Improving Writer Identification by Means of Feature Selection and Extraction

Andreas Schlapbach and Vivian Kilchherr and Horst Bunke
Institute of Computer Science and Applied Mathematics
University of Bern, Neubrückstrasse 10, CH-3012 Bern, Switzerland
{schlpbch, vkilch, bunke}@iam.unibe.ch

Abstract

To identify the author of a sample handwriting from a set of writers, 100 features are extracted from the handwriting sample. By applying feature selection and extraction methods on this set of features, subsets of lower dimensionality are obtained. We show that we can achieve significantly better writer identification rates if we use smaller feature subsets returned by different feature extraction and selection methods. The methods considered in this paper are feature set search algorithms, genetic algorithms, principal component analysis, and multiple discriminant analysis.

Keywords: writer identification, feature selection, feature extraction.

1. Introduction

Writer identification is the task of determining the author of a sample handwriting from a set of writers [18]. Surveys covering work in automatic writer identification and signature verification until 1993 are given in [12, 18]. Recently, a number of new approaches to writer identification have been proposed. Said et al. [22] 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. [3, 29] address the problem of writer verification by casting it as a classification problem with two classes, *authorship* and *non-authorship*. Zois et al. [31] base their approach on single words by morphologically processing horizontal projection profiles. Edge-based directional probability distributions and connected-component contours as features for the writer identification task are proposed in [2, 24]. Bensefia et al. introduce graphemes as features for describing the individual properties of handwriting [1]. Leedham et al. [13] present a set of eleven features which can be extracted easily and used for the identification and verification of documents containing handwritten digits. Hidden Markov Model (HMM) based recognizers are used for the identification and verification of persons based on their handwriting in [23].

In previous work, we have cast the writer identification problem as a classification problem. In Hertel et al. [7] a system for writer identification is presented that extracts 100 features from a handwritten text line. The features extracted include basic features such as slant and skew angle, features computed from the connected components or the enclosed regions of a handwriting, features extracted from the lower and upper contour of a text line, and fractal features. All the extracted features are then used by a k-nearest-neighbor classifier that compares the extracted feature vector to a number of prototype vectors coming from writers with known identity.

Good results were reported when using all 100 features to determine the identity of a handwritten text line [7]. However, it is an open question whether the extracted features are optimal or near-optimal. Actually, the features may not be independent of each other or even be redundant. Moreover, there may be features that do not provide any useful information for the task of writer identification. We conclude that there may exist subsets of features that perform as well as, or even better than, the original set of features. Furthermore, using a smaller set of features results in a more efficient classifier with respect to both computation time and memory requirements.

Feature extraction is the process of deriving a subset of the original set of features in order to increase classifier efficiency and/or allow higher classification accuracy [11]. There are two different approaches to obtaining a subset of features: *feature selection* and *feature extraction*. In feature selection, a subset of the original set of features is selected while in feature extraction, the features of the original feature set are first combined and then projected onto a space of lower dimensionality.

In this paper, we apply common feature selection and extraction methods to obtain a subset of features from the original set of 100 features described in [7] to identify the author of a handwriting. The methods considered are feature set search algorithms, genetic algorithms, principal compo-

nent analysis and multiple discriminant analysis. We compare the writer identification rates and the dimensionality of the feature subsets returned by the various methods.

The remainder of the paper is structured as follows. In the next section, we give an overview of the feature selection and extraction methods we evaluated. The experimental setup is described in Section 3. In Section 4 our results are presented and discussed. Section 5 concludes the paper.

2. Feature Selection and Extraction

We apply two groups of methods to our writer identification problem. The first group consists of feature set search and genetic algorithms, while the second group includes methods which linearly combine the given features and then project them onto a space of lower dimensionality. In this section, we give a short overview of the three groups of methods and present related work.

