Automatic Handwritten Character Segmentation for Paleographical Character Shape Analysis

Théodore Bluche
A2iA SAS
39 rue de la Bienfaisance
75008 Paris, France
tb@a2ia.com

Dominique Stutzmann
IRHT - CNRS
Paris, France
dominique.stutzmann@irht.cnrs.fr

Christopher Kermorvant
Teklia SAS
Paris - France
kermorvant@teklia.com

***Abstract*—Written texts are both physical (signs, shapes and graphical systems) and abstract objects (ideas), whose meanings and social connotations evolve through time. To study this dual nature of texts, palaeographers need to analyse large scale corpora at the finest granularity, such as character shape. This goal can only be reached through an automatic segmentation process. In this paper, we present a method, based on Handwritten Text Recognition, to automatically align images of digitized manuscripts with texts from scholarly editions, at the levels of page, column, line, word, and character. It has been successfully applied to two datasets of medieval manuscripts, which are now almost fully segmented at character level. The quality of the word and character segmentations are evaluated and further palaeographical analysis are presented.**

I. Introduction

Written texts are both physical and abstract objects: ideas, signs and shapes, whose meanings, graphical systems and social connotations evolve through time. Beyond authorship and writer identification or palaeographical dating of textual witnesses, the materiality of text and the connexion between the ideas and their written instantiations are a matter of cultural history, historic semiology, and history of communication and representations. In the context of large, growing digital libraries of texts and digitized medieval manuscripts, the question of the cultural significance of script and the "dual nature" of texts may at last be addressed. Several research projects, interfaces and software allow for a closer text-image association during the editing process (TILE, T-PEN, MOM-CA) and for data visualisation (Mirador, DocExplore), with interoperable annotations schemas (SharedCanvas). But in most cases, the finest granularity is at line-level with an alignment being done by hand on small amounts of text (a few pages).

Automatic handwritten text recognition (HTR) systems based on Hidden Markov Models (HMM) can be used to segment manuscript images at line, word and character level. When the transcript of a line image is known, HMMs can compute the so-called *forced alignments*, *i.e.* force the recognition output to correspond to the actual sequence of characters. It simplifies the recognition process, and focuses only on the retrieval of characters positions. When we do not know the lines or words positions or transcript, but only the document transcript, lines can be automatically detected and mapped to the transcripts. Such methods also relax the annotation effort to build the corpora since only the page transcript is needed.

In this paper, we propose to use a handwriting recognition system to segment historical manuscript images at line, word and character level. We show that this segmentation can be done even if only a page transcript is available, without the need of manually locating the text lines on the images. It has been successfully applied to two datasets, "Graal" and "Fontenay". Partial results of alignment at word and character level are already online[1].

Our contention is that such an alignment method is the only way to gain access to new questions on a large-scale basis and massively transfer the results of traditional Humanities and textual scholarship into Digital Humanities. The automation causes not only a change of scale (larger corpora), but also a change in granularity (page by page or line by line alignment to word and character alignment). It avoids the tedious task of drawing boxes by hand around characters and allows a systematic analysis of the data. It outmatches the preceding attempts made to more closely associate both aspects of text and also opens perspectives on automated transcription.

This paper is divided as follows. In Section II, we give a brief review of text-to-image alignment methods. In Section IV, we present how HMMs can produce character alignments by modelling the sequences of characters for transcript at line or page level. Then, we present the two corpora in Section III, the design and training of the recognition system used to produce the alignment in Section V, and evaluate the quality of the word and character segmentations in Section VI.

II. Related work

The problem of mapping transcript to images has already been studied in the past decade. In 2004, Kornfield et al. [1] argued that handwriting recognition systems were not good enough to help the mapping of text to historical document images. They matched words of the transcript with automatically segmented word images using Dynamic Time Warping (DTW) with features for both image and text. We claim that handwritten recognition systems are now good enough to process historical manuscripts.

[1]http://oriflamms.a2ialab.com/Charsegm/

Feng et al. [2] proposed a method for aligning OCR output with a book transcription using Hidden Markov Models (HMMs). They focus on a text-to-text alignment rather than a text-to-image mapping, in order to evaluate the OCR accuracy for book transcription.

