Identifying, Classifying and Searching Graphic Symbols in the NOTAE System

Maria Boccuzzi[1], Tiziana Catarci[2], Luca Deodati[3], Andrea Fantoli[3], Antonella Ghignoli[1], Francesco Leotta[2(✉)], Massimo Mecella[2], Anna Monte[1], and Nina Sietis[1]

[1] Dipartimento di Storia Antropologia Religioni Arte Spettacolo, Sapienza Università di Roma, Rome, Italy
{maria.boccuzzi,antonella.ghignoli,anna.monte,nina.sietis}@uniroma1.it
[2] Dipartimento di Ingegneria Informatica, Automatica e Gestionale, Sapienza Università di Roma, Rome, Italy
{catarci,leotta,mecella}@diag.uniroma1.it
[3] Facoltà di Ingegneria dell'Informazione, Informatica e Statistica, Sapienza Università di Roma, Rome, Italy
{fantoli.1467336,deodati.1488696}@studenti.uniroma1.it

Abstract. The use of *graphic symbols* in documentary records from the 5th to the 9th century has so far received scant attention. What we mean by graphic symbols are graphic signs (including alphabetical ones) drawn as a visual unit in a written text and representing something other or something more than a word of that text. The Project NOTAE represents the first attempt to investigate these graphic entities as a historical phenomenon from Late Antiquity to early medieval Europe in any written sources containing texts generated for pragmatic purposes (contracts, petitions, official and private letters, lists etc.). Identifying and classifying graphic symbols on such documents is a task that requires experience and knowledge of the field, but software applications may come in help by learning to recognize symbols from previously annotated documents and suggesting experts potential symbols and likely classification in newly acquired documents to be validated, thus easing the task. This contribution introduces the NOTAE system that, in addition to the aforementioned task, provides non expert users with tools to explore the documents annotated by experts.

Keywords: Graphic symbols · Paleography · Image processing · Clustering

This work is supported by the ERC grant *NOTAE: NOT A writtEn word but graphic symbols*, funded by the European Union's Horizon 2020 research and innovation programme (grant agreement no. 786572, Advanced Grant 2017, PI Antonella Ghignoli). See also http://www.notae-project.eu.

The original version of this chapter was revised: The original version of the chapter 12 was previously published non-open access. It has now been changed to open access under a CC BY 4.0 license and the copyright holder has been updated to 'The Author(s).' The book has also been updated with the change. The correction to this chapter is available at https://doi.org/10.1007/978-3-030-39905-4_19

M. Ceci et al. (Eds.): IRCDL 2020, CCIS 1177, pp. 111–122, 2020.
https://doi.org/10.1007/978-3-030-39905-4_12

1 Introduction

With the gradual introduction of signature and the increasing use of papyrus from the 4th century, the presence of *graphic symbols* became widespread in legal documents as it already was in other written records, and continued in post-Roman kingdoms as part of the same historical process of reception of the late antique documentary practice. The sources of this practice - records written both in greek and latin, on diverse supports as papyrus, wooden tablet, slate, parchment - are expression of the so called "pragmatic literacy", defined in [8] as the "literacy of one who has to read or write in the course of transacting any kind of business".

A new approach to studying documents meant as complex systems of written texts and graphic devices was introduced in the 1990s by Rück [10,12]: morphology, semantics, syntax, pragmatic function and changes over time of the symbolic elements of a document were explicated involving results and concepts of other disciplines, e.g., archaeology, numismatics, semiotics, anthropology. The well-studied subjects in the new field of Diplomatics promoted by Rück have been however the striking graphic features of the charters issued by rulers and elites or written by public notaries of high medieval Europe (10th–12th century), and the few comparative analyses have so far been conducted mainly on high and late medieval western sources. Recent years have seen a renewed interest in the graphic aspects of early medieval written sources, and some works have shown also the connection between Late Antiquity and early Middle Ages; they have dealt, however, only with some specific signs (crosses, christograms and monograms) selected in advance as graphic signs of identity, faith, and power, disseminated in diverse media, but not in documentary records. The project NOTAE represents therefore the first attempt to conduct a research on graphic symbols in documentary records from Late Antiquity to early medieval Europe from a novel perspective: in the long historical period in question, drawing symbols had a major social impact, because, provided it was done in one's own hand, it placed on the same footing professional scribes, basic literates and illiterates. For illiterates, it certainly meant, both in the late Roman state (a Greek-Latin graphic and linguistic community) and in the post-Roman kingdoms (as long as Latin functioned as language of vertical communication) a way of taking an active part in the writing process. A thorough investigation of this 'other side' of the written world can therefore provide precious insights about the spread of literacy as a whole.

