LOCATING TEXT IN COLOR DOCUMENT IMAGES

Erel Ortaçağ, Bülent Sankur
Department of Electrical and Electronic Engineering
Boğaziçi University, İstanbul, TURKEY
Khalid Sayood
Department of Electrical Engineering, University of Nebraska at Lincoln, USA

ABSTRACT

A novel text extraction algorithm from cluttered color document images is developed and tested. The algorithm consists of a color segmentation stage followed by rule-based filtering of non-text regions. Extraction of text segments algorithm uses the measurement of geometrical properties as well as characterness properties and a set of heuristic rules. The algorithm includes a fusion cycle of three different segmentation maps, and a restitution cycle to restore any deleted characters and/or their diacritical marks. The proposed method, proven successful in extraction of texts from many color document images, has applications in color image indexing and retrieval.

1 Introduction

Extraction of text parts of color document images is an important task in the automatic indexing and retrieval of archival source material [1][2]. Document segmentation algorithms tuned to black-and-white documents do not perform satisfactorily for color documents. These algorithms mostly assume a regular document structure and they proceed by exploiting horizontal and vertical projections and text column/block separations [3]. On the contrary the color documents such as brochures, CD covers, book covers do not have such a regular layout, but they can have text portions of widely different fonts, sizes and typefaces arbitrarily superimposed on a background with many image details. The proposed algorithm, as in Fig. 1, consists of two major steps:

- Color image segmentation and connected component analysis.
- Elimination of non-text segments based on the properties of text segments.

The novelty of the algorithm consists in the consideration of judiciously chosen geometric, topological and color features, and the fusion of these evidences in a rule based scheme. A second fusion step follows to merge decisions from more than one segmentation channel. The algorithm has also a restitution cycle that restores back characters or marks erroneously deleted due to their irregular shape or size as in the case of diacritical marks.

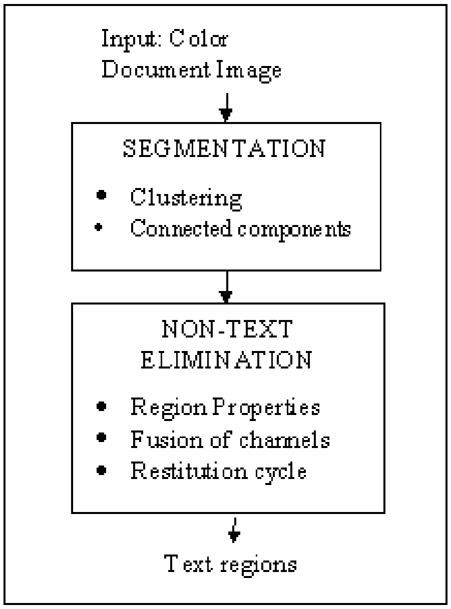

Figure 1: The main algorithm to locate text regions in complex color images

2 Color image segmentation

For text region extraction purposes, proper initial color segmentation is critically important in starting with a good estimate of text regions. In our algorithm, the initial segmentation is instrumented via vector quantization and/or octree quantization with given number of classes. Thus the document image with such a color quantization mapped into a much smaller set of colors which enables region analysis and eventual text segmentation [4].

We have designed VQ codebooks of low dimensions (4 to 8 classes) using a distance function that penalizes both the Euclidean distance to cluster centroids and the ratios of colors (i.e, R/G, R/B, G/B). In [5], it is shown that VQ codebooks designed using this criterion preserves image details better. A second scheme we have

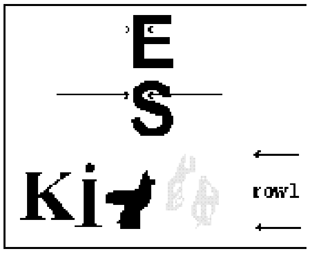

Figure 2: Illustrations of the stroke widths definition and of text row collinearity

concurrently used is the octree quantization which is actually a color palette design algorithm [6].

