Similarity-based Image Retrieval for Revealing Forgery of Handwritten Corpora

Ilaria Bartolini
DISI, Università di Bologna, Bologna, Italy

Keywords: Similarity-based Image Retrieval, Low-level Features, *k*-Nearest Neighbor, Handwritten Corpora, Forgery.

Abstract: Authorship attribution is a problem with a long history and a wide range of applications. Recent works in non-traditional authorship attribution contexts demonstrate the practicality of automatic analysis of documents based on authorial style. However, such analyses are difficult to apply and few "best practices" are available. In this paper, we show how quantitative techniques based on image similarity search can be profitably exploited for revealing forgery of handwritten corpora. More in details, we explore the case where a document is represented by means of the image of the document itself. Preliminary experimental results conducted on real data demonstrate the effectiveness of the proposed approach.

1 INTRODUCTION

Authorship attribution is the science of inferring identity of an author from the characteristics of documents written by the author itself. It represents a problem with a long history and a wide range of applications. Recent works in non-traditional authorship attribution domains demonstrate the practicality of *automatic* analysis of documents based on authorial style. Such analyses are however difficult to apply, little is known about types or rates of error, and few "best practices" are available (David and Karl, 2014).

The specific type of automatic document analysis depends on the meaning we convey to the word "document". The variance of this concept, in fact, implies that different methods could be exploited for automatically managing corpora with the final aim to reveal forgery. Computer scientists, mathematicians, philologists, quantitative linguists and digital humanists have different points of view on what a document is. This entails different approaches that could be constructively integrated in order to help with complex problems such as forgery (Tomasi et al., 2013).

In the following, we will focus on how to exploit traditional *computer science* practices in order to reveal forgery of handwritten corpora and show how quantitative techniques based on *image similarity search* can be profitably exploited for this purpose. More in details, we explore the case where a handwritten document is represented by means of the image of the document itself, that is, the digital representation of a manuscript page. The reference problem is thus turned into a *similarity-based image retrieval* one (Smeulders et al., 2000).

We demonstrate how the automatic characterization of the content of document pages, in terms of "low-level features", can effectively contribute to authorship attribution. In doing this, we need to investigate a set of relevant related issues. In particular, (1) Which low-level features should be used for representing pages? (2) How to compare such automatically extracted features? (3) How to asses if two pages are "(dis)similar", thus establishing whether a manuscript corpus is authentic or not with respect to a specific author?

In tackling above issues, we also need to consider that similarity could not be always a satisfactory criterion in order to attribute authorship in the case of suspected forgery: it is not surprising that a forgery is similar to the author's work. The problem is how to verify if a document is too similar to the author's work, and if such types of similarity cannot be found elsewhere in the remaining corpus.

Further, we have to consider that manuscript pages could be analyzed at different levels of granularity, each one defining an image *element*: syllables, words, sentences, paragraphs, whole pages, etc.

From each element, visual characteristics, able to define specific graphic aspects of an author (thus, able to differentiate her/his handwriting), are extracted. Among them, well known examples coming from the traditional *paleography* context are *leading*, *writing*

Bartolini I..
Similarity-based Image Retrieval for Revealing Forgery of Handwritten Corpora.
DOI: 10.5220/0005564401040112
In *Proceedings of the 12th International Conference on Signal Processing and Multimedia Applications* (SIGMAP-2015), pages 104-112
ISBN: 978-989-758-118-2

body, *writing angle*, *density*, *word's spacing*, and *ductus*. Although effective, however, all such features represent a completely *manual* solution to the problem, since this requires *human experts*, possibly supported by some image editing tools (David and Karl, 2014), to extract them.

In order to automate the management of handwritten corpora, we propose a completely automatic representation of elements based on the notion of *handwriting shape*. To model the writing shape, a set of effective visual characteristics (called *local* features) are extracted from each element using specific image analysis techniques like, for example, SIFT (Lowe, 1999) and SURF (Bay et al., 2008). So-obtained local features are then *compared* in order to establish their degree of (dis)similarity, with the final aim to establish whether a corpus is authentic or not with respect to a specific author. This implies a *preprocessing* phase where the analysis of some handwritten pages of *authentic* writings is executed in order to build a "ground truth" reference information for comparing suspicious writings to the authentic ones. Given an input page, composed by a set of target elements, and an element distance function that measures the (dis)similarity of a given pair of elements using their local features, we want to determine automatically if the query manuscript page could be considered authentic with respect to a specific author. The (dis)similarity between pages is numerically assessed by way of a page distance function that somehow "combines" the single element distances into an overall value. Preliminary experimental results conducted on a software implementation of the proposed solution, namely WRITINGSIMILARITYSEARCH (WSS), and using real data demonstrate the effectiveness of our method and encourage further investigations on this direction.

