Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR'95), Montreal, Canada, August 1995, IEEE Computer Society Press, pp. 1132-1135.

Representation and Classification of Complex-Shaped Printed Regions Using White Tiles

A. Antonacopoulos and R.T. Ritchings

Systems Engineering Group, Department of Computation,
University of Manchester Institute of Science and Technology (UMIST),
P.O. Box 88, Manchester, M60 1QD, U.K.

Abstract

There is an increasingly pressing need to develop document analysis methods that are able to cope with images of documents containing printed regions of complex shapes. Contrary to the bounding-box representation used in most past page segmentation and classification approaches which assume rectangular regions, there is a need for a more flexible description which also retains most of the functionality of the representation by rectangles. In the first part of this paper, the practical considerations of describing and handling the complex-shaped regions are examined and an appropriate representation scheme is proposed. For page classification, a new approach based on the description of white space inside regions is presented. In contrast to previous page classification approaches, skewed and complex-shaped regions are handled efficiently and the features are derived with no need for time-consuming accesses of the pixel-based image data.

1. Introduction

The freedom of layout design in the document generation processes is resulting in a very fast growing number of publications containing printed regions of complex shapes. Hence, there is an increasingly pressing need to develop document analysis methods that are also able to cope with the images of these documents.

The majority of the previous approaches to page segmentation and classification were designed based on the assumption that printed regions are rectangular. This has led to a simplified region-description scheme by bounding boxes and has also facilitated the application of the various page classification algorithms. However, the transition from the recognition of rectangular blocks to that of regions of varying shapes calls for a change in the representation of the printed regions and, consequently, to the algorithms that act upon it.

A flexible and practical method for the segmentation of images of documents that contain non-rectangular regions has already been presented [1]. In the first part of this paper, an appropriate representation scheme which is both flexible and also retains most of the functionality of the representation by rectangles is outlined. In the second part, a new page classification method based on the *description* of the white space inside regions is presented.

The page classification approaches in the literature can be divided into two main categories: these that take place synchronously with the page segmentation process (classification of connected components) and those which are applied after the page image has been segmented into the areas of interest. The effectiveness of the different segmentation approaches has been discussed in [1] and an efficient page segmentation method of a global character has, accordingly, been developed. Therefore, in this paper the focus of attention will be on the category of classification methods that are applied after page segmentation, on the already identified areas of interest.

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 areas of complex 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.

Wahl et al. [2] use as features the height of the block, its eccentricity, the ratio of its black area to enclosing box area and the mean horizontal length of black runs inside the area of the block in the original image. Fisher et al. [3] use, in addition to the dimensions of the block and its black pixel density in the original image, the perimeter length, perimeter to width ratio and perimeter squared to area ratio.

0-8186-7128-9/95 $4.00 © 1995 IEEE

Wang and Srihari [4] classify blocks in newspaper images into small, medium and large letter blocks, graphics or halftones. Three textural features are used, derived by examining the occurrences of black-white pair run lengths as well as those of black-white-black combination run lengths in each block.

Pavlidis and Zhou [5] distinguish between halftone regions and text/line-drawings based on observations of the cross-correlation of each scan line in a block with the ones below it. A black pixel density criterion is used to separate text from line-drawings. Finally, Haralick et al. [6] separate text from non-text regions using first and second order run-length moments computed for the black and the white pixels in each block in the horizontal, vertical and the two diagonal directions.

In the above approaches, at least one extra access of the pixels of the regions is needed specifically for the computation of classification features. 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. Furthermore, the above methods have not been designed to classify regions with shapes other than rectangular (in the case of Pavlidis and Zhou [5] the rectangular areas may be at a skew angle).

In this paper, a new practical page classification method is presented which uses the data acquired during page segmentation. No time consuming accesses to the vast amount of image data are required. Features are derived by simple computations from the description of image regions, and that of the space inside them which is available as a by product of the segmentation process. There is also no need for skew detection nor correction at any stage, thus, making the classification process much faster. Furthermore, no assumption is imposed on the shape of printed regions, which may vary considerably.

In the following section, the idea for the analysis of document images using white tiles is outlined. In Section 3, the practical considerations of describing and handling complex-shaped regions are examined and an appropriate representation scheme is proposed. The classification features and method are presented in Section 4. Finally, some example results and a short discussion of classification issues can be found in Section 5.

