Design Algorithm for Text and Figures Separating From Printed Image Document

Ali A.Yassen
Basrah University, Education College, Computer Sciences
AliAdel2005Alamre@Yahoo.com

ISSN –1817 –2695
(Received 3/9/2006, Accepted 6/3/2007)

Abstract

The aim of this research is to create algorithm that has ability to separate figures and texts which are exist in document.

The foundation of this algorithm, will deal with document as an image, then separate all figures which have many colors differ from the text color. But there is an effective problem that we might be false, there is some gray-levels values, which be near from gray-levels that exist inside the error text, whereas the error doesn't contain on black color absolutely.

Algorithm will make table consist of two fields; the first, the colors which exist in document, and the second number repetition for each color in document, algorithm will deal in intelligence with separate white, black colors and color (gray-levels) that mixed with texts in the writing, after that insulate all gray-levels which exist in the figures. At the final the outputs of algorithm are two documents:

The first, contains only writing (texts) and the second contains figures without making some distortion on writing documents.

Key Words : Fourier transform Pixels gray-levels Threshol Detection-Threshold Error Ratio Clear-Ratio RGB (Red-Green-Blue) FCM (Fuzzy c-means algorithm)

1. Introduction

In spite of the wide separate that uses of computers and other digital facilities, paper document keeps occupying a central place in our every day life. Conversely to what was expected, the amount of paper produced presently is larger than even important institutions like administrations, libraries, archive services...etc, which are heavy paper producers and consumers. To some point of view, paper is one of the most reliable information supports. Unlike numerical records, it is not constrained by format compatibility question, or device needs. On the other side document storage for safety or accessibility considerations is a very tricky problem. Research is presently done in such a direction. To digitize paper document, one must retrieve an image (For instance, via a scanner) and work on it. In this research, we are interested in the important step of document analysis, is Separating the text from many figures in the document.

2. Previous Works

Separate of figures in document images is part of document analysis and page segmentation. Traditionally, page segmentation methods are divided in three groups: top-down, bottom-up and hybrid approaches [1-3].

In top-down techniques, documents are recursively divided from entire images to smaller regions. These techniques are often fast, but the efficiency depend on a priori knowledge about the class of documents to be processed. Developments have been produced in early times. The most well known are projection methods [4-5], histogram analysis, rule based systems [6], or space transforms (Fourier transform, Hugh transform, etc).

Bottom-up methods start with the thinnest elements (pixels), merging them recursively in connected components or regions, and then in larger structures. Most popular bottom-up techniques are mathematical morphology [7-8], run length smoothing algorithm, and region growing-based methods [9]. The suggestion algorithm is element in bottom-up methods.

3. Document Representation

The document is represented by image which is input through scanner device or drawing it by using paint program. This document contains a set of differences figures such as circle, square, etc, which may be contains a texts in side of it, as well as the texts written outside of these figures .The document that will deal with is gray-level, whereas the process will be performed. After that, The separate step of figures from text. This step has two output documents; the first contains text and without distortion and the second contains just figures with exact clear degree on these figures. The figure (1) is clear the logical series of suggestion separate system:

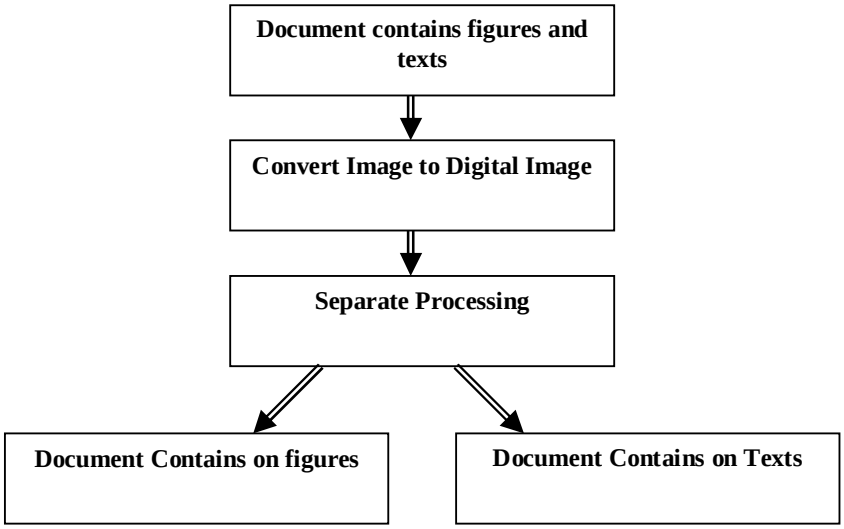

Fig. (1) Show the sprat system

4. Algorithm Suggestion

We suggested algorithm for Separate differences figures and differences texts detection. The aim of algorithm is not only Separate texts from figures in documents, but texts and figures are very clear exact in documents contains figures and texts, So that, we make this algorithm ability for doing Arabic and English text. The aim of algorithm deals with document after converting it to the gray-level image and then makes perform separate operation. The processes operation is detection the value of each image pixel and then computes the numbers of repeating of it, the results of this operation are saved as in table to be used in next step. After that, sort operation will be coming to applied on table which create in previous step, the purpose of these steps to threshold detect, where this step is very important, because the threshold contains all values which are not associated in background of document (white), color of texts (black) and the gray-level that we can find it associated with characters of texts. We know that the texts of documents not always black color, but often contain gray-levels these levels inside it; the figure (2) explains some letters that have some of gray-levels:

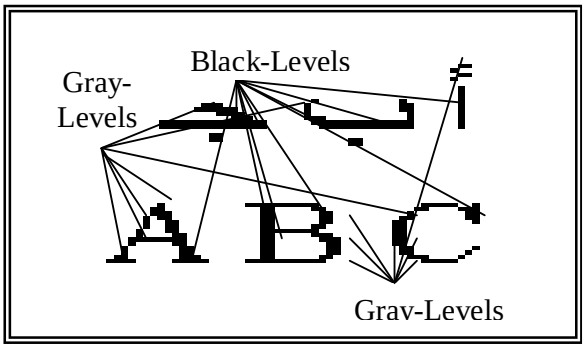

Fig. (2) Show the gray-levels and black-levels in this document

Now after threshold is specie, algorithm will be scanning on each pixel in image and remove it, if the following condition is true:-

If (Pixel) = Threshold then
NewDoc =255
Else
NewDoc = Pixel

Where NewDoc is meaning text-document which has text only.
255 represent value of white color.

Finally, algorithm will detect the figures area and then remove all areas that contains text when compare original document with text document, and then save this result in new document called Figures document.

The suggestion algorithm is the following shape:

```
Suggestion Separate Algorithm
    Input Document
    Convert Document to Image / image is gray-level
    [N M] = Size (image)
    K = 0
    For i = 1 To N
        For j = 1 To M
            Compute Gray-level-Repeat (image (i, j), Rep)
            K = K + 1
            Table (K, 1) = Image (i, j)
            Table (K, 2) =Rep
        End for
      End For
    Detection-Threshold (Table, K, Threshold, R)
  For i = 1 To N
    For j = 1 To M
        Flag = 1
        For F = 1 To R
            If Image (i, j) = = Threshold (f) Then
                Flag = 0
            End if
            If image (i, j) <> 255 and image (i, j) <> 0 and flag = = 1 Then
                Doc-Text (i, j) = 255
             Else
                Doc-text (i, j) = image (i, j)
             End if

          End for
        End for
      End for
End Suggestion Separate Algorithm
```

N: number of rows in image.
M: number of columns in image.
K : index is detection the numbers of colors in image and then using it in the table which contains each colors in documents in field table(K,1), as well as, the counts of repetition of these colors in table(K,2) that assign it by REP variable.

Threshold: Matrix contain on the colors (gray-levels) which is more repeated in documents and the R represent size of this Matrix.

Detection-Threshold : Procedure is detecting the threshold for using to text separate from document.

Flag: A variable controls removing figures from document or not, depending its value:

If flag value = 1 Then
Figure is removing
Else
This region contains text or background

Doc-Text : output document from Input document (Original) that is containing on text only.

Now, how we can get the figures document? To answer this question by the following algorithm:

```
Separate-Figures Algorithm
   S=0
    For i = 1   To    N
      For j = 1   To    M
          If Image (i, j) = = Doc-text (i, j)   Then
               S = S +1
               Doc-fig (i, j) = 255
          Else
              Doc-fig (i, j) = image (i, j)
           End if
        End for
     End for
  R = (S/ (N*M))* 100
End Separate-Figures Algorithm
```

Where S : Variable for detecting perfect of text in doc-text, and then get to ratio of matching between text in input(original) document and text document in variable R.

Doc-fig: Document has figures.

