ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

New Approach of Document Pattern Recognition Using Bi-Cubic Interpolation

[I]M.BOUTAOUNTE, [II]A.ELBALAOUI, [III]Y.OUADID, [IV]A.MERBOUHA
[I,II,III,IV]Information Processing and Telecommunication Teams, Faculty of Science and Technology, Sultan Moulay Slimane University, Beni-Mellal, Morocco

Abstract
The purpose of document layout analysis is to segment the different parts of the image document in order to locate the text and the non-text components, for this aim we proposed a new method based in resizing the image using the interpolation of bi-cubic to yield a value to the new pixels calculated from the neighbors to performing a smoothing to the image document in order to connect the various parts. The following steps involve to groups the components using as parameters the distance and the size to take the decision for merge the components or not.

Keywords
Pattern recognition, document layout, bi-cubic interpolation, smoothing image

I. Introduction

Document layout is the way in which the different documents components are arranged in the image of document. These components can be text divided to two type title and the body text, or non text components (pictures, graphic…) .the document layout analysis involves operations that divide a document into text composed by grouping characters and the non text components to determine the physical structure of a document image. It includes page segmentation and zone classification, so the goal of page segmentation is to segment a document image into homogeneous regions

II. Background

The various layout analysis methods proposed thus far can be categorized as top-down, bottom-up [1], or hybrid approaches. In the top-down methods it start with the highest level i.e. the page down a level to another until it reaches the level of connected components or pixels level. An example of algorithm using top-down strategy is the famous XY cut algorithm [2] that relies on projection profiles to cut a textual region into sub-regions; however, it may fail in text regions that lack a fully extended horizontal or vertical cut. The assumption is based on the fact that the structured elements of the page are usually presented in rectangular blocks. But also out that the blocks may be divided into groups so that the blocks which are adjacent to each other, in a group, have a size in common.

The bottom-up approaches start with lowest level and move up one level to another to complete the page. In fact they are based on the analysis of connected components. The latter are obtained by scanning an image pixel by pixel and combining the pixels of the components based on the connectivity of pixels that can be 4 or 8 neighbors.

The principle of the bottom-up methods is the following: they start by the lowest level merge, forming the words from the connected components, and then back to a higher level by combining words into lines, lines into blocks, etc ... until the page is completely restored. Examples are the document spectrum method, the minimal-cost spanning tree method, and the component-based algorithm [3].

We also have some work that combine between bottom-up and top-down methods and add some correcting processing to ameliorate the result such as, L.CINQUE [4] have proposed a method where they are using multi-resolution approach that give the essential information for structure analyses but this method are destine for document of simple structure and have low result in decomposition the documents that have text in form of colons close to each other

Jic XI [5] have proposed another method based on bottom-up analysis using the horizontal project –profile of document image to estimate the size of body text in order to determine the threshold used in RLSA[5] the problem that RLSA algorithm need to run through the image pixel by pixels also this method need extra treatment to combine the components resultant of RLSA other work from Fu CHING [6] have proposed another approach specially for Chinese document consist in although component from textlines with their nearest neighbors, this approach have the same problem of RLSA that need to run through the image pixels by pixels and verify the neighbors .

The contribution of Chung-Chin WU [7] using the graphical object of image and for the remaining areas the use contour tracing technique to find label components in the documents so to form the textlines the use this components .The neuro-fuzzy method has been proposed by L.COPOLIN [8] this method has giving good result but the problem are in training and using the neuro-fuzzy for document of different size and type. So in order to resolve the majority of those problems, we have looked for a method that will be effective and quickly , and the most important thing that it will be independent of the language of the documents.

