# Text lines and snippets extraction for 19th century handwriting documents layout analysis

Vincent Malleron, Véronique Eglin, Hubert Emptoz
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205,
F-69621, France
vincent.malleron@liris.cnrs.fr

Stéphanie Dord-Crouslé, Philippe Régnier
Université de Lyon, CNRS
LIRE, UMR 5611
F-69007, France
Stephanie.Dordcrousle@ens-lsh.fr

## Abstract

*In this paper we propose a new approach to improve electronic editions of human science corpus, providing an efficient estimation of manuscripts pages structure. In any handwriting documents analysis process, the text line segmentation is an important stage. The presence of variable inter-line spaces, of inconstant base-line skews, overlapping and occlusions in unconstrained ancient 19th handwritten documents complexifies the text lines segmentation task. In this paper, we only use as prior knowledge of script the fact that text lines skews can be random and irregular. In that context, we model text line detection as an image segmentation problem by enhancing text line structure using Hough transform and a clustering of connected components so as to make text line boundaries appear. The proposed approach of snippets decomposition for page layout analysis lies on a first step of content pages classification in five visual and genetic taxonomies, and a second step of text line extraction and snippets decomposition. Experiments show that the proposed method achieves high accuracy for detecting text lines in regular and semi-regular handwritten pages in the corpus of digitized Flaubert manuscripts (”Dossiers documentaires de Bouvard et Pécuchet”, 1872-1880).*

## 1. Introduction

Our work takes place in an human science project which aims at the realization of an electronic edition of the ”dossiers de Bouvard et Pécuchet” corpus. This corpus is composed of French 19th century manuscripts gathered by Gustave Flaubert in order to prepare the redaction of a second volume to his Novel ”Bouvard et Pécuchet”. Corpus contents are diversified in term of sense as well as in term of shape (different writers, styles and layouts). Besides, the corpus is mainly composed by text snippets (Newspapers extracts, various notes, etc.) put together by Flaubert. To produce the electronic edition we must consider the particular framework of the corpus : structure informations must be known in order to reproduce as well as we can its primary state, and restore snippets mobility. In order to retrieve structure informations we propose a three-step algorithm : at first, we roughly determine the page layout, then we extract the lines of the manuscript, and we finally rebuild the page structure. This paper is organized as follows : section 2 will detail previous works on text line extraction and structure recognition, section 3 and 4 will present our approach for page layout analysis, text lines and snippets extraction, section 5 provides results and perspectives, section 6 gives concluding remarks.

## 2. Related works

Handwritten text line segmentation is still a challenging task in ancient handwritten document image analysis. All segmentation techniques that must be applied to handwritten documents provide information for skew correction, zone segmentation, and sometimes character recognition. As related in [3], some preprocessing techniques must be performed before text line extraction so as to enhance text lines. Here we only focus on text lines and zones extraction techniques.

In works on handwritten documents it is often assume that text lines are already segmented or easy to be segmented so that conventional methods can be applied. But in practice handwriting variability is omnipresent in ancient handwritten documents and it is not trivial to extend machine printed documents algorithms to handwritten documents, especially when handwritten text lines are curvilinear and when neighboring handwritten text lines may be close or touch each other. In handwriting drafts for example, the page layout cannot be handled by simple rules, such as grouping based on geometric relationships of neighboring components, [1, 8, 4, 5]. So it is a real challenge to analyze freestyle handwritten document also due to non standard and irregular layout at the page level.

At lines level, conventional approaches based on linear approximation and regression can not always be efficient. At words level, due to the frequent connectivity between characters and words, it is often complex to process the connected components. A connected component can represent character of different sizes that is considered as an important parameter for the bottom-up connected component based text line segmentation algorithms, [6].

Most of works based on text line segmentation can be roughly categorized as bottom-up or top-down approaches. In the top-down methodology, a document page is first segmented into zones, and a zone is then segmented into lines, and so on. Projection based methods is one of the most successful top-down algorithms for printed documents and it can be applied on handwritings only if gaps between two neighboring handwritten lines are sufficient [9].

