2011 International Conference on Document Analysis and Recognition

A mixed approach for handwritten documents structural analysis

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

***Abstract*—In this paper we propose a new method for document pages segmentation. First dedicated to handwritten documents, our method is designed to extract the different text zones, paragraph and fragment in unconstraint documents.**

The proposed approach is a mixed one, using both the advantages of top-down and bottom-up approaches.

In this paper we proposed and evaluation of our methods on a 183 documents database, taken from a 19th century handwritten corpus : the « dossiers de Bouvard et Pécuchet » from Flaubert. With this evaluation we demonstrate that the combination of the top-down and the bottom-up approach allow to improve the obtained results.

***Keywords*-handwritten ; logical structure ; segmentation ; physical structure**

I. Introduction

Our work takes place in an humanities 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 fragments (Newspapers extracts, various notes, etc.) put together by Flaubert. To produce the electronic edition, structure informations must be known to reproduce the primary state of the corpus and restore fragments mobility.

In order to retrieve structure information, we propose an improvement of our previously published method [1]. We describe here a mixed approach using a top down approach to extract text zones clearly separated by white spaces and a bottom-up approach to extract more complex text zones. Our evaluation dataset is composed of 183 unconstraint handwritten pages from our corpus. The dataset is described in details in section V.

Definition of some concepts is required. Physical structure designates all structural elements corresponding to layout properties and script codification. In our study, documents are written in French : handwritten text lines can be gathered into structured sections to compose a page. Physical structure elements considered here are words, lines and paragraphs.

The notion of logical structure designates significant style and layout effects. The elements which can be included in this category are variable from one page to another. On our corpus, we can identify three categories of stable logical structure elements. The logical structure is then composed by underline words that designate relevant keywords and section headings. Margin words, that introduce text sections, and page title words. Basically, they can be reduced to to groups : keywords and associated fragments.

This paper is organized as follows : section 2 and 3 details previous works on text line extraction and structure recognition, section 4 details the proposed approach, section 5 provides evaluation results and in section 6 we give concluding remarks and perspectives.

II. State of the Art technics

For documents, layout analysis is an important pre-processing step for many applications such as OCR, appearance-based document retrieval, information retrieval or specific navigation applications. In this section we introduce technics dealing with problems close with ours. We separate lethods dealing with constraint documents and unconstraint ones.

A. Segmentation of constraint documents

On machine-printed documents, two different classes of methods can be found : top-down methods, generally based on white space analysis [2] and bottom-up methods, based on image patterns analysis.

In [3] T.M.Breuel summarize algorithms for machine-printed documents layout analysis. Presented algorithms are noise resistant and adapted to different layouts and languages. They perform text-line extraction, layout analysis and provide additional information such as reading order. Nevertheless, presented algorithms mainly based on connected components adjacency rules are only efficient on regular layouts.

Most of the remarkable works are based on neighbourhood information. Neighbourhood information are generally extracted from the connected component information [3] or on edge features and Voronoi tessellation [4],[5].

B. Segmentation of unconstraint documents

It is not trivial to extend machine printed documents algorithms to handwritten documents, especially when handwritten text lines are curvilinear and when neighbouring

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.62

handwritten text lines may be close or touch each other. In [6], L.O'Gorman's Docstrum, dedicated to machine printed pages can be applied to retrieve layout structure of handwritten pages with regular layouts.

S.Nicolas et al. in [7] propose an Hidden Markov Model based algorithm for unconstrained manuscript page segmentation : 6 classes are extracted from neighbourhood information, representing page background, textual components, erasures, diacritics, interline and interword spaces.

All of the presented works extract text zone using neighbourhood information only. Generally, handwritten documents are not regular and additional information are needed for layout extraction and document understanding : that is why we use to build a method based on a combination of structural and logical rules.

III. Preliminary results

To perform structure extraction we use results and algorithms previously described in [1].

