AUTOMATIC HANDWRITING FEATURE EXTRACTION, ANALYSIS AND VISUALISATION IN THE CONTEXT OF DIGITAL PALAEOGRAPHY

Y. LIANG[A], M.C. FAIRHURST[B], R.M. GUEST[C1], M.ERBILEK[D]

[a,b,c,d1]*School of Engineering and Digital Art, University of Kent, Canterbury, Kent CT2 7NT, UK*

Email: {M.C.Fairhurst[b], R.M.Guest[c], M.Erbilek[d1]}@kent.ac.uk, yl67@kentforlife.net[a]

Digital palaeography is an emerging research area which aims to introduce digital image processing techniques into palaeographic analysis for the purpose of providing objective quantitative measurements. This paper explores the use of a fully automated handwriting feature extraction, visualisation and analysis system for digital palaeography which bridges the gap between traditional and digital palaeography in terms of the deployment of feature extraction techniques and handwriting metrics. We propose the application of a set of features, more closely related to conventional palaeographic assesment metrics than those commonly adopted in automatic writer identification. These features are emprically tested on two datasets in order to assess their effectiveness for automatic writer identification and aid attribution of individual handwriting characteristics in historical manuscripts. Finally, we introduce tools to support visualisation of the extracted features in a comparative way, showing how they can best be exploited in the implementation of a content-based image retrieval (CBIR) system for digital archiving.

Keywords: Digital palaeography, manuscript exploration, image analysis.

1 Introduction

Palaeographic analysis concerns the study of ancient writing, of which the most prominent applications include identifying the date, place of origin, writer(s) and other information about a specific script. The importance of using digital image processing techniques in the analysis of digitised cultural archives has been noted by both the scientific and humanities-based communities [1]. In this respect, the desirability of introducing image processing techniques into palaeographic analysis for the purpose of providing objective quantitative measurements has been identified by a number of researchers [2-6]. A palaeographer's work commonly involves the analysis of handwriting features with regard to three important questions about the written text: i) "When was this written?", ii) "Where was this written?" and iii) "Were these different texts written by the same person?" [4]. A number of quantitative technology-based metrics have been reported in the literature, but the technology is not yet mature [4], with most of the work in this field to date having been carried out by small groups of inter-disciplinary researchers [2-7].

In order to address these important issues, in this paper we investigate methods inspired by conventional non-ICT-based palaeographic evaluation of documents specifically to assess the question of writer attribution (question iii in the above list). Our methodology is novel in that, by using recognised methods, we aim to bridge the gap between traditional and digital palaeography. We initially trial our methods within the historical document domain to demonstrate the range of information that can be extracted, but the fully automated processes of feature extraction and analysis allow for easy adoption by different research communities. Furthermore, visualisation enables intuitive interpretation of results between different end-use scenarios and discplines.

2 Related Work

2.1 *Traditional palaeography analysis*

The analysis of the handwriting style of a written segment constitutes an important part of palaeographic analysis. Over the years, considerable effort has been made to standardise a particular methodology for this purpose. In 1952, Jean Mallon described handwriting with regard to seven aspects [5]: (1) Form - "the morphology of the letters", (2) Pen angle - "in relation to the

[1] Corresponding author. Tel.: +44 (0)1227 823717; fax: +44 (0)1227 456084.

base line", (3) Ductus - "the sequence and direction of a letter's different traces", (4) Modulus - "the proportions of the letters", (5) Weight - "the difference in thickness between the hair lines and the shadow lines", (6) Writing support and (7) Internal characteristics - "the nature of the text".

Although quantitative and objective descriptions of these aspects are still undefined, this list of features provide the basis for a palaeographer's work, hence these particular methods have been adopted widely over the years, and the validity of his theory has been accepted for palaeographical examination [5]. Following the Mallon proposal, several other attempts have been made to extend the defined list with new elements. For instance, new criteria were proposed by Lothar Michelin in 1982 and Jan Burgers in 1995. These criteria, as described in [4, 8], are illustrated in Table 1.

From an assessment of these features and methods, a trend towards detailed, quantitative and objective metrics has been observed in the development of tools for palaeography analysis over recent years. In the light of the popularity of computer assisted tools and the availability of digital representation of historical documents, the development of palaeography is taking another step forward with the adoption of digital technology.

Table 1: Proposed Criteria

Lothar Michelin's Criteria [9]	**Jan Burgers's Criteria [10] [4][34]**
• Quality of the strokes • Change of pressure exerted on the pen • Writing speed • Movement and form • Direction of movement: a) ductus, b) angle of inclination, c) form and direction of the base line • Vertical proportions • Horizontal proportions • Vertical division • Horizontal division • Other characteristics	• Slant • Writing angle • Weight • Modulus • Format • Width of the margins • Ruling and irregularities of the base • Flourishes and other decoration • "Text structure", punctuation and use of majuscules and capitals • Abbreviations • Cursiveness between letters • Cursiveness within letters • Characteristic letter forms

2.2 Digital palaeography

The adoption of statistical and mathematical methods in palaeography has been proposed in a number of studies in recent years [2, 3, 7, 11-14]. Two different approaches can be identified in these studies with respect to the type of measurements taken from the manuscripts: *local* and *global*.

The approaches in the local category attempt to replicate the work of palaeographers, which requires the analysis of some particular letters, or the descender/ascender parts of letters. Without successful character segmentation and handwriting recognition methods, human interaction is necessary in these approaches. In contrast to these local approaches, the methods in the global category are identified by the automatic extraction of global characteristics of the text within the manuscript.

For local analysis, the 'System for Palaeographic Inspections' (SPI) tool was developed in [3, 7], in order to assist the manual segmentation of individual characters and ligatures from handwritten text. Following segmentation and labelling, quantitative measurements of the characters (e.g. width, height, vertical histogram, etc.) are taken and 'letter models' established. The letter models, in combination with other domain information collected by the user, assisted in the process of a palaeographer's work. Although semi-automated approaches are welcomed by the palaeography community due to character-specific feature extraction, not all features considered in traditional palaeography are available in the SPI tool. Moreover, there is a lack of systematic support by the SPI tool for the various abstraction stages within the letter modeling process (including the segmentation process, model generation, setting of morphological parameters, comparisons and measurements), which led to an interruption of the development of the tool itself, as acknowledeged by the authors [7].

In the case of the global category, the work in [11], for example, aimed to differentiate the main categories of writing styles using quantitative measurements obtained from the manuscripts in a trained classification system. The result of [11] suggested that the handwriting style, if well presented in terms of statistical measurements, can be connected to the historical period and/or the geographical localisation independently of the writer's personal style. Automatic feature extraction and machine learning methods are also adopted in [12] for the purpose of scribe distinction. The main feature adopted in [11] was a joint probability of observing the same intensity value between two different pixels in relation to their spatial location, named Spatial Gray-Level Dependence, while those in [12] relate to the page margins, inter-column and row spacing.

2.3 Automatic writer identification

Forensic document analysis is a topic directly relevant to palaeography with respect to the question "Were these different texts written by the same person?". The handwriting features used in the forensic document analysis community include class of allograph (the writing form of a grapheme - a character or a part of a character), design of allographs, dimensions, slant, intra-word and inter-word spacing, baseline alignment and line continuity [15]. Similar to the study of palaeography, the work in forensic document examination suffers from possible subjectivity and the lack of standardisation. Also, for legal reasons, researchers are motivated to validate the individuality of handwriting through the implementation of objective quantitative measurements [15-17]. Automatic writer identification systems using pattern recognition methods have been implemented in a number of studies [18-24], with direct applications in historical manuscripts in [12, 19, 25]. The approaches established in the studies related to automatic writer identification are in fact similar to the solutions to digital palaeography in the global category.

2.4 Analysis of the handwriting features

The extracted features may not all show the same level of distinctiveness across the entire range of writers. For example, two writers may show differences in the word proportion but share common characteristics in word spacing, while the word proportion may be a distinctive pattern of a third writer. Instead of developing a universal model for identifying scribal hands as in previous works in automatic writer identification/verification [12, 18-25], we suggest that it is more useful to analyse the handwriting features individually.

As previously discussed, scribal hand identification/verification is only one of the questions in palaeographic analysis, and the handwriting features can be analysed in relation to many different questions. When comparing two handwriting samples, for instance, the question can be abstracted to "Are these two samples similar enough that they can be considered as coming from the same population?". The "population" represents each of the subgroups under observation. In order to evaluate the usefulness of each individual feature, however, it is essential to analyse them within a well-defined context. For the purpose of benchmarking the usefulness of individual features compared with those described in the literature, we relate our analysis to the question of scribal hand verification.

2.5 Objectives

It can be concluded from the above discussion that the adoption of computer-aided methodologies by traditional palaeographers is hindered by the fact that most of the handwriting features are not available in a digital platform. Local approaches have attempted to address this problem by providing character-specific handwriting features, but they have required intensive human interaction in the process. Automation of the process being a primary goal, global approaches are intended to provide a rather different perspective on palaeographic analysis.

