9-18-2011

Extending Page Segmentation Algorithms for Mixed-Layout Document Processing

Amy Winder
Boise State University

Tim Andersen
Boise State University

Elisa Barney Smith
Boise State University

 DOI: 10.1109/ICDAR.2011.251

DOI: 10.1109/ICDAR.2011.251

Extending Page Segmentation Algorithms for Mixed-Layout Document Processing

Amy Winder
CS Department
Boise State University
Boise, ID, USA.

Tim Andersen
CS Department
Boise State University
Boise, ID, USA.

Elisa H. Barney Smith
ECE Department
Boise State University
Boise, ID, USA.

***Abstract*—The goal of this work is to add the capability to segment documents containing text, graphics, and pictures in the open source OCR engine OCRopus. To achieve this goal, OCRopus' RAST algorithm was improved to recognize non-text regions so that mixed content documents could be analyzed in addition to text-only documents. Also, a method for classifying text and non-text regions was developed and implemented for the Voronoi algorithm enabling users to perform OCR on documents processed by this method. Finally, both algorithms were modified to perform at a range of resolutions.**

Our testing showed an improvement of 15-40% for the RAST algorithm, giving it an average segmentation accuracy of about 80%. The Voronoi algorithm averaged around 70% accuracy on our test data. Depending on the particular layout and idiosyncracies of the documents to be digitized, however, either algorithm could be sufficiently accurate to be utilized.

***Keywords*- page segmentation, RAST, Voronoi, open source OCR**

I. Introduction and Background

Numerous historical documents in book and other forms have yet to be digitized. Historical books can be too fragile to be scanned, but today's inexpensive digital cameras can produce images comparable in quality and resolution to those generated by a flatbed scanner. Thus, it is now feasible to safely and cheaply digitize historical documents.

Amy Winder is now with Hewlett Packard, Boise.
Contact: EBarneySmith@BoiseState.edu

Once digitized it is desirable to convert document images into text documents for readability and searchability. A first step is to analyze the image to determine which areas are text and which are not, so that only text regions are sent to the OCR engine. This process is called page segmentation. Page segmentation algorithms can be categorized as top-down, bottom-up or hybrid methods [1]. Top-down methods involve operating on the document as a whole and subdividing it, whereas bottom-up methods start at the pixel-level and recursively merge constructs into segmented regions. Hybrid methods may include a little of both.

The Recursive X-Y Cut (RXYC) and Run-Length Smearing Algorithms (RLSA) are top-down methods. RXYC[2] uses vertical and horizontal projections of the binarized image where the white areas correspond to low elevations and the black areas to high elevations. Valleys in the projections then delineate candidate segmentations. The algorithm recursively subdivides the document around the largest valley(s), maintaining the data in a structure called an X-Y tree. RLSA[3] operates like RXYC, but classifies the regions as well. It examines each of the pixels in a row-by-row and column-by-column fashion and changes each white pixel to black if it is surrounded by enough black pixels, after which the generated row and column bit maps are ANDed together to form a single bit map. This is then smoothed horizontally to connect words in text lines. Block features (such as numbers of black and white

DOI: 10.1109/ICDAR.2011.251

pixels, etc.) determine block classification.

OCRopus' version of RAST [4], [5] was designed for text-only documents and consists of three steps: finding the columns, finding the textlines, then determining the reading order. It finds columns using a whitespace rectangle algorithm [6] similar to RXYC. The largest whitespace rectangles (covers) delimited by the connected components of the image are determined and sorted by how many connected components touch each major side. Covers are then merged iteratively as long as the combined cover obeys a given rule on how many components are incident upon it. Reading order is determined by considering pairs of lines such that either the line below or the line to the right at the top of the page (e.g. in the next column) goes next, followed by sorting these pairs to give the final reading order.

The Voronoi method [7] is a bottom-up approach that extracts sample points along the boundaries of the connected components to construct a Voronoi point diagram, which initially creates a large number of superfluous edges. Edges are deleted based on shortness and whether they are connected to other lines, converting the diagram to an area Voronoi diagram representing regions.

