

Toward an AI-based Tool for Paleographic Analysis – A Global Approach to the Scribal Characterization Step

Stefano Ferilli^{1,*†}, Domiziana Piscopo^{2,†}

¹Dipartimento di Informatica, Università degli Studi di Bari, Via E. Orabona, 4, Bari, 70125, Italia

²Dipartimento di Ricerca e Innovazione Umanistica, Università degli Studi di Bari, Piazza Umberto I, 1, Bari, 70121, Italia

Abstract

Optical Character Recognition has been traditionally used for transcription of texts. A solved problem for modern printed scripts, it is still challenging for ancient scripts and documents. Paleographic research tasks are very different, and pose very different problems and challenges, with a strong emphasis on the deep study of single characters. The available solutions for transcription are of little use for these tasks. There is a need for flexible and interpretable tools that can compare characters and return all relevant recognition options, so that researchers and scholars may explore, validate, or even generate research hypotheses. This paper takes up one such task: testing a method for analyzing graphic variability within a normative medieval script. It proposes an AI-based approach that derives this measure directly from the corpus, comparing every occurrence of a graphic element against all others, and evaluates it on an eleventh-to-twelfth-century manuscript, written in the Beneventan ‘Bari-type’ and attributed to a single hand. The results confirm the complementarity of computational and paleographic analysis, and their capacity to produce, in dialogue, interpretations that neither could reach alone.

Keywords

Paleography, Character Recognition, Document Processing, Machine Learning

1. Introduction

All of the historical documents of interest for Digital Humanities come in the form of digitized legacy physical documents, i.e., (sequences of) images representing scanned pages. However, in a scanned page, even characters are just images. Due to the importance of text in a document, much effort was devoted in the past decades to Optical Character Recognition (OCR), aimed at identifying and classifying character symbols in these images, usually with transcription purposes. Handwritten Text Recognition (HTR) specifically focuses on handwritten text. The nearly 100% accuracy of OCR on contemporary printed documents cannot be straightforwardly ported to historical documents, where recognition is much more challenging, for several reasons: linguistic resources (e.g., dictionaries) to fix misrecognitions may not be available, instances of the same character are different (especially in calligraphic writings), there are non-standard characters for abbreviations, and there is a much wider set of fonts/scripts. Also, recognition is not the only objective of researchers and scholars. For paleographic research, an insight in the single characters and their features is often fundamental, e.g. for scribal hand identification.

Such insight is harder to secure, and more consequential, precisely where a script is bound by a shared graphic canon. A scribe working within a normative system still produces, from one instance of a grapheme to the next, differences fine enough to carry information about a hand’s identity, or about how a single hand’s practice shifts across a manuscript — and fine enough, too, that even an

4th Italian Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH 2026, <https://ai4ch.di.unito.it/>), co-located with the 24th International Conference of the Italian Association for Artificial Intelligence (AIXIA 2026). 6–9 October 2026, Perugia, Italy

*Corresponding author.

†Sections 1, 2 and 6 were co-authored; Section 3 was written by D. Piscopo; Section 4 was written by S. Ferilli; in Section 5, S. Ferilli set up the recognition tool and ran the experiments, D. Piscopo analyzed and interpreted the results.

✉ stefano.ferilli@uniba.it (S. Ferilli); domiziana.piscopo@uniba.it (D. Piscopo)

🌐 <http://ara.di.uniba.it/~ferilli/ufficiale/ferilli.html> (S. Ferilli); https://www.uniba.it/it/dottorati/patrimoni-storici-filosofici/corso/dottorandi/xxxix-ciclo/piscopo-domiziana_bio-bibliografia_dottorato_uniba.pdf/ (D. Piscopo)

🆔 0000-0003-1118-0601 (S. Ferilli)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

expert eye rarely judges with the same consistency once the instances to compare multiply, whether within a manuscript or across several. The more tightly a script is governed by its canon, the sharper this difficulty becomes: less room is left for variation, and what remains is correspondingly harder to isolate without purposely-built tools. This contribution belongs to a doctoral research project in Latin paleography at the University of Bari, focused on the Beneventan ‘Bari-type’, a highly normative medieval book script, never previously approached computationally. The aim of this research is to test, through a case study, how far paleographic expertise and computational analysis can inform each another in measuring graphic variability within such a system. The first step of this program addresses the graphic behavior of a single scribe within a single manuscript. An initial phase [1] measured this behavior by selecting, for each grapheme under study, one occurrence identified through paleographic judgment as representative of its graphic synthesis, and comparing every other occurrence against this single reference prototype. That approach produced a coherent picture of the scribe’s overall adherence to the canon, but exposed a structural limitation: a prototype selected in this way is itself only one realization among many, with no guarantee that it is truly representative of the scribe’s hand. The resulting measure of variability depends on which occurrence was chosen as the term of comparison. The present contribution addresses this limitation directly, seeking a prototype that emerges from the data themselves rather than one selected in advance — an output the system returns, not a category the analyst induces, derived from the relations among all occurrences considered together. This paper describes how the recognition system used in that first phase was extended from prototype-based comparison to a fully pairwise one, in which every occurrence of a grapheme is measured against all the others rather than against a single fixed model. The extended method is applied to the same manuscript, the Casanatense 1101 (Roma, Biblioteca Casanatense), focusing on the assimilated *ti* ligature, as a preliminary trial of the method itself: the present contribution does not attempt to conclusively assess the graphic variability of a single scribe across the manuscript, but to test, within a deliberately narrow scope, whether this corpus-driven approach offers a viable path toward that broader goal.

