A Document Image Segmentation System Using Analysis of Connected Components

Conference Paper · August 2013
DOI: 10.1109/ICDAR.2013.154

4 authors, including:

Abdel Ennaji
Université de Rouen
69 PUBLICATIONS **540** CITATIONS
SEE PROFILE

Stéphane Nicolas
Université de Rouen
47 PUBLICATIONS **263** CITATIONS
SEE PROFILE

D. Mammass
University Ibn Zohr - Agadir
137 PUBLICATIONS **335** CITATIONS
SEE PROFILE

A document image segmentation system using analysis of connected components

F.Zirari, A.Ennaji, S.Nicolas
LITIS Laboratory
University of Rouen
Rouen, France
zirari_fattah@yahoo.fr, {Abdel.Ennaji, stephane.nicolas}@univ-rouen.fr

D.Mammass
IRF-SIC Laboratory
Ibn Zohr University
Agadir, Morocco
driss_mammass@yahoo.fr

***Abstract*—Page segmentation into text and non-text elements is an essential preprocessing step before optical character recognition (OCR) operation. In case of poor segmentation, an OCR classification engine produces garbage characters due to the presence of non-text elements. This paper presents a method to separate the textual and non textual components in document images using a graph-based modeling and structural analysis. This is a fast and efficient method to separate adequately the graphical and the textual parts of a document. We have evaluated our method on two well-known subsets: the UW-III dataset and the ICDAR 2009 page segmentation competition dataset. Comparisons are led with two methods of state-of-the-art; these results showing that our method proved better performances in this task.**

Keywords—text/non-text separating; connected components; graph; structural analysis; document image.

I. INTRODUCTION

Segmentation is a crucial basic step in a document image processing and analysis workflow, because it actually precedes other operation of identification or classification. This step depends on the type of image that differs from both the acquisition system and the image formation process. In the case of document images, the problem is to classify the content of a document image into a set of text and non-text classes. The non-text class consists in the following categories: halftone, drawing, mathematical formulas, logos, tables, etc.

The document image segmentation techniques can be classified into the following groups: pixel based classification [1] [2], connected component based classification [3], block or zone based classification [4] [5] [10] and multiresolution morphology based segmentation [6].

Bukhari et al. [3] presented segmentation approach introducing connected component based classification, thereby not requiring a block segmentation formerly. Here they train a self-tunable multi-layer perceptron (MLP) classifier to differenciate text and non-text connected components using shape and context information as a feature vector.

The review paper by Okun et al. [5] succinctly sums up the main approaches used for document zone classification in the 1990s. The main feature type is based on connected components and run-length statistics. Wang et al. [10] presented the block classification system, each block with a 25 dimensional feature vector and use an optimized decision tree classifier to classify each block into one of different target classes. Keysers et al. [4] presented the most recent and detailed block classification method, a document block classification system can be constructed to use run-length histogram feature vector alone. Generally speaking, the approaches that classify blocks highly depend on the result of page segmentation into blocks. The blocks may be segmented in a wrong way which leads to a mis-classification.

Bloomberg [6] presented an approach to page segmentation based on multiresolution morphology. The multiresolution morphology-based method [6] was especially designed to separate halftones from document images. It works well for halftone segmentation, but it doesn't for other types of non-text elements like drawings, maps, etc. Furthermore, an open source implementation is available as part of the Leptonica library. It is a simple approach, based on the assumption that the size of non-text elements is larger than text elements in document images.

As part of this paper, we suggest a segmentation method based on a modeling of the document image by graphs and an applying structural layout rules. The main advantage of using graphs to represent images is the integration of spatial information in the model. Indeed classical representations provide no information on the way regions of interest of the image are organized. On the contrary, the representation made by the graph was to describe the structure of image as the way in which the areas are laid out in the forms of the one compared to the others. Our method is insensitive to low skew and adapted to the text / non-text segmentation.

This paper is organized as follows. In Section 2, we first describe the steps used by our approach. Then, the

1520-5363/13 $26.00 © 2013 IEEE
DOI 10.1109/ICDAR.2013.154

experimental results are given and discussed in Section 3. Finally, the last section is our conclusion.

II. SUGGESTED METHOD

The approach we suggest consists in modeling the document image by using a graph that will allow us to establish the close relations and connexity according to a homogeneity criterion based on the pixels intensity. This will be the first step to extract the connected components of the document image using a graph modelling approach. For the second step, the connected components thus formed will be categorized into graphical regions and text areas by applying structural layout rules.