Another possible reason for the slow adoption of computer-aided methodologies is that there is little support for the analysis of the features irrespective of the application, i.e. writer identification, dating the manuscript, or identifying the source of origin of the manuscript. The work reported in the literature to date has focused on one specific application of palaeography. The goal of digital palaeography, as identified in [13], is not to replace, but to enhance and extend the traditional methodology. Therefore, it is useful to provide a tool to compare the individual features between two different samples regardless of the end-use scenario.

In order to improve the acceptance of computer-aided methodology in the palaeographic community, we will introduce a range of handwriting features that have been successfully implemented in automatic writer identification into digital palaeography and also propose a set of new features that directly or indirectly measure some of the conventional features used by forensic document examiners, while at the same time stressing the importance of a fully automated extraction and visualisation process. The visualisation tool allows the palaeographer visually to compare a single feature between two samples. Additionally, a set of experiments is designed to assess the features for the task of scribe distinction, namely identifying the number of writers in a manuscript. Conventionally, a training process is required to establish an automatic writer identification/verification system, wherein the number of existing writers is known a-priori and a model is generated for each writer by comparing the inter-writer and intra-writer variation in the feature values. However, this training-testing paradigm is brought into question in the context of this study, due to the fact that the number of existing writers is not generally available a priori. Hence, an alternative training-free paradigm for establishing an automatic writer verification system will be presented and discussed in the following sections.

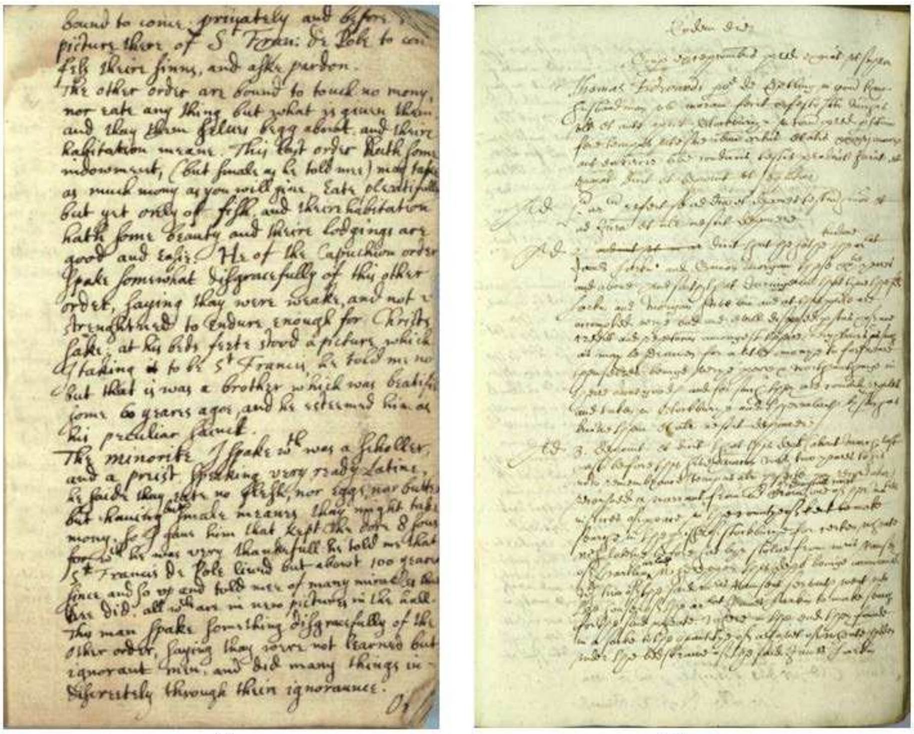
3 Datasets and Pre-processing

3.1 Datasets

Two datasets are adopted for the experimental study described in this paper. The first dataset consists of four Medieval/Early Modern manuscripts of the 16^{th} century, provided by the Canterbury Cathedral Archives. These manuscripts include John Bargrave's travel diary (in English, known writer), records of the Canterbury French Huguenot Church (in French, unknown scribes) and a court deposition (in Latin, unknown scribes). For ease of subsequent reference, we designate this dataset MEMDB.

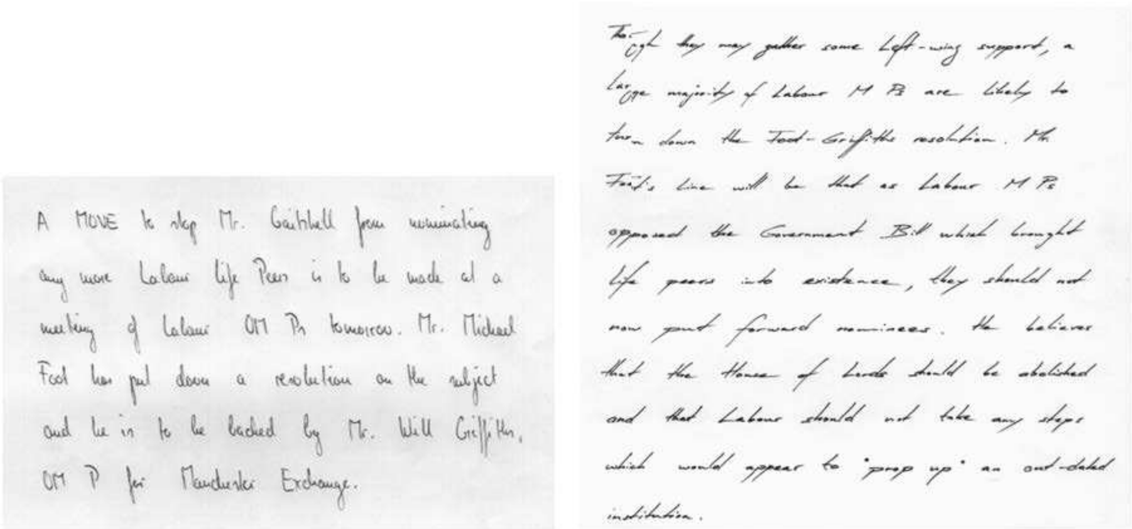
Although the context of the current study relates to historical documents, in order to report comparative results with respect to to previous studies, a commonly adopted and publicly available contemporary multi-writer database IAMDB [26] is utilised as the second dataset for our investigation.

The MEMDB and IAMDB databases consists respectively of five (as verified by an expert with detailed knowledge of the documents adopted) and 200 writers, and provide 24 and 400 pages of text respectively. Samples from each dataset are illustrated in Figure 1 and Figure 2. As can be seen, the two datasets demonstrate very different characteristics in terms of image quality, colour, type of ink, etc. These characteristics are largely irrelevant to handwriting styles, and hence, for the purposes of this particular study, are removed during the pre-processing stage. Additionally, since, typically, the number of scribal hands may not be known, we develop, using the MEMDB dataset, a new paradigm for a writer verification system that does not directly depend on a training process using samples written by known writers. Our groundtruth knowledge for this database, noted above, is important in supporting this development.



(a) (b)

Figure 1: (a) Sample 1 and (b) Sample 2 from MEMDB dataset



(a) (b)

Figure 2: (a) Sample 1 and (b) Sample 2 from IAMDB dataset

3.2 Pre-processing

The pre-processing stage of the manuscript images consists of three tasks: binarisation, line segmentation and directional-chain-code [18, 27] extraction. The localised binarisation method described in [35] is adopted in this work for the first task. The methods for the second and third tasks will be described in more detail in this section.

3.2.1 Line segmentation

A simple Fourier Transform on the horizontal projection of ink pixels is employed in order to find the body of a text line and the gaps between text lines. The horizontal projection of ink pixels of the manuscript image in Figure 1 (a) is shown in Figure 3, where the peaks correspond to the bodies of the text line and the troughs are the likely gaps between lines. The stages of the text line segmentation are as follows;

1) The number of text lines are found by analysing the frequency components of the profile of the horizontal projection as the result of Fourier transformation [28]. Except for the zero frequency component, the frequency component with highest energy is the number of text lines, because it represents the number of times the most significant pattern repeats itself within the image. Indeed, there are 33 text lines in the manuscript image shown in Figure 1 (a).

2) The average height of a text line is calculated as the height of the writing block within the page divided by the number of text lines.

3) The local minima (the main body of text lines) and maxima (the separation between text lines) within a neighbourhood of the average height of a text line on the horizontal projection are calculated.

4) The page is initially segmented into lines according to the local maxima. An example text line from the MEMDB as the result of the initial segmentation is shown in Figure 4 (a) and, as can be observed, some text lines are slanted. In order to adjust the slant, the text lines are rotated to an angle at which the local maxima above and below the current text line are maximised. The result after slant adjustment is shown in Figure 4 (b).

5) The local minima and maxima are recalculated due to the change induced by the rotation. The local maxima above and below the text line are considered as the separations from the lines above and below. The connected components, of which the centroid is outside the boundary set by the local maxima, are considered as writing from the lines above and below, and hence are erased. The result is shown in Figure 4 (c).

With this methodology, the bodies of text lines in our test datasets were identified with 100% accuracy. However, due to the cluttered handwriting in the MEMDB dataset, some residue of writing from the adjacent lines can be found in all of the text lines after the segmentation, which, in all likelihood would probably be a contributory factor to the degradation of overall feature performance. This method does, however, represent a generic solution to assess the number of lines thereby giving consistent application of assessment rules. This may, however, need tuning to particular manuscript cases, forming the basis for further experimentation with this method.