The rest of this paper is organized as follows: after discussing basics and related work in the next section, Section 3 describes the problem. Then, Section 4 describes our approach, and Section 5 evaluates it on a real use case, before concluding the paper in Section 6.

2. Background & Related Work

Recognizing characters in scanned document is the task of Optical Character Recognition (OCR). For an OCR system to be effective and flexible, it should be invariant or tolerant to change in intrinsic variables of the image (color, luminance or contrast), scale (the character size is irrelevant), translation (the character location in the image is irrelevant), rotation (even in a straight page image, characters may be skewed, especially in ancient documents) [2]. Efficiency (basically, speed) is important to process large amounts of data. OCR was born to handle modern documents, typically in Latin scripts, where it is considered as a solved problem. More recently, the need for automated processing of ancient documents raised new issues and opened new research avenues. Here, the task is much more difficult [3], especially for handwritten text recognition [4], due to the great variability in layout and fonts/calligraphy, where models learned on modern digital fonts are ineffective. Most tools developed in this field are for transcription, which is the immediate counterpart of traditional OCR: turning the text into a sequence of digital, standard codes. Worth mentioning are Transkribus [5] and eScriptorium [6]. The former allows users to train HTR/OCR machine learning models, offering two methods for HTR: one based on Hidden Markov models (HMMs) that integrates Pattern Recognition techniques and a language model (LM), and one based on Recurrent Neural Networks (RNNs) that does not have the out-of-vocabulary limitation of the HMM-LM tool and uses algorithms such as Mask R-CNN or U-Net for detecting text regions.

For the specific OCR/HCR task, the Tesseract - OCRopus - Kraken - Calamari lineage of neural-network based open source tools marked state-of-the-art for practical applications. Tesseract [7] is the forefather: its classifier can be trained using digital fonts that resemble historical fonts or images of

character sets cut from real historical texts. OCRopus [8] introduced RNNs [9]. Kraken [10] is based on a Long Short-Term Memory (LSTM) architecture. Calamari [11], based on advanced Convolutional Neural Networks (CNNs) and LSTM, supports high-precision OCR/HTR with pre-trained models and uses data augmentation techniques to improve training. In more recent research on HTR, data-driven techniques gained momentum. The HisDoc project [12] applied character-based recognition with HMMs to keyword detection [13][14]. The In Codice Ratio project [15] used a character classifier based on a CNN to identify clipped fragments as characters, and a language model based on a HMM for Word prediction.

Due to the underlying techniques, these tools need large amounts of high-quality training data (usually not available for ancient documents and scripts, especially handwritten ones), a manually created ground truth (typically very expensive), and expertise in machine learning. The object recognition technique described in [16] has complementary pros and cons with respect to traditional subsymbolic approaches. It is prototype-based, requiring just one (or very few) example(s) of each character. It is interpretable, meaning that it can explicitly describe why a character was recognized and assigned to some class, and what features of the character image were more relevant to the classification. It returns a ranking of classifications rather than a single option. It is incremental, meaning that the set of recognized characters can be expanded on-the-fly while processing a document, without requiring a global recomputation that starts from scratch. So, it is a natural candidate for the research task approached in this paper.

Script analysis is a completely different task than transcription. Tools such as Transkribus, however sophisticated, remain oriented from the paleographic point of view toward recognizing letters as instances of a script type for the purposes of transcription, and thus of access to textual content, rather than toward the analysis of letterforms at the level of graphic detail that paleographic research itself requires. Computational approaches typically compare a query instance of a graphic form against some fixed representation of its class (see [17] for an extensive survey). Most works in this direction focus on the categorization of scripts into typological classes rather than the behavior of individual hands. This is the case of the System for Paleographic Inspections (SPI), which generates an interpretable prototype for each character as the centroid of semi-automatically segmented occurrences [18], later applied to Caroline and Gothic manuscripts to reflect on the relationship between computational and traditional method [19]. The identification of hands through a fixed reference is comparatively rare, and has mainly been explored outside the Latin tradition, as in the comparison of ancient Greek inscriptions against ‘platonic’ ideal letterforms constructed for each letter, with results consistent with expert epigraphists’ judgment [20]. Whether a fixed reference of this kind is itself representative of the variation it is meant to capture has, to the best of our knowledge, never been tested for the graphic behavior of an individual hand in a Latin book script. An earlier phase [1] of the work described in this paper used the technique in [16] for single-prototype comparison to a single scribe of the Beneventan ‘Bari-type’, obtaining a coherent measure of graphic consistency. Still, it left this question open, and this contribution takes it as its starting point to move on.

3. The Problem

The experiment presented here belongs to a broader research program concerned with the relationship between paleographic method and computational analysis, applied to the study of graphic variability within highly normative scribal systems — in this case, the ‘Bari-type’ Beneventan script. Although governed by a shared graphic canon, a normative script is never entirely static: minimal but measurable differences persist between and within its concrete realizations, and it is precisely these differences that allow scholars to trace a script’s chronological development, distinguish internal typifications, associate a witness with a given *scriptorium*, or establish whether one hand or several were at work on the same manuscript. The present research program can be described, in its most general terms, as an attempt to give systematic and cumulative treatment to this tension between stability and variability.

