Engineering Applications of Artificial Intelligence 72 (2018) 99–110

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
0952-1976/© 2018 Elsevier Ltd. All rights reserved.

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

Download English Version:

https://daneshyari.com/en/article/6854171

Download Persian Version:

https://daneshyari.com/article/6854171

Daneshyari.com