Je-LKS
Journal of e-Learning and Knowledge Society
The Italian e-Learning Association Journal

PEER REVIEWED PAPERS
STEPS TOWARD THE DIGITAL AGENDA: OPEN DATA TO OPEN KNOWLEDGE
Vol. 10, n.2, 2014
ISSN: 1826-6223 | eISSN: 1971-8829

TOWARDS THE DEVELOPMENT OF USER TOOLS FOR KNOWLEDGE ACQUISITION IN DIGITAL DOCUMENT ANALYSIS

Michael Fairhurst, Meryem Erbilek
School of Engineering and Digital Arts, University of Kent,
Canterbury, Kent CT2 7NT, UK.
{M.C.Fairhurst and M.Erbilek}@kent.ac.uk

Keywords: document processing, biometric prediction, handwriting analysis

Handwritten documents provide a rich source of data and, with the growth in the availability of digitised documents, it becomes increasingly important to improve our ability to analyse and extract “knowledge” from such sources. This paper describes an approach to the provision of tools which can extract information about the writer of handwritten documents, especially those which were written in earlier times and which constitute key elements in our heritage and culture. We show how the constraints inherent in such documents influence our analytical approach, and we also show how developing appropriate “knowledge extraction” techniques can also be essential in other, more general, important application scenarios.

for citations:
Fairhurst M., Erbilek M. (2014), *Towards the development of user tools for knowledge acquisition in digital document analysis,* Journal of e-Learning and Knowledge Society, v.10, n.2, 35-52.
ISSN: 1826-6223, e-ISSN:1971-8829

1 Introduction

The "Digital Agenda" is driving change in a variety of ways and through many different applications. One area of increasing importance is in the manipulation and analysis of digital documents, with application scenarios ranging from making information more accessible to a wider constituency than would otherwise benefit, through genealogical studies, to forensic investigation and analysis. In this paper we are interested especially in the handling of handwritten documents from a range of possible application scenarios. These might include, for example, using handwriting as a biometric modality to provide personal authentication/encryption options, or to offer secure access to virtual or physical spaces, and so on, while the whole area of forensics can benefit from more extensive and reliable analysis of handwritten documents. However, in the present context, we focus on historical manuscripts as an illustrative practical domain for investigation. Specifically, we note the rapidly increasing availability of such documents - now digitised and accessible - which hitherto were not generally available, or to which access was severely limited because of susceptibility to damage, fragility, or inherent value. In such circumstances, while the data embedded in such documents were sometimes available to researchers and scholars, this was by its nature on a restricted basis, and with considerable effort required to analyse the data contained. The availability of such documents in digitised form now offers opportunities for more data to be made available, and on a much wider basis. However, to turn such data into knowledge requires analytical tools, first to complement the professional skills of scholars based on observation and visual examination, and secondly to make it possible for a wider range of interested individuals to extract knowledge from the raw data. If this can be achieved, then we can take some important steps to supporting more widespread and easier access to our cultural heritage but, also, we can increase the extent to which anyone - from specialist scholars to the general public - can benefit from the knowledge potentially embedded in the data which is now more widely available.

The document analysis and recognition community has shown increasing interest in the processing of historical documents in recent years. As reported in (Antonacopoulos & Downton, 2007), there are several topics which are of particular interest in the analysis of historical documents. Examples include *digitising* (scanning or capturing) historical documents, the digital *enhancement* of the captured documents, and the analysis of the *layout* and that of individual regions of interest. Subsequently, once the historical documents are digitised more research topics arise, such as; the *recognition* of the content of different entities (for instance: does the handwriting belong to a single or to multiple writers, who the writer is, whether the writer is male or female?, and so on), the possibility of *word retrieval* and *word spotting* (words in a collection are matched as images and grouped into clusters which contain all instances of the same word (Rath & Manmatha, 2003)) and so on.

In (Bulacu *et al.*, 2007), for instance, the construction of a layout analyser was described as the first step towards automatic content-based retrieval of document images in historical documents. Word spotting and/or retrieval methods for historical manuscripts are widely investigated (see, for example (Rath & Manmatha, 2003; Liang *et al.*, 2012; Louloudis *et al.*, 2012). Automatic writer recognition using pattern recognition methods have been implemented in a number of studies (Bensefia *et al.*, 2005; Bulacu & Schomaker, 2007; Bulacu & Schomaker, 2007; Bulacu, 2007; Schlapbach & Bunke, 2007; Brink *et al.*, 2008; Brink *et al.*, 2012), and especially with respect to applications in historical documents in (Bulacu & Schomaker, 2007; Schomaker *et al.*, 2007; De Stefano *et al.*, 2011). In these studies, words are usually first segmented into characters, results being combined to form words. However, in (Lavrenko *et al.*, 2004; Madhvanath & Govindaraju, 2001), a holistic word recognition approach is described.

