IMEKO International Conference on
Metrology for Archaeology and Cultural Heritage
Lecce, Italy, October 23-25, 2017

Measuring Layout Features in Mediaeval Documents for Writer Identification

C. De Stefano[1], F. Fontanella[1], M. Maniaci[1], A. Scotto di Freca[1]

[1]*Università di Cassino e del Lazio Meridionale – ITALY*
{destefano,fontanella,m.maniaci,a.scotto}unicas.it

***Abstract* – Palaeography is the study of ancient handwriting, aiming not only at deciphering, reading, and dating historical manuscripts, but also at reconstructing and interpreting the history of writing techniques and styles from the Antiquity to the end of the Middle Ages. Palaeographers are therefore engaged in discovering when a manuscript was written, where it was written and how the writing was technically executed; they are also interested in characterizing features and habits of individual scribes and in distinguishing them from one another.**

We present a pattern recognition system which tries to solve a typical palaeographic problem: to distinguish the different scribes who have worked together to the transcription of a single medieval book. In the specific case of a high standardized book typology (the so called Latin "Giant Bibles"), we wished to verify if the extraction of certain specifically devised features, concerning the layout of the page, allowed to obtain satisfactory results. The experiments, performed on a large dataset 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.

I. INTRODUCTION

In traditional palaeographic studies handwriting analysis and recognition are performed by human experts who are able to identify the peculiarities of a style or of a particular scribe, in order to answer the following questions: where, when and by whom was a manuscript written; on what grounds is it possible to distinguish among different scribes; how should the development of writing styles or single graphic characteristic be followed and described [1].

Digital palaeography, which has received growing attention in recent years (also thanks to the increasing availability of electronic reproductions), explores new and more "objective" ways of characterising medieval handwritings and distinguishing between scribal hands [2, 3]. The application of computer-based techniques of analysis – originally developed in the field of forensic analysis – to the study of ancient scripts raises delicate and very much debated theoretical issues, only partially overlapping with those concerning modern writing. Various methodologies have been proposed to solve such issues. At a simpler level, the digital approach can be used to replace qualitative measurements with quantitative ones, for instance for the evaluation of parameters such the angle and width of strokes, or the comparison among digital examples of letter-forms. Another application consists in enhancing image quality and presentation, both for teaching and research purposes. In the above mentioned cases, technology is employed to perform "traditional" observations more rapidly and systematically than in the past. In contrast to this, there are some entirely new approaches emerged in the last few years, which have been made possible by the combination of powerful computers and high-quality digital images.

All methods require the gathering of information from selected corpora of pre-processed digital images, aimed at the generation of quantitative measurements: these will contribute to the creation of a statistical profile of each sample, to be used for finding similarities and differences between writing styles and individual hands. In the treatment of the graphic sequence, possible methodologies range from the automatic recognition and characterisation of single words and signs, to the reduction of the *ductus* to its basic profile, to the extraction of a more global set of "texture" features, depending on the detection of recurrent forms on the surface of the page.

However promising, all these approaches haven't yet produced widely accepted results, both because of the immaturity in the use of these new technologies, and of the lack of real interdisciplinary research: palaeographers often missing a proper understanding of rather complex procedures, and scientists being unaware of the specificity of medieval writing and tending to extrapolate software and methods already developed for modern writings. However, the results of "digital palaeography" look promising and ought to be further developed. This is particularly true for what concerns "extra-graphic" features, such as those related to the layout of the page, which are more easily extracted and quantified. For instance, in the case of highly standardized handwriting and book typologies, the comparison of some basic layout features, regarding the organization of the page and its exploitation by the scribe, may give precious clues for distinguishing very similar hands even without recourse to palaeographical analysis.

Moving from these considerations, we propose a pattern recognition system for distinguishing the different scribes who have worked together to the transcription of a single medieval book. The proposed system is based on the use of specifically devised features, directly derived from the analysis of the layout of the page, and performs classification by using a standard Multi Layer Perceptron (MLP) network, trained with the Back Propagation (BP) algorithm [4]. It is worth noticing that the main goal of our study is to verify that the use of such features allows obtaining satisfactory results, rather then that of building a top performing recognition system.

A particularly favourable situation to test the effectiveness of this approach is represented by the so-called "Giant Bibles", a hundred or more "serially produced" Latin manuscripts each containing the whole sacred text in a single volume of very large size (up to 600 x 400 mm and over). The Bibles originated in Central Italy (initially in Rome) in the mid-11th century, as part of the political programme of the "Gregorian Reform", dealing with the moral integrity and independence of the clergy and its relation to the Holy Roman Emperor. 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", which was written in Italy by at least nine scribes within the third decade of the 12h 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 [5]. 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 scribal hands.

The remainder of the paper is organized as follows: Section 2 presents the architecture of the proposed system, while in Section 3 the experimental results are illustrated and discussed.

II. THE SYSTEM ARCHITECTURE

The architecture of the proposed system is illustrated in Fig. 1. The proposed system receives as input RGB images of single pages of the manuscript to be processed, and performs for each page the following steps: *pre-processing*, *segmentation*, *feature extraction*, and *scribe distinction*. These steps are detailed in the following subsections.