In this novel perspective graphic symbols are meant as graphic entities (composed by graphic signs, including alphabetical ones or signs of abbreviation [11]) drawn as a visual unit in a written text and representing something other or something more than a word of that text. The message they carry on is to be discovered, because there is no intrinsic prior relationship between the message-bearing graphic entity and the information it conveys. Examples of frequently employed graphic symbols are shown in Fig. 1.

Identifying and classifying graphic symbols on documents is a task that requires experience and knowledge of the field; specific software applications may support this task, by learning to recognize symbols from previously annotated documents and suggesting experts potential symbols and

Fig. 1. Examples of graphic symbols. *(a)* graphic symbol in a complex structure at the end of the autograph subscription of a greek notary. *(b)* autograph symbol (greek cross and diagonal cross crossing each other) of an illiterate seller. *(c)* graphic symbol in a complex structure at the end of the autograph subscription of a witness. *(d)* staurogram and $\chi\mu\gamma$-group at the end of the final datation written by a notary. *(e)* autograph diagonal cross (or letter χ) of an illiterate man. *(f)* graphic symbol in complex structure at the end of the autograph subscription of a bishop. *(g)* autograph greek cross of an illiterate clerk.

likely classification in newly acquired documents to be validated. This contribution introduces the NOTAE system that, in addition to the aforementioned task, provides non expert users with tools to explore the documents annotated by experts. The system is part of the NOTAE project [5], which broadly studies the employment of graphic symbols in documents, in relation to historical and geographical contexts.

The paper is organized as it follows. Section 2 relates this work in the wider area of digital paleography and digital humanities. Section 3 initially describes the approach with a bird-eye view, and then specifically describes each component of the system. Section 4 finally concludes the paper also describing the ongoing validation and future research directions.

2 Related Works

Paleography is the study of ancient and historical handwriting. Included in this discipline, it is the practice of deciphering, reading, and dating historical manuscripts, and the cultural context of writing.

The approach proposed in this paper falls into the category of computing systems applied to paleography [6], which are part of the digital paleography [3] area belonging to the wider field of digital humanities [1]. Digital humanities represent a field of study that raised a growing interest from the research community and funding agencies as witnessed by the number of project recently funded at national and European level (e.g., D-Scribes[1], NEPTIS [7], etc.).

In the field of digital paleography, our task is to identify and classify graphic symbols. At the best of our knowledge, this represents a completely new research area. This task is much different from handwritten text recognition, which is the goal of many recent works such as [4] and of research projects like the above mentioned D-Scribe, whose aim is to make papyrologists able to look for similar, or identical, handwritings to a given papyrus and for typical samples of writing for a given period. Anyway, symbols have indeed several characteristics making them different than words. For example, it is impossible to employ row spacing and spaces between words to identify them, being usually placed in casual position with respect to the text, overlapping to it and covering several lines.

Commonly to many other tasks in digital humanities, a prior step for graphic symbols identification and classification is preprocessing the digital reproduction of the documents and, in particular, binarizing it. The binarization of documents is a particularly hot topic as witnessed by the availability of challenges such as DIBCO [9]. In the case of the NOTAE system, the binarization is particularly difficult as physical supports varies both in terms of material (e.g., papyrus, slate) and preservation conditions.

3 Proposed Approach

Figure 2 depicts the approach proposed in this paper. Documents considered in the NOTAE project are available either in museums' showcases or, more and more frequently, through digital reproductions in public web repositories and aggregators[2]. An expert who wants to identify and study graphic symbols in a specific document, usually inspects its digital reproduction together with associated bibliography. Visual inspection is, generally, an hard task, especially if executed on several documents in a working session, due to the state of conservation of the document, the color of the background, and the possible overlapping between graphic symbols and normal text.