The rest of the paper is organized as follows: Section 2 reports the related work; Section 3 details the proposed similarity-based image retrieval approach. In Section 4 we describe WSS, whereas in Section 5 we comment some preliminary experimental results based on real document collections. Finally, Section 6 concludes the paper.

2 RELATED WORK

In this section we report state-of-the-art solutions to authorship attribution with respect to the specific field of image analysis.

OCR is a traditional approach based on pattern recognition techniques that enable a computer to read texts (i.e., scanned images of a texts) (Bunke and Wang, 1997). However, if this is a feasible solution on printed text, its use for manuscripts is rather problematic. In general, OCR applied to handwritten texts is far from being perfect because of the issue of "variations". For example, the same letter drawn by the same person is slightly different each time, as well as letters drawn by different hands. These variations make it hard for the computer to read the writing correctly and to make a successful match in the context of authorship attribution.

Due to poor recognition results provided by OCR, handwritten document image retrieval remains a very challenging problem: keeping documents as image format is a more economical and flexible alternative than converting image documents into text format by OCR; furthermore, it is more robust for different variations and degradations (David and Karl, 2014).

In (Aiolli and Ciula, 2009), the authors propose the tool System for Paleographic Inspections (SPI). SPI solves the problem of variations by training and working on prototypes of letters, i.e., collecting abstracted models of a single person's handwriting. The prototype comes with a predefined set of limits between which the letter belonging to the unidentified document may deviate from the prototype. The main limit of SPI is that the segmentation process focuses on the shape of individual letters only. Thus, the overall appearance of the manuscript page, at the different levels of granularity (i.e., sentences, words, etc.), and its immediate context are completely ignored.

In (Rath et al., 2004), a probabilistic annotation model for word matching in written documents is presented. Word images are represented by means of Fourier coefficients. A learning model is trained to map any given word image to a specific word from a vocabulary with a probability. At query time, the model estimates the probability of a query word and a sequence of feature vector occurring together. The method is pure text-based retrieval and achieves multiple-words query tasks; however, it suffers from queries which do not appear in the training set.

Finally, (Cao et al., 2011) propose an adapted vector model for word retrieval, where documents and queries are represented by means of a vector space of term frequency (TF) and inverse document frequency (IDF) for each term in the vocabulary. TFs and IDFs are estimated by means of word segmentation and recognition likelihood. Retrieval is achieved by measuring the similarity between vectors of query and data documents with a ranked list. Similarly to (Rath et al., 2004), also this approach is impracticable when queries do not belong to the vocabulary.

To the best of our knowledge, WRITINGSIMILARITYSEARCH is the first attempt to provide a thorough solution to the problem of revealing forgery of handwritten corpora based on image similarity.

3 THE SIMILARITY-BASED IMAGE RETRIEVAL APPROACH

In this section, we detail our content-based image similarity solution to the authorship attribution problem. In our model, each document is represented by means of the *image* of the document itself, that is, through the digital representation of the manuscript page. By applying image analysis methods and (dis)similarity search techniques, we automatically characterize the content of each page through "low-level features" and easily retrieve the most (dis)similar pages to the target (i.e., a specific author) one, following the *k-Nearest Neighbor* (*k*-NN) search paradigm (Baeza-Yates and Ribeiro-Neto, 1999).

The basic idea of our approach is summarized in Figure 1: we build a database of *training* features extracted from some handwritten pages of authentic writings (what we call "ground truth"). Then we extract the same features from a (*suspicious*) test page and compute the (dis)similarity between above features in order to establish the paternity of the test page with respect to the target author.