Feature set search algorithms address the problem of finding a subset of d features of a given set of D features. Since an exhaustive search is not possible if many features are involved, a number of feature set search algorithms have been proposed. An overview of early work on feature set search algorithms is given in [10]. In [19] floating search methods are introduced, which dynamically change the number of features included or excluded in the feature set. Extensions to these algorithms, such as adaptive floating search or oscillating search algorithms for feature selection, are presented in [27] and in [26], respectively.

A Genetic Algorithm (GA) is a stochastic algorithm inspired by natural evolution. GAs maintain a set of solutions (called chromosomes) in a population. The chromosomes evolve through successive iterations (called generations). In each generation, the chromosomes are evaluated using a fitness function and the fitter a chromosome is the higher is its chance to be selected for inclusion in the next generation. To simulate the process of evolution, genetic operations such as crossover and mutation are applied to the chromosomes. A detailed introduction to genetic algorithms is given in [15]. GAs have been first applied to feature selection problems in [25] and [5]. In [20] work that combines feature selection and data classification using genetic algorithms is summarized.

Another approach to reducing the dimensionality of feature sets consists in linearly combining the original features and then projecting the high-dimensional data onto a space of lower dimensionality. There are two classical methods to find an effective linear transformation. The first method is Principal Component Analysis (PCA) which seeks a projection that best *represents* the data. The second method is Multiple Discriminant Analysis (MDA) which seeks a projection that best *separates* the data. MDA is an extension of Fisher's linear discriminant analysis from a two-class to a c-class classification problem. PCA and MDA are discussed in detail in [4].

To evaluate the merits of various feature selection and extraction methods, a number of comparative studies have been conducted. Ferri et al. [6] compare methods for large-scale feature selection. In [11], algorithms that select features for pattern classifiers are studied on small (0–9), medium (20–49) and large scale (50–∞) feature sets. A recent comparative study of different feature selection algorithms is given in [16].

Related to writer identification, genetic algorithms have been used to select feature subsets for handwritten character [9] and handwritten digit recognition [17]. An overview of recent work that uses genetic algorithms in character recognition system is given in [8]. Zhang et al. present a novel method for feature dimensionality reduction for the recognition of handwritten numerals [30].

3. Experiments

3.1. Methodology

For all experiments, we use an Euclidean-distance based 5-Nearest-Neighbor (5-NN) classifier. This classifier determines the five nearest neighbors to each input feature vector and opts for the class that is most often represented. In case of a tie, the class with the smallest sum of distances is chosen. The number of nearest neighbors has been empirically determined. The advantage of this classifier is its conceptual simplicity and the fact that no classifier training is needed. In the baseline experiment, all 100 features are used as prototypes for the 5-NN classifier to determine the writer identification rate.

We evaluated the following sequential search methods: Sequential Forward Search (SFS), Sequential Backward Search (SBS), Sequential Floating Forward Search (SFFS), and Sequential Floating Backward Search (SFBS) [10, 19]. SFS starts with the empty set of features. Then, at each step one single feature to be added to the feature set is selected from the remaining features so that the new, enlarged set of features yields the highest writer identification rate determined by the 5-NN classifier on a validation set. The counterpart of SFS is SBS, which starts with the full set of features and iteratively removes one feature so that the new reduced set of features yields the highest writer identification rate [10]. Both algorithms add (in the case of SFS) or remove (in the case of SBS) one single feature at a time and no backtracking is possible. In contrast, in SFBS and SFFS the number of features to be included or removed at each step dynamically changes. Thus, the resulting dimensionality in respective stages of the algorithm is not changing monotonously but is actually "floating" up and down [19]. The search in the forward direction is referred to as SFFS

and starts with the empty set of features. The search in the backward direction is called SFBS and starts with the complete set of features.