The NOTAE system simplifies the work of experts in the NOTAE project by providing a Web app that assists them by providing useful functions to ease their job by leveraging on previously acquired annotations.

When a NOTAE expert, also referred to as curator, is studying a document, s/he can upload a digital reproduction (see step (1) in Fig. 2), taken from a public repository, in the NOTAE Curator Web app. The Web app communicates with the intelligent core of the NOTAE system, the Symbol Engine, through the Classification API. The Symbol Engine is a python service which exploits

[1] See https://d-scribes.philhist.unibas.ch.
[2] See, for example, https://papyri.info/.

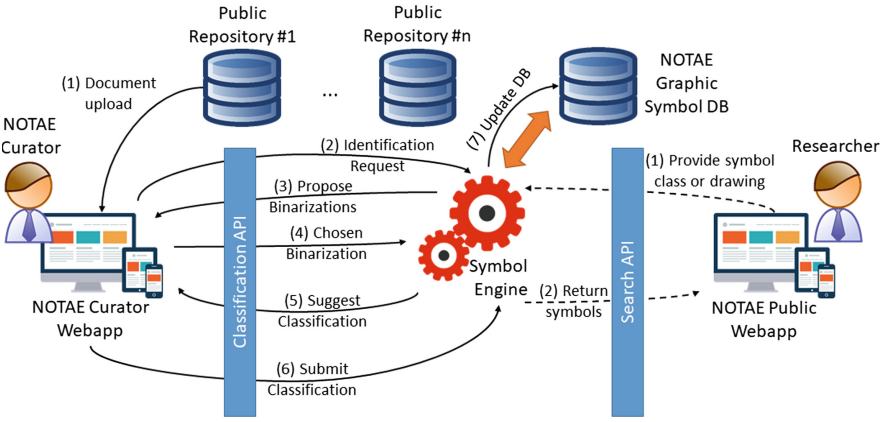

Fig. 2. The proposed approach.

OpenCV [2] for image processing tasks. Once a document has been uploaded, the curator Web app shows five different possible binarizations of the document (steps (2) and (3), see Sect. 3.1. At this point, the curator chooses the binarization version that preserves more the original content (step (4)). Once a binarization has been selected, the Symbol Engine returns a picture containing all the graphic symbols contained in the document (step (5), see Sect. 3.2).

The symbols identified by the Symbol Engine are not supposed to be perfect, as the main goal is to provide experts with potential symbols, leaving to them and to their expertise the burden of classifying symbols. As a consequence, curators send a feedback about identified and classified symbols (step (6, see Sect. 3.3). As a last step, expert classification of symbols is stored inside the NOTAE Graphic Symbol DataBase (step (7)).

The NOTAE Graphic Symbols DataBase stores information about graphic symbols contained in the documents within the scope of the project. Documents are referenced by using identifiers that are globally recognized in the research community. Information does not only include their presence in a specific document, but also additional details such as comments about their usage. In the NOTAE system, one of the goal of the NOTAE Graphic Symbols DB is to be used as a reference to detect symbols in newly uploaded documents. At the bootstrap, this database is empty and, as a consequence, the very first identification tasks simply provide no results to the expert user. With experts continuously introducing new classifications of symbols, the database is progressively populated (step (7)) with graphic symbols, and this allows to increasingly refine the identification and classification skills of the Symbol Engine.

Beyond providing a mean to identify and automatically classify symbols in documents, the Symbol Engine also serves as a search engine. The final users of this engine are researchers (this category includes the NOTAE curators themselves), who will access through the NOTAE Public Web app This one employs a specifically designed API, called Search API, to navigate the content of the

Fig. 3. Color clusters obtained from selected documents. (Color figure online)

NOTAE Graphic Symbol DB by performing the research activity using symbols category or event drawing (see Sect. 3.4).