III. Proposed method

Our method is based on bottom-up analysis. Our aim is to create homogeneous blocks by applying a smoothing on the binary document image. in order to achieve this goal we apply a method of image resizing based on bi-cubic interpolation algorithm which will generated for each pixel a value calculated from neighboring pixels so the method is to reduce the size of the image to removes spaces between characters, words and even lines then reset the image, to resize the image without exceeding the volume of components, we followed the steps below, we will start by describe the Bi-cubic interpolation before explain our approach:

A. Bi-cubic interpolation

Bi-cubic interpolation [9] beyond the bilinear method, taking into account the 16 pixels closest to the point to be interpolated. The idea is to fit a polynomial model of the 16 gray levels of the source image, and then deduce the level of the interpolated point value calculation taken by model represented in equation 1.
The model has the following form:

$$M_g(x,y) = \sum_{p=0}^{p=3}\sum_{q=0}^{q=3} \propto_{pq} x^p y^q \qquad \text{eqn (1)}$$

The neighborhood is centered around the point A (same technique of positioning to rounded value coordinates A (x, y))(Fig.1).

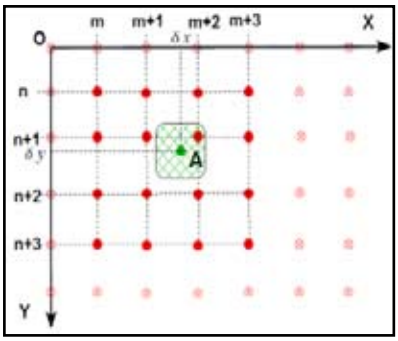

Fig. 1 : The 16 neighborhood of point A

We see that the bi-cubic model has 16 coefficients; neighborhood adjustment from 16 control points, the adjustment is correct.
So just write 16 equations and solving them (eg Cramer). The resolution gives the expression of α_{pq} coefficients according to grayscale image source. The value taken by the point A will be calculate using the equation 2.

$$NG(A) = \sum_{p=0}^{p=3}\sum_{q=0}^{q=3} \propto_{pq}\delta x^p \delta y^q \qquad \text{eqn (2)}$$

At the end we fixed a threshold to conserve it as a binary image so after calculate the value of the new pixel are changed to 0 or 1 according to the sill.

B. Connect horizontal components

In this stage we apply the Bi-cubic method to resize the image as follow:
The height of document image is reducing by divided on a constant M the result are an image with height equals to :

$$H = Hi/M \qquad \text{eqn (3)}$$

So the white space between character, words and also line generally, all the horizontal space will reduce or eliminate in the same time we expands the width to:

$$W = W_i * N \qquad \text{eqn (4)}$$

Increase the vertical space (we can also keep it without expands but that can cause some problems if they are noisy in the picture that may regroup two horizontal blocks) and keep the horizontal component separated
We will take this example of Arabic journal illustrate in Fig. 2:

Fig. 2 : Document example

Applying the first phase we have the following result in Fig. 3.

Fig. 3 : Result of reducing the width

After this we restore the image to the real dimension the result is a smoothing image in which we conserve only the vertical space between the components as illustrate in the following Fig. 4.

Fig. 4: Restore the image to the original dimension

The components of the image are now connected but the most of vertical component are logged so to correct this we will use the same process yet this time vertically

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

C. Connect vertical components

In this step the same process is apply, yet this time the width are reduced:

$$W=Wi/M \qquad \text{eqn (5)}$$

And the height will increased according to the following equation (Equation 6)

$$H=Hi*N \qquad \text{eqn (6)}$$

The results of this procedure are in the following images (Fig. 5):

Fig. 5: Reducing the height of the image

Restoring the image to the original dimension the results in Fig. 6 represent the horizontally smoothing of image.