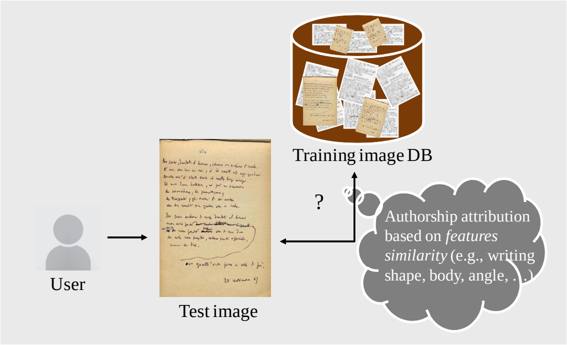

Training image DB
?
Authorship attribution based on *features similarity* (e.g., writing shape, body, angle, ...)
User
Test image

Figure 1: The basic idea of content-based image similarity search for authorship attribution.

Data and comparison models are inspired by the Windsurf ones (Bartolini et al., 2011)

With respect to the data model, images are composed by elements, that is, relevant parts of the handwritten page; each element is described by means of automatically extracted low-level features that represent, in an appropriate way, the content of the element itself (i.e., the handwriting style of an author with respect to a specific set of graphic aspects).

As for the comparison model, given an input (*query*) manuscript page, composed of m relevant elements, and an element distance function d_e that measures the dissimilarity of a given pair of elements using their features, we want to automatically determine if the query manuscript page could be considered authentic with respect to a specific author. The (dis)similarity between manuscript pages is numerically assessed by means of a page distance function d that somehow "combines" the single element distances into an overall value. The efficient resolution of comparisons over features is ensured by an index structure (e.g., M-tree (Ciaccia et al., 1997)) built on top of elements (for example, *words*).

Pragmatically, manuscript pages are first segmented in parts (e.g., syllables, words, sentences, and also the *whole* page). From each element, visual salient characteristics, able to define specific graphic aspects, and thus differentiate the handwriting of an author, are extracted. Image elements are then compared according to their visual features according to an ad-hoc distance metric d_e. Elements scores are finally appropriately matched to aggregate distance values of matched elements using the page distance function d (e.g., the average).

Among relevant visual features, we have characteristics traditionally used in paleography literature like (1) *leading*: vertical distance between a line and the next one of a document page; (2) *writing body*: height of the body of the letters (except ascenders and descenders); (3) *writing angle*: angle of inclination of the pen (except ascenders and descenders); (4) *density*: filling factor of selected area (black pixels vs. white ones); (5) *words' spacing*: horizontal distance between two consecutive words; (6) *ductus*: qualities and characteristics of writing instantiated in the flow of writing the text. Note that, some of the above features refer to the whole document page (e.g., leading, density, and ductus), while others work at a finer granularity level (e.g., writing body, writing angle, and words' spacing). Each feature is represented as a numerical value (1-*D* feature vector); comparison between elements is then assessed as the absolute value of the difference between single feature values.

Although effective, and thus used in paleography contexts, all above features share the limit of requiring a manual extraction process by human experts, eventually supported by image editing tools like Graphoskop.[1] In the latter case only, we can refer to paleography features as *semi-automatic* ones.

To *scale* and *integrate* traditional methods, we propose a completely automatic solution for the rep-

[1] Graphoskop library: www.palaeographia.org/graphoskop/

resentation of each element based on its *content* (i.e., *writing shape*) description. In details, we model writing shape through a set of local features (or salient points) automatically exacted using image analysis techniques like, for example, SIFT (Lowe, 1999) and SURF (Bay et al., 2008). So-obtained salient points are then *compared* in order to establish their degree of (dis)similarity, with the final aim of establishing whether a corpus is authentic or not with respect to a specific author. We refer to writing shape features as *automatic* features.

As above mentioned, SIFT and SURF techniques represent the image content by means of a (*large*) set of local salient points (e.g., corners, blobs, and T-conjunctions). Each point is usually represented by a 128-D feature vector. Similarity between images is then assessed by matching visual characteristics of their salient points based on Euclidean or quadratic distance functions and aggregating (using the average function) local scores to a global value.

