Automatic abbreviation detection in medieval medical documents

[1] Malagón, C., [2] Rizky, R., [3] Kim Y., [4] Marzal, F., [5] Izquierdo, L.
[1,2,4,5] *Nebrija University, Spain,* [3] *Purdue University*
[1] cmalagon@nebrija.es, [3] kim153@purdue.edu

Abstract

Automatic palaeographic transcription of medieval manuscripts is a challenging task. Transcribing and deciphering these manuscripts is considered complicated and problematic, especially due to particular characteristics of medieval handwriting and the composition and conservation of these documents. In this paper, a new problem in handwriting recognition is addressed: the detection of abbreviations in medieval medical manuscripts. We propose a histogram-based method to detect these specific kinds of characters. The results achieved with real medieval medical manuscripts are very promising, and it opens new research lines in handwriting recognition.

1. Introduction

Automatic palaeographic transcription of medieval manuscripts is a challenging task. Transcribing and deciphering these kinds of manuscripts is considered complicated and problematic, especially due to particular characteristics of medieval handwriting and the composition and conservation of these documents. Another difficulty is the restricted access to these, although many of them are beginning to be digitalized. For these reasons, automatic palaeographic transcription of old manuscripts has not yet been explored much.

Medieval manuscripts and especially, those which are previous to the invention of the printing press in the XV century, show specific characteristics, and their knowledge is essential for palaeographic transcription [1]. This knowledge influences the design and development of an automatic or even semiautomatic character recognition system. So, this challenge demands the collaboration between researchers and specialists in these kinds of texts, leading to a cutting edge interdisciplinary research [2].

Among various medieval manuscripts, we will focus on medical documents of the thirteenth to fifteenth centuries. These manuscripts, exclusively pertinent to Medicine, History of Medicine and Medieval Medicine, have special importance in medical and scientific fields because researchers in these fields seek practical knowledge in these texts to cure existing diseases.

Pattern recognition in medieval documents introduces additional difficulties. The most important are the following: wide variety of types of patterns (signs, characters, capitalization rules, alphabets, etc.) and the use of Latin as the language of the majority of the documents. But the main difficulty is the abundance of abbreviations, with different kinds of shapes and positions [3].

These abbreviations appear frequently in most of medieval documents, but its incidence is higher in medical texts. About six hundred abbreviations can be found in each page of our working document. So it is clear the importance of automatically detect and transcribe these abbreviations.

The working document is entitled *Liber morborum tam ulterium quam particularium*, by Gilbertus Anglicus, and it was written during the XIII century. This document was provided for this work by The Historical Library "Marqués de Valdecilla" in Madrid, Spain.

This paper will focus only on preprocessing documents and detecting abbreviations in medieval medical documents. In order to do this, the documents need to be preprocessed so that each captured page can be first segmented into columns, an then into lines and words.

Moreover, the abbreviations that are detected need to be related to a word. These abbreviations will mostly appear above some words in different geometric forms, and depending on its position and the letter over which it appears, they will have different meanings.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.184

Figure 1 is an example of a captured page that has been used in this paper.