2. About the white tiles approach

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 assume anything about the shapes of the regions.

A flexible way to describe the surrounding net of white space is by white tiles. Each 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 method for obtaining the white tiles and identifying the regions of interest in terms of their contours is described in [1].

Apart from the white tiles belonging to the region-delimiting streams of white space there are also those that are not used during the segmentation i.e. the white tiles inside the identified regions (Figure 3). These white tiles can be used for the classification of the segmented regions. The main principle is the exploitation of the textural characteristics of text and other areas using the white tile information in order to achieve fast classification of regions of various shapes.

The different types of regions in the image have different textural characteristics. During the processes that precede classification, the aim is to preserve and, wherever possible, enhance these distinguishing characteristics. At the first instance, the value chosen for the vertical smearing, performed before segmentation, preserves the white space between characters and between words inside text lines [1]. It should also be noted that, white tiles which are ignored during segmentation, being too narrow to be considered as part of the delimiting streams, are tagged as *'narrow'* and they provide a very useful basis for feature derivation. These tiles typically correspond to space between characters in words as well as to some instances of space between words.

3. Representation of regions

The representation of the segmented areas in a document image is a significant issue. It is important for the following reasons. Firstly, a representation scheme must be flexible enough to be able to describe regions in a variety of shapes. Secondly, it must be practical to use in terms of easy and fast access to the information about the regions or parts of them. Finally, a representation scheme must also incorporate adequate information to describe the regions.

As it was mentioned earlier, the bounding-box scheme lacks the flexibility to describe complex shaped or skewed (even rectangular) regions. In these cases, rectangles enclosing regions overlap and, therefore, may include parts of different regions. Representing the contour of a region by a polygon may be more accurate but not very efficient as it becomes more difficult to search inside the region. To calculate its area, for instance, will require time-consuming pixel accesses.

The principle of the representation scheme used here is to describe accurately the regions while retaining most

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE

of the functionality of a rectangular representation. The area of each region is divided into rectangular intervals that span the whole width of the region, each of different height, determined by the corners of the contour polygon. If the region has a bay at the top or bottom then the corresponding intervals will consist of more than one parts. The partition of a region into these rectangular intervals can be seen in Figure 1.

A one-dimensional array representation is used to hold the interval structure. There is one entry for each pixel in the vertical direction. For the start only of each vertical interval, the array entry contains information about the horizontal span of each of the parts of the interval and the position of the start of the next vertical interval. All other array entries contain just a label which is the vertical position of the start of the interval that contains them. Building this representation structure from the contour description is straightforward with no time consuming operations required. In fact, only one end point for each vertical segment only is needed.

With the above representation of the regions, the functionality of rectangles is achieved complemented by the flexibility of a polygonal description. One essential advantage is that this representation enables the easy and efficient checking for inclusion of white tiles in the regions. For this purpose no time consuming search is required. Instead, the array entry corresponding to the y-coordinate of the centroid of the tile can be accessed directly to determine whether the tile is inside the region or not. Another advantage is that the calculation of parameters of the region, like the area for example, becomes straightforward and much more efficient than accessing the pixels directly. Furthermore, if the polygonal description is required it is embedded in the representation.

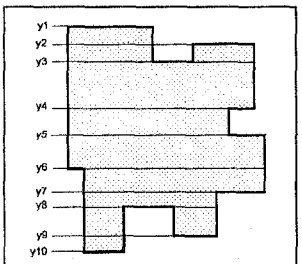

Figure 1. Partition of a non-rectangular region.

4. Classification by white tiles

The segmented areas of interest corresponding to different types of printed regions have distinct textural characteristics. These characteristics can be expressed using the information about the white tiles acquired during page segmentation. The following observations are made in terms of the white tiles inside regions:

i) *Text* regions contain a significant number of narrow (see Section 2) white tiles. These are distributed evenly inside the region and the white area covered by them is large in proportion to the total area of the region.

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

iii) *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.

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

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

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

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

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.

The following algorithm is used to classify the regions into text, graphics and line art:

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


Figure 2. Image with 15º skew.