To describe and caracterize a document page we first extract connected components using the algorithm proposed by F.Chang et al. dans [8]. This algorithms is based on contour tracking and performed well on our corpus. We build an adjacency graph of the connected components, using edge to edge distance.

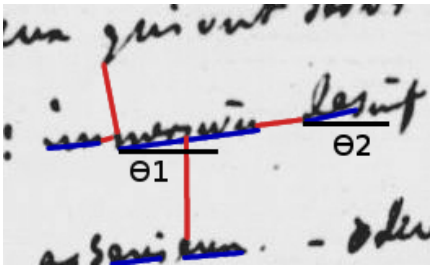

Figure 1. Connected components distance

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}) (figure 1).
Edge to edge distance is more robust than a classic Euclidean distance from centroid to centroid : the estimated distance is the true interline or the true interword distance.

To be consistent with orientation variation in handwritten documents we weight our measure with an orientation coefficient. This coefficient, based on the fact that orientation remains mainly constant in a line or a fragment is computed using orientation of the two connected components : $\theta 1$ and $\theta 2$. Thanks to this weight, two connected components with a significant $\Delta\Theta$ will have a higher distance than two components with the same d_{edges} and a $\Delta\Theta$ equal to 1.

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

The distance can be summarized by the following statement :

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

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 orientation ranges. Nearest neighbours are used to build the adjacency graph and perform lines and fragments extraction. Lines are extracted using the method described in [9]. The results obtained on our dataset for text lines extraction are provided in table I. We achieve line extraction with a sucess rate arround 88%, depending on the page categories. As we will see later, the extraction rate of 88% induce some limitation for a fragment extraction based on line fusion only.

Class	Title	Note	List	Collage	Print	Mixed	Total
Pages	3	64	9	104	8	7	195
Lines	8	1308	160	2225	x	97	3798
Recall	1	0.88	0.83	0.86	X	1	0.87
Prec.	0.89	0.87	0.91	0.877	X	1	0.88

Table I
LINE EXTRACTION EVALUATION RESULTS

IV. Our method

A. Bottom-up algorithm

We propose at first to use a bottom-up algorithm to extract fragments from the documents. The bottom-up approach is based on the results delivered by the adjacency graph, and particularely the results of text line extraction : we work on the line adjacency graph G', composed of n sub-graphs where each sub-graph stands for a line. Once line extraction is complete, lines are gathered in fragments using the set of rules described in the following part of this section.

Each node of the graph G' stand for line and has the following labels : height minimum position, height maximum position ,lenght, mean height, center of mass, orientation, space befor next line.

1) Physical structure extraction: To extract the physical structure with our bottom up algorithm we use the interline space variation information.

For each node, the space with the next node is computed as follows : we use a sample of points extracted from the line convex hull. We look for the closest point of the next line convex hull in the orthogonal direction to line orientation. Once distance value is extracted for each sample point, we computed the mean value, excluding extreme values ($d > m + 2 * \sigma$ and $d < m - 2 * \sigma$). The computation principle is

described on figure 2. It allows to consider the orientation variation between different fragments, which is an important marker for the extraction.

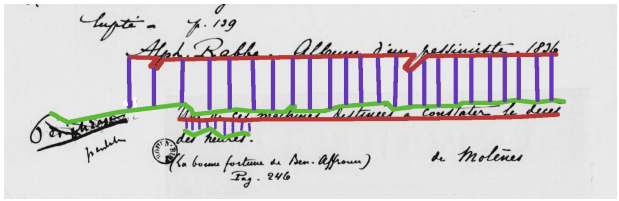

Figure 2. Interline space on a sample page

Once the distance is computed for each node, an interline-space histogramm (3 is computed and the threshold value for fragment extraction is extracted.

