A STUDY ON THE WRITER IDENTIFICATION TASK FOR PALEOGRAPHIC DOCUMENT ANALYSIS

Fabio Aiolli
DMPA, University of Padova, ITALY
Via Trieste, 63 - Padova - Italy
email: aiolli@math.unipd.it

Manuel Giollo
DMPA, University of Padova, ITALY
Via Trieste, 63 - Padova - Italy
email: manuel.giollo@gmail.com

ABSTRACT
The subject of paleography is the study of ancient documents. In particular, the paleographer's aim is to locate a document in a cultural environment and chronological interval in the past. Automatic writer identification is then a desirable tool for a paleographer as she/he gains useful information about the document at hand. However, the paleographer is often interested in methods which can be easily interpretable by humans. In this paper, we apply some state-of-the-art techniques devised for modern documents to the paleographic domain. Moreover, we propose new techniques and document representations with the aim at producing more understandable representation of a writing style. Experimental results have been performed on a large dataset of paleographic images and demonstrate the feasibility of the proposed approach, and the suitability of this tool on helping the paleographer's work.

KEY WORDS
Writer identification, medieval document analysis, pattern recognition, intelligent data analysis, biometrics.

1 Introduction

The main goal of a paleographer is to locate handwritten material, being it entire books or single scripts, within a particular chronological interval and within a specific cultural environment. Consequently, a very important problem in paleographic studies is the identification of the author of a given manuscript.

Among the few different tools available for paleographic analysis, in recent years, the software SPI (System for Paleographic Inspections) [1, 2] has been successfully exploited by experts for their job. The aim of that work was to give the paleographer a tool to make morphological comparisons and analysis of paleographic documents possible. This kind of analysis can help a paleographer to make a guess about the provenance of documents once the same information is known for similar documents. The assumption here is that morphologically similar documents will have a similar cultural environment and localization. This clearly gives some further information about a manuscript that can be useful to identify its writer as well. However, a more direct approach, which tries to learn typical writing styles of different authors, would be desirable and this is the subject of the present work.

In the last decade several works tried to study handwritten text with the aim at identifying the writer's identity, both for security and forensic purposes. This has been done for a variety of document types, using different alphabets and languages, including Hebrew and Chinese documents. However, much less work has been devoted to ancient Latin manuscripts.

The above-mentioned tasks can be cast in the so called *offline writer identification* task as dynamic features are not available (as opposed to *online writer identification* where dynamic features on the stroke are present at identification time). Two different approaches to the writer identification task can be followed. In *text-independent* methods, the whole text image, without any prior knowledge about the script content, is analyzed, while in *text-dependent* techniques such knowledge (sign verification is a classical task in this sense), is required. Clearly, text independent methods have many advantages, as different types of manuscripts can be promptly analyzed with the same techniques, while text dependent methods often need of a textual transcription and/or an OCR method to extract the text contained in the document.

Another important aspect of the paleographic domain is the crucial importance of the interpretability of the identification process. In fact, while in security and forensic domains the effectiveness of a method is the most important thing, in the paleographic domain the way these results have been obtained is also of great importance. Paleographers cannot be satisfied by a black box method even when it has ideal effectiveness. These domain experts generally prefer tools able to formally support their expertise in the field.

The contribution of this paper is two-fold. First of all, we studied the applicability of state-of-the-art text independent writer identification techniques to the paleographic domain. In particular, we focus on *grid microstructure features* [3] for document representation. The effectiveness and robustness of these techniques seem confirmed by our experiments but the interpretability of the obtained representations turns out to be not satisfying for the domain at hand. To overcome this problem, we propose two additional hybrid methods which borrow some character-level information ideas to improve the interpretability of the results. Experimental comparisons between ours and state-of-the-art approaches on a large paleographic dataset demonstrate that our method shows comparable effectiveness and robustness while improving the interpretability of the respective representation.

2 Related Work on Manuscript Representation

A common problem in Pattern Recognition is to get a good representation of the input data, which is crucial in the following recognition/decision step. Two main types of representations are used in a text-independent framework for document analysis. They differ for the type of information they use, *Character-level* or *Texture-level* information. Both of them have proven reasonable effectiveness in modern document analysis.

Character-level information Characters written by different authors differ largely by their shape in scripts. Since letters are quite simple symbols, it is possible to draw all their typical forms, and create a codebook C of known allographs: in [4] it is proposed to create an histogram which describes the probability of appearance of every $c \in C$. Clearly, this information is intuitably discriminative, but requires the capability of separating different characters in the manuscript, which is an hard task. Moreover, the typical allographs depend on the language (Latin letters differ from Japanese ones) and from the style (modern writings are quite different from medieval scripts), then some a-priori knowledge about text images is needed indeed. Works done in [5], [6] and [7], for instance, try to locate precise characters by means of *structuring elements* and OCR respectively, and extract simple information (like area ratio, center of gravity, concavity, etc) from the letters.

