A Graph Based Approach for Heterogeneous Document Segmentation

Fattah Zirari[1,2], Driss Mammass[1], Abdellatif Ennaji[2], and Stephane Nicolas[2]

[1] Laboratory IRF-SIC Agadir Maroc
[2]Laboratory LITIS Rouen France
{zirari_fattah,driss_mammass}@yahoo.fr,
{Abdel.Ennaji,stephane.nicolas}@univ-rouen.fr

Abstract. In the field of document image processing, the text/graphic separation is a major step that conditions the performance of the recognition and indexing systems. That involves identifying and separating the graphical and textual components of a document image. In this context, it is important to implement approaches that effectively address these problems. This paper presents a method for separating 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 areas of a document. Some examples obtained on technical documents and magazines issued from the databases approved by the community make it possible to validate the approach.

Keywords: Segmentation, text/no text Separation, Document Image, Graph, modelization, structural analysis.

1 Introduction

Segmentation is a crucial basic step in a document image processing and analysis workflow, because in fact it precedes any 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, this consists in locating and possibly identifying the elements constituting the document at different granularity levels. Thus, a first segmentation task may consist in locating and identifying the text areas and the areas of different natures. If we consider for example the document images presented in Figure 1, which are pages from technical documents and magazines, the first segmentation task may consist in differentiating the areas containing text from the areas representing tables, curves or graphs.

In the literature, three families of approaches are possible for document image segmentation: the bottom-up, the top-down and hybrid approaches.

In top-down techniques, document images are recursively divided to smaller regions. These techniques are often fast, but the efficiency depends on a prior knowledge about the class of documents to be processed. Among the developments

A. Elmoataz et al. (Eds.): ICISP 2012, LNCS 7340, pp. 424–431, 2012.

produced in early times, the most well known methods are projection methods [1] and space transforms [2] (Fourier transform, Hough transform, ect).

Though these top-down methods generally perform well, they have a major drawback which is the need to have a prior knowledge about the document class and layout (number of columns, width of margins, etc) for them to be effective.

Bottom-up methods start with the thinnest elements (pixels), merging them recursively in connected components or regions, and then in larger structures. They are more flexible but may suffer from accumulation of errors. They make use of methods like connected components analysis [3], [4], region growing methods, run-length smoothing (RLSA) [5], neural networks [6] and active contours [7].

The advantage of bottom-up methods is that they are very flexible. Another thing is that these methods make use of a lot of parameters that need to be adjusted precisely for good results.

Many other methods that do not fit into either of these categories are therefore called hybrid methods that combine and make use of both bottom-up and top-down approaches. For example, connected components analysis for shape information and block separation for background block map have been used in [8] in a hybrid segmentation approach. Classification of these blocks is achieved according to the scenarios defined by the user.

As part of this paper, we propose a segmentation method based on a modeling of the image document by graphs and 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 how the regions of interest of the image are organized. On the contrary, the representation by the graphs makes to describe the structure of image as the way in which the areas are laid out 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. Some experimental results are given and discussed in Section 3 and the last section concludes the paper.

Fig. 1. Examples of structured documents containing textual and graphical information

2 Proposed Method

The approach we propose consists in modeling the document image by a graph that will enable us to establish the neighborly relations and connexity according to a homogeneity criterion based on the pixels intensity. This will be a first step to extract the connected components of the document image using a graph modelling approach. In a second step, the connected components thus formed will be categorized into graphical regions and text areas by applying structural layout rules. First introduce the graph formalism that we have adopted in our system.

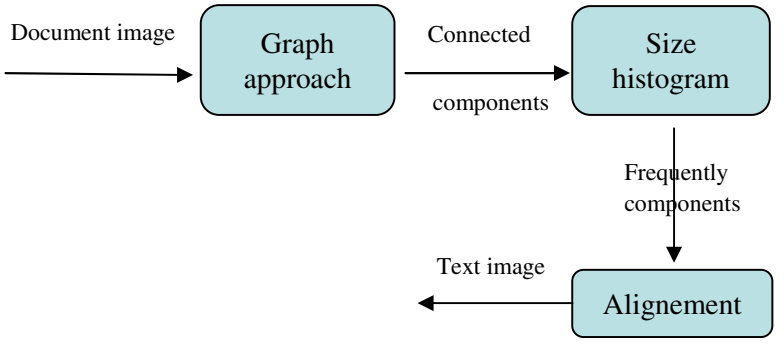


Fig. 2. Proposed Method

2.1 Used Formalism

The graphs constitute a mode of representation very frequently used in image processing and pattern recognition. They indeed make it possible to describe naturally in a unified formalism some objects and the relations between these objects.

A graph G consists of a set of nodes, denoted by V, linked by a set of edges, denoted by E:

$$G=(V,E)$$
$$V=\{v_1,v_2,\ldots,v_n\}$$
$$E=\{(v_i,v_j)\backslash v_i \in E,\ v_j \in E\}$$

Finally let us give for memory some definitions that we will need later.