After the VQ or Octree clusters are established, the connected components in the resulting reduced color palette (e.g. VQ palette or Octree palette) are calculated. Each connected component forms a candidate text region.

3 Region features

Various region features are computed for discrimination of text segments from non-text segments. These features are:

- Geometrical features consist of the area, and the diameter (the larger of the horizontal and vertical dimensions of the bounding box).
- Character features, are made up of the stroke width (Fig. 2), and of the characterness score based on the Fourier descriptors. The stroke width of a character is defined as the minimal horizontal runlength of black pixels.
- Topological features, such as collinearity [7], i.e when commensurate segments of similar color have their bounding boxes overlapping by more than 50% (Fig. 2).

A characterness score invariant to translation, rotation and scaling can be calculated based on the Fourier descriptors of the closed contours. A total of 10 harmonics are used to represent the closed contours of segments. We have selected a Radial Basis Function (RBF) network with input the Fourier descriptors and with the output, the characterness score.

4 Locating text regions

Identification of text segments is based on filtering rules of non-text regions using their measured features. The scheme is made more robust by fusion of the results of different "channels". Channels are defined in this context as the segmentation maps obtained with different parameter settings. The heuristic rules demand that the candidate segments on a row have regularity in color, size, stroke width, character width and height.

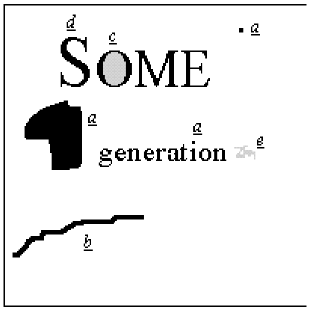

Figure 3: Illustration of first pass non-text elimination

First-pass: The regions are tested for their potential to be a text or a non-text region using their measured properties compared against threshold values. The following heuristic rules have been used in this stage:

1. Remove any region that is below a minimum (10) or above a maximum area size (800). (case a in Fig. 3)
2. Remove any region that exceeds a maximum diameter value. (case b in Fig. 3)
3. Any enclosed segment within a text candidate is removed if its color matches that of the surrounding removed area, e.g. the enclosure of "o". (case c in Fig. 3)
4. Restore back any region eliminated due to its size if it's collinear enough with other segments already declared on the same row (case d in Fig. 3)
5. Any region in a row is removed, if its color is distinctly different than of any other blob in the same text row. (case e in Fig. 3)

Second-pass: In the second stage, the regions surviving from the first stage are processed with a rating mechanism for their "characterness" properties. We assume that all text characters in a line or within a word should have similar areas, stroke widths, character widths, and character heights. The processing steps of this second pass are based on computing mutual similarity measures of segments coexisting on a row. Flow diagram of this process is given in Fig. 4.

Calculation of the similarity score: A character's mutual similarity score is a measure of its feature proximity to the other character candidates in a row. For each property it is calculated by summing its mutual similarity measures with all the other segments in the row. A fuzzy measure of similariy is used which enables one to spread the tolerance between 40 percent to 200 percent of each other 's measure. Thus for example character blobs on a row should not differ from each other in area or stroke width by the given percentages.

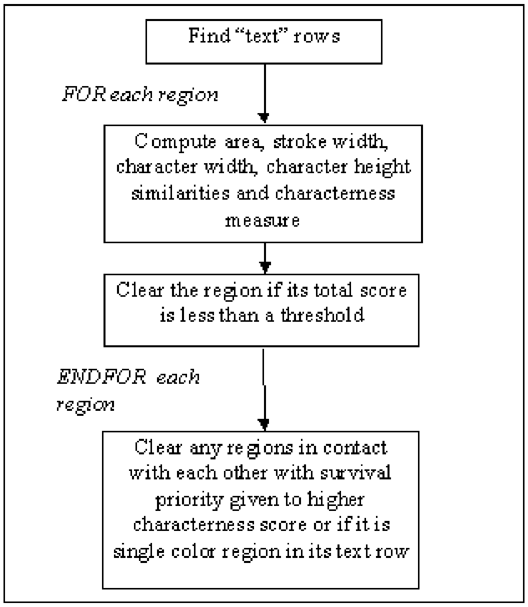
Find "text" rows
FOR each region
Compute area, stroke width, character width, character height similarities and characterness measure
Clear the region if its total score is less than a threshold
ENDFOR each region
Clear any regions in contact with each other with survival priority given to higher characterness score or if it is single color region in its text row

