SEGMENTATION AND CLASSIFICATION OF DOCUMENT IMAGES

A Antonacopoulos and R T Ritchings

1. Introduction

There is a significant and growing need to convert documents from printed paper to an electronic form. Document image analysis is concerned with the segmentation of the document image into regions of interest, their description, and the classification of the regions according to the type of their contents. In this paper, a new unified approach to page segmentation and classification, based on the description of the background with tiles, is presented. The segmentation method is flexible to successfully analyse and describe regions in complicated layouts where other methods fail. Images with severe skew are handled equally well with no additional computations. The classification is based on textural features which are derived by simple calculations from the representation of space in the regions, produced during the segmentation process. This is a considerable advantage over previous methods where extra image accesses and lengthy computations are necessary. Overall, the whole approach of segmentation and classification by white tiles is fast and efficient as no time-consuming processes are required.

2. Background

Page segmentation is a key component in a document analysis system. Its performance significantly influences the subsequent processes. Badly identified regions can lead to poor recognition results and increased run time in these later stages. The objectives of an efficient page segmentation method are as follows. Firstly, it must be flexible in that it must handle complicated layouts as well as the 'traditional' ones which contain rectangular printed regions. Secondly, it should provide a good description of the regions for the efficient performance of later stages. Thirdly, a page segmentation method must be fast because it deals with a vast amount of data (an A4 page digitised at 300 dpi is about 8.5 million pixels). Finally, the method must co-operate with the processes that follow it. This should be in terms of provision of features for page classification and faithful description of the regions.

The dominant earlier approaches to page segmentation have assumed that all printed areas in the page image are *rectangular*. This is limiting not only in the case where these areas are of different shapes, but also where skew has been introduced during the scanning of the document page. In the case of skew, computationally expensive methods may be used to correct it but, when areas of interest are not rectangular, segmentation fails completely. For a more detailed analysis of previous approaches the reader is referred to other papers[1,2,3].

A practical page classification method must complement the page segmentation approach chosen. It should utilise the data and measurements produced during segmentation before performing further measurements when computing features. Accessing the image data again is time consuming and should be avoided wherever possible. Furthermore, the classification process should be applicable to the same type of documents and under the same circumstances as the page segmentation is. For instance, it should be able to classify regions of complex

Dr A. Antonacopoulos is with the Department of Computer Science at the University of Liverpool, P.O. Box 147, Liverpool, L69 3BX, U.K. (aa@csc.liv.ac.uk) and Dr R.T. Ritchings is with the Department of Computation at UMIST, P.O. Box 88, Manchester, M60 1QD, U.K. (tim.ritchings@umist.ac.uk).

© 1995 The Institution of Electrical Engineers.
Printed and published by the IEE, Savoy Place, London WC2R 0BL, UK.

shapes and if skew detection and correction is not necessary for segmentation it would be an advantage if it is also not necessary for classification.

The vast majority of previous classification approaches assumed, as a consequence of the segmentation, that the areas of interest in the image are rectangular blocks. The ones that did not assume rectangular regions required computationally extensive calculations of image transforms. All the approaches required at least one additional time-consuming image access specifically for the computation of classification features. Furthermore, skew detection is necessary and, in most of them, a skew corrected image is mandatory. All these requirements make the classification process a lengthy one. Another important issue is their inapplicability to images of documents with non-rectangular regions.

Figure 1. A document page image.

In this paper, new approaches to page segmentation and classification are described. The *structure* of the background space is analysed in both of them to identify and classify the areas of interest in the document image. One of the benefits of these new approaches is the ability to process and analyse documents whose printed regions may not be rectangular. The methods are designed to be practical in order to be applied to real-world problems. With this in mind, apart from being able to successfully identify and classify regions of complex shapes, they are efficient and fast. Extra consideration has been given so that no time-consuming operations are required. Even the presence of considerable skew does not introduce any complications in terms of accuracy and processing time. Furthermore, the page segmentation and classification methods are tightly coupled under a unified approach so that the whole approach benefits from the added efficiency resulting from this co-operation. For instance, classification features are derived by simple computations from the description of image regions and other data available as by product of the segmentation process.