Rothfeder et al. [3] automatically segmented the G. Washington Database into words, using a linear HMM and the Viterbi algorithm to align images with the transcript. The system was trained on a manually corrected segmentation and annotation of this database. We show that the character alignment can be done without manual line or word segmentation.

In [4], experiments were conducted on the St. Gall database, where the transcription was known but inaccurate (missing text lines, line breaks not indicated, missing abbreviated or capitalized words). They concatenate all text lines and perform a Viterbi alignment with HMMs corresponding to the transcript, including spelling variants. The HMMs were trained on only one page of the database, yet the results were good, and the system could deal with inaccurate transcriptions. Similarly, our work is based on a forced alignment but also provides the text alignment at line level.

In [5], the authors used the Finite State Transducer (FST) framework for the ground-truth alignment in difficult historical documents, where OCR systems do not perform well. They modelled the transcript as an FST with variants, allowing the system to cope with OCR errors (substitutions, deletions, insertions), ligatures, and hyphenation. They extracted OCR lattices, and aligned them using an FST representing the line content plus its surroundings. Our method is also based on a FST representation of the transcript but we allow the algorithm to use multiple line segmentation hypothesis, and jointly search for the best segmentation and transcript mapping by enforcing some ordering constraints, both geometrical and textual.

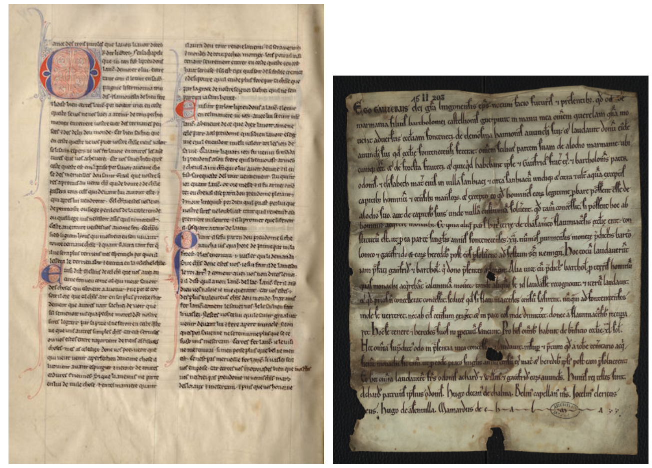
Finally, other methods rely on a combination of manual and automatic processing, for example to manually provide the bounding box of the character and automatically extract the shape [6].

III. The Graal and Fontenay corpora

We have applied the automatic character segmentation method to two different corpora of medieval manuscripts. Examples of pages from the two corpora are shown on Figure 1.

The Graal corpus is composed of 130 pages from the manuscript *Lyon, City Library, Palais des Arts 77*, written in the 13th century. It contains the La Queste del saint Graal, one of the oldest prose chivalric romance in Old French. The handwriting is a fairly regular book hand. The scholarly edition gives several layers of text. The text we used is the *diplomatic edition*, retaining the spellings, punctuation, capitalization, line divisions of the manuscript. The main difficulties in the image were the presence of decorated initials in color ranging from two to six lines in height and damaged areas.

The Fontenay corpus is composed of 104 single charters from the 12th and 13th centuries, that is 104 documents of one page each, preserved in the Archives départementales de Côte d'Or in Dijon (series 15H). In this second corpus, there are no decorated initials in color, but there are many difficulties in the image analysis: damaged areas, great variations in the height of text lines (from 5 mm to 20 mm) and text size on the line (from less than a sixth of the line height to the full height). The text difficulties are important even beyond the dimensional aspects. On the one hand, these documents were written by many different scribes, so that there are not only different allographs for each letter, but there are great variations for each allograph. This corpus has even more abbreviations and superscript letters (9193 abbreviations, that is 0,41 abbreviation per word) than the Graal corpus. On the other hand, the transcript does give indications on some allographs, but not in a systematic manner.



(a) Graal (b) Fontenay

Fig. 1: Examples from the Graal (Lyons, City Library, PA 77, fol. 187v) and Fontenay Database (Dijon, Archives départementales de Côte d'Or, 15 H 203 (Reproduced with permission).