The tension can be tested at different levels, according to whether attention falls on a single hand or on a plurality of hands, and on a single witness or on several: the internal consistency of one scribe

may be examined within a manuscript or across different manuscripts and copying campaigns, while the convergence or divergence between distinct hands may be examined within a single, collaboratively produced witness, or across several witnesses judged comparable on chronological or typological grounds; a further and related case is that of manuscripts suspected, on independent grounds, of sharing a single scribe. Framed in these terms, the problem engages two distinct fields of inquiry at once. On the paleographic side, it touches one of the discipline's constitutive questions: how an ideal graphic model — shaped by a shared norm and mediated by the individual scribe's training — comes to be realized, each time, in a way that is never a perfect replica. On the digital humanities side, it raises the question of what a computational tool can contribute to an analysis traditionally grounded in the paleographer's trained eye. The two proceed in substantially different ways: the paleographer's eye works by synthesis, weighing multiple cues at once — layout, ductus, rhythm — into a single qualitative judgment; the digital eye, by contrast, isolates discrete formal properties of a graphic sign and translates their distance from a reference model into a quantifiable value. Rather than asking which of the two sees better, the more productive question is what each is able to perceive that the other cannot, and whether bringing them into dialogue can advance both paleographic knowledge and the tools devised to pursue it. It is within this broader horizon — sketched here only in its general terms — that the present contribution situates a first, deliberately circumscribed experimental step, whose specific object and protocol are set out in what follows.

The Beneventan 'Bari-type' is attested, between the eleventh and the thirteenth century, in the Apulian 'Terra di Bari' and, across the Adriatic, in Dalmatia. Historiographical treatment of this typification has so far been episodic rather than systematic: the coordinates established by Elias Avery Lowe [21] and subsequently refined by Guglielmo Cavallo [22] remain, to date, the main points of reference, arrived at in the absence of a complete census of the surviving witnesses against which to test them. An ongoing survey directed at such a census has begun to reveal a graphic variability internal to the type that these earlier characterizations do not fully account for, a picture further complicated by the near-total silence of the sources regarding the identity of individual scribes: colophons, dates and indications of provenance are seldom found in 'Bari-type' manuscripts, leaving comparison among witnesses in the same script as often the only road open for dating and localizing them. It is against this twofold difficulty that a quantitative approach to its graphic behavior acquires genuine methodological weight. Among the witnesses in which the 'Bari-type' is attested, the present analysis is conducted on the Casanatense 1101: an evangeliary of 181 folios, datable to the turn of the eleventh and twelfth centuries, held in the Roman library from which it takes its name, whose contents paleographic examination assigns in their entirety to a single scribe. Attention is directed here to a single graphic element drawn from this manuscript: the assimilated *ti*. The ligature fuses a *t*, formed of two loops, with an *i* attached to it by a descending stroke that curves markedly as it falls. Its structure is thus more articulated than that of a single autonomous letter, the overall shape depending on the coordination of several distinct strokes rather than one; this makes it a comparatively demanding element on which to test the consistency of a scribe's graphic behavior.

A preceding analytical step [1] had measured the consistency with which this scribe reproduces the ligature by comparing twenty recorded occurrences against a single reference instance (f. 23v), selected on paleographic grounds as representative of the symbol's graphic synthesis. The twenty occurrences were sampled at a regular interval of approximately nine folios throughout the manuscript, with occasional deviations dictated by the availability of the grapheme on specific folios: f. 1r, f. 9v, ff. 20v–21r, f. 29v, f. 36v, f. 49r, f. 53v, f. 63r, f. 72v, f. 81v, f. 91v, f. 100r, f. 110r, f. 118v, f. 126v, f. 138r, f. 152r, f. 163v, f. 171r, and f. 179r. Each occurrence recorded was chosen on purely practical grounds, favoring instances graphically separated from neighboring letters, which simplifies their automatic extraction from the page image; paleographic judgment, which treats every occurrence of the grapheme as equally admissible evidence of the scribe's practice, played no part in this choice, so that the sample is not pre-selected for its resemblance to any expected or preferred form. The only occurrence chosen for its perceived representativeness, rather than for its position in the sequence or its ease of extraction, was the reference instance itself. That step returned an encouraging result — an extremely high degree of overall fidelity to the canon — but also exposed a limitation inherent to any design resting on a

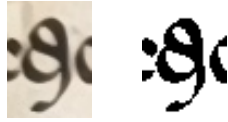


Figure 1: Prototype of assibilated ti (original document fragment and corresponding binarization)

single term of comparison: since the reference occurrence is itself only one realization among many, its particular position and context inevitably leave a trace in the measurements it generates, so that part of what appears as a property of the scribe’s hand may instead be an artifact of the form chosen as its standard. The present contribution takes this limitation as its specific object of study. Rather than designating any single occurrence as a fixed term of comparison, the twenty-one instances of the assibilated *ti* available for this manuscript are considered together: the twenty samples recorded in the preceding step, together with the occurrence that had there served as reference, now treated as one instance among equals. Each occurrence is compared against every other, generating a complete set of reciprocal distances from which a profile of relative centrality — rather than of proximity to a single, externally designated model — can be derived for every sample, including the one previously singled out as the reference. This reversal of design allows the question of bias to be addressed directly, within the deliberately narrow scope of a single ligature in a single manuscript: whether the form selected through paleographic judgment occupies a central or a peripheral position within the corpus it was meant to represent. More generally, the present step is conceived as a preliminary trial of this methodology — a first test of whether deriving a term of comparison from the internal structure of the data offers a viable path towards the broader question of a scribe’s graphic variability, one that a study of this scope can indicate but not establish, and that would require extension to further graphemes to be assessed in full.

