Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 1004–1013
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Extraction and analysis of forensic document examiner features used for writer identification

Vladimir Pervouchine[a,*], Graham Leedham[b]

[a]*Forensics and Security Lab, School of Computer Engineering, Nanyang Technological University, Block N4, Nanyang Avenue, Singapore 639798, Singapore*
[b]*University of New South Wales Asia, 1 kay Siang Road, Singapore 248922, Singapore.*

Received 21 December 2005; received in revised form 13 July 2006; accepted 8 August 2006

Abstract

In this paper we present a study of structural features of handwriting extracted from three characters "d", "y", and "f" and grapheme "th". The features used are based on the standard features used by forensic document examiners. The process of feature extraction is presented along with the results. Analysis of the usefulness of features was conducted via searching the optimal feature sets using the wrapper method. A neural network was used as a classifier and a genetic algorithm was used to search for optimal feature sets. It is shown that most of the structural micro features studied, do possess discriminative power, which justifies their use in forensic analysis of handwriting. The results also show that the grapheme possessed significantly higher discriminating power than any of the three single characters studied, which supports the opinion that a character form is affected by its adjacent characters.

Keywords: Handwriting analysis; Writer identification; Feature extraction; Feature selection

1. Introduction

Handwriting is a personal biometric that has long been considered to be unique to a person. For many centuries signature verification has been used for authentication purposes. Experts in forensic document analysis all around the world daily perform examination of handwritten documents to determine the authorship of a questioned document or detect evidence of forgery or disguise.

The methods used by forensic document examiners are based on a set of established and well documented techniques [1–3]. The examiners look at features of handwriting, which characterize shapes of letters, lines, and document as a whole. The techniques have been derived from experience and are generally accepted by the various forensic laboratories. However, whilst they are intuitively reasonable, basis. the methods of document analysis lack a scientific Because of this, a fundamental question has arisen in several recent court cases querying whether the results of forensic document analysis are scientific and acceptable as evidence [4,5].

1.1. Individuality of handwriting

The computer scientists approach to the problem is to apply various computer vision and pattern recognition techniques [6]. To represent a handwritten document, a set of features, extracted from the image of the document, is used. The features of handwriting are divided into two classes [7]. Those that are used by forensic experts are called document examiner features, whilst those that are measured by computer algorithms are called computational features. Not every document examiner feature can easily be represented as a computational feature and vice versa. Embellishments of handwriting [8] is an example of a document examiner feature which is hard to represent in numbers. On the other hand, purely computational features like gradient features hardly correspond to any document examiner features.

* Corresponding author. Tel.: +65 67904618; fax: +65 67926559.
E-mail addresses: vpervouchine@gmail.com (V. Pervouchine), G.Leedham@unswasia.edu.sg (G. Leedham).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.08.008

Recent studies of the problem of discrimination of writers by their handwriting have been carried out [9–12]. These studies used computational features of handwriting [13,14]. Such an approach is suitable to determine whether handwriting is indeed unique to a person, and from the studies the answer to this questions is affirmative. The approach is also promising for automatic writer identification systems.

1.2. Forensic experts vs. lay people

Recent investigations by Kam et al. [15–17] as well as by [18] has demonstrated that expert document examiners perform significantly better than lay people in classification of handwritten samples according to their authorship. Hence, it may indeed be the techniques of forensic document examination that explain the difference in accuracy between experts and lay people. Since forensic document examiners use certain features of handwriting to distinguish writers, it is necessary to study whether the features allow discrimination of writers or not as a step towards establishing scientific foundation of forensic document analysis.

2. Our approach

In our work we deal only with document examiner features expressed in a numerical manner and measured by computer algorithms. The purpose of the study is to determine whether it is possible to distinguish people by their handwriting using a subset of document examiner features. We assume that an experienced person, that is, an expert in forensic document analysis: (i) may effectively utilise more features than our system does, for it is very hard to express in strict mathematical terms many of the document examiner features, and (ii) the person may be able to determine which features should, and which should not, be used in a particular case (that is, having looked at handwritten samples, an expert is able to select only the important features for handwriting comparison). As a consequence, an expert may distinguish writers or establish authorship of questioned documents better on average, than our system.

Our research is thus aimed at determining whether some document examiner features are useful for writer discrimination and also determining a lower bound on the accuracy of writer discrimination when only some document examiner features are used (it is not a lower bound in strict terms, because the two assumptions above, although seemingly reasonable, are not proven). We consider only unconstrained genuine handwriting in our study. The problem of authorship identification in court cases usually involves forged and disguised handwriting, however, we have not considered such documents for two reasons. The main reason is that before studying complicated cases of deliberately changed handwriting it is important to study the general case of people's normal handwriting. If the results achieved from the study are negative, that is, the considered features are useless for writer discrimination, there is no point in applying the same approach to the more complicated cases. The other reason is a lack of data on forged and disguised handwriting. Initial studies of forgery detection in handwritten documents has recently been reported [19,20].

The rest of the paper is organised as follows. Section 3 describes the feature extraction: the choice of letters and graphemes to study, the choice of features to extract, the representation of the features in a numerical manner, and the methods of measuring the feature values. Section 4 provides information about the estimation of usefulness of the features: how we define usefulness, subsets of features of different usefulness, and a search for optimal feature subsets. Section 5 presents and discusses the results of our experiments. Section 6 presents the conclusions of this study.