IV. Automatic character alignment with HMMs

Most of the current HTR models are based on Hidden Markov Models (HMM) associated with a sliding window approach to segment the input image. These models produce a precise character segmentation of the "text image" as a by-product. When the transcript of a line image is known, the HMM focuses on the retrieval of characters positions and forces the output to correspond to the actual sequence of characters (so-called "forced alignment"). If only the whole document transcript is available, and not the positions or transcript of the words or lines, line detection methods can be used beforehand to map the transcripts to the detected lines.

In this paper, the character segmentation results from an incremental process previously proposed in [7] to improve the training of HTR models using images of paragraphs and their textual transcripts. The process was adapted to produce the text/image alignment at different levels (line, word, character) and is composed of the following steps:

Fig. 2: Forced alignment example : pre-processed image (top) and the character segmentation (bottom).

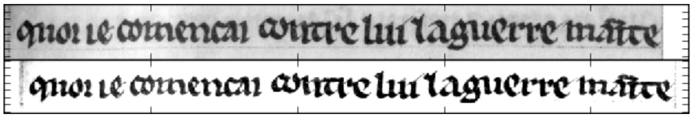


Fig. 3: Preprocessing example.

- apply a text line segmentation algorithm, adapted from [8] to the full page
- keep the pages for which the number of detected lines is equal to the number of lines in the transcript. At this stage, the line segmentation of remaining pages can be manually corrected.
- assign the line transcripts to the line images and use them to train a first HMM based on Gaussian Mixture Models (GMM-HMM)
- align the line transcription with the line images with the trained GMM-HMM as proposed in [7]
- based on this new alignment, train a new GMM-HMM recognizer.

This process is fully automatic and repeated (usually between 2 and 5 times) until all text lines are correctly aligned. The criterion to stop this process is still empirical and is based on a qualitative examination of the alignment current alignment. Afterwards, we train a final text recognizer based on deep neural networks HMMs. This model is trained with a discriminative criterion and yields better transcription results and segmentation accuracy than the standard GMM-HMM.

The character segmentation is obtained with a forced alignment of the text line transcript and the line image using the final hybrid HMM mode. An example of alignment result is shown on Figure 2.

V. Design and training of the HTR system

The handwritten text recognition system is based on a fairly standard HMM recognizer. The characteristics and the training of this system are described in this section. At this stage of the process, the text line image is supposed to be localized (manually or automatically, in our case,the text lines were automatically localized).

A. *Image processing and feature Extraction*

For each line of text, we first converted the color images into gray levels, applied a deskew and deslant correction, and mapped the darkest 5% of pixels to black and the lightest 70% to white, with a linear interpolation in between, to enhance the contrast. We added 20 columns of white pixels to the beginning and end of each line to account for empty context. To normalize the height of text lines, we detected three regions in the image (ascenders, descenders and core region), and scaled these regions to three fixed heights. An example of preprocessing result is shown on Figure 3. Finally, we computed handcrafted features described in [9] using a sliding window of width 3px with no overlap (shift of 3px).

B. *Word and character modelling*

Each character occurring in the transcript was modelled with an HMM. The transcript contained lower-cases and upper-cases letters, punctuation, abbreviation such as Tironian notes, superscript or subscript letters or symbols.

For the Graal corpus, following the analysis of the character morphology described in [10], we have modelled several writing variants of characters, presented on Figure 4:

- Conjunction: the last stroke of the first letter is superposed with the first stroke of the second one.
- Elision: the initial stroke of a letter is left out when the last stroke of the preceding letter ends at the upper limit of the minims. Both letters are in contact and their shapes may be modified.
- Ligature: two or more letters are joined as a single glyph. Their shapes may be largely modified.
- Allograph: the same letter can have different forms, depending on the position or the context.

The observation and analysis of these phenomena are of core interest for palaeographers and historians of written communication. They allow for the identification of scribes or dating of non dated documents as well as a broader understanding of the evolutions of the Latin script in the Middle Age[11].

Since the system was trained on a diplomatic transcription, the character forms, the conjunctions, elisions and ligatures were not given. Therefore, during the training, each word was modelled with all the possible variants, as shown on Figure 5 and the system decided which form was best for the alignment.

