

SIMILARITY MEASURES FOR WRITER CLUSTERING

JAYASHREE SUBRAHMONIA

*IBM T.J. Watson Research, P.O. Box 218 / Route 134,
Yorktown Heights, NY 10598, U. S. A.*

E-mail: jays@watson.ibm.com

This paper addresses the problem of improving the performance of an online, writer-independent, large-vocabulary, unconstrained, handwriting recognition system by clustering writers with similar writing styles. Recognition performance is enhanced by identifying the writer cluster that a test writer is closest to and using a model trained for the corresponding writer cluster in decoding. The recognition system is based on hidden Markov models. A common set of features are computed for all writers, which are then projected to a lower dimensional space that preserves most of the information in the original feature set. The reduced dimensional space varies from writer to writer. This paper describes two measures of similarity between writing styles. The first is based on the distance between the writer-dependent reduced dimensional feature subspaces. The second is based on the hidden Markov Model output probabilities.

1 Introduction

Accurate, automatic, writer-independent recognition of unconstrained, large-vocabulary, online handwriting remains elusive. One way to improve the performance of these systems is to make the system parameters writer-dependent. However, large-vocabulary systems tend to have a large number of parameters, and in order to robustly estimate these parameters, a large amount of training data is needed. This implies that the test writer will have to furnish a large amount of training data to train the system accurately for their writing. This is not always a practical solution. Hence, there is interest in designing approaches to improve recognition rates with minimal amounts of data.

One approach is to identify a set of writers from the training set whose writing style is similar to that of the test writer. A system trained on training data collected from this set can then be used to recognize the test writer's handwriting.

2 Background

One well known approach for writer clustering by Schomaker and others¹ identifies writing styles based on a distance between stroke related features like cursivity index. This approach suffers from the problem that writer clustering is done independent of the models trained for them and hence there

is no guarantee that writers with similar styles will have similar handwriting models. This in turn does not guarantee improved recognition from improved handwriting models after writer clustering.

The recognizer described in this paper knows about a writing style only through the trained writer dependent model. Hidden Markov models are used to model a writer's handwriting. The clustering scheme described in this paper attempts to cluster writers in the hidden Markov model space to ensure that clustering and classification are done in the same space. Hence, two writers are considered to have similar writing styles if their trained hidden Markov models are similar.

One commonly used approach for computing the similarity between trained hidden Markov models is based on the relative entropy between discrete output probability distributions^{3,2}. This approach, which has been used for identifying gross differences in speech, has not been successfully used for identifying writing styles in handwriting. This approach also does not use all the model parameters in clustering.

The approach presented in this paper uses more of the information from the hidden Markov models in writer clustering.

3 Recognition System Overview

Online handwriting can be considered to be a collection of strokes, where a stroke is a set of time ordered (x, y) points from when a pen is placed down to when it is picked up. The IBM Online handwriting recognition system uses Hidden Markov models (HMMs) to model each character. Algorithmic details are provided in previous papers.^{4 5}

In the experiments described below, 93 distinct characters are recognized: approximately those on a typical American-English computer keyboard. The writer-dependent version of our system uses a tied pool of 200 Gaussian distributions, with one hidden Markov model per character, the 93 models totaling approximately 500 states, with no state tying.

4 Writer Clustering

To compute the writer clusters, writer dependent models were first trained for each writer. These models were then used in conjunction with a measure of similarity between writing styles to identify writer clusters.^{4,6} contains algorithmic details of training the writer dependent models. Each writer's model includes a lower dimensional feature subspace that best represents most of the information in the writing, and hidden Markov models in this space

that model the variations in writing. Since, the lower dimensional feature space varies between writers, one measure of similarity between writing styles is the distance between the lower dimensional feature spaces. Section (4.1) describes one method for computing this distance using principal angles.

Another measure of similarity between writing styles is the difference in the likelihood of the test data given different writer dependent models. Section (4.2) describes one method of computing this measure.