Texture-level information In this case, the document image is analyzed in its wholeness, and characteristic information are extracted by means of frequency analysis or by domain dependant feature extractors. Frequency analysis is a general purpose technique in the image processing task: its application in the writer identification context was firstly introduced in [8]. The effectiveness of the method has laid the foundations for further works, as in [9], [10], [11], [12], [13]. Also Gabor wavelet, contourlet, complex wavelet, curvelets and Gabor filters were tested (more generally, any spatially located frequency analysis could work) as feature extractors, and appreciable results were always achieved. However, two main disadvantages afflict these approaches, i.e. the gained representations are meaningless for humans and they need of a prior normalization on the document image. On the other hand, domain dependent information can be measured, like text slantness, characters' size (by means of run-length), and letters' "roundness" (using, for example, the contour-hinge [4]). Besides the human understandability of these information, experimental results show that such features have an high discriminative power.

We mainly focus on this second type of measures. In particular, the family of *grid microstructure feature* proposed in [3] has been chosen, due to their effectiveness on earlier experimental work. Despite their simplicity (they only consider pairs or groups of three pixels), it has been demonstrated that these techniques are able to detect important information of the writing style. Character-level features have not been considered as they require a difficult character segmentation step, which would be quite ineffective in medieval documents, due to the high variability in their writing styles.
In the following sections, we also propose a new method which produces an hybrid representation standing in the middle between character and texture level type of information. Specifically, we propose to use a clustering algorithm to produce a set S of prototypes of $k \times k$ contour windows. We hence attempt to measure with an histogram the so called *shapes distribution (SD)*, which is the probability of use of a given $s \in S$ in an incoming document. A generalization of this technique, the *Shape Cooccurrence Distribution (SCD)*, is also considered.

Text dependent methods Since character level information are the most discriminative, several works have used OCR techniques to extract the text content. Hidden Markov Models was applied by [6] to obtain writer identification: for each considered writer, an individual HMM based handwriting recognition system is trained using data from that writer. Input lines are first normalized with respect to slant, skew, baseline location and height. In [1], character shapes of a given letter are extracted from the document image by means of letter segmentation techniques. Then, such shapes are directly compared to the "typical" character's shape of each known writer via tangent distance measurement (which compute a set of affine transformations among the two shapes in order to maximize their likelihood). In [7] character specific features are used: OCR locates letters and, for each one, gradient, structural and concavity information is computed. Such vector representation is used to compare the same characters by means of Euclidean distance. A similar idea is also described in [14], where chains of aligned pixels in the scripts are analyzed by means of run lengths observed along 8 different (quantized) directions. The result is a vector of 64 features, used as a sign for a given writer, which characterize the typical slant in the text. For each direction, 8 features describe the frequency distribution of different stroke's length. Finally, in [15] similar features have been used with a Fuzzy C-mean Clustering as classifier.

3 Text-independent statistical features

Three main activities, namely, preprocessing, feature extraction, and feature comparison, are involved in the whole writer identification process.

Preprocessing In this step, input images are normalized: since an uncontrollable variability could affect data, it is important to reduce the source of uncertainty. Many noise reduction methods can be employed in document and image preprocessing, including smoothing and spot removals [16], seeping ink elimination [17], or binarization [18]. Moreover, even the text itself can be modified, and characters' skewness and slantness of text lines [19] can be removed.

In the present work, we apply edge extraction techniques, only. For this, several methods are available, like the Moore's contour-following algorithm, which has been used in [4], the Sobel operator, which has been exploited in [3], [15], [7], or the Canny operator. The mentioned approaches are quite similar, and all of them use derivative information of the image itself to detect shapes borders. Additional considerations about these edge extractors will be reported in the next section.

Features extraction In this step, "relevant" information of the input data (typically the edges of an image) is detected: the aim is to capture a representation which should describe the most characterizing peculiarity in a document (its writing style) in a simple way. For example, if we want to infer the writer identity, it could be useful to locate atypical characters or analyze text skewness. In our work we extract the C1, C2 and C3 *Grid microstructure features*, and both *shapes distribution (SD)* and *Shape Cooccurrence Distribution (SCD)* representations. In the next section, such techniques will be more formally explained.