For both corpora, regular characters were modelled with 3-state HMMs, space with 2 states and the punctuations with 1 state. For the Graal corpus, the compound characters (elision, conjunction, ligature) were modelled with 5-state HMMs.

C. *Training*

We have trained the GMM-HMM system with the Maximum Likelihood criterion, following an Expectation-Maximization (EM) procedure with early-stopping (the training stopped when the error rate on the validation set started to increase). On the Graal corpus, the final GMM-HMM model was composed of 69 character models, 14,648 Gaussians in 211 mixtures. The Word Error Rate (WER) and Character Error Rate (CER) on the validation set were 23.0% and 6.9%. On the Fontenay corpus, the final GMM-HMM model was composed of 185 character models, 29,488 Gaussians in 929 mixtures. The Word Error Rate and Character Error Rate on the validation set were 21.9% and 8.4%. Note that recognition error rates are not good measures of the word or character segmentation quality since the recognizer may not need to

(a) (b) (c) (d) (e) (f)

Fig. 4: Examples writing variants modeled by the HMM. a. Conjonction (pa), b. Elision (eu), c. Ligature (st), d. 2 Allograph of s, e. 2 Allograph of v, f. Symbols

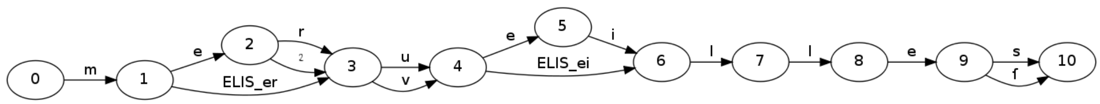


Fig. 5: Example of lexical modeling for the word *"merveilles"*

correctly segment the characters to recognize them, as for example with LSTM recurrent neural networks. Therefore, new metrics were designed specifically.

VI. Evaluation of the segmentation

To evaluate the quality of the segmentation, we performed the forced alignment of the 130 pages of Graal corpus and the 104 pages of the Fontenay corpus at line, word and character levels, resulting in the following numbers of new annotations to the corpora :

Level	Graal	Fontenay
Segmented lines	10,362	1,363
Segmented words	114,273	22,730
Segmented characters	504,5230	128,946

A. Word segmentation evaluation

For both corpora, the word positions have been manually corrected and validated for all words. We used this information to assess the quality of automatic alignment.

The left and right reference boundaries (vertical positions expressed in pixels) of a words **ref** are denoted $\mathbf{ref}_l$ and $\mathbf{ref}_r$ respectively and the hypothesis boundaries $\mathbf{hyp}_l$ and $\mathbf{hyp}_r$. For each word, the following boundary errors were computed : absolute error $= |\mathbf{hyp}_l - \mathbf{ref}_l| + |\mathbf{hyp}_r - \mathbf{ref}_r|$, left relative error $= \mathbf{ref}_l - \mathbf{hyp}_l$, right relative error $= \mathbf{ref}_r - \mathbf{hyp}_r$.

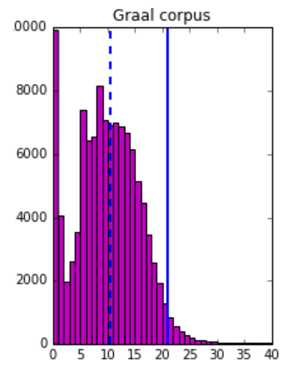
The system accuracy is defined as the proportion of word alignments considered as correct given a boundary tolerance (in pixels). The histograms of boundary error values are shown on Figure 6 and 7 . Regarding the absolute boundary error, this figure shows that in Graal, 63% of boundaries are correct with a 11 px tolerance (half a character width) and 99% are correct with a 23% tolerance (1 character width). In Fontenay corpus, 72% of boundaries are correct with a 22 px tolerance (half a character width) and 94% are correct with a 45px tolerance (1 character width).

Considering the right and left boundaries errors, give details on the type of segmentation errors. On both corpora, the left errors were negative and the right errors positive, which means that the automatic segmentation was to *tight* on the word. However, since the annotated word boundaries were given only at the middle of the blank between words, the error value is overestimated.

