Off-line Handwriting Identification Using HMM Based Recognizers

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

Abstract

In this paper, an off-line, text independent system for writer identification using Hidden Markov Model (HMM) based recognizers is described. For each writer we build an individual recognizer and train it on text lines written by that writer. A text line of unknown origin is presented to each of these recognizers. As a result we get, from each recognizer, a transcription including the log-likelihood score for the considered input. We rank all scores, and based on the assumption that the recognizer with the highest log-likelihood is the one that has been trained using text lines of this writer, we assign the text line to the writer whose score ranks first. We tested our system using over 2,200 text lines from 50 writers and have in 94.47% of all cases correctly identified the writer. Using a simple confidence measure to define a rejection mechanism, we achieved an error rate of 0% by rejecting 15% of the results.

Keywords: writer identification, off-line handwriting, HMM-based handwriting recognition.

1. Introduction

Writer identification is the task of determining the author of a sample handwriting from a set of writers [13]. Related to this task is writer verification, i.e., the task of determining whether or not a handwritten text has been written by a certain person. If any text may be used to establish the identity of the writer the identification task is *text independent*. Otherwise, if a writer has to write a particular predefined text to identify himself or herself the identification task is *text dependent*. Writer identification can be performed on-line, where temporal and spatial information about the writing is available, or off-line, where only a scanned image of the handwriting is available. In this paper we address the problem of text independent writer identification using off-line data. Generally it is believed that text-independent writer identification is more difficult than text-dependent writer identification. Moreover, it puts less constraints on the writer and is thus more flexible from the application oriented point of view.

There is a close relationship between the tasks of writer identification and general handwriting recognition. The field of handwriting recognition has significantly progressed in recent years [14]. For isolated word and general text recognition, Hidden Markov Models (HMMs) have become the predominant approach. In fact, HMM based recognizers have a number of advantages over other approaches [4]: First, they are resistant to noise and can cope with shape variations. Second, they allow to model characters of variable width occurring in the text. Third, HMM based recognizers are able to implicitly segment a text line into words and characters, a task that is difficult to perform explicitly [17]. Last, there exist standard algorithms for training and testing [15].

This paper is based on the idea of utilizing HMM-based handwriting recognition systems for the purpose of writer identification. For each writer in the considered population, an individual HMM based handwriting recognition system is trained using only data from that writer. Thus for n different writers we obtain n different HMMs. They all have the same architecture, but their parameters, i.e., transition and output probabilities, are different because they are trained on different data each. Intuitively, each HMM can be understood as an expert specialized in recognizing the handwriting of one particular person. Given an arbitrary line of text as input, each HMM based recognizer outputs a transcription of the input together with a recognition score. Assuming that correctly recognized words have a higher score than incorrectly recognized words, and assuming furthermore that the recognition rate of a system is higher on input from the writer the system was trained on than on input from other writers, we can use the scores produced by the different HMMs to decide who has written the input text line. We simply opt for the writer whose HMM produces the highest score.

To experimentally evaluate this approach, we use the same data as Hertel et al. [6] and thus can directly com-

0-7695-2128-2/04 $20.00 (C) 2004 IEEE

pare the results. A writer identification rate of 94.47% is obtained by our system which is superior to the 90.7% reported in [6]. Implementing a rejection mechanism based on a simple confidence measure, we can reduce the error rate to 0% at a rejection level of 15%.

The remainder of this paper is structured as follows. The next section presents related work. In Section 3 we describe our handwritten text line recognizers, and in Section 4 we show how we combine them to build a system to determine the writer of a text line. We present our experiments in Section 5 and Section 6 concludes the paper.

2. Related Work

Surveys covering work in automatic writer identification and signature verification until 1993 are given in [7, 13]. Writer identification can be understood as a classification problem where a word, text fragment, or text is to be assigned to one out of a number of possible writers. Recently, different approaches to writer identification have been proposed. Said et al. [16] treat the writer identification task as a texture analysis problem. They use global statistical features extracted from the entire image of a text using multi-channel Gabor filtering and grey-scale co-occurrence matrix techniques.