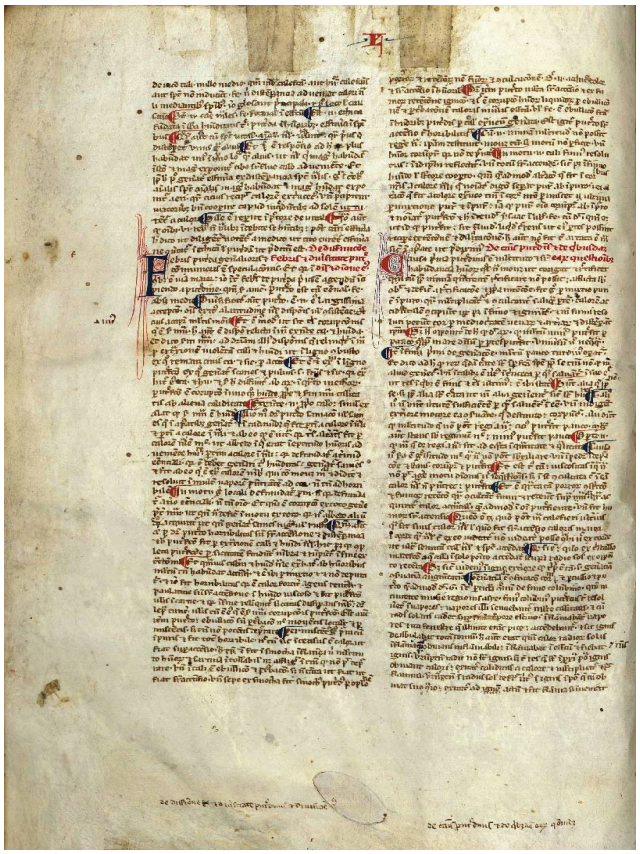

Figure 1. Example of a captured page of the working manuscript.

The remainder of this paper is as follows: first, we will describe the problem i.e. the abbreviations and their specific geometric characteristics in order to understand the problem, its difficulties and peculiarities. Secondly we will mention some of the related work that can be found in this field. In section 4, we will describe the preprocessing phase. In the next section we will describe the detection phase, and we will show the results obtained in our experiments. Finally we will present the conclusions of our work.

2. Description of the problem

One of the toughest challenges that a researcher faces in the field of palaeographic transcription is the abundance of abbreviations in medieval documents. Apart from these abbreviations that can be classified as generic, a great number of unusual contractions and peculiar acronyms in medieval medical manuscripts can be found [4].

The objective of this work is to automatically detect and identify these abbreviations within the text. This knowledge can be then applied and incorporated into the general palaeographic transcription by building compiled dictionaries of generic and peculiar abbreviations, which will effectively solve the particular difficulties of transcribing abbreviated words in medieval medical manuscripts.

Figure 2. Example of an abbreviation

Figure 2 shows an example of this kind of abbreviation. The stroke over letter “o” is a generic abbreviation, and it's equivalent to the letter “n”, so this word must be read as De diffinitio(n)e.

One of the major problems is the abundance of overlapping characters. One example of this can be seen in figure 3:

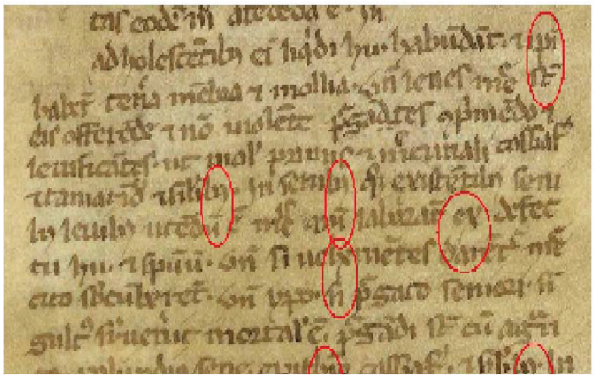

Figure 3. Example of overlapping characters

3. Related work

The specific problem of automatically detecting and identifying abbreviations in medieval documents has not received enough attention from researchers in the Pattern Recognition field, while it is a key problem in Paleography.

Some related work can be found in [5], where the author attempts to address general automatic recognition of medieval texts. The author applied classic pattern recognition techniques to automatically transcribe medieval documents with modest results. It should be noted that the documents used in the present work are very different to those because they do not

have any kind of abbreviations, and the author did not point out this problem.

Researchers from Pisa University developed SPI (System Palaeographic Inspection), a system which is capable of extracting, semi-automatically, single characters from the documents and to generate prototype models of characters belonging to the same document[6]. The project has not been continued, and it has not been normally used by the palaeographers [7].

There are other related work in automatic transcription of palaeographic texts [8]. This work does not focus on automatic transcription but on author identification. The applied method is based on the comparison with a database of digitalized and transcribed documents, according to its style and calligraphy.

4. Preprocessing

In an initial phase, the documents are preprocessed in three stages: firstly, a typical binarization process is carried out [9]. As it has been mentioned above, the major problems with these documents are their conservation, so the preprocessing tasks are very important in order to successfully detect the characters and abbreviations. These characteristics make the preprocessing process a challenging problem in itself.

A second preprocessing task is the column and line segmentation [10]. This task has been carried out by using histograms.

In figure 4, an example of line segmentation is shown.

Figure 4. Line segmentation

The third task in the preprocessing phase is the skew and slant correction. In order to do this, a second histogram-based method is used [9].

