From Text and Image to Historical Resource: Text-Image Alignment for Digital Humanists

Dominique Stutzmann
Institut de Recherche et d’Histoire des Textes (CNRS), France, dominique.stutzmann@irht.cnrs.fr

Théodore Bluche
A2iA, France, tb@a2ia.com

Alexei Lavrentev
ICAR, Interactions, Corpus, Apprentissages, Représentations (ENS de Lyon – UMR 5191), France, alexei.lavrentev@ens-lyon.fr

Yann Leydier
LIRIS Laboratoire d’Informatique en Image et Systèmes d’information (INSA de Lyon – UMR 5205), France, yann@leydier.info

Christopher Kermorvant
Teklia, France, kermorvant@teklia.com

Introduction

In the context of large, growing digital libraries of texts and digitized medieval manuscripts, the “dual nature” of texts may at last be analysed. Indeed, written texts are both abstract and physical objects: ideas, signs and shapes, whose meanings and graphical systems evolve through time. Attempts have been made to more closely associate both aspects, by textual scholars (TEI community) and palaeographers. A trending approach consists in a text-image association during the editing process (TILE, T-PEN, MOM-CA) and for data visualisation (Mirador, DocExplore), with interoperable annotations schemas (SharedCanvas). Issues at stake cover cultural history, history of communication and representations, materiality of text, authorship and writer identification, dating of textual witnesses, historic semiology.

Is all the previous work obsolete? It should not! New issues arise: transferring traditional Humanities into digital ones and (partially) automate the process.

This paper proposes a method to automatically align images of manuscripts with texts from scholarly editions, at the levels of page, column, line, word, and character. This method has been applied to two datasets “GRAAL” and “FONTENAY” and opens perspectives on automated transcription.

All challenges have to be met in the interdisciplinary work: granularity, scalability, verifiability, exhaustiveness, automation, representativeness, ergonomics, modelling, data format, visualisation, alignment of research questions.

Handwritten text segmentation through recognition

Handwritten text recognition (HTR) aims to automatically transcribe the image of a handwritten text (“text image”) into an electronic text. Most of the current HTR models are based on Hidden Markov
Model (HMM) associated with a sliding window approach to segment the input image. These models may 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, we can use line detection methods and map the transcripts to the detected lines. Such methods also relax the annotation effort needed to produce character segmentation.

A method to automatically segment and annotate handwritten words from line images using forced alignments was proposed by (Zimmermann 2002). The problem has then shifted to mapping the whole transcription of historical documents to segmented words or lines (Kornfield 2004). When the word or line segmentation is not known, a global forced alignment of the full transcript is possible as proposed by (Fisher 2011) or at different levels (word, line, sentence, page) as proposed by (Al Azawi 2013). Our model is similar to the last one but adds yet another level of complexity since we deal with handwritten cursive text: the character segmentation cannot easily be found. OCR-free methods have been proposed (Hassner 2013, Leydier 2014) with lesser results.

Proposed method for handwritten text character segmentation
The character segmentation results of an incremental process (Figure 1): (1°) we convert the image to gray scale and remove black borders, (2°) we apply a text line segmentation algorithm, adapted from (Zahour 2007) to the full page, (3°) we keep the pages with a correct number of detected line, (4°) we assign the line transcript to the lines images and use them to train a first GMM-HMM recognizer, which is (5°) used to align the line transcription with the line images as described in (Bluche, 2014). (6°) Based on this result, we train a new recognizer. This process is repeated until all text lines are correctly transcribed. 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 (example of forced alignment on Figure 2).

