Database of fragments of medieval codices of the 11th-12th centuries – the uniqueness of requirements and data

Jakub Leszek Pach[1,2][0000−0002−3483−4340]

[1] Warsaw University of Life Sciences – SGGW
Nowoursynowska 166, 02-787 Warsaw, Poland
[2] The National Library of Poland
al. Niepodległości 213, 02-086 Warsaw, Poland
jakub_pach@sggw.edu.pl

Abstract. This paper presents a new offline dataset called the Fragments of Medieval Codices (FOMC). It contains medieval Latin handwritings coming from 11th-12th century and can be used to evaluate the performance of offline writer identification and to find the handwriting similarity between the writers, or to test the handwritten optical character recognition systems. It consists of 117 fragments of handwritten documents of medieval codices and contains in total over two thousand very high quality images. The collection was assembled using the IIIF standard. We describe the collecting and processing steps performed to develop the dataset and define several evaluation tasks regarding the use of this dataset.

Keywords: Latin manuscripts database · offline writer identification · optical character recognition.

1 Introduction

Databases of handwritten texts in image processing have a dual purpose, which are the identification of writers and the text recognition. For the first purpose, correct classification is possible without the need to recognize letters and, consequently, the content. In the case of the second purpose, it is necessary to correctly identify each character – a letter – to recognize the content of the document. OCR and writer identification can work together, because this makes it possible to analyze the similarity of specific characters. Both purposes already have many dedicated methods that perform the task with very high efficiency for modern writing.

However, the identification of Latin manuscripts is still a major challenge today, partly because of how medieval writing was done and what it was written on. Writing materials, e.g. parchment, were incomparably more expensive than paper used today is, and this resulted in a much tighter written text, in order to save every fragment of such a valuable material. The letters of the text often

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_8

overlapped, the margins contained comments by copyists and descendants, and some of the content was written in complicated abbreviations. The ability to decipher the content of such documents is now only within the reach of specialist historians. Hence, in order for such writer identification systems to implement the discussed topic, interdisciplinary cooperation of highly qualified specialists in the discipline of history and computer science is necessary. The second reason why there is still so much to be done in the discussed topic is the lack of databases on which these systems could be taught and refined, while for modern writings such databases are easier to find. Therefore, in this paper a database of Latin manuscripts from the 11th-12th centuries, which is supposed to help in this matter, is presented. The database is accessible at [15].

The remaining part of this paper is organized as follows. In the second section, an overview of existing databases in the field of writer recognition is provided. In section 3, a detailed description of the presented database and its analysis is given. Section 4 contains conclusions and discussion on further work with the database.

2 Existing databases

The criterion of selection of the databases described here is their high frequency of use. The author searched databases of papers in the field of author identification and handwriting recognition coming from the last two decades, with the largest scientific value measured by the number of citations, and selected those which are most often used to validate systems of writer identification. If a base is no longer available or it is very rarely cited by other authors, such base is omitted. Due to that the Latin scripts are the object of studies of the author, the use of the Latin alphabet script was the key to the selection of reference databases. Non-alphabetic script appears only as a reference.

The description of the databases is made in alphabetical order.

The BFL Database [4] was made in the first decade of 21th century in Latin America that has 315 writers, three samples per writer, in Portuguese language.

The CERUG-MIXED dataset [17] established in University of Groningen also in 21th century contains handwritten documents collected from 105 writers, four samples per writer that wrote in two different languages – Chinese and English.

The next database which is the second most commonly used database written in modern English and German is CVL [8]. It contains 311 different writers where one writer wrote between five and seven samples, mainly in English but one sample is in German. This database was created in 21th century as well.

On the last year of the second millennium, the Firemaker [1] database was created that contains 1000 images of scanned handwritten text, containing pages of text written by 250 writers, four pages per writer, in Dutch.

The GDRS dataset [11] is next database generated by a research group in the Computational Intelligence Laboratory at the National Center for Scientific Research "Demokritos", Greece, in 21th century. 26 writers made eight sample texts written in four languages, that is, English, German, Greek and French.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_8

The following database IAM [14] is probably the most commonly used dataset in writer identification and contains 657 writers, where writers wrote samples between one to almost sixty. The part of this database which is used mainly is known as MIAM (Modified IAM).

Another popular historic database is IAM-HistDB [3]. It contains three medieval manuscripts from 9th, 13th and 18th century and has 127 pages.

The next benchmarking dataset ICDAR2013 [10] was created in 21th century with the help of 250 writers who wrote four parts of text in two languages: English and Greek. This base contains the whole ICDAR2011.

