International Workshop on Document Analysis Systems

Automatic Handwritten Character Segmentation for Paleographical Character Shape Analysis

Théodore Bluche, Dominique Stutzmann, Christopher Kermorvant

April 12, 2016

a2ia

IRHT
Institut de recherche
et d'histoire des textes

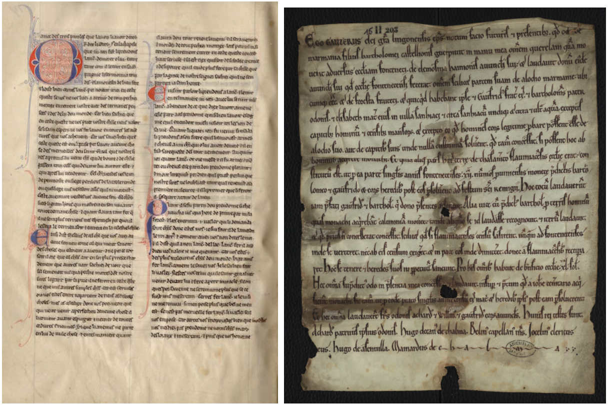
Paleographical Character Shape Analysis

- **Paleography** = study of ancient and historical handwriting
- Goal of character shape analysis: gather occurences of each character and identify different forms or graphical events
- **Digital Humanities:** use automatic approaches (computer vision, HTR) to leverage the large quantity of transcribed data
- Result: about **700M segmented characters** = the biggest database for paleographers

The ORIFLAMMS Project

- **O**ntology **R**esearch, **I**mage **F**eatures, **L**etterform **A**nalysis on **M**ultilingual **M**edieval **S**cripts
- Funded by French National Research Agency (ANR)
- Gloal: **Evolution and variability of handwriting**
  - Latin manuscripts from Europe
  - 12th-15th centuries
  - Inscriptions, books, registers, charters...

Data

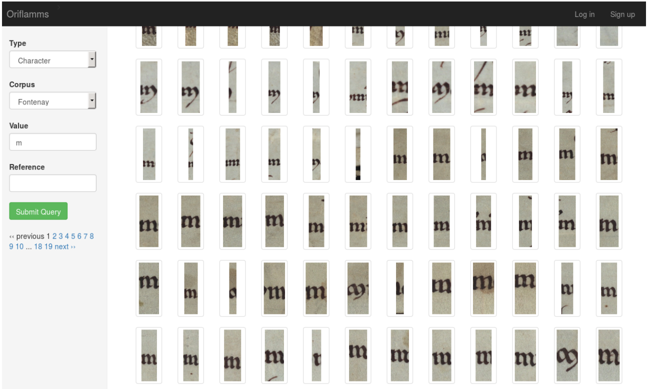


(a)Graal

(b)Fontenay

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

Open Visualisation of Results



Type: Character

Corpus: Fontenay

Value: m

Reference:

900 pages have been automatically segmented into 21241 lines, 198219 words and 694100 characters! ⟶ `http://oriflamms.teklia.com`

Overview

Introduction

Method

Results

Conclusion

Method

Introduction

Method

Results

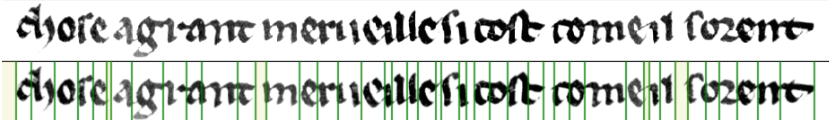
Conclusion

Related work

Text-image alignment / Ground-truth mapping:

- Rothfeder et al. (2006) : G. Washington database : word alignments from text line with HMMs
- Fischer et al. (2011) : St. Gall database : alignment of inaccurate transcriptions from text line images with HMMs
- Kornfield et al. (2004); Stamatopoulos et al. (2010); Leydier et al. (2014) : based on image and transcription features
- Gatos et al. (2014) semi-supervised
- Feng & Manmatha (2006) : align OCR results with ground-truth (text-to-text)
- Al Azawi et al. (2013); Bluche et al. (2014) : using FSTs