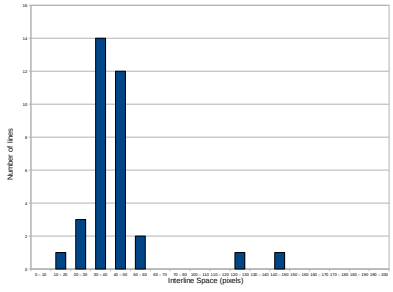

Figure 3. Interline space histogramm

Once this step is performed, we obtain an estimation of the page physical layout. We now have to extract logical structure to get the full layout.

2) Logical structure extraction: To extract some of the logical structure of the document page we use style and positionning informations given by the writer. 3 kinds of informations are extracted : margin words and associated text zones, underline words and title words. In this study we focus only on margin words which provide the most important informations for logical structure analysis. The figure 4 shows the expected result of text zone extraction on a complex note page.

The approach proposed here is designed using *a priori* knowledge on the documents. It can be adapted to any corpus were keywords and associated containers can be identified.

Margin word extraction: In our corpus, margin is the first marker of logical structure. To extract components located inside the margin we use the following properties : margin is a low density zone and is located at the left of the page. In practice we compute a projection of all black pixel of the binarized version of our image in the direction orthogonal to the mean orientation of the page. The same approach can be applied on fragments considering them as sub-images.

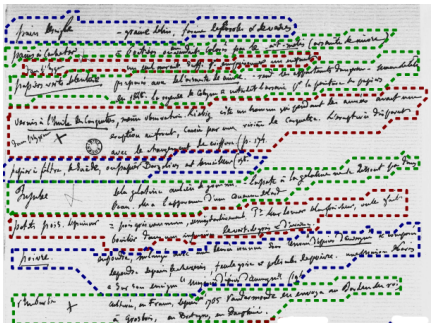

Figure 4. Text zone extraction results on a complex note page

If a lower density zone appear in the left third of the profile, each connected component located inside this zone is labelled as margin component.

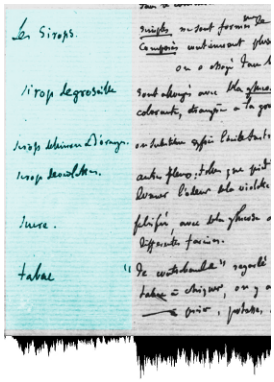

Figure 5. Note page and associated projection profile

Figure 5 illustrates the extraction result. Margin appears clearly thanks to the profile.

Associated text zones extraction: Once margin words are identified, associated text zone can be found.

The extraction is based on the text line extraction results. Lines are grouped into logical fragments depending on the presence of margin words.

For each line of the page, if any component of the line is labelled as margin word, the line is labelled as start line. The process is performed for each line. Line are then grouped into fragments using a simple procedure : the top line of the document start a fragment. If a start line is encountered, a new fragment is initiated and the previous fragment is closed.

B. Top-down algorithm

In addition to the bottom-up algorithm we choose to develop a top-down method knowing that some textual fragments can be extracted quickly and more efficienctly

by a top-down approach with no a-priori knowledge on the documents. Only the physical layout structure can be extracted this way.

Our top-down method is based on the result of the image distance transform. The distance transform is a representation of the image associating each pixel of the image to the nearest obstacle pixel. In our context, the obstacle are black pixels of the document image binarized version .

The figure 6 shows the result of distance transform applied to a natural image were obstacle points are contour points. On the distance image, different regions appear. The distance transform is a good candidate to build a quick and efficient method for image segmentation in general and document image segmentation in particular.

Figure 6. Distance transform result on Lena image

On a digitize handwritten document, presented figure 7, we note that interline spaces, interword spaces and inter paragraphs spaces appear with different gray levels values. To perform the desired segmentation a thresholding of distance transform values if performed. The threshold is determined experimentally, its value is set to 25 pixels for handwritten pages digitized in 200dpi.

Page segmentation results are post-processed in order to remove all detected fragments containing no connected components.

The thresholding result allow us to extract regions of the image. Extraction is quick, less than 1s for a standard 1952×2714 pixels image and allows to extract every zone clearly identified by white spaces.