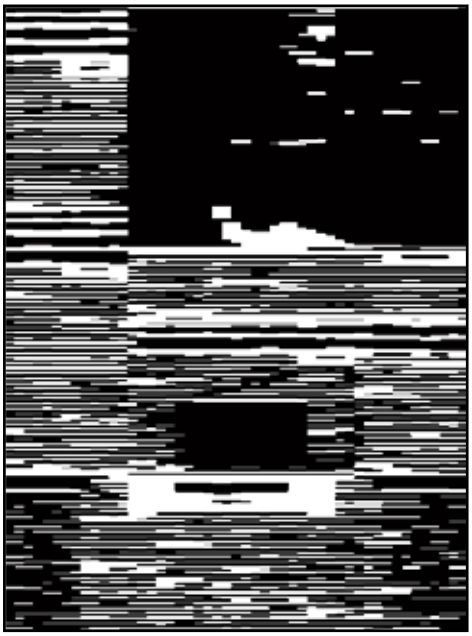

Fig. 6: Restore the image to the original dimension

As already been said vertical gaps are almost removed, against the horizontal still exist. So the next step is to use the first (vertical smoothing) and the second (horizontal smoothing) result

D. The final step

The two images resulting from recent processes (vertical and horizontal smoothing) will be used in this section to extract the layout of the document image for this purpose the two images will amalgamate of wielding to reconstruct an image which regroups the components respecting the horizontal and vertical spaces at the same time the result in Fig. 7.

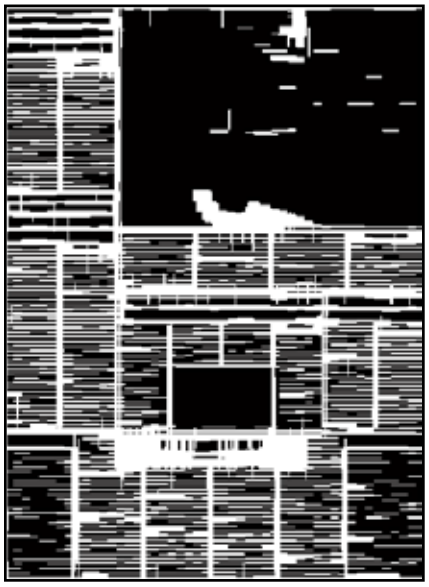

Fig. 7 : Result of combining the two of smoothing image resultant

Now we will introduce a new concept the components, a component [10] is a range of black pixels surrounded by white pixels
In our case the black pixels are represented by 0 and white pixels by 1 (Fig. 8):

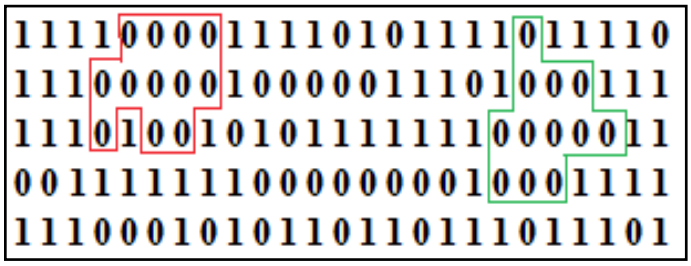

Fig. 8 : Components structure in image

For each component we create a rectangle that surrounds it, to visualize the results of segmentation and the following figure (Fig.9) represent the result of this step.

Fig. 9 : Result after detecting components

IV. Correction phase

After carrying out smoothing on the picture, we should apply an algorithm to groups the component having the same size and representing the same objects (lines of text in the same column ...). For this, we used the distance between the components as well as their size.

A. Distance

The distance [5] is the space between two components either vertically or horizontally as it is present in the following Figure (Fig.10).

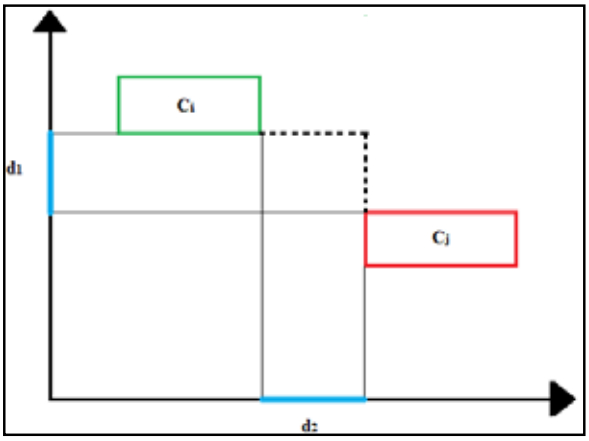

Fig. 10 : Distances between components

For any component Ci and Cj . i, j∈ [1, n] and i≠j. where d1 is the vertical distance as the horizontal distance d2

$$d_1 = \begin{cases} 0 & if\ L < R \\ |L - R|\ if\ not \end{cases}$$ eqn (7)

L represent the maximum point in the lift for one of the two components Ci and Cj and the R is the minimum point in the right.

