HORAE: an annotated dataset of books of hours

Mélodie Boillet, Marie-Laurence Bonhomme, Dominique Stutzmann, Christopher Kermorvant

Teklia SAS, Paris, France
LITIS, Rouen-Normandie University, France
IRHT-CNRS, Paris, France

HIP 2019, 20th September 2019
Horae project

- Book of hours, the medieval *best-seller*: more than 10,000 witnesses
- Personal prayer books, owned by rich laypersons
- Content:
  - perpetual calendar of the Church feasts
  - texts for each of the eight canonical hours (payer times) of the day
  - rich illustrations
- 300 pages, complex organization
- Surprisingly, no complete transcriptions of books of hours
- HORAE Project: automatic text recognition and structuration of book of hours
Les Très Riches Heures du duc de Berry
Project overview

ICDAR2019 → Page classification → Layout Analysis → Handwriting recognition → Text identification

fontes lacerinarum tuaum
ut cum omnibus sanctis
et aeterni dei veni et fessina
in auxilium et consilium
meum in omnibus ora
cionibus et requestis mei
In omnibus augusclis
et necessitibus meis et
in omnibus ilius rebus in
quisus ego sum facturus
locuturus aut cogitatura
omnibus diabus ac nec
tibus horis atque nomen
tis vite me et nicho fa
mulo tuo N imperpetra a
## Manuscripts collection

| Provider          | City      | Manuscripts |
|-------------------|-----------|-------------|
| UGent             | Gent      | 1           |
| ≤ 10              |           | 124         |
| Angers            |           | 21          |
| Autun             |           | 12          |
| Beaune            |           | 15          |
| Chantilly         |           | 30          |
| Nantes            |           | 18          |
| Paris             |           | 17          |
| Rennes            |           | 23          |
| Toulouse          |           | 15          |
| Gallica           | Paris     | 183         |
| Harvard           | Cambridge | 32          |
| UBC               | Vancouver | 1           |
| Stanford University | Stanford | 6           |
| WDL               | Baltimore | 2           |
| **Total**         |           | **500**     |
Layout examples I
Layout examples II
How to select the most representative set of pages?

✗ Randomly : overrepresentation of the text pages and the large manuscripts;

✓ Selection process.
Selection process schema

Manuscript 1

Manuscript 2

...
Selection process schema

Manuscript 1

Manuscript 2

Manuscript 500

Classification

- Binding
- White-page
- Calendar
- Text
- Miniature
- Miniature-and-text
- Text-and-miniature
Selection process schema

Classification
- Binding
- White-page
- Calendar
- Text
- Miniature
- Miniature-and-text
- Text-and-miniature

Filtering
- Binding
- White-page
- 2 max by class
Selection process schema

Classification
- Binding
- White-page
- Calendar
- Text
- Miniature
- Miniature-and-text
- Text-and-miniature

Filtering
- Binding
- White-page
- 2 max by class

Clustering
- HDBSCAN
- 141 clusters
- 2 200 outliers
Selection process schema

Classification
- Binding
- White-page
- Calendar
- Text
- Miniature
- Miniature-and-text
- Text-and-miniature

Filtering
- Binding
- White-page
- 2 max by class

Clustering
- HDBSCAN
  - 141 clusters
  - 2 200 outliers

Selection
- 141 centroids
- 459 outliers
Random selection

Mostly text pages
Our selection

More illustrations
Distribution of the annotated elements using Transkribus

- Pages: 797
- Text region: 843
- Text line: 12512
- Miniature: 284
- Border: 1010
- Initial: 3349
- Line-filler: 1112
- Ormamentation: 5
- Music notation: 4
Annotation examples

- Illustrated border
- Historiated initial
- Miniature
- Simple initial
Annotation examples

- Decorated border
- Text line
- Line filler
How many documents to annotate?

Line and region detection with dhSegment

| Training size | Task                  | IoU with post-processing |
|---------------|-----------------------|---------------------------|
| 220           | Line detection        | 0.88                      |
|               | Layout analysis       | 0.71                      |
How many documents to annotate?

Line and region detection with dhSegment

| Training size | Task                  | IoU with post-processing |
|---------------|-----------------------|--------------------------|
| 220           | Line detection        | 0.88                     |
|               | Layout analysis       | 0.71                     |
| 510           | Line detection        | 0.88                     |
|               | Layout analysis       | 0.72                     |

More data not needed with dhSegment model
Visualization of the predictions I
Visualization of the predictions II
Conclusion and future work

- Introduction of a new dataset Horae including a large variety of types of pages;
- First reference results for line segmentation and layout analysis;
- Satisfactory results that can be improved using more complex neural networks.

- Classification for double-pages → only one class assigned;
- Ambiguity considering the initials → Inside or outside the text lines;
- Confusions between the initials;
- Problem with the post-processing step → Only rectangles are created for now.
| Horae project | Pages selection process | Annotation results | Document layout analysis |
|---------------|-------------------------|--------------------|-------------------------|
|               |                         |                    |                         |

Freely available

https://github.com/oriflamms/HORAE
Bibliography

- Dominique Stutzmann et al. “Integrated DH. Rationale of the HORAE Research Project”. In: Digital Humanities. July 9, 2019. published.

- Emanuela Boros et al. “Automatic page classification in a large collection of manuscripts based on the International Image Interoperability Framework”. In: International Conference on Document Analysis and Recognition. Sept. 1, 2019. published.

- Leland McInnes, John Healy, and Steve Astels. “HDBSCAN: Hierarchical density based clustering”. In: The Journal of Open Source Software 2.11 (2017). DOI: 10.21105/joss.00205. URL: https://doi.org/10.21105%2Fjoss.00205.

- Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. “dhSegment: A generic deep-learning approach for document segmentation”. In: Frontiers in Handwriting Recognition (ICFHR), 2018 16th International Conference on. IEEE. 2018, pp. 7–12.