5. Experiments

We will deal with a group of documents that contains different figures with much color, which may be more than gray-levels, then using in texts that have different types and sizes with two languages: Arabic and English. After that accounting clear degree of document text which is contains on texts only, so that the document will be ready for any recognition operation that may be perform on it. In addition, appears the table that has each value and number repetition in input document, as well as, detection threshold that do not influence of text distortion. In the fact ,We are exam algorithm on thirty documents which are different in figures, colors, sizes, fonts. In the last we are get the best results on all documents. Now, we are display two experiments as the following shape:

5.1 First Experiment

In this experiment, we have English texts and some figures that have many colors. The figure (3)clear these shapes and texts, as well as, the table (1) contains on each pixel value, repeating numbers, numbers of thresholds, clear ratio and error ratio. Error ratio is compute depending up on

$$New = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Doc-text}(i, j) + \text{Doc-fig}(i, j) \longrightarrow (1)$$

the difference between original documents and summation of two output documents by using the following equations:

$$Ratio = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Original}(i, j) + \text{New}(i, j) \longrightarrow (2)$$

Where: n, m size of documents.

Doc-texts: Image contains on texts only.
Doc-fig : Image contains on figures only.
Ratio : ratio of error.

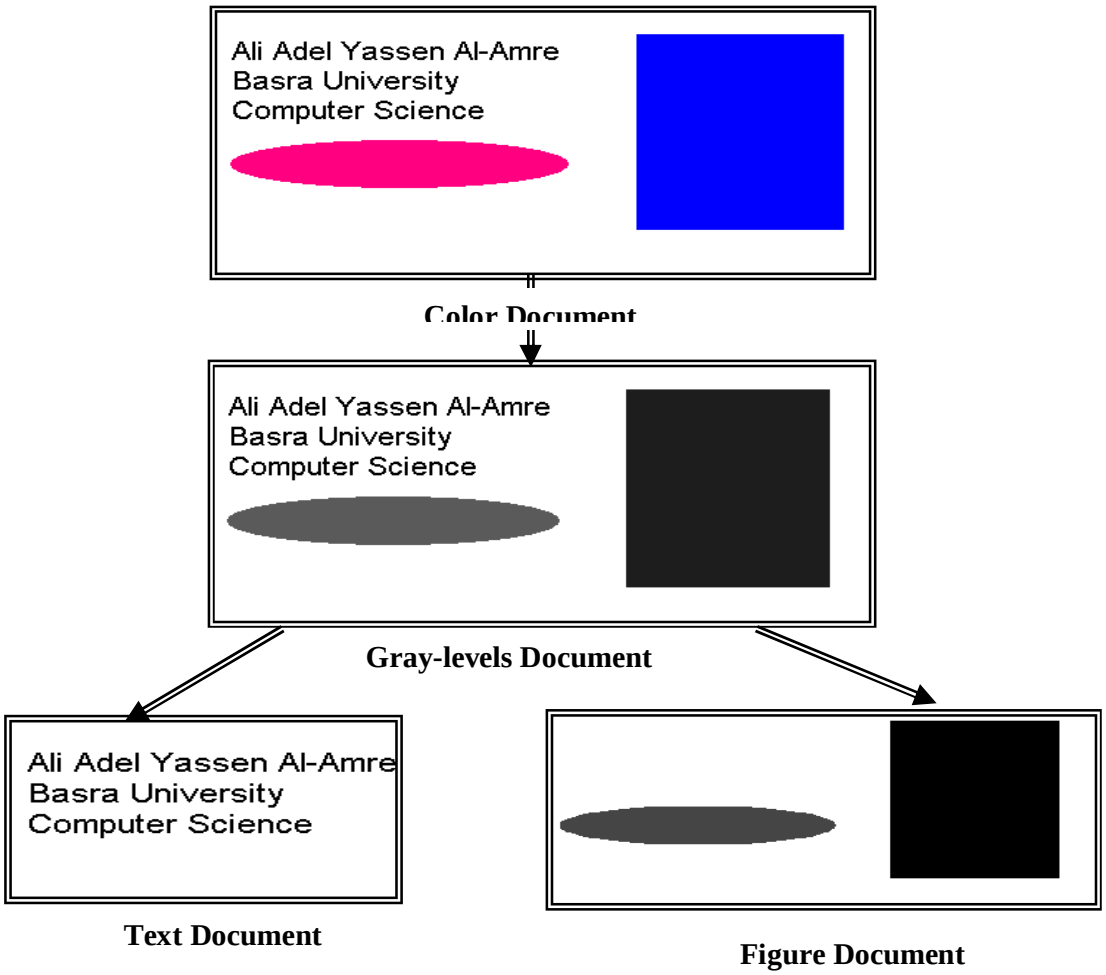

Fig: (3) Clear the steps of algorithm applied on certain document

Table (1) clear experimental results of suggestion algorithm on document in experiment (5.1)

pixel-value	Repeating	Error-Ratio	Clear-Ratio
29	22418	8.8797	91.12
90	7080		
0	2610		