C. Mixed approach

To improve the results of page layout extraction we choose to combine both approaches previously described. The top-down approach is quick and allows to extract paragraphs with a better efficiency than the bottom-up approach. The bottom-up approach allows to extract logical zones and finest physical text zones.

The top-down approach is run first and the bottom-up approach is run on each extracted zone. The summary of our approach is described on figure 8

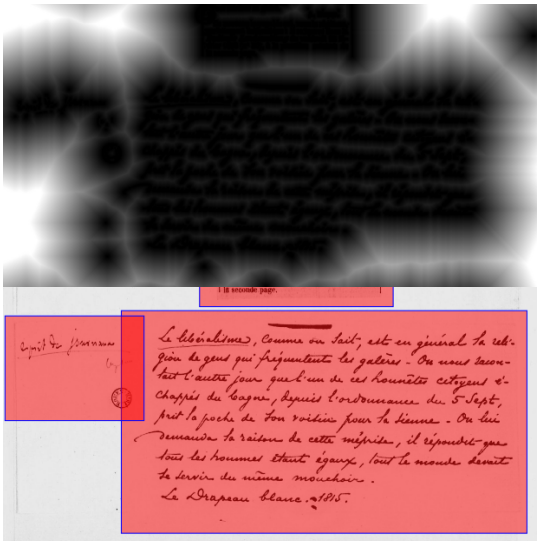

Figure 7. Distance transform and associated segmentation

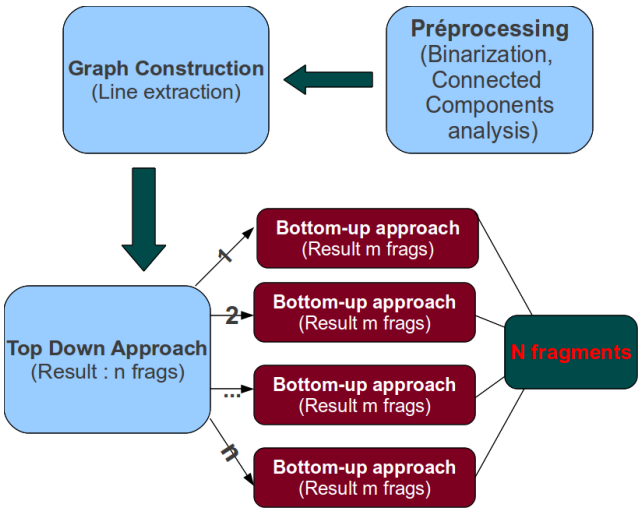

Figure 8. Synoptic of our mixed approach

V. Evaluation

A. Database

To evaluate our algorithms, we choose to use images from the « dossiers de Bouvard et Pécuchet ». We select a set of 195 pages, representing 130 documents. The selected pages are all taken from the same thematic volume. This point guarantee that all the variety of pages of the corpus are represented. The set is composed of 195 pages, with 3798 text lines and 550 fragments. Pages belong to 6 differents classes : Title pages, Note pages, Lists, Gathered information pages (Collages), Machine printed pages, Handwritten and Machine printed mixed pages. (Identification is automati-

cally performed using [9] method)

The table II summarize the content of the test database. Evaluation is performed on all pages and results will be provided for each page category.

Class	Title	Note	List	Collage	Print	Mixed	Total
Pages	3	64	9	104	8	7	195
Lines	8	1308	160	2225	x	97	3798
Frags	18	154	11	319	21	27	550

Table II
SUMMARY OF PAGE DATASET

B. Top-down algorithm

Table III shows the results of top down segmentation algorithms. It separates well the differents text zones, when they are clearly identified with withe space. However, on note pages, the fragments are not well extracted due to the necessity of a low level analysis of the page : most of the logical fragments are not extracted.