The horizontal distance has the same definition as the vertical

B. The size

At the end of the phase of calculating distances, this is the part of the merger, but we introduced another parameter the sizes, whose purpose is to compare the sizes of the components, where the distance does not exceed a fixed threshold. Either we compared the height of the components if the distance minimum is the horizontal distance or width if it is the vertical distance [6].

The following Fig. 11 show explains what has just been said

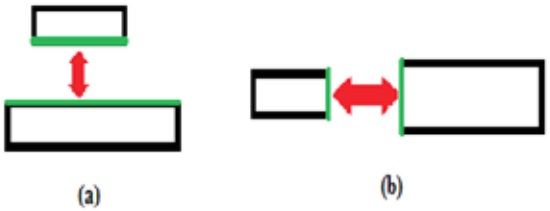

Fig.11: Comparing the two side components

To merge two components, we provided a condition in the sizes as follows where the Wi represent the width of a component C i and Wj the width of component Cj:

$$Wi < Wj * 1.20$$ eqn (8)

The width of the component Ci must be lower than of the Cj. This condition is to fuse the components that built the same column and also the Hi represent the width of a componentC i and Hj the width of component Cj

$$Hi < Hj * 1.20$$ eqn (9)

The height of the component Ci must be lower than of the Cj, this condition is to fuse the components that built a ling of text in order to reduce the number of components .

The Figure12 represent the final result of using this method.

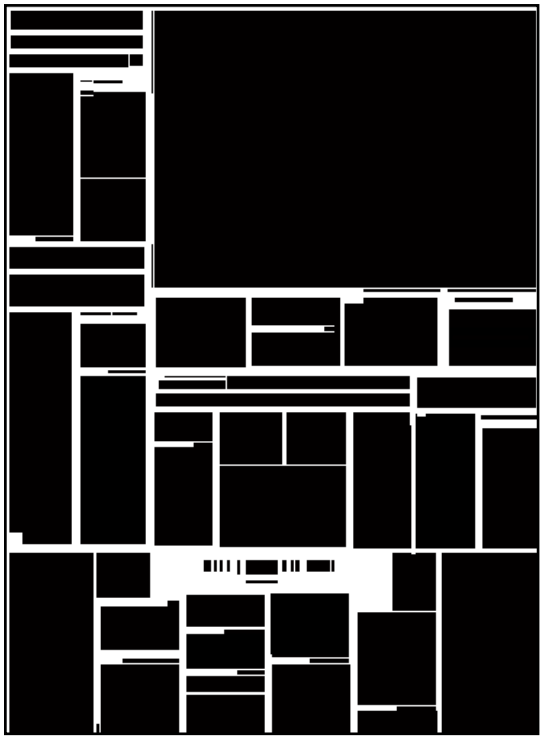

Fig. 12: Final Result

C. Experimental results

In this section we will give the results of applying our method on pages of newspapers of different languages and sizes to test the performance of our method

ISSN : 2347 - 8446 (Online)
ISSN : 2347 - 9817 (Print)

In order to test the performance we have used a data base composed of ten documents of various size and languages as follow:

Table. 1: The document size

Type Document	Size (Pixels)
1	750x1200
2	750*900
3	1000*2000

In segmentation we have two major type of problem first the fragmentation[11] means that a region of a single document component is erroneously divided into two or more regions. Second means that the regions of multi-document components are erroneously merged into one region. The first problem is not a serious one because most time it caused by size of text and fragments only the text area although the second one can fusion text and image blocks that request more processing to detect this problem and segment it.

