Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Reliable writer identification in medieval manuscripts through page layout features: The “Avila” Bible case

C. De Stefano [a], M. Maniaci [b], F. Fontanella [a,*], A. Scotto di Freca [a]

[a] Dipartimento di Ingegneria Elettrica e dell’Informazione, Università di Cassino e del Lazio meridionale, Via G. Di Biasio 43, 03043 Cassino (FR), Italy
[b] Dipartimento di Lettere e Filosofia, Università di Cassino e del Lazio meridionale, Via Zamosch 43, 03043 Cassino (FR), Italy

ARTICLE INFO

Keywords:
Palaeography
Medieval handwritings
Feature selection
Classification
Reject option
Writer identification

ABSTRACT

In the field of manuscript studies (palaeography and codicology), a particularly interesting case is the study of highly standardized handwriting and book typologies. In such cases, the analysis of some basic layout features, mainly related to the organization of the page and to the exploitation of the available space, may be very helpful for distinguishing similar scribal hands. In this framework, we have defined a set of layout features to develop a pattern recognition system for identifying the scribes who collaborated to the transcription of a single medieval Latin book. We have also experimentally characterized the discriminative power of each considered feature and we have verified whether the selection of an appropriate subset of features for each scribe, specifically devised for distinguishing him from all the others, could allow us to achieve better results. This approach allowed us to introduce in a very simple way a reject option for rejecting unreliably classified samples, namely those not assigned to any scribe or assigned to more scribes. The experiments, performed on a large database of digital images from the so called “Avila Bible” – a giant Latin copy of the whole Bible produced during the XII century between Italy and Spain – confirmed the effectiveness of the proposed method. Finally, we made publicly available the data set extracted from the Avila Bible images.

1. Introduction

Palaeography, or the study of medieval handwritings, aims, among its main tasks, to ascertain when (and possibly where) a manuscript was written, how many people participated to the handwriting process and how they shared the work among them (Stokes, 2009). In traditional palaeographic studies this analysis is performed by human experts who are able to identify the peculiarities of a single scribe or the characteristics of a school of copyists. In this context, there has been in the last years a growing scientific interest in the use of computer-based techniques, whose aim is to provide new and more objective ways of characterizing medieval handwritings and distinguishing between scribal hands (Ciula, 2009; Gurrado, 2009). The application of such techniques, originally developed in the field of forensic analysis, gave rise to a new research field generally known as *digital palaeography*.

At a simpler level, the digital approach can be used to replace qualitative measurements with quantitative ones, to perform “traditional” observations more rapidly and systematically than in the past. In contrast to this, recently emerged approaches combine powerful pattern recognition algorithms and high-quality digital images of medieval manuscripts. These methodologies range from the automatic recognition and characterization of single words and signs, to the reduction of the *ductus*[1] to its basic profile, to the extraction of “texture” features, depending on the detection of recurrent forms on the surface of the page (Antonacopoulos and Downton, 2007).

More generally, the contributions of pattern recognition experts to the analysis of historical documents, focused either on the “local” characterization of the handwritten trace, or on the “global” observation of the written page.

The first approach is based on the analysis of individual letters and signs as well as of their composing strokes. In this context, run-length based features have been proposed in the literature to represent local binary patterns, such as the information about slant and curvature of handwritten texts, while grapheme-based ones have been exploited for extracting local structures and map them into a common space. These techniques have been widely used in document analysis applications

[1] *Ductus*: the shape, the direction and the order of the strokes used to form each letter.

* Corresponding author.
E-mail addresses: destefano@unicas.it (C. De Stefano), mmaniaci@unicas.it (M. Maniaci), fontanella@unicas.it (F. Fontanella), a.scotto@unicas.it (A. Scotto di Freca).

https://doi.org/10.1016/j.engappai.2018.03.023
Received 13 October 2016; Received in revised form 16 March 2018; Accepted 26 March 2018
Available online 12 April 2018

for both binarized and gray scale images, and have also been applied to historical documents (Dinstein and Shapira, 1982; He et al., 2016a; Schomaker et al., 2007).

In Yosef et al. (2007) scribe identification is performed by means of comparisons with a database of characters automatically extracted from a set of fourteenth to sixteenth century Hebrew manuscripts, written by 34 different scribes. More recently, novel approaches for writer identification have been devised. In Dahllof (2014), the author proposes a procedure for identifying early medieval hands based on the comparison with a set of segmented and classified letter shapes extracted from pages written by already known scribes. In Papaodysseus et al. (2014), the authors present a novel methodology to automatically identify writers of ancient inscriptions and Byzantine codices. The method initially estimates the normalized curvature of letter contours. Then a number of statistical criteria are used for the automatic identification of the writers. In Wahlberg et al. (2014) binarization was used to find the ink strokes. Then statistics on these ink strokes are used as features for writer identification. Moreover, in Sampath (2016) the author presents a novel approach for analyzing scribal behavior by using information about the handwriting of characters. The author also proposes some metrics to quantitatively evaluate the behavior of the writers. The proposed approach can potentially be used for writer identification and document dating. It should be noted however that, because of the unsatisfying results obtained by character segmentation, specially on severely degraded documents, "holistic" methods, based on word spotting (En et al., 2016; Louloudis et al., 2012; Pintus et al., 2015; Rath and Manmatha, 3–6 August, 2003) and/or retrieval techniques (Lavrenko et al., 2004; Liang et al., 2012; Wei and Gao, 2014), have attracted increasing interest and have been tested on documents of different periods and origins, from the Middle Ages to modern times.

The second approach focuses on the global, automated observation of the handwritten page by using texture features and/or layout analysis. In Bulacu and Schomaker (2007) the authors underline the limit of the local approach, due to the difficulty of applying the segmentation at the level of individual characters (especially when using text-independent methods), and shift the focus on allograph and texture level. In Joutel et al. (2007) a segmentation free approach based on curvlets features is proposed for revealing morphological properties of handwriting such as curvature and orientation. Such a texture-level information allows the authors to distinguish the scribes collaborating to the transcription of two samples of Middle Age and eighteenth century manuscripts. Another segmentation free approach is presented in Al-Aziz et al. (2011), where authors apply a Spatial Gray Level Dependence (SGLD) technique to analyze Arabic manuscripts of different ages. Texture-level information and SGLD method have been also used in Moalla et al. (27–28 April, 2006) for studying old Latin writings of the eight–sixteenth centuries. More recently, novel approaches have been devised both for writer identification (Dhali et al., 2017; Liang et al., 2016) and document dating (He et al., 2016b; Wahlberg et al., 2015). In Dhali et al. (2017), Dhali et al. present a preliminary study for the identification of the writers of the dead sea scrolls. As features, they adopt texture-based statistical information about slant and curvature of the handwritten characters. Once the features are extracted, writers are classified by means of the nearest neighbor approach. In Liang et al. (2016) authors presents a fully automated handwriting feature extraction, visualization and analysis system, whose aim is to design and test script and layout features more closely related to conventional palaeographic metrics than those commonly adopted in automatic scribe identification. In Wahlberg et al. (2015), the authors adopt shape statistics for manuscript dating. The proposed strategy use stroke width transform and a statistical model of the gradient image to find ink boundaries. Then for each manuscript, a distribution over common shapes is produced. Finally, in He et al. (2016b) the authors introduce a family of features extracted from contour fragments and stroke fragments. Then, for each page, the statistical distributions of these fragments are used for capturing the handwriting style.

However promising, all these approaches have not yet produced results widely accepted by palaeographers, because of both the immaturity in the use of these new technologies, and the lack of real interdisciplinary research: manuscript historians often missing a proper understanding of rather complex image analysis procedures, and scientists being unaware of the specificity of medieval writing and tending to extrapolate software and methods already developed for modern writings (Conti et al., 2015). A further difficulty derives from the fact that not all the approaches are applicable to manuscripts of any historical period, because of the specific problems originating from the heterogeneous nature of handwritings of different ages, languages and styles. The main challenge, however, is represented by the application of convincing models of digital representation and analysis to the characteristics of medieval handwriting, given its extreme variability and the difficulty of taking account of the gestures from which it originated. Thus, *digital palaeography* is increasingly used and the research activity in this field ought to be further developed (Stokes, 2015). In this framework, a particularly interesting case is the study of highly standardized handwriting and book typologies, for which the analysis of some basic layout features, regarding the organization of the page and its exploitation by the scribe, may give precious information for distinguishing very similar hands even without recourse to palaeographical analysis. This kind of features are more easily and finely extracted and quantified by using standard image processing algorithms and, therefore, could be very helpful for implementing automatic classification systems.

