

## Automatic Discrimination between Confusing Classes with Writing Styles Verification in Arabic Handwritten Numeral Recognition

Chun Lei He Louisa Lam Ching Y. Suen

Centre for Pattern Recognition and Machine Intelligence

Department of Computer Science and Software Engineering, Concordia University

Montreal, Quebec, H3G 1M8, Canada

Emails: {cl\_he, llam, suen} @ cenparmi.concordia.ca

### Abstract

In handwriting recognition, confusing/conflicting writing styles can result in irreducible errors, so the study of writing style consistencies is important for applications. In Arabic Handwritten Numeral Recognition, most errors occur between samples of classes two and three due to their very similar shapes in some writing styles. In this paper, an automated writing style detection process is effectively implemented in the pair-wise verification of samples in these two classes. As a result, the recognition results have improved significantly with a reduction by 25% of previous errors. With rejection, when the LDA (Linear Discriminant Analysis) measurement rejection threshold is adjusted to maintain the same error rate, the recognition rate increases from 96.87% to 97.81%.

### 1. Introduction

Ambiguous shapes that result in confusing pairs of handwritten characters often cause irreducible errors in the recognition process. In particular, the Arabic numerals two and three can be written in almost the same form as shown by some real samples in Figure 1. This confusion may account for the lower performances in Arabic Handwritten Numeral Recognition (AHNR) when compared with handwritten numeral recognition in general [1].

In handwriting recognition, some researchers have applied different strategies to distinguish between confusing pairs. For example, Zhang et al. [2] designed a method based on multi-modal discriminant analysis to reduce the feature dimensionality in order to verify the recognition result of handwritten numerals within

confusing pairs, while Rahman *et al* [3] applied combinations of multiple experts to the confusing pairs. However, these methodologies cannot solve the problems in AHNR effectively due to the overlapping of shapes between classes 2 and 3. More information besides shapes would need to be extracted so that samples in these two classes could be classified correctly.

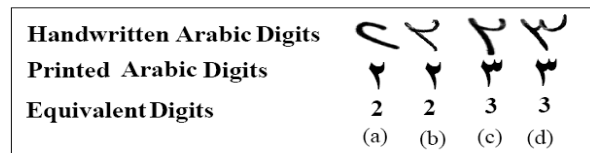


Figure 1. Samples of Handwritten Arabic numerals “2” and “3”

We assume that a writer would not confuse him/herself by writing samples of two different classes with the same shape or style (such as the handwritten samples of 2 and 3 shown in Fig. 1 (b) and (c), which should have been produced by different writers). This means writers could be grouped based on their writing styles in this confusing pair, and this prior knowledge can result in more accurate recognition performance. In fact, some researchers have applied the writer/writing information in handwriting recognition. In [4], a writer adaptive training is proposed with a character dependent Hidden Markov Model (HMM) in Offline Arabic Word Recognition so that writers’ writing can be learned in training and utilized in testing. Huang *et al* [5] utilized a writer-dependent system in online handwriting recognition with Incremental Linear Discriminant Analysis (ILDA), while Vuori [6] clustered writing styles in online mode on over 700 objects with a self-organizing map.

However, most of these researchers adapted their systems for each writer in the training procedure. We note that writers can be grouped based on their writing styles, so it would not be necessary to learn the style of each writer in training. Instead, writing styles can be categorized and this knowledge can be applied to correctly classify the ambiguous shapes encountered.

It is possible to implement this process by recording some writer information during the data collection process. This is the case for the Isolated Arabic Numeral Database at CENPARMI, in which an ID had been assigned to each writer. This enabled us to design an unsupervised learning (clustering) process that makes use of the Writing Styles (WS) Information to validate the recognition results.

We first briefly describe the supervised learning processes for the CENPARMI Arabic Isolated Numerals Database (Section 2). In Section 3, we define a Confusing Pair (CP) of clusters and a Writing Style (WS), and devise a methodology to automatically detect a CP and WS with unsupervised learning. The framework of the process is summarized in Section 4, the experiments and error analysis based on writing style verification are described in Section 5, and a conclusion is given in Section 6.