3. Extraction of features

Features of handwriting can be divided into micro and macro feature classes with respect to the scale at which they are extracted. Micro features are those extracted from strokes, character elements, characters, and short character combinations (graphemes). Features that are extracted from words, lines, and bigger aggregates are macro features [7]. In this study we concentrate on structural micro features extracted from characters and graphemes. These represent a subset of the features used by forensic document examiners.

3.1. Choice of features

In our study different features are extracted from different characters and graphemes. As it is impossible to consider all characters and graphemes that occur in handwriting, only three characters and one grapheme were chosen for the study. The choice was based on several considerations. A character (grapheme) must occur frequently in handwriting samples to obtain reliable feature values for it. Frequencies of occurrence of some characters and graphemes are shown in Fig. 2. The frequencies were measured by analysis of the content of several novels, available in free online libraries.

Several studies have shown that characters are not equal in their discriminating power [21,22,11]. Capital letters as well as letters that consist of several strokes, like those with ascenders or descenders, bear more individual information than simple characters like "i" or "c". In addition we suggested that frequent graphemes like "th" can be very useful for writer identification purposes because they reflect spacing between characters, relative sizes of characters etc. [1, Chapter 8].

Based on these considerations characters "d", "y", "f", and grapheme "th" were chosen for study. The character images were extracted manually from 600 samples of the CEDAR letter [23] representing 200 writers. To decrease variation of a character form caused by the preceding and the following characters, samples of characters "d" and "y" were

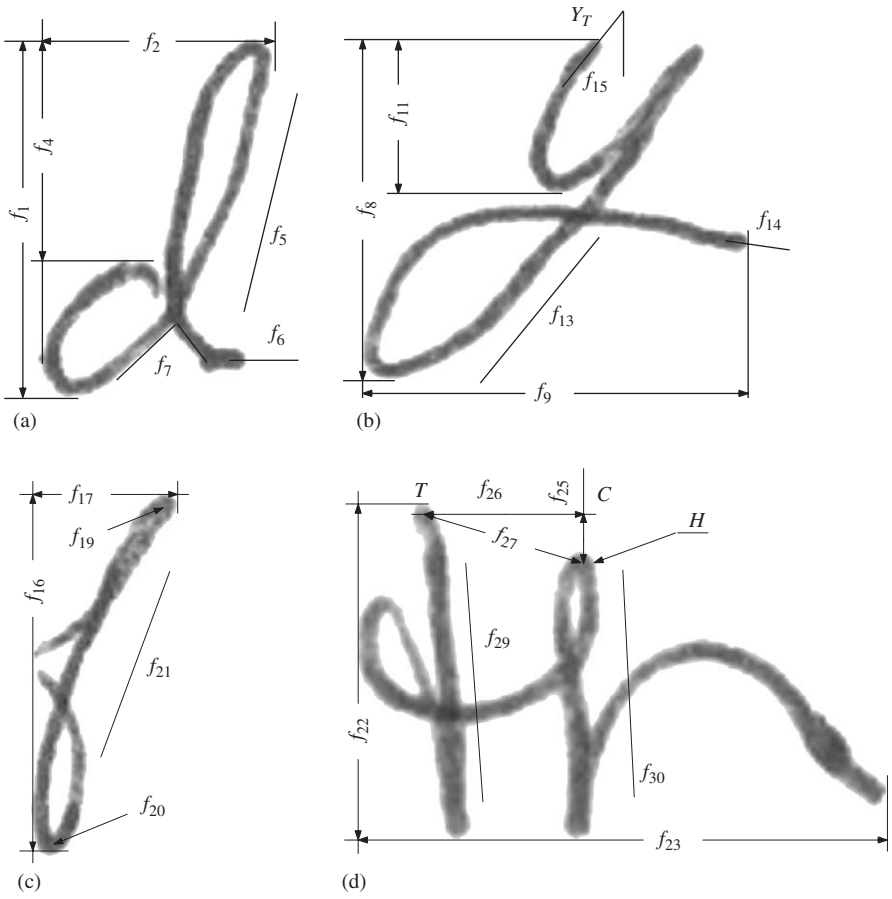

Fig. 1. Some of the features extracted from the four characters.

extracted only from the end of words. These two characters are the third and sixth most frequent characters at the end of words correspondingly [24]. Samples of grapheme "th" were extracted only from the beginning of words. All samples of character "f" were extracted because there were only eight occurrences of this character in the letter. There were at most 10 samples of character "d", eight samples of "y", eight samples of "f", and nine samples of "th" extracted from each of the 600 documents making total 30, 24, 24, and 27 samples of the corresponding character per writer. For some writers less samples were obtained because of missing words or sentences in some of their documents.

3.2. Feature extraction

Initial selection of the structural features to extract was motivated by studying the types of features described in books on forensic document examination [8,3,1]. Huber and Headrick compiled commonly used features into a list of 21 discriminating elements of handwriting [8]. Some of these elements, like height and width of a character, were strictly defined and thus were easy to extract from a character image. Other features, like initial and final strokes, were defined quite vaguely. We formalised part of them by introducing additional strictly defined features which corresponded to those from the list. Further experiments and analysis of extraction accuracy as well as feature usefulness via a filter approach [25,26] allowed us to reconsider the feature set and eliminate some of the features and add others. The final set comprised of 31 features ($f_1 \ldots f_{31}$). Most of the features are schematically represented in Figs. 1(a)–1(d). The features extracted from each character are listed in Table 1 (Fig. 2).

