HORAE: an annotated dataset of books of hours

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

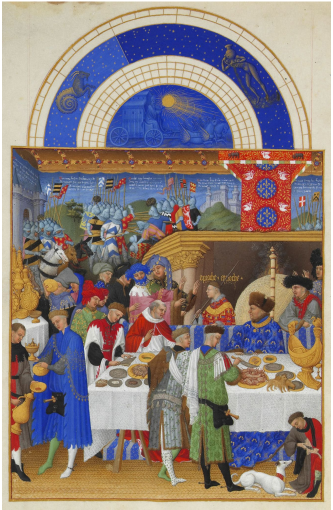
Teklia SAS, Paris, France
LITIS, Rouen-Normandie University, France
IRHT-CNRS, Paris, France

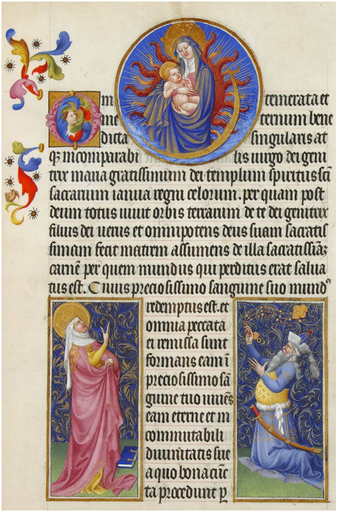
HIP 2019, 20th September 2019

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

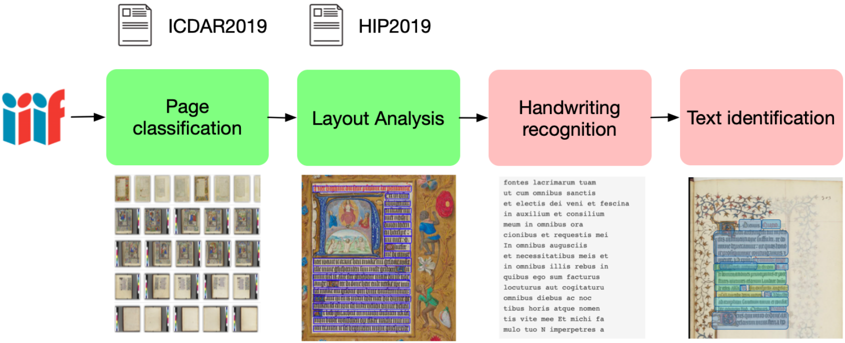
Les Très Riches Heures du duc de Berry





m temerata et
ine ternum bene
dicta ſingularis at
q3 incomparabi lis uirgo dei geni
trix maria gratiſſimum dei templum ſpiritus ſci
ſacrarium ianua regni celorum. per quam poſt
deum totus uiuit orbis terrarum de te dei genitrix
filius dei uerus et omnipotens deus ſuam ſacratiſ
ſimam fecit matrem aſſumens de illa ſacratiſſimã
carnẽ per quem mundus qui perditus erat ſalua
tus eſt. Cuius preciosiſſimo ſanguine ſuo mund⁹
redemptus eſt et
omnia peccata
et remiſſa ſunt
formans carni
preciosiſſimo ſã
guine tuo inueſ
tam eterne et in
commutabili
diuinitatis ſue
a quo bona cũc
ta procedunt p

Project overview

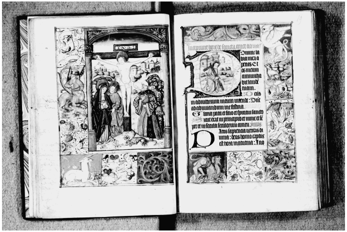


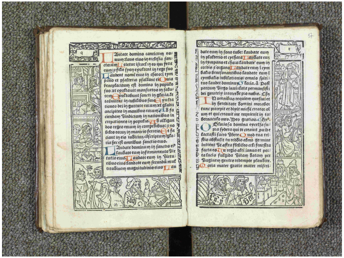
fontes lacrimarum tuam
ut cum omnibus sanctis
et electis dei veni et fescina
in auxilium et consilium
meum in omnibus ora
cionibus et requestis mei
In omnibus augusciis
et necessitatibus meis et
in omnibus illis rebus in
quibus ego sum facturus
locuturus aut cogitaturu
omnibus diebus ac noc
tibus horis atque nomen
tis vite mee Et michi fa
mulo tuo N imperpetres a