Cha et al. [5] address the problem of writer verification, i.e., the problem of determining whether two documents are written by the same person. In order to identify the writer of a given document, they model the problem as a classification problem with two classes, *authorship* and *non-authorship*. Given two handwriting samples, one of known and the other of unknown identity, the distance between two documents is computed. Then the distance value is used to classify the data as positive or negative.

Zois et al. [19] base their approach on single words by morphologically processing horizontal projection profiles. The projections are derived and processed in segments in order to increase the discrimination efficiency of the feature vectors which are then classified using either a Bayesian classifier or a neural network.

In Hertel et al. [6] a system for writer identification is described. The system first segments a given text into individual text lines and then extracts a set of features from each text line. The features are subsequently used in a k-nearest-neighbor classifier that compares the feature vector extracted from a given input text to a number of prototype vectors coming from writers with known identity.

Bulacu et al. [3] use edge-based directional probability distributions as features for the writer identification task. They introduce edge-hinge distribution as a new feature. The key point of this feature is to consider two edge fragments in the neighborhood of the central pixel and compute the joint probability distribution of the orientations of the two fragments. This feature performs better than other features they evaluated.

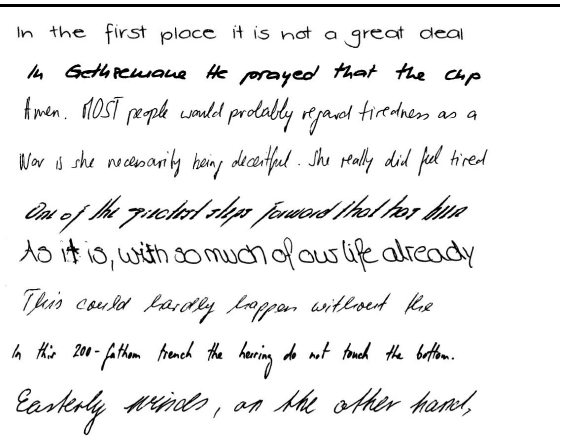

Figure 1. Text Line Examples.

In a set of papers [1, 2, 12] graphemes as features for describing the individual properties of handwriting are proposed. Furthermore, it is shown that each handwriting can be characterized by a set of invariant features called the writer's invariants. These invariants are detected using an automatic grapheme clustering procedure.

Leedham et al. [8] present a set of eleven features which can be extracted easily and used for the identification and verification of documents containing handwritten digits. These features are represented as vectors and by using the Hamming distance measure and determining a threshold value for the intra-author variation a high degree of accuracy in authorship detection is achieved.

3. Handwritten Text Line Recognizers

The system for writer identification described in this paper is based on HMM recognizers designed for the task of handwritten text line recognition. Each text line (some examples are given in Fig. 1) presented to the system is first normalized with respect to slant, skew, baseline location and height (see [9] for more detail). A sliding window is used to transform a normalized handwritten text line into a sequence of feature vectors. The window is one pixel wide and shifted from left to right over a line of text. At each position of the window, nine geometrical features are extracted. Hence the input to our HMM is a sequence of nine-dimensional feature vectors of variable length. The features represent the following geometrical quantities: 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. At first glance it may appear counterin-

tuitive to apply normalization operations because they decrease a handwriting's individuality. However, any HMM that uses a sliding window for feature extraction needs careful normalization to achieve reasonable recognition performance, and reliable text recognition is essential for successful writer identification. Hence any potential reduction in handwriting individuality is compensated for by a gain in recognition performance.

For each upper and lower case character, an individual HMM is built. Additionally, we model frequent punctuation marks, such as full stop, colon and space, and map infrequent ones, such as semi-colon and question mark, to a special garbage model. Each HMM consists of 14 states that are connected in a linear topology. The character models are concatenated to word models, which share the individual character models. Furthermore, to model a complete line of text the word models are concatenated. Because of the continuous nature of the extracted features, the output probability distribution is continuous using a Gaussian mixture with four components. More details can be found in [9].

The system has been implemented using the HTK toolkit [18], originally developed for speech recognition. The toolkit employs the Baum-Welch algorithm for training and the Viterbi algorithm for recognition [15]. The output of a HMM classifier is a sequence of words together with the log-likelihood score of each word. The score of a text line is the sum of the log-likelihood of all words.