Some comments need to be made here. Besides height (f_1, f_8, f_{16}, f_{22}) and width (f_2, f_9, f_{17}, f_{23}) features, height to width ratio was measured for each character (f_3, f_{10}, f_{18}, f_{24} for "d", "y", "f", and "th", respectively; not shown on Fig. 1). For character "d" feature f_4, the relative height of ascender, was measured as a/f_1, where a was the height of the ascender, and for character "y" feature f_{11}, the relative height of descender, was measured as d/f_8, where d was the height of the descender.

Final strokes f_6, f_{14} as well as slants f_5, f_{13}, f_{15}, f_{21}, f_{29}, f_{30} were measured as angles between a vertical line and the line representing a stroke or slant. A slant was represented by a line fitted with the least squares method into the set of points belonging to an ascender, descender, or stem. A final stroke was represented by a tangent to the curve of the final stroke drawn at its end point.

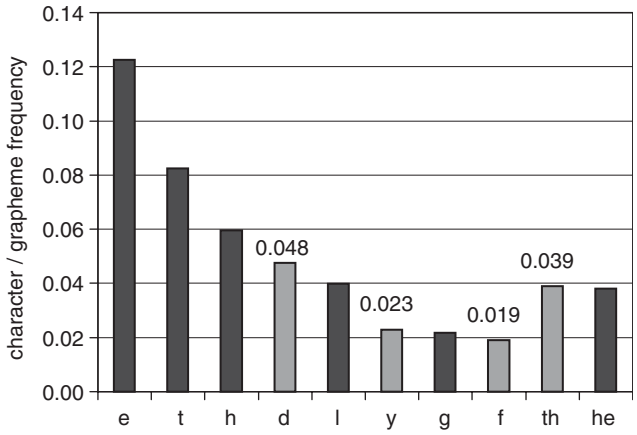

Fig. 2. Frequencies of occurrence of characters and graphemes in texts.

For character "f" presence of loops at the top and bottom points F_T and F_B of the stem was represented by binary features f_{19} and f_{20}. Loop features are claimed to be very useful in forensic analysis of handwriting [8]. However, due to stroke thickness as well as image thresholding and thinning loops are often lost and the features are extracted incorrectly. Completeness of the descender loop of character "y" f_{12} was measured as $1 - x/L$, where x was the distance between loop ends and L was the distance along the loop curve. Along with distances f_{25}, f_{26}, f_{27} for grapheme "th" the angle between line TH and a horizontal line was measured (f_{28}). Relative position of a t-bar on t-stem was represented as f_{31}.

Algorithms for feature extraction consisted of a main program and subroutines for extraction of particular features. The input to the algorithms was a character image, the binarised image, and the skeleton, and the output was the feature vector along with additional information which was later used to verify correctness of the feature values. The algorithms, which were different for each of the three characters and grapheme, analysed the shape of a given character to determine which parts of the image correspond to which parts of the character. The feature extraction was done stage by stage. First, simple features like height and width were measured. Then end points or top/bottom points of skeleton were analysed and the branches were traced. For example, for character "y" the bottommost point of the skeleton was detected and the skeleton was traced to the left and to the right from that point to extract the descender. After that the descender features were measured. Depending on what whether an end point or a junction was encountered at the ends of the descender, the algorithm chose an appropriate way to extract the final stroke and measure its features. After extraction of the descender and final stroke the base part of "y" was analysed in a similar manner (Fig. 3) . Also, for some features, feature extraction was differently for different character shapes: for example, to find a horizontal line that separates ascender of "d" from its base part, one

Table 1
Features extracted from each character

(a)	
f_1	Height
f_2	Width
f_3	Height to width ratio
f_4	Relative height of ascender
f_5	Slant of ascender
f_6	Final stroke angle
f_7	Fissure angle
(b)	
f_8	Height
f_9	Width
f_{10}	Height to width ratio
f_{11}	Relative height of descender
f_{12}	Descender loop completeness
f_{13}	Descender slant
f_{14}	Final stroke angle
f_{15}	Slant at point Y_T
(c)	
f_{16}	Height
f_{17}	Width
f_{18}	Height to width ratio
f_{19}	Presence of loop at F_T
f_{20}	Presence of loop at F_B
f_{21}	Slant
(d)	
f_{22}	Height
f_{23}	Width
f_{24}	Height to width ratio
f_{25}	Distance HC
f_{26}	Distance TC
f_{27}	Distance TH
f_{28}	Angle between TH and TC
f_{29}	Slant of t-stem
f_{30}	Slant of h-stem
f_{31}	Position of t-bar