Figure 4: Flow diagram of the second-pass non-text elimination

The final score for a segment is the sum of scores:

$$C = \sum_{i=1}^{5} C_i \tag{1}$$

where C_1, C_2, C_3, C_4 are, respectively, the scores from area, stroke width, character width, character height similarities and C_5 is the neural network based characterness score. If the final score of a segment is above a threshold, then this region is declared to be a character segment.

4.1 Fusion cycle

In this stage, the text candidate regions from three separate segmentation maps are fused together using a variation of majority voting. Given the wide diversity of document types it has been found advantageous to obtain more than one segmentation map with different parameter settings of the clustering algorithm with the expectation that where one map fails by deleting a character, or conversely, by giving a false alarm, other maps may turn out to be more correct. We have found experimentally that the best text extraction performance is obtained by a threesome combination of the binary segmentation maps obtained by 1) VQ with 4 classes, 2) VQ with 8 classes, 3) Octree with 8 classes.

4.2 Restitution cycle

In this stage, a final check on deleted characters (near misses) and on diacritical marks is done, based on the characters extracted in the fusion stage.. Thus removed segments that might potentially correspond to erroneously deleted characters and marks have a second and new chance to be restored back on the basis their color similarity as well as row and column information. Several cases are illustrated in Figs. 5e and 5f.

5 Experimental results

A variety of 30 test images consisting of CD and book covers have been processed with the proposed algorithm. Two quantization algorithms with different parameter settings have been tested: The VQ algorithm with four different codebook sizes (4,8,16,32) and the Octree algorithm was tested with three different color sizes (8,16,32). Both the VQ and Octree algorithms performed significantly better with low numbers of color clusters.

In addition to subjective judgments, an objective measure for "goodness" of segmentation was used as follows : Q = 0.5 (Number of missed characters) + 0.5 (Number of false alarms). Out of 30 documents tested, the Q score was 0 for 25 of them, that is, perfect text region extraction was achieved; one sample was having the characters varying in color, the other three samples got false alarms after restitution cycle, and the non-text characters in the other samples could not be removed with the algorithm. Thus their Q scores were, respectively, (6, 2, 1, 0.5, 5.5). The various thresholds that need to be set in the algorithm have been determined experimentally:

- Maximum area, Amax = 800 pixels, or 2% of the image
- Minimum area, Amin = 10 pixels
- Maximum horizontal diameter, HD = image size / 3
- Minimum characterness score, C = 3.5

The first example is a book cover (Fig. 5a). The result of the first pass, i.e., elimination based on region properties is shown in Fig.5b for a VQ-4 channel only. The results of the second pass for the octree channel is shown in Fig 5c. The fusion stage eliminates (Fig.5d) all the remaining non-text blobs, while the restitution cycle restores back the erroneously removed characters or marks in Fig.5f. Among several restored parts notice the umlaut of u, the l,r,i,r characters in the forth row, etc. Fig. 6 illustrates another successful run, but with two false alarms.

6 Conclusion

An algorithm is proposed and tested to extract text from complex color consisting of a color segmentation stage followed by systematic elimination of non-text blobs. Experiments have indicated that combinations of small sized VQ and Octrees, i.e.,4 to 8 classes are adequate to initiate the segmentation. The initial segmentation map is refined by classifying text and non-text regions based on their geometrical properties (area, diameter), their topological property (string of regions on a row), and their similitude to a character accessed both by the analysis of the contour as well as the co-similarity in height, stroke width, color, and width of the string of candidates on a row.

