Off-line Writer Identification Using Gaussian Mixture Models

Andreas Schlapbach and Horst Bunke
Department of Computer Science, University of Bern
Neubrückstrasse 10, CH-3012 Bern, Switzerland
{schlpbch, bunke}@iam.unibe.ch

Abstract

Writer identification is the task of determining the author of a sample handwriting from a set of writers. In this paper, we propose Gaussian Mixture Models (GMMs) to address the task of off-line, text independent writer identification of text lines. The resulting system is compared to a system that uses a Hidden Markov Model (HMM) based approach. While the GMM based system is conceptually much simpler and faster to train than the HMM based system, it achieves a significantly higher writer identification rate of 98.46% on a data set of 4,103 text lines coming from 100 writers.

Keywords: writer identification, off-line handwriting, Gaussian Mixture Model.

1. Introduction

Significant progress has been achieved in writer identification in recent years. Surveys on early work in automatic writer identification and signature verification are given in [9, 15]. New approaches to writer identification have been proposed recently. Said et al. [18] treat the writer identification task as a texture analysis problem using multi-channel Gabor filtering and grey-scale co-occurrence matrix techniques. Srihari et al. [4, 24] address the problem of writer verification by casting it as a classification problem with two classes, *authorship* and *non-authorship*. Zois et al. [25] base their approach on single words by morphologically processing horizontal projection profiles. Hertel et al. [7] describe a system for writer identification that extracts a set of features from a text line and uses a k-NN classifier to determine the author. Edge-based directional probability distributions and connected-component contours as features for the writer identification task are proposed in [22, 23]. Bensefia et al. introduce graphemes as features for describing the individual properties of handwriting [1, 2]. Leedham et al. [10] present a set of eleven features which can be extracted easily and used for the identification and verification of documents containing handwritten digits.

In previous work we have used Hidden Markov Model (HMM) based recognizers for writer identification [20, 19]. For each writer, an individual recognizer is built and trained on text lines of that writer. This results in a number of recognizers, each of which is an expert on the handwriting of exactly one writer. The recognizers are built from character models using HMMs with 14 states that are connected in a linear topology. A mixture of five Gaussian components forms the output probability distribution. The character models are connected to word models which themselves are connected to form line models. In the identification phase, a text line of unknown origin is presented to each of these recognizers and each one returns a transcription that includes the log-likelihood score for the generated output. Using this system, a writer identification rate of 97.03% in a 100 writers experiment using 4,103 lines of text is reported in [19].

In this paper, instead of HMM based recognizers, we use Gaussian Mixture Models (GMMs) to model a person's handwriting. While GMMs have been used in the speech recognition community [16, 17], they have not been applied, to the best of our knowledge, to off-line writer identification. A GMM can be viewed as a single-state HMM with a Gaussian mixture observation density. The advantages of using GMMs over HMM based recognizers are manifold. First, GMMs are conceptually less complex than HMMs consisting of only one state and one output distribution function, which leads to significantly shorter training times. Second, in GMMs only the parameters of the output distribution function have to be estimated during training compared to HMMs where the state transition probabilities have to be estimated as well. Third, neither words nor characters have to be modeled using GMMs, because every writer is represented by exactly one model. Finally, no transcription of the text lines are needed during training.

The rest of this paper is structured as follows. In the next section we shortly introduce GMMs. In Section 3 we present our GMM based system. Section 4 presents and discusses the experimental results and Section 5 concludes the paper and presents future work.

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006

2. Gaussian Mixture Models (GMMs)

We use GMMs to model the handwriting of each person of the underlying population. The distribution of the feature vectors extracted from a person's handwriting is modeled by a Gaussian mixture density. For a D-dimensional feature vector, $\mathbf{x}$, the mixture density for a specific writer is defined as

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i p_i(\mathbf{x}). \quad (1)$$

The density is a weighted linear combination of M unimodal Gaussian densities, $p_i(\mathbf{x})$, each parameterized by a $D \times 1$ mean vector, μ_i, and a $D \times D$ covariance matrix, C_i. The parameters of a writer's density model are denoted as $\lambda = \{w_i, \mu_i, C_i\}, i = 1, \ldots, M$ where the mixture weights, w_i, sum up to one. While the general model supports full covariance matrices, we use diagonal covariance matrices in this paper as experiments have shown that they perform better than full covariance matrices [17].