Moving from these considerations, in previous studies (De Stefano et al., 2011b; De Stefano et al., 2011a), we proposed a pattern recognition system for distinguishing the different scribes who worked together to the transcription of a single medieval Latin book. In these preliminary works we used a specifically devised set of features, directly derived from the analysis of page layout according to the suggestions of palaeographic and codicological researchers, and performed classification by using a standard Multi Layer Perceptron (MLP) neural network. Even if the results were interesting, the experiments highlighted two main problems. On the one hand, the number of samples collected for the different scribes was considerably different, due to a very uneven distribution of the transcription task between them. This is a frequent problem to deal with in palaeographic analysis, since it is very unlikely that all the scribes who collaborated to the production of a single manuscript wrote in the average the same amount of text. The effect is that such unbalanced data distribution bias the "learning" procedure, better classifying the scribes whose samples are more frequent in the training set. On the other hand, there were cases in which even the scribes adequately represented in the training set were not effectively classified. This suggests that the considered set of features may not have enough discriminative power for distinguishing scribes writing in very similar ways.

In the present contribution, we propose a new classification system which tries to solve the main drawbacks previously discussed. As regards the first problem, we performed a large experimental analysis for selecting a better classification scheme. As it will be shown in the experimental results, Decision Tree (DT) classification method demonstrated to be particularly effective in managing the complexity of the problem at hand and the unbalanced distribution of data among the classes. Thus we used DTs for implementing our system. As regards the second problem, we performed an experimental investigation for verifying the discriminant power of the proposed features. To this aim we considered a set of *univariate* measures and combined their results by using a Borda count based technique.

We also aimed at verifying whether the selection of an appropriate subset of features for each scribe, specifically devised for distinguishing a single scribe from all the others, could allow us to achieve more satisfactory results. Following this idea, we implemented two different classification schemes. The first one is a single classifier using all the available features. The second one was obtained by decomposing the original classification problem (recognition of the parts of text written

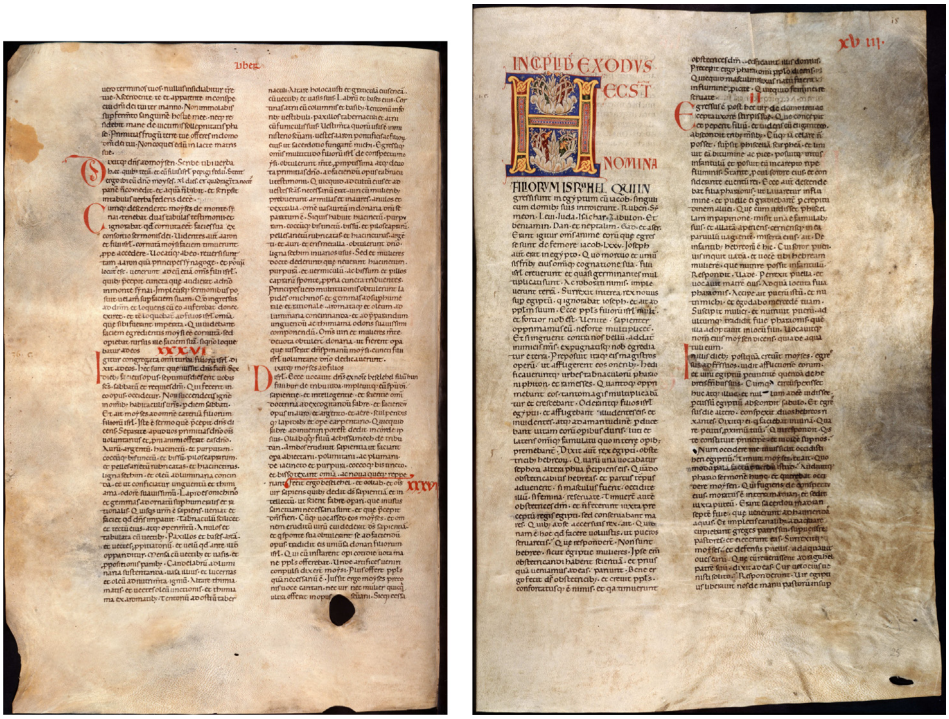

Fig. 1. Two examples of pages from the Avila Bible.

by each of N considered scribes) into N two-class problems, each consisting in discriminating the text produced by a scribe from the text produced by all the others. For each two-class problem, a feature selection phase is performed in order to identify the subset of features able to characterize the peculiarities of that particular scribe. In this way, features having poor discriminative power, or even misleading ones, are automatically discarded. Finally, the proposed approach allowed us to easily implement a reject option: samples not assigned to any scribe, or sample assigned to more scribes are rejected. It is worth noticing that the analysis of the layout features selected for distinguishing each scribe from all the others may be very interesting for palaeographers and codicologists, since it allows them to characterize the peculiarities of each scribe's hand and his attitude towards the distribution of writing onto the page, with respect to the rules and the standards imposed by the writing tradition or school he belonged to.

Summarizing, with respect to our previous studies reported in De Stefano et al. (2011b) and De Stefano et al. (2011a), we performed the following improvements:

- a deeper analysis of the discriminant power of the proposed features, using a set of *univariate* measures and combining their results through a Borda count based technique;
- the design of an high performing classification scheme based on the Decision Tree (DT) method, which demonstrated to be much more effective than the BP classifier used in the previous studies;
- the design of a new classification architecture obtained by decomposing the original classification problem in as many sub-problems as the number of scribes to be distinguished;
- the definition of a feature selection method to be used in each sub-problem. According to this approach, an appropriate subset of features may be selected for each scribe able to capture the his distinctive aspects, so as to better identify his hand from those of all the others;
- the definition of an effective reject option, able to discard unreliably classified patterns with a very slight reduction of the correct recognition rate.

A particularly favorable situation to test the effectiveness of our approach is represented by the so-called "Giant Bibles", a hundred or more of serially produced Latin manuscripts each containing the whole sacred text in a single volume of very large size (up to 600 × 400 mm and over). The Bibles originated in Central Italy (initially in Rome) in the mid-11th century, very likely as part of the political program of the "Gregorian Reform" (Maniaci and Orofino, 2000). Very similar in shape, material features, decoration and script, the Bibles were produced by groups of several scribes, organizing their common work according to criteria which still have to be deeply understood. The distinction among their hands often requires very long and patient palaeographical comparisons. In this context, we have used for our experiments the specimen known as "Avila Bible" (Madrid, Biblioteca nacional, ms. Vitr. 15.1), which was written in Italy by at least nine scribes within the third decade of the 12th century and soon sent (for unknown reasons) to Spain, where its text and decoration were completed by local scribes; in a third phase (during the 15th century) additions were made by another copyist, in order to adapt the textual sequence to new liturgical needs (Maniaci and Orofino, 2012). The Bible offers an "anthology" of contemporary and not contemporary scribal hands, thus representing a severe test for evaluating the effectiveness and the potentialities of our approach to the distinction of the scribes. To the best of our knowledge, this is the first study in which digital palaeography techniques have been applied to Romanesque Bibles, and particularly to the "Avila Bible". Examples of pages from the Avila Bible are shown in Fig. 1. The remainder of the paper is organized as follows: Section 2 presents the architecture of the system, Section 3 details the proposed features, Section 4 illustrates the methods used for feature analysis, while Section 5 describes the two proposed classification architectures. The experimental results are presented and discussed in Section 6. Finally, Section 7 is devoted to some discussions and conclusions.

2. The pattern recognition system

The architecture of the proposed system is illustrated in Fig. 2. The system receives as input RGB images of single pages belonging to the manuscript to be processed, and performs for each page the

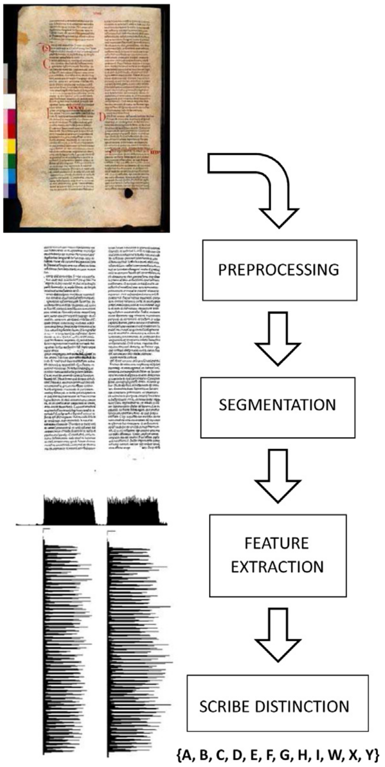

Fig. 2. The proposed pattern recognition system for scribes distinction.

following steps: *pre-processing*, *segmentation*, *feature extraction*, and *scribe distinction*.

The appearance of input RGB images may be damaged by holes or stains on the parchment. In the pages several red out-scaling capital letters may be used by the copyists for starting new paragraphs.

