Hierarchical decomposition of handwritten manuscripts layouts

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

Hierarchical decomposition of handwritten manuscripts layouts

Vincent Malleron[1,2], Véronique Eglin[1], Hubert Emptoz[1],
Stéphanie Dord-Crouslé[2], and Philippe Régnier[2]

[1] Université de Lyon, CNRS,
INSA-Lyon, LIRIS, UMR5205,
F-69621, France
vincent.malleron@liris.cnrs.fr
[2] 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 literary corpus, providing an efficient estimation of manuscripts pages structure. In any handwriting documents analysis process, structure recognition is an important issue. The presence of variable inter-line spaces, of inconstant base-line skews, overlappings and occlusions in unconstrained ancient 19th handwritten documents complicates the structure recognition task. Text line and fragment extraction is based on the connexity labelling of the adjacency graph at different resolution levels, for borders, lines and fragments extraction.*

Key words: text lines and fragments extraction, graph, handwriting

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 by french 19th century manuscripts gathered by Gustave Flaubert who intends 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 fragments (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 possible its primary state. The main goal of this work is to retrieve as many structural information as possible in order to provide a good estimation of handwritten document pages structure. The document structure is mainly composed by pages, fragments, lines, words and characters. Our proposition consists in representing the overall page layout with and oriented adjacency graph that contains all information relative to the page content. This

paper is organized as follows : section 2 details previous works on text line extraction and structure recognition, section 3 presents our approach for text lines and fragments extraction, section 4 provides results and perspectives and section 5 gives concluding remarks.

2 Related Works

Handwritten text line segmentation and structure recognition are still challenging tasks in ancient handwritten document image analysis. 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 algorithm for printed documents and it can be applied on handwritings only if gaps between two neighboring handwritten lines are sufficient [1]. Connected component based methods is a popular bottom-up method : connected components are grouped into lines, and lines into blocks. In [2] and [3] the algorithms for text lines extraction are based on connected components grouping coupled with Hough Transform are exposed. Likforman-Sulem et al. [5] give an overview of all text lines segmentation methods. In [6], Yi Li et al. propose a curvilinear text lines extraction algorithm based on level-set method. This algorithm uses no prior knowledge, and achieves high accuracy text line detection.
Most of works on document structure recognition are performed on machine-printed texts. In [7] T.M.Breuel presents algorithms for machine-printed documents layout analysis. Theses algorithms are noise resistant and adapted to different layouts and languages. Kise et.al [8] propose an algorithm for machine printed pages layout analysis based on Voronoi diagrams and successfully achieve segmentation of document components in non-manhattan documents. Lemaitre et. al in [4] propose an application of Voronoi tessellation for handwritten documents recognition. 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 [9], L.O'Gorman's Docstrum allows to retrieve layout structure of handwritten pages with regular layouts. S.Nicolas et al. in [10] and [11] propose an Hidden Markov Model based algorithm for manuscript page segmentation : 6 classes are extracted on a page representing background, textual components, erasures, diacritics, interline and interword spaces.

3 Our graph based approach

3.1 Connected components distance measure

We introduce at first our connected components distance measure : let's consider two handwritten shapes A and B that can represent words, word fragments or characters. The distance between A and B is given by the smallest edge to edge

distance (d_{edges}). This measure is more representative of page structure than a simple Euclidean distance from center to center and provide an estimation of interwords and interline spaces (figure 1).

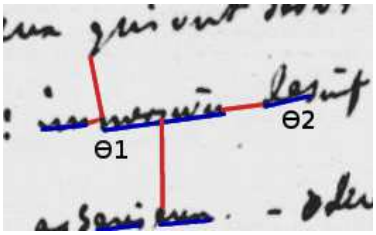

Fig. 1. Connected components distance

To be consistent with orientation variation in handwritten documents we weight our measure with and orientation coefficient. This coefficient, based on the fact that orientation remain mainly constant in a line or a fragment is computed using orientation of the two connected components : $\theta 1$ and $\theta 2$.

$$\Delta\Theta = \alpha * (1 + |\frac{|\theta 1| - |\theta 2|}{|\theta 1| + |\theta 2|}|) \tag{1}$$

Orientation is estimated using Hough transform at low resolution. We compute an orientation map of extracted hough lines and each connected component gets its orientation from the associate line in orientation map. Our distance can be summarize with the following statement :

$$D(A, B) = \Delta\Theta * min(d_{edges}(A, B)) \tag{2}$$

Figure 1 shows baselines in blue and minimal edge to edge distances between connected components in red. As $D(A, B)$ represents the distance between two connected components and not between their contour edges, the three distance properties can be simply demonstrated. For the graph construction and labelling we use this distance to find nearest neighbours of each connected component in different orientations ranges to build the adjacency graph and perform lines and fragments extraction.