The GMMs are trained using the iterative Expectation-Maximization (EM) algorithm [6]. The EM algorithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for the observed feature vectors. We apply variance flooring to impose a lower bound on the variance parameters [14].

During decoding, the feature vectors of $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ are assumed to be independent. The log-likelihood of a model λ for a sequence of feature vectors X is defined as

$$\log p(X|\lambda) = \sum_{t=1}^{T} \log p(\mathbf{x}_t|\lambda),$$

where $p(\mathbf{x}_t|\lambda)$ is computed according to Eq. 1 [17]. The GMMs are implemented using the Torch library [5].

3. A GMM Based Writer Identification System

For each writer, a GMM is built and trained with data coming from that writer only. To train the GMM, a set of features are extracted from a text line. Before feature extraction, we apply a series of normalization operations to each text line. First, contrast is enhanced. Then the writing is vertically scaled (see [21] for a detailed description of the vertical scaling operation).

Different pens of different width have been used to write the text lines. In order to eliminate the effect of the pen width on the writer identification rate, all text lines are thinned using the MB2 thinning algorithm [3]. After thinning, all lines in a text line image are at most two pixels wide.

In the next step, features are extracted using a sliding window. The window moves from left to right one pixel per step. For every column of pixels in the sliding window, nine geometrical features are extracted. The features represent the following geometrical quantities [11]: number of black pixels in the window, center of gravity, second order moment, position and contour direction of the upper and lower-most pixel, number of black-to-white transitions in the window, and fraction of pixels between the upper and lower-most pixel. The feature vectors of every column in the sliding window are averaged to produce the final feature vector. In a final step, the feature vectors which do not contain any upper and lower-most black pixels are deleted.

The window width of the sliding window was optimized in an independent validation experiment consisting of 571 text lines from 20 writers. These 20 writers were not used in the subsequent experiments. A fixed number of 100 Gaussian mixture components and a variance threshold of 0.001 were used for training. The window width was varied from 2 to 32 by steps of two. The highest writer identification rate of 99.05% was achieved using a window width of 14 pixels. This window width was used in all subsequent experiments to extract the features from a text line.

The sequence of nine-dimensional feature vectors thus obtained from each line of text is used to train the GMMs. As a result of the training procedure, we obtain for each writer a GMM that is specially adapted to the individual handwriting style of that writer. During identification, a text line to be classified is presented to the GMM of each writer. Each GMM outputs the log-likelihood score and the standard deviation for the given text line. The log-likelihood scores are sorted in decreasing order and the text line is assigned to the best ranked writer.

In applications where a wrong identification implies a high cost, it is advisable to implement a rejection mechanism [13]. For the GMM based system, we define two simple confidence measures. The first confidence measure is the difference between the log-likelihood of the first and the second best ranked writer, normalized by the length of the text line, i.e., $cm_1 = (l_1 - l_2)/length$. For the second confidence measure, instead of the length of the text line, we use the standard deviation of the first ranked writer, s_1, to normalize the score, i.e., $cm_2 = (l_1 - l_2)/s_1$. Using these confidence measures, we define the following rejection rule. If the confidence measure is above a certain threshold, we assign the text line to the best ranked writer. Otherwise, no decision about the identity of the text line is made.

4. Results and Discussion

The experiments are based on pages of handwritten text from the IAM database [12][1].The database contains over

[1]The database is publicly available at: www.iam.unibe.ch/~fki/iamDB

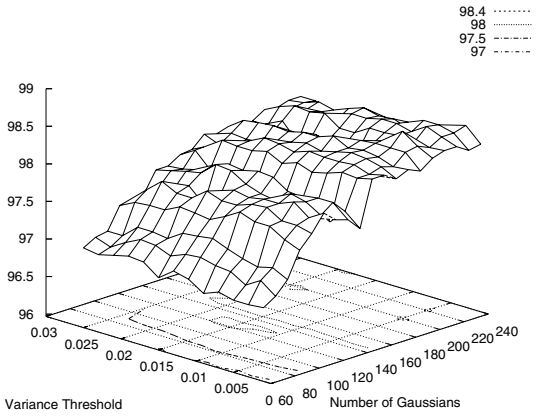

Figure 1. Writer identification rate as a function of the number of Gaussians and the variance threshold.