In the pre-processing step noisy pixels, such as those corresponding to stains or holes on the page, or those relating to a miniature, are detected and removed. Red out-scaling capital letters are also removed since they are often written by a single scribe, specialized for this task, who may be different from those responsible for the transcription of the text. Finally, RGB images are transformed into gray level ones and then in binary black and white images. This last step has been performed by choosing a fixed threshold value, equal to 135, heuristically determined and used for all the images.

In the segmentation step, columns and rows in each page are detected. As anticipated in the Introduction, this study focuses on the "Avila Bible" written in a two-column format where each column contains a slightly variable number of rows (i.e. text lines), approximately equal to sixty. The detection of both columns and rows is performed by computing pixel projection histograms on the horizontal and the vertical axis, respectively.

The feature extraction step has been developed following the suggestion reported in Stokes (2009). A detailed description of the feature extraction process is reported in Section 3.

Finally, the scribe distinction step consists in identifying text sections written by each of the copyists that collaborated to the transcription of the manuscript. Further details about this process are given in Section 5.

3. The extracted features

The set of features designed for scribe distinction can be divided into three groups, mainly concerning the layout of the page or the interaction between the handwriting and the available space on the page:

- the first set relates to geometrical properties of the whole page;
- the second group concerns the exploitation of each column of the written area;
- the third group characterizes the way the scribe distributed the text in each row, according to the ruling.

The first set of features includes the *upper margin* of the page, the *lower margin* and the *intercolumnar distance*. Such features are not very distinctive for an individual copyist, but may be very useful to highlight chronological and/or typological differences;

In the second set of features, we have considered the *number of rows* in the column and the column *exploitation coefficient* (Bozzolo et al., 1982). The exploitation coefficient is a measure of how much the column is filled with ink, and it is computed as:

$$exploitation\ coefficient = \frac{N_{BP}(C)}{N_P(C)} \tag{1}$$

where the functions $N_{BP}(C)$ and $N_P(C)$ return the number of black pixels and the total number of pixels in the column C, respectively. Both features vary according to different factors, among which the expertise of the scribe. In the case of highly standardized handwriting, such as the "carolingian minuscule" used in the "Avila Bible", such features may be considered as a measure of the writer's skill and may be very helpful for scribe distinction.

As regards the third set of features, we have analyzed the rows in each column, considering the following features: *weight*, *peaks*, *modular ratio*, *interlinear spacing* and *modular ratio/interlinear spacing ratio*.

The weight measures how much a row is filled with ink. Thus, its definition is very similar to that of Eq. (1), but considers row pixels instead of column pixels:

$$weight = \frac{N_{BP}(R)}{N_P(R)} \tag{2}$$

This feature may be considered as a qualitative measure of thickness of the handwritten trace: the higher the *weight*, the higher the number of black pixels in a row and thus the thicker the ink trace produced by the scribe.

The information regarding the inter-character space and the number of characters in a row, may be indicative of the scribe's behavior. Unfortunately, such information is very difficult to obtain, since it requires to extract the single characters (including abbreviation signs) contained in each word: as it is well known, this problem is very complex and far to be solved in the general case. Given the diversity of writing styles and of their individual executions, as well as the different thickness that writing could assume according to the morphology of the instrument used to trace it and of the pressure exerted on it by the scribe, we believe that a reasonable estimate of the number of characters in a row may be obtained by counting the number of peaks in the pixel projection histogram on the horizontal axis for that row (see Fig. 3). Therefore, we have added this measure (denoted as $peaks$) to the set of features.

Modular ratio (or $module$) is a typical palaeographic feature, which estimates the ratio between the horizontal and vertical size of a character (width/height ratio). Such measure may be used by palaeographers for distinguishing a copyist from another. Following this definition, we have computed the modular ratio for each row considering the ratio between the average width of characters in a row (denoted as X_{dim}) and the height of the "center zone" of the words in that row (denoted as Y_{dim}). X_{dim} is obtained by dividing the row width (in pixel) by the value of the feature $peaks$ for the same row, while Y_{dim} has been estimated by using the pixel projection histograms on the horizontal and vertical axes.

Once the "center zone" of each row in a column has been estimated, the interlinear spacing between two rows is the distance in pixels between their "center zones". Modular ratio, interlinear spacing and modular ratio/interlinear spacing ratio characterize not only the way of writing adopted by a single scribe in a specific context, but might also hint geographical and/or chronological distinctions. In Maniaci

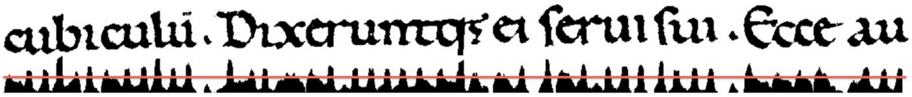

Fig. 3. The number of peaks in the pixel projection histogram on the horizontal axis for a row.

Table 1
The whole set of considered features.

	Feat. id.	Feat. name
SET 1	F1	Intercolumnar distance
	F2	Upper margin
	F3	Lower margin
SET 2	F4	Exploitation
	F5	Row number
	F6	Modular ratio
	F7	Interlinear spacing
SET 3	F8	Weight
	F9	Peak number
	F10	Modular ratio/Interlinear spacing

and Orofino (2012), for instance, the distance among layout lines in rows and the dimension of letters significantly differentiate Spanish and Italian minuscule. The whole set of considered features is summarized in Table 1.

4. Features analysis

In order to identify the set of features having the highest discriminant power, we have used five standard *univariate* measures. Each of them ranks the available features according to a criterion which evaluates the effectiveness of each single feature in discriminating samples belonging to different classes. The final ranking of all the features has been obtained by using the Borda count rule. According to such a rule, a feature receives a certain number of points corresponding to the position in which it has been ranked by each univariate measure. In our study, we have considered the following univariate measures: Chi-square (Liu and Setiono, 1995), Relief (Kononenko, 1994), Gain ratio, Information Gain and Symmetrical uncertainty (Hall, 1999).

The *Chi-Square* measure estimates feature merit by using a discretization algorithm based on the χ^2 statistic. For each feature, the related values are initially sorted by placing each observed value into its own interval. The next step uses the chi-square statistic χ^2 to determine whether the relative frequencies of the classes in adjacent intervals are similar enough to justify the merge. The formula for computing the χ^2 value for two adjacent intervals is the following:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{C} \frac{(A_{ij} - E_{ij})^2}{Eij} \tag{3}$$

where C is the number of classes, A_{ij} is the number of instances of the jth class in the ith interval and E_{ij} is the expected frequency of A_{ij} given by the formula: $E_{ij} = R_i C_j / N_T$ where R_i is the number of instances in the ith interval and C_j and N_T are the number of instances of the jth class and total number of instances, respectively, in both intervals.

The extent of the merging process is controlled by an automatically set χ^2 threshold. The threshold is determined through attempting to maintain the consistency of the original data.[2]

The second considered measure is the Relief, which uses instance based learning to assign a relevance weight to each feature. The assigned weights reflects the feature ability to distinguish among the different classes at hand. The algorithm works by randomly sampling instances from the training data. For each sampled instance, the nearest instance of the same class (nearest hit) and different class (nearest miss) are found. A feature weight is updated according to how well its values distinguish the sampled instance from its nearest hit and nearest miss. A feature will receive a high weight if it differentiates between instances from different classes and has the same value for instances of the same class.

Before introducing the last three considered univariate measures, let us briefly recall the well known information-theory concept of entropy. Given a discrete variable X, which can assume the values $\{x_1, x_2, \dots, x_n\}$, its entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2(x_i) \tag{4}$$

where $p(x_i)$ is the probability mass function of the value x_i. The quantity $H(X)$ represent an estimate of the uncertainty of the random variable X. The entropy concept can be used to define the conditional entropy of two random variables X and Y taking values x_i and y_j respectively, as:

$$H(X|Y) = -\sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)} \tag{5}$$

where $p(x_i, y_j)$ is the probability that $X = x_i$ and $Y = y_j$. The quantity in (5) represents the amount of randomness in the random variable X when the value of Y is known.

The above defined quantities can be used to estimate the usefulness of a feature X to predict the class C of unknown samples. More specifically, such quantities can be used to define the *information gain* (I_G) concept:

$$I_G = H(C) - H(C|X) \tag{6}$$

I_G represents the amount by which the entropy of C decreases when X is given, and reflects additional information about C provided by the feature X.