Clearly, high dimensionality can be an issue working with both SIFT and SURF: e.g., 200/1000 salient points, each represented by a 128-D vector, can be used for representing a single image element. To overcome this problem, approximate solutions to the k-NN search problem in high-dimensional spaces are applied: the “best bin first” 1-1 matching (instead of exact one) is adopted together with a simple Euclidean distance function as similarity criterion. The rationale of the approximate matching algorithm is to match each salient point of the query element, starting from the first one, to the “best” (i.e., most similar) salient point of the target DB element.[2]

From the complexity point of view, using the “best bin first” approximation allows us to bound the global cost to $O(N^2)$, denoting with N the (average) number of salient points in an image (Rui et al., 1998) (the complexity of 1-1 exact matching, e.g., by means of the Hungarian algorithm, is $O(N^3)$).

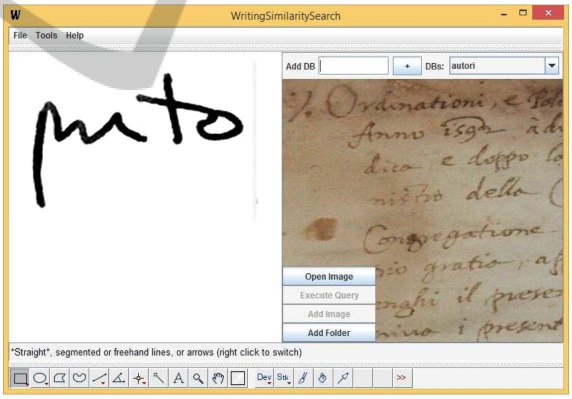
4 WRITINGSIMILARITYSEARCH

This section describes WRITINGSIMILARITYSEARCH (WSS), the software architecture that implements the proposed approach and that is able to automatically analyze and classify manuscript pages. WSS is basically composed of two phases:

Training Phase: the ground truth is built on a representative subset of the documents collection (i.e., authentic writings) by extracting features from elements at different level of granularity; each image is also labeled by means of the reference “author class” that, in the more general case, is represented by a keyword. The class can be also modeled at a finer level of granularity; this is possible by associating to each image a path of an “author taxonomy”, (i.e., “author/date”, “author/word”, “author/word/date”, etc.).

Test Phase: at run time, given a suspicious image element, its features are extracted and compared to the training features in order to compute their (dis)similarity scores; paternity of the test element is established based on a k-NN classifier, as we will present in the following.

In doing this, WSS exploits *both* “automatic” extracted features (i.e., writing shape) and “semi-automatic” extracted paleography characteristics (i.e., leading, body, angle, density, words’ spacing, margins, etc.), and provides the user with several functionalities, such as persistent (MySQL-based) features extraction, features comparison, k-NN querying and classification, that are available through an intuitive and user-friendly GUI (see Figure 2).



"Straight", segmented or freehand lines, or arrows (right click to switch)

Figure 2: WRITINGSIMILARITYSEARCH’s interface.

WSS is based on the Windsurf software library (Bartolini et al., 2011), freely available at URI http://www-db.disi.unibo.it/Windsurf/. Further, it exploits the graphical functionalities offered by Graphoskop to deal with traditional paleography features in a semi-automatic way.

Windsurf provides a general framework for *region-based* image retrieval that we have opportunely instantiated and extended for the specific context of handwritten documents collections. We also have included a software module that offers to the user specific graphical tools to extract paleography features.

Focusing on writing shape, given an image of

[2] We note that our approach is completely independent from specific features, distance functions, and matching algorithms; this is due to the nice property of the Windsurf framework to be parametric with respect to *all* such dimensions.

a document element (for example, the Italian word "punto" depicted in Figure 2), SIFT/SURF features are automatically derived and persistently maintained.[3] The process can be recursively repeated for whole folders of images by using the *Add Folder* button in the GUI instead of the *Add Image* one.

The basic ingredients of the Windsurf framework are instantiated within our new context as follows: image regions correspond to elements features (e.g., SURF salient points); the region (dis)similarity function d_e is the Euclidean distance; the matching problem is solved by means of the best bin first 1-1 matching and using the average aggregation function for computing (dis)similarity between images.

To complete the WSS description, we detail how the *k*-NN classifier works (Baeza-Yates and Ribeiro-Neto, 1999). This is based on the notion of *k*-NN similarity search, where, given a query (test) image and a numeric value *k*, the *k* most similar images of the training set with respect to the query are returned in descending order of similarity (or, equivalently, the *k* less dissimilar, in ascending order of dissimilarity) (see Figure 3, for a concrete example).