Finally, section (4.3) contains a description of a method for combining the two measures of similarity that gives slightly better performance compared to using the two measures of similarity in isolation.

4.1 Distance Between Feature Spaces As A Measure of Similarity Between Writing Styles

The idea of principal angles between two subspaces is well known in numerical linear algebra. We use this concept to define the distance between two subspaces.

Let F and G be two subspaces of dimension p . Then the *principal angles*, $\theta_1, \theta_2, \dots, \theta_p \in [0, \frac{\pi}{2}]$ between F and G are defined recursively by

$$\cos(\theta_k) = \max_{u \in F} \max_{v \in G} u^T v = u_k^T v_k \quad (1)$$

subject to: $\|u\| = \|v\| = 1$; $u_k^T u_i = 0, v_k^T v_i = 0 \quad i = 1 : k - 1$

The principal angles satisfy $0 \leq \theta_1 \leq \theta_2 \leq \dots \leq \theta_p \leq \frac{\pi}{2}$. The vectors $\{u_1, \dots, u_p\}$ and $\{v_1, \dots, v_p\}$ are called the *principal vectors* between the subspaces F and G .

The largest principal angle is related to the notion of distance between equidimensional subspaces, as shown below

$$\text{dist}(F, G) = \sqrt{1 - \cos(\theta_p)^2} = \sin(\theta_p) \quad (2)$$

For θ_p between 0 and 180 degrees, $\text{dist}(F, G)$ is positive definite and symmetric. Since $\text{dist}(F, G)$ does not satisfy the triangle inequality, it is not a distance measure in the strict sense.

4.2 Likelihood Ratio As a Measure of Similarity Between Writing Styles

Let K denote the number of writers in the training set. Let (M_1, M_2, \dots, M_K) denote the writer dependent models trained for the writers. Let \mathbf{X}_i and \mathbf{Y}_i

denote the training and test data for writer i . Model M_i is trained using all the training data from writer i , i.e. \mathbf{X}_i .

One measure of the similarity between writing styles for writers i and j can be computed using the likelihood ratio as follows

$$L(i, j) = \frac{p(\mathbf{Y}_i | M_j) p(\mathbf{Y}_j | M_i)}{p(\mathbf{Y}_i | M_i) p(\mathbf{Y}_j | M_j)} \quad (3)$$

$L(i, j)$ is positive definite and symmetric. Since $L(i, j)$ does not satisfy the triangle inequality, it is not a distance measure in the strict sense.

4.3 Combined Measure of Similarity Between Writing Styles

Since (2) and (3) vary between 0 and 1, we can combine the two using the following linear combination of the two distances. We have not explored other combination methods in this paper.

$$dist(i, j) = \alpha_1(1 - dist(F_i, F_j)) + \alpha_2 L(i, j) \quad (4)$$

α_1 and α_2 are parameters that normalize the range of $(1 - dist(F_i, F_j))$ and $L(i, j)$ to be between 0 and 1.

5 Experimental Results

All the handwriting data for the experiments were collected using the CrossPadTM product from the A. T. Cross Company. CrossPad consists of a digitizer on which any regular pad of paper can be placed. When one writes on the paper with an electronic pen, a copy of the writing is also stored in the memory of the digitizer. The digitizer stores the writing as a sequence of time ordered (x, y) points sampled at 133Hz. The resolution of the digitizer is 254 dpi.

Handwriting data were collected from 113 writers. Writers were instructed to write in their own natural style whether discrete, cursive or mixed; to write the words in order; and to complete one word (e. g. crossing t's and dotting i's) before beginning another. Each writer was asked to write 120 short phrases to be used as the training data and 120 short phrases to be used as test data.

For the experiments described in this paper, 13 writers were held out as a test writer pool and the rest were used for computing writer clusters. All the training data from 100 (out of the 113) writers were used to build a writer independent system. For each of the remaining 13 writers in the held out set, five recognition experiments results are presented in Table 1.