Features comparison Once the description of a document in terms of its representation has been built, it is possible to compare a query document with a set of preclassified documents on the basis of a distance between their representations. It is assumed that, if the representations $r_i, r_j \in R$ are very similar, then r_i, r_j are extracted from document written by the same author. Comparison are typically performed using Euclidean distance or the χ^2 distance, which is defined as follow:

$$\chi^2(q, r_n) = \sum_k \frac{(q^{(k)} - r_n^{(k)})^2}{(q^{(k)} + r_n^{(k)})}. \qquad (1)$$

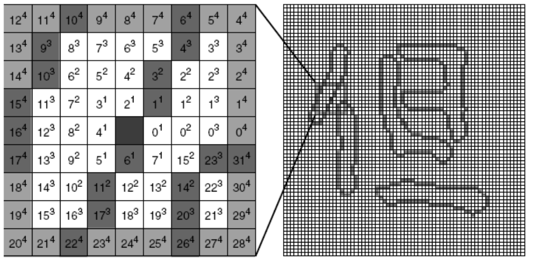
Here q is the vector representation of the query document and r_n are vector representations of known writers. In addition, two improved versions of these metrics can also be considered. They can be obtained by simply adding $\sigma^{(k)}$ (the standard deviation) in the denominator of the respective formulas. These distance measures are typically used in writer identification, and have proven their effectiveness in several works (see for example [4] [3]).

4 Features extraction for the analysis of writing styles

Since the goal of this paper is to inspect the effectiveness of different writing style representations, we will focus on the feature extraction step only. Future work will be devoted to both preprocessing and features comparison techniques. In the following, we present the implemented grid microstructure features.

4.1 Grid microstructure feature based representation

In this section we introduce the *Grid microstructure feature* [3] which represents the state-of-the-art on modern document writing style representation. This family of features will be tested against medieval documents in the experimental section. This kind of features can be obtained as follows. Edge information are retrieved (by means of an edge operator, like the Sobel one) and used to build a Probability Distribution Function (PDF) called *grid microstructure features*. As shown in Figure 1, this information is computed by means of a floating grid of fixed size (13x13 in our implementation). During the feature extraction, the center square of the grid traverse all the edge pixels. At every position in such a grid is associated an univocal symbol i_m, where m is the bigger distance in the horizontal and vertical distance between the square and the center (Figure 1 exhibits the i_m values of each square). As long as the grid moves on, related pairs of pixels $\langle i_m, j_l \rangle$ are stored in an accumulator $h(i_m, j_l)$, which increases by one its value when a relationship is met. Three kinds of relationships between pairs of pixels are managed:



12^4	11^4	10^4	9^4	8^4	7^4	6^4	5^4	4^4
13^4	9^3	8^3	7^3	6^3	5^3	4^3	3^3	3^4
14^4	10^3	6^2	5^2	4^2	3^2	2^2	2^3	2^4
15^4	11^3	7^2	3^1	2^1	1^1	1^2	1^3	1^4
16^4	12^3	8^2	4^1		0^1	0^2	0^3	0^4
17^4	13^3	9^2	5^1	6^1	7^1	15^2	23^3	31^4
18^4	14^3	10^2	11^2	12^2	13^2	14^2	22^3	30^4
19^4	15^3	16^3	17^3	18^3	19^3	20^3	21^3	29^4
20^4	21^4	22^4	23^4	24^4	25^4	26^4	27^4	28^4

Figure 1. Example of the floating grid lying over the edges of the writings.

C1 consecutive pairs $\langle p_1, p_2 \rangle$ having distance $d_{p_1} = d_{p_2}$ from the center. For example, in Figure 1, $\langle 1_1, 6_1 \rangle$, $\langle 3_2, 11_2 \rangle$, $\langle 11_2, 14_2 \rangle$

C2 pairs $\langle p_1, p_2 \rangle$, belonging to the same edge and having $d_{p_2} - d_{p_1} = 1$. For example, in Figure 1, $\langle 1_1, 3_2 \rangle$, $\langle 6_1, 11_2 \rangle$, $\langle 3_2, 4_3 \rangle$

C3 pairs $\langle p_1, p_2 \rangle$ belonging to the same edge, having $d_{p_2} - d_{p_1} = 2$. For example, in Figure 1, $\langle 1_1, 4_3 \rangle$, $\langle 6_1, 17_3 \rangle$, $\langle 3_2, 6_4 \rangle$

Once stored this relationship in the accumulator $h(i_m, j_l)$, a normalization step provides us the grid microstructure feature.

