Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR'99), Bangalore, India, August 1999, IEEE Computer Society Press, pp. 193-196.

Region Description and Comparative Analysis Using a Tesseral Representation

A. Antonacopoulos[1] and F.P. Coenen
Department of Computer Science, University of Liverpool,
Peach Street, Liverpool, L69 7ZF, United Kingdom

Abstract

This paper presents a region-representation scheme and comparative analysis methods based on a tesseral addressing system. The proposed scheme is described in the context of performance analysis of Page Segmentation methods, a Document Image Analysis area that is particularly sensitive to both a successful region-description scheme and efficient methods for comparative analysis. The proposed tesseral representation is more economical in storage than other Cartesian-based approaches and can be advantageous for comparative analysis.

1. Introduction

Region Description schemes play a very important role in Document Image Analysis and Understanding. In particular, an accurate description that enables efficient processing of regions is critical to Layout Analysis [1].

A good description scheme should be able to precisely represent regions in complex situations and, at the same time, include enough region information to enhance the efficiency of a method operating on them.

A pixel-based description (e.g. set of pixels with same label) is the ultimate in detail. However, a large number of time-consuming image accesses are required for most operations on regions, making this kind of description unattractive for most tasks. For the vast majority of applications, a higher-level description is required.

Bounding rectangles are by far the most widely used means of describing regions. These can be described by the coordinates of only two diagonally opposite vertices, rendering the comparison of regions (e.g. to establish spatial relationships) straightforward. However, although certain regions (especially smaller ones, such as individual characters and words) can be described reasonably accurately by rectangles, the presence of skew and complex-shaped regions require more elaborate descriptions. To overcome these problems, polygonal representations have been used [2]. All these description schemes are, traditionally, based on the Cartesian coordinate system.

This paper presents a region-representation scheme and comparative analysis methods based on a *tesseral* addressing system. It is argued here that the proposed tesseral representation is more economical in storage and can be advantageous for comparative analysis. The new tesseral representation scheme is presented in the context of analysing the performance of Page Segmentation algorithms, where region description and efficient comparative analysis are crucial.

The characteristics of the performance analysis problem and the proposed framework are examined in the following section. Section 3 defines the context within which the evaluation is performed. The basics of the tesseral addressing system and its application to the representation and comparative analysis of regions are presented in Section 4. Finally, the paper concludes with a discussion of the issues pertaining to the use of the tesseral approach in Section 5.

2. Performance analysis of Page Segmentation methods

Large-scale objective evaluation is essential not only for OCR [3] but for all subsystems involved in DIA, such as the identification of regions of interest in the document page image (*page segmentation*). This is a significant stage that seriously affects the performance of subsequent DIA stages (e.g., OCR, Document Image Understanding).

The analysis of the performance of Page Segmentation methods is not straightforward since regions, not characters, are to be compared (the ground-truth description with the results of page segmentation). A string-matching approach based on a comparison of ground-truth characters with the results of OCR applied to the segmented document image has been proposed [4]. Although this approach will give a global score on the performance of a complete OCR system, it does not provide sufficient information for page segmentation algorithm developers.

Methods that perform region comparisons can be divided in two categories: those that use a *pixel-based* description and those that use a *geometric* one. In the first category, a method developed at Xerox [5] creates region-maps (two reduced-resolution images) to represent the

---

[1] Corresponding author (e-mail: aa@csc.liv.ac.uk)

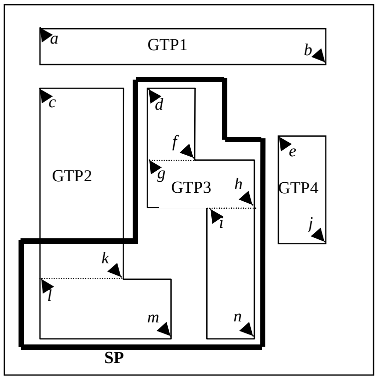


Figure 1. Example of GTPs and a SP (emphasised line).

ground-truth and the segmentation regions. Overlaps between the two types of regions can be identified by scanning the region maps. Although this is a flexible approach, it requires two more instances of an image (albeit reduced) and building region maps in run-time requires time-consuming pixel-level image accesses.

In contrast, approaches based on geometric comparison are more efficient, as they require less memory to describe the image and do not require pixel-level image accesses. However, the geometric comparison approach is not straightforward as, to be successful, there must be a very accurate description of regions without excess background and the region-representation schemes of the result and the ground-truth must be directly comparable.

The developers of the University of Washington (UWASH) document database [3] have made provisions for ground-truth region-description using bounding rectangles. As was previously mentioned, rectangles are efficient for storage and comparative analysis but, while many types of documents have rectangular regions, any evaluation approach based on this database will not be applicable to methods dealing with complex layouts [1].