1,500 pages of hand written text from over 650 different writers. Each page contains between five and eleven text lines. For each writer we use five pages of text from which between 27 and 54 text lines are extracted.

The data set used in this experiment contains 4,103 text lines from 100 different writers. This data set is identical to the one used in [19]. For each writer, the set of available text lines is split into four disjoint subsets. We perform full four-fold cross validation experiments. Three subsets are used for training and the fourth subset is used for testing.

We first conducted an experiment to measure the influence of the number of Gaussians and the variance threshold on the writer identification rate. In this experiment, we systematically varied the number of Gaussian mixture components from 60 to 240 by steps of 10 and varied the variance threshold from 0.001 to 0.025 by steps of 0.002.

The results of this experiment are presented in Fig. 1. The highest writer identification rate of 98.46% is achieved using 200 Gaussians and a variance threshold of 0.005. This result is significantly better (at the statistical significance level of 99% using the K-Fold Cross-Validation Paired t-Test [8]) than the 97.03% presented in [19]. An identification rate higher than 97.03% is achieved using 70 Gaussians or more.

In Fig. 2, an n-best list is shown, where the writer identification rate based on the first n ranks is plotted. The error rate of the GMM based system drops below 1% if the first two ranks, and below 0.5% if the first five ranks are considered. For comparison, the n-best list of the HMM system

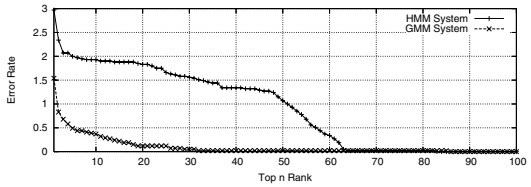

Figure 2. n-best list for the HMM and the GMM based system.

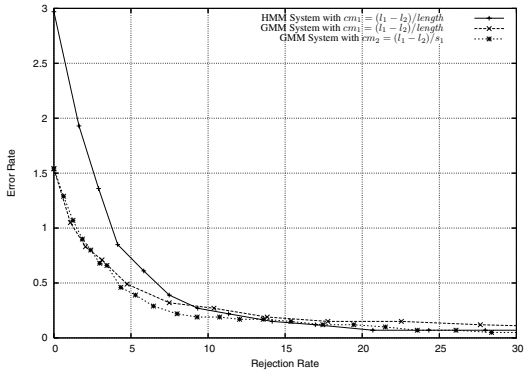

Figure 3. Error-rejection curve for the HMM and the GMM based system.

on the same data set is shown in Fig. 2 as well. The n-best list of the GMM based system clearly outperforms the one of the HMM based system.

The error-rejection curves of the HMM and the GMM based system are shown in Fig. 3. If 4% of the text lines with the lowest confidence score are rejected, then a writer identification rate higher than 99.5% is achieved by the GMM based system. The error-rejection curve of the GMM based system using the confidence measure cm_2 to reject the text lines performs slightly better than cm_1. Additionally, in Fig. 3 the error-rejection curve for the HMM based system using cm_1 is shown for comparison. The error-rejection curve produced by the GMM based system using cm_2 performs better than, or equal to, the error-rejection curve of the HMM based system.

5. Conclusions

In this paper we use Gaussian Mixture Models (GMMs) to address the task of text-independent writer identification. A GMM models the distribution of a person's handwriting for each writer. When presented with a text line of unknown origin, each GMM outputs the log-likelihood score

and standard deviation for the given input. We rank the log-likelihood scores of each model and choose the highest ranked writer.

We have tested our system on a data set consisting of 4,103 text lines coming from 100 writers. This data set is identical to the one used in [19], where a writer identification system built from Hidden Markov Model (HMM) based recognizers is presented. While the GMM based system is conceptually much simpler and takes substantially less time to train than the HMM based system, it achieves a significantly higher writer identification rate of 98.46% on this data set. Furthermore, if we consider not only the first, but the four highest ranked writers, in over 99.5% of all cases the writer of the text line is correctly identified. Applying a simple confidence measure, we obtain a writer identification rate higher than 99.5% if we reject 4% of the text lines with the lowest confidence score.

In order to compare the performance of the GMM system with the HMM based system we have used three pages of text from each writer to train the system. In future work, we plan to measure the influence of using less data to train the GMMs. A possible approach is to use an universal background model [17] and then adapt this model to each specific writer model using fewer data. Furthermore, we plan to address the task of writer verification.