The last three considered univariate measures use the information gain defined in (6). The first one is the information Gain itself. The second one, called *Gain Ratio* (I_R), is defined as the ratio between the information gain and the entropy of the feature X to be evaluated:

$$I_R = \frac{I_G}{H(X)}. \tag{7}$$

Finally, the third univariate measure taken into account, called *Symmetrical Uncertainty* (I_S), compensates for information gain bias toward attributes with more values and normalizes its value to the range $[0, 1]$:

$$I_S = 2.0 \times \frac{I_G}{H(C) + H(X)} \tag{8}$$

5. Scribe distinction

The last block performs the recognition task, which has the effect of identifying in each page the rows written by the same copyist. In our study, we assumed the availability of a training set of labeled data, produced through the traditional palaeographic analysis, in which N different copyists have been identified. We also assumed that a test set of labeled data is available, which constitutes the ground truth for assessing the performance of our system: this set of data was used only during the testing phase of our system.

Moreover, we assumed that each single pattern to be classified is formed by a group of M consecutive rows, described by using the previously defined features. More specifically, patterns belonging to the

[2] Two patterns are inconsistent if they have the same (discretized) values but belongs to different classes.

same page share the same features of both the first and the second set, while feature values of the third set are averaged over the M rows forming each group. Thus, each pattern is represented by a feature vector containing 10 real values.

Finally, we assumed that each pattern was written by a single copyist: in principle, this assumption may be not always verified but, choosing an appropriate value for the parameter M and avoiding to include in the same pattern rows belonging to different pages or to different paragraphs, the risk of obtaining patterns written by more copyists should be very limited.

In order to identify the text written by each of the N copyists who participated to the handwriting process, we used two different classification architectures: in the first one, a single N class classification system employing all the available features has been designed for associating each input pattern to one of the N copyists. In the second one, a specific classification system has been developed for each copyist, to distinguish his handwritten text from that produced by all the others. This implies that the initial N-class classification problem is split up into N two-class problems, each associated to one of the scribes to be identified (see Fig. 4). As shown in the figure, the architecture is organized in two stages: in the first one, each single classifier is associated to a given copyist and decides, on the basis of specifically selected features, whether an input pattern has been produced or not by that copyist. In the second one, on the basis of the whole set of responses provided in the first stage, the input pattern is attributed to one of the N copyists, or it is rejected.

The rationale behind this approach is the assumption that some of the available features may be more effective in capturing the distinctive aspects of each class. i.e. the hand of a particular scribe, with respect to the others (Cordella et al., 2010; Marrocco and Tortorella, 2016). As a consequence, during a learning phase, a feature selection procedure identifies for each scribe the feature subset which better discriminates that scribe from all the others: the corresponding two-class classifier is then trained by using such features. During the operative phase, an unknown sample is supplied as input to all the classifiers in the first stage and a binary decision is taken by each of them. Thus, the first stage produces a binary vector, in which a value “1” in the ith position means that the input sample has been associated by the ith two-class classifier to the ith scribe, while a value “0” means the opposite. In the second stage the input sample is classified only if the binary vector contains just a single value “1”, otherwise it is rejected. This implies that an input sample may be rejected either if no classifiers in the first stage associated it to one of the scribes, or if that sample has been associated to more scribes. Such a second case typically happens for scribes exhibiting very similar handwritings.

In the Experimental results section, we tested the proposed features for both classification architectures.

6. Experimental results

As anticipated in the Introduction, we have tested our system on a large dataset of digital images obtained from a giant Latin copy of the whole Bible, called “Avila Bible”[3] . It is worth noticing that because giant Bibles are very large and difficult to handle, they have not been often digitized in recent times. Moreover, available microfilms are not good enough to perform pattern recognition analysis and new digital reproduction are not easy (and very expensive) to obtain.

The Avila Bible consists of 870 two-column pages, but we considered only 800 pages and the palaeographic analysis has individuated the presence of 12 scribal hands. The pages written by each copyist are not equally numerous (they range from 1 to 143) and there are cases in which parts of the same page are written by different copyists. This implies that the classification problem to be solved is characterized by a strongly imbalanced distribution of the number of samples per class. Since the rubricated letters might be all the work of a single scribe, we have removed them during a pre-processing step, thus considering only 12 copyists to be identified. Rubrication, in fact, was typically performed by specialized scribes in red ink, following the instructions given by the text scribes. Usually, red headings were added to emphasize the end of one section of text and the beginning of another, or to introduce the subject of a section.

Table 2
Number of samples and recognition rates for the values of M tested.

M	#samples	rec. rate
2	37,611	95.53%
3	25,471	96.03%
4	20,246	**98.25%**
5	15,559	95.94%
6	12,959	94.30%

Since the parameter M plays an important role in the feature extraction process of our approach, we performed a set of experiments to find its best value. To this aim, we tested the following values: $\{2, 3, 4, 5, 6\}$. For each value, we performed the feature extraction process obtaining a dataset with a different number of samples. Then, each dataset was normalized by using the Z-normalization method.[4]

Once a dataset has been built and normalized, we classified its samples by using the decision tree classifier. For each dataset, we performed 20 runs and evaluated the classification results using the 10-fold cross-validation with different random seeds. The obtained results, averaged over the 20 runs, are shown in Table 2. From the table it can be seen that the best performance was achieved with $M = 4$. This result confirms that the above value represents the best compromise between two opposite requirements: on the one hand, high M values allows us to obtain, for each extracted sample, significant average values for the devised features. On the other hand, small M values ensure that the rows included in the pattern are written by a single scribe, thus allowing us to capture the distinctive aspects of his hand.

The dataset obtained for $M = 4$ has been divided in two subsets: the first one (TR hereafter) containing 10430 samples, has been used as training set for both the N class classification system, and the classifiers in the first stage of Fig. 4, while the second one (TS hereafter), containing the remaining 10,437 samples, has been used for testing both the considered classification systems. For each class, the samples have been randomly extracted from the database in such a way to ensure that, approximately, each class has the same number of samples in both the training and the test set.

6.1. Feature evaluation

In order to ascertain the effectiveness of the whole set of features defined in Section 3, as well as the discriminant power of each of them, we performed several experiments using the dataset TR and the univariate measures described in Section 4. For each univariate measure, a ranked list of all the available features was provided.

In a first set of experiments, the univariate measures were computed considering the N-class classification problem. This implies that we evaluated the ability of each single feature in separating TR samples belonging to different classes. The results of this analysis are reported in Table 3, which shows the ranking related to each measure. Although the different measures produced quite different results, they give a good insight about the best and worst features.

[3] The dataset extracted from the Avila Bible images is publicly available at the following page: http://webuser.unicas.it/fontanella/avila.html.

[4] For each feature f_i, we first computed the mean μ_i and the standard deviation σ_i, then we applied the following transformation:

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i}$$

where x_{ij} is the value of the ith feature extracted from for the jth sample and z_{ij} is the corresponding normalized value.

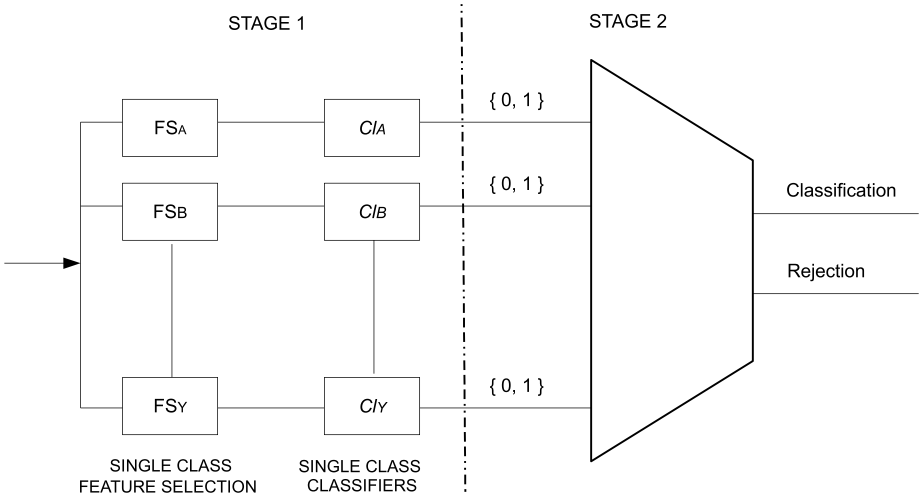

Fig. 4. Scribe detection System.

Table 3
Feature ranking according to the five considered measures.

Measure	Ranking									
C_S	F4	F3	F2	F1	F5	F9	F7	F6	F10	F8
I_R	F4	F5	F1	F3	F2	F9	F7	F6	F10	F8
I_G	F4	F3	F2	F1	F5	F9	F7	F6	F10	F8
I_S	F4	F3	F5	F1	F2	F9	F7	F6	F10	F8
R_F	F5	F4	F1	F9	F3	F7	F6	F10	F8	F2

In order to combine the results provided by the different rankings taken into account, we computed a global score for each feature according to the following equation, derived by the Borda count rule:

$$Os_i = \sum_{j=1}^{5} 10 - pos_{ij} \tag{9}$$

where Os_i is the global score of the ith feature, while pos_{ij} is the position of the ith feature in the j-ranking.