Evaluation of alignment accuracy (GRAAL, FONTENAY)
Two methods were applied to evaluate the automatic alignment at word level. First a tabular view to validate/reject each occurrence of a word, evaluate average accuracy, and spot problematic lines. Then a complete validation by a palaeographer. The accuracy is computed by the distance in pixel between the automatic segmentation and the ground-truth word boundaries (distribution on Figure 3 and 4). In GRAAL, 95%, resp. 89%, of left, resp. right, boundaries are correct with a 10px tolerance. In FONTENAY, 91.8%, resp. 89.4% with a 30px tolerance and 85%, resp. 74.4% with a 15px tolerance, less than half of the average character width (23px in GRAAL and 45px in FONTENAY): this is a great achievement.

The alignment was also performed at a character level on GRAAL and FONTENAY. The immense number of characters make a validation difficult. Samples show a very high accuracy rate for complex graphic structures (e.g. 100% for st-ligature), but further tools are needed to measure the accuracy. The results of evaluation partly depend on the validator’s skills in reading ancient scripts.

Evaluation of transcription accuracy (GRAAL)
The HTR-based system was also tested for recognition and evaluated according to the word and character error rate criterion (WER/CER) by splitting the corpus into a training set (101 pages) and a test set (29 pages). For the recognition, we used a lexicon of 7,005 unique words and 4gram statistical language model estimated on the train set. The standard GMM-HMM achieved 23.0% WER and 6.9% CER, whereas the hybrid deep neural network HMM achieved 19.0% WER and 6.4% CER, that is more than 80% of words and 93% of characters are accurately recognised.
Granularity and scalability
This method unlocks a new level of granularity and allows to model different letterforms (“allographs”, e.g. s/l/S). In GRAAL, palaeographers provided the analysis of the graphical chain and the system had to choose between identified possible solutions (Figure 5). In FONTENAY, the system had to create several character models without any previous knowledge, resulting in allograph clustering.

Even at this level of granularity, this system is scalable to large corpora. GRAAL includes 130 pages, 10’700 lines, 114’268 words and more than 400’300 characters; FONTENAY includes 104 pages, 1’341 text lines, 22’276 words and more than 99’900 characters. In comparison, the historical databases “Saint-Gall” and “Parzival” comprise 60 pages, 1’410 text lines, 11’597 words, resp. 47 pages, 4’477 text lines and 23’478 words. The 4-year DigiPal project produced a database encompassing 61’372 manually annotated images of letters, without text transcriptions.

The system is furthermore robust (book hands and diplomatic scripts) and the data format is fully TEI compliant.

Human in the loop: Evaluation, Verifiability and Ergonomics
In interdisciplinary research, Humanities and Computer Sciences scholars must articulate their respective systems of proof and uncover underpinnings and pre-assumptions, in order to produce efficient systems that present data in a way that scholars on all sides can understand, evaluate, and trust. During this research, we observed many times that the so-called ground-truth was not 100% accurate or set too high standards (e.g. transpositions of words) and we explored new paths and hesitations. This is a challenge for future developments: large resources will all comprise inaccuracies or not automatable information. Likewise, some inconsistencies in evaluation may appear. Increase of interactivity in software tools is a solution, not only to overcome shortcomings of strictly automatic approaches, but also to correct the ground-truth and improve the tools and the data models. Therefore this project also developed a user-friendly software (Leydier 2014). Agile development and interoperability concern software creation, but also corpus enhancement, to use Humanist, Computer-Scientist and Machine competences at their maximum. The alignment was performed with 3 person-months. This is obviously less than by manually drawing boxes around words (or even letters). Our tools and system open a large-scale, standardized, interoperable approach of historical scripts. The human in the loop is part of an interdisciplinary work and process avoiding tautological approaches and allowing better results, user-friendly tools and a better understanding on all sides.

References


**Figures**

Figure 1: HTR-alignment

![Flowchart for HTR-alignment](image1)

Figure 2: Forced alignment

![Flowchart for Forced alignment](image2)
Figure 3: GRAAL alignment accuracy (number of occurrences / correction on left or right boundaries, in pixels)

Figure 4: FONTENAY alignment accuracy (number of occurrences / correction on left or right boundaries, in pixels)
Figure 5: Modelling allographs and graphical connexions