3.2 Graph construction

For each connected component we search for the nearest neighbour in four direction of space : top, down, left and right. Exact directions are provided by the orientation estimation described in 3.2 : nearest neighbour research is performed around hough direction, orthogonal hough direction, inverse hough direction and inverse orthogonal hough direction. Once the four neighbours are computed for each connected component we build a weighted directed graph $G = (V, A)$.

$V = \{v_1, v_2, ..., v_n\}$ with v_i representing a connected component of our page. Outdegree of G is 4 : each vertex is the tail of up to four arcs ($e_i = (v_i, v_j)$), representing the link between the connected component and its neighbours : $A = \{e_1, e_2, ..., e_n\}$ is the arc set of our graph. Arc weights are provided by the real distance between connected components. This graph can be reprojected on a manuscript image as shown on figure2. Right arcs (direct hough direction) are colored in cyan, left arcs (inverse hough direction) are colored in black, top arcs (orthogonal hough direction) are colored in red or blue depending of the weight and down arcs (inverse orthogonal hough direction) are colored in blue. When two arcs are superposed only tops and rights arcs are represented. We can also observe that several arcs can converge to a single connected component. This single connected component is generally a long connexity (a long connected word, a straight line or underline).

3.3 Graph labelling for a multi level layout structure analysis

Arcs weight analysis In order to set thresholds for borders, text lines and fragments extraction we compute an histogram of arcs weights only in top and down directions. Figure 3 represents the histogram of cumulated distances. Interline spaces clearly appears between 30 and 120 whereas interfragment spaces are above 120. Threshold values for text lines, borders and fragments extraction are computed from this histogram by considering the highest local gradient of the histogram.

Text borders extraction The purpose of this step is to label each vertex of G with his corresponding label in the five classes described in figure 4. Adjacency function is a simple function, computed on an arc which returns a value corresponding to the direction of the arc (DirectHough = 1, InverseHough = 2, OrthHough = 3, InvOrthHough =4, NoArc = 0). To extract text borders, adjacency function also used the maximum range value computed above. In practice, the graph labelling is based on a simple evaluation of the neighbourood of a vertex. If the outdegree of the vertex is equal to 4, it represents an inner-text connected component. If the outdegree is less than 4, the label is computed given the result of adjacency function on each arc. We use the following color scale on figure 4 to show the results of border extraction : yellow for left components, red for rights, blue for tops, green for downs and black for inner-text ones.

Text lines extraction The text line is an important structural element : the knowledge of text-lines positions and shapes is an essential step that allows to perform the alignment between handwritten images fragments and ASCII transcriptions. It also gives us a partial knowledge of page structure and therefore a good a priori for fragments extraction. The result of border extraction is used as an initialization. To be consistent with latine script direction a line starts with a left border component and ends with a right border component. Line extraction is performed using a shortest path research between left and right