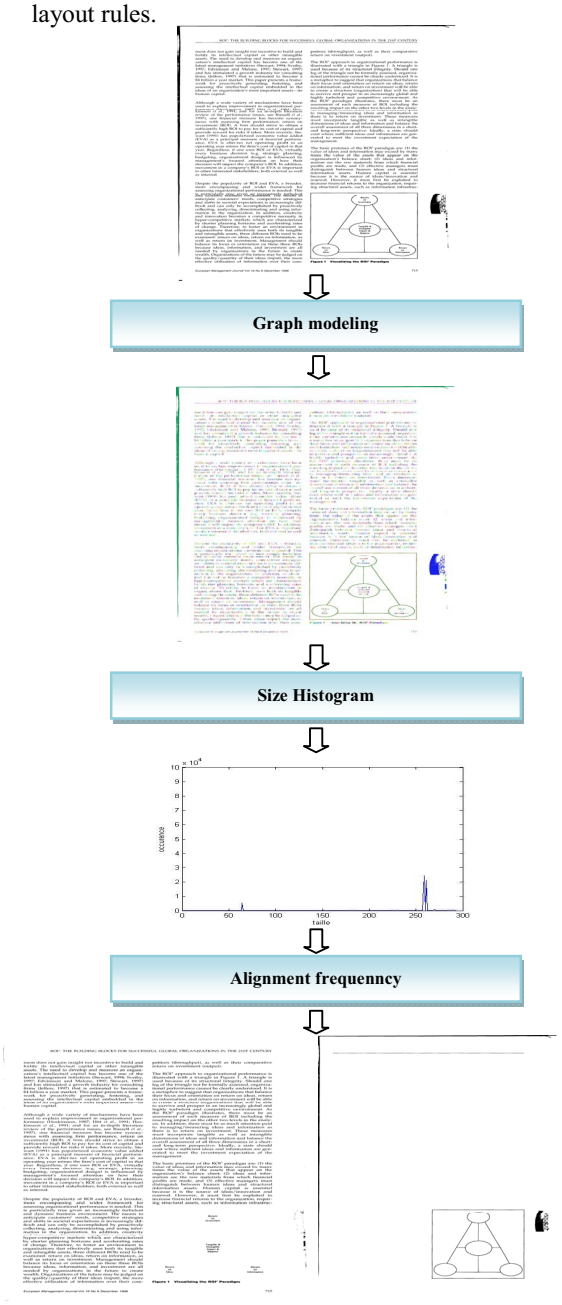

Fig. 1. Proposed method

A. Extraction of the connected components

In our approach we model the image by a non oriented Graph $G = (V,E)$ with vertices $v_i \quad V$, the set of elements to be segmented, and edges $(v_i, v_j) \in E$ corresponding to pairs of neighboring vertices. Each edge $(v_i, v_j) \in E$ has a corresponding weight $p(v_i, v_j)$. The nodes of the graph represent the pixels of the image and will be balanced by their intensity, whereas the edges represent the relationships of connexity and are balanced by the sum of the pixels' intensities in the end.

We seek to extract the connected components of images by combining the pixels according to the intensity value and then label these regions. To measure the homogeneity of intensity, we adopt the concept of internal difference of a component defined in [7] as follows:

The internal difference of a component $C \quad V$ is the largest weight in the Minimum Spanning Tree of the component, $MST(C,E)$. That is,

$$Int(C) = max_{e \in MST(C,E)}\, p(e) \qquad (1)$$

with $$p(e) = \frac{I(p_i)+I(p_j)}{2}\ ,\ e(v_i, v_j) \in E \qquad (2)$$

and $I(p_i)$ pixel intensity.

The justification is that, since the MST spans a region C through a set of edges of minimal cost, any other connected set of same cardinality will have at least one edge with weight more superior to that of *Int(C)*.

Initially, a graph is constructed from the entire document image, each pixel p being its own unique region {p}. Subsequently, the regions are merged by traversing the edges in a sorted order by increasing weight and evaluating whether the edge weight is smaller than the internal variation of both regions incident to the edge. If true, the regions are merged and the internal variation of the compound region is updated.

Now we explain the algorithm, presented in the next figure (figure 2), to find a partitioning of the image in homogeneous regions.