The idea of concurrently running the above rule-based scheme on more than one segmentation map (called also channel) furnishes robustness to the algorithm. The restitution cycle at the end has also found to be beneficial in restoring back the erroneously deleted characters and their marks.

References

[1] Zhong, Y., K. Karu, A.K. Jain, "Locating Text in Complex Color Images", Pattern Recognition, vol28, no10, pp 1523-1535,1995

[2] Barber, R., M. Flickner, J. Hafner, W. Niblack, D.Petkovic, C.Faloutsos, W.Equitz, "Efficient and effective querying by image context", J. Intell. Inform. Systems 3, 231-262, 1994

[3] T. Pavlidis, J. Zhou, Page Segmentation and Classification, CVGIP, 54, 484-496, 1992.

[4] Xiang, Z., G. Joy, "Color Image Quantization by Agglomerative Clustering", IEEE Comp. Graph. and App., pp 44-48, May 1994

[5] Skarbek, W.,A. Koschan, "Color Image Segmentation - A Survey", Technical Report 94-32,Technical University of Berlin, 1994

[6] Clark, D., "Color Quantization using Octrees", Dr. Dobb's Journal, pp. 54-57, Jan 1996

[7] Feltcher, L.A., R. Kasturi, "A Robust Algorithm for Text String Separation from Mixed Text/Graphics Images", IEEE Trans. Pattern Analysis and Machine Intell., vol. 10, no. 6, pp 910-918, Nov. 1988

[8] Kuhl, F.P., C. R.Giardina, "Elliptic Fourier Features of a Closed Contour", Computer Graphics and Image Processing, vol. 18, pp 236-258, 1982

[9] Trier, O., A.K. Jain, "Feature Extraction Methods for Character Recognition - A Survey", Pattern Recognition, Vol. 29, no 4, pp.641-662, 1996

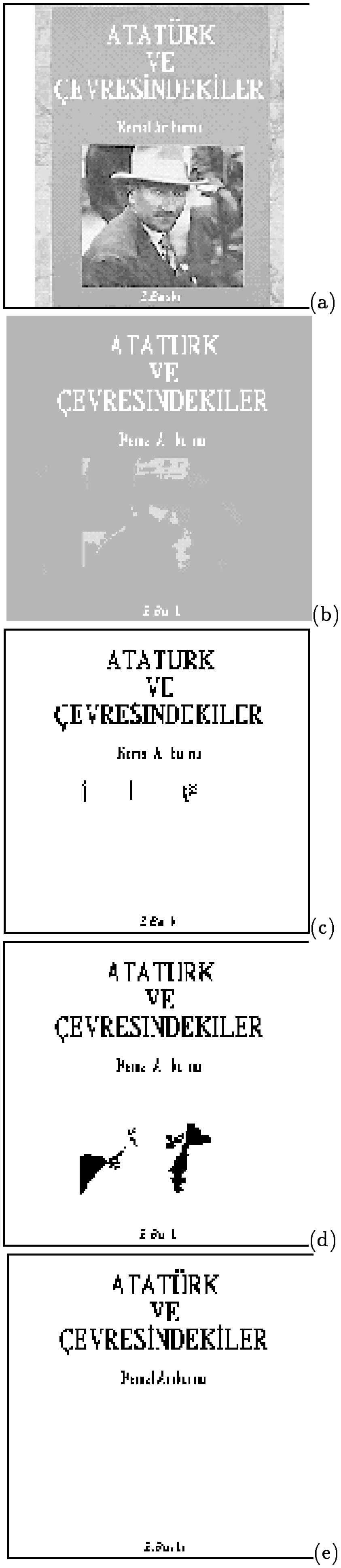

Figure 5. Text extraction of "Kemal Arıburnu - Atatürk ve Çevresindekiler" book cover; a) Original color image; b) First pass in the VQ-4 channel; c) Second pass in the octree channel; d) Fused image of the images from VQ-4, VQ-8 and Octree e) Resulting image after restitution cycle