This paper presents improvements to the RAST and Voronoi segmentation algorithms found in OCRopus. Before improvement, RAST was not able to accurately segment and determine reading order of mixed content documents. Voronoi tended to oversegment documents that contained images and did not classify regions as text or non-text so could not be used for OCR. While a number of methods have been proposed for region classification [8], [9], [10], we developed a simple yet robust approach for classification of voronoi regions. In addition, both segmentation algorithms were modified to remove hard coded resolution dependent parameter settings. Sec. II describes RAST and our improvements. Sec. III likewise describes the Voronoi algorithm and its enhancements. Sec. IV gives empirical results for the original and enhanced RAST and Voronoi performance. Performance of a commercial system is also given for comparison. The paper concludes in Sec. V.

II. Mixed-Content RAST Algorithm

While the RAST algorithm performed well on text-only documents, it was not designed to process documents containing images and graphics or documents that were scanned or photographed at different resolutions. To overcome these limitations, the authors leveraged information extracted from the most commonly occurring object of the document: the letters. RAST begins by extracting the connected components of the document, which consist of letters and isolated components of the images and graphics. Since letters make up the bulk of the components, statistics are gathered for them, such as their heights and widths. These dimensions are then used to determine the columns, the text lines and to process the remaining connected components for segmentation and classification.

The orginal version of OCRopus gathered these statistics; however, we found that the dimensions obtained from them were often erroneous. Therefore, a method was developed to refine this data to extract the height and width of the characters more reliably. The main problem with the original method was that it was not taking into account the noise of the data (i.e. puncuation marks like periods, commas, and so on). So, it was setting the height of the letters smaller than they actually were, which negatively impacted the rest of the process.

The histogram shown in Figure 1a contains the heights of a series of connected components sampled along a horizontal scan of the document. Applying a Gaussian smoothing function to the histogram reduces the number of peaks dramatically. After the second round of smoothing the histogram has three clearly defined peaks, which correspond to puncuation marks, the heights of short letters and the heights of tall letters, Figure 1c. By selecting the rightmost peak, the height of a text line can be accurately determined.

RAST was further modified to better handle image and graphical data. Since images can contain

DOI: 10.1109/ICDAR.2011.251

components much smaller or larger than letters, these objects are retained, merged and classified as non-text. Also, images and graphics can contain text or components with sizes similar to letters, which means that previously defined text lines lie within non-text regions. So, these components are merged with their overlapping non-text regions and reclassified as non-text. Similarly, graphics regions can overlap each other, so they are merged as well.

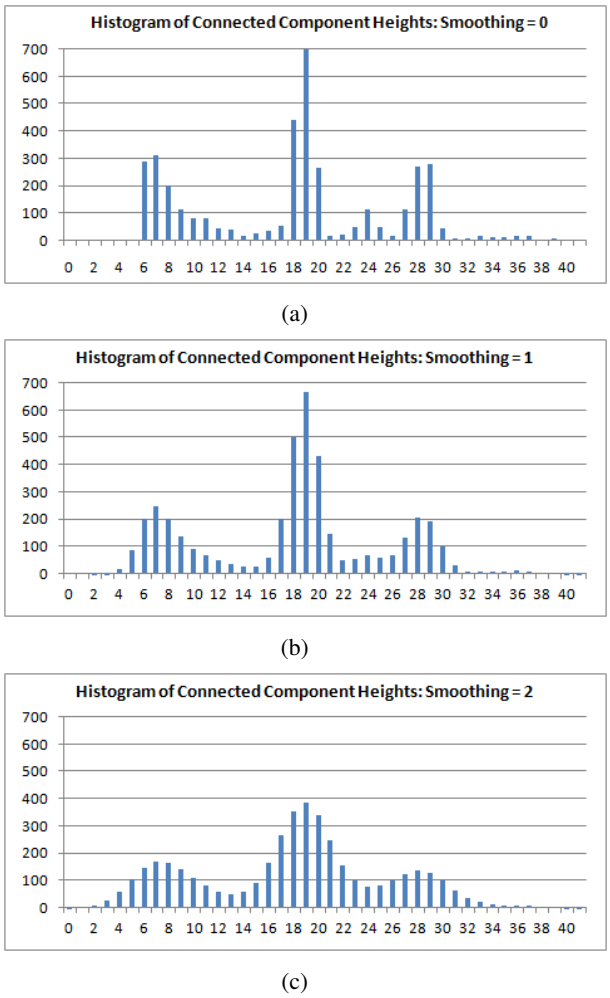
Finally, small, isolated non-text regions are morphologically closed (rectangularly dilated, merged then eroded using a fraction of a text line height) to combine them. Since all of this processing can result in new non-text areas, it is continued until no new areas are created. This ensures that the text and non-text regions are complete and distinct.