C. Bottom-up algorithm

Table IV shows the results of bottom-up segmentation approach. The algorithm as a low mean-recall value : when line extraction or margin extraction fails it leads to the fusion of two consecutive fragments. Lots of fragments are therefore not detected. Nevertheless, most of the undetected fragments are physical ones : logical fragments not detected using the top-down method are detected here.

D. Mixed approach

Table V shows the results of the mixed approach. We benefits of the advantages of both approaches : the down down approach allow to perform the segmentation on under-segmented text zones by the top-down method. We note that the improvement is mainly shown on the note pages : this is the category which contains the most important number of logical fragments.

Class	Title	Note	List	Collage	Print	Mixed	Total
Pages	3	64	9	104	8	7	195
Frags	18	154	11	319	21	27	550
Recall	0.94	0.71	1.00	0.86	1	0.96	0.83
Precision	1.00	0.72	0.91	0.82	1	1	0.82

Table III
EVALUATION OF TOP-DOWN SEGMENTATION METHOD

Class	Title	Note	List	Collage	Print	Mixed	Total
Pages	3	64	9	104	8	7	195
Frags	18	154	11	319	21	27	550
Recall	0.72	0.65	0.45	0.5	0.76	0.59	0.59
Precision	0.92	0.84	0.71	0.81	0.5	0.84	0.80

Table IV
EVALUATION OF BOTTOM-UP SEGMENTATION METHOD

Class	Title	Note	List	Collage	Print	Mixed	Total
Pages	3	64	9	104	8	7	195
Frags	18	154	11	319	21	27	550
Recall	0.88	0.84	1.00	0.92	0.95	0.96	0.90
Precision	0.84	0.79	0.92	0.84	0.8	1	0.83

Table V
EVALUATION OF MIXED APPROACH SEGMENTATION METHOD

VI. CONCLUSION AND PERSPECTIVES

In this paper we propose a new approach to perform handwritten page layout segmentation. Result are promising on complex handwritten pages. The approach can be applied to any document to extract the physical structure.

Improvement are still to be made for both parts of the method. Concerning bottom-up algorithm, we currently investigate new algorithms for graph extraction without connected components extraction in order to improve the line extraction rate. Modification of margin words extraction method is also planned : the current heuristic is to be replaced with a clustering approach.

Concerning top-down method, automatic or semi-automatic initialisation of threshold have to be implemented in order to improve the performance on the method on documents of different size and resolutions.

REFERENCES

[1] V. Malleron, V. Eglin, H. Emptoz, S. Dord-Crouslé, and P. Régnier, "Hierarchical decomposition of handwritten manuscripts layouts," in *CAIP*, S. LNCS, Ed., Sep. 2009, pp. 221–228.

[2] G. Nagy and S. Seth, *Hierarchical Representation of Optically Scanned Documents*. IEEE Computer Society, 1984, pp. 347–349.

[3] T. M. Breuel, "High performance document layout analysis," 2003.

[4] K. Kise, A. Sato, and M. Iwata, "Segmentation of page images using the area voronoi diagram," *Comput. Vis. Image Underst.*, vol. 70, no. 3, pp. 370–382, 1998.

[5] A. Lemaitre, B. Coüasnon, and I. Leplumey, "Using a neighbourhood graph based on voronoï tessellation with dmos, a generic method for structured document recognition," *Graphics Recognition*, vol. 3926, pp. 267–278, 2006.

[6] L. O'Gorman, "The document spectrum for page layout analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1162–1173, 1993.

[7] S. Nicolas, "A markovian approach for handwritten document segmentation," in *ICPR06*, 2006, pp. 292–295.

[8] F. Chang, C.-J. Chen, and C.-J. Lu, "A linear-time component-labeling algorithm using contour tracing technique," *Computer Vision and Image Understanging*, 2003.

[9] V. Malleron, V. Eglin, H. Emptoz, S. Dord-Crouslé, and P. Régnier, "Text lines and snippets extraction for 19th century handwriting documents layout analysis," in *ICDAR*, Jul. 2009.