Algorithm:

The input is a graph $G = (V, E)$, with n vertices and m edges. This graph is formed according to the rules of 8-connected neighborhood classically used to model images. The output is a segmentation of V into components $S = (C_1, ..., C_r)$. The proposed algorithm is iterative:

1- Sort E into $\Pi = (o_1, ..., o_m)$, with $o_q = (v_i, v_j)$, by non-decreasing edge weight.

2- Start with a segmentation S^0, where each vertex v_i corresponds to exactly one unique component.

3- Construct S^q given S^{q-1} as follows:

Let v_i and v_j denote the vertices connected by the q^{th} edge in the order list P, i.e., $o_q=(v_i, v_j)$. If v_i and v_j are in disjoint components of $S^{q-1}$and $p(o_q)$ is small compared to the internal mean of both components, then merge the two components otherwise do nothing. More formally, let C_i^{q-1}be the component of S^{q-1}containing v_i and C_j^{q-1} the component containing v_j. If $C_i^{q-1} \neq C_j^{q-1}$ and $p(o_q) \leq MInt(C_i^{q-1}, C_j^{q-1})$ with $MInt(C_i^{q-1}, C_j^{q-1}) = \min(Int(C_i^{q-1}), Int(C_j^{q-1}))$, then S^q is obtained from S^{q-1} by merging C_i^{q-1} and C_j^{q-1}. Otherwise $S^q = S^{q-1}$.

4- Repeat the step 3 for $q = 1, ..., m$.

5- Return $S = S^m$.

Fig. 2. Algorithm of connected components extraction

At the end of the algorithm we get a set of connected components (Figure 3) we have to classify them as graphical or textual elements. We describe this classification process in the next section.

B. Classification of the connected components

The aim is to label the components resulting from the segmentation obtained at the previous step, in two classes: "text" (or textual components) or "non-text" class (graphics, tables, lines ...). To do so, we sought to exploit the fact that the text zones are often characterized by an alignment of characters of very similar size. Thus, we developed a simple approach to identify text areas based on the filtering of the components provided by our first stage, based on a size criterion and then the overlapping between components is analyzed. Thus, to detect the textual components we apply the following two steps:

The first step consists in calculating the histogram of the frequencies of the component sizes. Only the components belonging to the most significant peaks of this histogram are retained (figure 4). To do so, we use a detection threshold set empirically up to now. This threshold can also be determined by a machine learning procedure. The idea of this first step is to filter the majority of the non textual components.

The second phase consists in removing the frequent noise and graphical components that were not filtered in the preceding step. To do so, we use the notion of component alignment. This alignment is determined by the vertical overlap among the components according to a given threshold, allowing a certain inclination.

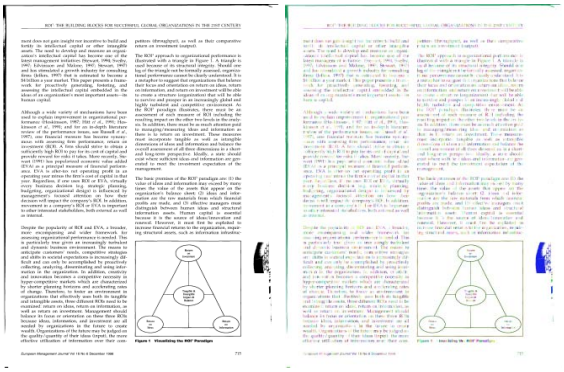

Fig. 3. Original document; segmented document.

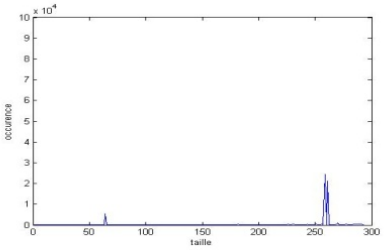

Fig. 4. Frequency histogram of the components size of the document result in Figure 3.

III. EXPERIMENT AND RESULTS

We have evaluated our approach by using the Washington University III (UWIII) dataset [8] and the ICDAR-2009 page segmentation competition dataset [9]. The main reason for using different datasets is to check the accuracy of our approach on different types of images which have not been used in training as well as to have a variety of text and non-text components. For example, majority of the document images in UW-III dataset have Manhattan-layout but ICDAR 2009 dataset also contains documents with non-Manhattan layout.