III. Voronoi Page Segmentation with Classification

The Voronoi algorithm starts out the same way as RAST by extracting connected components to determine the heights and widths of the letters. The algorithmic improvement in RAST were leveraged to more accurately calculate the height of the text line for use later in processing the zones. The Voronoi algorithm was extended to include classification of zones, merging of non-text zones, and clean up of overlapping non-text regions ("zones" corresponds to geometries created by Voronoi and "regions" corresponds to page segments).

While a number of methods have been proposed for region classification [8], [9], [10], a simple yet robust approach was developed to classify the voronoi zones. The classification algorithm concentrates on the connected components, which have been labeled as characters. First, it finds their locations along the y-axis and uses this information to identify text lines. This is done with a technique similar to the one used for determining the character dimensions, except in this case the y-values of letters extending below the line need to be ignored rather than puncuation marks.

Once the text lines have been found the density of the characters along it is determined by dividing the summed width of the character boxes along the line by the length of the line itself, which is derived from their locations. If the density of the line exceeds 80% it is classified as text; otherwise, it is classified as non-text. Then if the number of lines in a zone exceeds 50% the zone is classified as text.



(a)

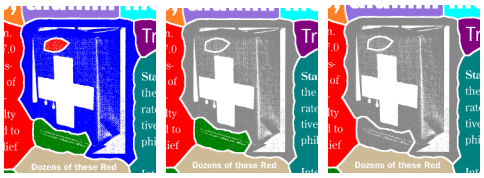
(b)

(c)

Figure 1. Frequency of occurance of connected component heights (a) before smoothing, (b) after first round of smoothing and (c) after second round of smoothing.

The next step in the process is to merge the non-text zones since Voronoi segmentation typically oversegments them due to the existence of large white spaces in images and graphics as shown in Figure 2. This is done by identifying the perimeters of the zones then selecting one of the non-text zones randomly and exploring its neighbors by

DOI: 10.1109/ICDAR.2011.251

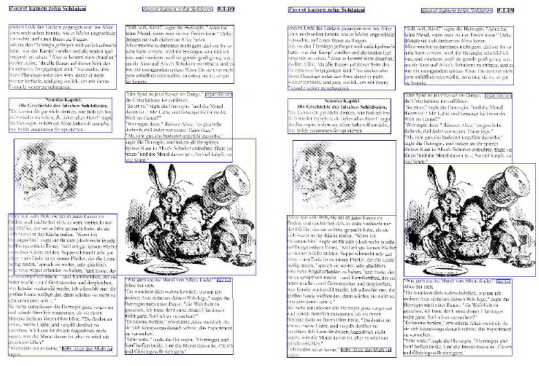


(a) (b) (c)

Figure 2. Progress in merging graphic regions.

crossing its perimeter. Once a non-text neighbor is found it is labeled the same as the original zone and its neighbors are examined. When all of the original zone's non-text neighbors have been relabeled, a non-text zone that has not been visited, yet, is selected and the process is started over again. This continues until all non-text zones have been considered.

Once the zones have been classified as either text or non-text they are converted into regions using the values of the outermost pixels (i.e. north-east, southeast, etc.). While zones may not overlap, the regions often do as shown in Fig. 3. This type of issue is addressed by deleting text regions completely overlapped by non-text regions; re-classifying text regions, which completely overlap non-text regions as non-text (and deleting the over-lapped non-text region); spliting non-text zones, which cross column dividers and segmenting text regions, which partially overlap non-text regions.