A. Pre-processing and Segmentation

In the pre-processing step noisy pixels, such as those corresponding to stains or holes onto the page or those included in the frame of the image, are detected and removed. Red out-scaling capital letters are also removed since they might be all written by a single scribe, specialized for this task. Finally, the RGB image is transformed into a grey level one and then in a binary black and white image. In the segmentation step, columns and rows in each page are detected. The Bible we have studied is a two column manuscript, with each column composed by a slightly variable number of rows. The detection of both columns and rows is performed by computing pixel projection histograms on the horizontal and the vertical axis, respectively.

B. Feature Extraction

The feature extraction step is the most relevant and original part of our work and it has been developed in collaboration with experts in palaeography. We have considered three main sets of features, mainly concerning the layout of the page. The first set relates to properties of the whole page and includes the upper margin and the lower margin of the page and the intercolumn. Such features are not very distinctive for an individual copyist, but they may be very useful to highlight chronological and/or typological differences. The second set of features concerns the columns: we have considered the number of rows in the column and the column exploitation coefficient [6]. The exploitation coefficient is a measure of how much the column is filled with ink, and may be is computed as:

$$ExploitationCoefficient = \frac{N_{BP}(C)}{N_p(C)} \qquad (1)$$

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 writer. In the case of very standardized handwritings, such as the "carolingian minuscule" shown by the Bible of Avila, the regularity in the values assumed by such features may be considered as a measure of the skill of the writer and may be very helpful for scribe distinction. The third set of features characterizes the rows, and includes the following features: weight, modular ratio, interline spacing, modular ratio/interline spacing ratio and peaks. The weight is the analogous of the exploitation coefficient applied to rows, i.e. it is a measure of how much a row is filled with ink. It is computed as in equation (1) but considering row pixels instead of column pixels. The modular ratio is a typical palaeographic feature, which estimates the dimension of handwriting characters. According to our definition, this feature is computed for each row measuring the height of the "centre zone" of the words in that row. Once the cen-

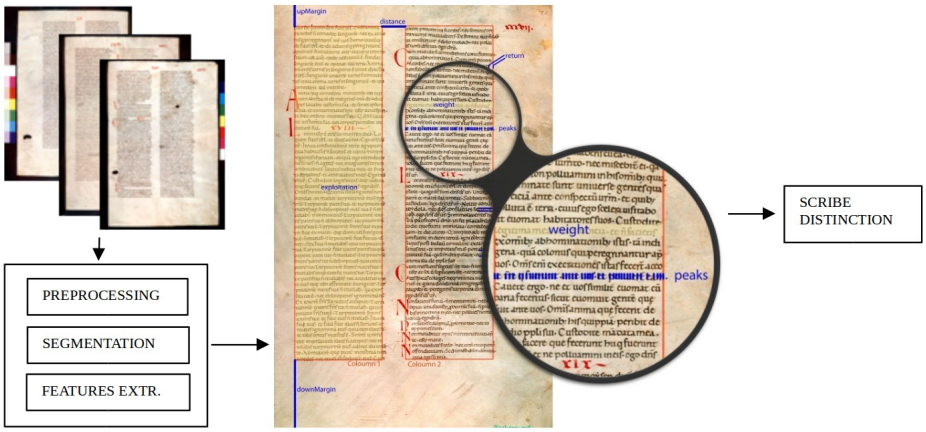

Fig. 1. The architecture of the system.

tre zone has been estimated, the interline spacing is the distance in pixels between two rows. Modulus, interline spacing and modular ratio/interline spacing ratio characterize not only the way of writing of a single scribe, but may also hint to geographical and/or chronological distinctions. In [7], for instance, the distance among layout lines in rows and the dimension of letters significantly differentiate Spanish and Italian minuscule. Highly discriminating features, such as the inter-character space and the number of characters in a row, imply the very difficult task of extracting the single characters contained in each word, which is far to be solved in the general case. Therefore, we have chosen to estimate the number of characters in a row by counting the number of peaks in the horizontal projection histogram of that row.

C. *Scribe Distinction*

The last block performs the recognition task, which has the effect of identifying the rows in each page written by the same copyist. In our study, we have assumed that the manuscript has been produced by N different copyists, previously identified through the traditional palaeographical analysis. We have also 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 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. Summarizing, each pattern is represented by a feature vector containing 10 values. Finally, each pattern is attributed to one of the N copyist by using a Neural Network classifier: in particular we used a MLP trained with the Back Propagation algorithm [4]. We chose such a classifier because of both its effectiveness and its simplicity.

III. 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". The palaeographic analysis of such a manuscript has individuated the presence of 13 scribal hands. Since the rubricated letters might be all the work of a single scribe, they have been removed during the pre-processing step; we have therefore considered only 12 copyists to be identified. The pages written by each copyist are not the equally numerous and there are cases in which parts of the same page are written by different copyists.