4. The AI-based Tool

The general character recognition workflow consists of three steps [23], described in detail in the next subsections: *Pre-processing*, aimed at transforming the source image to enhance important details and/or diminish less significant ones, such as the noise introduced during acquisition, and identify the characters; *Feature Detection*, aimed at extracting high-level, characterizing elements of the character image, that can describe it (e.g., contour [24]); *Recognition*, aimed at predicting, based on the extracted features, the type (‘class’) of a character by comparing it to models of the different classes built in a preliminary training phase by analyzing the features of characters whose class is known. The approach we use in this paper adopts and adapts the object recognition technique originally presented in [16] to the character recognition task.

4.1. Pre-processing

We will describe in this section how we approached the pre-processing phase on the document we used for our experiment. All the technical details of the techniques we mention can be found in [25], while a sample of the input and output of this phase is shown in Figure 1.

The first step in our pipeline to extract the single characters from a scanned document image consisted in turning the original color image to grayscale. We applied the basic standard approach of assigning to all channels of a pixel the average value of the three original RGB components. Then, the result was *binarized*, i.e., transformed into black&white by deciding, for each pixel, if its gray level is sufficiently dark to turn it into a black one, or otherwise turn it into a white one, in such a way that the foreground (relevant pixels, turned to black) is neatly distinguished by the background (irrelevant pixels, turned to white). Thresholding techniques are based on the pixel’s gray level being smaller or larger than a luminance threshold. The higher the threshold, the less levels of gray are set to white, and the darker the

image. Thus, identifying a proper threshold that preserves the interesting details without resulting in meaningless black blobs is an issue that deserves attention. Our system provides both global approaches (apply the same threshold to all pixels in the image) and local ones (that determine a different threshold for each pixel based on its neighborhood).

The acquired images can be of different size, depending on the acquisition procedure. We resize them for reducing the computational cost of processing, when the images are too large. The pages and character size may vary significantly among documents and within documents, which prevents precise standardization at the page level. After several attempts, we realized that, at the moment, a resize ratio such that the size of the characters to be processed is about 50×50 pixels still preserves the character details which are relevant for our experimental purposes. However, checking that this assumption holds will be part of the experiment.

The acquired document images are often skewed or warped. After applying automated deskewing techniques, we realized that the rotation affected the character contours, which are the main source of information for our approach. Since we are interested in processing single characters, and the recognition procedure may take care of making up for noisy rotations, we decided not to apply a global deskewing/dewarping to the document page, and to deal with this issue at recognition time.

For singling out the characters we used a Connected Component (CC) Extraction approach. A CC is a set of black pixels in an image such that each pixel in the set is adjacent to at least another pixel in the set. For finding CCs we used *flood filling*, a bottom-up strategy inspired by the pouring of a liquid into a container. Starting from a black pixel, that becomes the ‘seed’ of the CC, the black pixels adjacent to it are also included in the CC, and the same strategy is applied to each of them until no new pixel can be added. Adjacency can consider only the orthogonal pixels in the image (*4-neighborhood*) or also the diagonal ones (*8-neighborhood*). After identifying a CC, its pixels are removed from the image and the algorithm is restarted from another black pixel in order to obtain the next component, and so on until no more black pixels are available.

After choosing the characters to study in the original manuscript, the paleographer selected the corresponding CCs automatically extracted in the binarized image, and the tool will further process them to form the shape and carry out the character analysis.

4.2. Feature Extraction / Representation

We call a *shape* the internal description of a blob/symbol/character. It is based on blob border as a very indicative feature for object recognition [26]. Of course the internal strokes of a character, with their concavities and convexities, also carry much relevant information, but for this work we decided to use only the border, to have an idea of the accuracy of our approach under very stressful conditions.

Differently from approaches that describe the full contour of the blob, we use Fourier descriptors, and sample the distances at which 256 radian lines at equally spaced angles, originated in the shape centroid and starting from the positive X axis in a coordinate system centered in the blob centroid, meet the blob contour. Such distances are normalized to the integer interval $[0 \dots 255]$ (which has the advantage of being representable using just one byte) and collected in a histogram. This representation ensures invariance with respect to translation (no information on blob location in the page is involved), scale (that does not affect the data structure, but just the values it contains), rotation (by rotating the histogram) and mirroring (by mirroring the histogram).