A new performance analysis and evaluation framework is under development at the University of Liverpool, focusing mainly on Layout Analysis subsystems [6]. The framework is designed to enable the evaluation of algorithms under an increased number of significant conditions that were not catered for under past approaches. Such conditions include complex layouts with non-rectangular regions, colour and textured backgrounds and non-uniform region orientation. Furthermore, the evaluation methods can provide information at various levels of detail. At the local level, detailed information is available for each region on a number of conditions. This detailed information is aimed at the algorithm developers. At the global level, information is available for the performance of an algorithm on a whole page or set of pages. A global score is also given at this point for end-users to compare and contrast different algorithms.

Central to the new framework is a very flexible region representation scheme using isothetic polygons (having only horizontal and vertical edges). The flexibility is due to the ground-truth polygons describing regions very accurately without excess background space. The White Tiles page segmentation method [1], which can identify and describe regions very accurately even in the presence of complex layouts and severe skew, is used as a first stage of the ground-truthing process. With a small number of point-and-click operations to correct the results the final ground-truth polygons are obtained. The description of each region resulting from a page segmentation method under consideration (e.g., set of bounding boxes) is also converted into a minimum-enclosing isothetic polygon.

With respect to the evaluation of Page Segmentation, a new methodology for description-based comparative analysis of regions has been proposed [6]. In that approach, for efficient representation and comparative analysis, each region polygon is represented as a set of horizontal rectangular intervals. All the regions on a page are represented in a global interval structure. The global interval structure is very efficient for direct comparison of regions as it approximates the efficiency of rectangular representation and, at the same time, it retains the accuracy of the polygonal description [6]. In the pursuit of efficiency, alternative representation schemes have also been investigated. Before proceeding to describe the tesseral approach of this paper, the general context in which it operates is explained below.

3. Region representation and comparative analysis

A region is defined here to be the smallest logical entity on the page. For the purpose of assessing Page Segmentation and Classification, a region is a single paragraph, in terms of text (body text, header, footnote, page number, caption etc.), or a graphic region (halftone, line-art, images, horizontal/vertical ruling etc.). Composite elements of a document, such as tables or figures with embedded text, are considered each as a single (composite) region.

For ground-truth description, each region is represented by the closest-fitting isothetic polygon around it, a *ground-truth polygon* (*GTP*). The regions resulting from the application of a page segmentation method are referred to as *segmentation polygons* (*SP*). Figure 1 shows an example of GTPs and a SP (the horizontal divisions in the polygons will be explained later).

The goal of the comparative analysis is to identify, given the GTP and SP description, the following situations (or combinations thereof): (1) a SP correctly describes a GTP, (2) a GTP is split, (3) a GTP is merged with other GTPs, (4) a GTP is partially missed, (5) a GTP is totally missed, (6) A SP does not describe any GTP (wrongly introduced by page segmentation, possibly due to noise in the image). For more information on the treatment of these cases by the proposed framework see [6].

4. The tesseral method

Tesseral addressing is a method for representing space by repeatedly dividing a given space into isohedral sub-spaces until some predefined resolution is reached [7].

The representation proposed here has similarities with existing forms of tesseral addressing (such as quad-tesseral addressing [8]) in the sense that it is a mechanism for addressing spaces decomposed into isohedral subspaces. However, this representation assumes a predefined resolution—there is no explicit inclusion of the concept of hierarchical decomposition. A more detailed explanation of the derivation of the spatial representation used here can be found in [9].

4.1 Tesseral description of regions

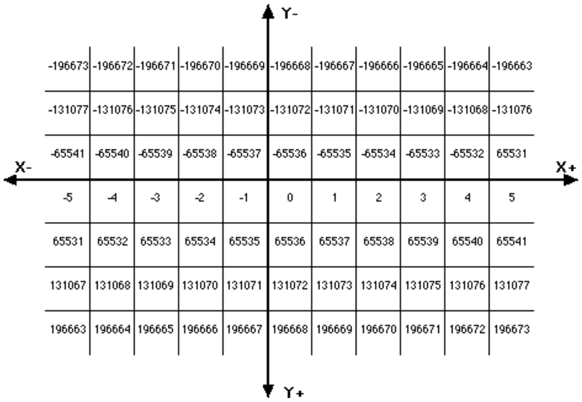
From the foregoing, the regions in a document image can be described as isothetic polygons, each described in terms of vertices (typically defined as X-Y coordinate pairs). In the tesseral representation, however both X and Y coordinates are compressed into a single integer referred to as a *tesseral address* (or simply an *address*. We assume that this integer is a 32 bit signed integer where the least significant 16 bits are allocated to the X coordinate and the remaining 16 bits to the Y coordinate.