In this paper we will explore some specific tools which might address issues such as these, offering practical and relevant support to those who seek to analyse and understand historical documents in a practical way. The core of the paper is the development of some tools which will help to analyse handwriting. In particular, we will explore the predictive power of handwriting in several relevant contexts, and show how such tools can help to identify the writer, to compare writing objectively, and to predict other, less absolutely defining characteristics of a writer such as gender, handedness, age, and so on. Tools such as these can be seen as broadly generic (in the sense that they should easily be transferrable to other task domains) but, especially, will provide an explicit link between data and knowledge in historical document analysis scenarios.

2 A framework for a knowledge-based toolkit for writer-oriented document analysis

The theme of this paper is that the increasingly vast amounts of data now becoming available from easier access to documents are providing opportunities for such data to be transformed into "knowledge" by means of automated processing. While this idea is generic, in the sense that the principle is sufficiently broad to be directly relevant to a wide variety of document types, this is a particularly interesting concept in relation to the study of historical documents. There are two primary reasons for this. First, such documents are most likely to be handwritten, even though some may be written in a very stylised hand, and this makes analysing the writing more challenging than in other circumstances. Second, the wider availability of such documents necessarily stimulates interest across a greater variety of individuals, and thus the benefits of producing automated analytical tools become correspondingly greater and of more significant potential impact.

While a wide variety of analytical tools can be considered, we will focus here on the extent to which characteristics of the writer can be predicted from the handwriting itself and, since this is an area less fully investigated than other possibilities, we will take as an illustrative characteristic of interest the task of predicting the gender of the writer, although we will also demonstrate that other important writer characteristics can also be predicted using similar tools. In order to support a wide-ranging study of the important issues relevant to this type of task we will use handwritten data generated in a contemporary writing scenario, but we will then demonstrate how such techniques might translate to historical documents, and how limitations to current predictive techniques are imposed in implementing this change of target data.

We initially report an experimental study based on a recently acquired database of handwriting samples captured under various different conditions. Subsequently, we will provide some information about related studies using different data and writing environment to illustrate the range of analytical tools which can be developed. Our previous work (Liang *et al.*, 2013) has demonstrated that many analytical techniques developed for modern handwriting can be transferred to, or modified for, historical document analysis. Thus, we will show how the constraints on data representative of information extracted from digitised versions of historical manuscripts restrict the available analytical options, and what the practical effect of this is likely to be for generating accurate knowledge about the writer, and we will draw some conclusions about the relationship between the data -> knowledge transformation in different types of document and operational environment, linking the fields of palaeography with more contemporary concerns such as biometrics and digital forensics.

3 Experimental infrastructure

In the work reported in this paper the database described in (Fairhurst *et al.*, 2014) is used for the experimental study. Handwriting data in this database were collected for four different task types from 100 subjects (one sample each per task), with a 55:45 gender balance (male:female). In our study, two different types of task (in terms of content) will be adopted in the experiments, defined as follows;

Fixed task: subjects were asked to copy a list of pre-defined words (43 words, 243 characters). The words were chosen to encompass execution of all the most common character-to-character-transitions in English, providing a rich data generation environment for subsequent analysis.

Variable task: subjects were asked to look at a picture and to write a brief description of this in their own words. Hence, in this task, the number of words written was different for each writer.

Acquisition of handwriting samples in this database was carried out in a standard office environment using a Wacom Intuos 5 graphics tablet, under the guidance of a supervisor.

24 dynamic and 26 static features, a total of 50 features in all, which the literature shows are commonly used in signature and handwriting processing (Fairhurst *et al.*, 2014; Lee *et al.*, 1996; Lei & Govindaraju, 2005; Guest, 2006; Chapran *et al.*, 2008; Abreu & Fairhurst, 2009; Erbilek & Fairhurst, 2012; 2013) are extracted from the handwriting data for our experimental study. Classification is performed using three classifiers of differing functionality and complexity: a very simple KNN classifier (with K=1) and also a Naïve Bayes and an SVM classifier, using a hold-out validation methodology (the first 25% of subjects are used in testing, following a training process using the other 75%). For this process the Weka software is used with default settings. It is important to note that the same writer is not included in both the testing and training set at the same time.

4 Prediction from constrained handwriting samples

Although, because of the convenient availability of data which provides information sources of relevance and in sufficient volume for acceptable experimentation, the experiments reported here will be performed on samples from the database of present-day handwriting described above, we will focus on defining the experimental conditions and, in particular, the nature and characteristics of the parameters investigated which are especially pertinent to the conditions most likely to be encountered in the analysis of handwritten historical documents. This is important, because a suitable database of historical writing samples with appropriate metadata to allow us to develop these analytical algorithms directly is not currently available to us. We will also discuss this further in Section 6.

In the present context, we will also narrow the focus of what we attempt to predict from the samples available. We select a single attribute to predict which falls short of specific individual identification of the writer, mainly because this may not always be easy to determine in very old documents, and also because the number of samples per writer would, in a practical situation, often preclude attempting this. Initially, we will explore the prediction of the gender of the writer, under various conditions which reflect in different ways tasks representative of analysis of historical writing, and we will then show some further results which broaden the discussion somewhat by exploring more briefly the prediction of writer age in different circumstances.