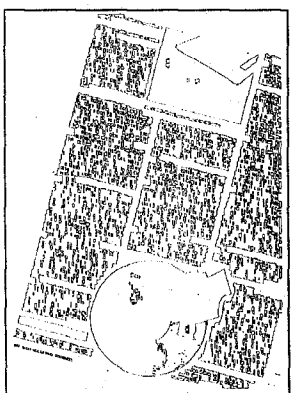
Figure 3. The white tiles inside regions.

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE

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.

5. Results and discussion

The text regions extracted from the image of Figure 2 can be seen in Figure 5, while the regions classified as graphics are shown in Figure 6. The text regions of the image in Figure 4 are illustrated in Figure 7 and the line drawing regions are depicted in Figure 8. Note that solid black regions (those containing no white tiles) are also shown in Figure 8.

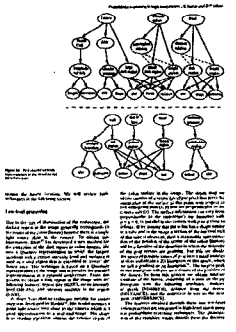

Figure 4. A document containing line art.

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. However, in a rare case where a graphics region had a very similar texture to that of text, it has been classified as text. Another type of error is when isolated characters (relatively quite large) do not contain any white space (e.g. *I, T, L,* etc.) are classified as graphics. In this case post-processing is needed to examine the context.

Overall, the page classification process is a practical one. Using an efficient region representation scheme and the 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] A. Antonacopoulos and R.T. Ritchings, "Flexible Page Segmentation Using the Background", *Proceedings of the 12th ICPR*, Vol. II, Jerusalem, Israel, October 1994, pp. 339–344.

[2] F.M. Wahl, K.Y. Wong and R.G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents", *Computer Graphics & Image Processing*, **20**, 1982, pp. 375-390.

[3] J.L. Fisher, S.C. Hinds and D.P. D'Amato, "A Rule-Based System for Document Image Segmentation", *Pattern Recognition, I, Proceedings of the 10th ICPR*, Atlantic City, U.S.A., June 1990, pp. 567-572.

[4] D. Wang and S.N. Srihari, "Classification of Newspaper Image Blocks Using Texture Analysis", *Pattern Recognition*, **47**, 1989, pp. 327–352.

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

[6] R.M. Haralick, I. Phillips, S. Chen and J. Ha, "Document Zone Hierarchy and Classification", *Proceedings of the IAPR Int. Workshop on Structural and Syntactic Pattern Recognition (SSPR'94)*, Nahariya, Israel, October 4–6, 1994.

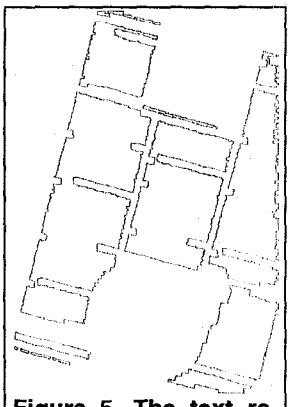

Figure 5. The text regions of Figure 2.

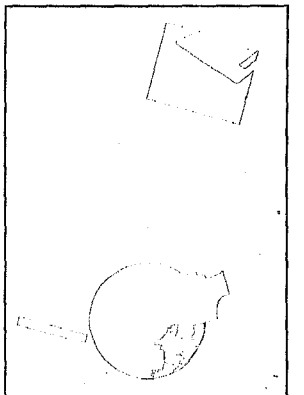

Figure 6. The graphics regions of Figure 2.

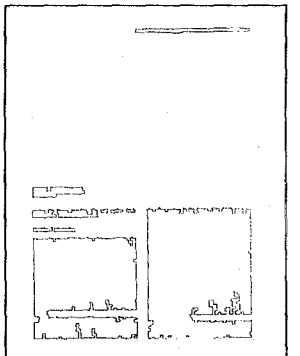

Figure 7. The text regions of Figure 4.

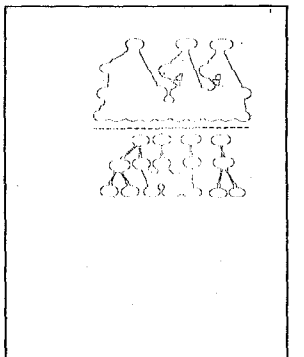

Figure 8. The line art regions of Figure 4.

Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR '95)
0-8186-7128-9/95 $10.00 © 1995 IEEE