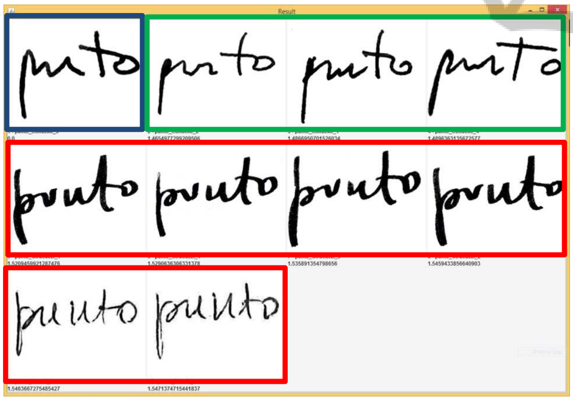


Figure 3: *k*-NN similarity search example in WSS: 9-NN results for the query "punto" (top-left image) depicted in Figure 2.

Images in the above figure have been highlighted with different colors to emphasize their relevance with respect to the query image. The first three images (colored in green) have been written by the same author of the query, while other images (colored in red) have been written by different authors. In this case, therefore, we are able to automatically (and correctly) predict that the author of the query image is the author of the green-colored images (the most similar to the query).

Since the *k*-NN similarity query is performed on images that are part of the ground truth, the classifier simply analyzes the paternity classes associated to the *k* returned images, assigning to the (suspicious) test image the most frequent class over the *k* results. Of course, the choice of the "optimal" value of *k* is an issue that requires to be experimentally investigated for our *k*-NN classifier in the proposed context.

We note that the same approach can also be exploited at a *cluster* level of training features, instead that using *individual elements*, as described above. In this case, a clustering phase is integrated within the training process in order to derive clusters of elements at fixed levels of paternity classes (e.g., at level of author/, author/word, author/date, author/date/word, etc.). For each of so-derived cluster, a centroid is computed as the average of the included elements features. Since all images belonging to a single class are now compacted into the corresponding centroid, the *k*-NN search is now reduced to a 1-NN search, and the predicted class is the one of the most similar centroid. As a main consequence, here the selection of the numeric value *k* is no longer an issue; furthermore, the comparison process becomes simpler than before, since it now involves a lower number of elements (clusters) instead of the whole dataset of individual training elements. On the other hand, it has to be noted that the effectiveness of a cluster-based classifier is strongly influenced by the "quality" of centroids in representing *all* cluster elements, thus in the *variance* of the involved features, i.e., classes having a high variance in elements' features are expected to be less effective than homogeneous classes.

In the following, we will refer to the two classifying modalities as *cluster-based* and *element-based* classifiers, respectively.

5 EXPERIMENTAL RESULTS

We include here the results of preliminary experiments performed on real handwritten corpora consisting of about 200 JPEG image documents (320x240 pixels in size). The reference corpora includes documents by a specific target author *X* and suspicious documents written in a very similar handwriting style to the target author, that we refer as *X'*.

In details, paleography features (i.e., leading, body, angle, and density) were computed on the whole documents using WSS provided graphical tools, while writing shape features were automatically extracted from specific word elements characterizing the writing style of the target author *X*. In our specific image context, we experimentally found similar re-

[3] We experimentally found that a good number of salient points for representing the shape of a document element, such as a word, ranges from 200 to 500 points.

sults in effectiveness using SIFT and SURF features. Thus, for efficiency reason, we decided to base our feature extraction process on the SURF algorithm. In particular, about 200 salient points were extracted for each word. Several tests demonstrated that 200 salient points represented a good compromise between effectiveness and efficiency for representing the content of single words. Finally, in order to add noise to our initial corpora, a large number of handwritten documents (representing both whole document pages and same special words as above), written by different authors (named *A*, *B*, and *C*), were added to the reference DB. A sample of representative query images, not belonging to the training set and for which paternity classes (at different levels of granularity) were associated for evaluation purposes, was randomly selected to test our classifier.

We start our investigation by trying to answer to a preliminary important question: “Does the *way* we obtain the digital representation of an original handwritten page, e.g., scan of the original page, photo of the original page, scan/photo of a photocopy of the original page, etc., affect the target features representation?”