The aim of the classification step is that of associating each pattern, corresponding to a group of M consecutive rows, to one of the $N = 12$ copyists: in our experiments we have assumed $M = 4$, thus obtaining a database of 20867 samples extracted from the set of the 800 pages which are in two column format (the total number of pages in the Bible is 870). The database has been normalized (by using Z-normalization) and divided in two subsets: the first one, containing 10430 samples, has been used as training set for the neural network classifier, while the second one, containing the remaining 10437 samples, has been used for testing the system. The training set 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 training and test set. Preliminary experiments have been performed for setting MLP parameter values: in particular, we have obtained the best results with 100 hidden neurons and 1000 learning cycles.

The performance achieved by our system on training and test set is 94.16% and 91.51%, respectively. These results are very interesting since they have been obtained by considering only page layout features, without using more complex information relative to the shape of each sign: such information would be typically analyzed by palaeographers, but the process for automatically extracting them

Table 1. The confusion matrix of the 12 copyists computed on the test set; the second column reports the number of samples for each copyist

Scribe	#Sam.	A	B	C	D	E	F	G	H	I	W	X	Y
A	4286	**97.30**	0.0	0.0	0.10	0.30	1.20	0.50	0.40	0.10	0.20	0.0	0.0
B	5	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
C	103	13.60	5.80	**77.70**	0.0	0.0	0.0	0.0	0.0	0.0	1.00	1.90	0.0
D	353	7.10	0.60	2.00	**74.20**	11.10	0.90	0.30	2.30	0.30	0.0	1.10	0.0
E	1095	4.20	0.0	0.0	0.20	**91.80**	1.20	0.30	0.30	0.70	0.50	0.90	0.0
F	1962	9.60	0.30	0.10	0.0	0.70	**88.30**	0.40	0.40	0.0	0.0	0.20	0.20
G	447	4.50	0.0	0.0	0.0	1.30	2.20	**86.80**	0.50	0.0	0.70	0.50	3.40
H	520	12.70	1.00	1.00	0.40	4.00	6.20	0.80	**72.70**	0.60	0.0	0.40	0.20
I	832	2.50	0.0	1.30	0.0	0.10	0.10	0.0	0.0	**94.80**	0.80	0.10	0.0
W	45	6.70	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**57.80**	33.30	0.0
X	522	1.70	0.0	0.0	0.20	2.10	0.0	0.0	0.0	0.20	0.0	**94.60**	1.20
Y	267	1.90	0.80	0.0	3.80	0.80	1.10	0.0	0.0	0.0	3.00	5.20	**83.20**

from the original images is very complex and not easy to generalize. Table 1 reports the confusion matrix of the 12 copyists obtained on the test set; the classification results for each copyist are bolded. The data in the table show that the worst performance is obtained for copyists represented by a reduced number of samples (the number of samples for each copyist is reported in the second column): in these cases, in fact, it is difficult to adequately train the MPL classifier. This happens, for instance, for the copyists B and W. In particular, the copyist B, for which only 5 samples may be included in the training set, is completely confused with the copyist W (belonging to a completely different period and context). On the contrary, the best performance is obtained for the copyist A, which has the highest number of samples in the training set.

On the basis of these preliminary results, the proposed system seems to be very promising and may constitute an effective first classification step. In fact, exploiting the information about the classification reliability, which are also provided by the MLP classifier (but not used in this work), palaeographers may find further confirmation of their hypothesis or concentrate their attention on those sections of the manuscript which have not been reliably classified on the basis of layout features. These topics will be addressed in a future work.

REFERENCES

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

[2] Ciula, A., "The Palaeographical Method Under the Light of a Digital Approach", in Kodikologie und Paläographie im digitalen Zeitalter / Codicology and Palaeography in the Digital Age, ed. by M. Rehbein, P. Sahle, T. Schaßan, in coll. with B. Assmann, F. Fischer, Ch. Fritze, Norderstedt, pp. 219-236, 2009.

[3] M. Gurrado, "graphoshop, uno strumento informatico per l'analisi paleografica quantitativa". In: Codicology and Palaeography in the Digital Age. pp. 251–259, 2009. (2009)

[4] Rumelhart, D. E., Hinton, G. E. and Williams, R. J., " Learning representations by back-propagating errors", Nature, 323:533?536, 1986.

[5] Maniaci, M. and Orofino, G., "Prime considerazioni sulla genesi e la storia della Bibbia di Ávila", in Miscellanea F. Magistrale, Spoleto, 2010.

[6] Bozzolo, C., Coq, D., Muzerelle,D., Ornato, E., "Noir et blanc. Premies résultats d?une enquête sur la mise en page dans le livre médiéval", Proc. of Il libro e il testo, Urbino, pp. 195-221, 1982.

[7] Maniaci, M. and Orofino, G., "Dieci anni di Bibbie atlantiche a Cassino", in Les Bibles atlantiques. Le manuscrit biblique à l?époque de la réforme ecclésiastique du XIe siècle, éd. par N. Togni, 2011.