4.2 Shape distribution based representation

In order to gather additional features from the input images and increase the interpretability of the obtained representations, we developed a new technique. We do not consider pure *Character-level information* as described in Section 2, since it requires either the ability to segment letters, or style and language dependent information in order to recognize the written text. However, allograph-level information [4] are quite interesting, despite their limitations in the proposed implementation. We suggest therefore to focus the analysis on more simple shapes, which can be located without problems, and hence, without the need of character segmentation. Specifically, since letters are made of simple lines (straight lines, curves or sharp corners), we would like to extract the typical strokes used by writers. This is a natural choice for modern cursive writings, but becomes a suitable idea also in medieval documents, if input images are convolved with edge extractor filters. Our aim is then the following: computing, for each "simple shape" (examples can be seen in Figure 2 and 3), its probability of appearance in incoming images. This is what we compute with the *shapes distribution (SD)* feature. Note that, focusing on simple lines and curves should be better since character borders (obtained, for instance, with the Sobel operator) of every language are entirely formed by such type of graphical elements. Moreover, edges in the images summarize the most important information in a small set of relevant features: this is especially true in document image processing, since text is typically binarized, so just black and white pixels can be considered. Again, it is well shown by *image registration* researches that it is simpler to find a match between different images if we just consider *interesting points* (edges in our case) instead of the whole image. On the other hand, considering complex allographs (which in addition account for big blob of ink) makes their detection in the image a burdensome job, and requires an high computational effort to provide a reliable recognition. In order to obtain a good set of "simple shapes", a set S of simple images are chosen to describe the known basic primitives. Such primitive shapes are computed from training document images by resorting to the k-means algorithm on a set of $k \times k$ windows extracted from the training documents and centered on a edge pixel. Using a clustering algorithm seems to be a good choice, since these primitives will represent typical lines in the author's writing style. Such set S is then used to measure the *shapes distribution (SD)* during the document analysis in each new incoming script. Obviously, this set is always the same for each future analysis, so comparison among SDs representation of different images is reasonable and possible. Both during the samples provision to the k-means algorithm and during the following *shapes distribution analysis*, the center of the sliding window (of size $k \times k$) moves only through edge pixels (hence avoiding background pixels). The shapes obtained by this strategy are generally simpler and this makes k-means algorithm output more reliable. In particular, samples given to the clustering algorithm are extracted considering a piece of a document of each author's training documents.

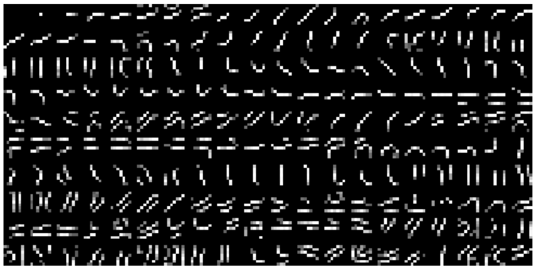


Figure 2. A set of 200 extracted shapes, each one of size 7×7 pixels

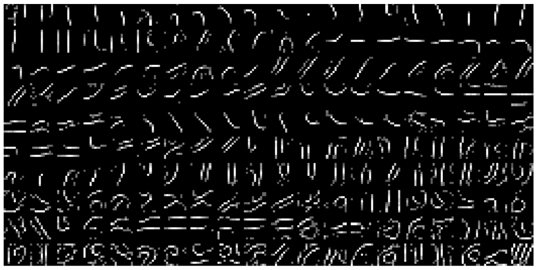


Figure 3. A set of 200 extracted shapes, each one of size 11×11 pixels

SD calculation is very simple: the sliding window extracts all characters borders R from the edge images, and the most similar prototype $s \in S$ w.r.t. any $r \in R$ is found by means of Euclidean distance. For each $s \in S$ the total number of occurrence s_o in R is so stored. At the end, the probability of appearance of a primitive shape can be estimated with the following formula:

$$p(s') = \frac{s'_o}{\sum_{s \in S} s_o} \quad (2)$$

hence giving $\sum_{s \in S} p(s) = 1$ at the end of the analysis. We stress that no character extraction is needed in this process, and from the human point of view, the knowledge of how much these basic lines are used in a certain writing can give us precious and understandable information (differently from what we typically extract from frequency analysis). This idea may seem to be similar to the one exposed in [20], but in our method we don't need either online information (like speed and pressure of the pen) nor shapes simplification.