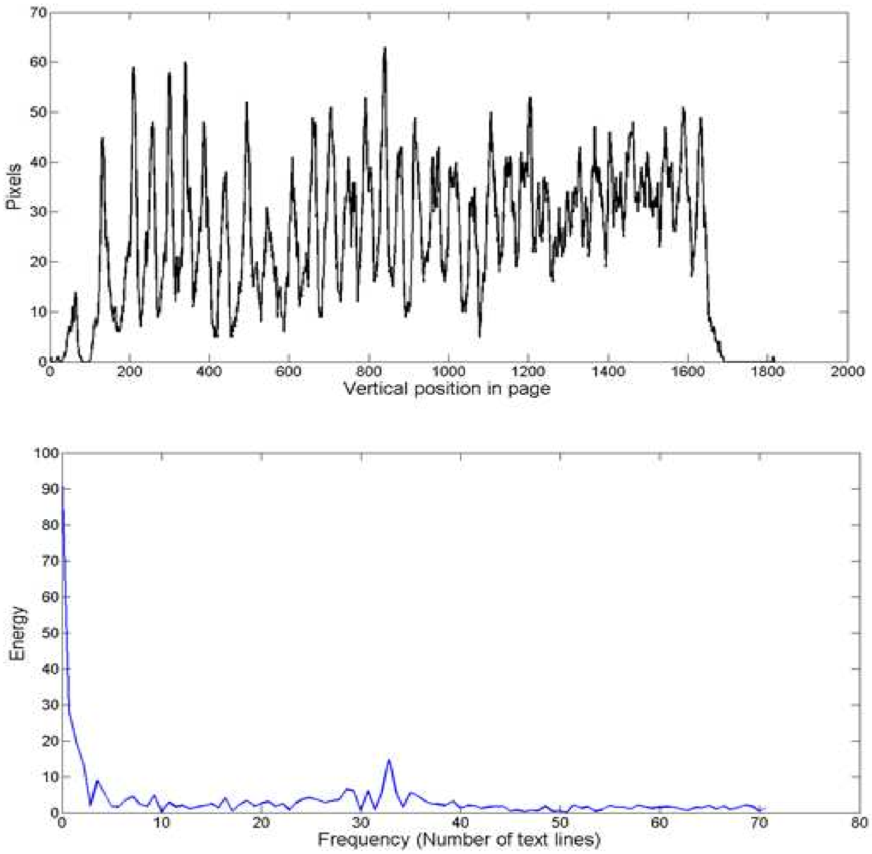
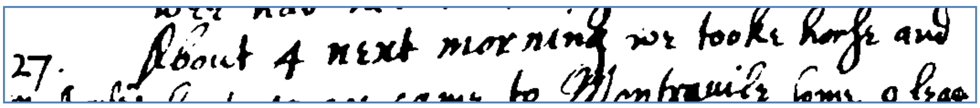
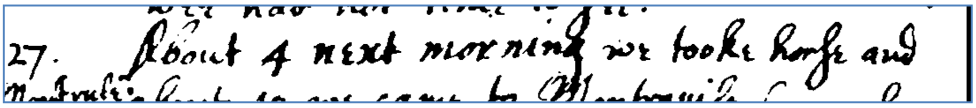


Figure 1: Horizontal projection of inky pixels



a) Initial segmentation



b) Text line is rotated to adjust the slant

c) Ink from lines above and below is erased

Figure 4: Text line segmentation

3.2.2 *Directional-chain-code*

The Freeman directional-chain-code (DCC) has been utilised in representing the contour of handwriting for the purpose of feature extraction [18, 27]. The handwriting contour is initially represented as a chain of consecutive pixels. For example, a handwritten sample of the word “Duke” is shown in Figure 5 (a), and the contour of this sample in Figure 5 (b). The chain of consecutive pixels, starting from the topmost left pixel is shown in Figure 5 (c). This sub-figure shows the letter ‘e’ of the word “Duke” from Figures 5 (a) and (b). Angles $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are defined in Section 4.2.

The contents of the DCC indicate the direction between the current pixel and the subsequent pixel. In the original design, there are only eight possible values for the direction, because there are eight pixels in the immediate neighbourhood. In order to refine the range of directions, it is common to define the direction by n (n>1) consecutive pixels, i.e. denoting the *i*th pixel in the chain of consecutive pixels as P_i, the DCC is expressed by the angle between the vector starting from p_i and ending at p_{i+n} and the vector starting from p_i pointing to the right horizontally. When n=2, for example, the DCC at the 4^{th} pixel is corresponding to α_1 in Figure 5 (c).

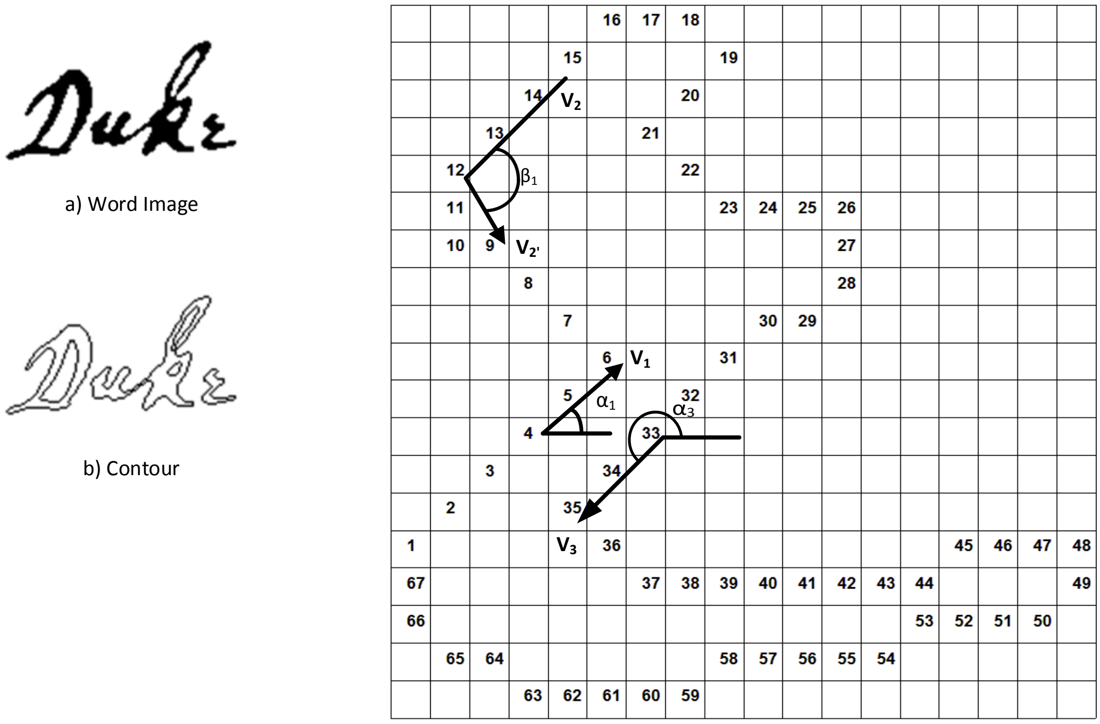
In many cases, where the categorisation of the direction is more useful than the exact value, the angle value will be replaced by a numeric code. The term “code” in DCC refers to the coding of the directions. Under the configuration n=2 there are 16 possible DCC directions, signifiying the number of perimeter pixels a distance of two pixels away from a pixel location under investigation.

It is important to note that the order of pixels in the DCC does not reflect the real writing motion because the sequence of the writing cannot be reconstructed from a ‘static’ image of the completed sample. From Figure 5 (c) taking the two vectors v_1 and v_3 as examples, the counter-clockwise directions in relation to a horizontal vector pointing from left to right are 45^o and 225^o respectively. They are, however, the left and right boundaries of the same stroke, and hence were produced at the same time. Therefore, the actual writing sequence at these two vectors must be the same. To facilitate this, pixel writing angles are all mapped to the range of 0^o to 180^o, i.e. from 0^o inclusive to 180^o exclusive. Therefore, the direction for v_3 in this case is considered as $225^o - 180^o = 45^o$. The total number of discrete codes is also reduced by half. For instance, when n=2, there are eight unique discrete codes within the range of 0^o to 180^o mapped into codes 0 to 7.

4 Handwriting Features

Having defined our datasets, in this section a number of newly proposed handwriting features will be described that aim to address aspects of traditional analysis which are not yet generally implemented in digital palaeography, together with some tools to visualise the results of feature analysis.