3.1 Binarization

As previously stated, the binarization of documents of interest for NOTAE curators is hard due to issues related to the nature of the physical supports and their conservation status. A standard way to binarize an image is to choose a threshold and to convert every pixel of the image to 1 if a specific combination of the components of the color of the pixel is above/below the threshold, to 0 otherwise. The choice of the threshold depends on the specific document and is not possible to find a one-fits-all solution. As the approach followed by the NOTAE system consists in involving the curators in a man-in-the-loop process, we decided to follow an approach that leaves the choice of the threshold to the curator. Instead of allowing him to choose any value though, we applied an approach that computes five possible thresholds based on the employment of K-means. It is a centroid-based clustering algorithm that, given a set of points in a coordinate space, the RGB color space in our case, groups them into K sets, where the parameter K is given as an input to the algorithm according to some distance function, usually the euclidean distance. The output of the algorithm is a set of K centroids providing the center of each group in the coordinate space. These centroids will be in our case specific colors. Figure 3 shows the result of the execution of K-means with K set to 4 on four different documents. Colors are ordered from left to right according to an increasing red component.

Once the clustering step has been executed, the five thresholds are computed as *(i)* the red component of the first color extracted which is different from black, *(ii)* the red component of the second color extracted which is different from black, *(iii)* the average value between the first two threshold values, *(iv)* the average between the first and third thresholds, and *(v)* the average between the second and third thresholds.

Once the thresholds have been determined, we move on to the effective removal of the background and generation of the binary images. To this aim, we compare the red component of every pixel in the image to a specific threshold. If the value is above the threshold we turn the pixel to white, leaving the value unchanged otherwise. At this point a very simple filtering, i.e., erosion, is

Fig. 4. The proposed binarization options.

applied to bring to the foreground the marginal components of the writing. At this point, we turn every pixel different from 1 into 0. As shown in Fig. 4, the obtained results are five different versions of the document, among which the curator will choose the best one for detecting symbols.

One issue with binarization is that digital reproductions of documents in public repository are usually at very high resolution. This makes the above described process very slow, as clustering must be applied to millions of pixels. To limit the binarization time, we decided to resize the original image, working with a small-sized version of the document and the original-sized version. The small-sized version is obtained by downsizing the original reproduction to 10% of its size. This version is employed for clustering and choosing the threshold. The original-sized version is instead used for actual binarization, which is instead a quick task.

3.2 Symbol Identification and Classification

Once the curator has chosen the best binarization of the document, the NOTAE Symbol Engine can identify symbols. In order to perform this task, we apply template matching. This is a method for searching and finding the location of a template image in a larger image. It simply slides the template image over the input image and compares the template with the current part of the original image. Several comparison methods are possible. Template matching returns a grayscale image, where each pixel denotes how much the neighborhood of that pixel matches the template. Template matching is performed using all different symbols available in the NOTAE System DB.

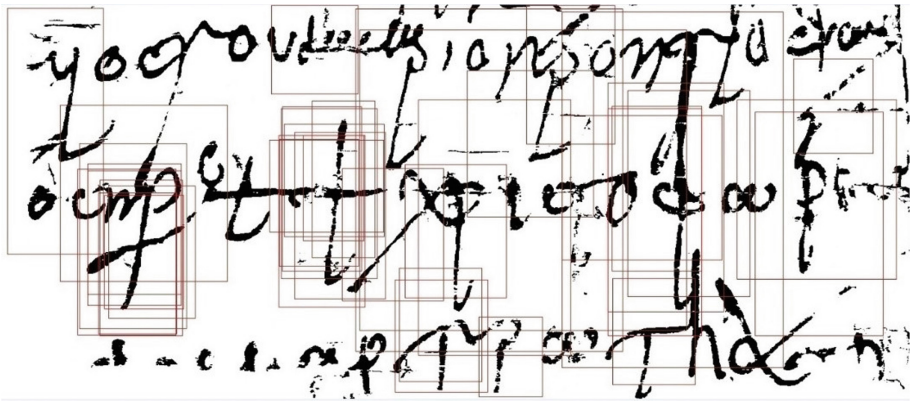

Fig. 5. Results of symbol identification without clustering.

