A Method for Scribe Distinction in Medieval Manuscripts Using Page Layout Features

Claudio De Stefano[1], Francesco Fontanella[1],
Marilena Maniaci[2], and Alessandra Scotto di Freca[1]

[1] Dipartimento di Automazione, Elettromagnetismo, Ingegneria dell'Informazione e Matematica Industriale
Via G. Di Biasio, 43 – 03043 Cassino (FR) – Italy
[2] Dipartimento di Filologia e Storia
Via Zamosch, 43 – 03043 Cassino (FR) – Italy
University of Cassino
{destefano,fontanella,mmaniaci,a.scotto}@unicas.it

Abstract. In the framework of Palaeography, the use of digital image processing techniques has received increasing attention in recent years, resulting in a new research field commonly denoted as "digital palaeography". In such a field, a key role is played by both pattern recognition and feature extraction methods, which provide quantitative arguments for supporting expert deductions. In this paper, 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 Bible"), we wished to verify if the extraction of certain specifically devised features, concerning the layout of the page, allowed to obtain satisfactory results. To this aim, we have also performed a statistical analysis of the considered features in order to characterize their discriminant power. 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.

1 Introduction

In the context of palaeographic studies, there has been in the last years a growing scientific interest in the use of computer-based techniques of analysis, whose aim is that of providing new and more objective ways of characterizing medieval handwritings and distinguishing between scribal hands [3,4]. The application of such techniques, originally developed in the field of forensic analysis, originated 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, for instance for the evaluation of parameters such as the angle and the width of strokes, or the comparison among digital examples of letter-forms. In the above mentioned cases, technology is employed to perform "traditional" observations more rapidly and systematically than in the past. In contrast to this,

G. Maino and G.L. Foresti (Eds.): ICIAP 2011, Part I, LNCS 6978, pp. 393–402, 2011.

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. These new approaches include the development of systems for supporting the experts' decisions during the analysis of ancient handwritings.

All the above approaches 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 characterization 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 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. 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 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 considers a set of features typically used by palaeographers, which are directly derived from the analysis of the page layout. The classification is performed by using a standard Multi Layer Perceptron (MLP) network, trained with the Back Propagation algorithm [10]. We have chosen MLP classifiers for two main reasons: on the one hand MLP classifiers are very simple, quite effective and exhibit a good generalization capability; on the other hand the main goal of our study is not that of building a top performing recognition system, but rather to verify that the use of page layout features allows obtaining satisfactory results. Finally, we have also performed a statistical analysis of the considered features in order to characterize the discriminating power of each of them. The results reported in Section 4 confirmed that the proposed method allowed us to select the feature subset which maximizes classification results.

A particularly favorable situation to test the effectiveness of this 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 x 400 mm and over). The Bibles originated

in Central Italy (initially in Rome) in the mid-11th century, as part of the political program 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 [8]. 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 system, Section 3 illustrates the method used for feature analysis, while in Section 4 the experimental results are illustrated and discussed. Finally, Section 5 is devoted to some conclusions.

2 The System Architecture

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*.

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.

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 and following the suggestion reported in [11]. 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 intercolumnar distance. 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

Fig. 1. The number of peaks in the horizontal projection histogram of a row

exploitation coefficient [2]. The exploitation coefficient is a measure of how much the column is filled with ink, and 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 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, interlinear spacing, modular ratio/interlinear 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 (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 centre zone has been estimated, the interlinear spacing is the distance in pixels between two rows. Modular ratio, interlinear spacing and modular ratio/interlinear spacing ratio characterize not only the way of writing of a single scribe, but may also hint to geographical and/or chronological distinctions. In [8], 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 (see Fig. 1). The whole set of considered features is summarized in Table 1 reporting, for each of them, the associated identification number.

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

Table 1. The considered features and the corresponding identification number (id)

1	intercolumnar distance	6	modular ratio
2	upper margin	7	interlinear spacing
3	lower margin	8	weight
4	exploitation	9	peak number
5	row number	10	modular ratio/interlinear spacing

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 [10].

3 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 measure which evaluates the effectiveness 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 [7], Relief [6], Gain ratio, Information Gain and Symmetrical uncertainty [5].