4. Writer Identification System Based on Text Line Recognition

For each writer, a text line recognizer as described in the previous section is built and trained with text lines of this writer only. In this way we obtain, for each writer, one recognizer that is specially adapted to the individual handwriting style of that writer.

A text line to be classified is presented to the HMM of each writer. Each HMM outputs a transcription of the input text line together with its log-likelihood score. We sort all log-likelihoods in a descending order and assign the input text line to the writer whose score is on the first rank.

In applications where a wrong identification implies a high cost, it may be advisable to implement a rejection mechanism. The rejection mechanism used in this paper is similar to the one described in [11]. We calculate the difference between the log-likelihood of the best and the second best ranked writer and normalize it by the length of the text line. This length is equal to the width of the text line in pixels. This results in the following confidence measure for a given text line:

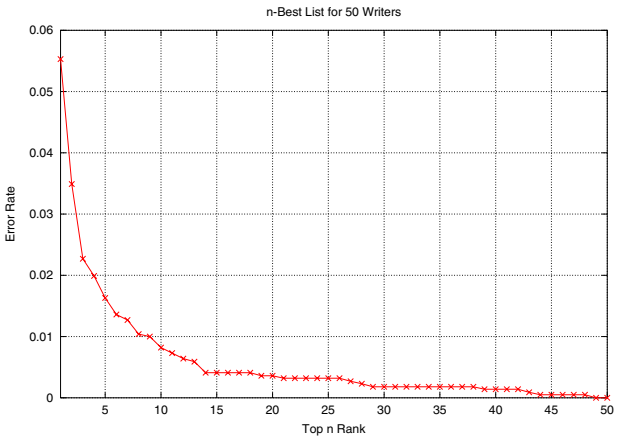

Figure 2. n-Best List

$$cm_{text\ line} = \frac{l_1 - l_2}{text\ line\ length} \quad (1)$$

where l_1 is the log-likelihood score of the highest ranked writer and l_2 is the log-likelihood score of the second highest ranked writer.

Once this confidence value has been computed, we reject an input text line if the value of cm is smaller than a given threshold.

5. Experiments

5.1. Database

Our experiments are based on the IAM database [10][1] which consists of handwritten pages containing between five to ten text lines per page. We use a subset of the database that is identical to the data set used in [6]. A total of 50 writers have contributed to this data set. For each writer we use five pages and extract between 26 and 63 text lines per writer. In total, we have 2,207 text lines from 50 different writers including 4,210 different words. For each writer, the set of available text lines is split into four disjoint subsets to perform full four-fold cross validation experiments. In each of the four runs, we use three subsets to train each of the handwriting recognition systems. The remaining sets of text lines form the test set. This procedure guarantees that text lines in the training sets do not appear in the test set. Hence our experiments are text independent.

5.2. Results

The correct writer identification rates obtained in the four-fold cross-validation experiment are shown in Table 1.

1 Publicly available at: www.iam.unibe.ch/~fki/iamDB

test set	1	2	3	4	avg.
writer identification rate	94.23	94.77	95.41	93.46	94.47

Table 1. Writer Identification Rates.

test set	1	2	3	4	avg.
word recognition rate *same*	67.87	67.01	68.35	65.92	67.30
word recognition rate *other*	15.53	15.23	15.93	15.28	15.49

Table 2. Word Recognition Rates.

The results are significantly better (at the statistical significance level of 99%) than the 90.7% reported in [4] on the same dataset. As our approach uses a handwriting recognition system and is based on the hypothesis that the word recognition rate is higher when the input comes from the writer whose data were used to train the system, we also report the correct word recognition rate of the handwritten text recognizer in Table 2. Here we distinguish between the case where training and input data come from the same writer (row *same*) and the case where they come from different writers (row *other*). Obviously, there is a huge difference in recognition performance between the two cases, which confirms our hypothesis.

Next, we measure the recognition rate not only based on the first rank, but based on the first n ranks. The corresponding n best list is shown in Fig. 2. We observe a drop of the error rate from 5.53% to 1.99% if the first four ranks are taken into account.