As the size in terms of pixels of the symbols in the NOTAE System DB depends on the original document, applying template match to a single version of a document would likely return several false negatives. As a consequence, we apply template matching to several different scales of the document, keeping track of the best match, and after looping over all scales, we take the global best match as our matched region. At each step, we collect the coordinates of each match for each template used, and we use these coordinates to draw different rectangles on the image.

The result on an example document is shown in Fig. 5. As can be easily understood, the obtained result is not satisfying, as it contains too many rectangles and too many false positives. To solve this issue, we used DBSCAN - Density-Based Spatial Clustering of Applications with Noise.

DBSCAN, differently from K-means employed for binarization, finds clusters of similar density, thus allowing to obtain clusters of any shapes. We fed all the centers of the symbols identified at the previous steps into DBSCAN. DBSCAN takes as input two parameters *min_samples* and ϵ, the first one representing the minimum number of points that form a cluster, and the second one representing the maximum distance between two samples for one point to be considered as in the neighborhood of the other. DBSCAN also returns points that were not included in any clusters. Then, iterating on the previously populated array of coordinates we have checked if the corresponding center was part of one of the clusters obtained thanks to the aforementioned function, and if this happens the user will see that rectangles on the image related to the papyrus.

As shown in Fig. 6, in this way the search for symbols returns the most plausible ones. Clearly, the more symbol templates the NOTAE System DB contains, the more accurate the search will be. This fits very well with a system intended to operate with a human-in-the-loop approach, as curators can provide feedback about the performed identification, making the system smarter and smarter.

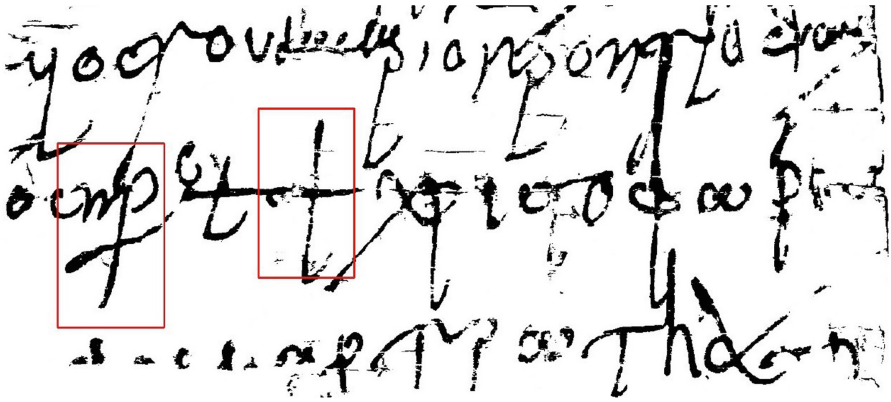

Fig. 6. Results of symbol identification with DBSCAN clustering.

As in the case of binarization, the bottleneck of this method is certainly the dimension of the digital reproduction of the document. Thus, the search of the symbols is performed on a version of the document scaled to 20% of the size. The result is instead shown on the original-sized version of the document.

Noteworthy, the very same mechanism employed for graphic symbol identification can be employed for symbol classification. As discussed, the identification of symbols is performed by looking for all symbols available in the database inside the digital reproduction of the document. For each symbol in the symbol database, a classification, provided by the experts as described in Sect. 3.3, is available reporting the type of symbol. By computing the type with the majority of occurrences in a DBSCAN cluster, the system assigns a class to each identified symbol. This class will be then confirmed or refused by the experts.

3.3 Manual Symbol Classification

Identification and classification results are shown to the expert using the window shown in Fig. 7, where identified symbols are marked with red boxes. This window provides the user with different functionalities.

In particular, given an identified and classified symbol, the expert can tell the system whether a box really contains a symbol and if the classification was correct, suggesting the system a different classification. Symbols validated by the expert are marked with a green box.

The curator can also select, using drag and drop, an area of the document where a symbol is present but nothing has been detected by the system. At this point, the symbol engine returns, for the given selection, the most likely classification that can be confirmed or corrected by the expert.

Noteworthy, the involvement of the expert in the process is deep. The system is initially dumb and becomes smarter and smarter thanks to the corrections and the labelling performed by the user. This allows us to overcome the lack of labelled datasets in the field of graphic symbols.