For each feature, two samples written by different hands are chosen to demonstrate how the feature can be visualised in a comparative view with regard to the meaning of the specific feature. As expained in Section 2.4, no single feature is universally distinctive across all scribes. Some writers may show distinctive values in one feature, while another writer might be distinctive in some other features. The purpose of illustrating our features in this section is simply to demonstrate the visualisation methods. They are not used to evaluate the overall effectiveness of any particular feature at this stage. Nevertherless, the comparative view is most useful for users to visually assess the difference between two samples in terms of a single feature.



a) Word Image

b) Contour

c) DCC example

Figure 2: Handwriting features based on directional-chain-code

4.1 Writing direction (**d**)

Definition: The histogram of writing directions is a frequency distribution of all writing directions expressed by the DCC. It is among the features described in [18], for example, for the purpose of writer identification. It is relevant to the overall slant associated with an individual writing style, i.e. the direction with the highest frequency within a writing style is likely due to the slant. Employing the same configuration as in [18], the number of successive pixels in the calculation of DCC is set to 3 (i.e. n=3), resulting in 12 discrete codes ranging from 0 to 11 within the range of 0^o to 180^o. In other words, a histogram of writing directions, denoted by *d*, is represented by a vector with 12 elements.

Visualisation: With a small range of discrete values, the distribution of writing directions is best viewed in a histogram form. Figure 6 shows this feature extracted from four text line samples, written by two different writers, identified as 'Sample 1' and 'Sample 2'. While similar patterns are observed within the samples written by the same writer, a clear distinction is also observed between the samples written by different hands.

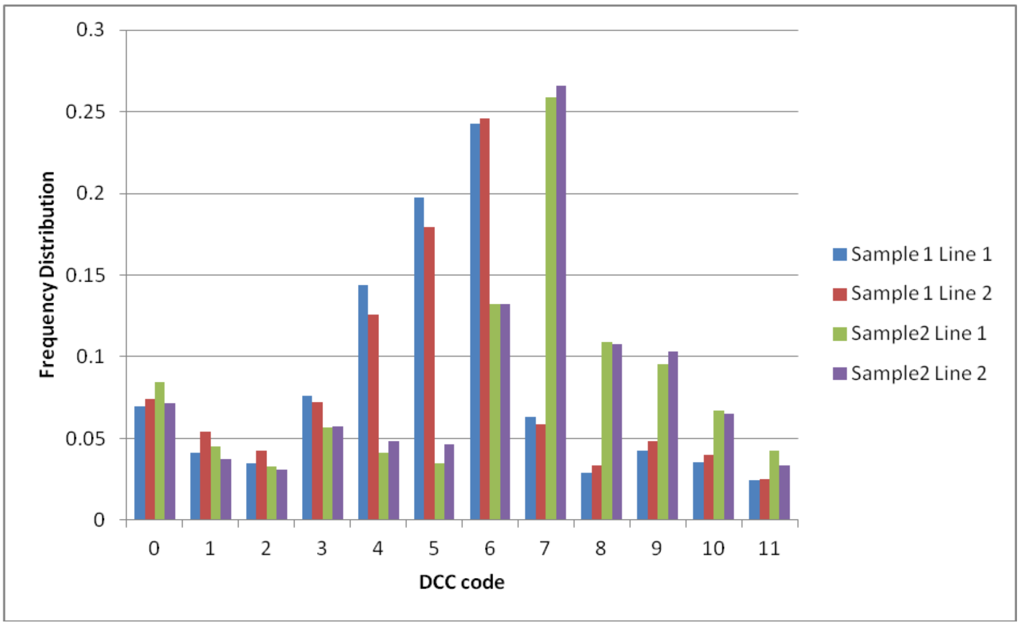


Figure 6: Writing directions

4.2 Roundness of handwriting (r)

Definition: Based on the writing directions described above, the roundness of handwriting was measured indirectly in [18] by the co-occurrence of two successive directional-chain-codes. In comparison, a more intuitive measurement is proposed for the roundness in this study; namely the angle between two successive sequences of *n* pixels (in this study *n*=3). For example, the roundness of writing at the 12th pixel in Figure 5 (c), denoted by β_1, is calculated as the angle between the two vectors starting from the 12th pixel and ending at a point three pixels away along the contour in opposite directions.

In order to distinguish between the concave and convex curves, the angles measuring the roundness of handwritings are assessed in the range of 0° to 360°. As explained in Section 4.1, when n=3, the angle values within the range of 0° to 180° are discretised to 12 unique codes, hence there are 24 discrete codes for the range of 0° to 360°. At the same time, as the roundness angle is measured by two consecutive vectors along the contour, it is impossible to result in an angle of 0°, therefore there are 23 discrete codes for the resulting roundness angles. Similar to the formulation of the writing direction histogram, the frequency distribution is implemented as a feature describing a general behaviour pattern of a single writer. Therefore, the roundness of handwriting is represented by a vector with 23 code bins.

Visualisation: The frequency values of the discretised roundness angles of handwriting are viewed in histograms, as shown in Figure 7. Though less distinctive than *d*, differences can be observed between the samples written by different hands.

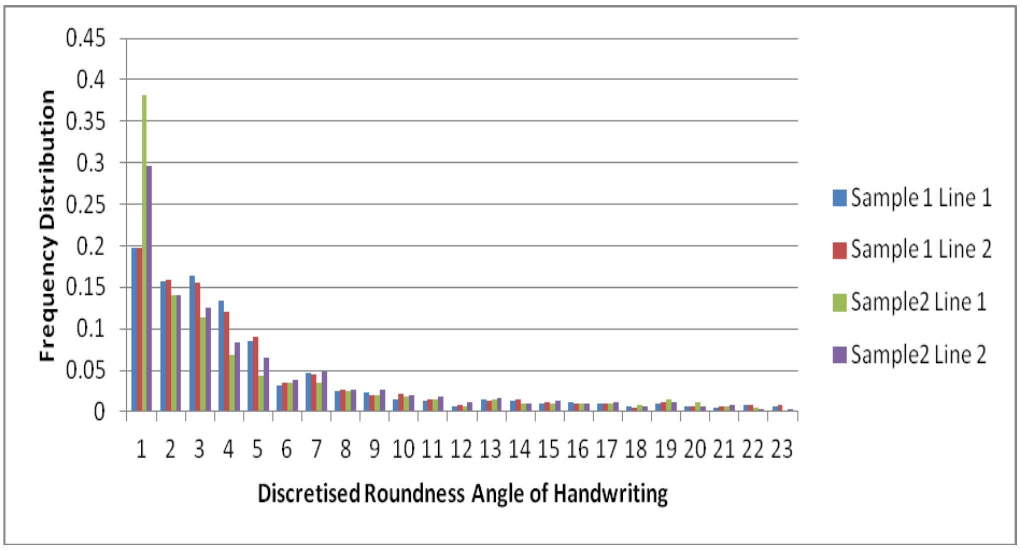


Figure 7: Roundness of handwriting

4.3 Thickness of strokes (w)

Definition: In addition to the type of pen being used, the thickness of strokes is indicative of the pen pressure [15]. Based on a binary image representation of the handwriting, the thickness of strokes can be measured by calculating the distance between the minima on the upper and lower contours in [25], or by the count of the number of foreground pixels in [15]. While the first method may be affected by degradation in the manuscript affecting ligatures or by the lift-off and pen-landing movements within a handwriting sample, the latter method is not a direct measurement of the thickness of strokes. Based on an evolution of the image dilation method adopted in [29], measuring the areas of the successive dilation as an indicator of writing quality, a new method is proposed in this study to measure the thickness of strokes.

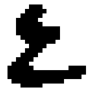
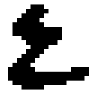
In our proposed method, the text lines are initally skeletonised. The skeletonised text lines are then subjected to a successive dilation process, where the diameter of the dilation disk increases by a minimum increment of two pixels at each iteration (i.e. the radius increases by one pixel). Taking the character "e" in the word "Duke" in Figure 5 (a) as an example, the original image along with the images at each iteration of dilation are shown in Table 2. The area within the original image that does not overlap with the dilated skeleton (variable *u* in Table 2) represents the pixels where the strokes are thicker than the diameter of the dilated skeleton. The percentage of area of a text line that is thicker than the diameter of the dilated skeleton and thiner than the diameter of the next iteration of dilation is recorded after each evolution (variable w_i in Table 2). The dilation process stops when the entirety of the original text line image overlaps with the dilated skeleton, indicating that all strokes are thinner than the diameter of the dilation disk at the current step.

Hence the thickness of strokes, denoted by *w*, is a vector of which the length is related to the thickest stroke in the sample, and the position of the elements relates to the thickness of strokes starting from one and increasing by two. The values of the elements represent the percentage of pixels where the strokes are thicker than the thickness related to the position of the element. For example, the values for *w* extracted from four text lines written by two different writers are shown in Table 3. Note that w_1 and w_2 are zero-padded to provide a comparative view with other samples.

Visualisation: A bar chart can be generated to visualise the values in Table 3 once the vectors are all adjusted to the same length. However, the small values in some elements in *w* result in very short bars that are hardly visible. Therefore, in order to maximise the visualisation effect, the values in each column in Table 3 are divided by the maximum value of that column. As a result, the bars in each group are scaled proportionally so that the longest bar reaches 100, as shown in Figure 8.

Note that the purpose of the scaling process is to maximise the visual effect, and the values in *w* remain the same for all samples.