The largest historical database written in Latin is ICDAR2017 [2] known also as HistoricalWI-2017. It contains 3600 handwritten pages originating from 13th to 20th century. It contains manuscripts from 720 different writers where each writer contributed five pages.

The IFN/ENIT database [16] contains Arabic handwriting from 411 writers, five samples each.

Database JEITA-HP [9] was prepared by Hewlett-Packard Japan and distributed by Japan Electronics and Information Technology Industries Association, in Japanese. It consists of two datasets: Dataset A and Dataset B, which store handwritten character patterns from 480 writers and 100 writers, respectively.

KHATT database [13] is a database of unconstrained handwritten Arabic Text written by 1000 different writers, four samples per writer.

The next offline dataset [12] is called the Qatar University Writer Identification dataset (QUWI). This dataset contains both Arabic and English handwritings. It consists of handwritten documents of 1017 writers, four samples each. The last database is RIMES [5] and was created to evaluate automatic systems of recognition and indexing of handwritten letters. The collection was a success with more than 1,300 people who have participated to the RIMES database creation by writing up to five mails (handwritten correspondence) in French and Bengali.

The information an all databases have been compiled in Table 1 for easier comparison. Let us analyze those of the databases with have a large historical component.

The IAM-HistDB database is composed of three codes that were created in different historical eras, styles and languages. This is profitable for a machine learning because there is a transcript of content and it is also possible to train a neural network to read content from the same age in other manuscripts. However, when it comes to author identification, three writers are not sufficient to properly validate the effectiveness of writer identification systems.

In addition, from the perspective of historical science, it is necessary to take into account the fact that two scientific disciplines, paleography and neography, deal with handwriting in this period of time. Paleography deals with handwriting until the creation of the first incunabula – the first printed books (until the 15th century), and since then, neography has been dealing with modern handwriting. The fact that the separation of two separate scientific disciplines was necessary

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_8

Table 1. Comparison between datasets.

Database acronym	Language	No. of writers	No. of images	Samples per writer	Type of a sample	Rounded width of a sample	Age of samples (century)
BFL [4]	Portuguese	315	945	3	page	2500	21
CERUG-MIXED [17]	Chinese; English	105	420	4	page	-	21
CVL [8]	English; German	310	1604	5-7	page	1800	21
Firemaker [1]	Dutch	250	1000	4	page	2500	20
GRDS [11]	English; German; Greek; French	26	208	8	page	-	21
IAM [14]	English	657	1539	1-57	page	-	21
IAM-HistDB [3]	Latin; Medieval German; English	3	127	60; 47; 20	page	3300	9; 13; 18
ICDAR2013 [10]	Greek	250	1000	5	page	2500	21
ICDAR2017 [2]	Latin	720	3600	5	page	<1000; 1500;	13-20
IFN/ENIT [16]	Arabic	411	2200	5	page; words	2600	21
JEITA-HP [9]	Japanese	480; 100	-	3306	character	-	21
KHATT [13]	Arabic	1000	4000	4	page	2000	21
QUWI [12]	Arabic; English; French	1017	5085	4	page	-	21
RIMES [5]	French; Bengali	1300	12093	9	line	<1000	21
FOMC [15]	Latin	117	2040	11-39	page	>2000	11-12

for the analysis of handwriting indicates that the changes in the handwriting itself and its evolution were so large that it is not justified to compare them together, because they are so different.

It is important that the oldest document from this database was written on parchment, which accounted for a very significant percentage of the cost of creating the codex, in contrast to the last one written on paper.

To emphasize how expensive parchment was in the Middle Ages, there was the practice of scraping off the contents of a document no longer in use and writing it down again (palimpsest). As writing has developed over time, writing materials have evolved. The replacement of parchment from use by its much cheaper equivalent – paper – has had profound consequences. One of them, which is important in the aspect of the analysis of manuscripts in the discipline of computer science, is the reduction of the text density in relation to the document format. Therefore, the second historical database ICDAR2017 contains 720 writers, each represented by five samples. It was created to validate the correct identification of the author of the manuscript. However, the problem of the huge time interval over which the samples were written carries the same problems as the previous database.