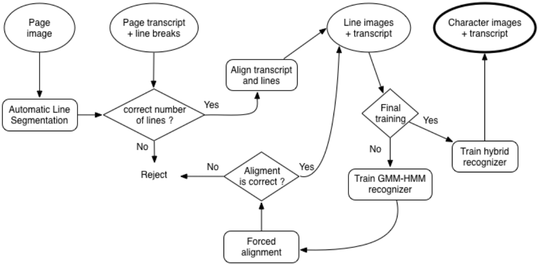
Goal: Retrieve character segmentation from unsegmented transcribed images



Forced alignment: Using HTR for Alignment

- Uses previous scholarly work
- Large corpora ⟶ automation
- Creates the training data for future HTR

Method



1. apply a text line segmentation algorithm to the full page
2. assign the line transcripts to the line images
3. use them to train a first HMM based on GMMs
4. assign the line transcription to the line images with the trained GMM-HMM
5. based on this new alignment, train a new GMM-HMM recognizer.

Finally, train a final text recognizer based on deep neural networks HMMs.

Details of the HTR System

Overview:

- Preprocessing - conversion to gray levels, deskew, deslant, contrast enhancement, height normalization
- Feature extraction - handcrafted features using a sliding window of width 3px with no overlap
- Model - Hidden Markov Models (HMM) associated with a sliding window approach $\longrightarrow$ segmentation of the "text image" as a by-product.

HMMs for characters, and for several writing variants:

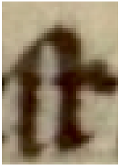
- **Conjunction**: last stroke of the first letter superposed with the first stroke of the second one
- **Elision**: initial stroke of a letter is left out
- **Ligature**: two or more letters are joined as a single glyph
- **Allograph**: the same letter can have different forms

$\longrightarrow$ these phenomena are of **core interest for palaeographers** (allow for identification of scribes, dating, broader understanding of the evolutions of the Latin script in the Middle Age)

Graphical Events Modeling With HMMs

(a)Conjonction (pa)

(b)Elision (eu)



(c)Ligature (st)

(d)Allograph of s

(e)Allograph of v

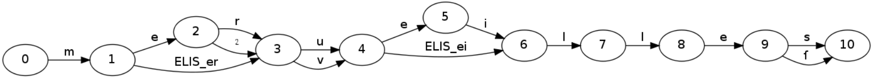


Figure: Example of lexical modeling for the word "merveilles"

Results

Introduction

Method

Results

Conclusion

Segmentation Results

A lot of data were automatically extracted:

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

⟶ how to evaluate the results?

Segmentation Evaluation

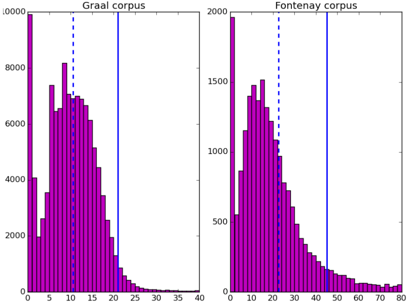
Word segmentation

- manually corrected word positions ⟶ assess automatic alignment quality
- corrected boundaries : $\mathbf{ref} = (\mathbf{ref}_l, \mathbf{ref}_r)$
- segmented boundaries : $\mathbf{hyp} = (\mathbf{hyp}_l, \mathbf{hyp}_r)$
- **Measures:**
  - 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 $= \mathbf{ref}_r - \mathbf{hyp}_r$.

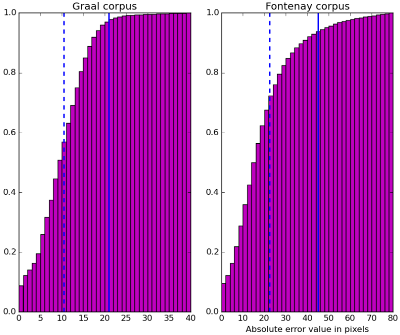
Character segmentation

- randomly selected 2% of these characters using a uniform distribution
- a palaeographer validated the segmentation
- **rejection if**
  - a structural stroke was missing
  - a structural stroke from a neighbour character was added