We experimentally found that both traditional paleography and writing shape features are *invariant* to the digital representation of the original handwritten page. In fact, the absolute value of the differences between single feature values representing paleography characteristics, when extracted from scan, photo and scan/photo of the original page, were almost equal to zero. The same behavior was confirmed for shape features: *k*-NN searches based on *any* of above digital representations of the original page (image query) provided the same ranking of returned images.

We are now ready to demonstrate the effectiveness of the WSS classifier in predicting the paternity of a suspicious test writing. In particular, we provide the results of both WSS element-based and cluster-based classifiers. In doing this, we use both traditional paleography features and writing shape ones.

We start by using traditional paleography features. In details, we clustered individual features so as to train WSS for each author; then we applied the cluster-based classifier.

Table 1 shows the effectiveness of WSS in distinguish handwritings from *different* authors: training author was in this case set to *B*, whereas the test image (query) was written by author *A*.

To check paternity, absolute differences in individual paleography features of test image and training class are compared to experimentally derived *threshold* values. In the proposed example, we postulate that the test image (written by author *A*) belongs to training class *B*. However, the table highlights high difference values for the following features: writing body (> 0.5 *mm*), all writing angles ($> 5^\circ$), space between words (> 0.5 *mm*), and density ($> 2\%$). Since the number of sufficiently dissimilar features between the test image and the training (average) image is high, we conclude that the test image does not belong to the test class *B*. In this case, WSS is thus able to infer that the test image has not the same handwriting than the training author.

Table 1: Discriminative power of WSS in distinguishing handwritings coming from different authors: training author is *B*; test image is associated to author *A* by the ground truth.

Target Features	Training (*B*?)	Test (*A*)	Abs. Diff.
Leading (*mm*)	5.01	5.00	0.01
Writing Body (*mm*)	4.45	2.90	1.55
Writing Ascender (*mm*)	1.50	1.30	0.20
Writing Descender (*mm*)	1.57	1.30	0.27
Writing Angle (°)	89.40	76.00	13.40
Writing Ascender Angle (°)	88.70	76.00	12.70
Writing Descender Angle (°)	86.90	77.00	9.90
Space between Words (*mm*)	2.65	3.40	0.75
Density (%)	10.10	6.00	4.09

If we apply the WSS classifier to the whole cluster dataset, predicted author for the test image is *A*, since the most similar centroid of all trained authors is the one corresponding to cluster *A* and the number of differences in features exceeding the threshold is low. Again, WSS is able to classify the given test image to the correct class.

In Table 2, we report the results of a similar experiment: this time, we trained WSS on a subset of writings (i.e., from a specific decade) of the author *X* and on a sample writings of the suspicious author *X’* coming from the same period. Then, we set author *X* as training and provide a test image (query) written by author *X’* to WSS. In this case, we want to investigate if WSS is able to distinguish (possibly existing) differences between the two handwritings corpora.

Table 2: Discriminative power of WSS in distinguishing handwritings coming from *similar* authors: training author is *X*; test image is associated to author *X’* by the ground truth.

Target Features	Training (*X*?)	Test (*X’*)	Abs. Diff.
Leading (*mm*)	7.70	7.10	0.30
Writing Body (*mm*)	2.23	2.10	0.13
Writing Ascender (*mm*)	2.07	2.40	0.33
Writing Descender (*mm*)	2.36	2.60	0.23
Writing Angle (◦)	71.00	73.00	2.00
Writing Ascender Angle (◦)	75.42	74.00	1.42
Writing Descender Angle (◦)	80.26	77.00	3.26
Space between Words (*mm*)	2.92	2.60	0.32
Density (%)	8.83	10.00	1.17

As we can observe from statistics shown in table, differences here are very small. The classifier would thus confirm the hypothesis (i.e., *X'*=*X*).

We now show the performance of WSS with respect to writing shape features, showing how the system is able to discriminate handwritings of different authors by exploiting *k*-NN search. In particular, we evaluate the effectiveness of *k*-NN search using the well known precision (P) and recall (R) metrics. Precision measures the number of *relevant* (i.e., the same word written by the same author) images to the query, over the *k* returned images, while recall is the number of relevant images to the query over the total number of relevant samples in the training set.