5.2 Second Experiment

In this experiment, we have English and Arabic texts and some figures that have many colors. The figure (4)clear these shapes and texts, as well as, the table (2) contains each pixel value, repeating numbers, numbers of thresholds, clear ratio and error ratio.

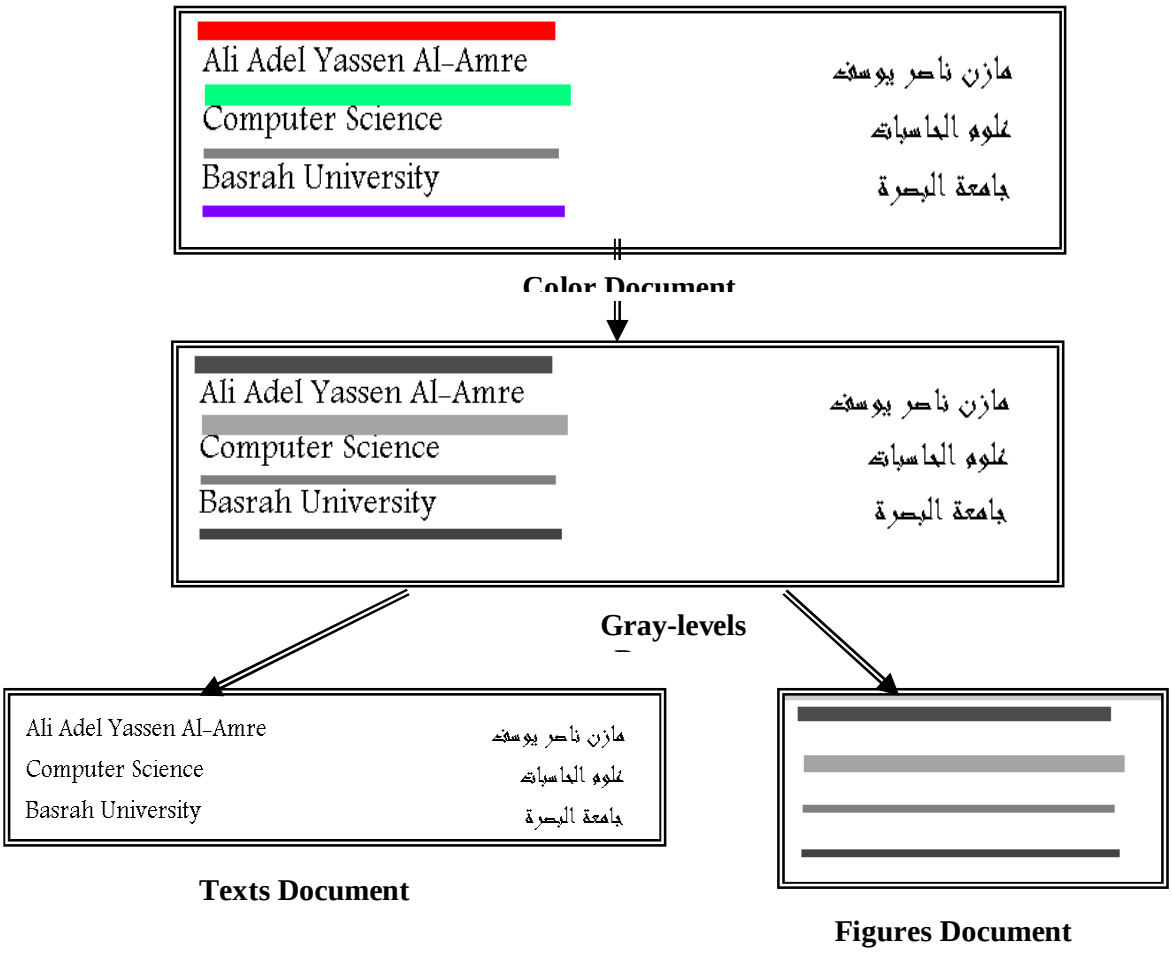

Fig. (4) Clear the steps of algorithm applied on certain document

Table (2) clear experimental results of suggestion algorithm on document in experiment (5.2)

pixel-value	Repeating	Error-Ratio	Clear-Ratio
164	4589	3.907	96.093
0	4110		
76	3886		
67	2423		
128	2078		

6- Conclusion and Future work

From the suggested algorithm, we can conclude a set of important things:

1- Algorithm has ability to separate differences figures and texts from document that has different types and sizes; moreover the texts may be in both languages Arabic and English.