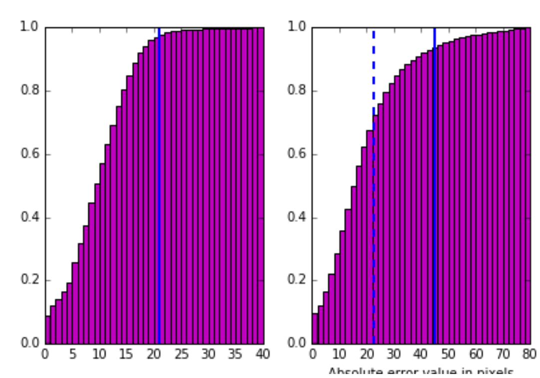
This word segmentation is precise enough to allow a large scale palaeographical analysis of these two corpora. The first consequence of this high accuracy is that the ground-truth, which was supposed to be 100% accurate, could to be corrected. A tabular view of words allowed to spot occurrences for which the human transcriber forgot to indicate that the word was abbreviated. At a word level, the palaeographical analysis encompasses the study of positional allographs (variant letter forms at the beginning or end of a word to help reading like in Arabic scripts), of semantic treatments of words (abbreviations and variant letter-forms in named entities) and of variability of overall shape. This word segmentation of large corpora provided, for the first time in palaeography, an objective measure of the spacing between words. A palaeographical study of this aspect is ongoing[12].

B. Character segmentation evaluation

The automatic alignment at character level of the two corpora produced more than 630,000 segmented characters. We have randomly selected 2% of these characters using a uniform distribution on all the characters and asked a palaeographer to validate the segmentation. Segmented character were rejected if a structural stroke was missing or if a structural stroke from a neighbour character was added (a structural stroke, that is a part of the letter which is a necessary component, which cannot be removed without creating ambiguities, e.g. three vertical minims in letter "m", while accessory strokes like feet or elongated endings are not necessary to the recognition of the letter). On average on all the characters, the segmentation error was 10.4% for the Graal corpus and 13.3% for Fontenay corpus. The detailed error rates per character are presented on Figure8. This criterion may prove difficult to use within a vertical alignment for letters generally attached or overhanging above the following one, such as "c" and "long s". For letter "e", the consequence was that many occurrences were rejected because the letter is cropped out on the left with the loss of a small part on the upper left section, causing the "eye" of the "e" to be open. The model for structural connexions between letters that give birth to new graphems such as ligatures proved

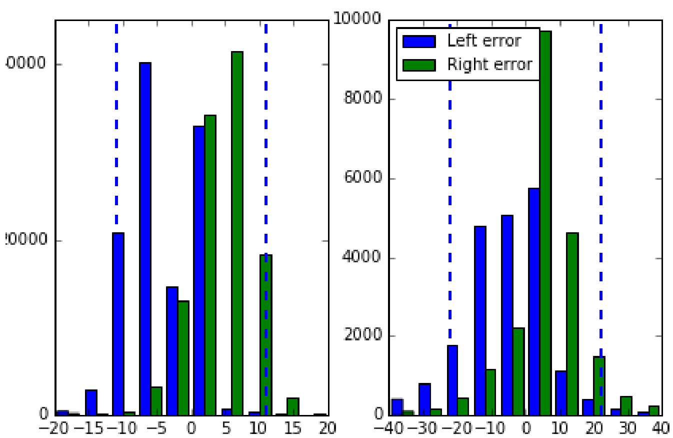


(a) Histogram of absolute word boundary errors in pixels



(b) Cumulative histogram of absolute word boundary errors in pixels

Fig. 6: For both corpora : (a) histograms and (b) cumulated histograms of absolute word segmentation errors. The average character width is shown with a continuous line and the average half character width with a dashed line.