Recognition experiments were run for test data set \mathbf{Y}_j using the writer-independent model. The error rate for this experiment is considered the baseline for the writer, and all error rates in Table 1 for the writer are percent differences relative to this baseline.

WD refers to the reduction in error rate when using a writer dependent model trained for the writer. WC(1), WC(2) and WC(3) refer to reduction in error rates when using a writer cluster model trained from the training data of 10 writers with similar writing styles, using (2), (3) and (4) respectively as measures of similarity.

Table 1. Error rates using different similarity measures for clustering writing styles.

Writer	1	2	3	4	5	6	7
WD	52.52	46.42	39.59	48.74	43.90	49.91	39.59
WC(1)	37.11	30.08	29.95	28.16	27.01	35.12	27.92
WC(2)	36.23	30.13	28.60	28.90	25.07	34.99	26.13
WC(3)	39.89	32.19	31.42	29.15	28.45	36.73	27.92
Writer	8	9	10	11	12	13	Avg
WD	30.11	33.64	48.81	41.69	55.66	07.88	41.42
WC(1)	20.90	25.01	31.09	26.99	44.36	04.08	28.29
WC(2)	21.21	23.90	30.41	25.08	42.01	04.08	27.44
WC(3)	21.37	25.91	32.02	26.99	44.36	04.08	29.27

For the next set of experiments, an attempt was made to cluster the 100 writers in the training set into a single set of 10 clusters using the combined similarity measure (4). The clustering scheme that was used was as follows:

First randomly divide the 100 writers into 10 clusters, each with 10 elements. Then iterate to refine the clusters. Refinement is done by minimizing the mean intracluster similarity measure. Finally train a writer cluster model for each cluster model using all the training data from writers in that cluster.

Results of this experiment are shown in Table 2. WC is the percent difference in error rate compared to the baseline, when decoding the writer's test data set using the writer cluster model that best models the test writer's writing. The best model is the one among the 10 cluster models that gives the lowest error rate.

6 Results and Future Work

The results in Tables 1 and 2 indicate that there is a significant reduction in error rate when using writer cluster models over writer independent models

Table 2. Error rates using single set of cluster models

Writer	1	2	3	4	5	6	7
WC	33.91	24.01	26.04	24.11	25.15	33.99	23.23
Writer	8	9	10	11	12	13	Avg
WC	17.41	21.11	28.28	21.92	38.26	02.19	24.58

for users who do not have the time to write the training data needed to train a writer-dependent system.

Future work involves designing a procedure for picking the best writer cluster model from small set of test phrases and methods to use them to better bootstrap writer dependent models.

7 Acknowledgement

The author would like to thank M. P. Perrone, E. R. Ratzlaff, J. F. Pitrelli and M. Miladinov for their contributions.

References

1. L. Schomaker, G. Abbink and S. Selen, "Writer and Writer-Style Classification in the Recognition of Online Handwriting", *Proc. of the European Workshop on Handwriting Analysis and Recognition : A European Perspective*, July 94.
2. M. Padmanabhan, L. R. Bahl, D. Nahamoo, M. A. Picheny, "Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems", *IEEE Trans. on Speech and Audio Processing*, Jan. 98, pp. 71-77.
3. J. F. Foote, H. F. Silverman, "A Model Distance Measure For Talker Clustering And Identification", *Proc. of ICASSP 94*, Apr. 94, v. 1.
4. K. S. Nathan, H. S. M. Beigi, J. Subrahmonia, G. J. Clary, and H. Maruyama, "Real-time on-line unconstrained handwriting recognition using statistical methods", *Proc. of ICASSP 95*, Michigan, May 95, v. 4.
5. J. Subrahmonia, K. S. Nathan and M. Perrone, "Writer dependent recognition of on-line unconstrained handwriting", *Proc. of ICASSP 96*, Georgia, May 96, v.6.
6. J. R. Bellegarda, D. Nahamoo, and K. S. Nathan, "Supervised Hidden Markov Modeling For On-Line Handwriting Recognition", *Proceedings of ICASSP 94*, Australia, April 94, v. 5.