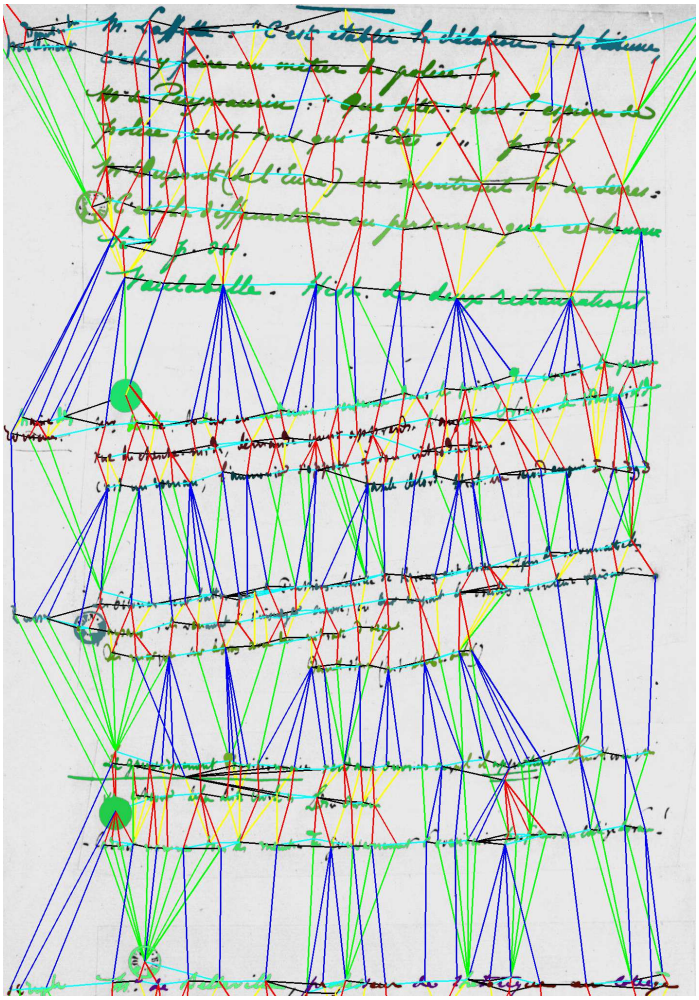

Fig. 2. Reprojected graph, 1f188r

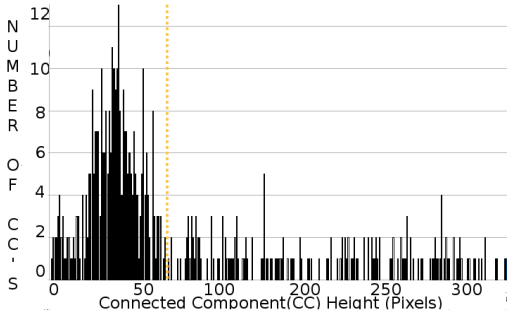

Fig. 3. Arcs weights histogram

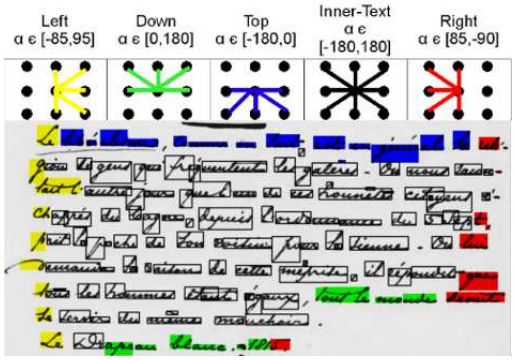

Fig. 4. Borders extraction results

borders on the previously described graph. A postprocessing step is performed to include a missed component positioned between two components of the same line.

Fragments extraction The text fragment is the highest structural element extracted in the page. Textual fragment extraction has been developed to fit with the particular framework of our corpus, mainly composed by text fragments. Text fragments extraction is done by line grouping with simple rules based on interline space and orientation variation coupled with fragment contours computation by a best path research between top, down, left and right vertices. The path starts and ends on the same vertice and describe the fragment contour. Cost function for best path computation is based on distance values and transition costs between tops, downs, lefts and rights vertices. Threshold values are given by the distance histogram described above. Figure 2 shows interline space variation : large interline spaces arcs are colored in blue and green whereas small interline spaces arcs are colored in yellow or red. Figure 5 presents the results of fragments extraction.

4 Results and Perspectives

The proposed algorithm for text lines extraction has been tested on a sample set of the "dossiers de Bouvard et Pécuchet" Corpus. A few number of ground-truthed images are currently available. Evaluation has been performed on pages of different shapes and layout in order to be representative of the corpus. A line is considered as wrong when less than 75 per cent of the line is included in the line bounding box. Overlapping between lines makes the two overlapped lines count as wrong. Table 1 shows the text lines extraction results for sample pages of the corpus.

Figure 5 shows results of text fragments extraction on a sample page of our corpus. Five of the six textual fragments of this page are correctly extracted.

Page	Wrong Lines	Correct Lines
1 f 179 r	1	15
1 f 007 v	4	27
228 f 020 r	2	21
4 f 234 r	0	24
1 f 188 r(fig 5)	1	24
Simple Layout Pages	14	198
Complex Layout Pages	39	216

Table 1. Line extraction results

Our algorithm based on interlinespace and baseline orientation comparison performs well on page of simple layout. Some limitations appears on page of more complex layouts : errors can occurs when two fragments are adjacent with a small orientation variation or when topological configuration of space cannot be describe with our distance. Those limitations can be seen on figure 5. Our graph based approach remains insensible to classic connected components approach limitations such as connected components overlapping or inclusions.
Graph based representation is an intuitive multiscale representation which allows us to describe and extract the layout of complex handwritten pages. It also allows to identify special configurations of space, such as underlines or included connected components. Theses configurations could be used to improve the results of lines and fragments extraction.