The *Chi-Square* measure estimates feature merit by using a discretization algorithm: if a feature can be discretized to a single value, then it can safely be removed from the data. The discretization algorithm, adopts a supervised heuristic method based on the χ^2 statistic. The range of values of each feature is initially discretized by considering a certain number of intervals (heuristically determined). Then, the χ^2 statistic is used to determine whether the relative frequencies of the classes in adjacent intervals are similar enough to justify the merging of such intervals. The formula for computing the χ^2 value for two adjacent intervals is the following:

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

where C is the number of classes, A_{ij} is the number of instances of the j-th class in the i-th 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 i-th interval and C_j and N_T are the number of instances of the j-th class and total number of instances, respectively, in both intervals. The extent of the merging process is controlled by a threshold, whose value represent the maximum admissible difference among the occurrence frequencies of the samples in adjacent intervals. The value of this threshold has been heuristically set during preliminary experiments.

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 $\{x_1, x_2, \ldots, x_n\}$, its entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2(x_i) \tag{3}$$

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:

$$H(X|Y) = -\sum_{i,j} p(x_i, y_j) \log \frac{p(y_j)}{p(x_i, y_j)} \tag{4}$$

where $p(x_i, y_j)$ is the probability that $X = x_i$ and $Y = y_j$. The quantity in (4) 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 [9]:

$$I_G = H(C) - H(C|X) \tag{5}$$

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 uses the information gain defined in (5). 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:

$$I_R = \frac{I_G}{H(X)}. \tag{6}$$

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 $[0, 1]$:

$$I_S = 2.0 \times \frac{I_G}{H(C) + H(X)} \tag{7}$$

4 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 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 the Z-normalization method[1], 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. 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 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 accuracy achieved by using the whole set of features on training and test set, averaged over 20 runs, is 95.57% and 92.46% 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 from the original images is very complex and not easy to generalize.

Table 2 reports the recognition rates obtained on the test set for each of the 12 copyists, together with the corresponding number of samples. The third and the fourth row of the table respectively report the average recognition rate and the variance obtained for each scribe over the 20 runs. Similarly, the fifth and the sixth row report the best and the worst recognition rate, respectively, over the 20 runs. 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 row): in these cases, in fact, it is difficult to adequately train the MLP classifier. This happens, for instance, for the copyists B and W. In particular, the copyist B, for which only 5 samples are included in the training set, is completely confused with other copyists represented by a

[1] Note that the Z-normalization transforms the distribution of the original data into standard normal distribution (the mean is 0.0 and standard deviation is 1.0).

Table 2. Test accuracies obtained for each scribe

Scribe	A	B	C	D	E	F	G	H	I	W	X	Y
Samples	4286	5	103	352	1095	1961	446	519	831	44	522	266
Av. Acc.	97.10	24.00	65.60	83.50	93.00	90.40	85.90	83.40	94.80	49.90	93.70	79.90
Variance	0.01	16.71	0.86	0.61	0.03	0.04	0.14	0.05	0.00	5.46	0.03	0.03
max Acc.	99.00	100.00	87.00	98.00	96.00	94.00	91.00	86.00	95.00	100.00	96.00	84.00
min Acc.	95.00	0.00	57.00	69.00	90.00	88.00	82.00	80.00	94.00	27.00	90.00	78.00

higher number of samples. On the contrary, the best performance is obtained for the copyist A, which has the highest number of samples in the training set.

Further experiments have been performed in order to evaluate the discriminant power of each feature and to find the feature subset which maximizes classification results. As discussed in Section 3, we have considered five univariate measures, each providing a ranked list of the considered features. The results of this analysis are reported in Table 3, which shows the ranking relative to each measures. Although the different measures produced quite different results, they give a good insight about the best and worst features.

In order to compute the overall ranking of the features, we used the Borda count rule [1]. According to such rule, the overall score of each feature is obtained by using the formula:

$$Os_i = \sum_{j=1}^{5} 10 - \mathrm{pos}_{ij} \tag{8}$$

where Os_i is the overall score of the i–th feature, while pos_{ij} is the position of the i–th feature in the j–ranking. Table 4 displays the overall ranking of the features obtained by using the Borda count rule.

Figure 2 shows the plot of the test set accuracy as a function of the number of features: for each number of features n_i, the first n_i features in the overall ranking have been considered, and 20 runs have been performed. Note that the most right bar refers to the results obtained considering the whole feature set. In the plot, for each number of features, the first and the third bar report the worst and the best recognition rate, respectively, while the second one reports the average

Table 3. Feature ranking according to the five considered measures. For each row, the most left numeric value indicates the best feature, while the most right value denotes the worst one.