## 2. Supervised Learning in AHRN

In the recognition process, the standard procedures of image pre-processing, feature extraction, and classification were implemented. In image pre-processing, we performed noise removal, grayscale normalization, and size normalization. Gradient features were extracted from the gray-scale images, and Support Vector Machines (SVM) was chosen as a classifier with a Radial Basis Function (RBF) kernel.

To implement the rejection, which can be considered to be a two-class problem of accepting the classification result or not, Linear Discriminant Analysis (LDA) is applied to determine the rejection threshold. A Linear Discriminant Analysis based Measurement (LDAM) [is designed to take into consideration the confidence values of the classifier outputs and the relations between them. Details are described in [7].

## 3. Writing styles design

In this study, we apply unsupervised learning (clustering) within each of the two confusing classes (2's and 3's), and the number of clusters can be determined automatically. Clusters of different classes containing samples with very similar shapes would form a confusing pair (CP). Accordingly, we can

define the writing styles based on the clusters in each class. All the writers can be assigned to a group with a known writing style or a group with an unknown writing style. Then when we know the writing style of a sample, this sample can be assigned to the correct class.

### 3.1 Definitions of confusing pairs & writing styles

For a classification problem with the two classes  $W_i (i = 1, 2)$ , only the samples close to the decision boundary of its class may be confused with the data from the other class. We propose to identify these confusing samples through unsupervised learning (clustering).

Suppose that for  $i = 1, 2$ , the data from class  $W_i$  is divided into  $k_i$  clusters (sub-classes)  $\{W_i^j\}$  each with center  $c_i^j$ , where  $j = 1, 2, \dots, k_i$ . The distance between any two clusters is defined as the Euclidean distance between their centers. For  $i = 1, 2$ , define the smallest intraclass distance between clusters in  $W_i$  as

$IAD_i = \min d(W_i^m, W_i^n)$  for all  $m, n$  in  $\{1, 2, \dots, k_i\}$ ,  $m \neq n$ .

We then determine the pair of clusters  $W_1^{ii}$  in  $W_1$  and  $W_2^{jj}$  in  $W_2$  (with  $1 \leq ii \leq k_1$ ,  $1 \leq jj \leq k_2$ ) such that

$d(W_1^{ii}, W_2^{jj}) = \min d(W_1^m, W_2^n)$ , for  $1 \leq m \leq k_1$ ,  $1 \leq n \leq k_2$ .

If this minimum interclass distance  $d(W_1^{ii}, W_2^{jj}) <$  minimum intraclass distance  $IAD_i$  for  $i = 1, 2$ , then  $W_1^{ii}$  and  $W_2^{jj}$  are considered to be a confusing pair (CP) of clusters. This is reasonable because if the distance between two clusters from different classes is smaller than the distances between clusters of each class, the former two clusters would be difficult for a classifier to distinguish.

On the other hand, if clusters  $W_1^m$  and  $W_2^n$  do not form a confusing pair, then they can be considered together as a consistent style of writing 2's and 3's and this is denoted by WS.

In the following section we describe the procedure for identifying a confusing pair (CP) of clusters.

### 3.2 CP search and WS detection with unsupervised learning

We apply the well-known K-means clustering to each class iteratively until a CP is located or a stopping criterion is satisfied. Initially, each class is divided into two clusters ( $k_1 = k_2 = 2$ ), and we search for a CP. As this is based on the minimum interclass distance,

the number of such pairs should be either 0 or 1. We search until the CP is found or all clusters have been considered. If no CP is found in the search, the search is repeated with the number of clusters increased by 1 for one class. This process can continue until a pre-defined criterion (such as the maximum number of iterations) is satisfied.

Once the CP is found, the consistent writing styles WS can be determined from the consistent pairs. The statistical results of each writer's writing in each sub-class (a sub-class is a cluster of a class) is then used to assign the writer to a WS, after which his/her writing of ambiguous shapes can be assigned to the correct classes. This process is described below.