Of course, one strong hypothesis for the application of such techniques is the existence of standard and regular page layout. Other researchers use different assumptions for specific script and specific contexts in mixed strategies, both top down and bottom up[7]. An achieved taxonomy of approaches reveals six major categories[3] : projection-based, smearing, grouping, Hough-based, repulsive-attractive network and stochastic methods. For the authors, most of these methods are able to face some image degradations and writing irregularities specific to historical documents. In that context, we propose an approach that is well adapted to authors handwritings.

## 3. Page layout analysis

Our corpus is composed by pages of various styles and shapes. We can classify most of the pages in the following categories, represented on figure 1 : (1-Note pages, 2-Gathered Informations, 3-Letters, Novel extracts, 4-Dictionary, 5-Printed text (not represented))

We introduce three simple rules to classify corpus pages into those fives categories : number of library stamps in the page, amount of data in page and profile dispersion.

For the original handwritten pages conservation, the library has appended a stamp near each pasted textual fragments. Therefore number of stamps in the page allow us to separate easily dictionary pages, gathered informations pages and others : a dictionary page has a least seven stamps, pages of gathered informations have between two and six stamps, other pages have one stamp or none. The number of library stamps in page is retrieved by template matching. Printed text pages are generally pages of high text density. We computed the amount of data in each page in order to separate printed text pages from others. The separation between letters, gathered informations and note pages is done by computation of profile dispersion :

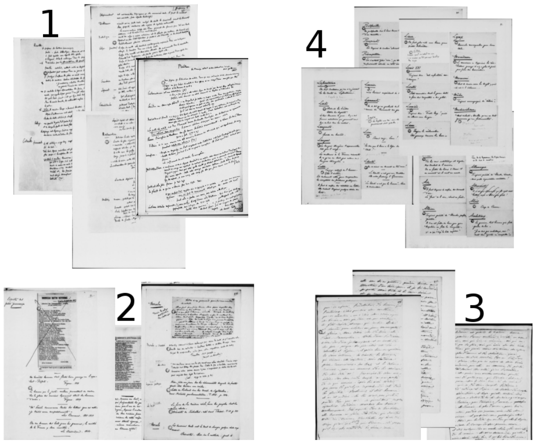

**Figure 1. Sample images of our corpus**

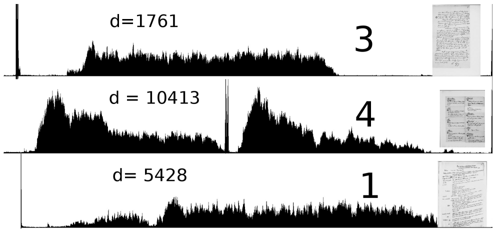

**Figure 2. Profile Dispersion**

$$d = \frac{\sum(Max(I) - I(y))}{X}$$

where I is the vertical projection profile value and X the width of the page.

Figure 2 shows that vertical profile projection characterizes the layout of the page. Typical profiles and dispersion values are provided.

The knowledge of page layout style helps us to choose algorithms and decision values for lines and snippets extraction. In the following sections we will only present our algorithm on exclusive handwritten pages of the corpus. Machine printed pages are easy to separate from the rest of the corpus and have been processed successfully with an OCR.

## 4. Text Line and Snippets Extraction

Text line extraction is the preliminary work for snippets extraction. Retrieving the lines of the text in the manuscript is also an important issue for transcription matching with the manuscript. Our algorithm relies on 5 steps : connected components extraction and neighborhood-fan computation, corner and borders detection, line orientation estimation, line construction and post-processing.

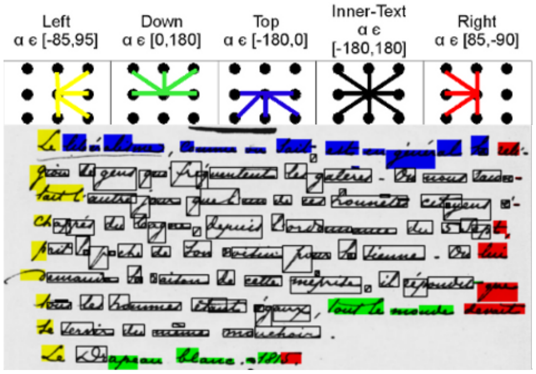

**Figure 3. Borders and NeighbouroodFan**

### 4.1 Connected components analysis

We extract connected components on binarized image version. Since images have a low quality (200 dpi microfilms) the connected components have various size. They could be a word, a syllable a diacritic sign or a letter. A post processing task separate diactritics from text and filter large components such as underlines. Those informations are stored separately for a further use.