For the number of radians, 256 corresponds to a binary system for angle measurement widely exploited in several application fields, called *brads* (from Binary RADianS)¹, that divides the round angle into 256 degrees. It is the power of 2 which is closest to the classical number of degrees in a round angle, 360, and corresponds to an angle displacement of 1.40625 degrees between consecutive radians. It was considered as a sufficient approximation for the task of character recognition, based on the shape and size of the characters. In this system, angles can be comfortably indexed using a single byte. This allows to straightforwardly exploit arithmetics modulo 256 when dealing with angles, so that in case

¹<http://web.archive.org/web/20080628051746>

of overflow just ignoring the extra carry bits provides the correct (intended) angle expressed in the basic $[0 \dots 255]$ range (which would not work using binary representations of integers expressing angles with the classical 360-degrees measure). Normalizing all the sampled values of a shape ensures scale invariance. We empirically found that a scale of 256 integer values from 0 (the centroid) to 255 (the distance of the farthest contour pixel from the centroid) provides a sensible tradeoff between accuracy and tolerance to noise in the blob contours (while requiring just a single byte in memory).

4.3. Recognition

The proposed representation allows us to compare two shapes by comparing their histograms: the more different the characters, the more different their histograms. Their distance is computed as the sum of the absolute differences between pairwise corresponding bars and normalizing the result to $[0, 1]$ by dividing it by $256 \cdot 255 = 65280$, i.e., the maximum for such difference. This provides the same information as variance with less effort, because both the number of values and the values are normalized. For rotation invariance, one such comparison must be carried out for each displacement (offset) of one histogram over the other. Then, we take the best case (i.e., the one with minimum distance among these displacements). In fact, character images might be skewed (e.g., due to the acquisition procedure), but considering too large rotations might be harmful (e.g., a ‘d’, if rotated by 180 degrees, becomes a ‘p’; a ‘b’ becomes a ‘q’, etc.). We considered a rotation tolerance of ± 5 degrees, which significantly reduces the amount of effort required for each comparison. Since each bar is associated to 1 brad, corresponding to 1.40625 degrees, a rotation tolerance of 5 degrees requires trying $\lceil 5/1.40625 \rceil = 4$ brads offsets both on the left and on the right, and so overall 8 histogram comparisons are needed. Since each histogram comparison requires computing 256 differences, the overall matching effort (number of differences computed) is $256 \cdot 8 = 2048$.

Classification involves a ‘query’ character to be classified and a number of ‘prototype’ characters stored as representatives of the classes in the form of shapes. So, our technique can be cast as a case-based, lazy learning approach in AI/ML terms. The described distance is computed between the query and each prototype, and a ranked list of the prototypes by increasing distance is returned. The ranking is useful because the user may see what are the candidate classifications, how close they are, and what specific parts of the shape are more relevant for the classification, rather than having to blindly trust a single outcome as in standard OCR. E.g., in Figure 2, the shape on the left might refer to a query character, and the shape on the right to a prototype character. Whenever the classification is not satisfactory, the shape of the misrecognized query character may be added to the prototypes, and from the very next moment the system will be able to recognize an extended set of characters, and presumably improve its accuracy and effectiveness. In ML terms, this is a form of totally incremental learning, i.e., learning may even start from scratch: the first query character will not be recognized and will be added to the set of prototypes, to form the first core of the recognition ‘model’. Since each class may be associated with many prototypes, even if the number of classes is kept constant, increasing the number of models in the database also increases the recognition effort.

4.4. Interface

Figure 2 shows at-a-glance the layout of the tool’s interface. On the left, the window for comparing a query character to prototypes. On the top, from left to right, the query character, its shape, and the prototype shape under consideration; on the bottom, under the shapes, the corresponding histograms, and under the query character the difference histogram (all histograms are aligned to the rotation/offset yielding the best matching). The overall difference value, and the rotation for which it was obtained, are also reported. Paleographers can appreciate where the specific differences in shape are located, both by looking at the side-by-side shape histograms, and at a glance by looking at the difference histogram, where the location and height of the peaks and valleys are evident. A heatmap-like system of colors in the shapes also allows them to see where the most similar or different shape borders are located. On the right of Figure 2, the global comparison of all prototypes to each other: the rows and columns of

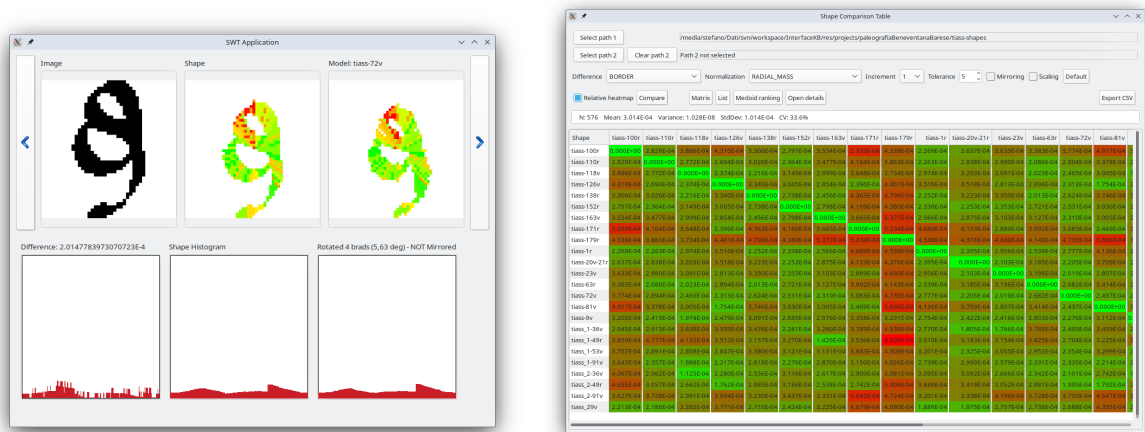


Figure 2: Tool interface showing the comparison between the query and a prototype for a character (left) and the pairwise similarities among prototypes for a character type (right)

the matrix are indexed by the prototypes, and each cell reports the difference between the prototypes in the corresponding row and column. The color of the cells, from green to red, immediately provides a heatmap-like assessment of the similarities and their distribution. Overall statistics are also reported above the matrix.