For the GA experiments, each of the 100 features extracted from a handwritten text line is represented by one bit in a chromosome. A chromosome thus consists of a binary vector of dimensionality 100. If a bit is set to one, the corresponding feature is selected, otherwise the feature is not selected. The operators we use to generate the populations of chromosomes are mutation and crossover. The mutation operator flips one bit at a random position of a chromosome to produce a new chromosome. The crossover operator splits two chromosomes at a random position and combines them to produce two new chromosomes. We use the roulette wheel selection to select the two chromosomes to perform the crossover operation, i.e., each chromosome in the population is assigned a sector on a virtual wheel whose size is proportional to its fitness value [15]. The fitness of a chromosome is the writer identification rate achieved on a validation set using the selected set of features in the 5-NN classifier.

The population consists of 50 chromosomes in the GA experiment. A run of the GA is terminated when the writer identification rate does not improve within 50 generations. The following, commonly used parameters are chosen. In the initial population, 50 out of the 100 features are randomly selected on each of the chromosomes. The mutation rate is set to 0.02 and the crossover rate is set to 0.6. The GAs are implemented using the GAlib package [28].

The Principal Component Analysis (PCA) algorithm first computes the mean and variance of all feature vectors and then normalizes mean and variance. Next the covariance matrix and its eigenvectors and eigenvalues are calculated. The eigenvectors corresponding to the M largest eigenvalues are retained and the input vectors are projected onto the subspace defined by these eigenvectors. The vectors of this lower dimensional space are then used in the 5-NN classifier.

Multiple Discriminant Analysis (MDA) is an extension of Fisher's linear discriminant analysis from a two class to a c-class classification problem. Fisher's linear discriminant projects high-dimensional data onto a line and performs classification in this one-dimensional space. The projection maximizes the distance between the means of the two classes and simultaneously minimizes the variance within each class. Again, the resulting vectors are used in the 5-NN classifier.

3.2. Database and Training

In this paper, we use five pages of handwritten text from each of 50 different writers. Our experiments are based on pages of handwritten text from the IAM database [14][1]. For each writer between 27 and 54 text lines are available. Overall 1,830 text lines are used. For each writer, the set of available text lines is randomly split into five disjoint subsets of approximately equal size.

In the baseline experiment, we use four of the five subsets as prototypes for the 5-NN classifier (training set) and the fifth set to determine the writer identification rate (test set).

For the feature set search algorithms and the GA experiments, one out of the five subsets is used as the test set. The other four subsets are used, first, to evaluate the fitness of the selected feature subsets and, subsequently, as prototypes in the 5-NN classifier. To evaluate the fitness of a selected feature subset, iteratively three out of the four subsets form the prototypes of the 5-NN classifier and the remaining set is used to measure the fitness of the feature subset under consideration. Once the algorithm terminates with the best feature subset, all of the four subsets are used as prototypes for the 5-NN classifier and the final writer identification rate is calculated on the test set.

In the PCA and the MDA experiments, we determine the optimal dimension of the transformed feature subspace as follows. Iteratively, we use three of the four subsets of the training set as prototypes in the 5-NN classifier and calculate the writer identification rate for a given dimension on the fourth set. The average of the four rates thus obtained is the writer identification rate for the dimension under consideration. We select the dimension which produces the highest writer identification rate. Using this dimension, we use all four subsets of the training sets in the 5-NN classifier and calculate the final writer identification rate on the test set.

4. Results and Discussion

In Table 1 the results of the different feature selection methods are shown (bold face indicates statistically significant improvements over the baseline experiment at the statistical significance level of 95%). Using all 100 features extracted from a text line in the 5-NN classifier, we obtain a writer identification rate of 92.08% in the baseline experiment. Among the four feature set search algorithms, SBS achieves the best writer identification rate of 94.26%. The two floating search algorithms, SFBS and SFFS, produce writer identification rates of 93.17% and 93.44% respectively. The SFS algorithm yields an identification rate of only 92.35%.