By rejecting text lines with a low confidence measure (see Eq. 1), we can reduce the number of misclassifications on the first rank. The resulting error-rejection curve is given in Fig. 3. The recognition rate of 94.47% reported in Table 1 corresponds to the case where no input text line is rejected. This is equivalent to an error rate of 5.53% shown in Fig. 3 at zero rejection rate. Increasing the confidence threshold as specified in Eq. (1) leads to an increasing number of rejections and a decrease of the error rate. At a rejection level of 15% the error rate drops to zero.

6. Conclusion

In this paper, we have presented a writer identification system that utilizes HMM based text line recognizers in the off-line mode. Our approach is text independent and uses text lines as basic entities, from which features are extracted. For each writer, we train a recognizer and present

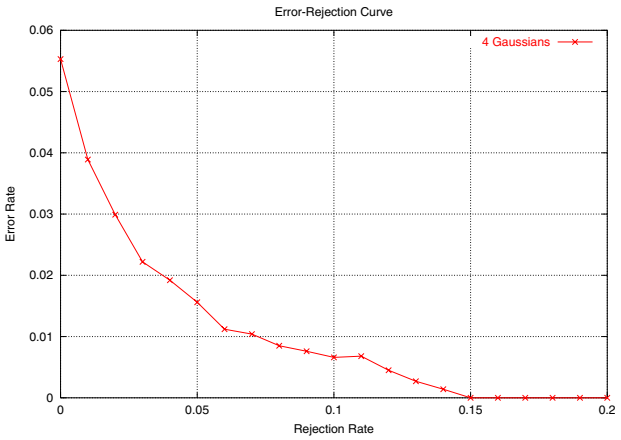

Figure 3. Error-Rejection Curve.

unknown input text lines to each recognition system. As output we get, for each recognizer, the log-likelihood of the input text line. We rank the output of each system and choose the writer whose HMM recognizer is on the first rank.

Using handwriting text line recognizers in this novel way, we achieve a writer recognition rate of 94.47%. A drop of the error rate to 1.99% is obtained if the first 4 ranks are considered. Using a rejection mechanism based on a simple confidence measure, we can reduce the error rate on the first rank to zero by rejecting 15% of the input data.

In our future work we plan to test the system on a larger database including a larger number of writers. Additionally, we want to apply the system to the task of writer verification.

Acknowledgments

This research is supported by the Swiss National Science Foundation NCCR program "Interactive Multimodal Information Management (IM2)" in the Individual Project "Access and Content Protection (ACP)".

References

[1] A. Bensefia, A. Nosary, T. Paquet, and L. Heutte. Writer identification by writer's invariants. In *International Workshop on Frontiers in Handwriting Recognition*, pages 274–279, 2002.

[2] A. Bensefia, T. Paquet, and L. Heutte. Information retrieval based writer identification. In *Seventh International Conference on Document Analysis and Recognition*, pages 946–950, 2003.

[3] M. Bulacu, L. Schomaker, and L. Vuurpijl. Writer identification using edge-based directional features. In *Seventh International Conference on Document Analysis and Recognition*, pages 937–941, 2003.

[4] H. Bunke. Recognition of cursive roman handwriting – past, present and future. In *Seventh International Conference on Document Analysis and Recognition*, pages 448–461, 2003.

[5] S.-H. Cha and S. Srihari. Multiple feature integration for writer verification. In L. Schomaker and L. Vuurpijl, editors, *Proc. Seventh International Workshop on Frontiers in Handwriting Recognition*, pages 333–342, 2000.

[6] 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.

[7] 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.

[8] G. Leedham and S. Chachra. Writer identification using innovative binarised features of handwritten numerals. In *Seventh International Conference on Document Analysis and Recognition*, pages 413–417, 2003.

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

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

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

[12] A. Nosary, L. Heutte, T. Paquet, and Y. Lecourtier. Defining writer's invariants to adapt the recognition task. In *Fifth International Conference on Document Analysis and Recognition*, pages 765–768, 1999.

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

[14] R. Plamondon and S. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22:63–84, 2000.

[15] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–285, 1989.

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

[17] T. Steinherz, E. Rivlin, and N. Intrator. Off-line cursive script word recognition – a survey. *International Journal on Document Analysis and Recognition*, 2:90–110, 1999.

[18] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK book*. 2002.

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