However, in order to draw some conclusions about the likely effectiveness of developing tools of particular applicability to heritage documents, we will initially consider the following factors and how they affect our ability to predict writer gender:

- The effect of *classifier choice*
- The effect of extracting *different feature types*, recognising the limitations imposed by the nature of the task of interest
- The effect of the *type of writing fragment* available
- The effect of the *size of the writing fragment* available for analysis

Gender prediction from handwritten documents consisting of a whole page of textual content is presented in (Bandi & Srihari, 2005), using eleven macro features described in (Srihari *et al.*, 2002), obtained from scanned images. Classification accuracy is determined to be 73.2%, 74.7% and 77.5% with a single neural network, ten neural networks combined with bagging and boosting respectively.

In (Liwicki *et al.*, 2007), gender prediction from handwriting data obtained from a conventional white board is presented. Gender is predicted by using both on-line (dynamic) and off-line (static) features types with an SVM and a GMM classifier. A database consisting of more than 200 writers with 8 handwriting text samples per writer is used for the experimental study. The accuracy achieved is found to be between 53.60% and 62.19% with different configurations of an SVM structure, and 67.06% using a GMM classifier. In (Liwicki *et al.*, 2011)the authors investigate gender prediction variability when using on-line (dynamic) features, off-line (static) features, and a combination of both features types, with a GMM classifier and the same database used in (Liwicki *et al.*, 2007). It is concluded that on-line features (accuracy of 64.25%) achieve better performance than off-line (accuracy of 55.39%) features while the best results are obtained with the combination of on-line and off-line features (accuracy of 67.57%).

Results obtained from a competition on gender prediction from handwriting, hosted on Kaggle, is presented in (Hassaine *et al.*, 2013). The dataset used for the experimental study consists of 475 writers with 4 handwritten data samples (1-Arabic handwritten text which is writer variable, 2-Arabic handwritten text which is not writer variable, 3-English handwritten text which is writer variable, 4-English handwritten text which is not writer variable). Only static features are used since the handwriting data are obtained via a scanner. The achieved accuracy (using a Random Forest classifier) is found to be around 70%.

Against this background, we here report a number of experiments to explore gender prediction from the paper-based handwriting samples described in Section 3.

Experiment 1: Here we investigate gender prediction performance when different classifiers are deployed for the prediction task. We use three different classifier platforms for experimentation, of varying complexity, as described above. We use handwritten data from the fixed task, and extract both static and dynamic features as basis for prediction, with the results shown in Figure 1.

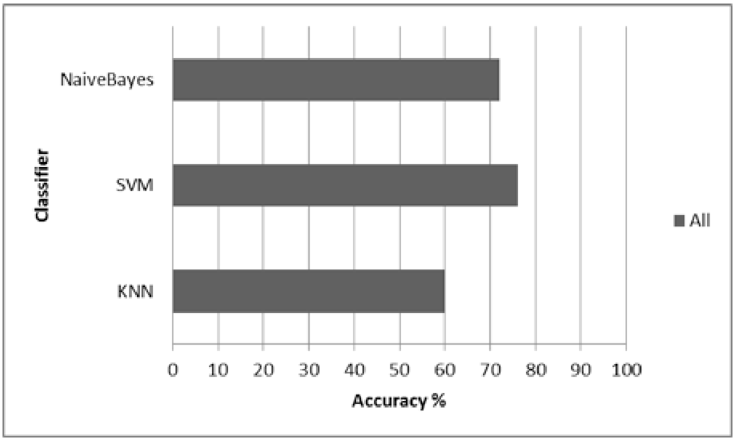

Fig. 1 - Prediction performance as a function of classifier type

Unsurprisingly, perhaps, it can be seen that the performance generated varies in a way which is broadly consistent with the "power" and sophistication of the classifier structure adopted. It is particularly useful, however, to record the quantitative performance metrics, and we see that the accuracy achieved ranges between 60% and around 76%, indicating that careful choice of classifier is an important factor in optimising performance.

Experiment 2: In the next experiment we again compare the same three classifiers, and process the data obtained in the fixed writing task, but now we consider the features used for classification in three different groups. These correspond to static features alone, dynamic features alone, and the full set of all extracted features, both static and dynamic (as in Experiment 1). The experimental results obtained are shown in Figure 2.

Here the results are less obviously predictable, but the attainable performance again shows considerable variation. Here we can see that the choice of feature type used (which, in fact, will generally be a function of the limitations of the data capture process, as will be discussed below) can have a very significant impact on performance, the magnitude of the difference depending on the classifier type adopted. For example, using the simplest K-NN classifier, the choice of features can change the achievable accuracy by as much as 20% but, if this choice is made, this configuration can return the best performance of all.

It is also notable that in these experiments it appears that using only the static features can sometimes generate the highest accuracy, an observation which is somewhat counterintuitive, in that in much handwriting analysis (especially in the field of automatic handwritten signature recognition) there is considerable evidence that the dynamic features provide optimal performance, although in such tasks it is often the balance between false positive and false negative metrics which is most informative, rather than an absolute error measurement, while here we are investigating more general handwriting rather than the handwritten signature alone. Writer identification from generalised text has its own specialised techniques too, and it should not be assumed that techniques found to be optimal in one task domain necessarily transfer to another domain without a degree of fine tuning.