A further evolution of this analysis is also considered which is inspired by the work in [14] where they tried to infer how text lines are written by just following them as a ball would roll on a groove in the sand. In particular, such information was used to directly compare such script lines in a text-dependent approach. Using the same idea, it is possible to compute shapes cooccurrence probabilities in the writing style of an author. For example, suppose that during the border following of text the primitive shape $s_{actual} \in S$ has been detected: what is the shape $s_{next} \in S$ that will be more likely to appear? This is what we formally denote as $p(s_{next}|s_{actual})$, and represent the so called *Shape Cooccurrence Distribution*. An additional parameter configurable in this estimation is the distance among s_{next} after s_{actual} in the text writing. Specifically, it can be inefficient and unuseful to estimate this correlation if the two shapes are almost identical, as in the case where all the consecutive pixels in the border are taken in account. For this reason, we suggest to update the statistics every 7 pixels found in the border. Shape Cooccurrence Distribution is a powerful information for several reasons: it allows to understand the relationship among different shapes used in writing. Differently from grid microstructure feature and other works described above (which simply consider two or three pixels together), in this way the cooccurrence handle pairs of sliding windows, and thus, complex groups of pixels. In addition, it allows the reconstruction of online information, i.e. movement engrained in the motor memory of the writer: causality of used strokes can therefore be studied and exploited. Moreover, is a generalization of the SD analysis. In fact, $\sum_{s_{next} \in S} p(s_{next}|s_{actual}) = p(s_{actual})$. Finally, this feature is easy to understand, and can be successfully exploited by paleographers.

4.3 Topological information

$C1$ and SCD can be graphically shown to a user. Interestingly, the C1 visualization shows a strong topologically related distribution, as can be seen in Figure 4. Evidently, different writing styles produce extremely diverse peaks in the $C1$ *Probability Distribution Function (PDF)*: at a first glance, even an untrained user can assert that such graphs are very different, due to the sharp valleys and peaks; nevertheless, Euclidean distance, χ^2 distance or other typical distance measures sometimes fail when trying to find the most similar representation of a new document. Of course, such failure comes from the fact that these distances can't opportunely cope with such a topological information, thus more investigation is needed to be done in this direction. Anyway, this is a good starting point: paleographers can exploit such graphical information, and use them as an additional tool for their work. In the same way, SCD visualization exhibit the same property of $C1$, which could provide useful information to an expert examiner (see Figure 5).

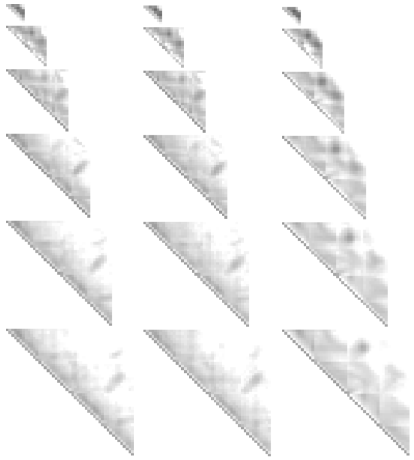


Figure 4. $C1$ representation of the documents in Figure 6. An additional document of the first author is also considered. As can be seen, representation of the same scriber (the former two) are different from the representation of a different scriber (the latter one).

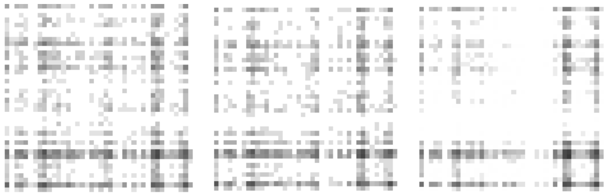


Figure 5. SCD representation of the documents in Figure 6. An additional document of the first author is also considered. As can be seen, representation of the same scriber (the former two) are different from the representation of a different scriber (the latter one)

5 Experiments

In this section, the paleographic dataset, the experimental setting, and the obtained results are presented.

5.1 Paleographic Data

The dataset used in our tests contains images from the "Early Manuscripts at Oxford University" [21], where almost one hundred of colored and high resolution medieval documents are published. For our purpose, for each of the 90 writers, 3 images are used as representative of the writing style of a given scriber. Therefore, 270 images are considered for our tests. Documents are really noisy, and always contains non textual information, due to graphical objects, pictures, copyright details, ruler and tools necessary during the scan, as can be seen in Figure 6. With the purpose of reducing such a noise, we simply have (manually) removed the non textual parts of the image around the scripts. We highlight that we don't correct line slantness, nor stains, infra-lines notes and possible graphical objects inside the text area. It is pretty important to avoid image correction, since real images extracted are full of noise, and we are looking for a fully automated noise-robust system.

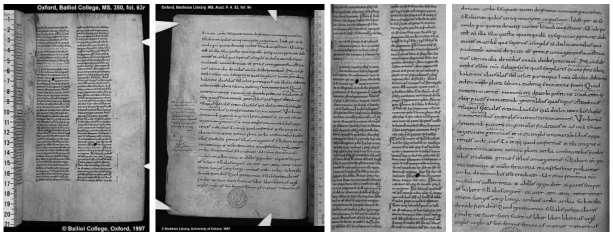


Figure 6. On the left: examples of images before the noise removal. On the right: the respective adjusted images

5.2 Experimental setting and results

Let us start by briefly reporting on preliminary experiments we have done for the tuning of the preprocessing and feature comparison phases. Specifically, quite different performances in the recognition can be observed using different edge extraction methods and/or distance measures. The best accuracy on these preliminary experiments has been achieved using the Sobel filter and the metric χ^2 distance (or its improved version). Canny filter gave worse results since it was prone to locate non existing characters borders, especially in noisy regions. On the contrary, the Sobel operator highlights a smaller set of borders which is more reliable. Concerning feature comparison, the Euclidean distance is resulted too simple to compare high dimensional data, and doesn't take in account the variance of the features.