Fig. 5 depicts the global feature ranking obtained by applying Eq. (9). The ranking shows that feature $F4$, the exploitation coefficient computed according to Eq. (1), is the most discriminant feature, while feature $F8$, which represents the exploitation coefficient computed on a single row, do not exhibit a significant discriminating power. This result is probably due to the fact that, in case of a high standardized book typology, the ability in exploiting all the available space along a whole column, uniformly filling each row with words, is highly distinctive of the copyists' behavior, while the same measure, performed on each single row, is much more variable and thus less distinctive.

In a second set of experiments, the above univariate measures were computed considering the N two-class problems in the first stage of the classification architecture reported in Fig. 4. In this case, we tried to characterize the ability of each feature in discriminating each single copyist from all the others. Following this approach, we obtained twelve different rankings, one for each two-class problem. Similarly to the previous case, in order to combine the results provided by the different rankings, we computed an average global score for each feature, according to the following equation:

$$\overline{Os_i} = \frac{1}{12} \sum_{k=1}^{12} Os_i^k \tag{10}$$

where $\overline{Os_i}$ is the average global score of the ith feature, while Os_i^k is the global score of the ith feature in the kth two-class problem, computed according to Eq. (9).

The values of the average global score are shown in Fig. 6.

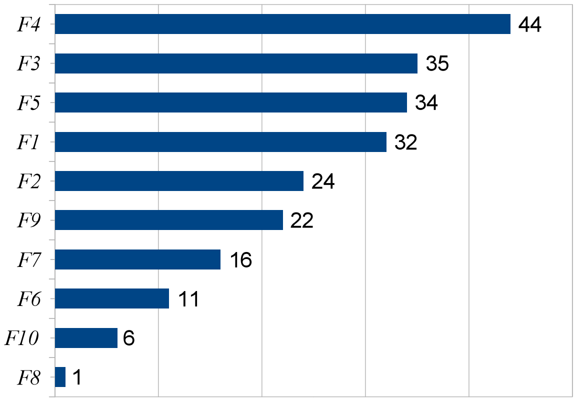

Fig. 5. The global feature ranking.

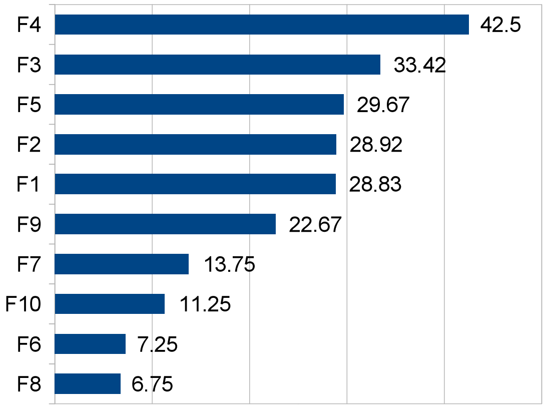

Fig. 6. The average global feature ranking.

The comparison between the values reported in the ranking of Fig. 5 with those reported in Fig. 6 confirms that the exploitation coefficient $F4$ is the most discriminant feature, while the exploitation coefficient computed on a single row (feature $F8$) is the worst. It is also interesting to note that almost all the features maintained their position in the rankings, with few not relevant local variations (see the features $F1$ $F2$ and $F6$ $F10$ in both rankings). The consistency of these results confirms

Table 4
Feature subset selection.

Scribe	Selected features
A	F1–F2–F3–F4–F9
B	F2–F9
C	F2–F4
D	F1–F4
E	F3–F4
F	F1–F2–F4
G	F4
H	F2–F4
I	F2–F4
W	F3–F4
X	F4
Y	F4

the effectiveness of our approach for evaluating the discriminative power of each single feature.

6.2. Feature selection

As regards the feature selection processes in the first stage of Fig. 4, we have considered feature-subset measures rather than the *univariate* measures used in the previous experiments. *Univariate* measures, in fact, have been designed to estimate the discriminative power of a single feature, while our aim is that of evaluating the discriminant power of groups of features, taking also into account the complementarity and redundancy of a single feature with respect to the others. More specifically, our feature selection method combines the Best First (BF) search strategy (based on the beam search heuristics) for searching feature subsets (Xu et al., Nov. 14–17, 1988), and the Consistency Criterion (CC) for evaluating their worth (Liu and Setiono, 1996). Such a criterion provides an effective measure of how well samples belonging to different classes are separated in a feature sub-space. The set of features selected for each class are summarized in Table 4.

The data in Table 4 show that both the type of features, and their number vary for each two-class problem, confirming our basic idea that specific feature subsets could allow us to better distinguish a single copyist from all the others. The data also show that there are simple cases in which a single feature completely characterize the writer's hand, while there are other copyists, who require a higher number of feature to be distinguished. For the scribe labeled 'A', for instance, 5 features were selected, while for the scribes 'G', 'X', 'Y' only the feature $F4$ was included in the corresponding feature subsets.

The histogram in Fig. 7 shows how many times each feature has been selected by our subset evaluation method. It is worth noticing that the features $F4$ is included in almost all the feature subsets, confirming that it is able to capture important aspects of scribal individualism, such as the ability in exploiting the available space, or the thickness of the strokes, due to the pressure on the quill. On the contrary, features $F5$, $F6$, $F7$, $F8$ and $F10$ are never selected. As regards the first three features, this may be probably due to the standardization imposed by the writing style (the so called "reformed Carolingian minuscule"), while feature $F10$ exhibited poor discriminant power also in the previous experiments. Finally, it is also confirmed that feature $F8$ does not add further information with respect to feature $F4$.

6.3. Choosing a classification scheme

In order to maximize classification performance, we tested four different classification schemes, namely Decision Trees, K-Nearest Neighbor, Neural Networks and Support Vector Machines, using for each of them the implementation provided by the WEKA tool (Hall et al., 2009). We chose these classifiers because: (i) they are widely used and standard implementations are available for each of them; (ii) they represent different paradigms of classification algorithms; (iii) they represent effective classification schemes.

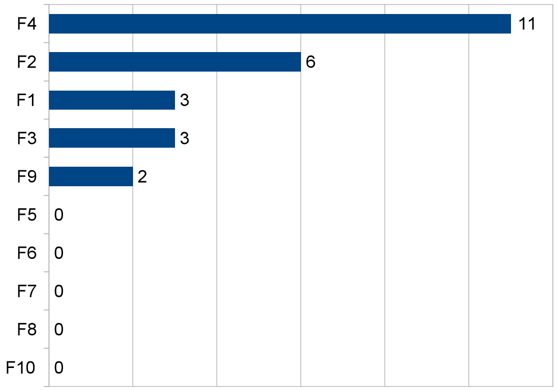

Fig. 7. Number of times each feature has been selected by feature subset selection modules.

Table 5
Results on TS for the considered classification schemes. The best result is in bold.

Class. scheme	*rec. rate*
DT	**98.25%**
k-NN	75.61%
NN	94.56%
SVM	82.67%

Table 6
Recognition rates (expressed in percentages) of the tested k values for the k-NN classifier. The best result is in bold.

k	*rec. rate*
1	**75.61**
3	73.34
5	73.16
7	71.96
9	70.97
11	70.48
13	69.32

In these experiments, we considered only the N-class problem applied to the feature space including all the available features. Since this problem is more complex than each of the two-class problems in the first stage of Fig. 4, the best selected classification scheme could be profitably used for implementing both the classification architectures presented in Section 5. For each scheme, we performed a preliminary set of experiments to find the related best set of parameters, using a "grid-search" approach and the data included in the set TR: various tuples were tried and the one with the best 10-fold cross-validation accuracy was chosen.

Once the best sets of parameters have been found, we performed 20 runs of the 10-fold cross-validation for each classification scheme, using different random seeds and the dataset TS. The obtained results, averaged over 20 runs, are reported in Table 5. The data in the table show that Decision Trees approach largely outperforms all the other classification schemes: thus, we adopted such a scheme for implementing the N-class classifier as well as each of the two-class classifiers in the first stage of Fig. 4. The performance of both classification architectures will be discussed in detail in the next section. The above results are in good accordance with the theory: it has been demonstrated, in fact, that Decision Trees exhibit good performance in case of imbalanced data even if no oversampling or undersampling is applied and, thanks to its tree structure, the learning procedure is effective even when only few features are used (Bria et al., 2016; Marrocco et al., 2011). As above noticed, this is exactly the case of our problem.

In the following, the parameter testing and the learning algorithms are detailed for each classification scheme.