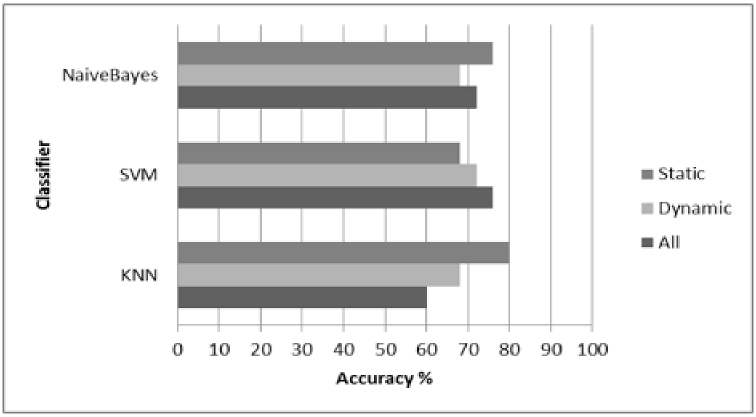

Fig. 2 - Performance as a function of feature type

What is particularly important here is that, while dynamic features are readily available when data capture is achieved using on-line writing acquisition, when we examine historical documents specifically, this is self-evidently not an option, and hence we are limited to the utilisation of only static features extractable from the images available of the original written material. There may be a possibility of inferring certain aspects of dynamic execution, but this would represent the limit of what can be determined about dynamic aspects of the data available.

Experiment 3: Our experimental database captured handwriting samples in two different task scenarios, first in a task where all writers wrote the same text (essentially based on a copying task where individuals copied a prescribed textual fragment) and, second, a task where individuals wrote a brief description of a visual scene, thereby ensuring that the textual fragment generated was variable across participants. This reflects many real-world situations where, sometimes, writers are likely to produce similar responses in a constrained task domain, in contrast to others where the writing content is free-form. This reflects, too, possible scenarios with historical manuscripts, some of which (for example in copying, say, biblical or liturgical text passages) will use essential the same material, and others of which will contain different textual material.

Figure 3 shows the results obtained in the two cases (fixed and variable task content), using the same three classifiers as previously tested, and for different feature types as before. Here the "fixed task" results are the same as those shown in Figure 2, but are included here for ease of comparison with those obtained in the "variable task" scenario. It can be seen that there is a substantial difference in performance depending on the nature of the task, with accuracy reduced for all classifiers when variable textual content is used as a basis for prediction, but also with a lower degree of variability observed as a function of feature type utilised. It is notable also that the differential in accuracy between the cases when static features alone are used compared with dynamic features, or a mix of both, is less than previously seen, although with variable content the adoption of static features alone generally results in poorer performance, and choice of classifier becomes a more sensitive issue.

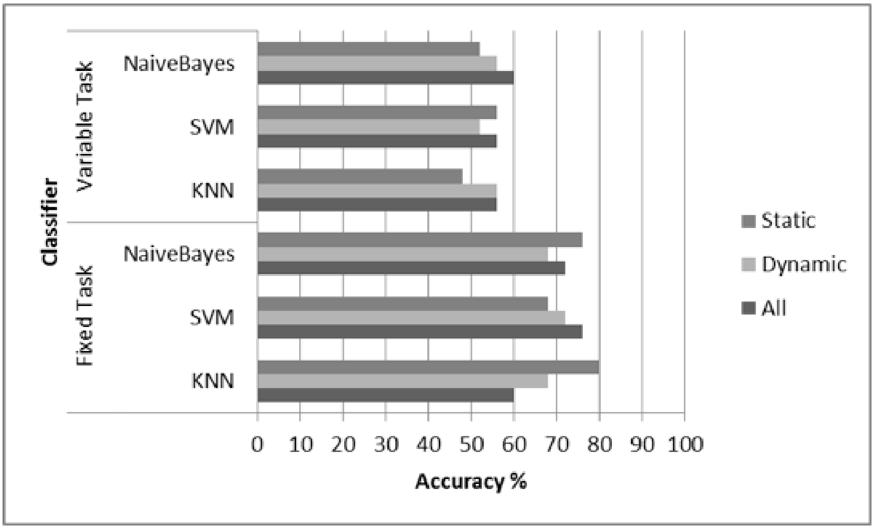

Fig. 3 - performance as a function of task domain characteristics

Experiment 4: It is apparent that documents for analysis are likely to vary considerably in terms of their general characteristics, both as noted above, and also in terms of their length. Specifically, in some cases analysis may be required in circumstances where only a small fragment of writing is available. While this is most acutely an issue perhaps in the analysis of modern documents (for example, in forensic document analysis for crime investigation), it is also a significant possibility in dealing with historical documents, perhaps especially where prediction of broad (soft) characteristics of the writer may be important.