The oldest documents from this database are from the 13th century and are written on parchment, in a completely different style and according to different rules than the last ones written in the 20th century on paper. For example, two

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_8

manuscripts from this database are shown, which are treated as equal in relation to each other (see Fig. 1). Figure (a) shows a manuscript that contains over 3 million pixels, and figure (b) contains only 700 000. If the amount of writing material between these two samples is taken into account, a significant disproportion appears, and it is not an isolated case in this database. Consider that some writers are represented with only five samples. For a classifier, one writer will be underrepresented and the other will be overrepresented, in the extreme case. Of all the databases included in Table 1, only ICDAR2017 and IAM-HistDB are

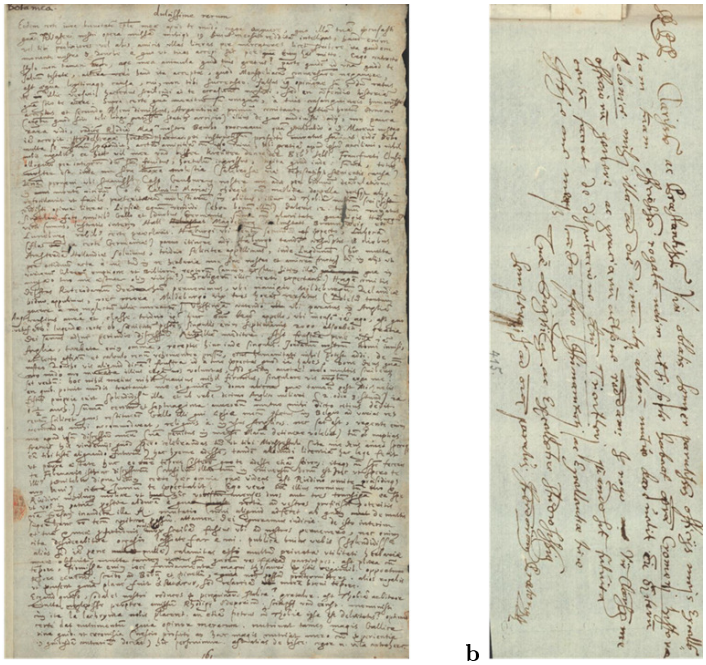

Fig. 1. Two samples of manuscripts from the ICDAR2017.

databases that have samples of Latin writing from the medieval period, while the rest are built from modern writing in modern languages. Therefore, they cannot be considered in the validation of manuscript identification systems dedicated to the Latin manuscripts, and can only be used as reference research.

3 Dataset description and analysis

In the 21st century, the most significant libraries in the world holding the heritage of medieval manuscripts make their materials available in a digitized form in excellent quality. In order to highlight the order of magnitude of the size of the collections we are talking about, let us highlight that the digital repository of

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_8

the National Library in Poland alone has over three thousand digitized sources in Latin of the period belonging only to the Middle Ages, and some of them are divided into over a hundred manuscripts. In addition, materials are being digitized all the time and this number is growing.

The digital representation of such a codex inherits all description features from its analogue counterpart. By this the identifier of such a source, called the signature, assigned according to the rules prevailing in a specific library, is meant.

This is where the first problem arises when using sources to try to create a larger database of manuscripts for research purposes from more than one library. In order for the source to be clearly identifiable, we need to know the information from which library it comes, and its signature, which forms a two-level key. In the IT aspect, it is necessary to create a single-level key, and in the case of access to sources by a normal user, this poses a specific difficulty if the user does not have knowledge in the field of historian's workshop.

Another problem is the standard of shared content in digital form – should the manuscripts be made available as raster images or as a PDF containing the entire codex? The images are of huge resolutions and consequently they take up huge amounts of space. The transfer of these data can significantly burden the network traffic of such a digital repository. In addition, downloading a specific image from the code means that the user has to leave the page being viewed and use the image viewing application. For a user/historian who is used to turning pages and analyzing content in a specific context, it is a huge impediment and makes it very difficult to become familiar with the content of the codex under examination.

Hence, a standard interface for scalar data access using the IIIF [18,6] standard was developed. The discussed digital library repositories, using the created Application Programming Interface (API), made it possible to view these documents in a web browser and download individual documents if necessary, and otherwise it enabled viewing the code as a whole. Thanks to this, it did not require the highest resolution of viewed documents and ultimately saved server resources. However, a manifest file has been developed for serial (scientific) data retrieval, but additional software is required to use it, which requires expertise in the discipline of computer science. Here, by solving one problem, we create a new one that makes it difficult to collect large amounts of data.

The problem of two-level access to sources is not a new thing in this field, it found its solution in a very interesting way.

Most of the medieval content in Latin is the liturgical content of the Catholic Church, which is related to the issues of singing, hymns, etc. Therefore, instead of using the internal library signatures that hold the sources for internal purposes, the RISM (Répertoire International des Sources Musicales) [7] is used as signatures. Hence, when using the discussed IIIF standard and its API, it was decided to collect digitized manuscripts available in excellent quality from digital libraries located in Bamberg, Cambridge, Düsseldorf-Gerresheim, Heidelberg, Köln, Lisbon, London, München, Oxford, Paris, Schaffhausen, Vendôme

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_8

and Warsaw. The double ID problem was solved by using RISM directory IDs. Those identifiers that did not exist in the database have been added.