5. Evaluation

To test the proposed approach, we ran a quantitative evaluation, using a preliminary implementation in Java. The size and resolution of the page images was such that characters were about 50 pixel high, so we did resize the pages in this experiment. For binarization we tried both global (with manual setting or automated computation of the threshold) and local approaches. Using Niblack’s local approach with different parameter settings, we realized that, while being more complex and computationally demanding, it did not return significant improvements over the global ones, possibly due to the uniform level of illumination of the pages. Then, for the global ones, we realized that the optimal manual threshold found after trying several values was basically converging to the automated one computed by Otsu’s approach, so we opted for this one. For CC extraction, we used 8-neighbors. Concerning recognition, in our experiment there is only one class (the assimilated *ti*); the selected characters are stored as prototypes, and compared to each other, resulting in a ranked list by similarity.

As said, the preceding step of this research program measured the graphic variability of a single scribe in reproducing the assimilated *ti* by comparing twenty occurrences of the ligature against a single reference occurrence (f. 23v), whose representativeness the present analysis sets out to test. To this end, each of the 21 available occurrences of the ligature — the 20 samples already examined, together with f. 23v, here treated as one realization among others — is compared against all 20 remaining occurrences, rather than against a single fixed reference, resulting in a symmetric matrix of 210 unique distances. For each occurrence, we computed its average distance from all 20 others (‘row mean’): a low mean indicates an occurrence close to the scribe’s general behavior in this manuscript; a high mean indicates one that departs from it. These results do not depend on the subjective choice of a fixed prototype: rather, they describe the scribe’s degree of variability based exclusively on the internal structure of his realizations.

All 210 distances are on the order of E-4 on a [0, 1] theoretical scale: prior to any internal comparison, this is a first indication of an execution faithful to the canon. The mean across the 210 pairs is 3.128E-4, with a standard deviation of 8.131E-5. Together, these figures describe a limited degree of variability: the scribe reproduces the ligature with high graphic consistency (differences between occurrences are on the order of a few parts in ten thousand), not substantial variations of form. Ranking the 21

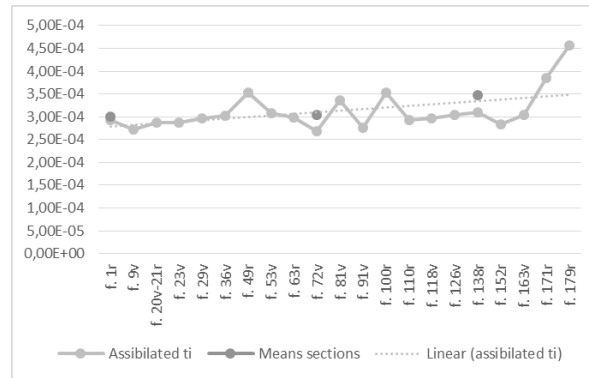


Figure 3: Row-mean distance values across the manuscript for character assimilated *ti*

occurrences by row mean yields a continuous distribution, not a sharp separation between forms: from the most central, f. 72v (2.681E-4), to the most peripheral, f. 179r (4.558E-4), nearly twice as far. The distribution is not uniform, however: the first 16 occurrences cluster within a narrow interval, spanning only 4.23E-5 in total, while the remaining 5 progressively draw apart, covering on their own an interval nearly three times as wide (1.20E-4). The dynamic of the row mean along the manuscript, folio by folio, is not entirely stable: it fluctuates noticeably even between consecutive folios — the relatively high value of f. 49r, for instance, is followed shortly afterwards by one of the lowest values in the entire corpus, at f. 72v — and the final part of the manuscript does not show uniformly high values either, since f. 152r falls among the more central occurrences while f. 171r and f. 179r are among the most peripheral. Extending the analysis to a larger number of samples, this instability would be worth examining in relation to possible resumptions of the copying campaign by the scribe. The trend line in Figure 3, calculated over the full series, shows an overall increasing direction. To identify possible anomalous occurrences, a threshold is computed as the mean of the row means plus twice the standard deviation (4.004E-4). The only occurrence exceeding it is, once again, f. 179r, confirming, by a second criterion, what the first had already indicated. Dividing the occurrences according to their position in the manuscript into three chronological blocks (ff. 1r-63v, ff. 72v-126v, ff. 138r-179r), the mean of the row means increases monotonically along the manuscript: 3.001E-4, 3.039E-4, 3.479E-4. Since f. 179r (third block) has already been identified as the corpus's sole anomalous occurrence, it is worth verifying how much of this increase depends on it: excluding it, the mean of the third block falls to 3.210E-4, still the highest of the three but by a reduced margin over the first two. The increase, therefore, is not attributable to a single anomalous sample alone, though it is measurably sharpened by it.