Table 3 shows the average results of P and R obtained on our test queries representing words. In this case, the *k*-NN search is performed over the query word written by different authors.

Table 3: *k*-NN average precision (P) and recall (R) values vs. *k*.

k	P	R
1	1	0.1937
2	1	0.3875
4	0.9375	0.7437
6	0.783	0.8562
8	0.5625	0.8562
10	0.475	0.8562
15	0.3162	0.9062
20	0.2567	0.9375

Considering that classes in our dataset have (on average) 5.5 relevant samples in the training set, we observe how the quality of the results is quite good: WSS is able to return most of the relevant images in the first positions; further, almost all relevant samples are localized by the system very soon (we get values of recall close to 1 in the first $15-20$ results). This demonstrates the effectiveness of both local features and comparison method adopted by WSS. Similar results were confirmed when searching for the author of *any* word. Even when a query word is not included in the training set, WSS is therefore able to correctly predict its author, thus obtaining good performance at different levels of granularity.

Next, we test the quality of WSS classifier. For evaluating the quality (Q) of the classifier, we assigned the binary value 1, when the class prediction for a test image was correct with respect to the ground truth, and 0 otherwise. With respect to the choice of the value *k*, we experimentally found that a reasonable solution is to bound *k* to the average value of the total number of relevant images with respect to the query in the training set.

For the element-based approach, our experiments produce 100% accuracy for any value of *k* in the range $1-8$. This is expected because of the very good performance of the *k*-NN search demonstrated earlier (see P and R values in Table 3).

Table 4 shows the average value of Q for the cluster-based working modality (Figure 4).

Table 4: Average quality results of WSS cluster-based classifier: Q vs. clustering level.

Clustering Level	Q (Cluster-based)
author	0.2
author/word	1

We observe how the quality decreases to 20% when working at the "author" level; as we argued in Section 4, this is due to the high variance in elements' features of author classes. In fact, as soon as the classifier is set to work at the "author/word" level, Q reaches optimal values, as for the case of the element-based classifier.

To complete our analysis, we report two visual examples of the WSS classifier at work, with the aim to show how WSS could be helpful in predicting the paternity of a suspicious test image by author *X'* (see Figures 4 and 5).

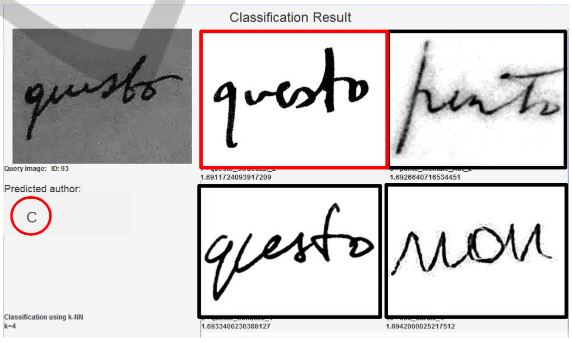
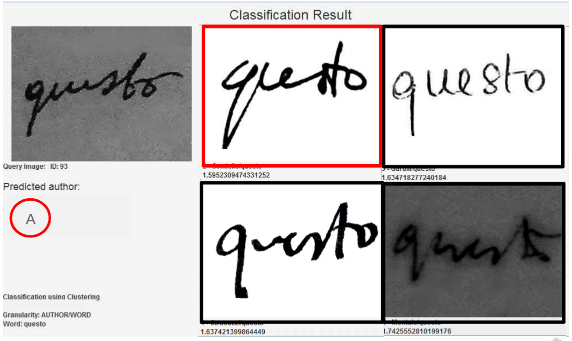


Figure 4: Visual example of *k*-NN classification based on elements ($k = 4$).

In details, Figure 4 shows the prediction according to the most frequent class among the results of a 4-NN individual elements query (i.e., element-based classifier). Since the four ranked results are all of distinct authors (*C*, *X*, *A*, and *B*, respectively), the classifier would select for the test image the class author associated to the most similar image (i.e., *C*). Thus, the classifier would have refused the thesis *X'*=*X*.

In Figure 5 the classification is based on the class of a 1-NN cluster query; note that, further results included in Figure 5 (with $k = 4$) show that WSS considers, in this case, the handwriting from *X'* (query image) very far from the handwriting from author *X* (which is ranked only at position 4). Even in this case, the classifier would have refused the thesis *X'*=*X*.