### 3.3 The CP and WS in AHR

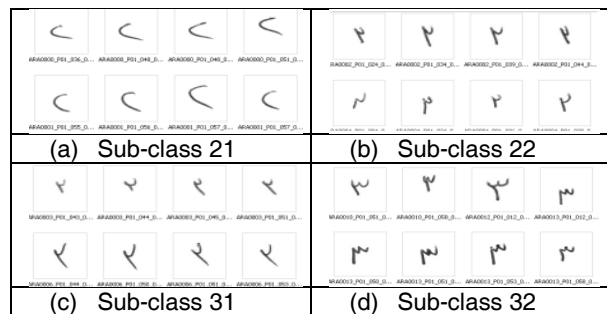
Since most recognition errors in AHR are due to confusions between samples in the Classes 2 and 3, we search for CP and determine the WS on these two classes. Initially,  $k_1 = k_2 = 2$  as stated in Section 3.2. If a CP is found in this search, the number of WS would become  $k_1 \times k_2 - 1 = 3$ . With two clusters in each class, the distances between each pair of centers are shown in Table 1, where Sub-class 21 (SC21) denotes cluster 1 of class 2, etc.

**Table 1. Distances between pairs of centers in Class 2 and Class 3**

Sub-class	21	22	31	32
21	0	6.62	6.46	7.56
22	6.62	0	<b>2.66</b>	3.63
31	6.46	2.66	0	4.42
32	7.56	3.63	4.42	0

In this case, distance  $d(SC22, SC31) = 2.66$  is the minimum interclass distance and it is also smaller than the two intraclass distances of classes 2 and 3. So SC22 and SC31 form a CP, and the search stops.

From our experiments, some randomly selected samples in each sub-class are shown in Figure 2. It is obvious that samples in SC22 and SC31 form a CP.



**Figure 2. Samples from four Sub-classes**

In this case, we can then categorize the writing of 2's and 3's to three valid combined writing styles (CWS) by eliminating the confusing combination of (SC22, SC31) with the assumption that a writer would not write 2's and 3's in almost identical shapes. Table 2 lists and shows examples of all three resulting CWS.

**Table 2. Combined Writing Styles for Classes 2 and 3**

	Class 2	Class 3
CWS I = (SC21, SC31)		
CWS II = (SC22, SC32)		
CWS III = (SC21, SC32)		
Case of Rejection	Unknown	Unknown

The cases for rejection arise when the writing styles cannot be determined due to insufficient samples from writers, or ambiguous styles are used by one writer in two classes, etc. These patterns are then rejected.

It follows that a major issue in AHR would be to distinguish between Class 3 in CWS I and Class 2 in CWS II. This issue can be resolved if the writer's CWS is known. For example, if the sample '2' originates from a writer with CWS I, then it would belong to Class 3, whereas if the writer has CWS II, then it would belong to Class 2. This means it is important to determine the CWS of a writer.

### 4. Training and testing procedures

Initially, we train a SVM classifier on the training set. For the two confusing classes of 2's and 3's, we apply the clustering process described in Section 3 to group all the samples in each class into two clusters. Then we search for the CP and detect the WS as described in Section 3. We assign the sub-class number to each pattern in the two confusing classes and re-train a sub-class classifier with all the samples in the two confusing classes. Each writer's CWS should be determined in the training step as well.

In the testing procedure, recognition and rejection with supervised learning is applied [7] as described in Section 2. The samples classified to one of the two confusing classes (2's and 3's) and rejected by the previous step should go through a verification by the sub-class classifier which returns one of the sub-classes shown in Figure 2. If this sub-class is SC22 or SC31, the sample would have ambiguous shape and can be a sample of either 2 or 3. In this case the writer's CWS can be applied to arrive at a classification as stated at the end of the last section. Whereas if the sub-class is SC21 (SC32), the sample should be assigned to class 2 (3) respectively.

## 5. Experiments and Error Analysis

Experiments with and without Writing Styles Verification have been conducted on the CENPARMI Arabic Isolated Numerals Database which contains 24,784 and 6,199 samples in the Training and Test sets, respectively. The results of the proposed method are compared with those of the algorithms presented in [7]. After applying the rejection measurement based on LDAM, when the error rate is 0.71%, the recognition rate increased from 96.87% to 97.81% with the implementation of Writing Style Verification while almost identical reliabilities are achieved as shown in Table 3. Without rejections, the recognition rate increased from 98.61% to 98.97% for the present method, and over 25% of previous wrongly classified samples can now be correctly recognized with WSV.