In the next section, the principles on which the page segmentation and classification methods are based are outlined. In Section 4, the page segmentation process is described with each step analysed in a separate subsection. The page classification stage is presented in Section 5. Finally, Some representative results are shown and discussed in Section 6.

3. The White Tiles Approach

The image of a page of a document can be seen in Figure 1. The printed regions on a page of a document are surrounded by white space which can be thought of as an irregular (because of the different shapes of the regions) net. The idea is that by reconstructing this net of white space one can identify and describe the holes i.e. the printed regions. This approach does not make any assumptions about the shapes of the regions. A flexible way to describe the surrounding net of white space is by white tiles. Each white tile represents the widest area of white space that can be represented by a rectangle. Hence, the whole net is represented as a set of white tiles of different sizes. The printed areas are identified by their contours. Each contour is a list of white tile edges that border with the region in question. Furthermore, an additional advantage of the white tiles approach is that those that are not used during the segmentation i.e. the white tiles inside the identified regions, can be used for the classification of these regions. The main principle is the exploitation of the texture characteristics of text and other areas using the white tile information in order to achieve fast classification of regions of varying shapes.

4. Page Segmentation

The method proceeds as follows. After pre-processing, a complete description of the white space areas (background) is constructed in terms of a series of tiles of varying sizes which follow the shapes of these areas very closely. Then, streams of these white tiles whose sides encircle printed regions are identified as belonging to the net of white streams. The surrounded regions are identified by tracing along the region-bordering edges of

the white tiles. A more detailed description of the stages of the method and the issues involved in each one is given in the following sub-sections.

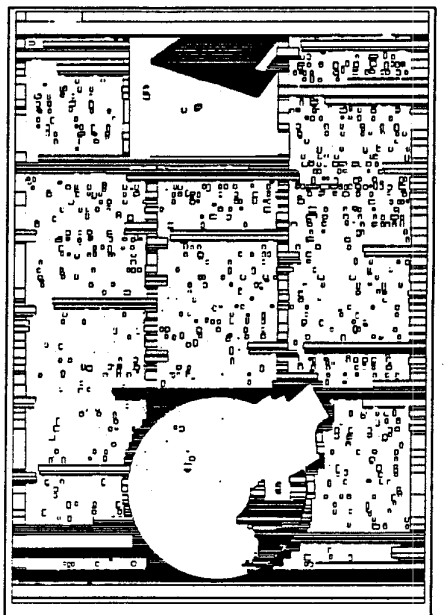

Figure 2: The white tile description.

4.1. Identification of White Tiles

In a page image, there is an abundance of white space apart from that belonging to these streams. There is also white space between text lines of the same paragraph, between words and characters of the same text line and inside characters themselves. The goal of the pre-processing step is to simplify the effort of subsequent steps in identifying the appropriate streams of white spaces. This is done by blocking or isolating white spaces that do not form part of the delimiting streams. To achieve this effect, a *vertical smearing* technique[1] is applied to the image.

After pre-processing, the next step is to describe all the remaining background space in the image by white tiles. A white tile is represented by a rectangle which is vertically and/or horizontally stretched or squashed to fit the longest possible white area in the horizontal direction. No restriction is imposed on the height of the rectangle. If the shape of the white space to be covered varies sharply in the vertical direction, the white tile can be reduced to a horizontal line (i.e. the horizontal sides will coincide).

It should be pointed out that white tiles whose width is narrower than a certain threshold are not considered as part of the description of the region delimiting streams. This threshold is not critical but it should be less than the width of the smallest white tile on the delimiting stream. The threshold used in this method is 1/3 of the baseline difference (computed during pre-processing). Each white tile identified is described by a data structure that holds information about the position of the white tile and about which white tiles are immediately above and below.