Measure	Ranking
Chi Squared (C_S)	4 3 2 1 5 9 7 6 10 8
Relief (R_F)	5 4 1 9 3 7 6 10 8 2
Gain Ratio (I_R)	4 5 1 3 2 9 7 6 10 8
Information Gain (I_G)	4 3 2 1 5 9 7 6 10 8
Symmetrical Uncertainty (I_S)	4 3 5 1 2 9 7 6 10 8

Table 4. Overall ranking of the features

id	feature	score
4	exploitation	44
3	lower margin	35
5	row number	34
1	intercolumnar distance	32
2	upper margin	24
9	peak number	22
7	interlinear spacing	16
8	weight	16
6	modular ratio	11
10	modular ratio/interlinear spacing	6

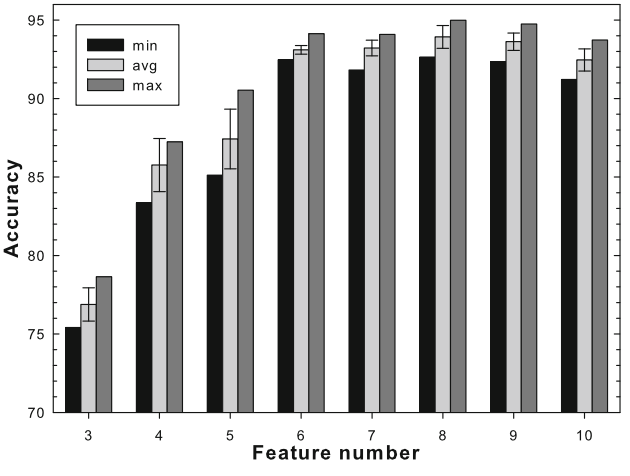

Fig. 2. Test set accuracy (averaged over 20 runs) vs feature number

recognition rate together with the corresponding variance. The results obtained by using one feature and two features have been omitted because they are too low (less than 60%). The data in the plot show that satisfactory results can be obtained considering at least the first six features in the overall ranking, while the best result has been obtained by using the first 8 features (95%), i.e. discarding the features representing the modular ratio and the "modular ratio/interlinear spacing". It is worth noting that while the performance difference between the use of eight and nine features is very small, the performance difference between the use of nine and ten features is more relevant. This means that the feature "modular ratio/interlinear spacing" is misleading.

5 Conclusion

We presented a novel approach for automatic scribe identification in medieval manuscripts. The task has been accomplished considering features suggested by paleograpy experts, which are directly derived from the analysis of the page layout. The experimental investigation has regarded two main aspects. The first one was intended to test the effectiveness of the considered features in discriminating different scribes. The second one was aimed at characterizing the discriminant power of each feature in order to find the best feature subset. The experimental results confirmed the effectiveness of the proposed approach.

Future work will include exploiting the information about the classification reliability. Such kind of information would allow palaeographers to find further confirmation of their hypothesis and to concentrate their attention on those sections of the manuscript which have not been reliably classified.

References

1. Black, D.: The Theory of Committees and Elections, 2nd edn. Cambridge University Press, London (1963)
2. Bozzolo, C., Coq, D., Muzerelle, D., Ornato, E.: 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 (1982)
3. 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, pp. 219–237 (2009)
4. Gurrado, M.: "graphoshop", 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, pp. 251–259 (2009)
5. Hall, M.: Correlation-based Feature Selection for Machine Learning. Ph.D. thesis, University of Waikato (1999)
6. Kononenko, I.: Estimating Attributes: Analysis and Extensions of RELIEF. In: European Conference on Machine Learning, pp. 171–182 (1994)
7. Liu, H., Setiono, R.: Chi2: Feature Selection and Discretization of Numeric Attributes. In: ICTAI, pp. 88–91. IEEE Computer Society, Los Alamitos (1995)
8. Maniaci, M., Ornato, G.: Prime considerazioni sulla genesi e la storia della bibbia di avila. Miscellanea F. Magistrale, Spoleto (2010) (in press)
9. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Series in Machine Learning. Morgan Kaufmann, San Francisco (1993)
10. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature 323(9), 533–536 (1986)
11. Stokes, P.: Computer-aided palaeography, present and future. In: Rehbein, M., Sahle, P., Schaßan, T. (eds.) Kodikologie und Paläographie im digitalen Zeitalter / Codicology and Palaeography in the Digital Age. pp. 309–338 (2009)