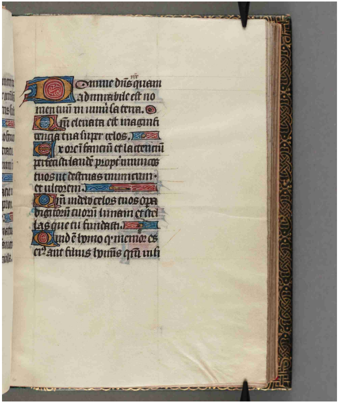
Manuscripts collection

Provider	City	Manuscripts
UGent	Gent	1
BVMM	≤ 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**

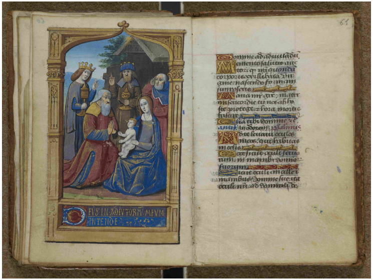
Layout examples I

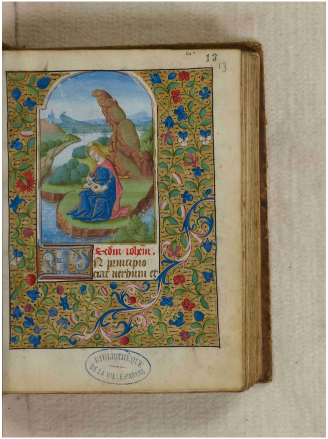






Layout examples II



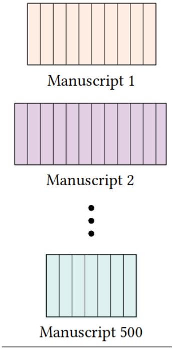


How to select the most representative set of pages ?

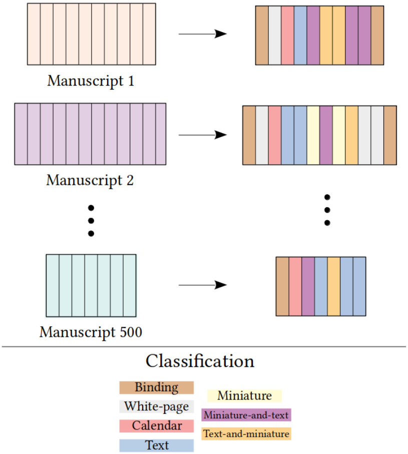
✗ Randomly : overrepresentation of the text pages and the large manuscripts;

✓ Selection process.