The white tiles identified in the smeared image in Figure 1 can be seen in Figure 2. Overall, the white tile determination process is fast, using only one sequential pass over the image data. It is also very accurate in covering the streams of white spaces with as few white tiles as possible. At the same time it produces information which is later used for the derivation of page classification features.

4.2. Segmentation

At the segmentation stage, the contours of the printed regions in the page are recognised from the tiles describing the streams of white spaces in the image. Each of these contours is a list of white tile edges that border with the corresponding printed region. Each list is cyclic and each element of a list is unique. The objective is to trace the appropriate edges of the white tiles that make up the delimiting white streams of the original image.

The white tile arrangement in the image can be traced as a graph. The white tiles will be the vertices, and the edges of the graph will represent the vertical adjacency between two white tiles. In this graph, then, the problem will be to trace the *minimum* cycles which encircle areas that *do not intersect*.

Tracing cycles in graphs is usually a time consuming task with a lot of computational effort spent in exhaustive search to identify all possible cycles. Moreover, choosing the needed cycles will also add to the overhead. In contrast, the approach described here is fast. It sequentially recognises only the correct cycles and traces each cycle only once. Furthermore, the search through the white tiles to identify a start of a cycle is performed only once. Once the first white tile is chosen for the tracing to start, there is no need for time consuming searches. All

starts of cycles will be identified while tracing other cycles. This makes use of the fact that the streams of white space that delimit regions in the page are connected.

The algorithm is described in more detail in a previous paper[1]. It starts by identifying a starting segment for the contour of the first region to be traced. It then proceeds from tile to tile around the region, guided by rules. For each tile, the appropriate parts of edges that border with the region being traced are appended to the contour. The tracing of a cycle ends when the starting contour segment is reached again. During the tracing of a cycle, segments that constitute potential starts of other cycles are recorded and put in a queue. A potential start may not always be the start of a cycle but simply part of it. The next potential start in the queue is tried each time a cycle has been traced.

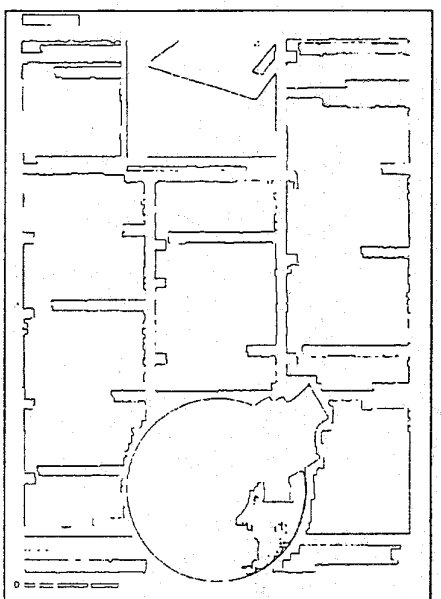

Figure 3: The segmented regions.

The contours of the segmented regions of the image in Figure 1 can be seen in Figure 3. Each of these edge lists follows the shape of the region very closely thus, constituting a quite accurate description of the region. Small segmented regions that belong to the same printed area on the page, can be grouped together if desired. These regions may result mostly from isolated single lines of text and titles. Overall, the segmentation process is efficient in that the cycles are traced sequentially without the need for exhaustive search. Only the needed cycles are traced and no cycle or part of it is traced more than once. Computation time depends linearly only on the amount of the white tiles covering the region-delimiting streams. Isolated white tiles inside the regions do not affect the process.

5. Page Classification

The different types of regions in the image have different textural characteristics. Having the information about the white tiles from page segmentation, the most appropriate textural property to exploit is the white space inside regions. Furthermore, it is natural that a practical page classification method should exploit this information instead of performing new computations on the image data. It should be noted that the description of the regions identified during page segmentation is adapted to allow for efficient search and indexing. Instead of the contour as a sequential list of edges, each region is described by a structure of rectangular intervals of various sizes to suit its shape. Searching inside regions for white tiles is very fast as the appropriate interval is indexed using the coordinates of the white tiles to be searched for. A more detailed description of this representation of the regions and its construction can be found in a previous paper[4].