It is intuitive to suggest that shorter fragments might be expected to carry a smaller amount predictive information than larger fragments, but it is less easy to suggest intuitively how this relation can be expressed quantitatively. In our next experiment we therefore investigate the relation between predictive performance and writing fragment size.

In this experiment we use all the available features and compare results on the basis of the size of the fragment used for gender prediction. In order to determine some quantitative data we divide the individual responses according to the number of words in the fragment, producing two predictive scenarios based on fragments of 10 and 30 words in length respectively.

The results are shown in Figure 4, where we have produced results for the same classifiers previously considered. We see that the predictive capacity across the two tasks varies in a similar way to that observed previously, with the fixed task supporting generally improved performance, but here the differential between the two scenarios is less marked overall.

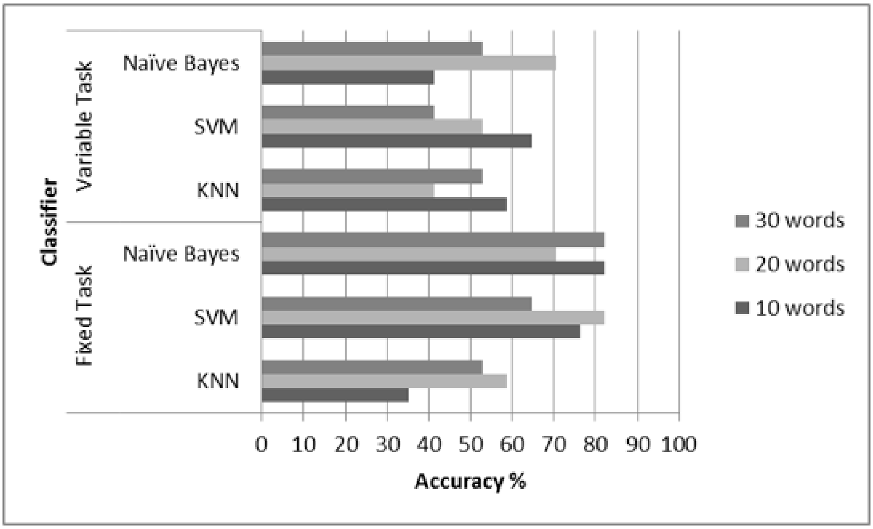

Fig. 4 - Performance as a function of sample fragment size

It can be seen, therefore, that increasing the amount of information available has a relatively small effect in the task where the writing is fixed, but becomes more significant when the target writing is not predetermined. It should be noted that the target fragment in the case of the fixed task is chosen specifically to include a rich range of typical digraphs in English, giving a robust source of discriminatory material, while with a target of unspecified content the performance change can benefit from more information improving both the number of discriminatory features available and better estimates of features variability in some cases, even if fewer different writing patterns are observable.

The restriction on the available database (only 67 participants wrote at least 30 words in the variable content task) limits the scope of this part of the investigation, and it is not possible to predict whether significant further changes would occur with the availability of much larger text fragments, but this is a useful preliminary investigation nevertheless.

The experimental results reported above have shown that the prediction of some very specific writer characteristics from handwritten data is possible, even when relatively small and/or content-variable textual fragments are available for analysis. We have shown how some of the factors which are controllable, and some which are not, are likely to affect prediction performance, and this is important in a variety of practical tasks. However, so far we have described only one example of a prediction task, that of predicting writer gender from freehand writing samples.

In order to demonstrate the generality of our ideas - and broaden the discussion to other data forms - we finally briefly report some further experiments on a different database. This database consists entirely of samples of the handwritten signature (providing an alternative type of data format) but also contains multiple samples for each writer which are age-tagged, enabling us to demonstrate both the extent to which the age of a writer can be predicted, and how the sort of tools proposed can be deployed in a different writing scenario.

For these final experiments we use the *Data Set 2 (DS2) of the BioSecure Multimodal Database (BMDB)* database (Biosecure http:\\www.biosecure.info, 2004). The biometric samples in this database were collected as part of an extensive multimodal database by 11 European institutions participating in the BioSecure Network of Excellence (Biosecure http:\\www.biosecure.info, 2004), and is a commercially available database. The database contains biometric samples from 210 users in the age range of 18-73 years. Acquisition of signature samples in this database were carried out in a standard office environment using a Wacom Intuos 3 A6 graphics tablet with a sample rate of 100 Hz, under the guidance of a supervisor (Ortega-Garcia *et al.*, 2010). 30 genuine signature samples were acquired from each user, collection split across two different sessions. In our experimental study, 15 signature samples were used per person. The same experimental infrastructure is used as that specified for the previous experiments, although after feature extraction, each feature is normalised using a mean and variance normalisation technique (Erbilek & Fairhurst, 2012; 2013).

Experiment 5: Here, then, we move from general handwriting, to the predictive properties of the more restricted handwritten signature. First we investigate the prediction of gender from this signature (rather than more general handwriting) data, with the results shown in Figure 5. In this experiment we have considered only static features in the prediction task.

Here we see that, again, performance is very sensitive to the classifier adopted, as before, but attainable accuracy is poorer than in the case when general handwriting samples are used as the predictor. This reflects the smaller amount of data available in this type of scenario, even though there is an argument to say that the signature is often more individually characteristic of the writer than general handwriting.