- A **Decision Tree** (DT in the following) is a decision support tool with a tree graph structure. In a DT, the internal nodes represent attribute tests, where each branch yields the outcome of the test, whereas leaf nodes represent class labels. The paths from the root node to the leafs represent classification rules. For the tree learning we used C4.5 algorithm. This algorithm builds a decision tree with a top down approach, by using the concept of information entropy. Given a training set S, it breaks down S into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. At each node of the tree, C4.5 chooses the attribute that most effectively splits the corresponding sample subsets. The splitting criterion is the normalized information entropy gain which measures how much the subsets are homogeneous, in terms of class labels, with respect to the split set. The algorithm then recurs on the smaller subsets.

  As concerns the algorithm parameter optimization, we found the best values for the confidence factor C_f and the minimum number of instances per leaf N_l. The first parameter is used to test the effectiveness of the post-pruning: lowering the confidence factor decreases the amount of post-pruning. The second parameter, instead, represents the minimum number of instances per leaf. We tested the following sets of parameters for C_f and N_l respectively: $\{0.1, 0.15, 0.2, 0.25, 0.3\}$ and $\{1, 2, 3, 4, 5\}$. The pair $(0.25, 2)$ obtained the best 10-fold cross-validation accuracy.
- the **K-Nearest Neighbor** algorithm (k-NN) is a well known non parametric method that can be used for both classification and regression. According to this approach, an unknown sample is labeled with the most common label among its k nearest neighbors in the training set. The rationale behind the k-NN classifier is that, given an unknown sample $\mathbf{x}$ to be assigned to one of the c_i classes of the problem at hand, the a-posteriori probabilities $p(c_i|\mathbf{x})$ in the neighborhood of $\mathbf{x}$ may be estimated by looking at the class labels of the k nearest neighbors of $\mathbf{x}$. Nonetheless its simplicity, k-NN has shown to be able to obtain good performances (Govindarajan and Chandrasekaran, 2010; Kumar et al., 2011; Pérez-Cortes et al., 2000). For this classification scheme, we sought for the best k value, testing the following values: $\{1, 3, 5, 7, 9, 11, 13\}$. The obtained results are shown in Table 6; From the table it can be seen that the value $k = 1$ significantly outperforms the higher values. This result suggests that, for many samples, higher value of k force the algorithm to consider far neighbors belonging to class different from the true class of the sample to be classified.
- A **Neural Network** (NN) is an information processing system made up of a number of simple, highly interconnected processing elements called *neurons*, each containing an activation function. NN topologies are usually organized in *layers*. The patterns are presented to the network via the "input layer", while the final answer is provided by means of an "output layer". Once the network topology has been chosen, a NN must be trained providing as input a set of labeled samples.

  In this study we considered a feed-forward completely connected network, trained by using the back-propagation algorithm, which is one the most popular training algorithm, successfully applied to a large variety of real world classification tasks. As concerns the NN training parameters, we searched the best values for four parameters: the learning rate l_r that affects the way in which connection weights are updated, the momentum m_w applied to the weights during updating, the number of training cycles N_c and the number of hidden neurons N_h. In this case, in order to reduce the number of experiments, we adopted a two-steps grid strategy. In the first step we searched the best pair (l_r, m_w), using for N_c and N_h the default values suggested by the Weka tool. For both parameters we tested the set of values $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and the pair $(0.1, 0.3)$ obtained the best accuracy. Once the best values were found for l_r and m_w, we performed a second step to seek for the best pair (N_c, N_h). For the parameter N_c we tested the set of values $\{50, 100, 200, 500, 1000, 1500\}$, while for the parameter N_h the set of values $\{10, 20, 50, 100, 200, 400, 500, 600\}$. The best accuracy was obtained by using the pair $(1000, 500)$.

Table 7
Precision and recall (expressed in percentages), for each writer, obtained by the four classification schemes tested.

Writer	DT		*k*-NN		NN		SVM	
A	prec.	rec.	prec.	rec.	prec.	rec.	prec.	rec.
A	98.0	99.6	78.7	79.3	77.2	86.9	82.6	88.8
B	66.7	80.0	60.0	50.0	0.0	0.0	66.7	80.0
C	95.2	97.1	73.6	51.5	77.8	47.6	83	75.7
D	97.2	97.5	68.6	62.6	74.0	62.9	85.6	77.6
E	96.7	97.4	72.2	66.7	76.2	78.2	85.5	83.4
F	98.9	98.9	64.4	68.5	70.3	64.1	76.9	71.0
G	98.4	97.5	57.7	59.5	67.5	55.3	73.8	72.5
H	98.0	95	56.5	60.6	79.5	62.5	84.6	76.2
I	100	98.8	98.3	96.4	99.4	92.2	99.6	97.8
W	100	97.8	80.0	71.1	67.6	55.6	91.7	97.8
X	96.9	88.9	91.0	87.2	88.4	86.4	93.6	89.7
Y	100	98.9	87.7	80.1	92.8	82.4	90.5	88.8

- **Support Vector Machines** (SVMs) are supervised learning models that are based on the concept of decision planes, which linearly separates, in the feature space, objects belonging to the different classes. Intuitively, given two classes to be discriminated in a given feature space, a good separation is achieved by the hyperplanes that have the largest distance to the nearest training points belonging to different classes; in general, the larger the margin, the lower the generalization error of the classifier. While the basic idea of the SVM applies to linear classifiers, they can be easily adapted to non-linear classification tasks by using the so called "kernel trick", which implies the mapping of the original features into an higher dimensional feature space. Differently from other classifier schemes, whose learning usually implies stochastic and suboptimal gradient descent procedures, e.g. NN and DT, given a training set, the SVM learning optimization problem can be solved by using quadratic programming algorithms which provide exact solutions.

  SVMs have shown to be effective in solving many real world applications like, for example, handwriting recognition, text categorization and image classifications.

  As suggested in Keerthi and Lin (2003), in our experiments we chose as kernel the radial basis function (RBF): $K(\mathbf{x}, \mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|^2}$

  As regards the SVM parameters, we sought for the best values for the penalty parameter C and for the γ term of the RBF kernel. We tested the following values: $C = \{2^{-5}, 2^{-3}, \dots, 2^{15}\}$ and $\gamma = \{2^{-15}, 2^{-13}, \dots, 2^{3}\}$. We achieved the best accuracy by using the pair $(2^{11}, 2^{-3})$.

As concerns the choice of the classification scheme, we conducted a further experiment to test if classifiers could be adapted to the writer. To this aim, for each writer, we measured the performance of each classifier in terms of precision and recall. The obtained results are shown in Table 7. From the table it can be seen that decision trees performs better than the other classification schemes, except for the writer B. For this writer, decision trees outperformed both k-NN and NN but performed equal to the SVM. However, for the sake of simplicity, we chose the decision tree classification scheme for the writer. This choice allowed us to use a unique classifier for all the writers.

6.4. *Classification results*

The classification results on dataset TS are summarized in Table 8. For the sake of clarity, the N-class classification system using all the

Table 8
Classification results. Error rates have been computed on the accepted samples.

	Rec. rate	Error rate	Reject rate
$System1$: all features N-class classifier	98.13%	1.87%	–
$System2$: feature selection two-class classifiers	99.26%	0.11%	0.63%
$System3$: all features two-class classifiers	96.74%	0.30%	2.97%

available features has been denoted as $System1$, while the classification architecture of Fig. 4 has been denoted as $System2$. Finally, we have developed a further classification system, denoted as $System3$, obtained by removing the feature selection modules in the architecture of Fig. 4, thus using all the available features in each two-class problem. We have considered such a system to evaluate if the performance of $System2$ is mainly due to splitting the original N-class problem in N two-class problems, or to the combined effect of the splitting process and the selection of appropriate features for each subproblem. As previously discussed, all the above systems have been implemented by using the Decision Tree classification scheme, whilst both $System2$ and $System3$ adopt the same decision rule in the second stage.

The analysis of the classification results shows that all the systems exhibited high recognition rates, confirming the effectiveness of both the proposed features, and the Decision Tree classification scheme, which was actually able to manage the complexity of the problem at hand and the unbalanced distribution of samples in the training data. These results are also significantly better than those obtained in previous works (De Stefano et al., 2011b; De Stefano et al., 2011a), in which a Neural Network classification scheme was used.

The comparison between the results of $System1$ and $System2$ shows that the idea of splitting the original N-class problem in N two-class problems, selecting appropriate features for each of them, allowed us to increase the overall recognition rate, from 98.13% of $System1$ to 99.26% of $System2$. Moreover, the introduction of a reject option in $System2$ produced a strong reduction of the error rate, whose value is 0.11%, with a reject rate equal to 0.63%. In practice, about 60% of errors of $System1$ were recovered, about 35% were rejected, while only the few remaining ones were wrongly classified.