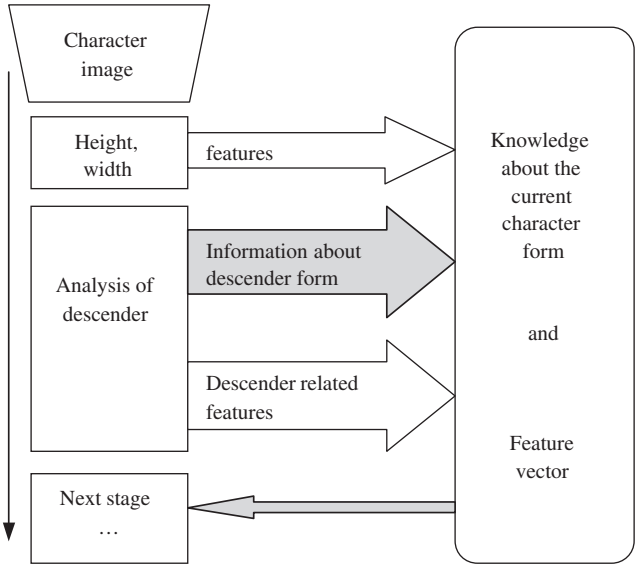

Fig. 3. Extraction of features of character "y".

algorithm was used in the case of thin ascender and the other one in the case of thick ascender. At each extraction stage more knowledge about the shape of the character sample was obtained and more features were extracted, until either all the features were extracted or extraction failed because of unexpected character shape.

Since it was not possible to take all imaginable character shapes into account, the feature extraction failed to extract some or all features in small number of cases. The correctness of feature extraction was checked manually for each sample. Binarised images that were the input to the skeletonisation algorithm were used for feature values check. A sample was marked as incorrectly processed if at least one feature was extracted incorrectly or if extraction failed at some stage. The average extraction accuracy according to this verification scheme was 85% for characters "d" and "y", 87% for grapheme "th" and 92% for character "f". There were two main sources of errors. The first was unusually shaped character samples that were not taken into account in the shape analysis algorithm. The second was the thinning-based skeletonisation which produced erroneous branches, especially for characters with many strokes, making it difficult to perform the shape analysis and find which branches correspond to which character strokes.

Several skeletonisation techniques were tried including those not based on erosion and it was decided to use the Matlab Image Processing Toolbox thinning algorithm [27] (p. 879, bottom of first column through top of second column) with post-processing to remove some artefacts like small disconnected components, short branches, and spurious loops that were introduced by thinning.

Since the purpose of the study was to select useful features rather than implement an automatic extraction of document examiner features, all samples marked as incorrectly processed were excluded from the set of patterns used in the feature usefulness study. Thus, only clean features were used to train and test the classifier.

4. Analysis of feature usefulness

First, all the features that may have discriminating power were extracted. Second, after extraction it was necessary to decide which features have discriminating power and which do not. It was also possible that a document examiner feature appears to be irrelevant for writer classification because the way it was formalised and measured was incorrect or insufficiently detailed.

4.1. Concept of feature relevance

John et al. [28] analysed several definitions of feature relevance which have been presented in the literature and proposed a definition that includes two degrees of relevance: strong and weak relevance. Strong relevance means that a feature cannot be removed from the feature set without loss of classification accuracy. Weak relevance means that a feature can sometimes contribute to classification accuracy. A feature is irrelevant if it is neither strongly nor weakly relevant, and thus can be excluded from the feature set without loss of classification accuracy.

Unfortunately, application of definitions of feature relevance is hard because it requires knowledge of the conditional probability densities $p(\vec{f}|\omega_j)$ for feature values $\vec{f}$ (ω_j is a class label). In our study we divided the features into three classes according to the results of feature subset selection experiments. After obtaining, presumably, all feature subsets which were equally good for writer discrimination, we simply considered how often each feature was selected. To prevent confusion in the terminology, we divided the features into indispensable, partially relevant, and irrelevant features. The last category coincides with the one defined by John [28]. A feature was indispensable if it was selected in each optimal feature set. A feature was irrelevant if it was not selected in any optimal set. If a feature was selected in some of the optimal sets, it was partially relevant. We believe such an approach is more practical although the division of features into the three categories can be classifier-dependent. We favour a three-category relevance division rather than the usual ranking of features because we only need to find which features are useful and which are not. Ranking the features may not provide any additional information since the features are a formalisation of document examiner features and could be ranked differently had they been differently formalised. Thus, ranking would be formalisation-dependent and will not reflect the situation with document examiner features.

4.2. Feature selection problem

In order to divide features into three sets according to their relevance it is necessary to find all possible feature subsets that can be formed from the initial set and which result in the highest classification accuracy. There are two approaches for selection of a subset of features: filter and wrapper [28]. In our study we chose the wrapper approach to find all the best feature subsets. In this approach an induction algorithm is used for evaluation of a feature set. A feature set is assigned a value proportional to its performance. Performance can be a classification accuracy achieved with the feature set. It can also be affected by the number of features, if one is interested in the minimal subset to be chosen.

In our study we used n-fold cross validation to evaluate a feature subset [29]. The training data was divided into n approximately equal partitions and the induction algorithm was then run n times each time leaving one subset for test and using the other $n-1$ parts for training. The classification accuracy obtained from n tests was then averaged and associated with the corresponding feature subset.