For each dataset, the pixel-level ground truth has been generated by using zone-level ground truth information. Each pixel in ground-truth images contains either text or non-text label. Different types of metrics have been used for the performance evaluation of document image segmentation method which are defined below, we used the same metrics as defined in [3] and we proposed a new metric (global accuracy) in order to compare our result with their results:

Non-text classified as non-text: percentage of intersection of non-text pixels in both segmented and ground truth image respecting the total number of non-text pixels in ground truth image.

Non-text classified as text: percentage of intersection of text pixels in segmented image and non-text pixels

in ground truth image respecting the total number of non-text pixels in ground truth image.

Text classified as text: percentage of intersection of text pixels in both segmented and ground truth image respecting the total number of text pixels in ground truth image.

Text classified as non-text: percentage of intersection of non-text pixels in segmented image and text pixels in ground truth image respecting the total number of text pixels in ground truth image.

Segmentation accuracy: average percentage of text classified as text accuracy and non-text classified as non-text accuracy.

Global accuracy:

$$Acc = 100\% - errors \quad (3)$$

errors = (non-text classified as text + text classified as non-text)%

Based on the matrices defined above, we have compared our approach with leptonica's page segmentation algorithm [6] and Bukhari approach [3]. The performance evaluation results of our method with leptonica and Bukhari methods on UW-III and ICDAR 2009 test datasets which contains only text and non-text components are shown in Tables 1.

It is obvious from the results that our method shows better text classification accuracy than the one of non-text classification and Bukhari method has better non-text classification accuracy than the one of text classification. Leptonica method misclassifies the small non-text components as the text components and the Bukhari method misclassifies some text components as the non-text components. Nevertheless, our method gives equal importance to both the text and non-text components. Unlike leptonica method, our method can also classify the small non-text and text components. The segmentation accuracy of our method is better compared to leptonica and Bukhari method. The segmentation results of our method can be improved nevertheless by increasing training samples and/or by using some post-processing operations.

The figures 5 and 6 illustrate the results obtained by our method on a sample of 2 documents chosen in the base of the documents treated so as to illustrate the capacities and the limits of our approach. These figures show in the order the original document image and the 2 images corresponding to the textual zones and the non-textual zones identified by our approach.

ACKNOWLEDGMENT

We would like to acknowledge the financial support of our project by the France-Morocco-Hubert Curien Program-PHC-Volubilis n° MA/10/233.

IV. CONCLUSION

We have presented a method of document image segmentation to identify the textual and the non textual zones in the form of either graphics or any other type of illustrations. This method is based on modeling of the various blocks document image by a graph approach. The blocks start from the step of modeling and then classified by a simple method which exploits the concept of alignment of the forms. Extensions of this approach to segment text blocks into words are underway for the development of a system in which the document is indexed by the content. Additional validations on more complex documents and/or degraded historical document are also underway. The exploitation of this whole information is considered thereafter for the treatment of the non textual zones.

ACKNOWLEDGMENT

We would like to acknowledge the financial support of our project by the France-Morocco-Hubert Curien Program-PHC-Volubilis n° MA/10/233.

V. CONCLUSION

We have presented a method of document image segmentation to identify the textual and the non textual zones in the form of either graphics or any other type of illustrations. This method is based on modeling of the various blocks document image by a graph approach. The blocks start from the step of modeling and then classified by a simple method which exploits the concept of alignment of the forms. Extensions of this approach to segment text blocks into words are underway for the development of a system in which the document is indexed by the content. Additional validations on more complex documents and/or degraded historical document are also underway. The exploitation of this whole information is considered thereafter for the treatment of the non textual zones.

TABLE I. PERFORMANCE EVALUATION OF OUR METHOD COMPARED TO THE LEPTONICA PAGE SEGMENTATION AND BUKHARI ALGORITHMS ON UW-III DATASET AND ICDAR 2009 PAGE SEGMENTATION COMPETITION TEST DATASET.