However, a secondary purpose of experimentation with this alternative database is to demonstrate that other “soft” characteristics are also predictable from handwritten samples. In this example we look to predict the age of the writer. However, age is a continuous function and it is well known that most age-related characteristics develop on a relatively slow timescale, and thus it is customary in work of this sort to predict a broad age grouping rather than a specific age. In these experiments we define three age groupings (namely age less than 25, age between 25 and 60, and age greater than 60), this division of a population being commonly reported in the literature, and has been shown (Erbilek & Fairhurst, 2012) to be justifiable from a theoretical point of view. The results are also shown in Figure 5.

These results show an accuracy poorer than that obtained for gender prediction. However, this is a more difficult task, not least because this is a three-class rather than a two-class problem, but the results are illustrative of generalising the idea of analytical tools to a variety of soft characteristics of a writer based on the extraction of features available when the target textual material is taken from a digitised representation of a historical document.

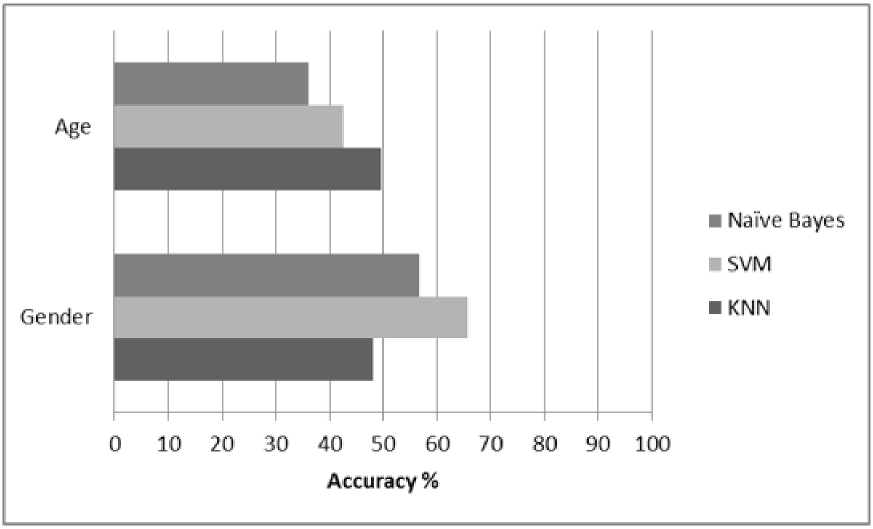

Fig. 5 - Gender and age prediction from the handwritten signature

5 Discussion

The investigation described above has shown some examples of an approach to develop tools for analysing handwritten script of varying forms, collected under a variety of different conditions. Since they provide an automatic means of specific information extraction, it is tools such as these which can easily facilitate the extraction of knowledge from high volume handwriting data. We have shown examples of how to build tools which might provide easy access to knowledge embedded in historical documents, thus making possible the discovery of a greater depth of understanding about aspects of culture and heritage, revealing features which have hitherto been largely unavailable, discernible only to specialist scholars, or extremely demanding to determine. As more and more data sources become available, this type of approach becomes increasingly important in allowing knowledge accumulation to keep pace with data availability.

It is clear that most of the work reported has been developed through experimentation based on handwriting data which pertain to modern writing. In the previous sections, however, we have identified some of the constraints which will apply when transferring the tools discussed to handwriting from the illustrative target scenario of interest, and we thus have shown that, although subject to some limitations, the sort of tools discussed are evidently transferrable to such applications. Indeed, processing the original handwriting data of our first experiments using a configuration suitable for processing historical manuscript data, but selecting optimal conditions as revealed by the results presented (using the K-NN classifier, a fixed content task, the largest available fragment size supported in the database, but restricting the feature set to static features), shows a gender prediction accuracy of 80%.

However, caution must be exercised here, since we must acknowledge that handwriting styles and document characteristics vary significantly over time, and thus we cannot assume that optimised performance on our current database will provide a direct benchmark for performance on other documents and writing styles. However, there is some evidence to suggest that we should be optimistic. In our own work, for example, as we have shown in (Liang *et al.*, 2013 (Internal report, as yet unpublished)), and supported by other studies (Bulacu & Schomaker, 2007), through experimentation on different databases (IAMDB (Marti & Bunke 2002) and an in-house, locally compiled database consisting of samples from four Medieval/Early Modern Manuscripts of the 16th Century), many extracted features are equally effective for both modern and older text and, where this is not the case, we now have a better understanding of manuscript-oriented optimal feature selection. We should not overlook either the fact that it is possible to infer some dynamic properties from static textual images which, although more limiting than using true dynamic features where they are available, provides a further source of information which, as our experiments above demonstrate, are likely to provide some improvement over static feature-based analysis alone.

The approach we have described, and the principle of providing analytical tools as a basis for extracting knowledge from data, thus appears to be sound, and remains an on-going motivation to drive this work forward in the future.