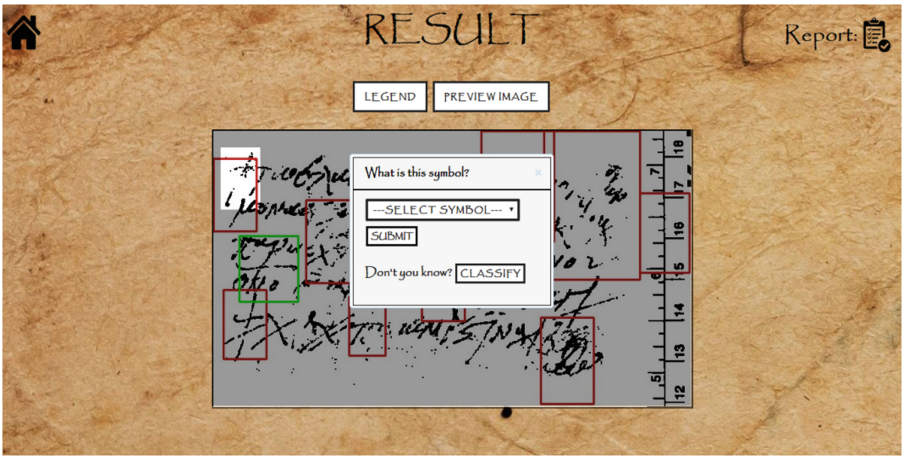

Fig. 7. The manual classification window. (Color figure onine)

3.4 Searching Symbols

Experts are not the only users of the NOTAE system. The aim of the NOTAE project is to provide researchers in the area of paleography with a useful tool to explore the world of graphic symbols and their usage. This consideration involves the necessity for implementing advanced query interfaces for search and analysis.

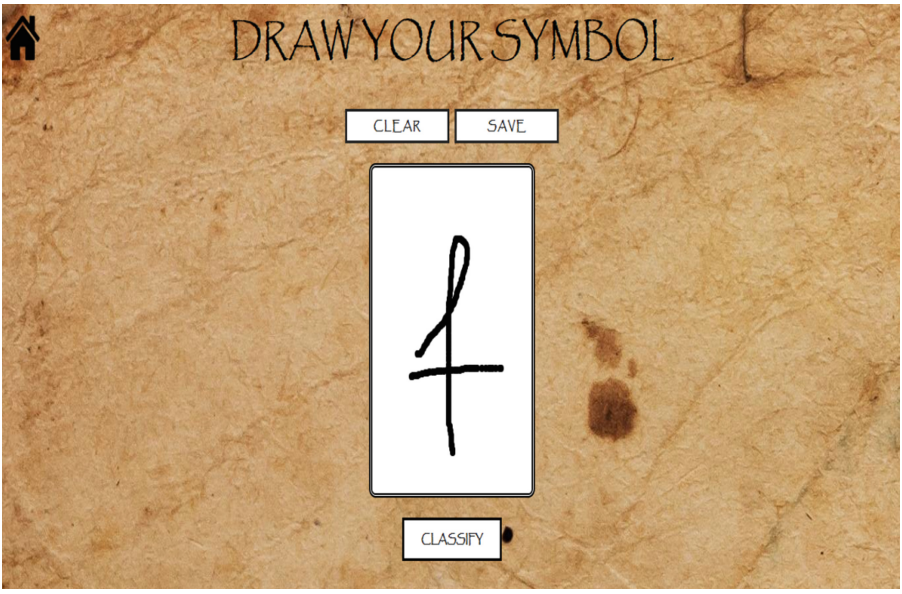

Fig. 8. The manual drawing input window.

At the current state, two search mechanisms have been implemented: search by type and search by drawing. *Search-by-type* is implemented as a simple query inside the database of symbols. In this case, the Symbol Engine retrieves from the database all the symbols labeled from expert users with the required category (notably, not using any image processing technique but simply by taking into account the classifying labels previously assigned).