Once the connected component extraction is performed we compute for each component a neighborhood-fan which contains euclidean distance and orientation of the component nearest neighbors in 18 direction of space ([-180,-170],[-170,-160],...,[170,180]). Distance range is restricted to 300 pixels. (Figure 3) This computation allow us to know the component neighborhood structure and prepare the further algorithms steps.

### 4.2 Corner and Borders extraction

Corner and Borders extraction provide a primary information of the document structure and an initialization for text line reconstruction. Extraction is realized using a k-means classification on a neighborhood fan based vector. We extract the 5 following classes : Left components, Right components, Top components, down components and inside-text components. We use a 18 value vector, each value standing for a spatial direction. The value is 1 if there is no neighbor in the direction of the space, and 0 is there is one. The distance between two vectors is computed as follow :

$$D(V_1, V_2) = \sum_{i=0}^{17} \rho_i |V_1(i) - V_2(i)|$$

with $\rho_i = 2\ if\ i \in 0, 4, 8, 9, 13, 17, \rho_i = 1\ elsewhere$
$\rho_i$ gives more weight to components located in the most frequent space directions. Figure 3 presents the result of border extraction on a sample snippet and associated neighborhood-fans.

### 4.3 Line orientation estimation

For each component we estimate the local line orientation with Hough transform. The transform is performed at low resolution after application of a canny edge detector. We compute an orientation map of extracted hough lines and each component gets its orientation from the associate line in orientation map. Out of bound orientations are filtered in order to avoid errors. We consider that line orientation in our document is in a -45/+45 degrees range. Hough based methods are mainly adapted to straight lines but the Hough transform has been computed here for each local connected component considered as small window of text inspection. It allows to solve some recurrent problems of line proximity, overlapping and fluctuating close lines. Some local considerations (i.e. in small inspection windows that can be included into connected components bounding boxes) could be thought to solve some other difficulties based on touching strokes.

### 4.4 Line construction

Line construction is realized using the results of the two previous steps for the initialization. A text line starts on a right-labeled component and stops on a left-labeled component. Research direction is given by the hough transform result and by the latine script direction. We consider that orientation has a small variation in a line. When hough transform result differs from more than 10 degrees to the current line mean orientation, mean orientation is used for the next computation step. Sample results of the line extraction are shown on figures 4,5,6. A post-processing step is performed to merge line detected at the same horizontal position and include a component positioned between two components of the same line in the line.

### 4.5 Snippets decomposition

Snippet extraction has been developed here to perform transcript mapping with non rectangular polygonal fragments. The method must fit the document layout :

Novel extracts or letters pages generally contain only one snippet, containing the whole page. Gathered Informations snippets are generally easy to extract on those pages, because the interline space between snippets is high. An example of a typical gathered informations page is illustrated on figure 4. Notes pages have the most complicated structure type : we must use relationships between sense and structure to perform snippet decomposition. We can also notice on figure 6 some errors occurring on the two text