All of the four feature set search algorithms achieve better writer identification rates than those obtained in the baseline experiment. However, only the result obtained by the

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

Experiment	N. of Features	Writer Id. Rate
Baseline	100	92.08%
SBS	42	**94.26%**
SFS	51	92.35%
SFBS	42	93.17%
SFFS	55	93.44%
GA	50	**95.08%**
PCA	59	92.35%
MDA	38	**98.36%**

Table 1. Writer identification rates achieved using different feature selection methods.

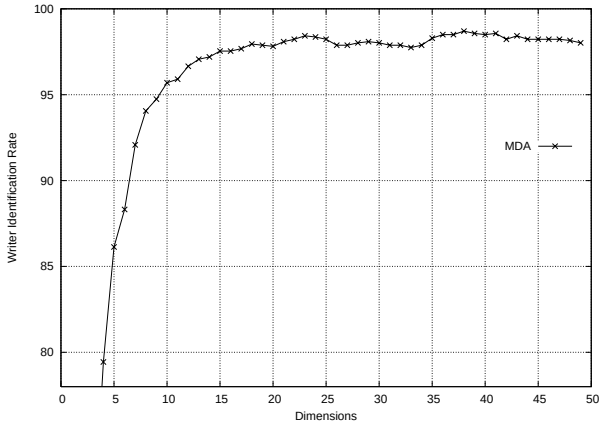

Figure 1. Writer identification rate as a function of the dimensionality using MDA.

SBS algorithm is statistically significantly better. In all four cases, approximately half of the original features are selected. This fact clearly shows that there are many dependent or irrelevant features in the original feature set, and by selecting a subset of these features we can improve the writer identification rate.

The GA using the parameters described in Section 3 yields a writer identification rate of 95.08%. The subset selected consists of 50 features. Note that the GA outperforms all feature set search algorithms. This result is consistent with the claim that GAs are suitable for large-scale problems and have high potential to find better solutions that cannot be found by search algorithms [11].

In Table 1 also the writer identification rates for the two feature selection methods, PCA and MDA, are shown. The PCA method achieves a writer identification rate of 92.35% with a subset that has 59 dimensions. However, the increase of the writer identification rate over the baseline experiment is statistically not significant. Using MDA, a writer identification rate of 98.36% using 38 dimensions is achieved. This improvement is statistically highly significant. It represents the highest writer identification rate of all the feature selection and extraction methods evaluated in this paper. Moreover, MDA yields the smallest set of features.

To illustrate the behavior of MDA, in Figure 1 the writer identification rate on the validation set is plotted as a function of the dimensionality of the transformed feature subspace. As can be seen from the plot, in this validation experiment, with seven features only a higher writer identification rate is achieved than with the 100 features in the baseline experiment on the test set, and an identification rate higher than 95% is achieved using ten dimensions only.

All methods evaluated in these paper produce substantially smaller feature sets. The feature set search and the GA methods reduce the feature set size by approximately 50%. PCA and MDA linearly combine all of the original features to obtain new feature sets. From these new feature sets, the best results are obtained using roughly one half (in the case of PCA) or one third (in the case of MDA) of the given 100 features.

5. Conclusion

In this paper we have experimentally evaluated various feature selection and extraction methods on the problem of writer identification. The task is to identify the author of a text line written by one of 50 writers. The original set of features extracted from a text line is of size 100. We show that feature selection and extraction methods can significantly improve the writer identification rate using a substantial smaller set of features.

Of the feature selection algorithms considered, GA outperforms all four feature selection algorithms and yields a significantly better result compared to the baseline experiment using approximately half of the original 100 features. Feature extraction using MDA achieves the best writer identification rate and a dimensionality reduction by two third of the original features.

In the present work, we have used a simple 5-NN classifier. In future work, we plan to use more complex classifiers such as Neural Network, Support Vector Machine or Bayes Classifier. Furthermore, we consider to implement other algorithms such as oscillating search and non parametric discriminant analysis (NDA), and compare them with the other algorithms. Finally, we can combine the results obtained by the different algorithms using a Multiple Classifier Systems (MCS) [21] to further improve the writer identification rate.