4.3. Experimental setup

DistAl, a constructive learning algorithm based on the multi-layer perceptron with spherical threshold units, was chosen as a classification system [30]. There were several reasons for this choice over other possibilities:

- DistAl does not require any a priori assumptions about network topology. Network topology is determined dynamically in the learning process.
- It is fast in learning because it does not use an iterative algorithm to compute perceptron parameters (weights, thresholds). The most time-consuming part is the calculation of inter-pattern distances for each pair of patterns. However, this needs to be performed only once.
- DistAl is based on a distance metric between patterns which means it can easily be adapted to handle patterns with missing feature values.
- Experiments conducted on both artificial and real data [30] demonstrated results of classification comparable to those obtained by other commonly used learning algorithms.

All feature values were treated as real numbers. Having performed several experiments we chose the normalised Manhattan distance as a distance measure for DistAl because this measure was shown to be suitable for the problem at hand

$$d(\vec{F}^1, \vec{F}^2) = \frac{1}{k}\sum_{i=1}^{k}\frac{|F_i^1 - F_i^2|}{\max_i - \min_i},$$

where k is the number of features, $\min_i$ and $\max_i$ are the minimum and maximum values of the ith feature in the data set, respectively.

A genetic algorithm (GA) was used to implement feature subset selection. From the studies of De Jong [31] GAs have been extensively used to solve problems of feature selection in pattern recognition [32–34]. Successful use of a GA together with the DistAl algorithm has also been demonstrated [35]. Use of a GA has several advantages to other commonly used methods:

- GAs have the capability of finding a "good" or even optimal solution for complex problems relatively quickly. They are less likely to get stuck in a local extremum than gradient-based techniques [36].
- GAs use only the fitness function itself and not any additional information such as the derivatives. In the case of the feature selection problem this is exactly what we need.
- Representation of a feature subset as a string in a GA is straightforward. Fixed length binary strings are used; the number of elements is equal to the total number of features; a value of 1 (0) in a string corresponds to the presence (absence) of the feature in the subset associated with the string.

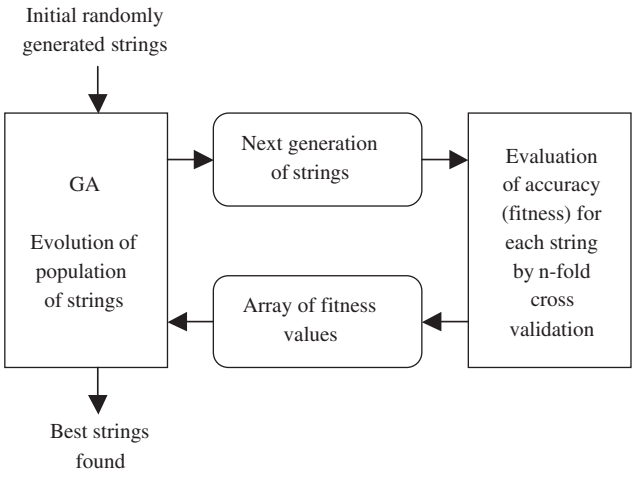

Fig. 4. Search of best feature subsets.

- GAs are not very sensitive to the values of their parameters. Even when the values are far from optimal a good solution can still be achieved although it may require a larger number of generations to reach the solution. This property is very useful because it means that one need not to be over-concerned about adjusting the values of several parameters which cannot be calculated a priori.

Fig. 4 shows the schema of feature selection which was used in the current study.

After trying out different types of crossover, replacement strategies, probabilities of crossover and mutation, the parameters of the GA used for feature selection were chosen as follows:

- population size: 50;
- uniform crossover [37] with probability 0.6;
- probability of mutation: 0.03;
- replacement strategy: best strings from offsprings and parents form the next generation;
- stop condition: maximum number of 100 generation is reached (the convergence was reached after about 50 generations in most cases).

Besides, linear scaling of fitness with factor 2 was used [36] to prevent domination of highly fit strings in early generations and random walk search in mature generations. One run of GA evaluated 5000 out of the possible 2^{31} feature subsets.

5. Results

Experiments were conducted to evaluate writer classification accuracy achieved using the DistAl neural network when only one character and a set of four characters were used. For some writers the amount of patterns obtained for

one of the four characters was too small because of errors in the feature extraction stage. To make the results of experiments comparable for all single characters and the four-character set, patterns from 165 different writers were used (writers who had less than 15 patterns for any of the character due to errors in feature extraction stage were excluded from the study). There were between 15 and 30 patterns per writer. Accuracy of classification was measured by five-fold cross validation on the DistAl classifier.

Feature vectors for four-character combinations were formed by merging feature vectors for separate characters together. Each character gave its feature vector for one and only one combination. Let the feature vectors obtained for a particular writer be $\vec{F}_1^d \dots \vec{F}_{n_d}^d$, $\vec{F}_1^y \dots \vec{F}_{n_y}^y$, $\vec{F}_1^f \dots \vec{F}_{n_f}^f$, $\vec{F}_1^{th} \dots \vec{F}_{n_{th}}^{th}$ extracted from the writer's samples of characters "d", "y", "f", and "th", respectively, n_d, n_y, n_f, and n_{th} being the number of vectors extracted for each character. In order to form the full vectors for the writer, four character-specific vectors were drawn from the set, without replacement, until any of the character-specific feature vector set was empty. The number of feature vectors formed for the four-character combination was thus $\min(n_d, n_y, n_f, n_{th})$. For example, a feature vector might be $\vec{F} = \{\vec{F}_1^d \vec{F}_{10}^y \vec{F}_3^f \vec{F}_{20}^{th}\}$.Several pattern sets differing by permutations of character-specific feature vectors for each writer were formed by this method.