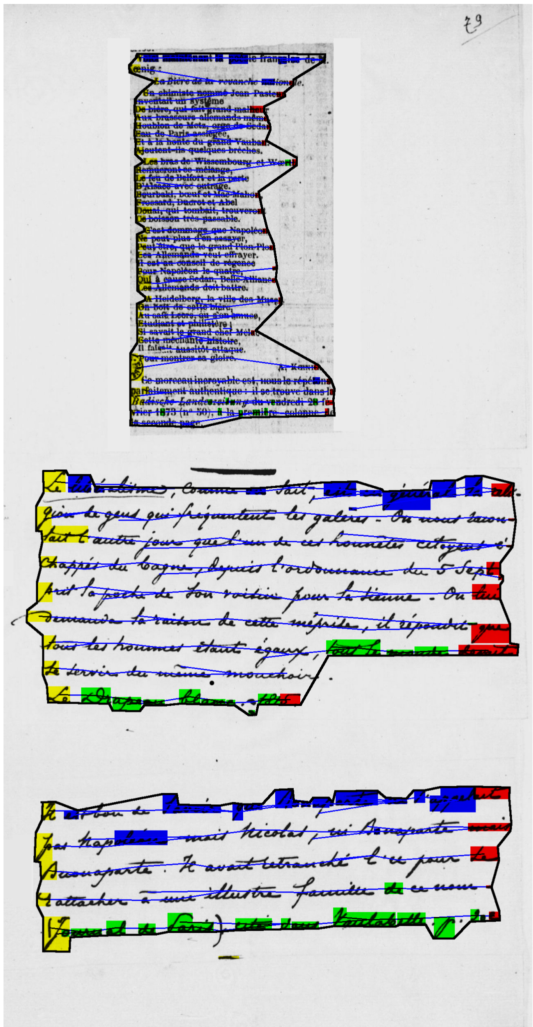

**Figure 4. Gathered Information structure**

lines at the bottom of the fragment. We intend to solve those difficulties using a classification based on interline space, margin analysis and line orientation variation to realize the page structure extraction. As a result, snippets are build by linking centroids of borders components. (figure 4)

## 5. Results and Discussion

Our page layout classification heuristic has been evaluated on a sample of 280 labelled images of the corpus, representative of the corpus diversity (about 3000 pages). Table 1 shows that we characterize pages layout with 3 simple characteristics. Addition of other characteristics could improve our results but a rough classification is enough to perform the followings steps of our algorithm. Figure 4 shows the result of snippets decomposition on a gathered informations page. The layout of this page is relatively easy

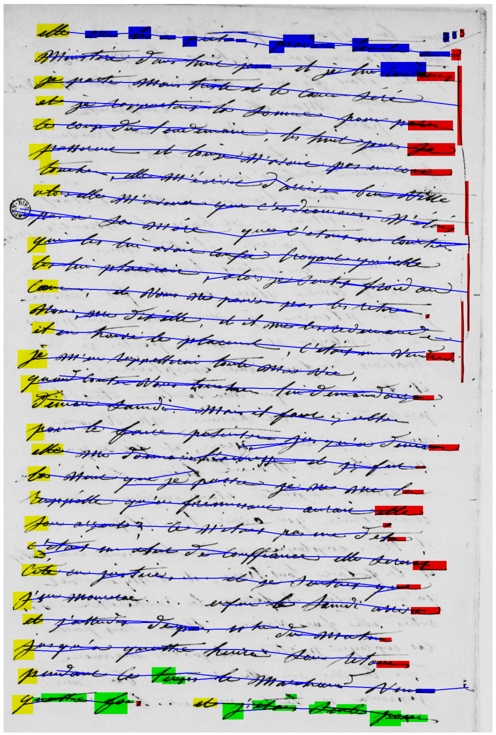

**Figure 5. Novel extract page structure**

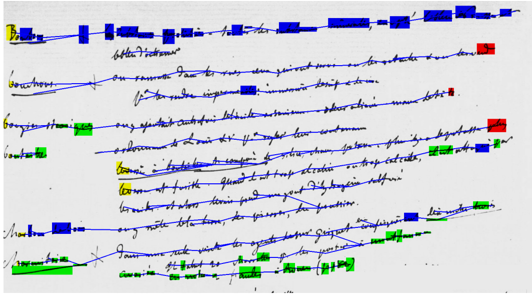

**Figure 6. Note page structure**

| Class | Images | Recall | Precision |
|---|---|---|---|
| Letters | 50 | 0.90 | 0.92 |
| Gathered Inf. | 45 | 0.95 | 0.86 |
| Printed Text | 50 | 0.90 | 0.96 |
| Notes | 70 | 0.71 | 0.78 |
| Dictionary | 52 | 0.96 | 0.88 |
| All | 267 | 0.884 | 0.88 |