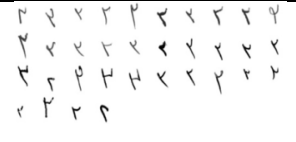
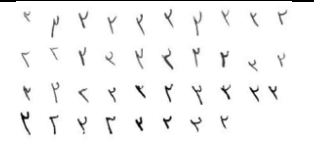
**Table 3 Performance comparisons**

	With Rejection		Without Rejection		
	[7]	Proposed Method	[1]	[7]	Proposed Method
# Correct Rate (%)	6005 (96.87)	<b>6063 (97.81)</b>	5802 (93.60)	6114 (98.63)	<b>6135 (98.97)</b>
# Errors Rate (%)	44 (0.71)	44 (0.71)	397 (6.40)	85 (1.37)	64 (1.03)
# Reject. Rate (%)	150 (2.42)	92 (1.48)	-	-	-

The two main sources of errors arise due to the fact that some writers do use contradictory styles that result in 2's and 3's being indistinguishable, and the difficulty in clustering the data accurately into sub-classes.

All the recognition errors that arose between Classes 2 and 3 in supervised learning are shown in Table 4.

**Table 4 Recognition Errors between Class 2 and Class 3 in supervised learning**

	
2 → 3	3 → 2

## 6. Conclusion

Since there is a high degree of confusion in shape between Classes 2 and 3 in AHNR, most errors in any recognition system in AHNR have been found to occur in these two classes. In this research, we designed a

verification system that can detect and correctly recognize the confusing pairs with the writing style information based on the rejections from a supervised learning process.

As a result, the recognition results improved significantly. Without rejections, 25% of the classification errors were eliminated by using writing style verification. When the rejection measurement is applied, the recognition and error rates are 96.87% and 0.71% respectively. After integrating the Writing Styles verification when we hold the error rate constant, the recognition rate increased to 97.81%.

While this approach has been motivated by and applied to the problem of Arabic numeral recognition, it can also be adapted for other pattern recognition contexts to distinguish between classes of highly similar patterns.

## Acknowledgement

The authors are grateful to Dr. Cheng-Lin Liu for his kind suggestions and comments.

## References

- [1] H. Alamri, J. Sadri, C. Y. Suen, and N. Nobile, "A Novel Comprehensive Database for Arabic Off-Line Handwriting Recognition," *Proc. 11<sup>th</sup> Int. Conf. on Frontiers in Handwriting Recognition*, Montreal, Canada, 2008, pp. 664-669.
- [2] P. Zhang, T. D. Bui, and C.Y. Suen, "Nonlinear Feature Dimensionality Reduction for Handwritten Numeral Verification," *Pattern Analysis and Applications*, Vol. 7, No. 3, 2004, pp. 296-307.
- [3] F. R. Rahman and M.C. Fairhurst, "A new hybrid approach in combining multiple experts to recognise handwritten numerals," *Pattern Recognition Letters*, vol. 18, No. 8, 1997, pp. 781-790.
- [4] P. Dreuw, D. Rybach, C. Gollan, and H. Ney, "Writer Adaptive Training and Writing Variant Model Refinement for Offline Arabic Handwriting Recognition," *Proc. 10<sup>th</sup> Int. Conf. Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 21-25.
- [5] Z. Huang, K. Ding, L. Jin, and X. Gao, "Writer Adaptive Online Handwriting Recognition Using Incremental Linear Discriminant Analysis," *Proc. 10<sup>th</sup> Int. Conf. on Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 91-95.
- [6] V. Vuori, "Clustering writing styles with a self-organizing map," *Proc. 8<sup>th</sup> Int. Workshop on Frontiers in Handwriting Recognition*, 2002, pp. 345-350.
- [7] C. L. He, L. Lam, and C. Y. Suen, "A Novel Rejection Measurement in Handwritten Numeral Recognition Based on Linear Discriminant Analysis," *Proc. 10<sup>th</sup> Int. Conf. on Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 451-455.