Identification of Text-Only Areas in Mixed Type Documents

C. Strouthopoulos, N. Papamarkos and C. Chamzas
Electric Circuits Analysis Laboratory
Department of Electrical & Computer Engineering
Democritus University of Thrace
67100 Xanthi, Greece

ABSTRACT - In this paper, we present a new method of text identification in mixed type documents that contain text, images, graphics, drawing, etc. The identification of text areas in a document is crucial for optical character recognition (OCR), image compression and image storage. The proposed method separates the document in text-blocks and no-text-blocks. To classify a block as text we use characteristics such as size, frequency, collinearity and vicinity. The performance of the method was tested on a variety of images. Its effectiveness is demonstrated by presenting three characteristic examples.

1 INTRODUCTION

One of the most important procedures in digitised documents analysis is the segmentation process [1-2]. The segmentation and classification of digitized documents into regions of text and images is a first pre-processing step in such a document analysis system. The distinction between graphics and text is quite difficult in many cases [3]. Another reason for this separation is that we can improve the compression ratio by encoding text and image areas with different methods. Archiving of mixed type documents using block segmentation and recognition requires also text-area definition. Finally, the identification of text areas is necessary and in other special applications (e.g. CAD/CAM). The important problem of document segmentation of text-only "recognition" is addressed here and we propose an effective, fast and practical method for finding such regions. In other words, we are trying to answer the question "where in a document we have only text?".

In the literature, there are two basic approaches to address the above problem. In one of them, Fletcher and Kasturi [1] propose a method that starts by first finding the connected components of an image and then separating graphics from text using the relative frequency of occurrence of components as a function of their areas. In the next step, they use an iterative procedure to improve the initial estimation by applying the Hough transform to all connected components. The second method is the run-length block segmentation algorithm (RLBSA), which can also be considered as a smoothing algorithm [4]. This is an effective and low complexity technique for block segmentation and automatic classification of a
digitized document into lines of text and regions of a. images. The RLBSA method imposes a smoothing on the document using two parameters defined in a b. heuristic way (one for the vertical and one for the horizontal direction). For the block classification, c. additional parameters are used leading to the necessity to train the system with documents d. having similar fonts or other morphological characteristics.

In this work we try to solve the above mentioned segmentation problem for a wide class of documents. We take advantage of the characteristic properties of the text and we develop a method that is more robust and has better performance. As text we define a large area with parallel lines and sufficient number of words or characters. The proposed technique is independent of the size and type of characters as well as the position of text and graphics in the document. Its basic stages are: binarization, block segmentation, surrounding rectangles creation, Hough transform, filtering, and extension of surrounding rectangles. The horizontal extension of text surrounding rectangles is one of the powerful new features of the method that makes it applicable to a wide class of documents. The outcome of the method is the identification of areas containing only text. The remaining part of the document might contain images, graphics, isolated words or characters, or text of significantly different point size. The method was tested with several documents containing text, line drawing, graphics or images. The experimental results confirm the effectiveness of the proposed method.

2 DESCRIPTION OF THE ALGORITHM

The proposed method for text string separation of mixed text graphics images, is based on the following general characteristics of a text line:

Text lines consist of characters and symbols of almost the same height.

Text blocks have centroids which are approximately collinear.

The distance between text blocks is about the size of their heights.

If we join the text blocks in a text line, then we can enclose them in elongated rectangles.

The method is composed of the following five stages.

2.1 Edge extraction and binarization

If the original image is not binary, then we must convert it in binary form. To do this, we first apply an edge extraction technique. It is very important for the success of the entire segmentation process to have a good edge extraction. After edge extraction, the document is converted into a binary form using a threshold algorithm [5].

2.2 Construction of document blocks

We define as a block a set of connected pixels. Therefore, a block includes all the pixels having at least one path leading to other pixels in the set.

2.3 Filtering of the rectangles according to their height

After the above stage, we filter the rectangles according to their heights and collinearity. We accept that the characters have almost the same height in any pure text area. Taking advantage of this, we continue with the finding of those rectangles which their heights appear very often in the document. To accomplish this, we create a histogram describing how many times each rectangle height appears in the document. Denoting Hmax the global maximum value of H(j),
we accept those rectangles with height $h_i$ that satisfy the following condition:

$$\frac{H_{\text{max}}}{2} < h_i < 2H_{\text{max}}$$

2.4 Filtering the rectangles using Hough Transform

It is obvious that the centroids of rectangles that correspond to text, are almost collinear to the horizontal direction. These centroids define horizontal lines which can be determined using the Hough Transform. From these lines, we accept those constructed from a large number of centroids.

2.5 Extension of the block rectangles

In the final stage of the method we check if the blocks of the document, which are enclosed in the rectangles, satisfy the next two pure text area properties:

- The distance between text blocks is approximately the size of their heights.
- If we join the text blocks in a text line then we can enclose them in elongated rectangles.

To take advantage of the above conditions we do the following:

Step A

We extend every rectangle to the left and to the right with a length equal to the rectangle height. As a result of this procedure, all the adjacent rectangles are connected along a text line, creating a chain of overlapping rectangles.

Step B

Considering as a new binary image the result of Step A, we surround each block (which is composed of extended rectangles) in new rectangles.

3 EXAMPLE

The proposed method was applied successively in many mixed type documents. Due to the space limit, we present here only the application of the new method to the document of Figure 1. For this document, Figure 2 shows the chains of overlapping rectangles while Figure 3 gives the final segmentation results.

4. CONCLUSIONS

We have presented a new method able to identify text-only areas in mixed type documents. This algorithm uses a variety of text characteristics to identify pure text areas. The method is independent of the size and type of the characters and the position of the text in the document. It works well even in the cases where graphics and text areas cannot be separated by vertical and orthogonal lines.

REFERENCES


Figure 1. Original mixed type document.

Figure 2. Chains of overlapping rectangles.

Figure 3. Final segmentation results.