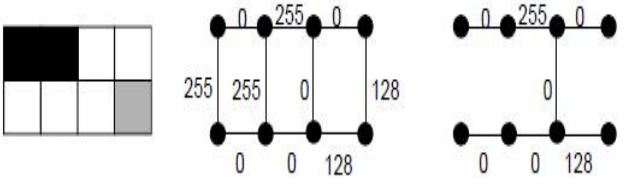
A connected graph is a graph where for any two nodes i and j we can find a walk which begins at i and ends at j.

An undirected graph is one in which the edges have no orientation. The edge (i, j) is identical to the edge (j, i), i.e., they are not ordered pairs.

A tree is a connected graph without cycles which connects a subset of all nodes. A spanning tree is a tree which connects all nodes.

A minimum spanning tree (MST) of an edge-weighted graph is a spanning tree whose weight (the sum of the weights of its edges) is not larger than the weight of any other spanning tree.

All these concepts are illustrated on the example of a simple graph modeling the image given in Figure 3.



a) b) c)

Fig. 3. a) initial image; b) associated graph ; c) Minimum spanning tree

In our approach we model the image by undirected related graph. 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 intensities of the pixels at the ends.

We seek to combine the pixels to form homogeneous regions and then label them. To measure the homogeneity of intensity we adopt the concept of internal difference of a component defined in [9] as follows:

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

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

with $$P(e) = (I(P_i) + I(p_j))/2 \ , \quad e = (v_i, v_j) \in E \tag{2}$$

and $I(P_i)$ pixel intensity.

The motivating argument 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 superior to Int(C).

Initially, a graph is constructed over the entire image, with each pixel p being its own unique region {p}. Subsequently, 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 that we have defined the formalism we will explain the algorithm to find a partitioning of the image in homogeneous regions related. This algorithm is:

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 $S = (C_1;\ldots; C_r)$.. The proposed algorithm is iterative:

1- Sort E into $\Pi = (o_1,\ldots, o_m)$, with $oq = (v_i; v_j)$, 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 $q = 1,\ldots, m$.

5- Return $S = S^m$.

At the end of the algorithm we obtain a set of homogeneous regions (Figure 4) we have to label as graphical or textual elements. We describe this labeling process in the next section.

2.2 Labelling of the Segmented 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 ...). For that, 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:

- A first step consists in calculating the histogram of the frequencies of the components size. Only the components belonging to the most significant peaks of this histogram are retained (figure 5). For that 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 eliminating the frequent noise components and graphics that were not filtered at the preceding step. For that, we use the notion of alignment components between them. This alignment is determined by the vertical overlap between the components according to a given threshold, allowing a certain inclination.

This first preliminary approach for the identification of text zones showed good performance despite its simplicity. It should nevertheless be completed in order to answer some weaknesses as shown in the results section.

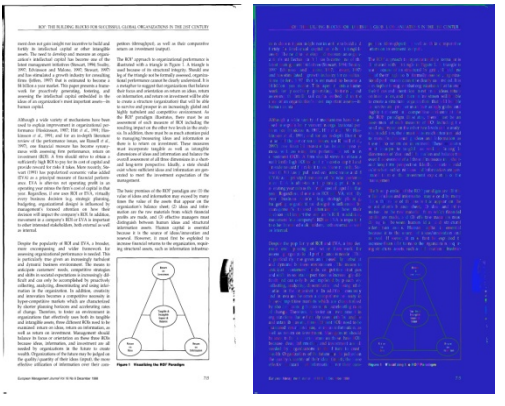


Fig. 4. original document ; segmented document

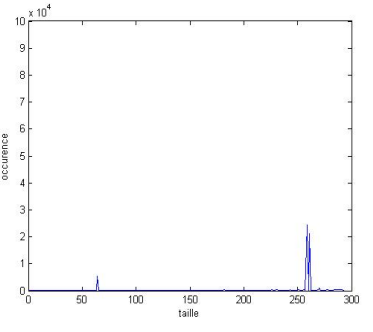


Fig. 5. Frequency histogram of the components size of the document result in Figure 3

3 Experiment and Results

To validate the effectiveness of our approach, we conducted tests on two databases of technical documents and magazines. These are documents that contain text, tables, charts, graphics, or inserts. These documents issued from the databases Prima [10] and de Washington University III (UWIII) [11]. These two databases are available with ground truth information. Thus, to evaluate our approach, we conducted a comparison between the pixel to pixel documents results and documents ground truth.

In the detection of textual components, we obtained a detection percentage of 96,7%. This rate of correct detection increases to 98,5% if one takes into account the textual components that are integrated with graphic blocks (Figure 6). Indeed, the ground truth provided for database for Prima does not consider this textual information as such but combines with graphic blocks in which they are understood. Similarly, we obtained a percentage of detection of 97% for non-textual components.

The error rate of 3% is due to the presence of textures in some graphic components that have similarities with textual components (Figure 7).

The figures 8 and 9 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, the image truth ground, and the 2 images corresponding to the textual zones and the graphic zones identified by our approach. In the truth ground the graphic zones are represented in red and the textual zones in blue.