In a first set of experiments, for each document q in the dataset D the documents $d \in D \setminus \{q\}$ are ranked by similarity with q. We then compute the Top k accuracy, for $k \in \{1, 2, 3, 5, 10\}$ (considering the k most similar documents $d_1, d_2 ... d_k \in D \setminus \{q\}$). In Table 1 (and Figure 7), the accuracy of the ranking produced by predictors based on different representations is shown. In particular, *Committee (Cmt)* represents a ranking system which use $C1$, $C2$, $C3$, SD and SCD classification list to provide more reliable results. The similarity score between a query document and a training document is given as follows. Each ranked document of a given feature has a certain weight in the committee; since top ranked writings should be the most relevant, the following formula has been used:

$$weight_{q,F}(d) = \frac{1}{R_{F,q}(d)} \tag{3}$$

So, the weight of a document d according to a feature F and query document q, follows an hyperbolic reduction bounded to his ranking position $R_{F,q}(d)$. The committee simply sum the $weight_{q,F}(d)$ of each known feature, and use the resulting weights to order the scripts. We observe that Cmt in this case is able to drastically improve on individual rankers.

	Top 1	Top 2	Top 3	Top 5	Top 10
C1	95.9	97.4	97.4	98.5	99.6
C2	93.7	97.8	98.9	99.6	99.6
C3	95.6	97.0	98.2	99.6	99.6
SD	95.9	97.4	98.2	99.6	99.6
SCD	90.0	93.0	96.3	98.9	99.3
Cmt	97.4	98.9	99.3	100	100

Table 1. Performances (in percentage) of each representation type using the χ^2 distance.

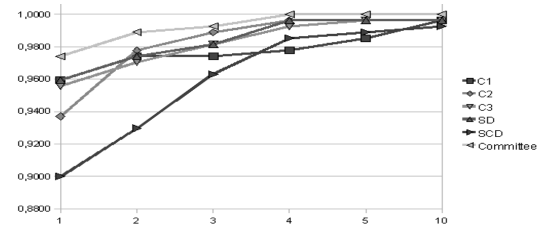


Figure 7. Graphical representation of the obtained results

The number of primitive shapes for SD and SCD has been set to 200 and 40, respectively. Each primitive shape is 7×7 or 11×11 (for SD only) pixels. Moreover, the distance among different sliding windows during SCD computation is set to 7.

In order to check the recognition power of each feature when higher variance in the data is present, we have divided each document in 4, leading to a new dataset (dubbed $D4$ in the following), which is composed of 1080 images. In Figure 8, it can be seen how such division is made. Then, we used $D4_w \subset D4$ as a set of query images, while $D4 \setminus D4_w$ represents the set of known writers examples. In particular, in $D4_w$, only a single quarter of a document for each scribe is present. We report the Top k accuracy results in Table 2.

This very high performance in almost all the cases (actually better than when using the original dataset D) could be misleading. In fact, what it is happening here is that the first retrieved document of a query document q is very often one of the parts of the same original document. In particular this happens in the 84.7% of cases for C1, in the 83.7% of cases for C2 and C3, in the 80.6% of cases for SD, and in the 53% of cases for SCD. Thus, it seems that the methods C1,C2,C3,SD keep also information not properly relevant to the task of identifying the writing style (e.g. the inclination of the writing line) which changes when different documents of the same author are considered. Interest-

Figure 8. Examples of images in $D4$, obtained from the pictures shown in Figure 6.

	Top 1	Top 2	Top 3	Top 5	Top 10
C1	100	100	100	100	100
C2	97.8	98.9	98.9	100	100
C3	96.7	97.8	97.8	98.9	100
SD	97.8	98.9	98.9	100	100
SCD	70.0	76.7	83.3	84.4	90.0
Cmt	100	100	100	100	100

Table 2. Performances (percentage) of each features using χ^2 distance on $D4$.

ingly, SCD seems to show a certain invariance to this kind of misleading information. On the contrary, not surprising, it suffers the size reduction of the training documents as the estimate of the conditioned probability get worse.