Figure 5. Words segmentation

It is well known that word segmentation is a complex task in automatic handwriting recognition [10]. In this case this task is necessary because both the abbreviation and the abbreviated word, have to be taken into consideration in order to get the complete meaning of the word.

5. Detection

Once the document is preprocessed, the abbreviation detection is carried out by applying a histogram based method along with the use of a lexicon.

The process is as follows: candidate abbreviations are proposed among the detected strokes. A histogram method based on pixel intensities is used, and so candidate abbreviations are pointed out.

The abbreviations are isolated from the word, and their skeleton and characteristic points are extracted [11].

This morphology is compared with those stored in the lexicon of abbreviations. If 85% of the points of their morphology match, the abbreviation is detected and marked.

The abbreviation is then returned as an image of the single abbreviation, the abbreviation over the letter, an abbreviation over the word, and the histogram of the abbreviation, as can be seen in figure 6.

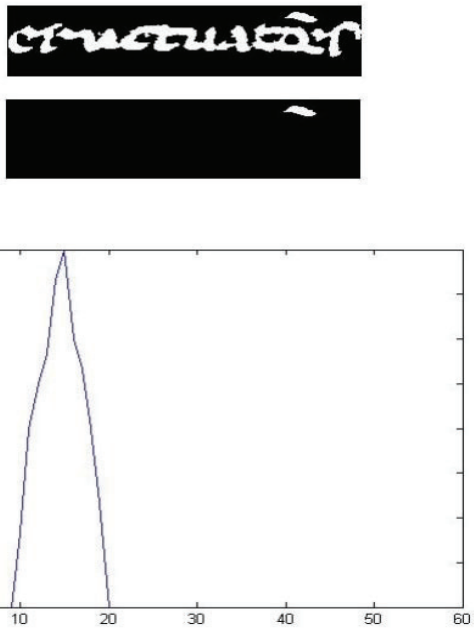

Figure 6. Histogram of an abbreviation

The same result is shown in figure 7, where two abbreviations have been isolated in two different words within a line.

Figure 7. Image of two isolated abbreviations corresponding to different words

6. Experimental results

A total of 16 pages have been used. As it has been said before, one of the main issues in this work was line segmentation, due to a particular difficulty in these kinds of documents.

The results for line segmentation are shown in Table 1.

Table 1: Results for Line Segmentation.

Page number	Number of lines column 1	Results column 1	Number of lines column 2	Results column 2
1	66	62	66	64
2	66	62	66	63
3	66	56	66	65
4	66	59	66	62
5	66	60	66	65
6	66	65	66	65
7	67	64	67	66
8	67	63	67	64
9	66	61	66	64
10	68	66	68	66
11	66	64	66	65
12	68	54	68	58
13	60	55	60	60
14	64	54	64	63
15	65	64	65	64
16	66	63	66	64
Total	**1053**	**972**	**1053**	**1018**

Table 1 shows that from a total of 2106 lines, the algorithm managed to successfully segment 1990 lines. In other words, the line segmentation task has a success rate of 94.5%.

Results for abbreviations detection is shown in Table 2. The table shows the real number of abbreviations for each captured page and the number of detected abbreviations. It has to be noted that some of them will be false positives. That's why the number of detected abbreviations could be greater than the number of abbreviations that really appears in the text. That would be then false positives.

Table 2: Results for Abbreviation Detection.

Page number	Number of abbreviations	Number of detected abbreviations
1	433	507
2	535	504
3	584	571
4	653	638
5	634	599
6	744	701
7	643	602
8	608	500
9	600	598
10	743	950
11	655	778
12	624	570
13	632	658
14	694	704
15	610	644
16	547	677
Total	**9939**	**10201**

Among these 10201 abbreviations detected, only 7852 are real abbreviations, so the accuracy is about a 76.97%.

In order to explain the errors, the results in one page are detailed in table 3. As can be seen in this table, the number of well detected abbreviations is 248 and 312 for column 1 and 2 respectively. So, in this case, the true positive rate are 0.77 and 0.74 respectively.

Regarding the errors, the number of non-detected abbreviations is 74 for column 1 and 119 for column 2, so the false negative rate (i.e. the rate of non detected abbreviations) is about 25%.