Table 2: Calculation of *w* – the thickness of strokes

Iteration	**Original**	**t_1=1**	**t_2=3**	**t_3=5**	**t_4=7**	**t_5=9**
Image						
o [a]	145	38	117	138	142	145
u [b]	0	107	28	7	3	0
w_i [c]		(107-28)/145=54%	(28-7)/145=14%	(7-3)/145=3%	3/145=2%	0

a) Overlapping area
b) Non-overlapping area
c) Percentage of area where the thickness of stroke is greater than t_i but smaller than t_{i+1}

Table 3: Thickness of strokes

a) Values of *w* calculated from four text lines

Thickness of strokes (t)	**1<t<3**	**3<t<5**	**5<t<7**	**7<t<9**	**9<t<11**	**11<t<13**
Writer 1 sample 1 ($w^{(1)}$)	49.58	28.04	3.40	0.10	0.00	0.00
Writer 1 sample 2 ($w^{(2)}$)	46.95	30.69	4.61	0.11	0.00	0.00
Writer 2 sample 1 ($w^{(3)}$)	42.10	31.71	9.54	1.60	0.17	0.02
Writer 2 sample 2 ($w^{(4)}$)	42.20	31.99	8.91	1.69	0.24	0.02

b) Values of *w* converted column-wise as the percentage of the maximum value within column

Thickness of strokes (t)	**1<t<3**	**3<t<5**	**5<t<7**	**7<t<9**	**9<t<11**	**11<t<13**
Writer 1 sample 1 ($w^{(1)}$)	100.00	87.65	35.61	5.73	0.00	0.00
Writer 1 sample 2 ($w^{(2)}$)	94.69	95.92	48.31	6.42	0.00	0.00
Writer 2 sample 1 ($w^{(3)}$)	84.91	99.11	100.00	94.47	71.65	98.89
Writer 2 sample 2 ($w^{(4)}$)	85.11	100.00	93.36	100.00	100.00	100.00

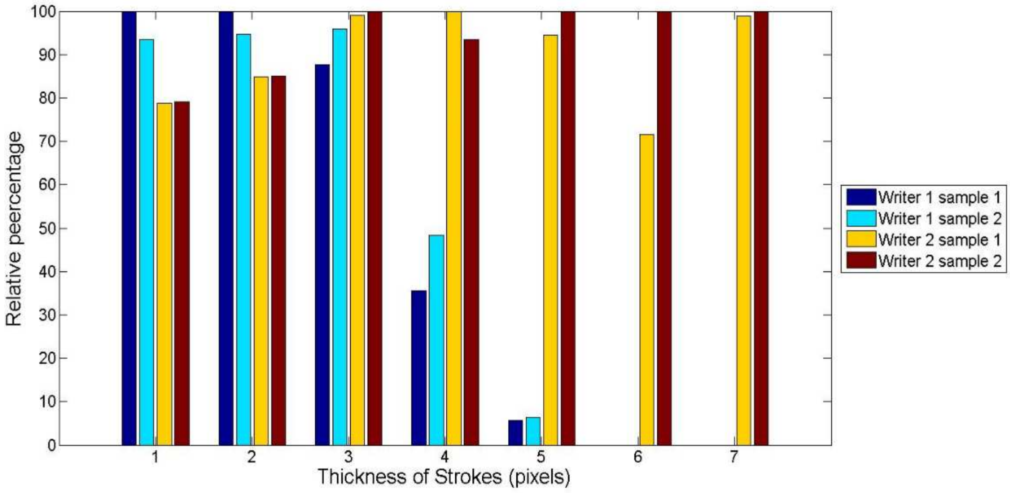


Figure 8: Thickness of strokes

4.4 Word proportions (p)

Definition: The word proportions, or proportions of letters, have been recognised as a distinctive feature of writing styles. In traditional palaeography or forensic document analysis, the proportions of individual letters are examined, i.e. the proportion of the ascender in the letter ‘b’, the descender in the letter ‘g’, etc. Since the automation of the process is a primary objective in our study, word proportions are averaged for all letters from the line proportions with regard to the upper partition (containing ascenders), main body and lower partition (containing descenders) as shown in Figure 9.

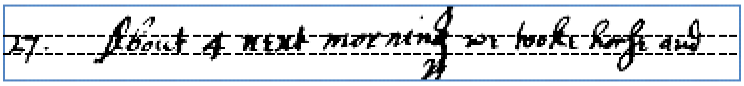


Figure 9: Separation of three line partitions: upper, main body, and lower partitions

The local extrema of the upper and lower contours of a text line are first of all extracted. Using a Fuzzy Clustering Metric (FCM) [30] algorithm, the local extrema are classified into three groups corresponding to the three partitions identified above. In order to be able to detect missing partitions, typically in a short text line where there is no ascender or descender, a validity measure on the optimal number of clusters (as described in [31]) is adopted to evaluate the possibility of missing upper and/or lower partitions. If, for example, the text line contains a single word “Duke”, it is obvious that the lower partition is missing. The extrema extracted from the contour of this text line are illustrated by the dots in Figure 10. The validity measurement assesses not only the compactness of individual clusters, but also the distance between clusters, and results in the two clusters as indicated by the two circles in Figure 10.

The result of separation of the three line partitions is shown in Figure 969. The proportion of each class is calculated as the height of each partition divided by the height of the text line. Therefore, the word proportions, denoted by *p*, are calculated for each text line and contain three scalar values.

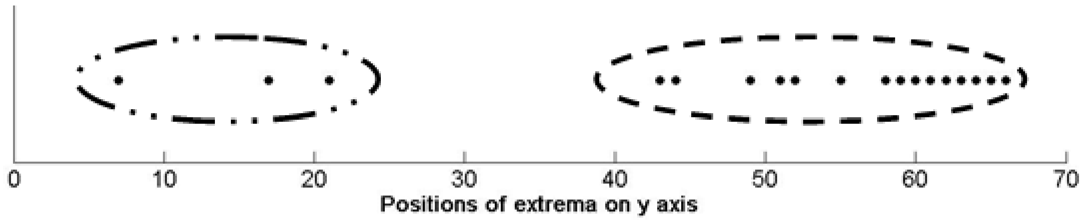


Figure 10: Clusters of extrema of a text line containing a single word “Duke”

Visualisation: Because the values of the line proportions sum to a value of one, one of the elements within *p* can be expressed linearly by the other two. Therefore, it is sufficient to visualise two of the elements contained in *p*. Viewed in a two-dimensional scatter plot, as shown in Figure 111, except for a few outliers, most of the values extracted from the samples written by the same hand form a cluster.

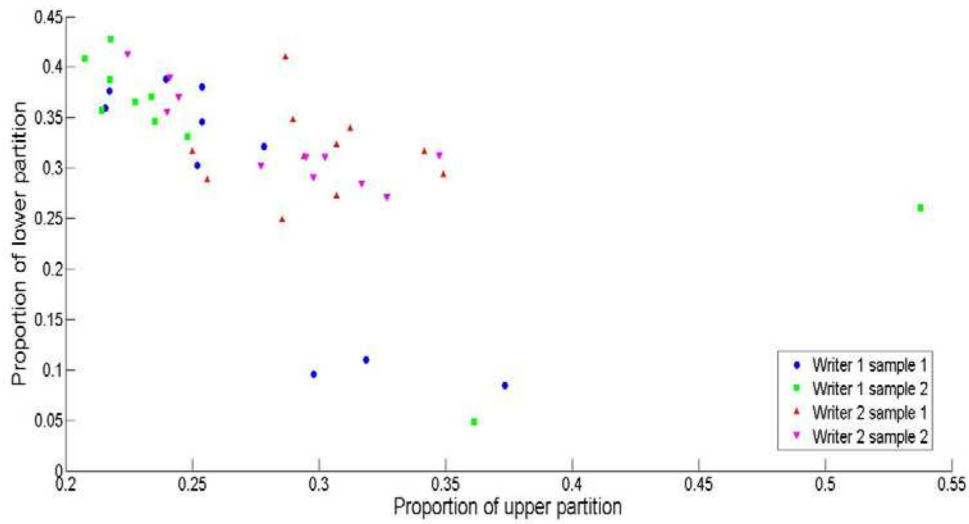


Figure 11: Word Proportions

4.5 Character and word spacing (s_c, s_w)

Definition: Without explicit word and character segmentation, character and word spacing are estimated from the horizontal distances between two ink-traces within the main body of the text line. Intuitively, a horizontal scan line is placed across a text line, and is slid vertically through the main body of the text line. Taking the text line in Figure 9 as an example, the horizontal scan line will be placed between the two dashed lines indicating the boundary of the main body of the text line. Along the scan line, the widths of all continuous white spaces along the scan line are recorded, including, for example, the white space inside the “o” in the word “About”, between “n” and “e” in the word “next”, as well as the space between words.

It is anticipated that widths shorter than average correspond to the space within individual characters or between characters within a word, and longer ones to the space between words. Therefore, the widths are taken as inputs to a *k*-means clustering algorithm (*k=2*) to distinguish between the word spacing and inter- or intra- character spacing within words. To better describe the range of values within each class, the following percentiles are chosen to express character and word spacing: 5th, 10th, 25th, 50th, 75th, 90th, and 95th. Therefore, the character and word spacing, denoted by s_c and s_w respectively, are formulated as a vector with seven elements.