(a) (b)

Figure 3. (a) Pictures in two different columns are merged. (b) Merged graphics zone is broken in two and text overlaps removed.

IV. Testing and Analysis

Performance was evaluated on a collection of 320 document images comprising eight different types (i.e. single column, double column, etc.) and a range of resolutions. The collection contains 200 hand made documents written in the Times New Roman 12 point font saved at four different resolutions (100, 200, 300 and 600 DPI). The documents contain the following layouts: single column text only (10x4), double column text only (10x4), single column text with half-tone images (10x4), double column text with half-tone images (10x4) and a mixture of single and double columns with half-tone images (10x4). The rest of the data set includes 80 pages of technical journals which contain graphs, figures, tables and a title/abstract combination (20x4) and 40 pages taken from magazines (10x4).

Ground truth XML files were generated for each of the documents from the TIFF files using True-Viz [11]. We also modified OCRopus to output its segmentation in XML format, and created a software utility for comparing the detected regions for each algorithm to the ground truth following the method used in the ICDAR page segmentation competitions [12], [13]. Following the testing of the improved RAST and Voronoi algorithms, ABBYY's FineReader OCR package was evaluated to see how well a commercial program could analyze these types of layouts.

A. RAST Algorithm Improvements

The performance improvement was assessed by comparing the original and improved algorithms on the test images. The output was compared to the ground truth using the metric from the ICDAR page segmentation contest with $w_1 = w_3 = w_4 = w_5 = 1$ and $w_2 = w_6 = 1.12$. The average for each class is plotted as a function of resolution in Fig. 4. The single, double and mixed column pages without half-tone images were segmented fairly accurately from 80-100%. However, for documents containing pictures the performance level peaked between 30-60% at 100 DPI then dropped at higher resolutions. The improved RAST algorithm not only displays better performance at 100 DPI,

DOI: 10.1109/ICDAR.2011.251

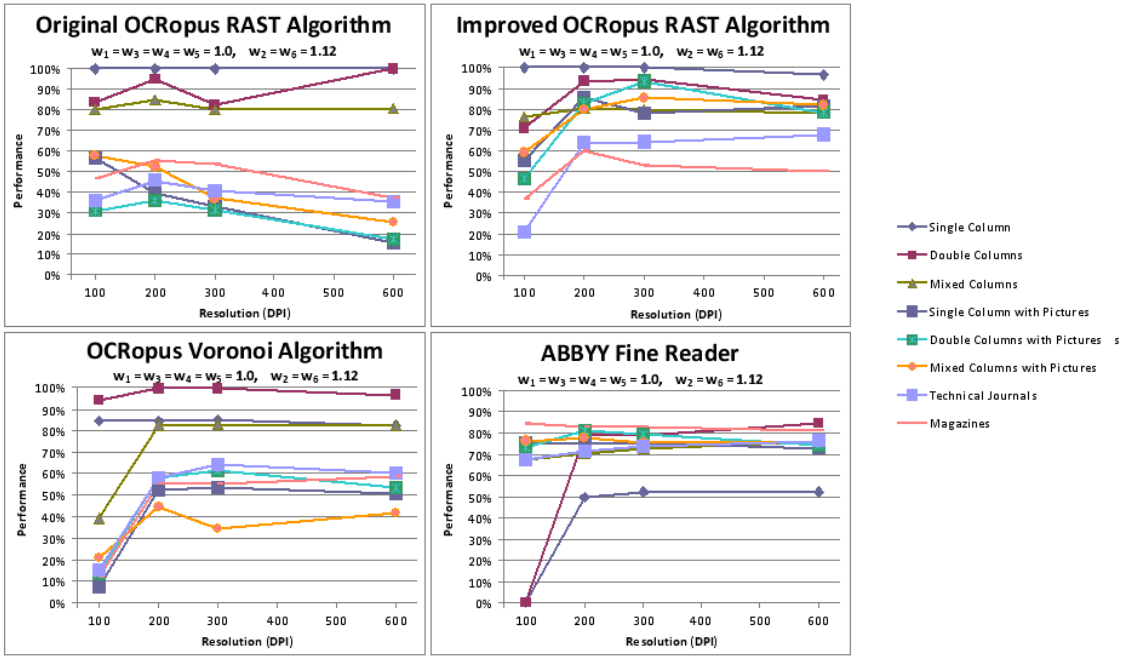


Figure 4. Performance: (a) Original RAST (b) Improved RAST algorithm (c) Improved Voronoi algorithm (d) ABBYY.