**Table 1. Recall/Precision Table of classification**

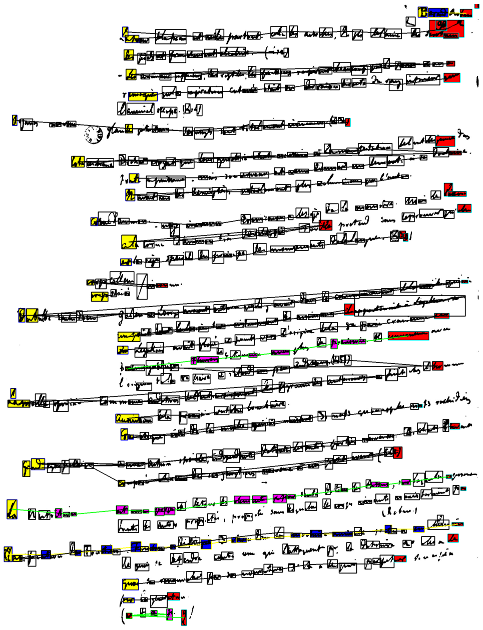

Figure 7. Complex Note Page

to extract and our method perform a successful extraction of both lines and snippets. The first line of a snippet is figured in blue while the last line is figured in red. Figure 6 displays our algorithm results on a note page. Text line extraction is successfully performed. Note that curvilinear lines are correctly extracted. Starting points of lines are also generally well computed and provide reliable structure information. The major difficulty on this kind of page consists in obtaining a precise text line, with all descenders and ascenders segmented for accessing isolated and non overlapping components. As segmentation and text line reconstruction are two dependent tasks, an exact segmentation of touching writing connected components pieces may need some knowledge about possible and regular touching configurations. Figure 5 exposes results for fourth category pages. We obtain similar results on dictionary pages. Line structure information is successfully retrieved. Fourth category pages structure are mainly composed by a single snippet : we only extract line structure. Dictionary pages are preprocessed using a vertical profile projection to separate the two columns. Afterward we use the same algorithm to extract line structure of the column. Figure 7 shows the current result of our approach on a more complicated page of the corpus. Current limitations are enlighted : a wrong border extraction leads to the omission of several lines. A false orientation information, line overlapping can occur. Bad snippets extraction could also appear when a line is not well detected. Our approach will be completed by using non textual elements (diacritics, stamps or underlines) to improve snippets decomposition.

## 6 Concluding remarks

In this paper, we proposed a dedicated text line segmentation approach for author's draft handwritings. Knowing that snippets extraction on an humanist corpus is usually a costly and time-consuming hand-made task, it is necessary to provide useful tools for the pre-extraction of snippets in a drafts documents. Experiments of our approach show that our proposition is really consistent regarding the complexity of many page layouts in the corpus. Our methodology can be compared to conventional text line segmentation methods, such as projection based top-down approaches and other connected component based bottom-up approaches and shows very promising results.

## 7 Acknowledgment

These studies had the support of the Rhone-Alpes Region in the context of cluster project. We also want to acknowledge the consortium of the "Bouvard et Pécuchet" ANR Project. Images are property of Rouen's town library.

## References

[1] L. A. Fletcher and R. Kasturi. A robust algorithm for text string separation from mixed text/graphics images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(6):910–918, 1988.

[2] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger. Script-independent text line segmentation in freestyle handwritten documents. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1313–1329, August 2008.

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

[4] R. Manmatha and N. Srimal. Scale space technique for word segmentation in handwritten documents. In *SCALE-SPACE '99*, pages 22–33, London, UK, 1999. Springer-Verlag.

[5] V. Martin and H. Bunke. Text line segmentation and word recognition in a system for general writer independent handwriting recognition. In *Proc. International Conf. Document Analysis and Recognition*, page 159163, 2001.

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

[7] U. Pal and P. P. Roy. Multioriented and curved text lines extraction from indian documents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34(4):1676–1684, 2004.

[8] J. L. Rothfeder. A scale space approach for automatically segmenting words from historical handwritten documents. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8):1212–1225, 2005. Member-R. Manmatha.

[9] B. Yu and A. Jain. A robust and fast skew detection algorithm for generic documents. *PR*, 29(10):1599–1629, October 1996.