The *search-by-drawing* mechanism is instead intended for less expert users. In this case, the user can draw through the window shown in Fig. 8 a symbol, maybe casually seen in a document, and search for it inside the symbol database. This feature allows the user to draw a symbol, with the trackpad or with the mouse on a canvas and to receive from the system the classification of what has been drawn. This classification is implemented likewise the identification and classification mechanism seen in Sect. 3.2. Here, instead of matching available symbols within an entire document, the comparison is performed against the user drawing.

4 Concluding Remarks

In this paper, we have presented the NOTAE system as the component of the NOTAE project in charge of automatically identify and classify graphic symbols in Late Antique and Medieval documents. The contribution, in particular, introduced all the different components of the system.

At the actual stage, the system is in use to the NOTAE project members who are conducting the evaluation of the system. In particular three experts are involved in evaluating the ability of the system at *(i)* identifying symbols in documents with respect to the amount of documents already labelled, *(ii)* classifying identified symbols, *(iii)* searching symbols by manual drawings. Evaluation will be also performed against the DIBCO challenge [9] in order to globally assess the performance of binarization.

The public Web app will be made available in the following months from the web site of the NOTAE project: www.notae-project.eu.

The current version of the system suffers from some limitations including the fact that negative feedbacks from experts, i.e., the fact that an expert mark an identified symbol as a false positive or wrongly classified, are not taken into account in the identification and classification process. A future research step will consist in including these negative feedbacks in the classification task, e.g., by including negative matching scores. Additionally, as the identification and classification process involves the comparison of all symbols in the database, this task is supposed to become slower and slower while the size of the symbol database grows.

Other future research directions include the possibility to automatically analyze symbols category by *(i)* highlighting geometric shapes patterns, i.e., recognizing the component shapes and their arrangement in graphic symbols categories, *(ii)* produce heat maps of the positions of graphic symbols categories, and *(iii)* define visual analytic tools to correlate the employment of symbols with the historical and geographical context.

References

1. Berry, D.M.: Introduction: understanding the digital humanities. In: Berry, D.M. (ed.) Understanding Digital Humanities, pp. 1–20. Palgrave Macmillan UK, London (2012). https://doi.org/10.1057/9780230371934_1
2. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, Inc., Newton (2008)
3. Ciula, A.: Digital palaeography: using the digital representation of medieval script to support palaeographic analysis. Digit. Medievalist **1** (2005)
4. Firmani, D., Maiorino, M., Merialdo, P., Nieddu, E.: Towards knowledge discovery from the vatican secret archives. In codice ratio - episode 1: machine transcription of the manuscripts. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, 19–23 August 2018, pp. 263–272 (2018)
5. Ghignoli, A.: The notae project: a research between east and west, late antiquity and early middle ages. Comp. Orient. Manuscript Stud. Bullettin **5/1**, 27–41 (2019)
6. Hassner, T., Rehbein, M., Stokes, P.A., Wolf, L.: Computation and palaeography: potentials and limits. Kodikologie und Paläographie im Digitalen Zeitalter 3: Codicology and Palaeography in the Digital Age **3**, 1 (2015)
7. Mecella, M., Leotta, F., Marrella, A., Palucci, F., Seri, C., Catarci, T.: Encouraging persons to visit cultural sites through mini-games. EAI Endorsed Trans. Serious Games **4**(14), e3 (2018)
8. Parkes, M.B.: Scribes, Scripts and Readers: Studies in the Communication, Presentation and Dissemination of Medieval Texts. Hambledon Press, London/Rio Grande (1991)
9. Pratikakis, I., Gatos, B., Ntirogiannis, K.: H-DIBCO 2010-handwritten document image binarization competition. In: 2010 12th International Conference on Frontiers in Handwriting Recognition, pp. 727–732. IEEE (2010)
10. Rück, P.: Graphische Symbole in mittelalterlichen Urkunden: Beiträge zur diplomatischen Semiotik, vol. 3. Jan Thorbecke Verlag (1996)
11. Sietis, N.: Abbreviations in Greek documentary texts. a case study of significant paleography. In: Conference on Novel Perspectives on Communication Practices in Antiquity, pp. 2–5 (2019)
12. Worm, P.: Ein neues bild von der urkunde: Peter rück und seine schüler. Archiv für Diplomatik **52**(JG), 335–352 (2006)