(a) Histogram of relative left and right word boundary errors in pixels

Fig. 7: For both corpora : histogram of relative left and right word segmentation errors. Theaverage half character width with a dashed line.

to be efficient (e.g. ligature ”st”, ligature ”ss”). Connexions that do not have structural consequences had a lower rate (e.g. elisions). Note that the evaluation does not give the recall for the modelled graphical phenomena. The punctuation signs were discarded from the evaluation, because both corpus were transcribed with different conventions. In the Graal corpus, the *punctus elevatus* (a reversed semi-colon) was not transcribed as such, but as the sequence of dot and apostrophe (”.” and ” ’ ”), causing the system to search for two different signs creating an ambiguity with dots instead of a specific one. In the Fontenay corpus, the punctuation was not transcribed in a diplomatic manner, but was regularized.

The low results for letter ”q” in the Graal corpus is explained by the position of this letter almost exclusively at word beginnings in Old French (6866 among 7282 occurrences). The system modelled the sequence ”space followed by letter q” not as blank space, but as blank with a curved stroke. Letters with ascenders are generally better aligned, since their silhouette is easier to model.

The accuracy of the alignment at letter level is very high. Rejected occurrences do not mean that the character has not been correctly spotted, but that the boundaries have been misplaced. Alignment at a letter level must face the challenge of addressing not vertically separated letters with vertical boundaries in the sliding window technique. As such, the alignment makes it possible to analyse from a human perspective the different letter shapes in a tabular view and draw conclusions which are not reachable in a linear reading and observation.

VII. Perspectives

This work was conducted within the interdisciplinary research project *Oriflamms* in which we faced several technical and Humanities driven challenges. We wanted to align text and image to be able to access our medieval Cultural Heritage both as meaning and as a world of shapes.

This became possible through an efficient alignment technique, based on machine learning which we have applied on more than 136’000 words and 630’000 characters, and we will continue with further corpora.

Above all, we encountered some of the trending issues of the Digital Humanities in a truly interdisciplinary project, rising new ideas, questions, and methods, and we are stepping out of our comfort zone. Part of the human skill of reading ancient script may be sooner or later replaced by HTR, but we have new research questions to imagine and new questions to answer regarding ergonomics and the human part in the digital humanities, about granularity and metrics, about the human in the loop and the communication. And if we measure the confidence of machine results, we also have to measure and build trust in interdisciplinary work.

From a technical point of view, there are two main challenges. The first one is to expand the datasets in order to

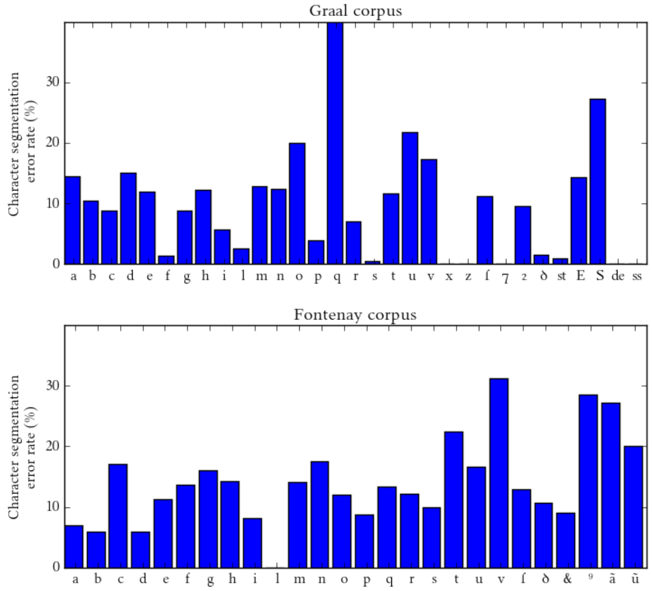


Fig. 8: Automatic segmentation error rate in % for the characters of the Graal and Fontenay corpora, evaluated on a random sample. Only characters with at least 10 occurrences are shown. For Graal : "st" is ligature "s" and "t", "de" is conjunction of "d" and "e" and "ss" is ligature of "s" and "s".

cover the diverse situation of the written communication. We already have worked on this task and produced alignments for dated and datable manuscripts (book hands) from the 12th to the 15th century, and we will continue with inscriptions, registers and cartularies. The expanded corpus, containing both images and transcripts in XML-TEI format, will be published under a Creative Commons license for future research work and reference.