5.1. The Features

After page segmentation all tiles that form part of the delimiting net of space have been marked as 'used' while they were traced in the white tile graph representation. The remaining white tiles are those that lie inside regions and are of two kinds: narrow (whose width is less than the threshold mentioned in Section 4.1) and wide. At this point the following observations can be made:

a) T*ext regions* contain a significant number of narrow white tiles. These are evenly distributed inside the region and the white area that they describe is large in proportion to the total area of the region.

b) *Graphics* regions usually contain less white space. There are more wide tiles than in text regions. The size of the tiles may vary significantly and they are not evenly distributed.

c) *Line art* regions are characterised from the relatively large amount of space they contain in the form of wide tiles. The size of the tiles may vary considerably in contrast with those in regions of text.

Based on the above observations, four features are derived by simple computations from the white tiles.

$$F_1 = \frac{\text{total area of region}}{\text{total white tile area}} \text{ , white tile number} > 0.$$

$$F_2 = \frac{\text{area from wide tiles}}{\text{area from narrow tiles}} \text{ , number of narrow} > 0.$$

$$F_3 = \frac{\text{mean area from wide tiles}}{\text{mean area from narrow tiles}} .$$

$$F_4 = \frac{\text{number of narrow}}{\text{number of wide}} \times F_1 \text{ , number of wide} > 0.$$

Text and line art regions have low F_1. Text regions also have low F_2. F_3 is a measure of complexity of the white tiles. Therefore regions of graphics or line art are likely to have low values (low mean area from wide and/or high mean area from narrow tiles) while text has higher F_3. Finally, F_4 is used to identify line art regions as they tend to have high values.

5.2. The algorithm

The first step is to identify which white tiles belong to which region by checking for inclusion aided by the region representation scheme. Once the white tiles of each region have been gathered together and their total area is estimated, the classification method proceeds for each region as follows.

if no white area **then** graphics
  else if $F_1 > T_{F1}$ **then** graphics
    else if no wide tiles or are insignificant **then** text
      else if ($F_2 < T_{F2}$) **AND** ($F_3 > T_{F3}$) **then** text
        else if $F_4 < T_{F4}$ **then** line art
          else graphics.

Small regions (having less area than an average word) are examined in a different way to increase the robustness of the method. If they are wide and very short (horizontal lines), or their shape does not resemble that of a character string, they are classified as graphics. The thresholds have been experimentally determined by applying the method to a variety of documents containing text regions with fonts of various sizes, graphics regions and line art. T_{F1} is set to 10. This is necessary to include text regions with small fonts when most of the text regions have larger fonts. T_{F2} is set more flexibly. If F_2 is less than 1 then it is set to 1. If F_2 is between 1 and 2 then T_{F2} takes the form of a straight line equation increasing slightly as the area of the candidate region increases. For F_2 values greater than 2, T_{F2} remains 1. F_3 is set to 1 and F_4 to 3.

The text regions extracted from the image of Figure 4 can be seen in Figure 6, while the regions classified as graphics are shown in Figure 7. The text regions of the image in Figure 5 are illustrated in Figure 8 and the line drawing regions are depicted in Figure 9. Note that solid black regions (those containing no white tiles) are also shown in Figure 9. For these regions it is not straightforward to decide whether they are graphics or parts of line drawings. Hence, greater emphasis was given to correct classification of text regions. In this respect, during the tests, the method did not classify any text region as non-text.

Figure 4. Image scanned with 15° of skew.

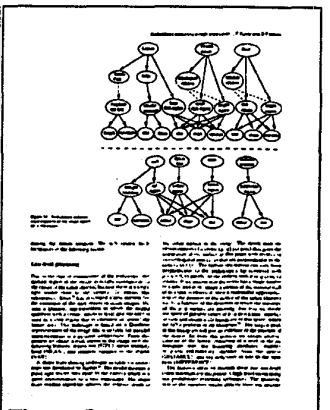

Figure 5. A document containing line art.