2- The clear and exact ratio of figures and texts separating is very high during algorithm applying to many documents.

3- Algorithm has some faults, one of these the threshold that is used restricted, whereas if matching it with document contains color image will need more than one iteration. The number of iterations depend on many color inside image and the gray-levels , more over the produced font will be un clear , then will be change operate of figures separate algorithm.

For increase in this field, we suggest the following:

1-Development this algorithm during built completing threshold system, in order to deal with documents that consist of figures, texts and pictures, more over using some of computer vision system such as RGB, Opponent, etc, The aim of using these systems is to increase perfect and exact results when dealing with color document.
2-The ability to using clustering analyses methods such as fuzzy C- Means (FCM) and K-Means for create some of clusters representing the values of colors which exist in figures are found, in order to be easy for texts detecting and choice threshold.

References

1- C. H. Chen ," Statistical Pattern Analysis ", Spartan Books, Hayden Book Company, Inc., Rochelle Park, New Jersey (USA), (1973).
2- L. O'Gorman and R. Kastrui, " Document Image Analysis", IEEE Computer society executive briefing, IEEE compute society, Los alamitos (California, USA), (1997).
3- J. Doung, H. Emptoz and C. Y. Suen," Extraction Of Text Areas In Printed Document Images", Center for pattern Recognition and machine intelligence suite Gm – 606 (CENPARMI), Concordia university,(2001).
4- J. R. Parker, " Algorithm For Image Processing and Computer Vision" , John Wiley and Sons ,Chichesten ,New York ,Brisbane, Toronto ,Singapore ,Weinheim ,design and measurement in electronic engineering edition,(1997).
5- T. Pavlidis, "Structural Pattern Recognition", Springier–Verlag, Berlin, Heidelberg, New York, Springier series in electro physic edition, (1997).
6- J. Serra, "Image Analysis and Mathematical Morphology (vol.1). Academic Press, New York, (1982).
7- V. Wu and R.Manmatha, "Document Image Clean-up and Binarization", Technical report, computer science department, university of Massachusetts, Amherst USA, DEC, (1997).
8- A. A. Alamre, "Image *Matching*", MSC Thesis, Science College of Basrah University, (2005).
9- X. Lu, D. Colbty and A.K .Jain , " *Matching 2.5D Scans for face Recognition* ", Department of computer science & Engineering , Michigan state university, East Lansing, I 48824,(1999).

علي عادل ياسين

البريد الاليكتروني : ***AliAdel2005Alamre@Yahoo.com***

المستخلص

يهدف هذا البحث إلى إنشاء خوارزمية لها القابلية على فصل الأشكال ذات التدرجات الرمادية المختلفة عن وثيقة تحتوي على مجموعة من النصوص. أساس عمل هذه الخوارزمية إنها سوف تتعامل مع الوثيقة على كونها صورة ثم بعد ذلك فصل كافة الأشكال ذات التدرجات الرمادية المختلفة عن اللون المكتوب فيه النص، لكن ما واجهناه من صعوبات هو إن بعض الأشكال تحتوي على تدرجات رمادية تكون قريبة من التدرجات الرمادية التي تكون ممزوجة مع بعض خطوط النصوص حيث إن الخط لا يكون باللون الأسود المطلق بل يحتوي في كثير من جوانبه على الألوان رمادية. الخوارزمية سوف تقوم بإنشاء جدول يتكون من حقلين الأول التدرجات الرمادية في الوثيقة والثاني عدد مرات تكرار كل لون في الوثيقة ، الخوارزمية سوف تتعامل بذكاء في فصل اللون الأسود والأبيض واللون الذي يختلط مع النصوص بالكتابة ثم بعد ذلك تعزل كافة الألوان التي تتواجد ضمن تلك الأشكال وبالنهاية مخرجات هذه الخوارزمية وثيقتين الأولى تحتوي على كتابة فقط و الثانية تحتوي على الأشكال فقط.