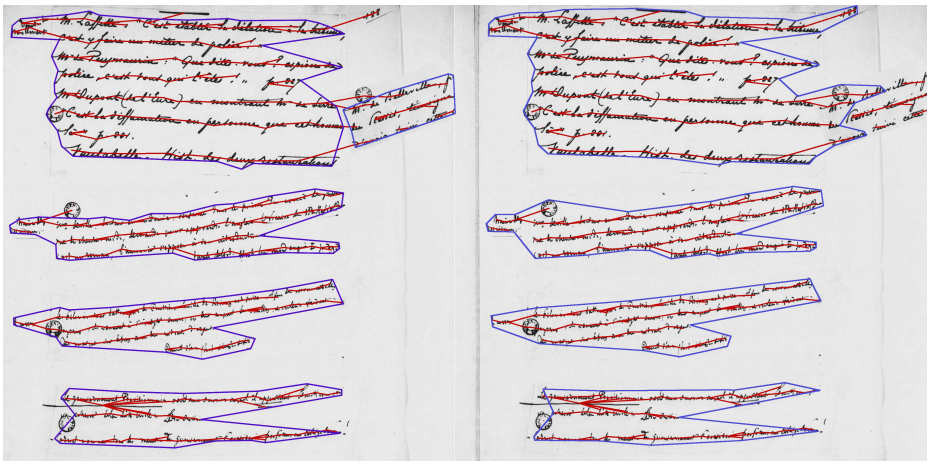

Fig. 5. Fragment extraction, 1f188r : Ground Truth (Left), Simulation (Right)

5 Concluding Remarks

In this paper, we proposed a dedicated text lines and fragments segmentation approach for author's draft handwritings. Knowing that fragments 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 fragments in drafts documents. Experiments of our approach show that our proposition is really consistent regarding the complexity of many page layouts in the corpus. Our edge to edge distance allow us to face some classic limitation of connected components based approach like the included connected component problem. Our methodology should be compared to conventional text lines segmentation methods, such as [6] or [5]. Due to the difference in segmentation goals, the comparison required some adaptations.

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.

References

1. B. Yu and A.K. Jain, "A robust and fast skew detection algorithm for generic documents," *PR*, vol. 29, no. 10, pp. 1599–1629, October 1996.
2. G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis, "Text line detection in handwritten documents," *PR*, vol. 41, no. 12, pp. 3758–3772, 2008.
3. L. Likforman-Sulem, A. Hanimyan, and C. Faure, "A hough based algorithm for extracting text lines in handwritten documents," in *ICDAR '95*, Washington, DC, USA, 1995, p. 774, IEEE Computer Society.
4. A. Lemaitre, B. Coüasnon, I. Leplumey, "Using a neighbourhood graph based on Vorono tessellation with DMOS, a generic method for structured document recognition" in *Graphics Recognition*, Volume LNCS 3926, Pages 267-278, 2006
5. L. Likforman Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *IJDAR*, vol. 9, no. 2-4, pp. 123–138, April 2007.
6. Y. Li, Y.F. Zheng, D. Doermann, and S. Jaeger, "Script-independent text line segmentation in freestyle handwritten documents," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 8, pp. 1313–1329, August 2008.
7. T.M Breuel, "High performance document layout analysis," *SDIUT '03*, 2003.
8. K. Kise, A. Sato, and M. Iwata, "Segmentation of page images using the area voronoi diagram," *CVIU*, vol. 70, no. 3, pp. 370–382, June 1998.
9. L. O'Gorman, "The document spectrum for page layout analysis," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1162–1173, 1993.
10. S. Nicolas, T. Paquet, and L. Heutte, "A markovian approach for handwritten document segmentation," in *ICPR06*, 2006, pp. III: 292–295.
11. S. Nicolas, T. Paquet, and L. Heutte, "Complex handwritten page segmentation using contextual models," in *DIAL06*, 2006, pp. 46–59.
12. K. Etemad, D. Doermann, and R. Chellappa, "Page segmentation using decision integration and wavelet packets," in *ICPR94*, 1994, pp. B:345–349.