Conclusion

The results provided here demonstrate that extremely useful analytical tools can be developed which can extract interesting and valuable data from documents under the sort of practical constraints which pertain when working with heritage documents. The experiments are illustrative, but point to opportunities for providing tools which are likely to be instrumental in providing scholars, historians, custodians of important documents and, indeed, ordinary citizens, with an opportunity to embrace and exploit the increasing availability of digitised documents, creating knowledge from the large volume of raw data by extracting important characteristics of the writer. By extension, we can also see how such basic ideas can be developed to extract other sorts of data too - information about writing style, which can provide clues to place and writing era, characteristics of the background surface on which the writing is superimposed, comparing documents, identifying common hands across different documents, differentiating the genuine from the artefact, and so on.

Thus, we have shown how a focus on the predictive information inherent in handwriting can be especially valuable in seeking knowledge from data. It is clear that much remains to be done in relation to the transfer of techniques such as those described here to the target documents themselves, but related work has already begun to show how it is possible to identify extracted features which are optimised in relation to document provenance, with differential optimisation properties when considering historical documents and modern handwriting. Key to this process will be the compilation of an appropriate database of historical document samples with validated metadata appended, which we are currently exploring.

It is important to emphasise that although we have presented results in this paper in the (illustrative) context of historical document analysis - manifestly an important and very worthwhile application of the principles proposed - the work reported here is, of course, of great relevance in many other applications too and, indeed, in the analysis of any handwritten document. Importantly, while focusing primarily on the prediction of the gender of the writer as an exemplar of the possibilities and providing a comparative reference point, we have also shown some examples of the prediction of other characteristics too, and used the context of a different type of handwritten target to show the breadth of opportunity the approach offers. Many other characteristics can also be considered, and some particularly interesting recent work is the prediction of mental state from handwriting samples, and some preliminary results can be found in (Fairhurst *et al.*, 2014).

More generally, this sort of work is pointing very clearly to the benefits of crossing the boundaries between traditionally separate disciplines. Working with characteristics such as writer age and gender, for example, is a key element in the exploitation of so-called soft biometrics, an important aspect of the development of techniques for biometrics-based identification of individuals, while the desirability of predicting such characteristics from handwriting samples, and indeed, other "biometric" measurements (see, for example, work on soft biometric prediction from iris patterning in (Lagree & Bowyer, 2011; Erbilek *et al.*, 2013; Sgroi *et al.*, 2013) is now recognised. Indeed, it is possible to build greater intelligence into the classification process and hence predict soft biometrics and subsequently better utilise this information in identification tasks (Abreu, 2010; Abreu & Fairhurst, 2011). This increasing integration of ideas and practical techniques which generate commonalities across apparently diverse fields of study is perhaps one of the most encouraging indicators of how we might most effectively manage "big data" in the future.

Acknowledgement

The authors would like to thank Cheng Li for his leading role in data collection.

REFERENCES

Weka 3. http://www.cs.waikato.ac.nz/ml/weka/.

Abreu, M. and M. Fairhurst (2009), *Improving Identity Prediction in Signature-based Unimodal Systems Using Soft Biometrics*. Biometric ID Management and Multimodal Communication. J. Fierrez, J. Ortega-Garcia, A. Esposito, A. Drygajlo and M. Faundez-Zanuy, Springer Berlin / Heidelberg. 5707: 348-356.

Abreu, M. and M. Fairhurst (2011), *Combining Multiagent Negotiation and an Interacting Verification Process to Enhance Biometric-Based Identification*. Biometrics and ID Management. C. Vielhauer, J. Dittmann, A. Drygajlo, N. Juul and M. Fairhurst, Springer Berlin Heidelberg. 6583: 95-105.

Abreu, M. C. D. C. (2010), *Enhancing classification structures for biometric applications*. Ph.D. dissertation, University of Kent.

Antonacopoulos, A. and A. Downton (2007), *Special issue on the analysis of historical documents*. International Journal of Document Analysis and Recognition (IJDAR) 9(2-4): 75-77.

Bandi, K. and S. N. Srihari (2005), *Writer demographic classification using bagging and boosting*. Proc. 12th Int. Graphonomics Society Conference.

Bensefia, A., T. Paquet, *et al.* (2005), *A writer identification and verification system*."Pattern Recognition Letters 26(13): 2080-2092.

Biosecure (http:\\www.biosecure.info, 2004), *Biometrics for Secure Authentication*, FP6 NoE, IST-2002-507634.

Brink, A., M. Bulacu, *et al.* (2008), *How much handwritten text is needed for text-independent writer verification and identification*, IEEE.

Brink, A., J. Smit, *et al.* (2012), *Writer identification using directional ink-trace width measurements*. Pattern Recognition.

Bulacu, M. and L. Schomaker (2007), *Automatic handwriting identification on medieval documents*. ICIAP, IEEE.

Bulacu, M. and L. Schomaker (2007), *Text-independent writer identification and verification using textural and allographic features*. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE: 701-717.

Bulacu, M. and L. Schomaker (2007), *Text-independent writer identification and verification using textural and allographic features*. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(4): 701-717.