As regards the comparison with the ideal prototype, the first result this analysis returns bears directly on the assumption underlying the preceding stage of the research. F. 23v, the reference prototype, is the sixth (out of 21) in the centrality ranking: not the most representative occurrence overall, but not a peripheral one either. The paleographic choice that had designated it as the reference thus proves substantially correct, and the variability measure obtained in the preceding step is not invalidated, but only requires a correction in degree: a more central occurrence exists (f. 72v), and the scribe's true degree of variability, measured against it, differs slightly from that measured against f. 23v — but the difference is limited in extent, not such as to overturn the earlier result. The point on which the comparison with the fixed prototype is more clearly revised is the distribution along the manuscript. A measurement against a single fixed point registers only each occurrence's distance from that point, not occurrences' distance from one another: two occurrences equally distant from f. 23v can still differ considerably between themselves, and a single-reference design cannot detect this. That is what happens in the third block: its occurrences grew more divergent from each other without moving correspondingly away from f. 23v — which, being neither the corpus's most central occurrence nor located near the manuscript's end, is structurally less sensitive to change. Under the all-against-all method, what the fixed-reference method recorded as a flat pattern is a genuine increase in mutual graphic divergence among occurrences in the final part of the manuscript.

The comparison with the previously adopted method also invites a reflection on the paleographer's judgment — and thus on the paleographer's eye — in selecting the reference prototype. That the selection of f. 23v, an act of qualitative assessment grounded in direct observation, should fall so close to the top of the ranking produced by computation — albeit not coinciding with it — deserves to be read for what it says about the relationship between the two tools used in this research program. In examining the manuscript, the paleographer has no means of checking his choice against the totality of the 210 mutual relations among the occurrences: he can identify, through visual synthesis, an occurrence that appears representative, but cannot verify that choice, let alone quantify his own margin of error. That the choice nonetheless falls close to the center of the distribution indicates that the paleographer's eye grasps a genuine property of the scribe's degree of variability, even without precise quantitative tools to calculate it; that it does not coincide exactly with the most central occurrence marks the very limit of that judgment, one that computation can help to overcome. Overall, the picture emerges of a scribe who reproduces the ligature with high graphic consistency across the entire manuscript, with a single, circumscribed exception toward the end of the copying campaign.

6. Conclusion & Future Work

Paleographic research needs more than just transcription, the traditional objective of OCR/HTR. A support for thorough investigation of single characters and their features is needed. This adds complexity to the OCR task, which is by itself still quite challenging on ancient documents. In fact, available approaches to OCR may not be applicable to this task, for several reasons: they require large training datasets which may not be available, they may require flexible management and incremental extension of the character set, they just return one outcome with no interpretation or explanation for the recognition. This paper proposed and evaluated the adoption and adaptation of a novel incremental, prototype-based, interpretable approach to paleographic analysis of ancient documents.

While a previous work evaluated it on the first task of a writer recognition pipeline in paleographic research, the present contribution addresses a further step of that process, testing the representativeness of the reference prototype adopted in the first: the resulting all-against-all comparison confirms that the chosen prototype occupies a central, though not the most central, position within the corpus, and provides more general indications on the relationship between paleographic judgment and computational analysis. By deriving its measure of variability from the corpus itself, instead of a single externally chosen standard, the method offers a more neutral basis for this comparison than the fixed-reference design it extends. Where the two agree, computation confirms, with a precision that direct observation cannot achieve, what the trained eye had already grasped without the means to prove it; where they diverge, it marks the exact point beyond which judgment alone cannot reliably reach.

These results remain preliminary and should be treated as such: they concern a single ligature of the manuscript, not the scribe's entire graphic repertoire. The all-against-all method does not eliminate the need for paleographic judgment, but replaces an unverifiable assumption — the representativeness of the chosen prototype — with a verifiable one, and thereby proves itself the more reliable procedure for any subsequent stage of this research program requiring the identification of a representative term of comparison.

As future work, on the paleographic side we will extend the analysis to the remaining folios of the manuscript and to the other graphic elements, and fully assess, at that stage, the scribe's behavior. For the automated tool, we will investigate how to further adapt it so as to obtain more accurate and meaningful results for the paleographers. This may involve improving the representation of shapes in order to capture in more detail the character models (including their internal variations, beyond their external shape), and identifying strategies and good practices to set the pre-processing parameters that can enhance the quality of the blobs input to the recognition function. Improvements to the interface, for providing more intuitive support and more operational functions, are also planned.

Acknowledgments

This research was partially supported by project HERITAS, funded by the Italian Ministry of Universities and Research under “Avviso D.D. n. 307 del 18 marzo 2025”.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