Table. 2: Recognition result

	Total number	Fragmented	Over-merged
Text	202	20(9%)	10(5%)
Title	125	10(8%)	2(1.5%)
Non-Text	30	2(6%)	0(0%)

Let N be the number of document components in a document image, Nf be the number of fragmented document components, and Nm be the number of over merged document components. We define fragmentation rate as Nf/N and over-merging rate as Nm/N We use them to evaluate our segmentation results.
Another advantage is that our method works on documents of different languages and not like other methods that are related to a specific language that one strong point our method
Another point concerning the results of fragmentation and over-merged our method gives good results on this point, specially at the problem of over-margin compared to other methods. More our method based on two values so just change them when you want to change the number of segmented regions which gives flexibility to our method

V. Conclusions

The results obtained in this paper can be used later in other works of document recognition, for example, a image of document converter to other document format (pdf, doc ...). Else the structure of this method is a strong point, because it can be used on various types of paper of different language. This method may later use an intelligent system like (neural networks, SVM ...) for the correction phase based on the distance and size also we can add more parameters to improve results.

References

[1] *Anat Levin,Yair Weiss 'Learning to Combine Bottom-Up and Top-Down Segmentation' School of Computer Science and Engineering The Hebrew University of Jerusalem*
[2] *Jean-Luc Meunier 'Optimized XY-Cut for Determining a Page Reading Order'Xerox Research Centre Europe 6, chemin de Maupertuis F-38240 Meylan*
[3] *T.Saitoh and T. Pavlidis. "Page Segmentation without Rectangle Assumption".Proceedings of the 11th International Conference on Pattern Recognition, The Hague, USA, 1992, pp. 277-280.*
[4] *L. Cinquea, S. Levialdia, L. Lombardib, S. Tanimotoc "Segmentation of page images having artifacts of photocopying and scanning" aDipartimento di Scienze dell'Informazione, University of Rome "La Sapienza", Via Salaria 113, 00198 Rome, Italy Received 18 June 1999; received in revised form 16 February 2001; accepted 26 February 2001*
[5] *Jie Xi, Jianming Hu, Lide Wu "Page segmentation of Chinese newspapers" Department of Computer Science, Fudan University, Shanghai, Fudan 200433, People's Republic of China ,Received 31 October 2001; accepted 31 October 2001*
[6] *Fu Chang, Shih-Yu Chu, Chi-Yen Chen "Chinese document layout analysis using an adaptive regrouping" Institute of Information Science, Academia Sinica, 128 Academia Road, Taipei, 115 Taiwan, Received 22 October 2003; accepted 27 May 2004*
[7] *Chung-Chih Wu, Chien-Hsing Chou, Fu Chang "A machine-learning approach for analyzing document layout structures with two reading orders" Institute of Information Science, Academia Sinica, 128 Academia Road, Section 2, Taipei 115, Taiwan ,Received 16 April 2007 Accepted 12 March 2008*
[8] *Laura Caponetti , Ciro Castiello , Przemysław Go´recki, "Document page segmentation using neuro-fuzzy approach", Universita` degli Studi di Bari, Dipartimento di Informatica, Via E. Orabona 4, 70126 Bari, Italy. Received 16 December 2005; received in revised form 2 November 2006; accepted 22 November 2006*
[9] *P.BONNET « Cours de Traitement d'Image USTL » Université des Sciences et Technologies de Lille 2008-2009*
[10] *Luigi Di Stefano, Andrea Bulgarelli "A Simple and Efficient Connected Components Labeling Algorithm" DEIS, University of Bologna Via Risorgimento 2, 40136 Bologna, Italy*
[11] *K. Kise, A. Sato, M. Iwata, Segmentation of page images using the area voronoi diagram, Computer Vision Image Understanding 70 (3) (1998)*