References

[1] A. Bensefia, T. Paquet, and L. Heutte. Handwriting analysis for writer verification. In *Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition*, pages 196–201, 2004.

[2] A. Bensefia, T. Paquet, and L. Heutte. A writer identification and verification system. *Pattern Recognition Letters*, 26(13):2080–2092, 2005.

[3] T. M. Bernard and A. Manzanera. Improved low complexity fully parallel thinning algorithm. In *Proc. 10th Int. Conf. on Image Analysis and Processing*, pages 215 – 220, 1999.

[4] S.-H. Cha and S. Srihari. Multiple feature integration for writer verification. In *Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition*, pages 333–342, 2000.

[5] R. Collobert, S. Bengio, and J. Mariéthoz. Torch: a modular machine learning software library. Technical report, IDIAP, 2002.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, 39:1–38, 1977.

[7] C. Hertel and H. Bunke. A set of novel features for writer identification. In J. Kittler and M. Nixon, editors, *Audio-and Video-Based Biometric Person Authentication*, pages 679–687, 2003.

[8] L. I. Kuncheva. *Combining pattern classifiers: methods and algorithms*. Wiley-Interscience, 2004.

[9] F. Leclerc and R. Plamondon. Automatic signature verification: The state of the art 1989–1993. In R. Plamondon, editor, *Progress in Automatic Signature Verification*, pages 13–19. World Scientific Publ. Co., 1994.

[10] G. Leedham and S. Chachra. Writer identification using innovative binarised features of handwritten numerals. In *Proc. 7th Int. Conf. on Document Analysis and Recognition*, pages 413–417, 2003.

[11] U.-V. Marti and H. Bunke. Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. *Int. Journal of Pattern Recognition and Artificial Intelligence*, 15:65–90, 2001.

[12] U.-V. Marti and H. Bunke. The IAM–database: An English sentence database for off-line handwriting recognition. *Int. Journal of Document Analysis and Recognition*, 5:39–46, 2002.

[13] S. Marukatat, T. Artières, P. Gallinari, and B. Dorizzi. Rejection measures for handwriting sentence recognition. In *Proc. 8th Int. Conf. on Frontiers in Handwriting Recognition*, pages 25–29, 2002.

[14] H. Melin, J. Koolwaaij, J. Lindberg, and F. Bimbot. A comparative evaluation of variance flooring techniques in HMM-based speaker verification. In *Proc. of the 5th Int. Conf. on Spoken Language Processing*, pages 2379–2382, 1998.

[15] R. Plamondon and G. Lorette. Automatic signature verification and writer identification – the state of the art. In *Pattern Recognition*, volume 22, pages 107–131, 1989.

[16] D. A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17:91–108, 1995.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10:19–41, 2000.

[18] H. E. S. Said, T. Tan, and K. Baker. Personal identification based on handwriting. *Pattern Recognition*, 33:149–160, 2000.

[19] A. Schlapbach and H. Bunke. A writer identification and verification system using HMM based recognizers. *Submitted*.

[20] A. Schlapbach and H. Bunke. Off-line handwriting identification using HMM based recognizers. In *Proc. 17th Int. Conf. on Pattern Recognition*, volume 2, pages 654–658, 2004.

[21] A. Schlapbach and H. Bunke. Writer identification using an HMM-based handwriting recognition system: To normalize the input or not? In *Proc. 12th Conf. of the Int. Graphonomics Society*, pages 138–142, 2005.

[22] L. Schomaker and M. Bulacu. Automatic writer identification using connected-component contours and edge-based features of uppercase western script. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26:787–798, 2004.

[23] L. Schomaker, M. Bulacu, and K. Franke. Automatic writer identification using fragmented connected-component contours. In *Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition*, pages 185–190, 2004.

[24] B. Zhang, S. N. Srihari, and S. Lee. Individuality of handwritten characters. In *Proc. 7th Int. Conf. on Document Analysis and Recognition*, volume 7, pages 1086–1090, 2003.

[25] E. N. Zois and V. Anastassopoulos. Morphological waveform coding for writer identification. *Pattern Recognition*, 33:385–398, 2000.

0-7695-2521-0/06 $20.00

© 2006