The second challenge is to set common practices to evaluate the transcription alignment. In the present study, the result of word alignment consists of word coordinates with a space for word separation while the ground-truth was given with merged boundaries. The very notion of ground-truth for word as well as for letter alignment and for the identification of graphical phenomena (ligatures, elisions, etc.) has to be deepened. From the Humanities perspective, the very accurate results of our text-image alignment techniques open new paths for exploring and understanding the dual nature of the text. We may now address in a new way the materiality of text and the connection between the ideas and their written instantiations are a matter of cultural history, and history of communication and representations. There are many scholarly editions of ancient and medieval texts, which provide diplomatic or non-diplomatic transcripts of medieval manuscripts. There also are thousands of digitized manuscript and millions of images.

Aligning the very many text transcripts available with the images of the manuscripts allow to explore the notion of variability and evolution in the fields of palaeography, linguistics, epigraphy, communication studies.Each graphical phenomenon may now be analysed according to its context and its particular rules of appearance, so as to gain a better understanding of the history of script and to overcome the issue of letter-form ambiguities, to correlate graphical and phonemic phenomena, and to link morphological evolution to cognitive aspects in reading. The text-image alignment also opens a new field of research for historians by unlocking new levels of granularity, which were unreachable before: the levels of letter and sign level, or even of blank spaces [12]. At the crossroad of Computer Science and Humanities, this joint research has evidenced that, in interdisciplinary research, Humanities and Computer Sciences scholars must "align their research questions" and doing research that is equally interesting and challenging for all partners [13]. One of the common goal is to establish a formal ontology of written signs and their instantiations. This entails confronting the results of automated classification of letter-forms and the criteria of expert driven analysis.

Acknowledgment

This work is supported by the French National Research Agency and Cap Digital through the Oriflamms project (ANR-12-CORP-0010).

References

[1] E. Kornfield, R. Manmatha, and J. Allan, "Text alignment with handwritten documents," in *Int. Workshop on Document Image Analysis for Libraries*, 2004.

[2] S. Feng and R. Manmatha, "A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books," *Joint Conference on Digital libraries*, 2006.

[3] J. Rothfeder, R. Manmatha, and T. M. Rath, "Aligning Transcripts to Automatically Segmented Handwritten Manuscripts," in *Document Analysis Systems*, 2006.

[4] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription alignment of Latin manuscripts using hidden Markov models," in *Workshop on Historical Document Imaging and Processing*, 2011.

[5] M. Al Azawi, M. Liwicki, and T. M. Breuel, "WFST-based ground truth alignment for difficult historical documents with text modification and layout variations," in *Document Recognition and Retrieval*, 2013.

[6] N. Howe, A. Yang, and M. Penn, "A character style library for syriac manuscripts," in *Proc. Workshop on Historical Document Imaging and Processing*. ACM, 2015.

[7] T. Bluche, B. Moysset, and C. Kermorvant, "Automatic line segmentation and ground-truth alignment of handwritten documents," in *International Conference on Frontiers in Handwriting Recognition*, 2014.

[8] A. Zahour, L. Likforman-Sulem, W. Boussellaa, and B. Taconet, "Text Line Segmentation of Historical Arabic Documents," in *International Conference on Document Analysis and Recognition*, 2007.

[9] A.-L. Bianne, F. Menasri, R. Al-Hajj, C. Mokbel, C. Kermorvant, and L. Likforman-Sulem, "Dynamic and Contextual Information in HMM modeling for Handwriting Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2011.

[10] I. Ceccherini and D. Stutzmann, "Analyse morphologique des écritures médiévales," in *IRHT-CNRS Research Report*, 2014.

[11] M. McGillivray, "Statistical analysis of digital paleographic data : what can it tell us ?" *Text Technology*, 2005.

[12] D. Stutzmann, A. Lavrentiev, and Y. Leydier, "Pauses courtes, pauses longues. espacement, ponctuation et sémantique dans la quest del saint graal (lyon, bm, pa 77) : apport de l'analyse d'image assistée par ordinateur," 2015.

[13] D. Stutzmann and S. Tarte, "Digital palaeography: New machines and old texts : Executive summary," vol. 4, no. 7, pp. 112–134.