These results illustrate the good performance of our approach.

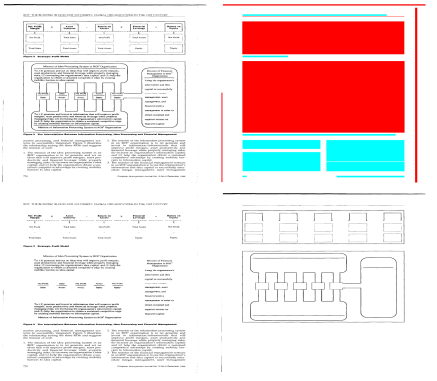


Fig. 6. Example of graphic elements with text included in the diagrams well detected by our method

Fig. 7. Example of detection error of the graphic parts produced by our method

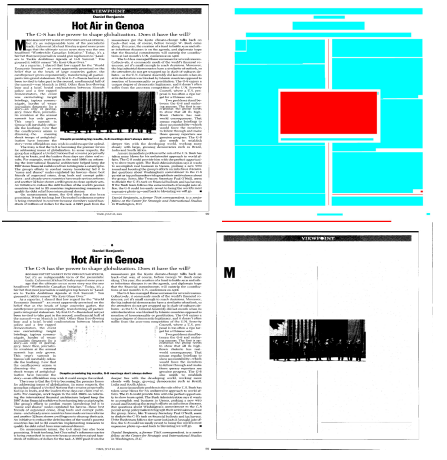


Fig. 8. Example of good results obtained by our method

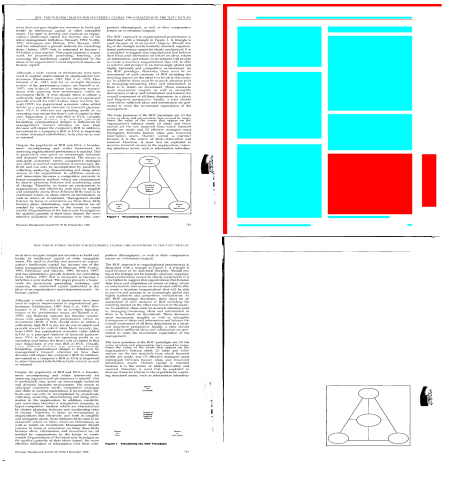


Fig. 9. Example of good results obtained by our method

4 Conclusion

We have presented a method for document image segmentation to identify the textual and the non textual zones being able to be either graphics or any other type of illustrations. This method is based on modeling of the various blocks of the document image by a graph approach. The blocks resulting from this step of modeling are then classified by a simple method which exploits the concept of alignment of the forms. Extensions of this approach for segmenting text blocks into words are underway for the development of a system for document indexing by content. Additional validations on more complex documents and/or degraded are also envisaged. The exploitation of the whole of this 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 "action intégrée Maroc-francaise" n° MA/10/233 and the AIDA project, program Euro Mediterranean 3+3 : n° M/09/05.

References

1. Antonacopoulos, A., Karatzas, D.: Semantics based content extraction in typewritten historical documents. In: 8th International Conference on Document Analysis and Recognition, pp. 48–53 (2005)
2. Jain, A.K.: Fundamentals of digital image processing. Prentice Hall (1989)
3. Mitchell, P.E., Yan, H.: Newspaper document analysis featuring connected line segmentation. In: Sixth International Conference on Document Analysis and Recognition, pp. 1181–1185 (2001)
4. Faure, C., Vincent, N.: Simultaneous detection of vertical and horizontal text lines based on perceptual organization. In: 16th Document Recognition and Retrieval Conference, DRR 2009, USA (2009)
5. Wong, K.Y., Casey, R.G., Wahi, F.M.: Document analysis system. IBM Journal of Research Development 26, 647–656 (1982)
6. Caponetti, L., Castiello, C., Gorecki, P.: Document page segmentation using neurofuzzy approach. Applied Soft Computing (2007) (in press, corrected proof)
7. Bukhari, S.S., Shafait, F., Breuel, T.M.: Segmentation of curled textlines using active contours. In: The Eighth IAPR Workshop on Document Analysis Systems (2008)
8. Ramel, J., Leriche, S.: Segmentation et analyse interactive de documents anciens imprimes. In: Traitement du Signal (TS), pp. 209–222 (2005)
9. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-Based Image Segmentation. International Journal of Computer Vision 59(2), 167–181 (2004)
10. Antonacopoulos, A., Bridson, D., Papadopoulos, C., Pletschacher, S.: Performance Analysis Framework for Layout Analysis Methods. In: Proceedings of The 10th International Conference on Document Analysis and Recognition (ICDAR 2009), Catalonia, Spain, pp. 296–300 (September 2009)
11. Guyon, I., Haralick, R.M., Hull, J.J., Phillips, I.T.: Data sets for OCR and document image understanding research. In: Bunke, H., Wang, P. (eds.) Handbook of Character Recognition and Document Image Analysis, pp. 779–799. World Scientific, Singapore (1997)