6. Results and Concluding Remarks

The methods described in this paper have been tried successfully on several documents containing printed regions of rectangular as well as of a variety of shapes. The methods have also been tested successfully on page images which were

scanned with skew such as the example of Figure 4. It should be noted that the methods were applied with no extra processing for skew estimation.

The main benefits of segmentation by white tiles are accuracy and speed. It can handle regions of non rectangular shapes, where other methods[5,6] will fail. It also produces more accurate descriptions of the printed areas on the document page. This is an improvement on the method of Pavlidis and Zhou[7] where non rectangular regions may appear as a set of disjoint rectangular blocks. Segmentation using white tiles is fast, in that it does not employ any time consuming computations. No skew correction is needed and it does not rely on successive grouping of components or regions. With the use of white tiles a great reduction in the amount of data is achieved. Finally, the cycle tracing algorithm sequentially identifies precisely the needed contours with neither backtracking nor exhaustive search. A further advantage of segmentation by white tiles is that the white tiles it identifies in the image can be used as the basis for the stages that follow in the document analysis and understanding processes. Most importantly, the white tile information is used to derive features for page classification.

Overall, the page classification process is a practical one. Through an efficient region representation scheme and the use of information produced by the preceding page segmentation process it does not perform any time consuming operations. In contrast with previous approaches, there is no need to access the pixels of the document image. The features are computed from the white tiles i.e. the representation of the white space in the segmented regions. An additional advantage is that the method is capable of classifying non-rectangular regions and regions which are skewed.

References

1. Antonacopoulos A. and R.T. Ritchings, "Flexible Page Segmentation Using the Background", *Proceedings of the 12th International Conference on Pattern Recognition (12th ICPR)*, Vol. II, October 9–12, 1994, Jerusalem, Israel, pp. 339–344.

2. Nadler M., "A Survey of Document Segmentation and Coding Techniques", *Computer Vision, Graphics and Image Processing*, **28**, 1984, pp. 240-262.

3. Srihari S.N. and G.W. Zack, "Document Image Analysis", *Proceedings of the 8th International Conference on Pattern Recognition*, Paris, France, 1986, pp. 434-436.

4. Antonacopoulos A. and R.T. Ritchings, "Representation and Classification of Complex-Shaped Printed Regions Using White Tiles", *Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR'95)*, August 14–16, 1995, Montreal, Canada, pp. 1132–1135.

5. Nagy G., S. Seth and S.D. Stoddard, "Document Analysis with an Expert System", *Pattern Recognition in Practice II*, North-Holland, 1986, pp. 149-159.

6. Baird H.S., "Background Structure in Document Images", *Advances in Structural and Syntactic Pattern Recognition*, H. Bunke (ed.), World Scientific, 1992, pp. 253-269.

7. Pavlidis T. and J. Zhou, "Page Segmentation and Classification", *CVGIP: Graphical Models and Image Processing*, **54**, no. 6, November 1992, pp. 484-496.

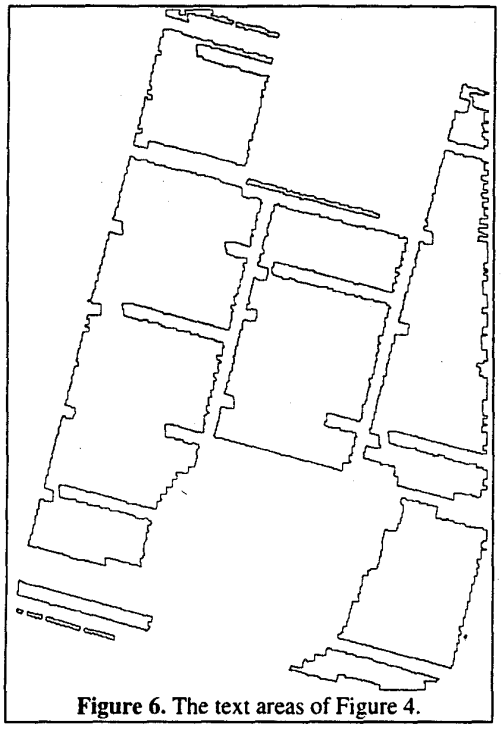

Figure 6. The text areas of Figure 4.