For experiments with one character the feature selection problem was solved using an exhaustive search (the largest amount of 10 features for grapheme "th" resulted in 2^{10} different feature sets). For experiments with all characters GA was used for feature selection. The GA was run several times on different pattern sets and all strings from the last generations of each run were copied into a set of strings—candidates for optimal solutions. For each of the candidates a test was performed to measure the classification accuracy of the associated feature subset on each of the available pattern sets. The classification accuracy was then calculated as the average of the accuracies achieved on each pattern sets. In addition, the standard deviation of the accuracy was calculated. The string with the highest average accuracy was then chosen and placed into a set of optimal solutions. For all other candidates a test on equal means was performed to determine whether their average accuracy values were insignificantly different from the highest one at a 1% significance level. The strings that passed the test were also included into the set of optimal solutions.

The results of the accuracy of classification achieved when a single character features and multi-character features ("all") were used are shown in Table 2. As seen from the chart, grapheme "th" has significantly larger discriminative power than the other characters. Use of all four characters results in a noticeable improvement in classification accuracy in comparison to that achieved when only one character is used. The highest achieved classification accuracy was 58% when the optimal subset of all four character features was used (that is, when a number of patterns belonging to different writers were fed to the classifier, around 58% of the patterns were assigned labels of their authentic writers). Optimal feature sets are presented in Table 3, where a is the average accuracy achieved when the associated feature set was used, and σ_a is the standard deviation of the accuracy value based on 30 experiments. In a string representing a feature set the value of 1(0) mean that the feature is included in (excluded from) the set. The position of a digit in a string in Table 2 corresponds to the index of the feature. Table 1 should be referred to get the name of the feature.

As seen from Table 2, different characters possess different discriminating power. Our experiments reveal that accuracy of writer identification obtained when three separate characters were used was the highest for character "f" and the lowest for character "d". This result is in agreement to that obtained by Zhang and Srihari [9] although they used a completely different feature set. Our experiments also showed that the discriminating power of grapheme "th" is significantly stronger than that of any of the three single characters considered.

As a result of feature selection experiments several optimal feature subsets were obtained (Table 3). Division of the features into three groups was performed as discussed in Section 4.1. The three categories of features are shown in Table 4, where feature indices are used to represent features (see Table 1 for feature names).

As is evident from Table 4, features of grapheme "th" constitute the greater part (8 out of 13) of indispensable features, which is another argument for use of frequent graphemes rather than single characters in writer identification. Four features, namely, final stroke angle (f_6) and fissure angle (f_7) of character "d" and presence of loop at the upper (f_{19}) and lower (f_{20}) points of f-stem proved to be irrelevant for writer discrimination.

It is important to note that the features, for which usefulness was assessed, were not exact document examiner features but rather our representation of them. If a feature was classified as relevant or partially relevant, that means that the document examiner feature has discriminative power. But if a feature appeared to be irrelevant we cannot conclude anything about the discriminating power of the document examiner feature, since it may be that our formalisation of it was not good enough. This might be the case for final stroke angle and fissure angle features.

The purpose of using features that represent the presence of loops at the upper and lower points of a stem of "f" was to distinguish between the hand-printed and cursive forms of character "f". From the results obtained we conclude that these two features do not help to discriminate between the two character forms effectively. This may be a surprising result since loop features are considered to be a good features to distinguish writers, and it is intuitively so. As was already mentioned, it is not correct to conclude that loop features are useless. Loops are often lost due to binarisation because

Table 2
Accuracy of writer identification using different characters

Character	d	y	f	th	all
Accuracy (%)	16	20	26	36	58

Table 3
Optimal feature sets

d-part	y-part	f-part	th-part	a	σ_a
1111100	01111001	111001	1111111111	0.58	0.04
1101100	10111011	111001	1111111111	0.57	0.04
1110100	11011010	111001	1111111111	0.55	0.04
1111100	11111001	111001	1101111111	0.54	0.05
1110100	11111101	101000	1111111111	0.54	0.05
1111000	11111010	111001	1111111111	0.53	0.04
1111100	11111001	111001	1111111111	0.53	0.04
1111100	11110011	111001	1011111111	0.53	0.04

Table 4
Division of features according to their relevance

Indispensable, f_i	Partially relevant, f_i	Irrelevant, f_i
1, 2, 11, 16, 18, 22, 25–31	3, 4, 5, 8, 9, 10, 12, 13, 14, 15, 17, 21, 23, 24	6, 7, 19, 20

of thick strokes. Also quite often loops are hidden: a person can easily tell it is a loop when looking at handwriting even though there is not gap between the strokes. Both hidden and lost loops were classified as non-loops and hence the feature values in such cases did not reflect the real situation. This is the most likely reason why loop features were classified as irrelevant.

6. Conclusions

Several structured features corresponding to those used by forensic document examiners were studied. Although we did not achieve identification accuracy close to 100%, a number of important issues were revealed from the study.

It was shown that the features we studied which correspond to those forensic document examiners use for their analysis indeed possess discriminating power and thus use of these features for the purpose of writer identification is justified.