Word Segmentation - Absolute Error



(a) Histogram of absolute word boundary errors in pixels

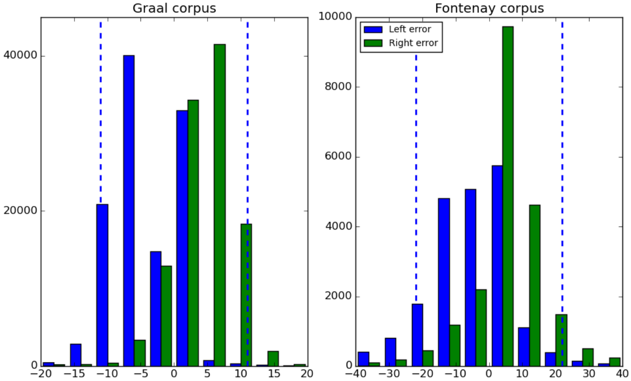


(b) Cumulative histogram of absolute word boundary errors in pixels

dashed line is half a character avg. width, plain line is 1 character avg. width

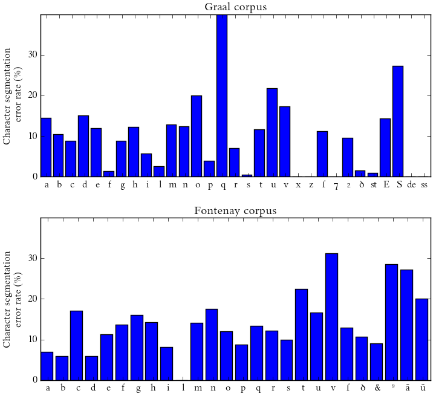
- **Graal** : 63% of boundaries are correct with a 11 px tolerance and 99% are correct with a 23px tolerance.
- **Fontenay** : 72% of boundaries are correct with a 22px tolerance and 94% are correct with a 45px tolerance.

Word Segmentation - Right and Left Errors



⟶ words tend to be cropped

Character Segmentation



On average on all the sampled characters, the segmentation error was

- 10.4% for the Graal
- 13.3% for Fontenay corpus

Conclusion

Introduction

Method

Results

Conclusion

Conclusion

- A lot of characters segmented automatically
- Despite errors, that quantity of alignment and segmentation helped paleographers for their analysis
- Next step (in progress) of automation: automatic clustering of character shapes
- ... also : extend this method to align more corpora, and even transcribe new material
- In the end: successful collaboration in interdisciplinary research
    - aligned corpora will be released publicly at the end of the project (2016)
    - continued collaboration on a new project

Thanks!

tb@a2ia.com

References

Al Azawi, M., Liwicki, M., & Breuel, T. M. (2013). WFST-based ground truth alignment for difficult historical documents with text modification and layout variations. In Document Recognition and Retrieval.

Bluche, T., Moysset, B., & Kermorvant, C. (2014). Automatic line segmentation and ground-truth alignment of handwritten documents. In International Conference on Frontiers in Handwriting Recognition.

Feng, S., & Manmatha, R. (2006). A hierarchical, HMM-based automatic evaluation of OCR accuracy for a digital library of books. Joint Conference on Digital libraries.

Fischer, A., Frinken, V., Fornés, A., & Bunke, H. (2011). Transcription alignment of Latin manuscripts using hidden Markov models. In Workshop on Historical Document Imaging and Processing.

Gatos, B., Louloudis, G., Causer, T., Grint, K., Romero, V., Sánchez, J.-A., Toselli, A. H., & Vidal, E. (2014). Ground-truth production in the transcriptorium project. In Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on, (pp. 237--241). IEEE.

Kornfield, E., Manmatha, R., & Allan, J. (2004). Text alignment with handwritten documents. In Int. Workshop on Document Image Analysis for Libraries.

Leydier, Y., Eglin, V., Bres, S., & Stutzmann, D. (2014). Learning-free text-image alignment for medieval manuscripts. In Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on, (pp. 363--368). IEEE.

Rothfeder, J., Manmatha, R., & Rath, T. M. (2006). Aligning Transcripts to Automatically Segmented Handwritten Manuscripts. In Document Analysis Systems.

Stamatopoulos, N., Louloudis, G., & Gatos, B. (2010). Efficient transcript mapping to ease the creation of document image segmentation ground truth with text-image alignment. In Frontiers in Handwriting Recognition (ICFHR), 2010 International Conference on, (pp. 226--231). IEEE.