The results obtained by $System3$ also confirmed the effectiveness of the reject option, which allowed us to considerably reduce the errors, rejecting about 85% of samples wrongly classified by $System1$. The introduction of such an option in $System3$, however, also produced the undesired effect of rejecting a percentage of samples correctly classified by $System1$, thus reducing the overall classification rate of about 1.4%. It is worth noticing that the large majority of rejections in the second stage of $System3$ are due to the fact that the corresponding samples have been associated to more copyists by the classifiers in the first stage. As above discussed, these cases typically happens for copyists exhibiting very similar hands.

Many of these complex situations have been recovered in $System2$, where the feature selection phase, performed for each two-class problem, allowed us to select the subset of features really distinctive for each copyist: the effect is a considerable reduction of the error rate with respect to $System1$ and a considerable reduction of the reject rate with respect to $System3$.

Table 9 summarizes the results provided by the two-class classifiers in the first stage of $System2$. For each of them, the following values are reported:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN};$$

$$Precision = \frac{TP}{TP+FP};$$

$$Recall = \frac{TP}{TP+FN};$$

where TP indicates the *true positive* samples, TN the *true negative*, FP the *false positive* and FN the *false negative* ones.

Table 9
Analysis of the results (expressed in percentages) in the first stage of $System2$.

Classifier	I Stage — two-class classifiers			
	Selected Features	Accuracy	Precision	Recall
Cl_A	F1–F2–F3–F4–F9	99.37	99.48	98.97
Cl_B	F2–F9	99.98	80	80
Cl_C	F2–F4	99.96	96.26	100
Cl_D	F1–F4	99.97	100	99.15
Cl_E	F3–F4	99.98	100	99.82
Cl_F	F1–F2–F4	100	100	100
Cl_G	F4	100	100	100
Cl_H	F2–F4	99.96	100	99.23
Cl_I	F2–F4	100	100	100
Cl_W	F3–F4	99.99	97.78	100
Cl_X	F4	99.97	99.43	100
Cl_Y	F4	99.95	99.88	99.25

Table 10
Analysis of the overall results of $System2$ for each copyist.

Copyst	# Samples	$System2$		
		Rec. rate	Error rate	Reject rate
A	4286	98.97	0.02	1.01
B	5	80	20	0
C	103	100	0	0
D	352	98.86	0.85	0.29
E	1095	99.36	0	0.64
F	1961	100	0	0
G	446	99.78	0	0.22
H	519	98.07	0.77	1.16
I	831	99.52	0	0.48
W	44	100	0	0
X	522	99.24	0	0.76
Y	266	99.25	0.75	0

The comparison between these data and those reported in Table 8 shows that there is, on average, a slight difference between the overall recognition rate and the accuracies of the classifiers in the first stage of $System2$. This is due to the fact that the number of false positive and false negative errors provided by each two-class classifier in the first stage is small and, in the large majority of cases, such wrongly classified samples do not generate errors, but are rejected in the second stage.

Finally, Table 10 reports the overall results for each copyist provided by $System2$. Note that for the copyist labeled as B the percentages shown in the table are not meaningful since there are only 5 samples in the set TS: thus 80% of recognition rate means that only one of these samples has been wrongly classified.

7. Conclusions

We presented a novel approach for automatic scribe identification in medieval manuscripts, which is based on the use of a set of features directly derived from the analysis of page layout. One of the aims of this study, in fact, is that of verifying if, in case of a high standardized book typology (such as the Giant Bibles, to which the "Avila Bible" considered in our experiments belongs), the use of information concerning only the layout of the pages allows us to obtain satisfactory results.

The experimental investigation has regarded several aspects. First we assess the effectiveness of the proposed features by using some standard univariate measures. The exploitation coefficient resulted the most discriminant feature while the weight, which represents the exploitation coefficient computed on a single row, was the worst. This result is probably due to the fact that, in case of a high standardized book typology, the ability in exploiting all the available space along a whole column, uniformly filling each row with words, is highly distinctive of the copyists' behavior, while the same measure, performed on each single row, is much more variable and thus less distinctive.

We also evaluated the classification performance of our system. The obtained results confirmed that the basic ideas of our system, i.e. splitting up the initial N-class problem in N two-class problems and selecting an appropriate subset of features for characterizing each single scribe, is very promising and allowed us to obtain a very high recognition rate using only simple layout features. Moreover, the introduction of a reject option produced a strong reduction of the error rate, without significantly affecting the overall recognition rate. Summarizing, the experimental results suggest that the high performance of our system is due to the combined effect of both the use of powerful features, and the definition of the above classification architecture, able to capture the distinctive aspects of each particular scribe with respect to the others.

It is worth to remark that the proposed approach has been tested on a very complex manuscript, the "Avila Bible", which was produced in different countries, Italy and Spain, along about three centuries, thus offering an "anthology" of contemporary and not contemporary scribal hands, ranging from the 11th to the 15th or 16th century. Thus the layout data extracted from this manuscript represent a severe test bed for evaluating the effectiveness and the potentialities of our approach for distinguishing scribal hands independently from their date and origin.

It is also worth to remark that we have not compared our results with those obtained by other researchers because, to the best of our knowledge, this is the first study in which the "Avila Bible" has been used for automatically identifying the different copyists who participated to the writing process.

From the experimental results we can draw the following conclusions:

- the use of layout features for the identification of the scribe shows an interesting potential in the case of highly standardized hands, trained to the faithful reproduction of the same letter forms, ligatures and abbreviations; in such cases, even small differences in the distribution of the writing trace within the written area, being more difficult for the scribe to control, may help to detect the difference between two otherwise very similar hands.
- the introduction of a reject option may be very useful for palaeographers, since they may concentrate their attention on those sections of the manuscript which have not been reliably classified on the basis of the proposed features. In principle, the possibility of rejecting unreliable samples may be also used for identifying, in other medieval manuscripts, one or more copyist whose characteristics have been learned by our system, rejecting all the patterns corresponding to unknown copyists.
- the high number of scribes collaborating to the writing of the Avila Bible, and the fact that the method provides reliable results in almost all cases with reference to their distinction is in itself significant for paleographers, whose research objectives are not limited to the automatic identification of the hands, but include the in-depth characterization of individual scribes, relying on the measurement of conscious and unconscious features other than those considered by traditional analysis.

As concerns the future work, it will include:

- testing the ability of our system in identifying in other manuscripts, copyists previously identified;
- studying more complex rejection rules, which exploit information about classification reliability. Such kind of information, in fact, is provided by Decision Trees classifiers together with the classification result;
- testing the proposed method on other Bibles or also on other expressions of Caroline minuscule or other script style. It is worth noticing that it will require a lot of preliminary work for the definition of the ground truth.

References

Al-Aziz, A.M.A., Gheith, M., Sayed, A.F., 2011. Recognition for old Arabic manuscripts using spatial gray level dependence (SGLD). Egypt. Inform. J. 12 (1), 37–43.

Antonacopoulos, A., Downton, A.C., 2007. Special issue on the analysis of historical documents. IJDAR 9 (2–4), 75–77.

Bozzolo, C., Coq, D., Muzerelle, D., Ornato, E., 1982. Noir et blanc. Premiers résultats d'une enquête sur la mise en page dans le livre médiéval. In: Il Libro E Il Testo, Urbino, pp. 195–221.

Bria, A., Marrocco, C., Molinara, M., Tortorella, F., 2016. An effective learning strategy for cascaded object detection. Inform. Sci. 340, 17–26.

Bulacu, M., Schomaker, L., 2007. Text-independent writer identification and verification using textural and allographic features. IEEE Trans. Pattern Anal. Mach. Intell. 29 (4), 701–717.

Ciula, A., 2009. The palaeographical method under the light of a digital approach. In: Rehbein, M., Sahle, P., Schaßan, T. (Eds.), Kodikologie und Paläographie im Digitalen Zeitalter-Codicology and Palaeography in the Digital Age. Bod, Norderstedt, pp. 219–237.

Conti, A., Da Rold, O., Shaw, P. (Eds.), 2015. Writing Europe, 500-1450: Texts and Contexts. Boydell and Brewer.

Cordella, L.P., De Stefano, C., Fontanella, F., Marrocco, C., Scotto di Freca, A., 2010. Combining single class features for improving performance of a two stage classifier. In: 20th International Conference on Pattern Recognition, ICPR 2010. IEEE Computer Society, pp. 4352–4355.

Dahllof, M., 2014. Scribe attribution for early medieval handwriting by means of letter extraction and classification and a voting procedure for larger pieces. In: Proceedings of the 22nd International Conference on Pattern Recognition. IEEE Computer Society, pp. 1910–1915.