It was demonstrated that not only characters are different in their discriminating power but also there exist a noticeable difference between characters and graphemes. Use of features of grapheme “th” resulted in significantly more accurate identification of writers than the use of any of the features of the single characters. This supports the suggestion that form of a character can be greatly affected by the adjacent characters—the suggestion which is used in forensic document analysis.

The most important (indispensable) features as well as irrelevant features were identified under the assumption that the data is all genuine unconstrained handwriting. The identification of indispensable features provides the information about which features should be selected when faced with the problem of distinguishing authors of several handwritten samples. Frequency of appearance of partially relevant features in the determined optimal sets may also be used as an indicator of their usefulness, although it is possible that under conditions different from ours the ranking of these features may be different. For example, height to width ratio of a character is thought to be very important when handwriting is constrained as in the case where handwriting is extracted from forms.

Formalisation of more document examiner features extracted from frequent graphemes and short words will likely result in achieving high accuracy of writer identification. Also, analysis of forged and disguised handwriting is necessary in order to determine how stable document examiner features are in such cases.

Acknowledgements

The authors would like to thank Professor Sargur Srihari, Director of the Center of Excellence for Document Analysis and Recognition at The University at Buffalo, New York, for allowing us to access data to carry out this analysis.

References

[1] O. Hilton, Scientific Examination of Questioned Documents, CRC Hall, Florida, USA, 1993.
[2] E.W. Robertson, Fundamentals of Document Examination, Nelson-Hall, IL, USA, 1991.
[3] W.R. Harrison, Suspect Documents, Their Scientific Examinations, Nelson-Hall, IL, USA, 1981.
[4] Daubert, et al., v. Merrell Dow Pharmaceuticals, 509 U.S. 579, 1993.
[5] United States v. Starzecpyzel, 880 F. Supp. 1027, 1046 (S.D.N.Y. 1995), 1995.
[6] S.N. Srihari, C.G. Leedham, A survey of computer methods in forensic document examination, in: H.L.Teulings, A.W.A. Van Gemmert (Eds.), Proceedings of 11th Conference International on Graphonomics Society (IGS2003), Scottsdale, AZ, USA, Avon Books, New York, 2003, pp. 278–281.
[7] S.N. Srihari, S.-H. Cha, S. Lee, Establishing handwriting individuality using pattern recognition techniques, in: Proceedings of Sixth International Conference on Document Analysis and Recognition (ICDAR'2001), Seattle, USA, 2001, pp. 1195–1204.
[8] R.A. Huber, A.M. Headrick, Handwriting Identification: Facts and Fundamentals, CRC Press, LCC, 1999.
[9] S.N. Srihari, S.-H. Cha, H. Arora, S. Lee, Individuality of handwriting, J. Forensic Sci. 47 (4) (2002) 1–17.
[10] S.N. Srihari, C.I. Tomai, B. Zhang, S. Lee, Individuality of numerals, in: Proceedings of Seventh International Conference on Document Analysis and Recognition (ICDAR'2003), Edinburgh, UK, 2003, pp. 1096–1100.
[11] B. Zhang, S.N. Srihari, S. Lee, Individuality of handwritten characters, in: Proceedings of Seventh International Conference on Document Analysis and Recognition (ICDAR'2003), Edinburgh, UK, 2003, pp. 1086–1090.
[12] B. Zhang, S.N. Srihari, Analysis of handwriting individuality using word features, in: Proceedings of Seventh International Conference on Document Analysis and Recognition (ICDAR'2003), Edinburgh, UK, 2003, pp. 1142–1146.
[13] G. Srikantan, S.W. Lam, S.N. Srihari, Gradient-based contour encoding for character recognition, Pattern Recognition 29 (7) (1996) 1147–1160.
[14] S.-H. Cha, S.N. Srihari, Multiple feature integration for writer verification, in: Proceedings of Seventh International Workshop on Frontiers in Handwriting Recognition (IWFHR-7), Amsterdam, The Netherlands, 2000, pp. 333–342.
[15] M. Kam, J. Wetstein, R. Conn, Proficiency of professional document examiners in writer identification, J. Forensic Sci. 39 (1) (1994) 5–14.
[16] M. Kam, G. Fielding, R. Conn, Writer identification by professional document examiners, J. Forensic Sci. 42 (5) (1997) 778–786.
[17] M. Kam, K. Gummadidala, G. Fielding, R. Conn, Signature authentication by forensic document examiners, J. Forensic Sci. 46 (6) (2001) 884–888.
[18] B. Found, J. Sita, D. Rogers, A pilot study for validating forensic handwriting examiners 'expertise': Signature examinations, in: Proceedings of Ninth International Conference on Graphonomics Society (IGS99), Singapore, 1999, pp. 209–212.
[19] S.-H. Cha, C.C. Tappert, Automatic detection of handwriting forgery, in: Proceedings of Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR-8), Ontario, Canada, 2002, pp. 264–267.
[20] H.-C. Chen, S.-H. Cha, Y.-M. Chee, C.C. Tappert, The detection of forged handwriting using a fractal number estimate of wrinkliness, in: H.L.Teulings, A.W.A. Van Gemmert (Eds.), Proceedings of 11th Conference International on Graphonomics Society (IGS2003), Scottsdale, AZ, USA, 2003, pp. 312–315.
[21] M.A. Eldridge, I. Nimmo-Smith, A.M. Wing, R.N. Totty, The variability of selected features in cursive handwriting—categorical measures, J. Forensic Sci. Soc. 24 (1984) 179–219.
[22] C.M. Greening, V.K. Sagar, C.G. Leedham, Automatic feature extraction for forensic purposes, in: Proceedings of Fifth IEE International Conference on Image Processing and its Applications, Edinburgh, UK, 1995, pp. 409–414.
[23] S. N. Srihari, S.-H. Cha, H. Arora, S. Lee, Handwriting identification: Research to study validity of individuality of handwriting and develop computer-assisted procedures for comparing handwriting, Technical Report CEDAR-TR-01-1, SUNY at Buffalo, NY, USA, 2001.
[24] English letter frequencies, ⟨http://www.central.edu/homepages/LintonT/classes/spring01/cryptography/letterfreq.html⟩.
[25] C.G. Leedham, V. Pervouchine, W.K. Tan, A. Jacob, Automatic quantitative letter-level extraction of features used by document examiners, in: H.L.Teulings, A.W.A. Van Gemmert (Eds.), Proceedings of 11th Conference International on Graphonomics Society (IGS2003), Scottsdale, AZ, USA, Avon Books, New York, 2003, pp. 291–294.
[26] C.G. Leedham, V. Pervouchine, W.K. Tan, A. Jacob, Assessment of the stability and usefulness of some handwriting features used by document examiners to identify authorship, in: H.L.Teulings, A.W.A. Van Gemmert (Eds.), Proceedings of 11th Conference International on Graphonomics Society (IGS2003), Scottsdale, AZ, USA, Avon Books, New York, 2003, pp. 316–319.
[27] L. Lam, S.-W. Lee, C.Y. Suen, Thinning methodologies—a comprehensive survey, IEEE Trans. Pattern Anal. Machine Intell. 14 (9) (1992) 869–885.
[28] G.H. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: Proceedings of 11th International Conference on Machine Learning (ML94), Rutgers University, New Brunswick, NJ, 1994, pp. 121–129.
[29] S.M. Weiss, C.A. Kulikowski, Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems, Morgan Kaufmann, Los Altos, CA, 1991.
[30] J. Yang, R. Parekh, V. Honavar, Distal: An inter-pattern distance-based constructive learning algorithm, Technical Report ISU-CS-TR 97-06, Department of Computer Science, Iowa State University, in: Proceedings of International Conference on Neural Networks, IEEE, Piscataway, NJ, 1998 (1997).
[31] H. Vafaie, K. De Jong, Genetic algorithms as a tool for feature selection in machine learning, in: Proceedings on Fourth International Conference Tools with Artificial Intelligence (TAI'92), IEEE Computer Society Press, Arlington VA, 1992, pp. 200–203.
[32] F.Z. Brill, D.E. Brown, W.N. Martin, Fast genetic selection of features for neural network classifiers, IEEE Trans. Neural Networks 3 (2) (1992) 324–328.
[33] J. Bala, J. Huang, H. Vafaie, K. De Jong, H. Wechsler, Hybrid learning using genetic algorithms and decision trees for pattern classification, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada, 1995.
[34] S. Chen, S. Smith, C. Guerra-Salcedo, D. Whitley, Fast and accurate feature selection using hybrid genetic strategies, in: Proceedings on Congress on Evolutionary Computation (CEC99), Washington DC, USA, 1999.