Selection process schema



Selection process schema



Manuscript 1

Manuscript 2

Manuscript 500

Classification

Binding

White-page

Calendar

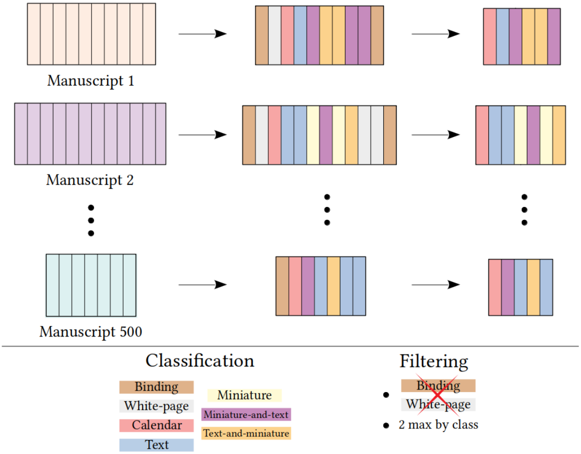
Text

Miniature

Miniature-and-text

Text-and-miniature

Selection process schema



Manuscript 1

Manuscript 2

Manuscript 500

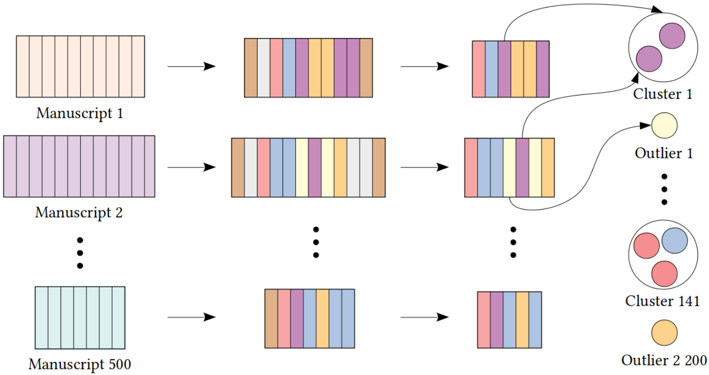
Classification

- 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

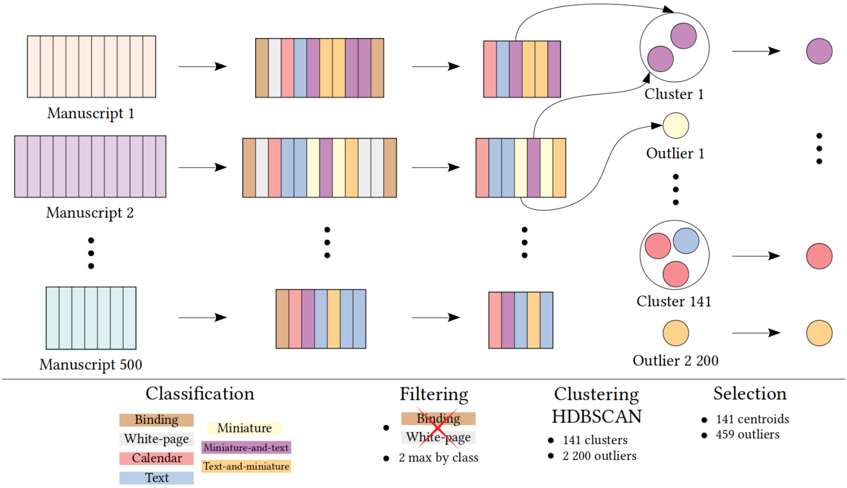
Filtering

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

Clustering HDBSCAN

- 141 clusters
- 2 200 outliers

Selection process schema



Manuscript 1

Manuscript 2

Manuscript 500

Cluster 1

Outlier 1

Cluster 141

Outlier 2 200

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

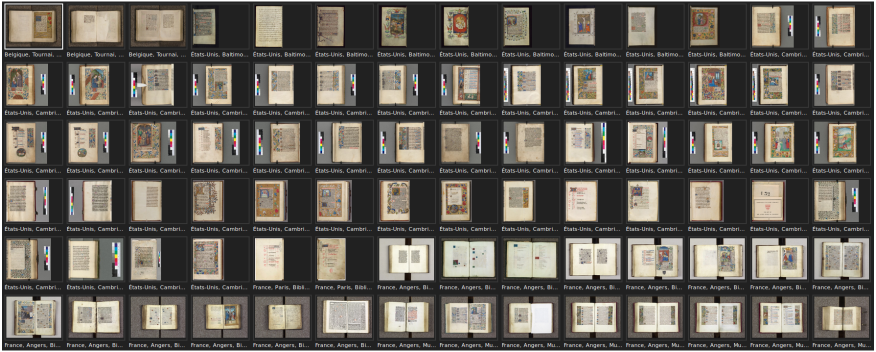
Selection

- 141 centroids
- 459 outliers

Random selection

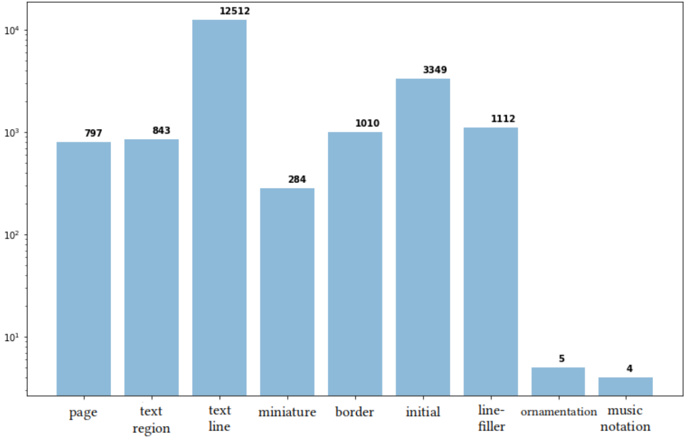