Using this representation we define a maximal space (the *address space*) measuring 2^{16} by 2^{16} (the origin address is in the centre), the positive (bottom-right) quadrant of which is used to describe the document. The 32-bit resolution is adequate for most document image analysis applications as an A4 size document presented at a resolution of 300 dpi is only about 2500x3500 pixels.

Given any X-Y coordinate pair the associated address a $(-2147483647 \leq a \leq +2147483647)$ can be calculated using the formula: $a = x + 65536y$. To convert an address back to a set of coordinates the two halves of the 32-bit integer are separately extracted e.g., using appropriate 'bit-shifts.' A small section of the numbering system is given in Figure 2. Note also that the origin has the address 0 and that the numbering is symmetric about this origin.

The main advantage of the representation is that it provides a "left-to-right" linearisation of space (as shown in Fig. 2). This in turn offers advantages of efficient data storage, using ideas founded on "run-length" encoding strategies, and computationally effective comparison and translation of sets of addresses (using straightforward integer addition and subtraction). In addition to the above there are further advantages of this representation over other tesseral ones [9].

4.1.1. An example document. Referring back to Fig. 1, the ground-truth (GT) document is described by 4 GTPs, some of which can be defined as single *rectangles* others as a series of rectangles. Either horizontal or vertical decomposition can be used to describe an isothetic polygon using rectangles (whichever gives the minimum number of rectangles).



-196673	-196672	-196671	-196670	-196669	-196668	-196667	-196666	-196665	-196664	-196663
-131077	-131076	-131075	-131074	-131073	-131072	-131071	-131070	-131069	-131068	-131076
-65541	-65540	-65539	-65538	-65537	-65536	-65535	-65534	-65533	-65532	65531
-5	-4	-3	-2	-1	0	1	2	3	4	5
65531	65532	65533	65534	65535	65536	65537	65538	65539	65540	65541
131067	131068	131069	131070	131071	131072	131073	131074	131075	131076	131077
196663	196664	196665	196666	196667	196668	196669	196670	196671	196672	196673

Figure 2. Fragment of address space indicating cell numbering.

Each rectangle is described by a pair of opposing corner pixels. In this case each corner is represented by a single address. The lowercase letters in Fig. 1 give a sequence of tesseral addresses where the alphabetic ordering describes the linearisation of the addressing system: *{a, b, c, d, e, f, g, h, i, j, k, l, m, n}*.

Each rectangle is then defined using a '..' infix operator whose prefix operand is the address nearest the origin and whose postfix operand is the address furthest away from the origin. Thus the GTPs given in Fig. 1 may be described as follows: **GTP1:** *{a..b}*, **GTP2:** *{c..k, l..m}*, **GTP3:** *{d..f, g..h, i..n}* and **GTP4:** *{e..j}*. The GT document can be described in terms of a set of sets as follows: *GT = { {a..b}, {c..k, l..m}, {d..f, g..h, i..n}, {e..f} }* i.e. the entire GT document is reduced to a small number of integers.

4.2 Comparative analysis

From the above it is evident that both GTPs and SPs can be encoded in terms of sets of tesseral addresses. To compare a GT document with a segmentation (S) document two set functions are required, $f_{numIntersection}$ and $f_{numSuperset}$. Both take two arguments, the first is a set P describing a polygon taken from one of the documents and the second is a set of sets I describing the entire other document. Both functions return a natural number. In the first case this describes the number of times that P intersects with some element in I (each element is a set describing a polygon in I). In the second case the result is the number of times that P is a superset of some element in I. The two functions are defined as follows:

$$Dom(f_{numIntersection}) = \{\langle P, I \rangle\}$$

$$Cod(f_{numIntersection}) = Nat$$

$$Gr(f_{numIntersection}) = \left\{ \begin{array}{l} \langle P, I, n\rangle \mid n = f_{cardinality}(K), \\ K = \{k \mid k = i \cdot \forall i \in I \Rightarrow f_{intersection}(P, i) = True\} \end{array} \right\}$$

$$Dom(f_{numSuperset}) = \{\langle P, I\rangle\}$$

$$Cod(f_{numSuperset}) = Nat$$

$$Gr(f_{numSuperset}) = \left\{ \begin{array}{l} \langle P, I, n\rangle \mid n = f_{cardinality}(K), \\ K = \{k \mid k = i \cdot \forall i \in I \Rightarrow f_{superset}(P, i) = True\} \end{array} \right\}$$

The set *Nat* represents the set of natural numbers, $f_{cardinality}$ returns the number of elements in its argument (a set), $f_{intersection}$ returns *true* if its two arguments intersect, otherwise the function returns *false* (note that both arguments are sets of tesseral addresses). $f_{superset}$ returns *true* if its first argument is a superset of the second argument, otherwise the function also returns *false*.