[35] J. Yang, V. Honavar, Feature subset selection using a genetic algorithm, IEEE Intell. Syst. 13 (1998) 44–49.
[36] D.E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Reading, MA, 1989.
[37] G. Syswerda, Uniform crossover in genetic algorithms, in: Proceedings of Third International Conference on Genetic Algorithms, Morgan Kaufmann, George Mason University, USA, 1989, pp. 2–9.

About the Author—VLADIMIR PERVOUCHINE born October 16, 1980 in Orenburg, Russia, received his B.Sc. and M.Sc. degrees from Moscow Institute of Physics and Technology (MIPT), Russia, in 2001 and 2003 respectively. He completed his final year of master degree in the School of Computer Engineering (SCE) in Nanyang Technological University (NTU), Singapore, after joining the NTU-MIPT student exchange in 2002. He received his Ph.D. from NTU in 2006. Currently, he is a Research Associate in the School of Computer Engineering at NTU, Singapore.

About the Author—GRAHAM LEEDHAM graduated in Electrical and Electronic Engineering from Leeds University, UK and obtained his M.Sc. and Ph.D. in Electronics from Southampton University, UK. He was a member of academic staff of Essex University (UK) from 1984 to 1994 and Nanyang Technological University (Singapore) from 1994 to 2006. He has also held visiting research positions in the USA, Canada and Australia. His research interests are in real-time image processing and pattern recognition with particular interest in the analysis and processing of handwriting. He has published over 120 refereed research papers in these areas and organised, chaired or served on the committees of many international conferences. Dr. Leedham is a Fellow of the Institution of Electrical Engineers and is currently Professor of Computer Engineering and Information Technology at the University of New South Wales (Aisa) in Singapore.