These errors (both false negatives and false positives) are mainly due to overlapping and the presence of noise.

Table 3: Details of results in one page.

Pagee number 6	Number of abbreviations	Abbreviations well detected	TPR	FNR
Column 1	322	248	0.77	0.22
Column 2	422	312	0.74	0.28
Total	744	560	0.75	0.25

At this point it is important to mention some of the challenges that have been encountered. Abbreviations that are in contact with words are almost impossible to detect. Figure 8 illustrates two abbreviations: the one on the left part of the figure is in contact with a word, while the other one, on the right, is not.

Figure 8. Overlapping of abbreviations and other normal characters

The problem of the overlapping strokes can explain most of the false positives that have been found, and new methods have to be explored in order to solve this.

Among other problems, the overlapping of lines and the presence of spots greatly affect the performance of the algorithm. Figure 9 illustrates overlapping of lines and a spot, marked in red, which is detected as an abbreviation (i.e. a false positive).

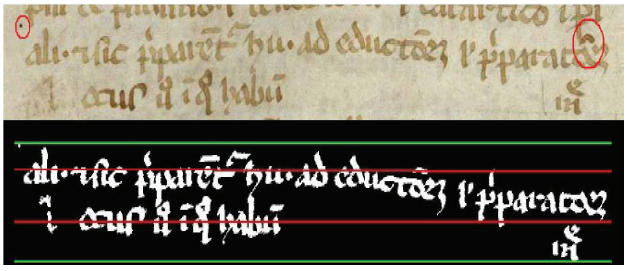

Figure 9. Overlapped lines and the influence in detecting a spot as an abbreviation.

Regarding the computing performance, the system used is a Pentium IV with 1.6 GHz and 2 GB of RAM. Approximately 120 minutes were needed to process a captured page. It has to be noted that about 80% of this time is consumed in the line segmentation phase.

7. Conclusions

In this paper, a system for preprocessing medieval manuscripts and detecting abbreviations is proposed. The methods put forward are based on typical computer vision functions such as image binarization and intensity histograms for lines and words segmentation. Another histogram-based method, along with a lexicon matching process has been used for abbreviations detection.

Using these functions we are able to create a system capable of extracting abbreviations from medieval medical manuscripts.

The proposed system has been successfully implemented in experimental settings that are representative of many practical simulations. The experimental results demonstrate that the proposed technique has a high success rate for line segmentation, with an accuracy of about 95%.

Regarding the main issue of this work, the abbreviations detection in medieval medical documents, the accuracy achieved in this task is about a 79%.

Further work can be explored to extend this proposed system not only to detect but automatically transcribe abbreviated words and their meanings in medieval medical manuscripts.

References

[1] Bischoff, B. Latin Palaeography: Antiquity and the Middle Ages, Cambridge University Press, 1989.

[2] Rehbein, M., Sahle, P., Schaßan, T. (eds.): Codicology and Palaeography in the Digital Age. BoD, Norderstedt, 2009

[3] John J., Latin Paleography, in J. Powell, Medieval Studies. Syracuse University Press, 1992.

[4] Cappelli, A. The Elements of Abbreviation in Medieval Latin Paleography. University of Kansas Library, 1982.

[5] Stokes, P. A. Palaeography and Image Processing: Some Solutions and Problems, 2008.

[6] Aiolli, F., Ciula, A. SPI: A System for Paleographic Inspections Proceedings of the 2009 conference on Computational Intelligence and Bioengineering: Essays in Memory of Antonina Starita. The Netherlands, 1999.

[7] Ciula, A. Digital palaeography: using the digital representation of medieval script to support

palaeographic analysis. Digital Medievalist, 2005.
[8] Bulacu, M, Schomaker, L. Automatic handwriting identification on medieval documents. Proceedings of the 14th International Conference on Image Analysis and Processing, 2007.
[9] Vinciarelli, A. A survey on off-line cursive word recognition. Pattern Recognition, 35(7):1433-1446, 2002
[10] Casey, R. G. , Lecolinet, E. A Survey of Methods and Strategies in Character Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 18 [7], 690-706, 1996.
[11] Liu, C., Kim, J., Kim, J. Model-based stroke extraction and matching for handwritten Chinese character recognition, Pattern Recognition 34 2339–2352, 2001 .