Visualisation: An empirical cumulative frequency distribution function (ECDF) is suitable to describe any type of distribution function. In fact, the percentiles of a distribution (values in s_c and s_w) can be observed in an ECDF. Therefore an ECDF graph is chosen to illustrate the approximation of the distribution of word and character spacing.

Figure 12 shows the word and character spacing extracted from four samples written by two writers. The values in s_c for each sample correspond to the x-coordinates of the ECDF when x-coordinates take the value of 0.05 (the 5th percentile), 0.1 (the 10th percentile), and so on. It can be observed from Figure 12 that there is a notable distinction between samples written by different writers both in word spacing and in character spacing.

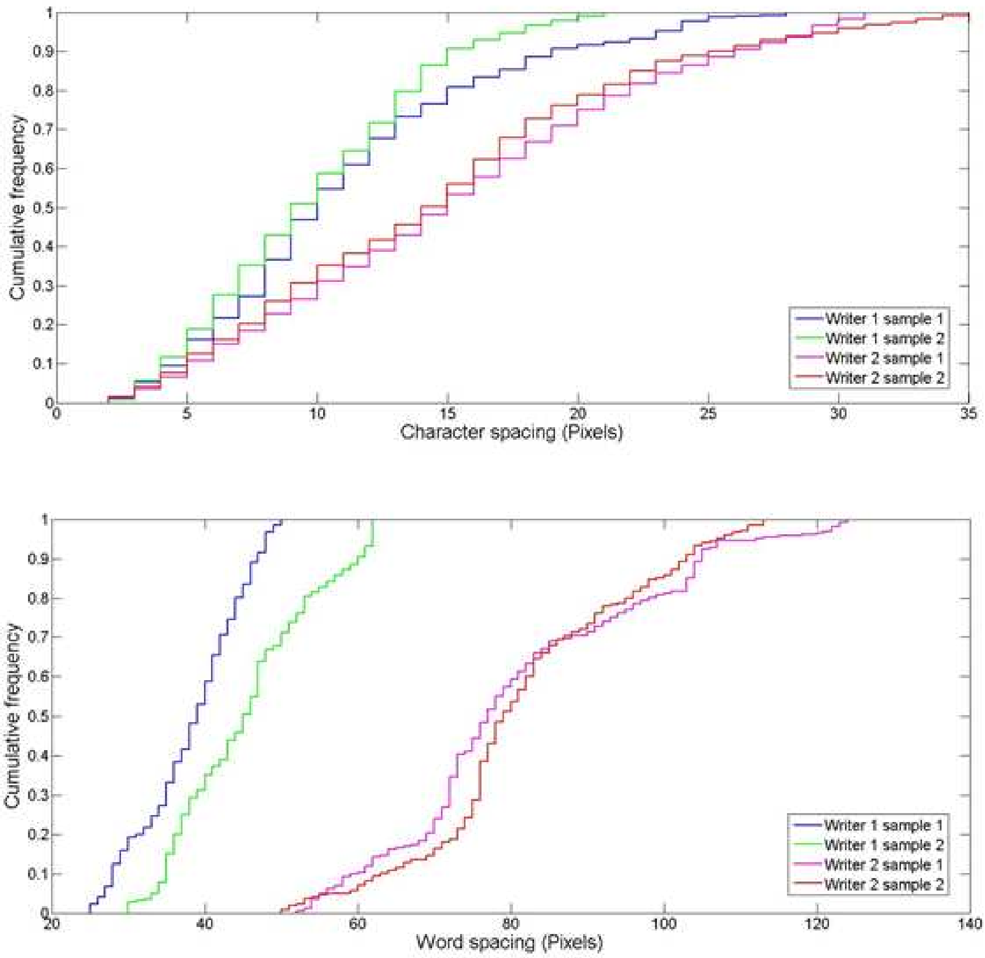


Figure 12: Character and word spacing (in pixels)

4.6 Line spacing (s_l)

Definition: A direct outcome from the line segmentation as described in Section 3.2 is the local maxima that are classified as the main body of text lines. The vertical distances between these local maxima are measured as the line space, denoted by s_l. The feature s_l is extracted from a page or a region containing multiple text lines, resulting in a vector of scalar values. The length of s_l is not consistent between different samples because it is equal to the total number of text lines minus one.

Visualisation: An ECDF is suitable to visualise the similarities in line spacing between samples. However, when there are only a few lines available for the analysis, the ECDF shows significant step effects. Assuming that there is some consistency in the line spacing within each writing style, the values of s_l should follow a normal distribution. Therefore, a theoretical cumulative distribution function (TCDF) can be estimated based on the parameters of a normal distribution calculated from the values of s_l, as shown in Figure 13.

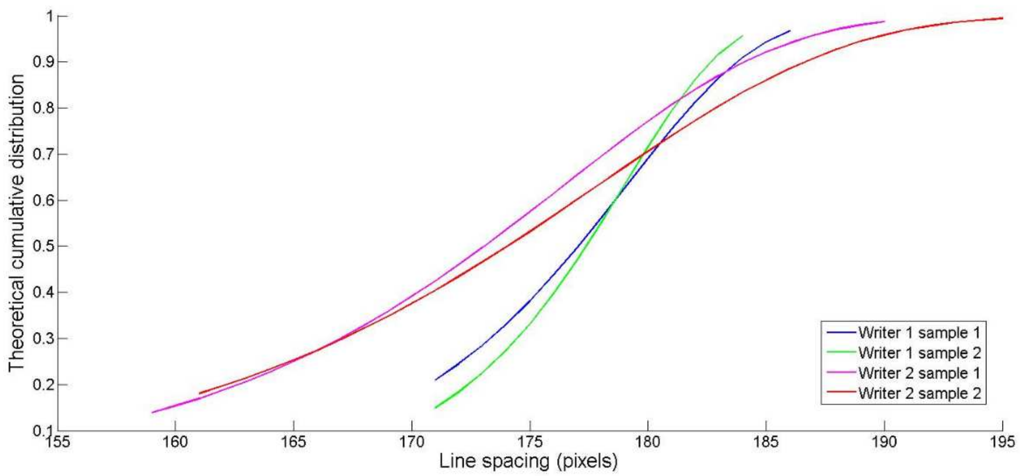
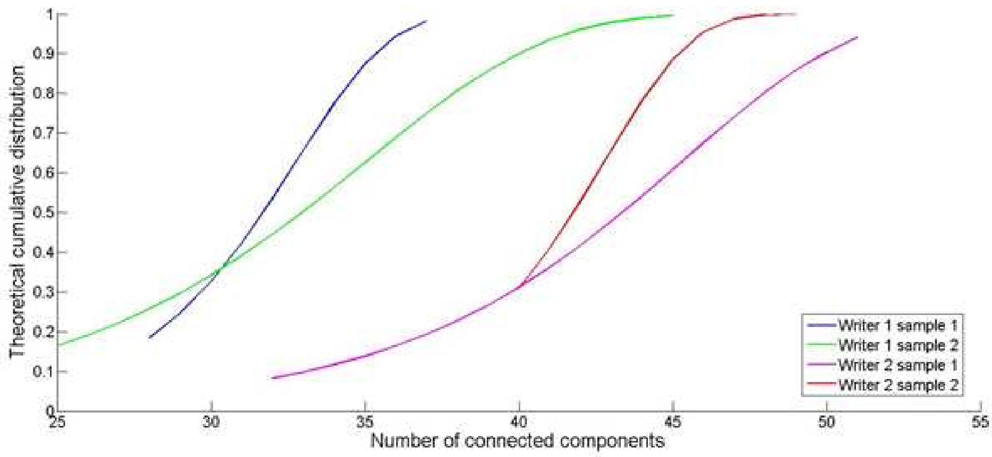


Figure 13: Line spacing

4.7 Connectivity (c, o)

Definition: The average number of interior and exterior contours [15] are indications of connectivity in cursive handwriting samples. In this work, the connectivity is expressed by two features; 1) the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 95^{th} percentiles of the width of connected components in a text line (denoted by c), and 2) the ratio of the number of loops to the number of connected components in a text line (denoted by o). The possible range of values for c is unknown and the values for o vary within a continuous range between 0 and 1.

Visualisation: Similar to the method of visualising the line spacing, in order to avoid the step effect in an ECDF graph, a TCDF is calculated using the parameters of a normal distribution estimated from the sampled values. As can be observed from Figure 14, the samples written by the same hand show more similarities than those written by a different hand.