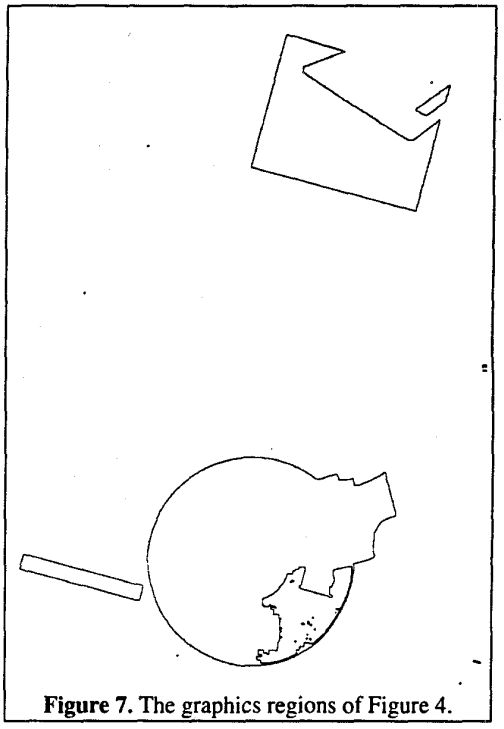

Figure 7. The graphics regions of Figure 4.

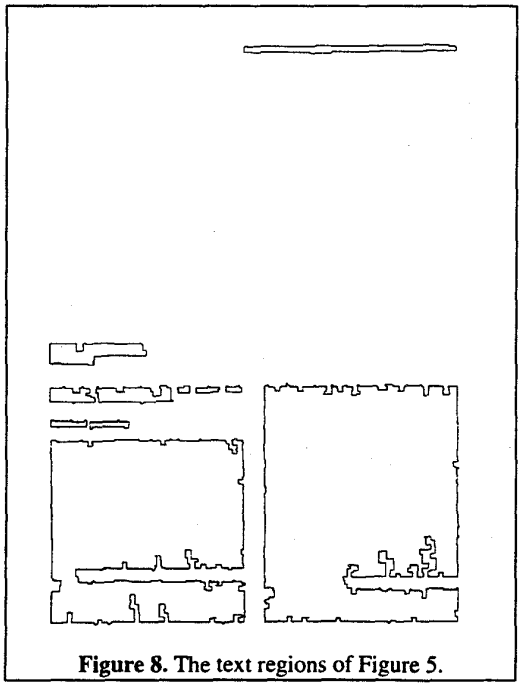

Figure 8. The text regions of Figure 5.

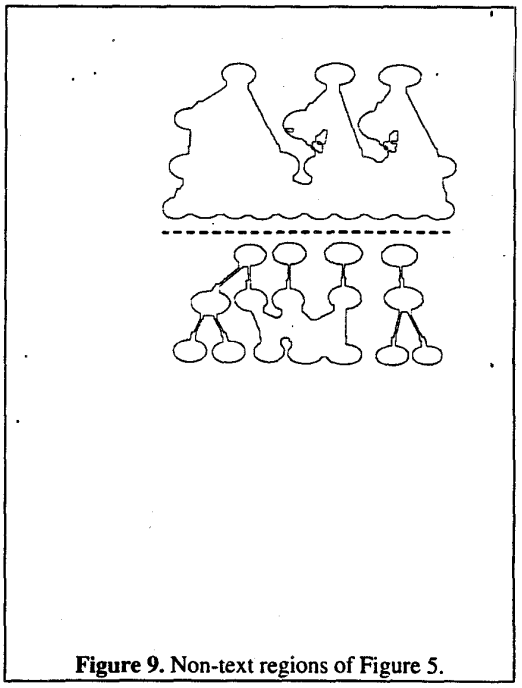

Figure 9. Non-text regions of Figure 5.