Bulacu, M., R. van Koert, *et al.* (2007), *Layout Analysis of Handwritten Historical Documents for Searching the Archive of the Cabinet of the Dutch Queen*. Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on.

Bulacu, M. L. (2007), *Statistical pattern recognition for automatic writer identification and verification*, PhD thesis.

Chapran, J., M. C. Fairhurst, *et al.* (2008), *Task-related population characteristics in handwriting analysis*. Computer Vision, IET 2(2): 75-87.

De Stefano, C., F. Fontanella, et al. (2011), *A method for scribe distinction in medieval manuscripts using page layout features*. Image Analysis and Processing–ICIAP 2011: 393-402.

Erbilek, M. and M. Fairhurst (2012), *Framework for managing ageing effects in signature biometrics*. Biometrics, IET 1(2): 136-147.

Erbilek, M. and M. Fairhurst (2012), *A Methodological Framework for Investigating Age Factors on the Performance of Biometric Systems*. The 14th ACM Workshop on Multimedia and Security, Coventry, UK.

Erbilek, M. and M. Fairhurst (2013), *Analysis of ageing effects in biometric systems: difficulties and limitations*. Age factors in biometric processing. M. Fairhurst., IET: 279-301.

Erbilek, M., M. Fairhurst, *et al.* (2013), *Age Prediction from Iris Biometrics*. 5th International Conference on Imaging for Crime Detection and Prevention (ICDP13).

Fairhurst, M. and M. Abreu (2009), *An Investigation of Predictive Profiling from Handwritten Signature Data*. Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on.

Fairhurst, M., M. Erbilek, *et al.* (27-28 March 2014), *Enhancing the forensic value of handwriting using emotion prediction*. 2nd International Workshop on Biometrics and Forensics (IWBF), Valletta, Malta.

Guest, R. (2006), *Age dependency in handwritten dynamic signature verification systems*. Pattern Recognition Letters 27(10): 1098-1104.

Hassaine, A., S. Al Maadeed, *et al.* (2013), *ICDAR 2013 Competition on Gender Prediction from Handwriting*. Document Analysis and Recognition (ICDAR), 2013 12th International Conference on.

Lagree, S. and K. W. Bowyer (2011), *Predicting ethnicity and gender from iris texture*. Technologies for Homeland Security (HST), 2011 IEEE International Conference on.

Lavrenko, V., T. M. Rath, *et al.* (2004), *Holistic word recognition for handwritten historical documents*. Document Image Analysis for Libraries, 2004. Proceedings. First International Workshop on.

Lee, L. L., T. Berger, *et al.* (1996), *Reliable online human signature verification systems*. Pattern Analysis and Machine Intelligence, IEEE Transactions on 18(6): 643-647.

Lei, H. and V. Govindaraju (2005), *A comparative study on the consistency of features in on-line signature verification*. Pattern Recognition Letters 26(15): 2483-2489.

Liang, Y., R. M. Guest, *et al.* (2012), *Implementing Word Retrieval in Handwritten Documents Using a Small Dataset*. Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on.

Liang Y., Fairhurst M., *et al.* (2013) ((Internal report, as yet unpublished)), *Automatic handwriting feature extraction, analysis and visualisation in the context of digital palaeography*.

Liwicki, M., A. Schlapbach, *et al.* (2011), *Automatic gender detection using on-line and off-line information*. Pattern Analysis and Applications 14(1): 87-92.

Liwicki, M., A. Schlapbach, *et al.* (2007), *Automatic detection of gender and handedness from on-line handwriting*. Proc. 13th Conf. of the Graphonomics Society.

Louloudis, G., A. L. Kesidis, *et al.* (2012), *Efficient Word Retrieval Using a Multiple Ranking Combination Scheme*. Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on.

Madhvanath, S. and V. Govindaraju (2001), *The role of holistic paradigms in handwritten word recognition*. Pattern Analysis and Machine Intelligence, IEEE Transactions on 23(2): 149-164.

Marti, U. V. and H. Bunke (2002), *The IAM-database: an English sentence database for offline handwriting recognition*. International Journal on Document Analysis and Recognition 5(1): 39-46.

Ortega-Garcia, J., J. Fierrez, *et al.* (2010), *The Multiscenario Multienvironment BioSecure Multimodal Database (BMDB)*. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32(6): 1097-1111.

Rath, T. M. and R. Manmatha (2003), *Features for word spotting in historical manuscripts*. Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on.

Schlapbach, A. and H. Bunke (2007), *A writer identification and verification system using HMM based recognizers*. Pattern Analysis & Applications 10(1): 33-43.

Schomaker, L., K. Franke, *et al.* (2007), *Using codebooks of fragmented connected-component contours in forensic and historic writer identification*. Pattern Recognition Letters 28(6).

Sgroi, A., K. W. Bowyer, *et al.* (June 2013), *The Prediction of Young and Old Subjects from Iris Texture*. IAPR International Conference on Biometrics.

Srihari, S. N., S.-H. Cha, *et al.* (2002), *Individuality of handwriting*. Journal of Forensic Sciences 47(4): 856-872.