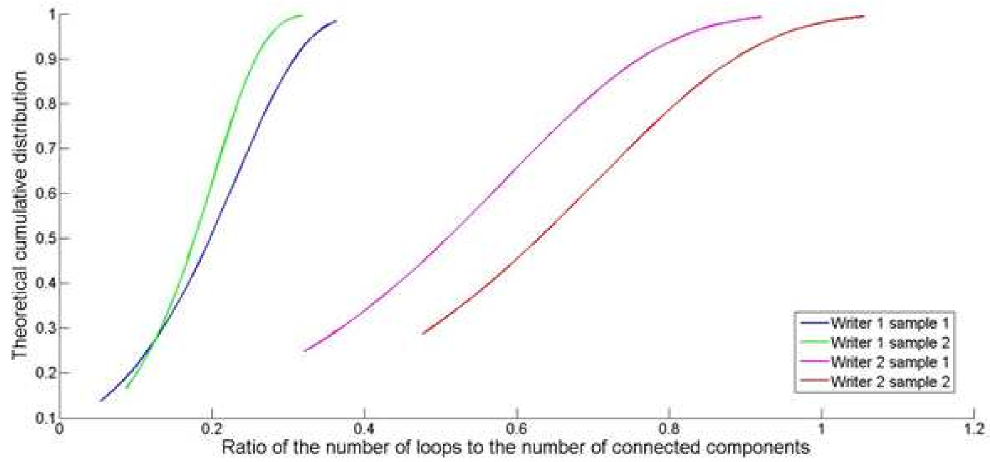


Figure 14: Connectivity

5 Distinctiveness of Handwriting Features

As noted earlier, the distinctiveness of the handwriting features will be evaluated in the context of scribal hand verification, and a new paradigm of writer verification system that does not require a training process will be introduced in this section.

5.1 Methodology

The features introduced in Section 4 are extracted from individual text lines, except line spacing (s_l) which is extracted from two adjacent text lines. Therefore, given a page containing multiple text lines, a small set of samples are available for each feature. The size of the feature set is *n-1* for line spacing and *n* for other features, where *n* is the number of text lines in the manuscript image(s).

A probability distribution function (PDF) can be estimated from the set of samples available for a single feature. It is assumed that if two sets of features are taken from handwriting samples written by the same writer, there is no significant difference between their PDFs. In other words, if two sets of feature samples show significant difference in their PDFs, the handwriting samples from which they are taken are likely to have been written by different writers.

A simple Kolmogorov-Smirnov (KS) test [32] can be employed to test the equality of one-dimensional PDFs, which applies to features expressed by scalar values. However, except for the two features representing the connectivity of a text line, which are expressed by scalar values, the features are all represented by vectors. In statistics, adapting the KS test to assess the equality of multidimensional PDFs is still a challenging task [33]. Several methods have been proposed, but they are computationally expensive. For a set of *n* samples, the computational complexity is at least at the order of n^2 [33]. Therefore, a method is proposed to convert the multidimensional feature space into a one-dimensional representation.

A number of techniques are available for the purpose of projecting a high dimensional dataset to a low dimensional space [34]. Several methods, including Isomap, Metric Multidimensional Scaling, and Maximum Variance Unfolding share the same underlying principle: preserving the pair-wise distances of the data samples [34]. The same principle is adopted in this study to reduce the multidimensional feature set into a one-dimensional space.

Following scaling the range of values within each dimension to between zero and one, a Euclidean distance metric is used to calculate the distance between a pair of feature samples. Regardless of the number of dimensions of the feature, for a set of *n* feature samples, the result of dimensionality reduction is a set of C_n^2 scalar values, and hence the KS-test is suitable for testing the equality of the PDFs of the dimensionally reduced feature sets.

When comparing the similarities between two sets of vector-based feature samples, not only the pair-wise distances within each feature set are calculated but also those cross feature sets. Let us denote the two feature sets as $s_1=\{f_{11}\ldots f_{1m}\}$ and $s_2=\{f_{21}\ldots f_{2n}\}$, where *m* and *n* represent the number of feature samples in each set respectively. The result of the dimensionality reduction conducted on s_1 and s_2 can be expressed by three sets of scalar values d_{11}, d_{22}, and d_{12}, where d_{ij} represents the pair-wise distances between the features within sets *i* and *j*.

The hypothesis on the equality of the PDFs of these two feature sets can now be expressed as follows: if s_1 and s_2 are taken from handwriting samples written by the same writer, there is no significant difference between the PDFs of d_{11} and d_{12} or between d_{22} and d_{12}. The overall hypothesis can be decomposed to two sub-hypotheses, both tested by the KS-test. The overall hypothesis is accepted only when both sub-hypotheses are accepted, and rejected when any one of the sub-hypotheses is rejected.

5.2 Experiment and performance evaluation

In the context of this study, each manuscript page constitutes a handwriting sample. For each individual feature, the writer verification method as described in Section 5.1 is employed to evaluate the equality of the sets of values extracted from any two samples. Therefore, for the handwriting samples within the MEMDB dataset, the total number of trials is 276. Future studies may be carried out to explore the relationship between the performance of the method using a smaller sample size. The current experiment is carried out using each dataset separately. Moreover, in order to evaluate the scalability of the method on assessing the distinctiveness of the handwriting features, the experiment is repeated on the subset of the IAMDB datasset. Also, for the purpose of providing a comparison of the performance between the two datasets with respect to problem scale, the size of the subset of IAMDB dataset is formed to contain the same number of writers as in MEMDB datasset. The subset is hence named as IAMDB-5 dataset, and the full dataset is referred to as IAMDB-200 dataset.

Similar to the situation with a conventional writer verification system, the acceptance rate (FAR) and the false rejection rate (FRR) can be adopted as the performance evaluation metrics, as shown in Eq. (1) and Eq. (2) respectively. In statistical trials, FRR is equivalent to type I error rate, and FAR to type II error rate. The evaluation relates to the hypothesis that two samples are written by the same writer. Type I error or FRR increases if the hypothesis is true but is rejected incorrectly. Similarly, type II error or FAR increases if the hypothesis is false but accepted incorrectly.

$$\text{FAR} = \text{number of times the hypothesis is false and accepted} \; / \; \text{number of trials when the hypothesis is false} \quad (1)$$

$$\text{FRR} = \text{number of times the hypothesis is true and rejected} \; / \; \text{number of trials when the hypothesis is true} \quad (2)$$

In relation to the FAR and FRR, two metrics giving a more intuitive interpretation of the feature performance are proposed in this paper: *inter-writer distinction* (D), and *intra-writer repeatability* (R). The inter-writer distinction measures how distinctive a feature is between handwriting samples produced by different writers, and it increases when the hypothesis is false and is rejected correctly. Conversely, the intra-writer repeatability assesses whether a feature is repeatable within the handwriting samples produced by a single writer, and the value increases when the hypothesis is accepted correctly. From the description above, the relations between FAR, FRR and D and R can be expressed as follows:

$$D=1\text{-}FAR \tag{3}$$

$$R=1\text{-}FRR \tag{4}$$

A third metric – the verification rate (V) – is introduced to assess the overall performance by averaging D and R. V can be expressed by D and R or by FAR and FRR as follows:

$$\text{V}= ((\text{D}+\text{R})/2) = 1 - ((\text{FAR}+\text{FRR})/2) \tag{5}$$

5.3 Results and discussion

The results obtained from the experiments on both datasets are shown in Table 4. The first observation is in relation to the first two features, which were introduced in [18]. The α in this study is a direct implementation of the same feature in [18], and the extraction of r is based on the writing directions of two adjacent sequence of pixels. The reported performance of α in writer verification in [18] based on the IAMDB dataset is an *EER* (equal error rate) of 7.1% or 92.9% in terms of correct verification rate. In our study, when tested on IAMDB-200, V is 89%. The two features d and r show the best performance within the modern dataset, and therefore confirming the findings in [18].

The second observation from the obtained results in Table 4 is with regard to the scalability of the method employed in scribal hand verification in our study. In comparison with the test on the subset of IAMDB, the decline in performance is less than 2%. Therefore, it can be concluded that, first of all, the performance of the features are consistent when the size of the dataset scales up.

Looking at the performance of individual features when tested using the full IAMDB dataset, we find that seven features: d, r, w, p, s_w, c, and o, show good performance (D, R, V>50%) while the other two features: s_c and s_l, show poor performance. The same phenomenon is observed when tested using the subset of IAMDB, i.e. IAMDB-5.

On the other hand, when reviewing the results obtained using the MEMDB dataset, a general decline in the performance can be observed in most of the features. However, a more interesting finding is that the two features associated with the worst performance in IAMDB: s_c and s_l are among the top three features showing the best performance in MEMDB. Also, the two features showing the best performance when tested on IAMDB, d and r, are among the less effective features when tested on IAMDB.

The last observation suggests that the handwriting features similar to those adopted in automatic writer identification/verification system described in [18], namely d and r, are less beneficial when applied to historical manuscripts. In comparison, the features proposed in this study are closer to the metrics adopted in conventional palaeography analysis. Therefore, the observation further indicates the importance of learning from the established metrics in the analysis of historical manuscripts.

Table 4: Results of writer verification

Feature	IAMDB-5			IAMDB-200			MEMDB		
	D	R	V	D	R	V	D	R	V
d	98.3%	81.8%	90.0%	97.8%	80.1%	89.0%	85.1%	52.1%	68.6%
r	96.9%	82.1%	89.5%	95.5%	81.5%	88.5%	94.3%	35.4%	64.9%
w	63.8%	88.8%	76.3%	60.2%	88.9%	74.6%	53.5%	95.8%	74.7%
p	64.1%	80.8%	72.5%	58.8%	84.7%	71.8%	63.6%	66.7%	65.1%
s_c	18.2%	100%	59.1%	18.0%	100%	59.0%	51.3%	97.9%	**74.6%**
s_w	67.0%	78.3%	72.6%	61.4%	82.2%	71.8%	69.7%	60.4%	65.1%
s_l	3.3%	99.8%	51.5%	3.8%	99.1%	51.4%	80.7%	77.1%	**78.9%**
c	75.7%	95.1%	85.4%	71.2%	96.7%	84.0%	77.6%	81.3%	79.4%
o	61.8%	91.1%	76.4%	61.8%	91.1%	76.4%	64.0%	64.6.0%	64.3%

6 Conclusion and Future Research

This paper has introduced a novel handwriting feature extraction, visualisation and analysis system for the purpose of digital palaeographic analysis. Initially, relevant studies in three different fields have been reviewed and discussed, including traditional palaeography analysis, digital palaeography, and automatic writer identification, leading to the opportunity for a synthesis of ideas and techniques.