In order to make the task harder, additional experiments have been made on the same D4 dataset. This time, for each available author w, two documents of this author $d_{1,w}, d_{2,w} \in D4$, with $d_{1,w} \neq d_{2,w}$, are randomly chosen and inserted in two different dataset. The set $D4_{query} = \bigcup_w d_{1,w}$ represents a set of incoming documents which should be classified, while $D4_{samples} = \bigcup_w d_{2,w}$ represents the known writer/document pairs. Therefore, we calculate the accuracy recognition rate as before. At the aim of obtaining a better estimation of accuracy, several tests like this have been done and results have been averaged. In Table 3, we report the Top k average accuracies for each method.

This last experimental setting exhibits much lower performances and this is consistent with our expectations, due to the higher variance in the data and the reduced size of the training samples. In particular, $C1$ exhibits the best performances, comparable to the one of SD (either when using the 7×7 window features or using the 11×11 window features) suggesting that even training data of reduced size can provide good probability estimation for these representations. Finally, the committee Cmt seems to suffer of the reduced mean performance of individual representations.

Some further consideration can be made on the basis of the previous experiments. As it is possible to see comparing Table 1 and Table 3 the performance of each representation deteriorates when images of reduced size are used. This was expected as the estimate of the statistics identifying the writing style of a single document is less accurate. In general, the more complex is the representation the more deteriorate its performance will be as the size of the training documents get smaller. This is confirmed by the fact that SD11x11 improves on SD7x7. In fact, we can note that the SD representations obtained with window $k \times k$ have a complexity which is inverse proportional to k. This can be explained by observing that when k get larger (keeping the cardinality of prototypes constant) the variance of the images which are assigned to a same prototype increases. Thus, different types of borders can more likely be assigned to the same primitive shape of the representation.

	Top 1	Top 2	Top 3	Top 5	Top 10
C1	61.4	70.0	75.2	82.6	90.7
C2	57.3	65.3	69.4	74.4	81.0
C3	54.7	62.1	65.0	69.3	76.2
SD7x7	59.7	68.0	72.4	77.2	86.1
SD11x11	61.3	69.6	75.4	80.9	88.6
SCD	45.3	55.5	60.3	65.2	72.6
Cmt	57.2	65.3	70.0	76.0	81.8
Cmt2(C1,C2,SD)	61.6	69.6	74.2	78.4	85.2

Table 3. Performances (percentage) of each features using χ^2 distance on $D4_{query}$ and $D4_{samples}$.

Finally, we have performed an analysis of the relationships between different representations. This turns out to be useful to better explain the reduced performances of the committee Cmt and to construct more performant committees. Specifically, for each pair of representations, we have computed the *Kendall Tau* rank correlation coefficient (a correlation measure between two rankings) between the rankings produced by these representations on each iteration of the last experiment we described.

Given two rankings of the sample documents $R_{F1} = R_{F1,q}(S)$ and $R_{F2} = R_{F1,q}(S)$, this correlation is computed as follow:

$$\frac{C - D}{\frac{1}{2}n(n-1)} \tag{4}$$

where $n = |S|$ and C and D represent the number of concordant/discordant order relations on pairs in the two rankings, respectively. It is easy to see that this correlation measure takes values in the interval $[-1, 1]$. Values in $[-1, 0)$ indicates anti-correlation between rankings, while in $(0, 1]$ lay correlated rankings.

As it is possible to observe in Table 4, all the representations are somehow correlated, and this was expected since they all try to solve the same problem; however, $C2$ and $C3$ seems to be quite similar and can be considered redundant. This also was expected, as the definition of C2 and C3 is quite similar. Moreover, $C1$, $C2$ and SD seems to have a noticeable relationship. Instead, SCD turned out to be less similar to the other representations, probably because

	C2	C3	SD	SCD
C1	0.03	0.03	0.05	0.02
C2		0.07	0.03	0.02
C3			0.03	0.02
SD				0.03

Table 4. Correlation among the used representations.

it considers more complex causal relations between them. Regarding the committees in Table 3, we recall that Cmt represents the performances using all the predictors. Motivated by the previous analysis we advocate that a committee constructed using the $C1$, $C2$, and SD, predictions only, should give better results ($C3$ is not considered, due to its similarity with $C2$, while SCD is not included as not sufficiently reliable). This has been confirmed in our experiments (see the results for $Cmt2$ in Table 3).