It should be remembered that only a small subset of the available codices from the digital repositories of the above-mentioned libraries was selected to create the database. Then, from the entire codes, handwriting samples (raster images) were carefully selected for the best preservation of the ink, the least amount of noise, absence of comments from other users of the codes (didascals), lack of miniatures and decorations. Another important aspect was the selection of samples that were digitized with lines of text kept horizontal, without any rotation.

The database currently has 117 fragments of codices with the number of samples between 11 and 39, with an average of 17, which gives 2040 writing samples. The resolution of the width of the manuscripts varies between two thousand pixels and almost nine thousand pixels (see Fig. 2). As it can be seen,

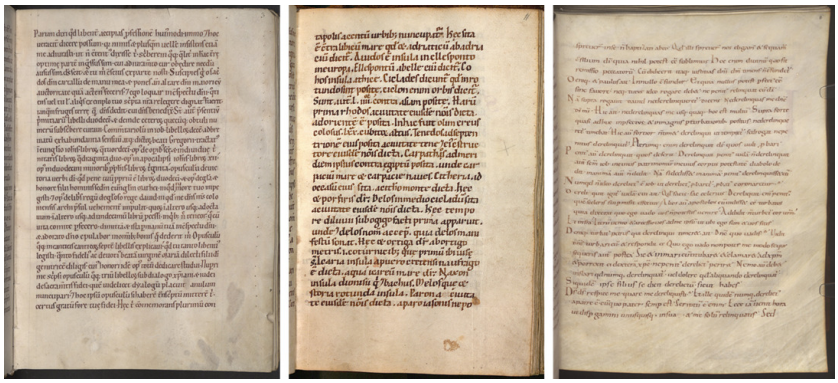

Fig. 2. Three sample manuscripts from the discussed database.

even within two centuries, the script differs significantly, let alone such a wide period of time of sources as in IAM-HistDB or ICDAR2017. Hence, in the future, a careful standardization of this database is possible along with its development. this is not easy, because the codes differ significantly in their sizes, and the distance between the digital camera and the code is also different, so the DPI in the file properties does not reflect the actual values. Therefore, it is necessary in the future to analyze, among others, the resolution, text density and font (duct) thickness.

4 Conclusion and Future Research

The database contains fragments from 117 medieval codices in Latin from the 11th-12th centuries. The manuscripts were collected using the IIIF standard, and the source identifier is the MISM signature. When selecting handwriting

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_8

samples, special care was taken to ensure that the documents were free of imperfections, comments, miniatures and decorations that may hinder the analysis of the database in question. The database is dedicated to verifying the effectiveness of author identification systems dedicated to Latin script or to universal use and is resistant to specific data, because systems dedicated to, for example, English script, were also validated with JEITA-HP databases with Japanese script or CERUG with Chinese script. The second use of the database may be not to identify the hand that wrote the manuscript, but the style in which the handwriting sample was written down, which could be used to show similarity between them. As a third application, the database can be used to teach neural networks to recognize content after obtaining a text layer. This content can be obtained from a specialist historian who can read the content, or the existing character recognition systems can be used to try to read the content.

In the future, it may be helpful to analyze the database from the IT aspect, i.e. to standardize the resolution of images in this database. Most databases discussed in Table 1 show that 2000 pixels is sufficient for correct handwriting recognition of A4 sheets at 300 DPI. Here the writing is very different from the modern one, and the sources are not of equal size. In addition, the writing, unlike the modern one, is stylistic, careful and the thickness of the writing (duct) may require greater resolution. Only a thorough analysis of the duct should approximate the resolution of the base standardization.

Historically, it may be helpful to analyze samples within a single codex to see if they were written by the same person. An analysis of the manuscript's contributions can help a lot with this. Another way to solve the same problem may be to validate multiple writer identification systems and analyze them against each other. On top of that, adding content transcription can help to use this database to train OCR handwritten models.

In the future, it is also possible to expand the database with new codices available in digital library repositories.

