Hierarchical decomposition of handwritten manuscripts layouts

Vincent Malleron, Véronique Eglin, Hubert Emptoz, Stéphanie Dord-Crouslé, Philippe Régnier

Laboratoire d'InfoRmatique en Image et Systèmes d'information

UMR5205 CNRS/Université de Lyon/INSA de Lyon

INSA de Lyon - Bâtiment Blaise Pascal, 7 avenue Jean Capelle - 69621 Villeurbanne Cedex, France

http://liris.cnrs.fr

Tel: +33 4 72 43 63 65; fax: +33 4 72 43 71 17; e-mail: vincent.malleron@liris.cnrs.fr

Context

- Handwritten manuscripts gathered by Flaubert
- 7 different writers
- 3200 pages, essentially made of text fragments
- Differents styles and layouts
- To make the corpus usefull, fragments have to be separated either manually or automatically

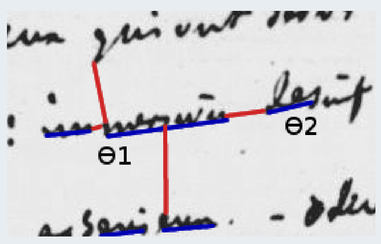
Objectives

Extract as many structural information as possible :

- To simplify the work of Human Sciences specialists
- To provide preliminary results for fragments classification
- To propose a new way of navigation in the corpus

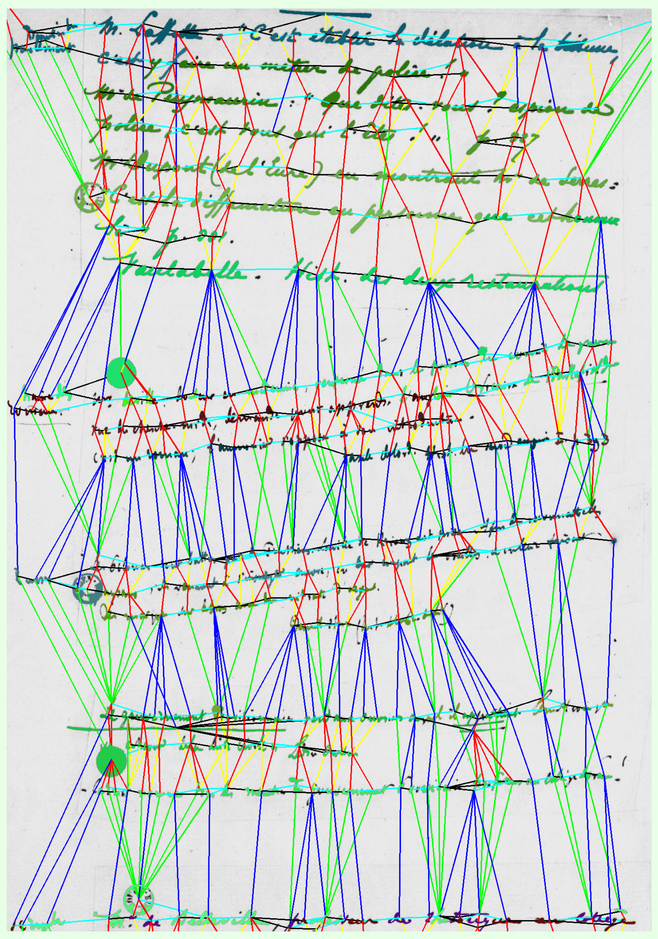
Our Method

- Graph-based approach
- Graph build using structural relationships between connected components
- Link orientation is estimated using Hough transform
- Edge to Edge connected components distance

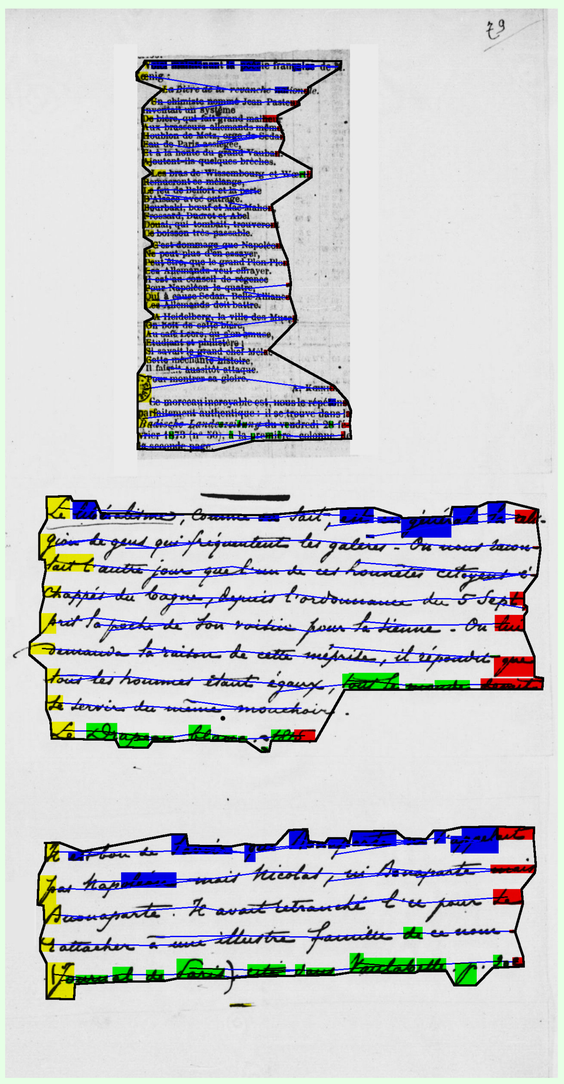


- Weighted Directed Graph G=(V,A)
- $V = \{v_1, v_2, ..., v_n\}$. v_i = a connected component
- $A = \{e_1, e_2, ..., e_n\}$ e_i = an edge linking 2 connected components
- Graph labelling to extract left, right, top and down borders of the page
- Text line extraction using shortest path research between a left and right border
- Fragment extraction using interline spaces rules

Graph representation



Fragment Extraction



Conclusions

- A dedicated approach for page structure extraction
- Hierarchical decomposition : page, fragments, lines...
- Usefull pre-processing task for Human Sciences researchers
- Approach can be adapted to any corpus composed of notes pages
- And to other documents...

Applications

- Image/Transcription matching
- Corpus Navigation improvement
- Corpus enrichment with page and fragment indexation

Results

Page	Wrong Lines	Correct Lines	%Correct Lines
Simple Note	3	27	90%
Complex Note	3	31	91%
Gathered Informations	1	22	95%
20 pages dataset	53	414	90%

References

L. Likforman Sulem, A. Zahour, and B. Taconet.
Text line segmentation of historical documents : a survey.
IJDAR, 9(2-4) :123–138, April 2007.

L. O'Gorman.
The document spectrum for page layout analysis.
IEEE Trans. Pattern Anal. Mach. Intell., 15(11) :1162–1173, 1993.