Addressing the gap between traditional and digital palaeographic approaches in terms of feature extraction techniques and appropriate handwriting metrics, firstly, a fully automated process has been implemented for feature extraction and visualisation, as well as analysis. Secondly, a number of new features have been introduced that are more closely related to the established metrics typically adopted in traditional palaeography than the features commonly adopted in automatic writer identification.

The set of features have been tested on two datasets: IAMDB which is a contemporary multi-writer handwriting database and MEMDB which contains four medieval/early modern manuscripts written by an unknown number of writers. The results of our experiments have confirmed the effectiveness of the features adopted in automatic writer identification, and at the same time showed that the new features are more revealing of the individualities of handwriting in historical manuscripts.

Compared to previous work in digital palaeography where manual feature extraction is required or where the goal is specific to the task of automatic writer identification, the advantage of the features introduced in this study is that they can best be exploited in the implementation of a content-based image retrieval (CBIR) [35] system in the context of digital archives. Future research building upon this current study can be developed to include an investigation into the image clustering problem for the purpose of accelerating a CBIR system. This may use the features introduced in this study as well as those adopted in previous work in automatic writer identification. Furthermore testing on additional datasets, the tuning of algorithms to sub-classes of documents and the multi-feature fusion of decision-making processess may lead to improved performance.

Acknowledgement

The authors gratefully acknowledge the EU INTERREG IVA France (Channel) – England Programme and the Canterbury Cathedral Archives for their support of this work. They also thank Dr Catherine Richardson for information provided on the documents comprising the MEMDB database.

References

[1] V. Cappellini, H. Maitre, I. Pitas, and A. Piva, "Guest Editorial Special Issue on Image Processing for Cultural Heritage," *IEEE TIP*, vol. 13, 2004.

[2] A. R. Rumble, "Palaeography, Scribal Identification and the Study of Manuscript Characteristics," in *Care and conservation of manuscripts 8: proceedings of the eighth international seminar held at the University of Copenhagen 16th-17th October 2003*. vol. 8, G. F. Jensen and P. Springborg, Eds., ed Copenhagen: Museum Tusculanum, 2005, pp. 217-228.

[3] A. Ciula, "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis," *Digital Medievalist*, vol. 1, 2005.

[4] P. Stokes, "Computer-aided palaeography, present and future," *Codicology and Palaeography in the Digital Age*, pp. 309-338, 2009.

[5] M. Aussems, "Christine de Pizan and the scribal fingerprint: a quantitative approach to manuscript studies," Utrecht University, 2006.

[6] M. P. Brown, *The Book of Cerne: prayer, patronage, and power in ninth-century England* vol. 1: University of Toronto Press, 1996.

[7] A. Ciula, "The palaeographical method under the light of a digital approach," *Codicology and Palaeography in the Digital Age*, pp. 219-237, 2009.

[8] M. Aussems, "Christine de Pizan and the scribal fingerprint: a quantitative approach to manuscript studies," 2006.

[9] L. Michel, *Gerichtliche Schriftvergleichung: eine Einführung in Grundlagen, Methoden und Praxis*. Berlin/New York: de Gruyter, 1982.

[10] J. W. J. Burgers, *De paleografie van de documentaire bronnen in Holland en Zeeland in de dertiende eeuw*. Louvain: Peeters, 1995.

[11] I. Moalla, F. Lebourgeois, H. Emptoz, and A. Alimi, "Image analysis for palaeography inspection," in *DIAL*, 2006.

[12] C. De Stefano, F. Fontanella, M. Maniaci, and A. Scotto di Freca, "A method for scribe distinction in medieval manuscripts using page layout features," *Image Analysis and Processing–ICIAP 2011*, pp. 393-402, 2011.

[13] S. Quirke, "Agendas for Digital Palaeography in an Archaeological Context: Egypt 1800 BC," in Franz Fischer et al.,Codicology and Palaeography in the Digital Age, BoD Norderstedt 2010, pp.279-294.

[14] P. A. Stokes, "Palaeography and Image-Processing: Some Solutions and Problems," *Digital Medievalist*, vol. 3, 2007.

[15] S. N. Srihari, S. H. Cha, H. Arora, and S. Lee, "Individuality of handwriting," *Journal of Forensic Sciences*, vol. 47, pp. 856-872, 2002.

[16] S. Srihari, G. R. Ball, and U. Commons, "Comparison of statistical models for writer verification," IS&T/SPIE Electronic Imaging, 2009, p. 72470E.

[17] S. N. Srihari and G. Leedham, "A survey of computer methods in forensic document examination," In Proc 11th Conference of the International Graphanomics Society, 2003, p. 279.

[18] M. Bulacu and L. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, pp. 701-717, 2007.

[19] M. Bulacu and L. Schomaker, "Automatic handwriting identification on medieval documents," in *ICIAP*, 2007, pp. 279-284.

[20] A. Brink, M. Bulacu, and L. Schomaker, "How much handwritten text is needed for text-independent writer verification and identification," 2008, pp. 1-4.

[21] A. Bensefia, T. Paquet, and L. Heutte, "A writer identification and verification system," *Pattern Recognition Letters*, vol. 26, pp. 2080-2092, 2005.

[22] A. Schlapbach and H. Bunke, "A writer identification and verification system using HMM based recognizers," *Pattern Analysis & Applications*, vol. 10, pp. 33-43, 2007.

[23] A. Brink, J. Smit, M. Bulacu, and L. Schomaker, "Writer identification using directional ink-trace width measurements," *Pattern Recognition*, 2012.

[24] M. L. Bulacu, *Statistical pattern recognition for automatic writer identification and verification*: PhD thesis, University of Groningen, 2007.

[25] L. Schomaker, K. Franke, and M. Bulacu, "Using codebooks of fragmented connected-component contours in forensic and historic writer identification," *Pattern Recognition Letters*, vol. 28, 2007

[26] U. V. Marti and H. Bunke, "The IAM-database: an English sentence database for offline handwriting recognition," *International Journal on Document Analysis and Recognition,* vol. 5, pp. 39-46, 2002.

[27] I. Siddiqi and N. Vincent, "A set of chain code based features for writer recognition," in *ICDAR*, Barcelona, Spain, 2009, pp. 981-985.

[28] K. R. Castleman, *Digital Image Processing*: Englewood Cliffs, N.J.: Prentice Hall, 1996.

[29] N. Vincent, V. Bouletreau, H. Emptoz, and R. Sabourin, "How to use fractal dimensions to qualify writings and writers," *FRACTALS-LONDON-,* vol. 8, pp. 85-98, 2000.

[30] J. C. Bezdek and R. Ehrlich, "FCM: The fuzzy c-means clustering algorithm," *Computers & Geosciences,* vol. 10, pp. 191-203, 1984.

[31] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 13, pp. 841-847, 1991.

[32] J. Frank and J. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association,* vol. 46, pp. 68-78, 1951.

[33] R. H. C. Lopes, I. Reid, and P. R. Hobson, "The two-dimensional Kolmogorov-Smirnov test," 2007.

[34] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee, "Spectral methods for dimensionality reduction," *Semisupervised learning,* pp. 293-308, 2006.

[35] E. Saykol, A. K. Sinop, U. Gudukbay, O. Ulusoy, and A. E. Çetin, "Content-based retrieval of historical Ottoman documents stored as textual images," *IEEE Transactions on Image Processing,* vol. 13, pp. 314-325, 2004.

Author Biography

Prof. Michael Fairhurst researches in handwriting analysis, medical image analysis and security/ biometrics. He is Editor-in-Chief of the ***IET Biometrics*** Journal, and an elected Fellow of the International Association for Pattern Recognition, recognising his contributions to the field. He has published some 350 papers in the scientific literature.

Dr. Richard Guest obtained a PhD in Electronic Engineering from the University of Kent, UK in 2000 where he is currently Senior Lecturer in the School of Engineering. Dr Guest's research interests include hand-drawn data analysis for forensic applications and security purposes, image processing and contents classification.

Dr. Yiqing Liang received her PhD degree at Kent in 2009 and worked in the China Telecom R&D Centre before coming to the UK. Her research interests are in Statistical Pattern Recognition, Image Analysis, and Medical Image Processing.

Dr. Meryem Erbilek obtained a PhD degree in Electronic Engineering from the University of Kent, UK in 2013. She is currently a Research Associate in the School of Engineering at the University of Kent. Her research interests are in pattern recognition, image analysis, handwritten signature analysis and biometrics.