1] J. Tsukumo, H. Tanaka, Classification of hand printed Chinese characters using non-linear normalization and correlation methods, in: Proceedings of the Ninth International Conference on Pattern Recognition, Roma, Italy, 1988, pp. 168–171.

[2] C.-L. Liu and K. Marukawa, Pseudo two dimensional shape normalization methods for handwritten Chinese character recognition, Pattern Recognition 38 (12) (2005), pp. 2242–2255.

[3] C.-L. Liu, High accuracy handwritten Chinese character recognition using quadratic classifiers with discriminative feature extraction, in: Proceedings of the 18th International Conference on Pattern Recognition, vol. 2, Hong Kong, 2006, pp. 942–945.

[4] K.-C. Leung, C.-H. Leung, Recognition of handwritten Chinese characters by combining regularization, Fisher's discriminant and distorted sample generation, in: Proceedings of the 10th International Conference on Document Analysis and Recognition, Barcelona, 2009, pp. 1026–1030.

[5] M. Suzuki, S. Ohmachi, N. Kato, H. Aso and Y. Nemoto, A discrimination method of similar characters using compound Mahalanobis function, Trans. IEICE Jpn J80-D-II (10) (1997), pp. 2752–2760.

[6] T.-F. Gao, C.-L. Liu, LDA-based compound distance for handwritten Chinese character recognition, in: Proceedings of the 9th ICDAR, Curitiba, Brazil, 2007, pp. 904–909.

[7] T.-F.Gao. C.-L.Liu, High accuracy handwritten Chinese character recognition using LDA-based compound distances, Pattern Recognition 41 (11) (2008) 3442-3451.

[8] K.C. Leung, C.H. Leung, Recognition of handwritten Chinese characters by critical region analysis, Pattern Recognition, Volume 43, Issue 3, March 2010, Pages 949-961, ISSN 0031-3203, DOI: 10.1016/j.patcog.2009.09.001.

[9] Bo Xu, Kaizhu Huang, Cheng-Lin Liu, "Similar Handwritten Chinese Characters Recognition by Critical Region Selection Based on Average Symmetric Uncertainty," icfhr, pp.527-532, 2010 12th International Conference on Frontiers in Handwriting Recognition, 2010

[10] F. Kimura, K. Takashina, S. Tsuruoka and Y. Miyake, Modified quadratic discriminant functions and the application to Chinese character recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 9 (1) (1987), pp. 149–153.

[11] T. G. Dietterich, R. H. Lathrop, T. Lozano-Perez. Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. Artificial Intelligence Journal, 89, 1997.

[12] Oded Maron , Tomás Lozano-Pérez, A framework for multiple instance learning, Proc. of the 1997 Conf. on Advances in Neural Information Processing Systems 10, p.570-576, 1998

[13] J. Wang and J.-D. Zucker. Solving the multiple-instance problem: a lazy learning approach. Proc. 17th Int'l Conf. on Machine Learning, pp. 1119-1125, 2000

[14] P. Auer and R. Ortner. A boosting approach to multiple instance learning. In Lecture Notes in Computer Science, volume 3201, pages 63–74, October 2004.

[15] Paul Viola, John Platt, and Cha Zhang. Multiple instance boosting for object detection. In Y. Weiss, B. Schölkopf, and J. Platt, editors, Advances in Neural Information Processing Systems 18, pages 1417–1424. MIT Press, Cambridge, MA, 2006.

[16] Cheng-Lin Liu, Kazuki Nakashima, Hiroshi Sako, Hiromichi Fujisawa, Handwritten digit recognition: investigation of normalization and feature extraction techniques, Pattern Recognition, Volume 37, Issue 2, February 2004, Pages 265-279, ISSN 0031-3203.

[17] Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. In: J. Comput. Syst. Sci. 55(1): 119−139 (1997).