but better performance at higher resolutions as well. Single, double and mixed column documents with half-tone images show the most improvement from 30-60% to 80-90%. The segmentation of the technical documents improved on the order of 25% from approximately 40% to 60-70%. The magazine class improved from 50% to 65%.

While the improved RAST algorithm performed better than the original, three types of errors emerge. The first one is the oversegmentation of text regions. This typically happened in areas where one text line was either much shorter or slightly longer than its neighboring text lines. The second type of error was the merging of text regions, which tended to occur with short columns. The reason why short columns were merged is because the function to find white spaces, some of which are later turned into column separators, examines their aspect ratios and rejects those below a certain threshold. So, short columns are not separated by white space covers. This could be fixed by reducing the expected aspect ratio. The last type of error involved merging text and non-text regions. This occurred in three different cases: when text wrapped around the figure in a non-linear fashion, when the column was very narrow, and when non-text was incorrectly detected in text regions.

B. *Voronoi Algorithm*

Since the original Voronoi algorithm in OCRopus does not classify zones, it was not possible to assess its segmentation accuracy. Fig. 4c shows the segmentation accuracy of the extended algorithm. The accuracy of text-only document segmentations averaged 90%; whereas, those of documents containing images and graphics averaged closer to 50

Examining the results of the document classes that included half-tone images, three types of errors dominate: one can be attributed to the data, another to Kise's Voronoi algorithm, and the third to the text classification algorithm. Starting with the first, a number of the documents contain half-tone images in very close proximity to text. For documents scanned at a resolution of 300 DPI, Kise's Voronoi algorithm failed to separate the images from text when they were separated by 23 or fewer pixels. The height of a tall letter

DOI: 10.1109/ICDAR.2011.251

at this resolution is 28 pixels, so if the image was positioned within this distance it might not be placed into its own region. After the Voronoi regions were defined it was impossible for the extension of the algorithm to further segment and classify them correctly.

The most frequently occurring zoning error is the oversegmentation of text. This can be seen in titles, headers, footers and occasionally in parts of outlying sentences in paragraphs. This problem relates more to reading order than region classification.

The third issue identified was that some text, namely italicized and bold text, tended to be classified as images rather than text. This was due to the fact that the bounding boxes of the characters overlapped and were not considered as text. This had substantial impact on performance only if the entire block was italicized.

C. Commercial Package

ABBYY's Fine Reader Engine 9.0 was used as a base comparison of overall performance. Fig. 4d shows its performance. For the most part, the performance is between 70% and 85% for all resolutions. Neither RAST nor Voronoi were able to segment as accurately at 100 DPI. Not only does Fine Reader have a flatter response as a function of resolution, but it also has a tighter response than RAST and Voronoi in that it performs equally well on all of the classes.

There were a couple of anomalies, though. At 100 DPI for the single column and double column classes the performance dropped to zero. This was because the regions were classified as pictures rather than text. Also, the single column class only performed at approximately 50% throughout the range of resolutions since the paragraphs were broken into individual regions, but were only represented by one region in the ground truth. Examining the output, the predominant error appeared to be overlapping regions, which depending on how you define the ground truth might not be considered an error.

V. Conclusions

We have modified the RAST and Voronoi segmentation algorithms in OCRopus to enable processing of mixed-content layouts at a variety of resolutions, making the digitization of standard format historical documents by low-budget organizations feasible. We tested the improved algorithms on a set of test images, with RAST showing an improvement of 40% on the hand-made documents with half-tone images, the technical document class 25% and the magazine class 15%, resulting in final overall accuracies of 90%, 65% and 55%, respectively.