Mostly text pages

Our selection

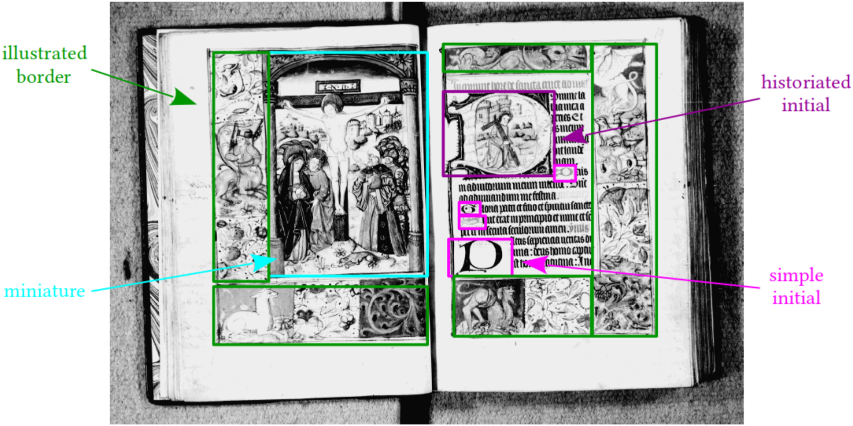


More illustrations

Distribution of the annotated elements using Transkribus



Annotation examples



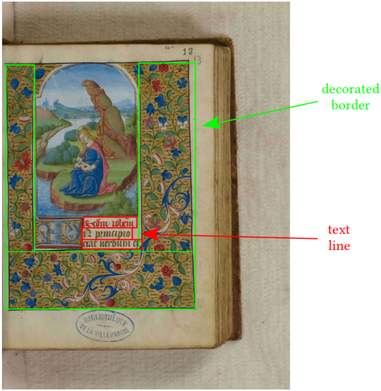
illustrated border

miniature

historiated initial

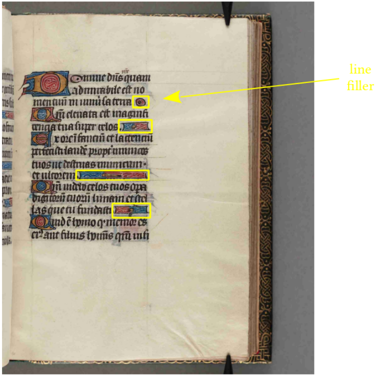
simple initial

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

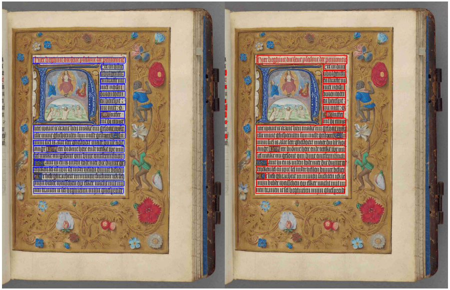
How many documents to annotate ?

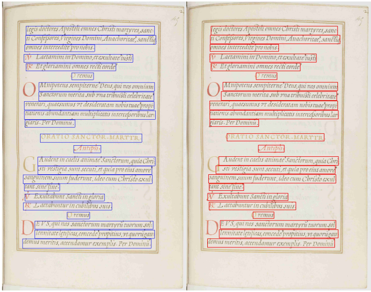
Line and region detection with dhSegment

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

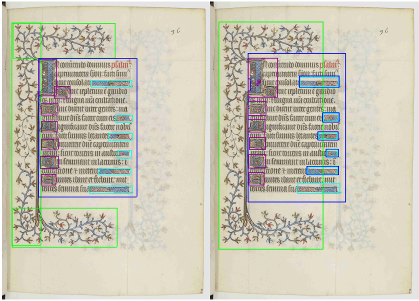
More data not needed with dhSegment model

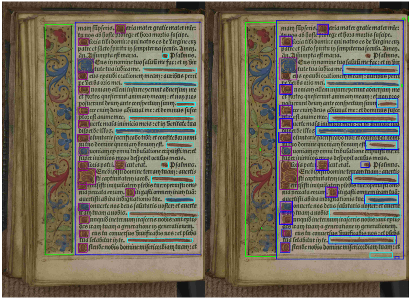
Visualization of the predictions I





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.

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.