Acknowledgments

This research is supported by the Swiss National Science Foundation NCCR program “Interactive Multimodal Infor-

mation Management (IM2)" in the Individual Project "Access and Content Protection (ACP)". The authors would like to thank Ben Zahler for his support in the feature set search experiments.

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] M. Bulacu, L. Schomaker, and L. Vuurpijl. Writer identification using edge-based directional features. In *Proc. 7th Int. Conf. on Document Analysis and Recognition*, pages 937–941, 2003.

[3] 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.

[4] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2001.

[5] F. Ferri, V. Kadirkamanathan, and J. Kittler. Feature subset search using genetic algorithms. In *Proc. of the IEE/IEEE Workshop on Natural Algorithms in Signal Processing*, volume 740, 1993.

[6] F. Ferri, P. Pudil, M. Hatef, and J. Kittler. Comparative study of techniques for large-scale feature selection. In E. S. Gelsema and L. S. Kanal, editors, *Pattern Recognition in Practice IV, Multiple Paradigms, Comparative Studies and Hybrid Systems*, pages 403–413. Elsevier, 1994.

[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] F. Hussein, N. N. Kharma, and R. K. Ward. Genetic algorithms for feature selection and weighting, a review and study. In *Proc. 6th Int. Conf. on Document Analysis and Recognition*, pages 1240–1244, 2001.

[9] G. Kim and S. Kim. Feature selection using genetic algorithms for handwritten character recognition. In *Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition*, pages 103–112, 2000.

[10] J. Kittler. Feature set search algorithms. In C. H. Chen, editor, *Pattern Recognition and Signal Processing*, 1978.

[11] M. Kudo and J. Sklansky. Comparison of algorithms that select features for pattern classifiers. In *Pattern Recognition*, volume 33, pages 25–41, 2000.

[12] 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.

[13] 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.

[14] 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.

[15] M. Mitchell. *An Introduction to Genetic Algorithms*. The MIT Press, 1996.

[16] I.-S. Oh, J.-S. Lee, and B.-R. Moon. Hybrid genetic algorithms for feature selection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1424–1437, 2004.

[17] L. S. Oliveira, N. Benahmed, R. Sabourin, F. Bortolozzi, and C. Y. Suen. Feature subset selection using genetic algorithms for handwritten digit recognition. In *Proc. 14th Brazilian Symposium on Computer Graphics and Image Processing*, pages 362–369. IEEE Computer Society, 2001.

[18] 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.

[19] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, 1994.

[20] W. Punch, E. Goodman, M. Pei, L. Chia-Shun, P. Hovland, and R. Enbody. Further reserch on feature selection and classification using genetic algorithms. In *Proc. of the Int. Conf. on Genetic Algorithms*, pages 557–564, 1993.

[21] F. Roli, J. Kittler, and T. Windeatt, editors. *Multiple Classifier Systems*, volume 3077 of *Lecture Notes in Computer Science*. Springer, 2004.

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

[23] 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.

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

[25] W. Siedlecki and J. Sklansky. A note on genetic algorithms for large-scale feature selection. *Pattern Recognition Letters*, 10(5):335–347, 1989.

[26] P. Somol and P. Pudil. Oscillating search algorithmus for feature selection. In *Proc. Fifteenths Int. Conf. on Pattern Recognition*, volume 2, pages 2406–2409, 2000.

[27] P. Somol, P. Pudil, J. Novovičová, and P. Paclík. Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, 20:1157–1163, 1999.

[28] M. Wall. *GAlib: A C++ Library of Genetic Algorithm Components*. Massachusetts Institute of Technology, 1996.

[29] 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.

[30] P. Zhang, T. D. Bui, and C. Y. Suen. Feature dimensionality reduction for the verification of handwritten numerals. In *Pattern Analysis Applications*, volume 7, pages 296–307, 2004.

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