- [1] S. Ferilli, D. Piscopo, Toward an AI-based tool for paleographic analysis – The scribal characterization step, in: *Proceedings of ISO LA 2026, Lecture Notes in Computer Science*, Springer, 2026 (to appear), p. 15.
- [2] R. Brause, B. Arlt, E. Tratar, Project Semacode: A Scale-invariant Object Recognition System for Content-based Queries in Images Databases, Technical Report 11/99 (FB20), Johann Wolfgang Goethe University, Computer Science Dept., Frankfurt/Main, 1999.
- [3] U. Springmann, A. Lüdeling, Ocr of historical printings with an application to building diachronic corpora: A case study using the ridges herbal corpus, *Digit. Humanit. Q.* 11 (2016).
- [4] J. Philips, N. Tabrizi, Historical document processing: A survey of techniques, tools, and trends, in: A. L. N. Fred, J. Filipe (Eds.), *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2020, Volume 1: KDIR*, Budapest, Hungary, November 2-4, 2020, SCITEPRESS, 2020, pp. 341–349.
- [5] P. Kahle, S. Colutto, G. Hackl, G. Mühlberger, Transkribus - A service platform for transcription, recognition and retrieval of historical documents, in: *1st International Workshop on Open Services and Tools for Document Analysis, 14th IAPR International Conference on Document Analysis and Recognition, OST@ICDAR 2017, Kyoto, Japan, November 9-15, 2017, IEEE*, 2017, pp. 19–24. doi:10.1109/ICDAR.2017.307.
- [6] B. Kiessling, R. Tissot, P. Stokes, D. Ezra, escriptorium: an open source platform for historical document analysis, in: *2019 international conference on document analysis and recognition workshops (icdarw)*, volume 2, IEEE, 2019, pp. 19–19.
- [7] R. Smith, An overview of the tesseract ocr engine, in: *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, 2007, pp. 629–633.
- [8] U. Springmann, D. Najock, H. Morgenroth, H. Schmid, A. Gotscharek, F. Fink, OCR of historical printings of latin texts: problems, prospects, progress, in: *Digital Access to Textual Cultural Heritage 2014, DATeCH 2014, Madrid, Spain, May 19-20, 2014, ACM*, 2014, pp. 71–75. doi:10.1145/2595188.2595205.
- [9] T. M. Breuel, A. Ul-Hasan, M. I. A. A. Azawi, F. Shafait, High-performance OCR for printed english and fraktur using LSTM networks, in: *12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, August 25-28, 2013, IEEE*, 2013, pp. 683–687. doi:10.1109/ICDAR.2013.140.
- [10] B. Kiessling, Kraken - a Universal Text Recognizer for the Humanities, 2019. doi:10.34894/Z9G2EX.
- [11] C. Wick, C. Reul, F. Puppe, Calamari - A high-performance tensorflow-based deep learning package for optical character recognition, *Digit. Humanit. Q.* 14 (2020).
- [12] A. Fischer, H. Bunke, N. Naji, J. Savoy, M. Baechler, R. Ingold, The HisDoc Project. Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries, De Gruyter, 2014, pp. 91–106. doi:doi:10.1515/9783110367317.91.
- [13] A. Fischer, A. Keller, V. Frinken, H. Bunke, Lexicon-free handwritten word spotting using character hmms, *Pattern Recognit. Lett.* 33 (2012) 934–942. doi:10.1016/J.PATREC.2011.09.009.
- [14] A. Fischer, M. Baechler, A. Garz, M. Liwicki, R. Ingold, A combined system for text line extraction

- and handwriting recognition in historical documents, in: 11th IAPR International Workshop on Document Analysis Systems, DAS, IEEE, 2014, pp. 71–75.
- [15] S. Ammirati, D. Firmani, M. Maiorino, P. Merialdo, E. Nieddu, In codice ratio: Machine transcription of medieval manuscripts, in: Digital Libraries: Supporting Open Science - 15th Italian Research Conference on Digital Libraries, IRCDL Proceedings, volume 988 of *Communications in Computer and Information Science*, Springer, 2019, pp. 185–192. doi:10.1007/978-3-030-11226-4_15.
 - [16] S. Ferilli, T. Basile, F. Esposito, M. Biba, A contour-based progressive technique for shape recognition, in: Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR-2011), volume 1, IEEE, 2011, pp. 723–727.
 - [17] D. Chapinal-Heras, C. Díaz-Sánchez, A review of ai applications in human sciences research, Digital Applications in Archaeology and Cultural Heritage 32 (2024) e00323. doi:10.1016/j.daach.2024.e00323.
 - [18] F. Aiolfi, M. Simi, D. Sona, A. Sperduti, A. Starita, G. Zaccagnini, Spi: A system for paleographic inspections, AIIA Notizie 4 (1999) 34–38.
 - [19] A. Ciula, Digital palaeography: Using the digital representation of medieval script to support palaeographic analysis, Digital Medievalist 1 (2005). URL: <https://journals.openedition.org/digitalmedievalist/773>.
 - [20] M. Panagopoulos, C. Papaodysseus, P. Rousopoulos, D. Dafi, S. Tracy, Automatic writer identification of ancient greek inscriptions, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 1404–1414. doi:10.1109/TPAMI.2008.282.
 - [21] E. Lowe, The Beneventan Script: A History of the South Italian Minuscule, Clarendon Press, Oxford, 1914.
 - [22] G. Cavallo, Struttura e articolazione della minuscola beneventana libraria tra i secoli x–xii, Studi Medievali 11 (1970) 343–368.
 - [23] R. Szeliski, Computer Vision: Algorithms and Applications, Springer, 2011.
 - [24] X. Shu, X.-J. Wu, A novel contour descriptor for 2d shape matching and its application to image retrieval, Image and Vision Computing 29 (2011) 286–294.
 - [25] S. Ferilli, Automatic Digital Document Processing and Management: Problems, Algorithms and Techniques, Springer, 2011.
 - [26] D. Zhang, G. Lu, A comparative study of curvature scale space and fourier descriptors, Journal of Visual Communication and Image Representation 14 (2003) 41–60.