	UW-III			ICDAR-2009		
	Our approach	**leptonica**	**Bukhari approach**	**Our approach**	**leptonica**	**Bukhari approach**
non-text classified as non-text	97.77%	95.36%	98.91%	95.43%	84.91%	96.70%
non-text classified as text	2.23%	4.64%	1.09%	4.57%	15.09%	3.30%
text classified as text	99.81%	99.79%	95.93%	99.92%	99.87%	93.31%
text classified as non-text	0.19%	0.21%	4.07%	0.08%	0.13%	6.69%
segmentation accuracy	98.79%	97.57%	97.42%	97.67%	92.39%	95.01%
global accuracy	97.58%	95.15%	94.84%	95.35%	84.78%	90.01%

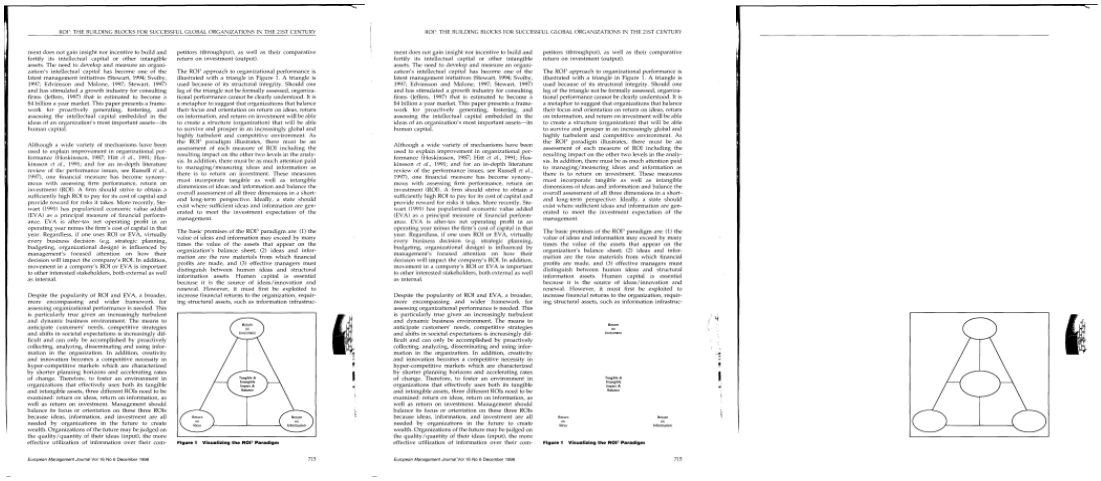

Fig. 5. Extraction of the text components integrated in graphical blocks.

Fig. 6. Example of good results obtained by our method.

REFERENCES

[1] M. A. Moll and H. S. Baird. "Segmentation-based retrieval of document images from diverse collections". In Document Recognition and Retrieval XV, Proc. Of the SPIE, volume 6815, pp. 68150L-68150L, 2008.

[2] M. A. Moll, H. S. Baird, and C. An. "Truthing for pixel-accurate segmentation". In Document Analysis Systems, the Eighth IAPR Int. Workshop, pp. 379-385, Sep. 2008.

[3] Bukhari, S. S., Shafait, F., and Breuel, T. M., "Document image segmentation using discriminative learning over connected components", in Proc. 9th IAPR Workshop on Document Analysis Systems, pp. 183-190, 2010.

[4] Keysers, D., Shafait, F., and Breuel, T. M., "Document image zone classification- a simple high-performance approach", in Proc. 2nd Int. Conf. Computer Vision Theory and Applications, pp.44-51, Mar. 2007.

[5] Okun, O., Doermann, D., and Pietikainen, M. "Page Segmentation and Zone Classification: The State of the Art". Technical Report LAMP-TR-036, CAR-TR927, CS-TR-4079, University of Maryland, College Park, 1999.

[6] Bloomberg, D. S., "Multiresolution morphological approach to document image analysis", in Proc. Int. Conf. Documnet Analysis and Recognition (ICDAR), pp. 963-971, 1991.

[7] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. "Efficient Graph-Based Image Segmentation". International Journal of Computer Vision, Volume 59: pp. 167–181, Number 2, September 2004.

[8] Guyon, I., Haralick, R. M., Hull, J. J., and Phillips, I. T. "Data sets for OCR and document image understanding research". In Bunke, H. and Wang, P., editors, "Handbook of character recognition and document image analysis", pp. 779–799. World Scientific, Singapore, 1997.

[9] A. Antonacopoulos, D. Bridson, C. Papadopoulos and S. Pletschacher "Performance Analysis Framework for Layout Analysis Methods", Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR), Catalonia, Spain, pp. 296–300, September 2009.

[10] Wang, Y., Phillips, I. T., and Haralick, R. M. "Document zone content classification and its performance evaluation". Pattern Recognition, 39 : pp. 57–73, 2006.