References

1. Bulacu, M., Schomaker, L., Vuurpijl, L.: Writer identification using edge-based directional features. In: Proc. 7th Int. Conf. on Document Analysis and Recognition ICDAR. pp. 937–941. IEEE (2003). https://doi.org/10.1109/ICDAR.2003.1227797
2. Fiel, S., Kleber, F., Diem, M., Christlein, V., Louloudis, G., Nikos, S., Gatos, B.: ICDAR2017 Competition on Historical Document Writer Identification (Historical-WI). In: Proc. 14th Int. Conf. on Document Analysis and Recognition ICDAR. vol. 01, pp. 1377–1382. IEEE (2017). https://doi.org/10.1109/ICDAR.2017.225
3. Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., Stolz, M.: Ground truth creation for handwriting recognition in historical documents. In: Proc. of the 9th IAPR Int. Workshop on Document Analysis Systems DAS. pp. 3–10 (2010). https://doi.org/10.1145/1815330.1815331
4. Freitas, C., Oliveira, L.S., Sabourin, R., Bortolozzi, F.: Brazilian forensic letter database. In: 11th International Workshop on Frontiers on Handwriting Recognition. Montreal, Canada (2008)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_8

5. Grosicki, E., Carre, M., Brodin, J.M., Geoffrois, E.: Rimes evaluation campaign for handwritten mail processing. In: Proc. ICFHR 2008: 11th Int. Conf. on Frontiers in Handwriting Recognition (2008)
6. International Image Interoperability Framework: Gain richer access to the world's image and audio/visual files, https://iiif.io, [Accessed: February, 2023]
7. Keil, K.: Repertoire international des sources musicales (RISM). Fontes Artis Musicae **59**(4), 343–346 (2012)
8. Kleber, F., Fiel, S., Diem, M., Sablatnig, R.: CVL-database: An off-line database for writer retrieval, writer identification and word spotting. In: Proc. 12th Int. Conf. on Document Analysis and Recognition ICDAR. pp. 560–564. IEEE (2013). https://doi.org/10.1109/ICDAR.2013.117
9. Liu, C.L., Sako, H., Fujisawa, H.: Handwritten Chinese character recognition: Alternatives to nonlinear normalization. In: Proc. 7th Int. Conf. on Document Analysis and Recognition ICDAR. vol. 3, pp. 524–528. IEEE (2003). https://doi.org/10.1109/ICDAR.2003.1227720
10. Louloudis, G., Gatos, B., Stamatopoulos, N., Papandreou, A.: ICDAR 2013 Competition on Writer Identification. In: Proc. 12th Int. Conf. on Document Analysis and Recognition ICDAR. pp. 1397–1401 (2013). https://doi.org/10.1109/ICDAR.2013.282
11. Louloudis, G., Stamatopoulos, N., Gatos, B.: Writer identification. In: Document Analysis and Text Recognition: Benchmarking State-of-the Art Systems, pp. 121–154 (2018)
12. Maadeed, S.A., Ayouby, W., Hassaïne, A., Aljaam, J.M.: QUWI: An Arabic and English handwriting dataset for offline writer identification. In: 2012 Int. Conf. on Frontiers in Handwriting Recognition. pp. 746–751 (2012). https://doi.org/10.1109/ICFHR.2012.256
13. Mahmoud, S.A., Ahmad, I., Al-Khatib, W.G., Alshayeb, M., Tanvir Parvez, M., Märgner, V., Fink, G.A.: KHATT: An open Arabic offline handwritten text database. Pattern Recognition **47**(3), 1096–1112 (2014). https://doi.org/10.1016/j.patcog.2013.08.009
14. Marti, U.V., Bunke, H.: The IAM-database: an English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition **5**(1), 39–46 (2002). https://doi.org/10.1007/s100320200071
15. Pach, J.L.: New database of fragments of medieval codices of the 11th-12th centuries (2023). https://doi.org/10.5281/zenodo.7569006
16. Pechwitz, M., Maddouri, S.S., Märgner, V., Ellouze, N., et al.: IFN/ENIT-database of handwritten Arabic words. In: Proceedings of the 7th Colloque International Francophone sur l'Ecrit et le Document (CIFED). vol. 2, pp. 127–136 (2002), http://ifnenit.com/download/CIFED_02_ifn-enit-database.pdf
17. Semma, A., Hannad, Y., Siddiqi, I., Lazrak, S., Elkettani, Y.: Feature learning and encoding for multi-script writer identification. International Journal on Document Analysis and Recognition **25**, 79–93 (06 2022). https://doi.org/10.1007/s10032-022-00394-8
18. Snydman, S., Sanderson, R., Cramer, T.: The international image interoperability framework (IIIF): A community & technology approach for web-based images. In: Proc. IS&T Archiving 2015. pp. 16–21 (2015), https://library.imaging.org/archiving/articles/12/1/art00005, article ID: art00005

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36027-5_8