The primary errors were caused by the oversegmentation of text areas with unusually long or short text lines, the merging of short columns due to the constraints used for the definition of column dividers, the merging of text and non-text regions in non-Manhattan layouts, and very narrow columns and stylized text being classified as non-text.

The performance of the Voronoi algorithm was similar to RAST for the text-only documents, but was lower for the documents containing half-tone images, graphs and tables. The double column text-only class fared the best at 95% and the double and mixed column text-only classes at 85%. The mixed column with half-tone images class was only segmented with an accuracy of 40%. The rest of the classes performed between 50% and 65%.

While the RAST and Voronoi algorithms performed well, there remain areas in which they could be improved. The robustness of RAST could be increased so that it can process text lines of varying widths as well as short and/or narrow columns. The processing of stylized text and, for Voronoi, italicized and bolded text, could also be improved. Also, the Voronoi algorithm could be enhanced by merging segmented titles and classifying italicized text properly.

In conclusion, the improved open-source RAST algorithm compares well to a widely used com-

Code and data are publically available at http://coen.boisestate.edu/EBarneySmith/SP_lab.

DOI: 10.1109/ICDAR.2011.251

mercial program in the case of documents that contain half-tone images rather than graphs and tables. The Voronoi algorithm did not perform as well as Fine Reader (by approximately 20%), but if the documents contain ample space between the figures and text, and there is no italicized or bolded text, it can perform adequately. Therefore, depending on the type of layout being digitized, either algorithm could potentially be employed.

REFERENCES

[1] Shi, Z. and Govindaraju, V. "Dynamic Local Connectivity and Its Application to Page Segmentation." *Proc. ACM Hardcopy Document Processing (HDP-04)*, Nov. 2004, pp. 47-52.

[2] Nagy, G., Seth S. and Viswanathan, M. "A Prototype Document Image Analysis System for Technical Journals." *Computer*, vol. 25(7), Jul. 1992, pp. 10-22.

[3] Wong, K.Y., Casey, R.G. and Wahl, R.M. "Document Analysis System." *IBM Journal of Research and Development*, vol 26(6), 1982, pp. 647-656.

[4] Breuel, T.M. "The OCRopus Open Source OCR System." *Proc. Document Recognition and Retrieval*, vol. 6815, 2008, pg. 68150F. Code at `http://code.google.com/p/ocropus`

[5] Breuel, T.M. "A Practical, Globally Optimal Algorithm for Geometric Matching under Uncertainty." *Electronic Notes in Theoretical Computer Science*, vol. 46, 2001, pp. 1-15.

[6] Breuel, T.M. "Two Geometric Algorithms for Layout Analysis." *DAS*, Aug. 2002, pp. 188-199.

[7] Kise, K., Sato, A. and Iwata, M. "Segmentation of Page Images Using the Area Voronoi Diagram." *Computer Vision and Image Understanding*, vol. 70(3), 1998, pp. 370-382.

[8] Andersen, T. and Zhang, W. "Features for Neural Net Based Region Identification of Newspaper Docs." *Proc. ICDAR*, Scotland, Aug. 2003, pp 403-407.

[9] Alginahi, Y., Fekri, D. and Sid-Ahmed, M.A. "A Neural-Based Page Segmentation System." *Circuits, Systems and Comp.*, vol. 14(1), 2005, pp 109-122.

[10] Keysers, D., Shafait, F. and Breuel, T.M. "Document Image Zone Classification - a simple high-performance approach." *Computer Vision Theory and Applications*, Spain, Mar. 2007, pp. 4451.

[11] Lee, C.H. and Kanungo, T. "The Architecture of TrueViz: A GroundTRUth/Metadata Editing and VIsualiZing ToolKit." *Pattern Recognition*, vol. 36, 2003, pp. 811-825.

[12] Antonacopoulos, A., Gatos, B. and Bridson, D. "ICDAR2007 Page Segmentation Competition." *Proc. ICDAR*, Brazil, Sept. 2007, pp. 1279-1283.

[13] Phillips, I.T. and Chhabra, A.K. "Empirical Performance Eval. of Graphics Recognition Systems." *TPAMI*, vol. 21(9), 1999, pp. 849-870.