De Stefano, C., Fontanella, F., Maniaci, M., Scotto di Freca, A., 2011a. A method for scribe distinction in medieval manuscripts using page layout features. In: Maino, G., Foresti, G. (Eds.), Image Analysis and Processing - ICIAP 2011. In: Lecture Notes in Computer Science, vol. 6978, Springer, Berlin –Heidelberg, pp. 393–402.

De Stefano, C., Fontanella, F., Maniaci, M., Scotto di Freca, A., 2011b. Exploiting page layout features for scribe distinction in medieval manuscripts. In: Proc. of the 15th International Graphonomics Society Conference. IGS 2011, pp. 106–109.

Dhali, M.A., He, S., Popovic, M., Tigchelaar, E., Schomaker, L., 2017. A digital palaeographic approach towards writer identification in the dead sea scrolls. In: Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods, ICPRAM, pp. 693–702.

Dinstein, I., Shapira, Y., 1982. Ancient hebraic handwriting identification with run-length histograms. IEEE Trans. Syst. Man Cybern. 12 (3), 405–409.

En, S., Petitjean, C., Nicolas, S., Heutte, L., 2016. A scalable pattern spotting system for historical documents. Pattern Recognit. 54, 149–161.

Govindarajan, M., Chandrasekaran, R., 2010. Evaluation of k-nearest neighbor classifier performance for direct marketing. Expert Syst. Appl. 37 (1), 253–258.

Gurrado, M., 2009. "Graphoskop", uno strumento informatico per l'analisi paleografica quantitativa. In: Rehbein, M., Sahle, P., Schaßan, T. (Eds.), Kodikologie und Paläographie im Digitalen Zeitalter-Codicology and Palaeography in the Digital Age. Bod, Norderstedt, pp. 251–259.

Hall, M., 1999. Correlation-based Feature Selection for Machine Learning. (Ph.D. thesis), University of Waikato.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explorations 11 (1), 10–18.

He, S., Samara, P., Burgers, J., Schomaker, L., 2016a. Image-based historical manuscript dating using contour and stroke fragments. Pattern Recognit. 58, 159–171.

He, S., Samara, P., Burgers, J., Schomaker, L., 2016b. Image-based historical manuscript dating using contour and stroke fragments. Pattern Recognit. 58, 159–171.

Joutel, G., Eglin, V., Bres, S., Emptoz, H., 2007. Curvelets based feature extraction of handwritten shapes for ancient manuscripts classification. In: Document Recognition and Retrieval XIV, San Jose, California, USA, January 30 - February 1, pp. 65000D 1–12.

Keerthi, S.S., Lin, C.-J., 2003. Asymptotic behaviors of support vector machines with gaussian kernel. Neural Comput. 15 (7), 1667–1689.

Kononenko, I., 1994. Estimating attributes: Analysis and extensions of relief. In: European Conference on Machine Learning, pp. 171–182.

Kumar, M., Jindal, M., Sharma, R., 2011. k-nearest neighbor based offline handwritten Gurmukhi character recognition. In: 2011 IEEE International Conference on Image Information Processing. ICIIP. IEEE Computer Society, Shimla, Himachal Pradesh, India, pp. 1–4.

Lavrenko, V., Rath, T.M., Manmatha, R., 2004. Holistic word recognition for handwritten historical documents. In: 1st International Workshop on Document Image Analysis for Libraries. DIAL 2004, 23–24 January 2004, Palo Alto, CA, USA, pp. 278–287.

Liang, Y., Fairhurst, M.C., Guest, R.M., Erbilek, M., 2016. Automatic handwriting feature extraction, analysis and visualization in the context of digital palaeography. IJPRAI 30 (4), 1653001 1–26.

Liang, Y., Guest, R.M., Fairhurst, M.C., 2012. Implementing word retrieval in handwritten documents using a small dataset. In: 2012 International Conference on Frontiers in Handwriting Recognition. ICFHR, Bari, Italy, pp. 728–733.

Liu, H., Setiono, R., 1995. Chi2: Feature selection and discretization of numeric attributes. In: Seventh International Conference on Tools with Artificial Intelligence. ICTAI. IEEE Computer Society, Washington, DC, USA, pp. 388–391.

Liu, H., Setiono, R., 1996. A probabilistic approach to feature selection - A filter solution. In: 13th International Conference on Machine Learning. ICML 96, Bari, Italy, pp. 319–327.

Louloudis, G., Kesidis, A.L., Gatos, B., 2012. Efficient word retrieval using a multiple ranking combination scheme. In: 10th IAPR International Workshop on Document Analysis Systems DAS, March 27–29. IEEE Computer Society, Gold Coast, Queenslands, Australia, pp. 379–383.

Maniaci, M., Orofino, G., 2000. Le bibbie atlantiche. Il libro delle Scritture tra monumentalità e rappresentazione. In: Catalogo Della Mostra, Centro Tibaldi, Milano.

Maniaci, M., Orofino, G., 2012. Prime considerazioni sulla genesi e la storia della Bibbia di Avila. In: P. Fioretti, with A. Germano, and M. A. Siciliani, Storie Di Cultura Scritta. Studi Per Francesco Magistrale, CISAM, Spoleto, 537–584.

Marrocco, C., Molinara, M., Tortorella, F., 2011. On linear combinations of dichotomizers for maximizing the area under the ROC curve. IEEE Trans. Syst. Man Cybern. B 41 (3), 610–620.

Marrocco, C., Tortorella, F., 2016. Exploiting coding theory for classification: An LDPC-based strategy for multiclass-to-binary decomposition. Inform. Sci. 357, 88–107.

Moalla, I., Alimi, A.M., Lebourgeois, F., Emptoz, H., 2006. Image analysis for palaeography inspection. In: Second International Workshop on Document Image Analysis for Libraries. DIAL 06, 27–28 April. IEEE Computer Society, Lyon, France, pp. 303–311.

Papaodysseus, C., Rousopoulos, P., Giannopoulos, F., Zannos, S., Arabadjis, D., , Panagopoulos, M., Kalfa, E., Blackwell, C., Tracy, S., 2014. Identifying the writer of ancient inscriptions and Byzantine codices. A novel approach. Comput. Vis. Image Underst. 121, 57–73.

Pérez-Cortes, J., Llobet, R., Arlandis, J., 2000. Fast and accurate handwritten character recognition using approximate nearest neighbours search on large databases. In: Ferri, F., Innesta, J., Amin, A., Pudil, P. (Eds.), Advances in Pattern Recognition. In: Lecture Notes in Computer Science, vol. 1876, Springer, Berlin –Heidelberg, pp. 767–776.

Pintus, R., Yang, Y., Gobbetti, E., Rushmeier, H., 2015. An automatic word-spotting framework for medieval manuscripts. In: 2015 Digital Heritage, Vol. 2. pp. 5–12.

Rath, T.M., Manmatha, R., 2003. Features for word spotting in historical manuscripts. In: 7th International Conference on Document Analysis and Recognition. ICDAR 2003. IEEE Computer Society, Edinburgh, Scotland, UK, 3–6 August, pp. 218–222.

Sampath, V., 2016. Quantifying scribal behavior: a novel approach to digital paleography. (Ph.D. thesis), University of St. Andrews. School of Computer Science.

Schomaker, L., Franke, K., Bulacu, M., 2007. Using codebooks of fragmented connected-component contours in forensic and historic writer identification. Pattern Recognit. Lett. 28 (6), 719–727 Pattern recognition in cultural heritage and medical applications.

Stokes, P., 2009. Computer-aided palaeography, present and future. In: Kodikologie und Paläographie im Digitalen Zeitalter Codicology and Palaeography in the Digital Age. Institut für Dokumentologie und Editorik, pp. 309–338.

Stokes, P.A., 2015. Digital approaches to paleography and book history: Some challenges, present and future. Front. Digit. Humanit. 2, 5.

Wahlberg, F., Mårtensson, L., Brun, A., 2014. Scribal attribution using a novel 3-d quill-curvature feature histogram. In: 2014 14th International Conference on Frontiers in Handwriting Recognition, pp. 732–737.

Wahlberg, F., Mårtensson, L., Brun, A., 2015. Large scale style based dating of medieval manuscripts. In: Proceedings of the 3rd International Workshop on Historical Document Imaging and Processing. HIP '15. ACM, New York, NY, USA, pp. 107–114.

Wei, H., Gao, G., 2014. A keyword retrieval system for historical Mongolian document images. IJDAR 17 (1), 33–45.

Xu, L., Yan, P., Chang, T., 1988. Best first strategy for feature selection. In: 9th International Conference on Pattern Recognition. ICPR 1988, Nov. 14–17, Vol. 2. IEEE Computer Society, Rome, Italy, pp. 706–708.

Yosef, I.B., Beckman, I., Kedem, K., Dinstein, I., 2007. Binarization, character extraction, and writer identification of historical Hebrew calligraphy documents. IJDAR 9 (2–4), 89–99.