Using the above, to compare a S document with a GT document the method proceeds as follows:

1. For each polygon *p* in S find N_s and N_i where $N_s = f_{numSuperset}(p, GT)$, $N_i = f_{numIntersection}(p, GT)$. Then compare N_s against N_i using the SP to GTPs comparison matrix presented below.
2. Check thecount *k* of how many times each GTP has been encountered by an SP is kept and if *k = 0* the GTP has been missed.

The SP to GTPs comparison matrix referred to above has the following form:

	$N_s = 0$	$N_s = 1$	$N_s > 1$
$N_i = 0$	"invented" SP	—	—
$N_i = 1$	Partial GTP miss	Match	—
$N_i > 1$	Partially missed and merged all GTPs involved	Partial miss and merge	if $N_s = N_i$ merge else merge and partial miss

5. Discussion and conclusions

The tesseral approach can describe ground-truth and segmentation regions (isothetic polygons) in terms of a set of rectangles. Each rectangle is represented by two diametrically opposite corner addresses.

The linearisation of space allows for computationally efficient comparison of sets of addresses. In addition the algorithms used (not included in this paper) operate without knowledge of the partitioning—horizontal, vertical or a mixture of the two.

Finally, the arithmetic supported by the representation allows for the computationally efficient calculation of area (useful for assessing accuracy of description by page segmentation)

With regard to the global interval representation proposed in [6] there is a similarity in the fact that in both approaches regions are described as sets of rectangles (horizontally aligned in [6]). Both approaches are considerably more efficient than the pixel-based description [5], however the implementation of the tesseral representation requires less storage than the global interval one. In terms of the efficiency of comparative analysis, both the tesseral and the global interval approaches compare favourably to the time-consuming pixel-based image accesses. The global interval approach however requires fewer operations to establish the correspondence between ground-truth regions and segmentation ones.

This paper has presented a new approach to region representation. This approach uses a tesseral addressing system to represent regions (isothetic polygons) efficiently as sets of rectangles. Preliminary results show that the tesseral approach is successful in the analysis of the performance of page segmentation methods. This approach compares favourably to Cartesian pixel-based region representations and requires less storage than other region descriptions.

References

[1] A. Antonacopoulos, "Page Segmentation Using the Description of the Background", *Computer Vision and Image Understanding*, Vol. 70, No. 3, 1998, pp. 350-369.

[2] A. Antonacopoulos and R.T. Ritchings, "Representation and Classification of Complex-Shaped Printed Regions Using White Tiles", *Proc. 3rd Int. Conf. Doc. Anal. and Recog. (ICDAR'95)*, Montreal, 1995, Vol. 2, pp. 1132-1135.

[3] I.T. Philips, S. Chen, J. Ha and R.M. Haralick, "English Document Database Design and Implementation Methodology", *Proc. 2nd Annual Symp. Doc. Analysis and Retrieval*, UNLV, USA, 1993, pp. 65–104.

[4] J. Kanai, S.V. Rice, T.A. Nartker and G. Nagy, "Automated Evaluation of OCR Zoning", *IEEE Trans. Patt. Anal. Machine Intell.*, Vol. 17, No. 1, January, 1995, pp. 86-90.

[5] B.A. Yanikoglu and L. Vincent, "Pink Panther: A Complete Environment for Ground-Truthing and Benchmarking Document Page Segmentation", *Pattern Recognition*, Vol. 31, No. 9, 1998, pp. 1191-1204.

[6] A. Antonacopoulos and A. Brough, "Methodology for Flexible and Efficient Analysis of the Performance of Page Segmentation Algorithms", *Proc. 5th Int. Conf. Doc. Anal. and Recog. (ICDAR'99)*, Bangalore, India, 1999.

[7] S.B.M. Bell, B.M. Diaz, F.C. Holroyd and M.J.J. Jackson, "Spatially Referenced Methods of Processing Raster and Vector Data", *Image and Vision Computing* **1**(4), 1983, pp. 211–220.

[8] H. Samet, "The Quadtree and Related Data Structures", *ACM Comput. Surveys,* **16** (2), 1984, pp. 187-260.

[9] F.P. Coenen, B. Beattie, M.J.R. Shave, T.J.M. Bench-Capon and B.M. Diaz, "Spatial Reasoning Using The Quad-Tesseral Representation", *Artificial Intelligence Review*, Vol. 12, 1998, pp. 321–343.