Classification Result

Query Image: ID: 93

Predicted author:

A

Classification using Clustering

Granularity: AUTHOR/WORD
Word: questo

1.5952309474331252

1.634718277240184

1.637421399864449

1.7425552810199176

Figure 5: Visual example of *k*-NN classification based on clusters (at author/word level).

Summarizing, provided results confirm the effectiveness of traditional *manual* practices of human experts in the context of paleography (i.e., manual extraction and human observation of specific corpora characteristics), practices here permitted by WSS in a semi-automatic way and with the final aim to classify suspicious writing to the proper class. Further, provided tests demonstrate how the same ability to discriminate between different handwritings can be obtained in a completely automatic way by exploiting writing shape features, meaning that the WSS classifier could be a very convenient and smart reference software architecture for paleography experts for tackling the authorship attribution problem. In particular, we envisage two possible uses of the WSS classifier: (1) The two classifiers (the one using paleographic features and the one using writing shape features) could be used separately to predict a class for a same query image; if the two predicted classes are different, then we alert the handwriting expert that the query image is suspect and that further investigation is needed. (2) Since the classifier using paleographic features requires additional parameters (i.e., the (dis)similarity thresholds) for which an appropriate value can be hard to be derived, we can use the classifier using writing shape features (which is completely automatic) to predict the class of the query image; then, the handwriting expert can restrict her search only on the author (or word) that has been predicted by WSS.

6 CONCLUSIONS

In this paper, we proposed a novel approach to the authorship attribution problem based on image similarity search. In details, we presented WSS, a software architecture able to automatically predict the paternity of a (suspicious) test document exploiting both automatic and semi-automatic features. Preliminary experimental results conducted on real data demonstrated the effectiveness of our classifier and encourage further investigations on this direction.

In the future, we plan to study and compare alternative representations for writing shapes, e.g., based on global features, like wavelet and Fourier coefficients. Finally, we advocate the creation of benchmarks consisting of large handwriting corpora allowing the comparison of existing approaches. In doing this, our intention is clearly to involve human experts in the validation process.

ACKNOWLEDGEMENTS

Part of this research has been supported by the FARB Project “Authorship, variants, style: textual analysis between philology, linguistics, mathematics and computer science”. A special thank to the students that participated in the WSS software development.

REFERENCES

Aiolli, F. and Ciula, A. (2009). A case study on the system for paleographic inspections (SPI): Challenges and new developments. In *Proc. Conf. Comp. Intell. and Bioeng.*, pp. 53–66, Amsterdam, The Netherlands.

Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley.

Bartolini, I., Patella, M., and Stromei, G. (2011). The Windsurf library for the efficient retrieval of multimedia hierarchical data. In *SIGMAP 2011*, pp. 139–148, Seville, Spain.

Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359.

Bunke, H. and Wang, P. S. P., editors (1997). *Handbook of character recognition and document image analysis*. World Scientific.

Cao, H., Govindaraju, V., and Bhardwaj, A. (2011). Unconstrained handwritten document retrieval. *IJDAR*, 14(2):145–157.

Ciaccia, P., Patella, M., and Zezula, P. (1997). M-tree: An efficient access method for similarity search in metric spaces. In *VLDB ’97*, pp. 426–435, Athens, Greece.

David, D. and Karl, T. (2014). *Handbook of Document Image Processing and Recognition*. Springer.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *ICCV ’99*, pp. 1150–1157.

Rath, T. M., Manmatha, R., and Lavrenko, V. (2004). A search engine for historical manuscript images. In *SIGIR 2004*, pp. 369–376, Sheffield, UK.

Rui, Y., Huang, T. S., Ortega, M., and Mehrotra, S. (1998). Relevance feedback: A power tool for interactive content-based image retrieval. *Trans. on Circuits and Systems for Video Technology*, 8(5):644–655.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *TPAMI*, 22(12):1349–1380.

Tomasi, F., Bartolini, I., Condello, F., Esposti, M. D., Garulli, V., and Viale, M. (2013). Towards a taxonomy of suspected forgery in authorship attribution field: A case: Montale’s diario postumo. In *DH-CASE ’13*, pp. 10:1–10:8, Florence, Italy.