6 Conclusions

This work is one of the first steps toward reliable writer identification on medieval documents. Differently from previous efforts on medieval texts, our implementation have been tested on a considerable dataset, containing scripts of different languages and writing styles, which provides trustworthiness on the effectiveness of our method. Moreover, results are so promising to suggest that the proposed techniques may possess an high discriminative power even in modern scripts. A notable property of the newly proposed method stands on its easy interpretability which makes it suitable for the paleographic task. Future works will try to establish how excessive noise influences the predictions. Moreover, we think it would be interesting to exploit distance measures different from the Euclidean distance for the comparison of SD/SCD representations. For this, a distance which is invariant to affine transformations, like tangent distance for example, can be useful. Finally, we could try to improve the recognition accuracy by means of character-level information and *features reduction and selection* methods, in order to improve even further the recognition accuracy of the whole system.

References

[1] F. Aiolli, M. Simi, D. Sona, A. Sperduti, A. Starita, and G. Zaccagnini. SPI: A system for paleographic inspections. *Notiziario AI*IA*, 1999.

[2] Fabio Aiolli and Arianna Ciula. A case study on the system for paleographic inspections (SPI): challenges and new developments. In F. Masulli et al., editor, *Computational Intelligence and Bioengineering*, pages 53–66. IOS Press, 2009.

[3] Xin Li and Xiaoqing Ding. Writer identification of chinese handwriting using grid microstructure feature. *Proceedings of the Third International Conference on Advances in Biometrics*, pages 1230–1239, 2009.

[4] Marius Bulacu and Lambert Schomaker. Text-independent writer identification and verification using textural and allographic features. *Transaction on Pattern Analysis and Machine Intelligence*, Volume 29:701–717, 2007.

[5] Itay Bar-Yosef, Isaac Beckman, Klara Kedem, and Itshak Dinstein. Binarization, character extraction, and writer identification of historical hebrew calligraphy documents. *International Journal on Document Analysis and Recognition*, Volume 9:89–99, 2007.

[6] Andreas Schlapbach and Horst Bunke. Off-line handwriting identification using hmm based recognizers. *Proceedings of the 17th International Conference on Pattern Recognition*, Volume 2:654–658, 2004.

[7] Bin Zhang, Sargur N. Srihari, and Sangjik Lee. Individuality of handwritten characters. *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Volume 2:1086, 2003.

[8] T. N. Tan H. E. S. Said, G. S. Peake and K. D. Baker. Writer identification from non-uniformly skewed handwriting images. *Proceedings of the 9th British Machine Vision Conference*, pages 478–487, 1998.

[9] Cong Shen, Xiao-Gang Ruan, and Tian-Lu Mao. Writer identification using gabor wavelet. *Proceedings of the 4th World Congress on Intelligent Control and Automation*, Volume 3:2061–2064, 2002.

[10] Zhenyu He, Yuan Yan Tang, and Xinge You. A contourlet-based method for writer identification. *IEEE International Conference on Systems (2005)*, Volume 1:364–368, 2006.

[11] Da-Yuan Xu, Zhao-Wei Shang, Yuan-Yan Tang, and Bin Fang. Handwriting-based writer identification with complex wavelet transform. *Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition*, pages 597–601, 2008.

[12] Guillaume Joutel, Vronique Eglin, Stphane Bres, and Hubert Emptoz. Curvelets based feature extraction of handwritten shapes for ancient manuscripts classification. *Document recognition and retrieval XIV*, Volume 6500, 2007.

[13] Z. Y. He and Y. Y. Tang. Chinese handwriting-based writer identification by texture analysis. *Proceedings of the Third International Conference on Machine Learning and Cybemetics*, Volume 6:3488–3491, 2004.

[14] Martin Gehrke, Karl-Heinz Steinke, and Robert Dzido. Writer recognition by characters, words and sentences. *International Carnahan Conference*, pages 281–288, 2009.

[15] Soheila Sadeghi Ram and Mohsen Ebrahimi Moghaddam. Text-independent persian writer identification using fuzzy clustering approach. *International Conference on Information Management and Engineering*, pages 654–658, 2009.

[16] D. A. Forsyth and J. Ponce. Computer vision: A modern approach. *Prentice Hall*, 2002.

[17] Chew Lim Tan, Ruini Cao, Qian Wang, and Peiyi Shen. Text extraction from historical handwritten documents by edge detection. *6th international conference on control, automation, robotics and vision*, 2000.

[18] Graham Leedham, Chen Yan, Kalyan Takru, Joie Hadi Nata Tan, and Li Mian. Comparison of some thresholding algorithms for text/background segmentation in difficult document images. *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, Volume 2:859–864, 2003.

[19] Richard G. Casey and Eric Lecolinet. Strategies in character segmentation: a survey. *Third International Conference on Document Analysis and Recognition*, 1995.

[20] Anoop Namboodiri and Sachin Gupta. Text independent writer identification from online handwriting. *Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.

[21] Oxford University. Early manuscripts at oxford university. `http://image.ox.ac.uk`, 1995.